

Engineering Math

May 30, 2023

CONTENTS

1	Introduction	13
2	Algebra and Notation	15
2.1	Sets And Set Notation	15
2.2	Well Ordering And Induction	16
2.3	The Complex Numbers	18
2.4	Polar Form Of Complex Numbers	21
2.5	Roots Of Complex Numbers	21
2.6	The Quadratic Formula	23
2.7	The Complex Exponential	24
2.8	Dividing Polynomials	25
2.9	The Fundamental Theorem Of Algebra	26
2.10	Exercises	28
3	Integrals, Functions of One Variable	31
3.1	Properties of the Integral	38
3.2	Uniform Convergence of Continuous Functions	42
3.3	Uniform Convergence And The Integral	44
4	Some Important Improper Integrals	45
4.1	Gamma Function	45
4.2	Laplace Transforms	47
I	Linear Algebra And Multivariable Calculus	51
5	Fundamentals	53
5.1	\mathbb{R}^n	53
5.2	Algebra in \mathbb{R}^n	55
5.3	Geometric Meaning Of Vector Addition In \mathbb{R}^3	56
5.4	Lines	57
5.5	Distance in \mathbb{R}^n	60
5.6	Geometric Meaning Of Scalar Multiplication In \mathbb{R}^3	63
5.7	Exercises	64
5.8	Physical Vectors	67
5.9	Exercises	71

6	Vector Products	75
6.1	The Dot Product	75
6.2	The Geometric Significance Of The Dot Product	78
6.2.1	The Angle Between Two Vectors	78
6.2.2	Work And Projections	79
6.2.3	The Dot Product And Distance In \mathbb{C}^n	82
6.3	Exercises	84
6.4	The Cross Product	85
6.4.1	The Box Product	89
6.5	Proof of the distributive law	90
6.5.1	Torque	91
6.5.2	Center Of Mass	92
6.5.3	Angular Velocity	93
6.6	Vector Identities And Notation	94
6.7	Planes	97
6.8	Exercises	100
7	Systems Of Equations	105
7.1	Systems Of Equations, Algebraic Procedures	105
7.1.1	Elementary Operations	105
7.1.2	Gauss Elimination	107
7.1.3	Balancing Chemical Reactions	117
7.1.4	Dimensionless Variables*	119
7.2	MATLAB And Row Reduced Echelon Form	122
7.3	Exercises	123
8	Matrices	129
8.1	Addition And Scalar Multiplication Of Matrices	129
8.2	Multiplication of Matrices	132
8.3	Linear Transformations and Matrices	135
8.4	Multiplication of Matrices	136
8.4.1	The Transpose	138
8.5	Some Examples of Linear Functions on \mathbb{R}^n	139
8.5.1	Rotations in \mathbb{R}^2	139
8.5.2	Projections	141
8.5.3	Rotations About A Particular Vector	142
8.6	The Inverse of a Matrix	144
8.6.1	The Identity And Inverses	144
8.6.2	Finding The Inverse Of A Matrix	145
8.7	MATLAB And Matrix Arithmetic	151
8.8	Exercises	152
9	Subspaces Spans and Bases	159
9.1	Subspaces	160
9.2	Exercises	165

CONTENTS

5

10 Eigenvalues and Eigenvectors	173
10.1 Definition of Eigenvalues	173
10.2 An Introduction to Determinants	174
10.2.1 Cofactors and 2×2 Determinants	174
10.2.2 The Determinant of a Triangular Matrix	177
10.2.3 Properties of Determinants	178
10.2.4 Finding Determinants Using Row Operations	179
10.3 Applications	181
10.3.1 A Formula For The Inverse	181
10.3.2 Finding Eigenvalues Using Determinants	183
11 Matrices and The Inner Product	185
11.1 Eigenvalues and Eigenvectors	186
11.2 Using Matlab	190
11.3 Distance and Unitary Matrices	190
11.4 Schur's Theorem	191
11.5 Diagonalization	196
11.6 Approximations	198
11.6.1 Fredholm Alternative	198
11.6.2 Least Squares	200
11.6.3 Regression lines	202
11.6.4 Identifying the Closest Point	204
11.6.5 Using MATLAB	207
11.7 The Singular Value Decomposition*	207
11.8 Exercises	209
12 Vector Valued Functions	217
12.1 Vector Valued Functions	217
12.2 Vector Fields	218
12.3 Exercises	220
12.4 Continuous Functions	221
12.4.1 Sufficient Conditions For Continuity	222
12.5 Limits Of A Function	223
12.6 Properties Of Continuous Functions	226
12.7 Exercises	227
12.8 Open And Closed Sets	229
12.9 Exercises	233
13 Some Fundamentals*	235
13.1 Combinations Of Continuous Functions	235
13.2 The Nested Interval Lemma	238
13.3 Convergent Sequences, Sequential Compactness	239
13.4 Continuity And The Limit Of A Sequence	241
13.5 The Extreme Value Theorem And Uniform Continuity	242
13.6 Convergence of Functions	243
13.7 Root Test	244
13.8 Convergence of Sums	245
13.9 Connected Sets	247

13.10 Exercises	251
14 Vector Valued Functions Of One Variable	253
14.1 Limits Of A Vector Valued Function Of One Real Variable	253
14.2 The Derivative And Integral	254
14.2.1 Geometric And Physical Significance Of The Derivative	256
14.2.2 Differentiation Rules	257
14.2.3 Leibniz's Notation	260
14.3 Exercises	260
14.4 Line Integrals	261
14.4.1 Arc Length And Orientations	262
14.4.2 Line Integrals And Work	265
14.4.3 Another Notation For Line Integrals	268
14.5 Exercises	268
14.6 Independence Of Parametrization*	270
14.6.1 Hard Calculus	270
14.6.2 Independence Of Parametrization	272
15 Motion On A Space Curve	275
15.1 Space Curves	275
15.1.1 Some Simple Techniques	278
15.2 Geometry Of Space Curves*	279
15.3 Exercises	282
16 Functions Of Many Variables	285
16.1 Review Of Limits	285
16.2 Exercises	287
16.3 The Directional Derivative And Partial Derivatives	288
16.3.1 The Directional Derivative	288
16.3.2 Partial Derivatives	289
16.4 Exercises	292
16.5 Mixed Partial Derivatives	293
16.6 Partial Differential Equations	294
16.7 Exercises	295
17 The Derivative Of A Function Of Many Variables	297
17.1 The Derivative Of Functions Of One Variable	297
17.2 Exercises	299
17.3 The Derivative Of Functions Of Many Variables	300
17.4 Exercises	305
17.5 C^1 Functions	306
17.6 The Chain Rule	310
17.6.1 The Chain Rule For Functions Of One Variable	310
17.6.2 The Chain Rule For Functions Of Many Variables	310
17.7 Exercises	315
17.7.1 Related Rates Problems	316
17.7.2 The Derivative Of The Inverse Function	318
17.7.3 Proof Of The Chain Rule	319
17.8 Exercises	320

CONTENTS

7

17.9 The Gradient	322
17.10 The Gradient And Tangent Planes	324
17.11 Exercises	326
18 Optimization	329
18.1 Local Extrema	329
18.2 Exercises	331
18.3 The Second Derivative Test	332
18.4 Exercises	335
18.5 Lagrange Multipliers	337
18.6 Exercises	342
18.7 Proof Of The Second Derivative Test*	345
19 The Riemannn Integral On \mathbb{R}^n	349
19.1 Methods For Double Integrals	349
19.1.1 Density And Mass	353
19.2 Exercises	354
19.3 Methods For Triple Integrals	355
19.3.1 Definition Of The Integral	355
19.3.2 Iterated Integrals	356
19.4 Exercises	359
19.4.1 Mass And Density	360
19.5 Exercises	362
20 The Integral In Other Coordinates	365
20.1 Polar Coordinates	365
20.2 Exercises	367
20.3 Cylindrical And Spherical Coordinates	368
20.3.1 Volume and Integrals in Cylindrical Coordinates	369
20.3.2 Volume And Integrals in Spherical Coordinates	371
20.4 Exercises	377
20.5 The General Procedure	379
20.6 Exercises	382
20.7 The Moment Of Inertia And Center Of Mass	384
20.8 Exercises	385
21 The Integral on Two Dimensional Surfaces In \mathbb{R}^3	389
21.1 The Two Dimensional Area In \mathbb{R}^3	389
21.2 Surfaces Of The Form $z = f(x, y)$	393
21.3 MATLAB and Graphing Surfaces	394
21.4 Piecewise Defined Surfaces	395
21.5 Flux Integrals	395
21.6 Exercises	396
22 Calculus Of Vector Fields	399
22.1 Divergence And Curl Of A Vector Field	399
22.1.1 Vector Identities	400
22.1.2 Vector Potentials	401
22.1.3 The Weak Maximum Principle	402

22.2	Exercises	403
22.3	The Divergence Theorem	404
22.3.1	Coordinate Free Concept Of Divergence	408
22.4	Some Applications Of The Divergence Theorem	409
22.4.1	Hydrostatic Pressure	409
22.4.2	Archimedes Law Of Buoyancy	410
22.4.3	Equations Of Heat And Diffusion	410
22.4.4	Balance Of Mass	412
22.4.5	Balance Of Momentum	412
22.4.6	The Reynolds Transport Formula	418
22.4.7	Frame Indifference	420
22.4.8	Bernoulli's Principle	422
22.4.9	The Wave Equation	423
22.4.10	A Negative Observation	423
22.4.11	Volumes Of Balls In \mathbb{R}^n	423
22.4.12	Electrostatics	425
22.5	Exercises	426
23	Stokes And Green's Theorems	429
23.1	Green's Theorem	429
23.2	Exercises	434
23.3	Stoke's Theorem From Green's Theorem	435
23.3.1	The Normal and the Orientation	438
23.3.2	The Mobeus Band	440
23.4	A General Green's Theorem	441
23.4.1	Conservative Vector Fields	442
23.4.2	Some Terminology	446
24	Moving Coordinate Systems	447
24.1	The Acceleration In Polar Coordinates	447
24.2	Planetary Motion	449
24.2.1	The Equal Area Rule, Kepler's Second Law	450
24.2.2	Inverse Square Law, Kepler's First Law	450
24.2.3	Kepler's Third Law	453
24.3	The Angular Velocity Vector	454
24.4	Angular Velocity Vector on Earth	455
24.5	Coriolis Force and Centripetal Force	457
24.6	Coriolis Force on the Rotating Earth	458
24.7	The Foucault Pendulum*	460
24.8	Exercises	462
25	Curvilinear Coordinates	465
25.1	Basis Vectors	465
25.2	Exercises	468
25.3	Curvilinear Coordinates	469
25.4	Exercises	472
25.5	Transformation of Coordinates.	474
25.6	Differentiation and Christoffel Symbols	475

CONTENTS

9

25.7	Gradients and Divergence	478
25.8	Exercises	481
25.9	Curl and Cross Products	482
26	Implicit Function Theorem*	485
26.1	More Continuous Partial Derivatives	489
26.2	The Method Of Lagrange Multipliers	490
26.3	The Local Structure Of C^1 Mappings*	492
II	Differential Equations	495
27	Determinants	497
27.1	Basic Techniques And Properties	497
27.1.1	Cofactors And 2×2 Determinants	497
27.1.2	The Determinant Of A Triangular Matrix	501
27.1.3	Properties Of Determinants	502
27.1.4	Finding Determinants Using Row Operations	503
27.2	Applications	505
27.2.1	A Formula For The Inverse	505
27.2.2	Finding Eigenvalues Using Determinants	509
27.2.3	Cramer's Rule	511
27.3	MATLAB And Determinants	513
27.4	The Cayley Hamilton Theorem*	513
27.5	Exercises	516
28	The Mathematical Theory Of Determinants*	525
28.0.1	The Function sgn	525
28.1	The Determinant	527
28.1.1	The Definition	527
28.1.2	Permuting Rows Or Columns	528
28.1.3	A Symmetric Definition	529
28.1.4	The Alternating Property Of The Determinant	529
28.1.5	Linear Combinations And Determinants	530
28.1.6	The Determinant Of A Product	531
28.1.7	Cofactor Expansions	531
28.1.8	Formula For The Inverse	533
28.1.9	Cramer's Rule	535
29	First Order Scalar ODE	537
29.1	First Order Linear Equations	537
29.2	Bernouli Equations	545
29.3	Separable Differential Equations, Stability	546
29.4	Homogeneous Equations	554
29.5	Exact Equations	555
29.6	The Integrating Factor	557
29.7	The Case Where M, N Are Affine Linear	562
29.8	Linear and Nonlinear Differential Equations	564
29.9	Computer Algebra Methods	567

29.9.1	MATLAB	567
29.10	Exercises	569
30	Laplace Transform Methods	581
30.1	Linear O.D.E. With Constant Coefficients	581
30.2	First Order Systems, Constant Coefficients	586
30.2.1	Some Technical Considerations*	587
30.2.2	Solving a First Order System	589
30.2.3	Using a Computer Algebra System	591
30.3	Homogeneous Particular and General Solutions	593
30.4	Higher Order Scalar Linear Equations	597
31	Numerical Solutions For Systems	601
31.1	A Few Numerical Methods	601
31.2	Using MATLAB to Find Solutions	603
31.3	Stability of Equilibrium Points	604
31.4	Periodic Orbits, Poincare Bendixon Theorem	608
31.5	Exercises	609
32	Solutions Near a Regular Singular Point	611
32.1	The Euler Equations	611
32.2	Some Simple Observations on Power Series	615
32.3	Regular Singular Points	615
32.4	Abel's Formula	618
32.5	Finding the Solution	619
32.6	The Bessel Equations	626
32.6.1	The Case where $\nu = 0$	626
32.6.2	The Case of ν Not an Integer	627
32.6.3	Case Where ν is an Integer	628
32.7	Other Properties of Bessel Functions	630
32.8	Exercises	633
33	Boundary Value Problems, Fourier Series	637
33.1	Boundary Value Problems	637
33.2	Eigenvalue Problems	638
33.3	Fourier Series	640
33.4	Mean Square Approximation	643
33.5	Pointwise Convergence of Fourier Series	647
33.5.1	Explanation of Pointwise Convergence Theorem	648
33.5.2	Mean Square Convergence	652
33.6	Integrating and Differentiating Fourier Series	654
33.7	Odd and Even Extensions	658
33.8	Exercises	659
34	Some Partial Differential Equations	669
34.1	Laplacian in Orthogonal Curvilinear Coordinates	669
34.2	Heat and Wave Equations	670
34.2.1	Heat Equation	670

CONTENTS

11

34.2.2 The Wave Equation	675
34.3 Nonhomogeneous Problems	678
34.4 Laplace Equation	683
34.4.1 Rectangles	683
34.4.2 Circular Disks	686
34.5 Exercises	689

III Fundamentals of Complex Analysis 695

35 Analytic Functions 697

35.1 Cauchy Riemann Equations	697
35.2 The Cauchy Riemann Equations	698
35.3 Contour Integrals	700
35.4 Cauchy Integral Theorem	705
35.5 Primitives and Cauchy Goursat Theorem	706
35.6 Functions Differentiable on a Disk, Zeros	709
35.7 Liouville's Theorem	717
35.8 Riemann Sphere	718
35.9 Exercises	719

36 Isolated Singularities and Analytic Functions 727

36.1 Open Mapping Theorem for Complex Valued Functions	727
36.2 Functions Analytic on an Annulus	730
36.3 Isolated Singularities	734
36.4 Meromorphic Functions	736
36.5 The Residue Theorem	737
36.6 Evaluation of Improper Integrals	738
36.7 Exercises	747

37 Some Fundamental Functions and Transforms 751

37.1 Gamma Function	751
37.2 Laplace Transform	752
37.3 Fourier Transform	754
37.4 The Inversion of Laplace Transforms	758
37.5 The Bromwich Integral	759
37.6 Exercises	764

IV Probability and Statistics 767

38 Probability 769

38.1 Improper Integrals	769
38.2 Combinations	771
38.3 The Binomial Theorem	772
38.4 Exercises	774
38.5 Counting and Basic Probability	774
38.6 Exercises	777
38.7 General Considerations Probability	779

38.8	Moment Generating Functions	787
38.9	Independence and Conditional Probability	791
39	Statistical Tests	797
39.1	The Distribution of nS^2/σ^2	798
39.1.1	Confidence Intervals for Variance	802
39.2	The T and F Distributions	805
39.2.1	The T Distribution	805
39.2.2	Confidence Intervals for the Mean	807
39.2.3	Testing For Two Different Means	810
39.2.4	The F Distribution	812
39.2.5	Confidence Intervals for the Ratio of Two Variances	814
39.3	Maximum Likelihood Estimates	816
39.4	Quadratic Forms	818
39.5	Linear Regression	825
39.6	Goodness of Fit	832
39.7	Contingency Tables	839
A	The Theory Of The Riemannn Integral*	843
A.1	An Important Warning	843
A.2	Basic Definition	843
A.3	Basic Properties	846
A.4	Which Functions Are Integrable?	849
A.5	Iterated Integrals	857
A.6	The Change Of Variables Formula	861
A.7	Some Observations	869
B	A Rigid Body Rotating About a Point	871
C	Lagrangian Mechanics	877
C.1	The Spinning Top and the Euler Angles	880

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.

Chapter 1

Introduction

This book is on multivariable calculus. It is to follow a calculus course which is devoted primarily to calculus of functions of one variable. It is not an advanced calculus course but is intended to serve as an elementary presentation of calculus of many variables. As part of this presentation, there is a short review of topics which are often omitted from single variable calculus courses. There is also a short treatment of linear algebra because multivariable calculus is dependent on linear algebra. For example, the derivative is a **linear transformation**. The determinant is used in change of variables formulas. You really don't understand Lagrange multipliers without some linear algebra concepts, the second derivative test is most easily remembered in terms of eigenvalues and one could go on like this. Indeed, multivariable calculus is really all about using linear algebra concepts to approximate nonlinear analysis ideas so if you don't have any concepts from linear algebra understood, then multivariable calculus can seem a little mysterious. The first part of the book, consisting of multivariable calculus and linear algebra will fit in one semester. The second part will fit in a second semester. I have left out the fluff which usually clogs our differential equations classes and replaced it with MATLAB. I taught differential equations way too often to pretend that it has anything sufficiently significant to justify the time spent on it. I think that one of the major difficulties people have with this subject is more about factoring polynomials than anything of mathematical significance.

While the book does contain a fairly complete presentation of linear algebra, it is not necessary to read all of this in order to do the multivariable calculus portion of the book. It suffices to consider that which includes matrices and linear transformations and eigenvalues. The first part of my book on calculus of one and many variables has the relevant material on \mathbb{R} so if one has read it, there will be no need to re read what is repeated in this book. Those interested in linear algebra would do better to read a linear algebra book like those on my web page.

I have called the book Engineering math because it is not limited to multivariable calculus and linear algebra. It also has the necessary material on differential equations and a part devoted to basic complex analysis. It concludes with a short introduction to probability and statistics. I have tried to emphasize those aspects of complex analysis which are in my opinion of most use. It is an elementary treatment of this subject, not the type of thing in a graduate text.

The material on probability and statistics is intended to explain some of the difficult topics in statistics. This is a subject which has been automated to a remarkable extent

and doing the applications amounts to using the right software at this point. I think that understanding why certain things are true needs to be presented and this is the emphasis in this book. However, really difficult mathematical issues are not included, especially those things which really need Lebesgue integration and measure theory to understand. Also, I am emphasizing moment generating functions rather than characteristic functions. I am emphasizing confidence intervals more than hypothesis testing. It seems to me that this is easier to understand with less jargon and is sufficient to draw conclusions. It is not a complete book on mathematical statistics, just an introduction to some of the important ideas. Not everything is proved in this section because some of the proofs are too long. However, I am making every effort to at least make it plausible.

Chapter 2

Algebra and Notation

The reader should be familiar with most of the topics in this chapter. However, it is often the case that set notation is not familiar and so a short discussion of this is included first. Complex numbers are then considered in somewhat more detail. Many of the applications of linear algebra and differential equations require the use of complex numbers, so this is the reason for this introduction.

2.1 Sets And 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, $\{1, 2, 3, 8\} \subseteq \{1, 2, 3, 4, 5, 8\}$. It is sometimes said that “ A is contained in B ” or even “ B contains A ”. 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 an element of 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. 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 $[a, b]$ where a and b are real numbers, denotes the set of real numbers x , such that $a \leq x \leq b$ and $[a, b)$ denotes the set of real numbers such that $a \leq x < b$. (a, b)

consists of the set of real numbers x such that $a < x < b$ and $(a, b]$ indicates the set of numbers x such that $a < x \leq b$. $[a, \infty)$ means the set of all numbers x such that $x \geq a$ and $(-\infty, a]$ means the set of all real numbers which are less than or equal to a . These sorts of sets of real numbers are called intervals. The two points a and b are called endpoints of the interval. Other intervals such as $(-\infty, b)$ are defined by analogy to what was just explained. In general, the curved parenthesis indicates the end point it sits next to is not included while the square parenthesis indicates this end point is included. The reason that there will always be a curved parenthesis next to ∞ or $-\infty$ is that these are not real numbers. Therefore, they cannot be included in any set of real numbers.

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

$$A \setminus B \equiv \{x \in A : x \notin B\}.$$

Set notation is used whenever convenient.

To illustrate the use of this notation relative to intervals consider three examples of inequalities. Their solutions will be written in the notation just described.

Example 2.1.1 Solve the inequality $2x + 4 \leq x - 8$

$x \leq -12$ is the answer. This is written in terms of an interval as $(-\infty, -12]$.

Example 2.1.2 Solve the inequality $(x + 1)(2x - 3) \geq 0$.

The solution is $x \leq -1$ or $x \geq \frac{3}{2}$. In terms of set notation this is denoted by $(-\infty, -1] \cup [\frac{3}{2}, \infty)$.

Example 2.1.3 Solve the inequality $x(x + 2) \geq -4$.

This is true for any value of x . It is written as \mathbb{R} or $(-\infty, \infty)$.

2.2 Well Ordering And Induction

Mathematical induction and well ordering are two extremely important principles in math. They are often used to prove significant things which would be hard to prove otherwise.

Definition 2.2.1 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 2.2.2 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. The symbol \mathbb{Z} denotes the set of all integers. Note that if a is an integer, then there are no integers between a and $a + 1$.

Theorem 2.2.3 (*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 consist 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, $a \notin T$. Thus $b \geq a + 1$, and so $b - 1 \geq a$ 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, $b - 1 \in T$ 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 every integer at least as large as a is also in S . ■

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

Example 2.2.4 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 2.2.5 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

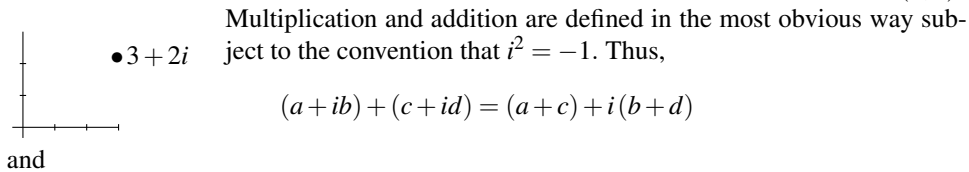
$$\left(\frac{1}{\sqrt{2n+3}} \right)^2 = \frac{1}{2n+3} > \frac{2n+1}{(2n+2)^2}$$

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 $m \geq a$.

2.3 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



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

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

$$\begin{aligned} (a+ib)(c+id) &= ac + iad + ibc + i^2bd \\ &= (ac-bd) + i(bc+ad). \end{aligned}$$

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 2.3.1 *The complex numbers with multiplication and addition defined as above form a field satisfying all the field axioms. These are the following list of 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 complex numbers or 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.

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

Observation 2.3.2 *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 2.3.3 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 $z = a + ib$,

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

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 2.3.4 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 2.3.5 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 2.3.6 : 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.

2.4 Polar Form Of Complex Numbers

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

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

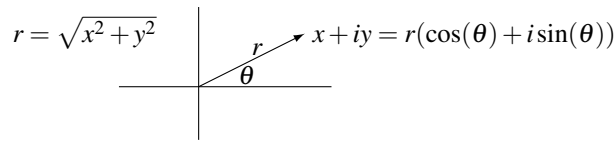
and so

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

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 $r = \sqrt{x^2 + y^2} \equiv |z|$.



2.5 Roots Of Complex Numbers

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

Theorem 2.5.1 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 2.5.2 *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 2.5.3 *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 2.5.4 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. You should show this is the case. To see how to do this, see Problems 17 and 18 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.

2.6 The Quadratic Formula

The quadratic formula

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

gives the solutions x to

$$ax^2 + bx + c = 0$$

where a, b, c are real numbers. It holds even if $b^2 - 4ac < 0$. This is easy to show from the above. There are exactly two square roots to this number $b^2 - 4ac$ from the above methods using De Moivre's theorem. These roots are of the form

$$\sqrt{4ac - b^2} \left(\cos\left(\frac{\pi}{2}\right) + i \sin\left(\frac{\pi}{2}\right) \right) = i\sqrt{4ac - b^2}$$

and

$$\sqrt{4ac - b^2} \left(\cos\left(\frac{3\pi}{2}\right) + i \sin\left(\frac{3\pi}{2}\right) \right) = -i\sqrt{4ac - b^2}$$

Thus the solutions, according to the quadratic formula are still given correctly by the above formula.

Do these solutions predicted by the quadratic formula continue to solve the quadratic equation? Yes, they do. You only need to observe that when you square a square root of a complex number z , you recover z . Thus

$$\begin{aligned} & a \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a} \right)^2 + b \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a} \right) + c \\ &= a \left(\frac{1}{2a^2} b^2 - \frac{1}{a} c - \frac{1}{2a^2} b \sqrt{b^2 - 4ac} \right) + b \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a} \right) + c \\ &= \left(-\frac{1}{2a} (b \sqrt{b^2 - 4ac} + 2ac - b^2) \right) + \frac{1}{2a} (b \sqrt{b^2 - 4ac} - b^2) + c = 0 \end{aligned}$$

Similar reasoning shows directly that $\frac{-b - \sqrt{b^2 - 4ac}}{2a}$ also solves the quadratic equation.

What if the coefficients of the quadratic equation are actually complex numbers? Does the formula hold even in this case? The answer is yes. This is a hint on how to do Problem 27 below, a special case of the fundamental theorem of algebra, and an ingredient in the proof of some versions of this theorem.

Example 2.6.1 Find the solutions to $x^2 - 2ix - 5 = 0$.

Formally, from the quadratic formula, these solutions are

$$x = \frac{2i \pm \sqrt{-4 + 20}}{2} = \frac{2i \pm 4}{2} = i \pm 2.$$

Now you can check that these really do solve the equation. In general, this will be the case. See Problem 27 below.

2.7 The Complex Exponential

It was shown above that every complex number can be written in the form

$$r(\cos \theta + i \sin \theta)$$

where $r \geq 0$. 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 $e^{\alpha + i\beta}$. Having done so, does it follow that the expression preserves the most important property of the function $t \rightarrow e^{(\alpha + i\beta)t}$ for t real, that

$$\left(e^{(\alpha + i\beta)t} \right)' = (\alpha + i\beta) e^{(\alpha + i\beta)t}?$$

By the definition just given which does not contradict the usual definition in case $\beta = 0$ and the usual rules of differentiation in calculus,

$$\begin{aligned} \left(e^{(\alpha + i\beta)t} \right)' &= \left(e^{\alpha t} (\cos(\beta t) + i \sin(\beta t)) \right)' \\ &= e^{\alpha t} [\alpha (\cos(\beta t) + i \sin(\beta t)) + (-\beta \sin(\beta t) + i\beta \cos(\beta t))] \end{aligned}$$

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.

2.8 Dividing Polynomials

It will be very important to be able to work with polynomials in certain parts of linear algebra to be presented later. It is surprising how useful this junior high material will be.

Definition 2.8.1 A polynomial is an expression of the form $a_n \lambda^n + a_{n-1} \lambda^{n-1} + \cdots + a_1 \lambda + a_0$, $a_n \neq 0$ 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 $a_n = 1$.

Note that the degree of the zero polynomial is not defined in the above. The following is called the division algorithm.

Lemma 2.8.2 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

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

and the degree of $r(\lambda)$ is $m \geq n$ where n is the degree of $g(\lambda)$. Say the leading term of $r(\lambda)$ is $b\lambda^m$ while the leading term of $g(\lambda)$ is $\hat{b}\lambda^n$. Then letting $a = b/\hat{b}$, $a\lambda^{m-n}g(\lambda)$ has the same leading term as $r(\lambda)$. Thus the degree of $r_1(\lambda) \equiv r(\lambda) - a\lambda^{m-n}g(\lambda)$ is no more than $m-1$. Then

$$r_1(\lambda) = f(\lambda) - (q(\lambda)g(\lambda) + a\lambda^{m-n}g(\lambda)) = f(\lambda) - \left(\overbrace{q(\lambda) + a\lambda^{m-n}}^{q_1(\lambda)} \right) g(\lambda)$$

Denote by S the set of polynomials $f(\lambda) - g(\lambda)l(\lambda)$. Out of all these polynomials, there exists one which has smallest degree $r(\lambda)$. Let this take place when $l(\lambda) = q(\lambda)$. Then by the above argument, the degree of $r(\lambda)$ is less than the degree of $g(\lambda)$. Otherwise, there is one which has smaller degree. Thus $f(\lambda) = g(\lambda)q(\lambda) + r(\lambda)$.

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, matching coefficients implies that $\hat{q}(\lambda) = q(\lambda)$. ■

2.9 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. The proof given here follows Rudin [31]. See also Hardy [20] for a similar proof, more discussion and references. The shortest proof is found in the theory of complex analysis. First I will give an informal explanation of this theorem which shows why it is reasonable to believe in the fundamental theorem of algebra.

Theorem 2.9.1 *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 $p(w) = 0$.*

To begin with, 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

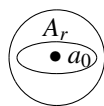
$$p(z) = z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

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,

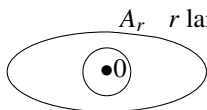
$$z^n = |z|^n (\cos(n\theta) + i \sin(n\theta))$$

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

$$(r(\cos t + i \sin t))^3 + r(\cos t + i \sin t) + 1 + i = 0 + 0i$$

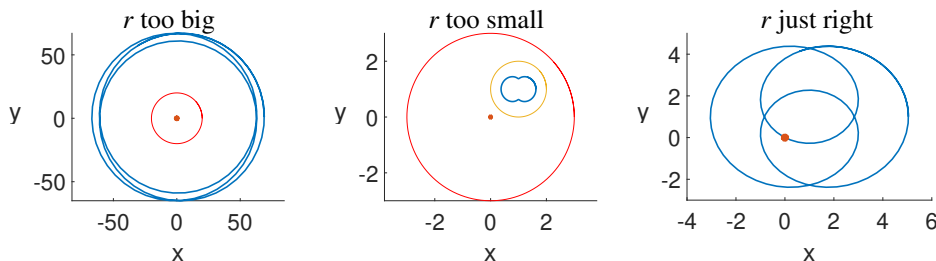
Expanding this expression on the left to write it in terms of real and imaginary parts, you get on the left

$$r^3 \cos^3 t - 3r^3 \cos t \sin^2 t + r \cos t + 1 + i(3r^3 \cos^2 t \sin t - r^3 \sin^3 t + r \sin t + 1)$$

Thus you need to have both the real and imaginary parts equal to 0. In other words, you need to have

$$(r^3 \cos^3 t - 3r^3 \cos t \sin^2 t + r \cos t + 1, 3r^3 \cos^2 t \sin t - r^3 \sin^3 t + r \sin t + 1) = (0, 0)$$

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 $r = 4$. 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 $p(z) = 0$.

Now here is a rigorous proof for those who have studied analysis.

Proof. Suppose the nonconstant polynomial $p(z) = a_0 + a_1 z + \cdots + a_n z^n$, $a_n \neq 0$, has no zero in \mathbb{C} . Since $\lim_{|z| \rightarrow \infty} |p(z)| = \infty$, there is a z_0 with

$$|p(z_0)| = \min_{z \in \mathbb{C}} |p(z)| > 0$$

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 $q(z) = 1 + a_k z^k + r(z)$ where $r(z)$ consists of higher order terms. Here a_k is the first coefficient which is nonzero. Choose a sequence, $z_n \rightarrow 0$, such that $a_k z_n^k < 0$. For example, let $-a_k z_n^k = (1/n)$. Then

$$|q(z_n)| = |1 + a_k z_n^k + r(z_n)| \leq 1 - 1/n + |r(z_n)| = 1 + a_k z_n^k + |r(z_n)| < 1$$

for all n large enough because $|r(z_n)|$ is small compared with $|a_k z_n^k|$ since $|r(z_n)|$ involves only higher order terms and $a_k z_n^k < 0$. This is a contradiction. ■

2.10 Exercises

1. Prove by induction that $\sum_{k=1}^n k^3 = \frac{1}{4}n^4 + \frac{1}{2}n^3 + \frac{1}{4}n^2$.
2. Prove by induction that whenever $n \geq 2$, $\sum_{k=1}^n \frac{1}{\sqrt{k}} > \sqrt{n}$.
3. Prove by induction that $1 + \sum_{i=1}^n i(i!) = (n+1)!$.
4. The binomial theorem states $(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k$ where

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \text{ if } k \in [1, n], \quad \binom{n}{0} \equiv 1 \equiv \binom{n}{n}$$

Prove the binomial theorem by induction. Next show that

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}, \quad 0! \equiv 1$$



5. Let $z = 5 + i9$. Find z^{-1} .
6. Let $z = 2 + i7$ and let $w = 3 - i8$. Find $zw, z+w, z^2$, and w/z .
7. Give the complete solution to $x^4 + 16 = 0$.
8. Graph the complex cube roots of 8 in the complex plane. Do the same for the four fourth roots of 16. ►
9. If z is a complex number, show there exists ω a complex number with $|\omega| = 1$ and $\omega z = |z|$.
10. De Moivre's theorem says $[r(\cos t + i \sin t)]^n = r^n (\cos nt + i \sin nt)$ for n a positive integer. Does this formula continue to hold for all integers n , even negative integers? Explain. ►
11. You already know formulas for $\cos(x+y)$ and $\sin(x+y)$ and these were used to prove De Moivre's theorem. Now using De Moivre's theorem, derive a formula for $\sin(5x)$ and one for $\cos(5x)$. ►
12. 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 , $r = |z|$.
13. Factor $x^3 + 8$ as a product of linear factors.
14. 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.
15. Completely factor $x^4 + 16$ as a product of linear factors.
16. Factor $x^4 + 16$ as the product of two quadratic polynomials each of which cannot be factored further without using complex numbers.

17. 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.
18. 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.
19. 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.

20. 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?

21. De Moivre's theorem 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 $1 = -1$ as in the previous problem. 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?

22. Review Problem 10 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.
23. 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 $a_{\alpha\beta} \in \mathbb{C}$. 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.
24. 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

$$p(x) = a_n (x-z_1)(x-z_2)\cdots(x-z_n).$$

25. Give the solutions to the following quadratic equations having real coefficients.

(a) $x^2 - 2x + 2 = 0$

(b) $3x^2 + x + 3 = 0$

(c) $x^2 - 6x + 13 = 0$

(d) $x^2 + 4x + 9 = 0$

(e) $4x^2 + 4x + 5 = 0$

26. 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) $x^2 + 2x + 1 + i = 0$

(b) $4x^2 + 4ix - 5 = 0$

(c) $4x^2 + (4 + 4i)x + 1 + 2i = 0$

(d) $x^2 - 4ix - 5 = 0$

(e) $3x^2 + (1 - i)x + 3i = 0$

27. 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.

Chapter 3

Integrals, Functions of One Variable

One cannot very well study integrals of functions of many variables without some knowledge of integrals of one variable.

I assume the reader is familiar with the usual techniques for finding antiderivatives and integrals such as partial fractions, integration by parts and integration by substitution. These topics are usually done very well in beginning calculus courses so I am not giving lots of exercises and examples related to formal symbol pushing techniques. However, the typical calculus book does not even give a complete explanation of why the integral of a continuous function exists.

I do not wish this book to be based on the kind of thing encountered in religion where we are asked to choose to believe without any good reason for doing so or even a very coherent description of what we are to believe. I do not wish to disparage religion since I am a religious man myself who chooses to believe many things with no solid evidence, even in the presence of obvious contradictions and patent absurdities, but math should not be this way. Nor should it in any way resemble magic. This is why I am attempting to give rational explanations. Sometimes these may fall flat, but at least I am giving it a try which is more than can be said of the typical undergraduate presentation of courses related to calculus. If you don't even understand why the integral exists, then what is the meaning of everything dependent on the integral? These things become nothing more than meaningless ritual and speculation (religion). This short chapter is on the fundamental questions related to the integral which are usually not discussed in undergraduate calculus. For a more complete treatment of Riemann integration which includes what is here, see my book on calculus of functions of one and many variable or the single variable advanced calculus book for a lot more.

The fundamental issues depend not on techniques of integration or some stupid geometric reasoning but on completeness of \mathbb{R} .

Definition 3.0.1 *One of the equivalent definitions of completeness of \mathbb{R} is that if S is any nonempty subset of \mathbb{R} which is bounded above, then there exists a least upper bound for S and if S is bounded below, then there exists a greatest lower bound for S . The least upper bound of S is denoted as $\sup(S)$ or sometimes as $l.u.b.(S)$ while the greatest lower bound of S is denoted as $\inf(S)$ sometimes as $g.l.b.(S)$. If there is no upper bound for S we say*

$\sup(S) = \infty$. If there is no lower bound, we say $\inf(S) = -\infty$.

The words mean exactly what they say. $\sup(S)$ is a number with the property that $s \leq \sup(S)$ for all $s \in S$ and out of all such “upper bounds” it is the smallest. $\inf(S)$ has the property that $\inf(S) \leq s$ for all $s \in S$ and if $l \leq s$ for all $s \in S$, then $l \leq \inf(S)$. In words, it is the largest lower bound and $\sup(S)$ is the smallest upper bound. Here the meaning of small and large are as follows. To say that x is smaller than y means that $x \leq y$ which also says that y is larger than x .

A consequence of this axiom is the nested interval lemma, Lemma 3.0.2.

Lemma 3.0.2 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

$$\lim_{k \rightarrow \infty} b^k - a^k = 0$$

then there is exactly one point in all of these intervals.

Proof: Since $I_k \supseteq I_{k+1}$, this implies

$$a^k \leq a^{k+1}, b^k \geq b^{k+1}. \quad (3.1)$$

Consequently, letting $k \leq l$,

$$a^l \leq a^k \leq b^k \leq b^l. \quad (3.2)$$

Thus

$$c \equiv \sup \{a^l : l = 1, 2, \dots\} = \sup \{a^l : l = k, k+1, \dots\} \leq b^k$$

because b^k is an upper bound for all the a^l . Then $c \geq a^l$ for all l . In other words $c \geq a^k$ for all k . Also $c \leq b^k$ for all k . Therefore, $c \in [a^k, b^k]$ for all k .

If the length of these intervals converges to 0, then there can be at most one point in their intersection since otherwise, you would have two different points c, d and the length of the k^{th} interval would then be at least as large as $|d - c|$ but both of these points would need to be in intervals having smaller length than this which can't happen. ■

Corollary 3.0.3 Suppose $\{x_n\}$ is a sequence contained in $[a, b]$. Then there exists $x \in [a, b]$ and a subsequence $\{x_{n_k}\}$ such that $\lim_{k \rightarrow \infty} x_{n_k} = x$.

Proof: Consider a sequence of closed intervals contained in $[a, b]$, I_1, I_2, \dots where I_{k+1} is one half of I_k and each I_k contains x_n for infinitely many values of n . Thus $I_1 = [a, b]$, I_2 is either $[a, \frac{a+b}{2}]$ or $[\frac{a+b}{2}, b]$, depending on which one contains x_n for infinitely many values of n . If both intervals have this property, just pick one. Let $x_{n_k} \in I_k$ and let $x_{n_{k+1}} \in I_{k+1}$ with $n_{k+1} > n_k$. This is possible to do because each I_k contains x_n for infinitely many values of n . Then by the nested interval lemma, there exists a unique point x contained in all of these intervals and $|x - x_{n_k}| < (b - a)/2^k$. ■

The next corollary is the extreme value theorem from calculus.

Corollary 3.0.4 *If $f : [a, b] \rightarrow \mathbb{R}$ is continuous, then there exists $x_M \in [a, b]$ such that*

$$f(x_M) = \sup \{f(x) : x \in [a, b]\}$$

and there exists $x_m \in [a, b]$ such that

$$f(x_m) = \inf \{f(x) : x \in [a, b]\}$$

Proof: From the definition of $\inf \{f(x) : x \in [a, b]\}$, there exists $x_n \in [a, b]$ such that

$$f(x_n) \leq \inf \{f(x) : x \in [a, b]\} + 1/n$$

That is, $\lim_{n \rightarrow \infty} f(x_n) = \inf \{f(x) : x \in [a, b]\}$. This is called a minimizing sequence. Therefore, there is a subsequence $\{x_{n_k}\}$ which converges to $x \in [a, b]$. By continuity of f it follows that

$$\inf \{f(x) : x \in [a, b]\} = \lim_{k \rightarrow \infty} f(x_{n_k}) = f(x)$$

The case where f achieves its maximum is similar. You just use a maximizing sequence. ■

Corollary 3.0.5 *If $\{x_n\}$ is a Cauchy sequence, then it converges.*

Proof: The Cauchy sequence is contained in some closed interval $[a, b]$. This is because, letting $\varepsilon = 1$, it follows that there exists N such that if $m, n \geq N$, then $|x_n - x_m| < 1$. In particular, for all $n \geq N$, $|x_n - x_N| < 1$. Therefore, $|x_n| \leq \max \{|x_N| + 1, |x_1|, |x_2|, \dots, |x_N|\}$ for all n . By Corollary 3.0.3, there is a subsequence of the Cauchy sequence, denoted as $\{x_{n_k}\}$ which converges to some $x \in [a, b]$. Since the original sequence is a Cauchy sequence, letting $\varepsilon > 0$ be given, there is N such that if $k, l \geq N$, then $|x_k - x_l| < \varepsilon/2$ and $|x_{n_k} - x| < \varepsilon/2$. Thus if $m \geq N$, then

$$|x - x_m| \leq |x - x_{n_m}| + |x_{n_m} - x_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

Indeed, if $m \geq N$, then $n_m \geq N$ because $\{x_{n_m}\}_{m=1}^{\infty}$ is a subsequence. Thus the original Cauchy sequence converges to x . ■

Actually, the convergence of every Cauchy sequence is equivalent to completeness and so it gives another way of defining completeness in contexts where no order is available. Recall completeness means that every nonempty set bounded above (below) has a least upper bound (greatest lower bound). More consideration of this issue is a good topic for advanced calculus courses. This standard definition involving least upper bounds depends on an order. One can prove that if you have the least upper bound property described in the above definition, then you also have the greatest lower bound property also described there and the other way around.

The Riemann integral pertains to bounded functions which are defined on a bounded interval. Let $[a, b]$ be a closed interval. A set of points in $[a, b]$, $\{x_0, \dots, x_n\}$ is a partition if

$$a = x_0 < x_1 < \dots < x_n = b.$$

Such partitions are denoted by P or Q .

Definition 3.0.6 A function $f : [a, b] \rightarrow \mathbb{R}$ is bounded if the set of values of f is contained in some interval. Thus

$$\sup \{f(x) : x \in [a, b]\} < \infty, \quad \inf \{f(x) : x \in [a, b]\} > -\infty$$

Letting P denote a partition,

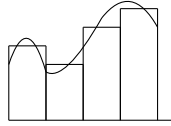
$$\|P\| \equiv \max \{|x_{i+1} - x_i| : i = 0, \dots, n-1\}.$$

A Riemann sum for a bounded f corresponding to a partition $P = \{x_0, \dots, x_n\}$ is a sum of the form

$$\sum_P f \equiv \sum_{i=1}^n f(y_i) (x_i - x_{i-1})$$

where $y_i \in [x_{i-1}, x_i]$. Then there are really many different Riemann sums corresponding to a given partition, depending on which y_i is chosen.

For example, suppose f is a function with positive values. The above Riemann sum involves adding areas of rectangles. Here is a picture:



The area under the curve is close to the sum of the areas of these rectangles and one would imagine that this would become an increasingly good approximation if you included more and narrower rectangles.

Definition 3.0.7 A bounded function defined on an interval $[a, b]$ is Riemann integrable means that there exists a number I such that for every $\varepsilon > 0$, there exists a $\delta > 0$ such that whenever $\|P\| < \delta$, and $\sum_P f$ is some Riemann sum corresponding to this partition, it follows that

$$\left| \sum_P f - I \right| < \varepsilon$$

This is written as

$$\lim_{\|P\| \rightarrow 0} \sum_P f = I$$

and when this number exists, it is denoted by

$$I = \int_a^b f(x) dx$$

One of the big theorems is on the existence of the integral whenever f is a continuous function. This requires a technical lemma which follows.

Lemma 3.0.8 Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous. Then for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $|x - y| < \delta, x, y \in [a, b]$, it follows that $|f(x) - f(y)| < \varepsilon$.

Proof: If not, then there exists $\varepsilon > 0$ and $x_n, y_n, |x_n - y_n| < 1/n$ but

$$|f(x_n) - f(y_n)| \geq \varepsilon.$$

By Corollary 3.0.3, there is a subsequence $\{x_{n_k}\}$ and point $x \in [a, b]$ such that $\lim_{k \rightarrow \infty} x_{n_k} = x$. Then it follows that also $\lim_{k \rightarrow \infty} y_{n_k} = x$ also because

$$|y_{n_k} - x| \leq |y_{n_k} - x_{n_k}| + |x_{n_k} - x|$$

and both of the terms on the right converge to 0. But then by continuity of f ,

$$0 = f(x) - f(x) = \lim_{k \rightarrow \infty} (f(x_{n_k}) - f(y_{n_k}))$$

which is impossible because $|f(x_{n_k}) - f(y_{n_k})| \geq \varepsilon$ for all k . ■

With this preparation, here is the major result on the existence of the integral of a continuous function.

Theorem 3.0.9 *Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous. Then $\int_a^b f(x) dx$ exists. In fact, there exists a sequence δ_m converging to 0 such that if $\|P\| < \delta_m$, and if $\sum_P f$ is a Riemann sum for P , then*

$$\left| \sum_P f - \int_a^b f dx \right| < \frac{2}{m} (b - a)$$

δ_m is defined to be such that if $|x - y| < \delta_m$, then $|f(x) - f(y)| < \frac{1}{m}$ and the sequence is decreasing.

Proof: Consider a partition P given by $a = x_0 < x_1 < \dots < x_n = b$. Then you could add in another point as follows:

$$a = x_0 < x_1 < \dots < x_{i-1} < x^* < x_i < \dots < x_n = b$$

Denote this one by Q . Then if you have a Riemann sum,

$$\sum_P f = \sum_{j=1}^n f(y_j) (x_j - x_{j-1})$$

You could write this sum in the following form.

$$\sum_{j=1}^{i-1} f(y_j) (x_j - x_{j-1}) + f(y_i) (x^* - x_{i-1}) + f(y_i) (x_i - x^*) + \sum_{j=i+1}^n f(y_j) (x_j - x_{j-1})$$

In fact, you could continue adding in points and doing the same trick and thereby write the original sum in terms of any partition containing P . If R is a partition containing P and if δ_m corresponds to $\varepsilon = 1/m$ in the above Lemma with $\dots > \delta_m > \delta_{m+1} \dots$ 3.0.8, then one can conclude that if $\|P\| < \delta_m$, then

$$\left| \sum_P f - \sum_R f \right| \leq \frac{1}{m} (b - a)$$

Now if $\|P\|, \|Q\| < \delta_m$, let $R = P \cup Q$. Then

$$\begin{aligned} \left| \sum_P f - \sum_Q f \right| &\leq \left| \sum_P f - \sum_R f \right| + \left| \sum_R f - \sum_Q f \right| \\ &\leq \frac{1}{m} (b-a) + \frac{1}{m} (b-a) = \frac{2}{m} (b-a) \end{aligned}$$

Let $M \geq \max \{|f(x)| : x \in [a, b]\}$. Then all Riemann sums are in the interval

$$[-M(b-a), M(b-a)]$$

Now let

$$S_n \equiv \left\{ \sum_P f : \|P\| < \delta_n \right\}$$

Then $S_n \supseteq S_{n+1}$ for all n thanks to the fact that the δ_n are decreasing. Let

$$I_n = [\inf(S_n), \sup(S_n)]$$

These are nested intervals contained in $[-M(b-a), M(b-a)]$ and so there exists I contained in them all. However, from the above computation,

$$\sup(S_n) - \inf(S_n) \leq \frac{2}{n} (b-a)$$

and so there is only one such I . Hence for any $\varepsilon > 0$ given, there exists $\delta > 0$ such that if $\|P\| < \delta$, then

$$\left| \sum_S f - I \right| < \varepsilon \blacksquare$$

We say that a bounded function f defined on an interval $[a, b]$ is Riemann integrable if the above integral exists. This is written as $f \in R([a, b])$. The above just showed that every continuous function is Riemann integrable.

Not all bounded functions are Riemann integrable. For example, let $x \in [0, 1]$ and

$$f(x) \equiv \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases} \quad (3.3)$$

This has no Riemann integral because you can pick a sequence of partitions P_n , such that $\|P_n\| < 1/n$ and each partition point is rational. Then for your Riemann sums, take the value of the function at the left end point. The resulting Riemann sum will always equal 1. But you could just as easily pick your point y_i in the Riemann sum to equal an irrational number and these Riemann sums will all equal 0. Therefore, the condition for integrability is violated for $\varepsilon = 1/4$.

If you can partition the interval $[a, b]$ into finitely many intervals $[z_{i-1}, z_i]$, such that a function f is continuous on each $[z_{i-1}, z_i]$, then the function will be integrable on $[a, b]$. This is roughly the claim of the next theorem.

Definition 3.0.10 A bounded function $f : [a, b] \rightarrow \mathbb{R}$ is called *piecewise continuous* if there are points z_i such that $a = z_0 < z_1 < \dots < z_n = b$ and continuous functions $g_i : [z_{i-1}, z_i] \rightarrow \mathbb{R}$ such that for $t \in (z_{i-1}, z_i)$, $g_i(t) = f(t)$.

Corollary 3.0.11 Let $f : [a, b] \rightarrow \mathbb{R}$ be piecewise continuous. Then f is Riemann integrable. Also

$$\int_a^b f dt = \sum_{i=1}^n \int_{z_{i-1}}^{z_i} g_i dt \quad (3.4)$$

Proof: Let P_i be a partition for $[z_{i-1}, z_i]$. Since there are only finitely many of these intervals, there exists $\delta > 0$ such that if $\|P_i\| < \delta$, then for each i ,

$$\left| \sum_{P_i} g_i - \int_{z_{i-1}}^{z_i} g_i dt \right| < \varepsilon$$

Let M_f be an upper bound for $|f|$ on $[a, b]$, M_g an upper bound for all $|g_i|$. Now let $\|P\| < \delta < \varepsilon$ where P is a partition of $[a, b]$, these points denoted as x_j . Let \hat{P}_i be those points of P which are in $(z_{i-1}, z_i]$ and let P_i consist of \hat{P}_i along with z_{i-1} and z_i . Thus $\|P_i\| < \delta$. Then for $y_i \in [x_{i-1}, x_i]$,

$$\left| \sum_{i=1}^n \int_{z_{i-1}}^{z_i} g_i dt - \sum_P f \right| \leq \sum_{i=1}^n \left| \int_{z_{i-1}}^{z_i} g_i dt - \sum_{x_j \in \hat{P}_i} f(y_j) (x_j - x_{j-1}) \right|$$

Now for $x_j \in \hat{P}_i$, $f(y_j) = g_i(y_j)$ except maybe at end points where these differ by no more than $2(M_f + M_g) \equiv 2M$. Thus the above is no more than

$$\begin{aligned} &\leq \sum_{i=1}^n \left| \int_{z_{i-1}}^{z_i} g_i dt - \sum_{x_j \in P_i} g_i(y_j) (x_j - x_{j-1}) \right| + \sum_{i=1}^n 4(M_f + M_g) \delta \\ &< n\varepsilon + 4Mn\delta < \varepsilon(n + 4Mn) \end{aligned}$$

Since ε is arbitrary, this shows that f is indeed Riemann integrable and equals 3.4. ■

Note that what has actually been shown is that if a bounded function f satisfies $f = g_i$ on $[z_{i-1}, z_i]$ except for possibly the end points, and g_i is Riemann integrable on $[z_{i-1}, z_i]$, then f is Riemann integrable on $[a, b]$. The above proof applies with no change.

It is important to notice that the integral is linear. That is, for α, β numbers and f, g piecewise continuous functions,

$$\int_a^b (\alpha f + \beta g) dx = \alpha \int_a^b f dx + \beta \int_a^b g dx$$

This is easy to see because such linearity holds for sums. Thus

$$\begin{aligned} \int_a^b (\alpha f + \beta g) dx &\equiv \lim_{\|P\| \rightarrow 0} \sum_P \alpha f + \beta g \\ &= \lim_{\|P\| \rightarrow 0} \alpha \sum_P f + \beta \sum_P g = \alpha \int_a^b f dx + \beta \int_a^b g dx \end{aligned}$$

I leave the details to you. Actually, this works under the assumption that f, g are Riemann integrable but in this case, you have to show that a linear combination of Riemann integrable functions is Riemann integrable. This is not hard but I don't want to waste time on it.

The above is the Riemann integral. There is another integral which can be proved to be equivalent to the above. It is called the Darboux integral.

Definition 3.0.12 For P a partition $a = x_0 < \cdots < x_n = b$ and

$$\begin{aligned} M_i &\equiv \{\sup f(x) : x \in [x_{i-1}, x_i]\}, \\ m_i &\equiv \inf \{\inf f(x) : x \in [x_{i-1}, x_i]\} \end{aligned}$$

for f a bounded function. Then the upper sum and lower sum are respectively

$$\begin{aligned} U(f, P) &\equiv \sum_{i=1}^n M_i (x_i - x_{i-1}), \\ L(f, P) &\equiv \sum_{i=1}^n m_i (x_i - x_{i-1}) \end{aligned}$$

$$\bar{I} \equiv \inf \{U(f, P) \text{ where } P \text{ is a partition}\}$$

$$\underline{I} \equiv \sup \{L(f, P) \text{ where } P \text{ is a partition}\}.$$

We say that f is Darboux integrable if $\bar{I} = \underline{I}$ and the Darboux integral is the common value of these.

Note that \underline{I} and \bar{I} are well defined real numbers and this definition of an integral really says that there is a unique number between all the upper sums and lower sums. If f is Riemann integrable, then it is not hard to see it is Darboux integrable. Indeed, from the definition, there exists P such that whenever you have a Riemann sum for P ,

$$\left| \int_a^b f dx - \sum_P f \right| < \varepsilon$$

In particular, this shows after a little consideration that

$$|U(f, P) - L(f, P)| < 2\varepsilon,$$

Thus, since ε is arbitrary, there can't be more than one number between all the upper and lower sums and this number must be the Riemann integral. One can also show that every lower sum is no larger than every upper sum, even if taken with respect to different partitions. In this book, we are mainly interested in piecewise continuous functions and so once you know these are Riemann integrable, it follows automatically that they are Darboux integrable. It can be shown that the two definitions are equivalent, but this is not needed in this book. I think it is a little more convenient to use the Darboux approach when dealing with the theory of the integral of a function of many variables which is done in the appendix. Either integral handles the functions of most interest and gives the same answer for these. However, both of these integrals have been obsolete for over a hundred years.

3.1 Properties of the Integral

To find the integral of a continuous function, one can often use another method which is much easier than taking the limit of Riemann sums. This other method is called the fundamental theorem of calculus. There are two forms to this theorem, one enabling the computation of the integral and another which gives the existence of a function whose derivative is a given function.

Theorem 3.1.1 Suppose $F'(x) = f(x)$ where f is a continuous function on $[a, b]$. Then

$$\int_a^b f(x) dx = F(b) - F(a) \quad (3.5)$$

Proof: Let $\varepsilon > 0$ be given and let P be a partition $a = x_0 < x_1 < \cdots < x_n = b$ such that whenever $\hat{y}_i \in [x_{i-1}, x_i]$,

$$\left| \int_a^b f(x) dx - \sum_{i=1}^n f(\hat{y}_i)(x_i - x_{i-1}) \right| < \varepsilon \quad (3.6)$$

Then from the mean value theorem, there exists $y_i \in (x_{i-1}, x_i)$ such that

$$\begin{aligned} F(b) - F(a) &= \sum_{i=1}^n (F(x_i) - F(x_{i-1})) \\ &= \sum_{i=1}^n F'(y_i)(x_i - x_{i-1}) = \sum_{i=1}^n f(y_i)(x_i - x_{i-1}) \end{aligned}$$

Let \hat{y}_i in 3.6 be equal to y_i just described. Then with the above, this shows that

$$\left| \int_a^b f(x) dx - (F(b) - F(a)) \right| < \varepsilon$$

Since ε is arbitrary, this verifies 3.5. ■

Example 3.1.2 Find $\int_0^2 \cos(t) dt$.

Note that $\cos(t) = \sin'(t)$ and so the above integral is $\sin(2) - \sin(0) = \sin(2)$.

Example 3.1.3 Find $\int_a^b \alpha dx$.

A function whose derivative is α is $x \rightarrow \alpha x$. Therefore, this integral is $\alpha b - \alpha a = \alpha(b - a)$.

The integral $\int_a^b f(t) dt$ has been defined when f is continuous and $a < b$. What if $a > b$? The following definition tells what this equals.

Definition 3.1.4 Let $[a, b]$ be an interval and let f be piecewise continuous on $[a, b]$ or more generally Riemann integrable on this interval. Then

$$\int_b^a f(t) dt \equiv - \int_a^b f(t) dt$$

Observation 3.1.5 With the above definition, $\int_a^b dx$ is linear satisfying

$$\int_a^b (\alpha f + \beta g) dx = \alpha \int_a^b f dx + \beta \int_a^b g dx$$

if $a < b$ or $b < a$. Also $\int_a^b \alpha dx = \alpha b - \alpha a$ if $a < b$ or $b < a$.

Note that this definition must hold if we want to continue to use Theorem 3.1.1. With this definition, one can give a convenient theorem. It holds for general Riemann integrable functions. However, I am stating it only for the case of most interest, piecewise continuous ones because I am basing the argument on Corollary 3.0.11. As noted, this corollary will end up holding in greater generality with very little change in the proof.

Theorem 3.1.6 *Suppose a, b, c are all points in some interval on which f is piecewise continuous. Then*

$$\int_a^b f(t) dt + \int_b^c f(t) dt = \int_a^c f(t) dt \quad (3.7)$$

Proof: case 1: $a < b < c$ In this case, 3.7 follows from Corollary 3.0.11.

case 2: $a < c < b$ In this case, Corollary 3.0.11 implies

$$\int_a^c f(x) dx + \int_c^b f(x) dx = \int_a^b f(x) dx$$

and so

$$\begin{aligned} \int_a^c f(x) dx &= \int_a^b f(x) dx - \int_c^b f(x) dx \\ &= \int_a^b f(x) dx + \int_b^c f(x) dx \end{aligned}$$

case 3: $c < a < b$ In this case, Corollary 3.0.11 implies

$$\int_c^a f(x) dx + \int_a^b f(x) dx = \int_c^b f(x) dx$$

so

$$\begin{aligned} \int_c^a f(x) dx + \int_a^b f(x) dx + \int_b^c f(x) dx &= 0 \\ \int_a^b f(x) dx + \int_b^c f(x) dx &= \int_a^c f(x) dx \end{aligned}$$

This includes all cases and proves the theorem. ■

Next is the triangle inequality.

Proposition 3.1.7 *Let a, b be in an interval on which f is piecewise continuous (or Riemann integrable). Then*

$$\left| \int_a^b f(x) dx \right| \leq \int_a^b |f(x)| dx$$

Proof: I will give the proof for almost the only case of interest in this book, piecewise continuous. If f is piecewise continuous, then so is $|f|$. Hence there is no problem with existence of the integral. Suppose first that $a > b$. Then, since $|f| - f, |f| + f$ are nonnegative functions, all Riemann sums are nonnegative and so their limit is also nonnegative. Hence

$$\begin{aligned} 0 &\leq \int_b^a (|f(x)| - f(x)) dx = \int_b^a |f(x)| dx - \int_b^a f(x) dx \\ 0 &\leq \int_b^a (|f(x)| + f(x)) dx = \int_b^a |f(x)| dx + \int_b^a f(x) dx \end{aligned}$$

and so

$$\begin{aligned}\int_b^a f(x) dx &\leq \int_b^a |f(x)| dx = \left| \int_a^b |f(x)| dx \right| \\ -\int_b^a f(x) dx &\leq \int_b^a |f(x)| dx = \left| \int_a^b |f(x)| dx \right|\end{aligned}$$

It follows that in this case where $a > b$,

$$\left| \int_b^a f(x) dx \right| = \left| \int_a^b f(x) dx \right| \leq \left| \int_a^b |f(x)| dx \right|$$

The argument is the same in case $a < b$ except you work with \int_a^b rather than \int_b^a . ■

With these basic properties of the integral, here is the other form of the fundamental theorem of calculus. This major theorem, due to Newton and Leibniz shows the existence of an “anti-derivative” for any continuous function.

Theorem 3.1.8 *Let f be continuous on $[a, b]$. Also let*

$$F(t) \equiv \int_a^t f(x) dx$$

Then for every $t \in (a, b)$,

$$F'(t) = f(t).$$

Proof: For $t \in (a, b)$ and $|h|$ sufficiently small, $t+h \in (a, b)$. Always let h be this small. Then, from the above properties of integrals in Proposition 3.1.7, and Theorem 3.1.6,

$$\frac{F(t+h) - F(t)}{h} = \frac{1}{h} \left(\int_a^{t+h} f(x) dx - \int_a^t f(x) dx \right) = \frac{1}{h} \int_t^{t+h} f(x) dx$$

Now from Observation 3.1.5,

$$\frac{1}{h} \int_t^{t+h} f(t) dt = f(t)$$

Therefore, by the properties of the integral given above,

$$\begin{aligned}\left| \frac{F(t+h) - F(t)}{h} - f(t) \right| &= \left| \frac{1}{h} \int_t^{t+h} f(x) dx - \frac{1}{h} \int_t^{t+h} f(t) dx \right| \\ &= \left| \frac{1}{h} \int_t^{t+h} (f(x) - f(t)) dx \right| \\ &\leq \frac{1}{|h|} \left| \int_t^{t+h} |f(x) - f(t)| dx \right|\end{aligned}$$

Now if $|h|$ is small enough, $|f(x) - f(t)| < \varepsilon$ by continuity of f at x . Therefore, for $|h|$ this small,

$$\left| \frac{F(t+h) - F(t)}{h} - f(t) \right| \leq \frac{1}{|h|} \left| \int_t^{t+h} \varepsilon dx \right| = \varepsilon$$

Since ε is arbitrary, it follows from the definition of the limit that

$$\lim_{h \rightarrow 0} \frac{F(t+h) - F(t)}{h} = f(t) \quad \blacksquare$$

Corollary 3.1.9 For $F(t)$ defined as above, it is also true that $F'(a) = f(a)$ and $F'(b) = f(b)$ provided the derivatives are taken respectively from the right and from the left.

Proof: You repeat the above argument paying attention to the sign of h . Otherwise there is no change. ■

Definition 3.1.10 When $F'(t) = f(t)$ for t on some interval, the function $t \rightarrow F(t)$ is called an anti-derivative for f . The set of all anti-derivatives is denoted as $\int f dx$. Thus $\int f dx$ is a collection of functions, not a number, while $\int_a^b f(x) dx$ is a number.

Proposition 3.1.11 Suppose $F, G \in \int f dx$ for x in some interval. Then there exists a constant C such that $F(x) + C = G(x)$.

Proof: It comes from the mean value theorem. By assumption $(G - F)' = 0$ and so if x_0 is a fixed point in the interval, then if x is another point, $(G - F)(x) - (G - F)(x_0) = (G - F)'(z)(x - x_0)$ for some z between x and x_0 . But by assumption, $(G - F)'(z) = 0$ and so $(G - F)(x)$ must equal $(G - F)(x_0)$ for all x in the interval. Let $C = (G - F)(x_0)$. ■

3.2 Uniform Convergence of Continuous Functions

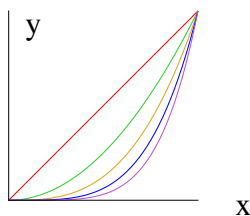
Suppose for each $n \in \mathbb{N}$, f_n is a continuous function defined on some interval $[a, b]$. Also suppose that for each fixed $x \in [a, b]$, $\lim_{n \rightarrow \infty} f_n(x) = f(x)$. This is called pointwise convergence. Does it follow that f is continuous on $[a, b]$? The answer is NO. Consider 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.



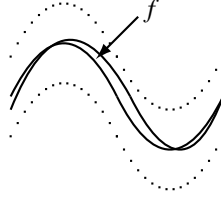
If you want the convergence to carry continuity with it you need something more than pointwise convergence.

Definition 3.2.1 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.

The following picture illustrates the above definition.



The dotted lines define sort of a tube centered about the graph of f and the graph of the function f_n fits in this tube for all n sufficiently large. The tube can be made as narrow as desired.

It is convenient to observe the following properties of $\|\cdot\|_\infty$, written $\|\cdot\|$ for short.

Lemma 3.2.2 *The norm $\|\cdot\|_\infty$ satisfies the following properties.*

$$\|f\| \geq 0 \text{ and equals 0 if and only if } f = 0 \quad (3.8)$$

For α a number,

$$\|\alpha f\| = |\alpha| \|f\| \quad (3.9)$$

$$\|f + g\| \leq \|f\| + \|g\| \quad (3.10)$$

Proof: The first claim 3.8 is obvious. As to 3.9, 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.

Theorem 3.2.3 *Let f_n be continuous on D and suppose $\lim_{n \rightarrow \infty} \|f_n - f\| = 0$. 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 ,

$$\begin{aligned} |f(y) - f(x)| &\leq |f(y) - f_n(y)| + |f_n(y) - f_n(x)| + |f_n(x) - f(x)| \\ &< \|f - f_n\| + \frac{\varepsilon}{3} + \|f_n - f\| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \end{aligned}$$

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. \blacksquare

This implies the following interesting corollary about a uniformly Cauchy sequence of continuous functions.

Definition 3.2.4 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 $m, k > n_\varepsilon$

$$\|f_m - f_k\| < \varepsilon$$

Corollary 3.2.5 Suppose $\{f_n\}$ is a uniformly Cauchy sequence of 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 . By completeness of \mathbb{R} , this sequence converges for each x . Let $f(x) \equiv \lim_{n \rightarrow \infty} f_n(x)$. Then by continuity of $y \rightarrow |y - f_n(x)|$, for each x ,

$$|f(x) - f_n(x)| = \lim_{m \rightarrow \infty} |f_m(x) - f_n(x)| \leq \lim_{m \rightarrow \infty} \inf_{m \rightarrow \infty} \|f_m - f_n\| < \varepsilon$$

provided n is sufficiently large. Since x is arbitrary, this shows that

$$\|f - f_n\| \equiv \sup_{x \in [a, b]} |f(x) - f_n(x)| \leq \varepsilon$$

if n is large enough. this says $\lim_{n \rightarrow \infty} \|f_n - f\| = 0$. Now the continuity of f follows from Theorem 3.2.3. How many such functions f are there? There can be only one because $f(x)$ must equal the limit of $f_n(x)$. ■

3.3 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 3.3.1 Let f_n be continuous and converging uniformly to f on $[a, b]$. 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 3.2.3. 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 $n \rightarrow \infty$. ■

Chapter 4

Some Important Improper Integrals

4.1 Gamma Function

This belongs to a larger set of ideas concerning improper integrals. I will just give enough of an introduction to this to present the very important gamma function. 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, then there is no hope for it at all. However, one can consider limits of Riemann integrals. The following definition is sufficient to deal with the gamma function in the generality needed in this book.

Definition 4.1.1 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 $\alpha > 0$.

Lemma 4.1.2 The limits in the above definition exist 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} . However,

$$\int_\delta^1 t^{\alpha-1} dt \leq \frac{1}{\alpha}$$

so the limit of these integrals exists. Also $e^{-t}t^{\alpha-1} \leq Ce^{-(t/2)}$ for suitable C if $t > 1$. This is obvious if $\alpha - 1 < 0$ and in the other case it is also clear because exponential growth exceeds polynomial growth. Thus

$$\int_1^R e^{-t}t^{\alpha-1}dt \leq \int_1^R Ce^{-(t/2)}dt \leq 2Ce^{(-1/2)} - 2Ce^{(-R/2)} \leq 2Ce^{(-1/2)}$$

Thus these integrals also converge as $R \rightarrow \infty$. It follows that $\Gamma(\alpha)$ makes sense. ■

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 4.1.3 *For n a positive integer, $n! = \Gamma(n+1)$. In general, $\Gamma(1) = 1, \Gamma(\alpha+1) = \alpha\Gamma(\alpha)$*

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 4.1.4 *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 .*

Theorem 4.1.5 *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 $|\phi(t)| \leq Ce^{\lambda t}, \lambda \geq 0$,

$$\frac{f^{(k)}(s+h) - f^{(k)}(s)}{h} = \int_0^{\infty} (-t)^k \frac{e^{-(s+h)t} - e^{-st}}{h} \phi(t) dt$$

$$= \int_0^\infty (-t)^k e^{-st} \left(\frac{e^{-ht} - 1}{h} \right) \phi(t) dt = \int_0^\infty (-t)^k e^{-st} \left((-t) e^{\theta(h,t)} \right) \phi(t) dt$$

where $\theta(h, t)$ is between $-ht$ and 0, this by the mean value theorem. Thus by mean value theorem again,

$$\begin{aligned} & \left| \frac{f^{(k)}(s+h) - f^{(k)}(s)}{h} - \int_0^\infty (-t)^{k+1} e^{-st} \phi(t) dt \right| \\ & \leq \int_0^\infty |t|^{k+1} C e^{\lambda t} e^{-st} \left| e^{\theta(h,t)} - 1 \right| dt \leq \int_0^\infty t^{k+1} C e^{\lambda t} e^{-st} e^{\alpha(h,t)} |ht| dt \\ & \leq \int_0^\infty t^{k+2} C e^{\lambda t} e^{-st} |h| e^{t|h|} dt = C|h| \int_0^\infty t^{k+2} e^{-(s-(\lambda+|h|))t} dt \end{aligned}$$

Let $u = (s - (\lambda + |h|))t$, $du = (s - (\lambda + |h|)) dt$. Then the above equals

$$\begin{aligned} & C|h| \int_0^\infty \left(\frac{u}{s - (\lambda + |h|)} \right)^{k+2} e^{-u} \frac{1}{(s - (\lambda + |h|))} du \\ & = \frac{C|h|}{(s - (\lambda + |h|))^{k+3}} \int_0^\infty e^{-u} u^{k+2} du = \frac{C|h|}{(s - (\lambda + |h|))^{k+3}} \Gamma(k+3) \end{aligned}$$

Thus, as $h \rightarrow 0$, this converges to 0 and so this proves the theorem. ■

The function $s \rightarrow f(s)$ in the above theorem is called the Laplace transform of ϕ .

4.2 Laplace Transforms

Suppose f is piecewise continuous on each interval $[0, R]$, meaning that it is bounded on that interval and equals a continuous function on each of finitely many closed subintervals except for the end points as described in Definition 3.0.10. Then from Corollary 3.0.11, $t \rightarrow f(t)$ is integrable. So is $t \rightarrow e^{-st} f(t)$. It is tacitly assumed that f is as just described in all that follows. It is much nicer to formulate this in terms of the Lebesgue integral however and use a condition of measurability instead of all this piecewise continuous nonsense.

Definition 4.2.1 We say that a function defined on $[0, \infty)$ has exponential growth if for some $\lambda \geq 0$, and $C > 0$,

$$|f(t)| \leq C e^{\lambda t}$$

Note that this condition is satisfied if $|f(t)| \leq a + b e^{\lambda t}$. You simply pick $C > \max(a, b)$ and observe that $a + b e^{\lambda t} \leq 2C e^{\lambda t}$.

Proposition 4.2.2 Let f have exponential growth and be continuous except for finitely many points in $[0, R]$ 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 is the definition of the following 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 [37]. 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 4.2.3 *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

$$\frac{1}{k!} \int_0^{\infty} e^{-t} t^k dt = \Gamma(k+1) \frac{1}{k!} = k! \frac{1}{k!} = 1$$

To see this, use integration by parts.

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 $[0, 1]$,

$$\begin{aligned} \left| \int_0^{1-\delta} \frac{k^{k+1}}{k!} (e^{-u}u)^k (\phi(u) - \phi(1)) du \right| & \leq K \int_0^{1-\delta} \frac{k^{k+1}}{k!} (e^{-u}u)^k du \\ & \leq K \frac{k^{k+1}}{k!} \left(e^{-(1-\delta)} (1-\delta) \right)^k (1-\delta) \end{aligned}$$

Now this converges to 0 as $k \rightarrow \infty$. In fact, for $a < 1$, $\lim_{k \rightarrow \infty} \frac{k^{k+1}}{k!} (e^{-a}a)^k = 0$ because of the ratio test which shows that for $a < 1$, $\sum_k \frac{k^{k+1}}{k!} (e^{-a}a)^k < \infty$ which implies the k^{th} term converges to 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 $|\phi(u) - \phi(1)| < \varepsilon$. Thus

$$\left| \int_{1-\delta}^{1+\delta} \frac{k^{k+1}}{k!} (e^{-u}u)^k (\phi(u) - \phi(1)) du \right| \leq \varepsilon \int_0^\infty \frac{k^{k+1}}{k!} (e^{-u}u)^k du = \varepsilon$$

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 4.1.5 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 $\phi(t)$. ■

Part I

Linear Algebra And Multivariable Calculus

Chapter 5

Fundamentals

5.1 \mathbb{F}^n

The notation, \mathbb{F}^n refers to the collection of ordered lists of n numbers. These numbers can be either real or complex numbers. More precisely, consider the following definition. If we mean real numbers, the symbol \mathbb{R}^n is used. If nothing is specified, assume the symbol means \mathbb{C}^n .

Definition 5.1.1 *Define*

$$\mathbb{F}^n \equiv \{(x_1, \dots, x_n) : x_j \in \mathbb{F} \text{ for } j = 1, \dots, n\}.$$

$(x_1, \dots, x_n) = (y_1, \dots, y_n)$ if and only if for all $j = 1, \dots, n$, $x_j = y_j$. When

$$(x_1, \dots, x_n) \in \mathbb{R}^n,$$

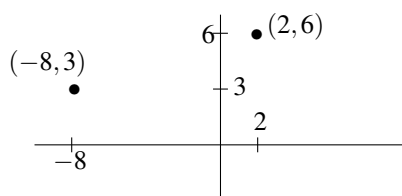
it is conventional to denote (x_1, \dots, x_n) by the single bold face letter \mathbf{x} . The numbers x_j are called the **coordinates**. The set

$$\{(0, \dots, 0, t, 0, \dots, 0) : t \in \mathbb{R}\}$$

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}^n 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 $n = 1$. 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 $n = 2$ 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 $n = 3$. 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 $n \leq 3$. 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 n 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 n dimensional space and its Cartesian coordinates.

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.

5.2 Algebra in \mathbb{R}^n

There are two algebraic operations done with points of \mathbb{R}^n . One is addition and the other is multiplication by numbers, called scalars. Yes, numbers =scalars.

Definition 5.2.1 If $\mathbf{x} \in \mathbb{R}^n$ and a is a number, also called a **scalar**, then $a\mathbf{x} \in \mathbb{R}^n$ is defined by

$$a\mathbf{x} = a(x_1, \dots, x_n) \equiv (ax_1, \dots, ax_n). \quad (5.1)$$

This is known as **scalar multiplication**. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ then $\mathbf{x} + \mathbf{y} \in \mathbb{R}^n$ and is defined by

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1, \dots, x_n) + (y_1, \dots, y_n) \\ &\equiv (x_1 + y_1, \dots, x_n + y_n) \end{aligned} \quad (5.2)$$

An element of \mathbb{R}^n $\mathbf{x} \equiv (x_1, \dots, x_n)$ is often 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 5.2.2 For \mathbf{v}, \mathbf{w} vectors in \mathbb{R}^n and α, β scalars, (real numbers), the following hold.

$$\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}, \quad (5.3)$$

the commutative law of addition,

$$(\mathbf{v} + \mathbf{w}) + \mathbf{z} = \mathbf{v} + (\mathbf{w} + \mathbf{z}), \quad (5.4)$$

the associative law for addition,

$$\mathbf{v} + \mathbf{0} = \mathbf{v}, \quad (5.5)$$

the existence of an additive identity

$$\mathbf{v} + (-\mathbf{v}) = \mathbf{0}, \quad (5.6)$$

the existence of an additive inverse, Also

$$\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}, \quad (5.7)$$

$$(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}, \quad (5.8)$$

$$\alpha(\beta\mathbf{v}) = \alpha\beta(\mathbf{v}), \quad (5.9)$$

$$1\mathbf{v} = \mathbf{v}. \quad (5.10)$$

In the above $\mathbf{0} = (0, \dots, 0)$.

You should verify these properties all hold. For example, consider 5.7.

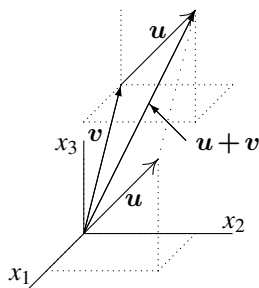
$$\begin{aligned} \alpha(\mathbf{v} + \mathbf{w}) &= \alpha(v_1 + w_1, \dots, v_n + w_n) = (\alpha(v_1 + w_1), \dots, \alpha(v_n + w_n)) \\ &= (\alpha v_1 + \alpha w_1, \dots, \alpha v_n + \alpha w_n) = (\alpha v_1, \dots, \alpha v_n) + (\alpha w_1, \dots, \alpha w_n) = \alpha\mathbf{v} + \alpha\mathbf{w}. \end{aligned}$$

As usual, subtraction is defined as $\mathbf{x} - \mathbf{y} \equiv \mathbf{x} + (-\mathbf{y})$.

5.3 Geometric Meaning Of Vector Addition In \mathbb{R}^3

It was explained earlier that an element of \mathbb{R}^n is an n 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 $n = 3$ and in two dimensional space, in the case where $n = 2$. This point was specified relative to some coordinate axes.

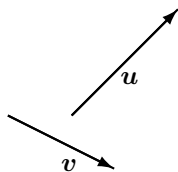
Consider the case where $n = 3$ for now. If you draw an arrow from the point in three dimensional space determined by $(0,0,0)$ to the point (a,b,c) with its tail sitting at the point $(0,0,0)$ and its point at the point (a,b,c) , this arrow is called the **position vector** of the point determined by $u \equiv (a,b,c)$. 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 .

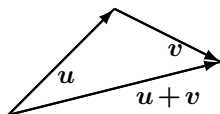
The following example is art.

Example 5.3.1 Here is a picture of two vectors u and v .

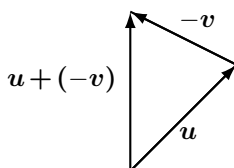


Sketch a picture of $u + v$, $u - v$, and $u + 2v$.

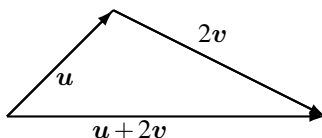
First here is a picture of $u + v$. You first draw u and then at the point of u you place the tail of v as shown. Then $u + v$ is the vector which results which is drawn in the following pretty picture.



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.

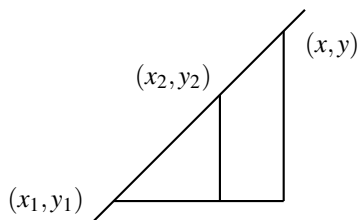


5.4 Lines

To begin with consider the case $n = 1, 2$. In the case where $n = 1$, 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

$$\begin{aligned} y &= y_1 + mt(x_2 - x_1) \\ &= y_1 + t(y_2 - y_1). \end{aligned}$$

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}^n .

Definition 5.4.1 A line in \mathbb{R}^n 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 5.4.2 Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ with $\mathbf{a} \neq \mathbf{0}$. Then $\mathbf{x} = t\mathbf{a} + \mathbf{b} \in \mathbb{R}^n$, 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 . ■

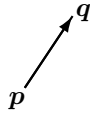
Definition 5.4.3 The vector \mathbf{a} in the above lemma is called a **direction vector** for the line.

Definition 5.4.4 Let \mathbf{p} and \mathbf{q} be two points in \mathbb{R}^n , $\mathbf{p} \neq \mathbf{q}$. The **directed line segment** from \mathbf{p} to \mathbf{q} , denoted by $\overrightarrow{\mathbf{p}\mathbf{q}}$, 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 $\overrightarrow{\mathbf{p}\mathbf{q}}$ 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 5.4.5 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 5.4.6 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

$$(x, y, z) = (1, 2, 0) + t(1, 2, 1), t \in \mathbb{R}. \quad (5.11)$$

Sometimes people elect to write a line like the above in the form

$$x = 1 + t, y = 2 + 2t, z = t, t \in \mathbb{R}. \quad (5.12)$$

This is a set of scalar parametric equations which amounts to the same thing as 5.11.

There is one other form for a line which is sometimes considered useful. It is the so called symmetric form. Consider the line of 5.12. 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 5.4.7 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

$$(2, 1, -3) + t(3, 2, 1) = (x, y, z), t \in \mathbb{R}.$$

5.5 Distance in \mathbb{R}^n

How is distance between two points in \mathbb{R}^n defined?

Definition 5.5.1 Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ be two points in \mathbb{R}^n . Then $|\mathbf{x} - \mathbf{y}|$ indicates the distance between these points and is defined as

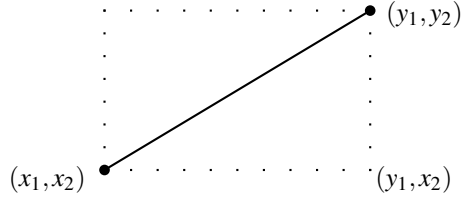
$$\text{distance between } \mathbf{x} \text{ and } \mathbf{y} \equiv |\mathbf{x} - \mathbf{y}| \equiv \left(\sum_{k=1}^n |x_k - y_k|^2 \right)^{1/2}.$$

This is called the **distance formula**. Thus $|\mathbf{x}| \equiv |\mathbf{x} - \mathbf{0}|$. The symbol $B(\mathbf{a}, r)$ is defined by

$$B(\mathbf{a}, r) \equiv \{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x} - \mathbf{a}| < r\}.$$

This is called an **open ball** of radius r centered at \mathbf{a} . It gives all the points in \mathbb{R}^n 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 $n = 2$.

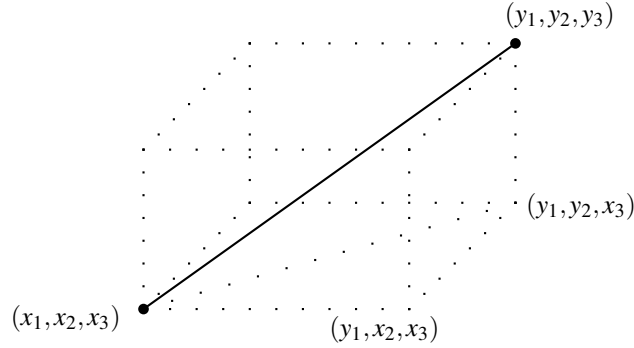


There are two points in the plane whose Cartesian coordinates are (x_1, x_2) and (y_1, y_2) respectively. Then the solid line joining these two points is the hypotenuse of a right triangle which is half of the rectangle shown in dotted lines. What is its length? Note the lengths of the sides of this 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 $n = 3$ 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 dotted 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 dotted line joining the following two points (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 5.5.2 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 in this book but sometimes this sort of thing is done.

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 5.5.3 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

$$2x + 2y + 2z = -9. \quad (5.13)$$

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 5.13 holds.

The following lemma is fundamental. It is a form of the Cauchy Schwarz inequality.

Lemma 5.5.4 *Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ be two points in \mathbb{R}^n . Then*

$$\left| \sum_{i=1}^n x_i y_i \right| \leq |\mathbf{x}| |\mathbf{y}|. \quad (5.14)$$

Proof: Let θ be either 1 or -1 such that

$$\theta \sum_{i=1}^n x_i y_i = \sum_{i=1}^n x_i (\theta y_i) = \left| \sum_{i=1}^n x_i y_i \right|$$

and consider $p(t) \equiv \sum_{i=1}^n (x_i + t\theta y_i)^2$. Then for all $t \in \mathbb{R}$,

$$\begin{aligned} 0 &\leq p(t) = \sum_{i=1}^n x_i^2 + 2t \sum_{i=1}^n x_i \theta y_i + t^2 \sum_{i=1}^n y_i^2 \\ &= |\mathbf{x}|^2 + 2t \sum_{i=1}^n x_i \theta y_i + t^2 |\mathbf{y}|^2 \end{aligned}$$

If $|\mathbf{y}| = 0$ then 5.14 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 zeroes or one repeated zero. If it has two zeroes, the above inequality must be violated because in this case the graph must dip below the x axis. Therefore, it either has no zeroes or exactly one. From the quadratic formula this happens exactly when

$$4 \left(\sum_{i=1}^n x_i \theta y_i \right)^2 - 4 |\mathbf{x}|^2 |\mathbf{y}|^2 \leq 0$$

and so

$$\sum_{i=1}^n x_i \theta y_i = \left| \sum_{i=1}^n x_i y_i \right| \leq |\mathbf{x}| |\mathbf{y}|$$

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

$$|\mathbf{x} - \mathbf{y}| = |\mathbf{y} - \mathbf{x}|,$$

$$|\mathbf{x} - \mathbf{y}| \geq 0 \text{ and equals 0 only if } \mathbf{y} = \mathbf{x}.$$

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 5.5.5 *Let \mathbf{x}, \mathbf{y} be points of \mathbb{R}^n . Then*

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|.$$

Proof: Using the Cauchy Schwarz inequality, Lemma 5.5.4,

$$\begin{aligned} |\mathbf{x} + \mathbf{y}|^2 &\equiv \sum_{i=1}^n (x_i + y_i)^2 \\ &= \sum_{i=1}^n x_i^2 + 2 \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2 \\ &\leq |\mathbf{x}|^2 + 2|\mathbf{x}||\mathbf{y}| + |\mathbf{y}|^2 \\ &= (|\mathbf{x}| + |\mathbf{y}|)^2 \end{aligned}$$

and so upon taking square roots of both sides,

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|$$

■

5.6 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 5.5.1. 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.

$$|\alpha \mathbf{x}| = |\alpha| |\mathbf{x}|.$$

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 \mathbf{x} and $\alpha \mathbf{x}$ are both points on the same line. Note that there is no change in this when you replace \mathbb{R}^3 with \mathbb{R}^n .

5.7 Exercises

1. Verify all the properties 5.3-5.10.
2. Compute the following
 - (a) $5(1, 2, 3, -2) + 6(2, 1, -2, 7)$
 - (b) $5(1, 2, -2) - 6(2, 1, -2)$
 - (c) $-3(1, 0, 3, -2) + (2, 0, -2, 1)$
 - (d) $-3(1, -2, -3, -2) - 2(2, -1, -2, 7)$
 - (e) $-(2, 2, -3, -2) + 2(2, 4, -2, 7)$
3. Find symmetric equations for the line through the points $(2, 2, 4)$ and $(-2, 3, 1)$.
4. Find symmetric equations for the line through the points $(1, 2, 4)$ and $(-2, 1, 1)$.
5. Symmetric equations for a line are given. Find parametric equations of the line.
 - (a) $\frac{x+1}{3} = \frac{2y+3}{2} = z + 7$
 - (b) $\frac{2x-1}{3} = \frac{2y+3}{6} = z - 7$
 - (c) $\frac{x+1}{3} = 2y + 3 = 2z - 1$
 - (d) $\frac{1-2x}{3} = \frac{3-2y}{2} = z + 1$
 - (e) $\frac{x-1}{3} = \frac{2y-3}{5} = z + 2$
 - (f) $\frac{x+1}{3} = \frac{3-y}{5} = z + 1$
6. Parametric equations for a line are given. Find symmetric equations for the line if possible. If it is not possible to do it explain why.
 - (a) $x = 1 + 2t, y = 3 - t, z = 5 + 3t$
 - (b) $x = 1 + t, y = 3 - t, z = 5 - 3t$
 - (c) $x = 1 + 2t, y = 3 + t, z = 5 + 3t$
 - (d) $x = 1 - 2t, y = 1, z = 1 + t$
 - (e) $x = 1 - t, y = 3 + 2t, z = 5 - 3t$
 - (f) $x = t, y = 3 - t, z = 1 + t$
7. 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) $(1, 2, 0), (1, 1, 0)$
 - (d) $(1, 0, -6), (-2, -1, 3)$
 - (e) $(-1, -2, -1), (2, 1, -1)$

(f) $(0, 0, 0), (2, -3, 1)$

8. 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) $x = 1 + t, y = 3 + 3t, z = 5 - t$

(c) $x = 7 + t, y = 3 + 4t, z = 5 - 3t$

(d) $x = 2t, y = -3t, z = 3t$

(e) $x = 2t, y = 3 + 2t, z = 5 + t$

(f) $x = t, y = 3 + 3t, z = 5 + t$

9. 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) $(0, 1, 1), (2, 5, 0)$

(c) $(1, 1, 0), (0, 1, 2)$

(d) $(0, 1, 3), (0, 3, 0)$

(e) $(0, 1, 0), (0, 6, 2)$

(f) $(0, 1, 2), (2, 0, 2)$

10. Draw a picture of the points in \mathbb{R}^2 which are determined by the following ordered pairs.

(a) $(1, 2)$

(b) $(-2, -2)$

(c) $(-2, 3)$

(d) $(2, -5)$

11. Does it make sense to write $(1, 2) + (2, 3, 1)$? Explain.

12. Draw a picture of the points in \mathbb{R}^3 which are determined by the following ordered triples.

(a) $(1, 2, 0)$

(b) $(-2, -2, 1)$

(c) $(-2, 3, -2)$

13. 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.

14. 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 .

15. 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 $y_0 > l$.
16. A sphere centered at the point $(x_0, y_0, z_0) \in \mathbb{R}^3$ having radius r consists of all points (x, y, z) whose distance to (x_0, y_0, z_0) equals r . Write an equation for this sphere in \mathbb{R}^3 .
17. 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.
18. Repeat the same problem except this time let the distance between the two points be $|x - x'| + |y - y'|$.
19. 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 17 ($|(x, y, z)| = \text{maximum of } |z|, |x|, |y|$)?
20. 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) .
21. 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.

22. 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 $\left(\frac{x}{\alpha}\right)^2 - \left(\frac{y}{\beta}\right)^2 = 1$. This is a nice exercise in messy algebra.

23. 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 $|(x, y) - (x_1, y_1)| = |(x, y) - (x_2, y_2)|$.
24. Show that $|\alpha x| = |\alpha||x|$ whenever $x \in \mathbb{R}^n$ for any positive integer n .

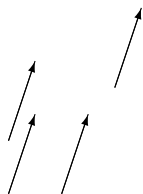
5.8 Physical Vectors

Suppose you push on something. What is important? There are really two things which are important, how hard you push and the direction you push.

Definition 5.8.1 *Force is a vector. The magnitude of this vector is a measure of how hard it is pushing. It is measured in units such as Newtons or pounds or tons. Its direction is the direction in which the push is taking place.*

Of course this is a little vague and will be left a little vague until the presentation of Newton's second law later. See the appendix on this or any physics book.

Vectors are used to model force and other physical vectors like velocity. What was just described would be called a force vector. It has two essential ingredients, its magnitude and its direction. Think of vectors as directed line segments or arrows as shown in the following picture in which all the directed line segments are considered to be the same vector because they have the same direction, the direction in which the arrows point, and the same magnitude (length).



Because of this fact that only direction and magnitude are important, it is always possible to put a vector in a certain particularly simple form. Let \vec{pq} be a directed line segment or vector. Then from Definition 5.4.4 it follows that \vec{pq} consists of the points of the form

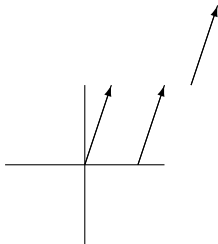
$$p + t(q - p)$$

where $t \in [0, 1]$. Subtract p from all these points to obtain the directed line segment consisting of the points

$$0 + t(q - p), t \in [0, 1].$$

The point in \mathbb{R}^n , $q - p$, will represent the vector.

Geometrically, the arrow \vec{pq} , was slid so it points in the same direction and the base is at the origin 0 . For example, see the following picture.



In this way vectors can be identified with points of \mathbb{R}^n .

Definition 5.8.2 *Let $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. The **position vector** of this point is the vector whose point is at x and whose tail is at the origin $(0, \dots, 0)$. If $x = (x_1, \dots, x_n)$ is called*

a vector, the vector which is meant, is this position vector just described. Another term associated with this is **standard position**. A vector is in standard position if the tail is placed at the origin.

It is customary to identify the point in \mathbb{R}^n with its position vector.

The magnitude of a vector determined by a directed line segment \overrightarrow{pq} is just the distance between the point p and the point q . By the distance formula this equals

$$\left(\sum_{k=1}^n (q_k - p_k)^2 \right)^{1/2} = |\mathbf{p} - \mathbf{q}|$$

and for \mathbf{v} any vector in \mathbb{R}^n the magnitude of \mathbf{v} equals $(\sum_{k=1}^n v_k^2)^{1/2} = |\mathbf{v}|$.

Example 5.8.3 Consider the vector $\mathbf{v} \equiv (1, 2, 3)$ in \mathbb{R}^n . Find $|\mathbf{v}|$.

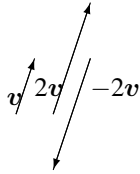
First, the vector is the directed line segment (arrow) which has its base at $\mathbf{0} \equiv (0, 0, 0)$ and its point at $(1, 2, 3)$. Therefore,

$$|\mathbf{v}| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}.$$

What is the geometric significance of scalar multiplication? If \mathbf{a} represents the vector \mathbf{v} in the sense that when it is slid to place its tail at the origin, the element of \mathbb{R}^n at its point is \mathbf{a} , what is $r\mathbf{v}$?

$$\begin{aligned} |r\mathbf{v}| &= \left(\sum_{k=1}^n (ra_k)^2 \right)^{1/2} = \left(\sum_{k=1}^n r^2 (a_k)^2 \right)^{1/2} \\ &= (r^2)^{1/2} \left(\sum_{k=1}^n a_k^2 \right)^{1/2} = |r| |\mathbf{v}|. \end{aligned}$$

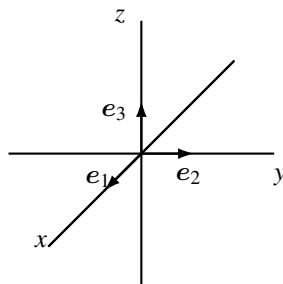
Thus the magnitude of $r\mathbf{v}$ equals $|r|$ times the magnitude of \mathbf{v} . If r is positive, then the vector represented by $r\mathbf{v}$ has the same direction as the vector \mathbf{v} because multiplying by the scalar r , only has the effect of scaling all the distances. Thus the unit distance along any coordinate axis now has length r and in this re-scaled system the vector is represented by \mathbf{a} . If $r < 0$ similar considerations apply except in this case all the a_i also change sign. From now on, \mathbf{a} will be referred to as a vector instead of an element of \mathbb{R}^n representing a vector as just described. The following picture illustrates the effect of scalar multiplication.



Note there are n special vectors which point along the coordinate axes. These are

$$\mathbf{e}_i \equiv (0, \dots, 0, 1, 0, \dots, 0)$$

where the 1 is in the i^{th} slot and there are zeros in all the other spaces. See the picture in the case of \mathbb{R}^3 .



The direction of e_i is referred to as the i^{th} direction. Given a vector $\mathbf{v} = (a_1, \dots, a_n)$, $a_i e_i$ is the i^{th} component of the vector. Thus $a_i e_i = (0, \dots, 0, a_i, 0, \dots, 0)$ and so this vector gives something possibly nonzero only in the i^{th} direction. Also, knowledge of the i^{th} component of the vector is equivalent to knowledge of the vector because it gives the entry in the i^{th} slot and for $\mathbf{v} = (a_1, \dots, a_n)$, $\mathbf{v} = \sum_{k=1}^n a_k e_k$.

What does addition of vectors mean physically? Suppose two forces are applied to some object. Each of these would be represented by a force vector and the two forces acting together would yield an overall force acting on the object which would also be a force vector known as the resultant. Suppose the two vectors are $\mathbf{a} = \sum_{k=1}^n a_k e_k$ and $\mathbf{b} = \sum_{k=1}^n b_k e_k$. Then the vector \mathbf{a} involves a component in the i^{th} direction, $a_i e_i$ while the component in the i^{th} direction of \mathbf{b} is $b_i e_i$. Then it seems physically reasonable that the resultant vector should have a component in the i^{th} direction equal to $(a_i + b_i) e_i$. This is exactly what is obtained when the vectors \mathbf{a} and \mathbf{b} are added.

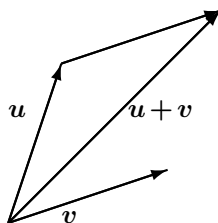
$$\mathbf{a} + \mathbf{b} = (a_1 + b_1, \dots, a_n + b_n) = \sum_{i=1}^n (a_i + b_i) e_i$$

Thus the addition of vectors according to the rules of addition in \mathbb{R}^n which were presented earlier, yields the appropriate vector which duplicates the cumulative effect of all the vectors in the sum.

What is the geometric significance of vector addition? Suppose \mathbf{u}, \mathbf{v} are vectors

$$\mathbf{u} = (u_1, \dots, u_n), \mathbf{v} = (v_1, \dots, v_n)$$

Then $\mathbf{u} + \mathbf{v} = (u_1 + v_1, \dots, u_n + v_n)$. How can one obtain this geometrically? Consider the directed line segment, $\overrightarrow{0\mathbf{u}}$ and then, starting at the end of this directed line segment, follow the directed line segment $\overrightarrow{\mathbf{u}(\mathbf{u} + \mathbf{v})}$ to its end $\mathbf{u} + \mathbf{v}$. In other words, place the vector \mathbf{u} in standard position with its base at the origin and then slide the vector \mathbf{v} till its base coincides with the point of \mathbf{u} . The point of this slid vector, determines $\mathbf{u} + \mathbf{v}$. To illustrate, see the following picture



Note the vector $u + v$ is the diagonal of a parallelogram determined from the two vectors u and v and that identifying $u + v$ with the directed diagonal of the parallelogram determined by the vectors u and v amounts to the same thing as the above procedure.

An item of notation should be mentioned here. In the case of \mathbb{R}^n where $n \leq 3$, it is standard notation to use i for e_1 , j for e_2 , and k for e_3 . Now here are some applications of vector addition to some problems.

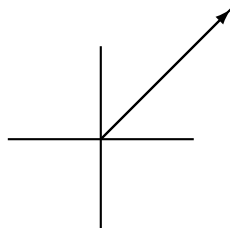
Example 5.8.4 *There are three ropes attached to a car and three people pull on these ropes. The first exerts a force of $2i + 3j - 2k$ Newtons, the second exerts a force of $3i + 5j + k$ Newtons and the third exerts a force of $5i - j + 2k$ Newtons. Find the total force in the direction of i .*

To find the total force add the vectors as described above. This gives $10i + 7j + k$ Newtons. Therefore, the force in the i direction is 10 Newtons.

As mentioned earlier, the Newton is a unit of force like pounds.

Example 5.8.5 *An airplane flies North East at 100 miles per hour. Write this as a vector.*

A picture of this situation follows.



The vector has length 100. Now using that vector as the hypotenuse of a right triangle having equal sides, the sides should be each of length $100/\sqrt{2}$. Therefore, the vector would be $100/\sqrt{2}i + 100/\sqrt{2}j$.

This example also motivates the concept of **velocity**.

Definition 5.8.6 *The **speed** of an object is a measure of how fast it is going. It is measured in units of length per unit time. For example, miles per hour, kilometers per minute, feet per second. The **velocity** is a vector having the speed as the magnitude but also specifying the direction.*

Thus the velocity vector in the above example is $100/\sqrt{2}i + 100/\sqrt{2}j$.

Example 5.8.7 *The velocity of an airplane is $100i + j + k$ measured in kilometers per hour and at a certain instant of time its position is $(1, 2, 1)$. Here imagine a Cartesian coordinate*

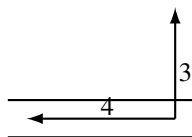
system in which the third component is altitude and the first and second components are measured on a line from West to East and a line from South to North. Find the position of this airplane one minute later.

Consider the vector $(1, 2, 1)$, is the initial position vector of the airplane. As it moves, the position vector changes. After one minute the airplane has moved in the i direction a distance of $100 \times \frac{1}{60} = \frac{5}{3}$ kilometer. In the j direction it has moved $\frac{1}{60}$ kilometer during this same time, while it moves $\frac{1}{60}$ kilometer in the k direction. Therefore, the new displacement vector for the airplane is

$$(1, 2, 1) + \left(\frac{5}{3}, \frac{1}{60}, \frac{1}{60} \right) = \left(\frac{8}{3}, \frac{121}{60}, \frac{121}{60} \right)$$

Example 5.8.8 A certain river is one half mile wide with a current flowing at 4 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 3 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?

Consider the following picture.



You should write these vectors in terms of components. The velocity of the swimmer in still water would be $3j$ while the velocity of the river would be $-4i$. Therefore, the velocity of the swimmer is $-4i + 3j$. Since the component of velocity in the direction across the river is 3, it follows the trip takes $1/6$ hour or 10 minutes. The speed at which he travels is $\sqrt{4^2 + 3^2} = 5$ miles per hour and so he travels $5 \times \frac{1}{6} = \frac{5}{6}$ miles. Now to find the distance downstream he finds himself, note that if x is this distance, x and $1/2$ are two legs of a right triangle whose hypotenuse equals $5/6$ miles. Therefore, by the Pythagorean theorem the distance downstream is

$$\sqrt{(5/6)^2 - (1/2)^2} = \frac{2}{3} \text{ miles.}$$

5.9 Exercises

1. The wind blows from the South at 40 kilometers per hour and an airplane which travels at 400 kilometers per hour in still air is heading East. Find the actual velocity of the airplane.
2. ↑ In the above problem, find the position of the airplane after two hours.
3. ↑ In the above problem, if the airplane is to travel due east, in what direction should it head in order to achieve this?
4. The wind blows from West to East at a speed of 50 miles per hour and an airplane which travels at 300 miles per hour in still air is heading North West. What is the velocity of the airplane relative to the ground? What is the component of this velocity in the direction North?

5. In the situation of Problem 4 how many degrees to the West of North should the airplane head in order to fly exactly North. What will be the speed of the airplane relative to the ground?
6. In the situation of 5 suppose the airplane uses 34 gallons of fuel every hour at that air speed and that it needs to fly North a distance of 600 miles. Will the airplane have enough fuel to arrive at its destination given that it has 63 gallons of fuel?
7. An airplane is flying due north at 150 miles per hour. A wind is pushing the airplane due east at 40 miles per hour. After 1 hour, the plane starts flying 30° East of North. Assuming the plane starts at $(0,0)$, where is it after 2 hours? Let North be the direction of the positive y axis and let East be the direction of the positive x axis.
8. City A is located at the origin while city B is located at $(300,500)$ where distances are in miles. An airplane flies at 250 miles per hour in still air. This airplane wants to fly from city A to city B but the wind is blowing in the direction of the positive y axis at a speed of 50 miles per hour. Find a unit vector such that if the plane heads in this direction, it will end up at city B having flown the shortest possible distance. How long will it take to get there?
9. A certain river is one half mile wide with a current flowing at 2 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 3 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?
10. A certain river is one half mile wide with a current flowing at 2 miles per hour from East to West. A man can swim at 3 miles per hour in still water. In what direction should he swim in order to travel directly across the river? What would the answer to this problem be if the river flowed at 3 miles per hour and the man could swim only at the rate of 2 miles per hour?
11. Three forces are applied to a point which does not move. Two of the forces are $2\mathbf{i} + \mathbf{j} + 3\mathbf{k}$ Newtons and $\mathbf{i} - 3\mathbf{j} + 2\mathbf{k}$ Newtons. Find the third force.
12. Three forces are applied to a point which does not move. Two of the forces are $\mathbf{i} + \mathbf{j} + 3\mathbf{k}$ Newtons and $\mathbf{i} - 3\mathbf{j} - 2\mathbf{k}$ Newtons. Find the third force.
13. The total force acting on an object is to be $2\mathbf{i} + \mathbf{j} + \mathbf{k}$ Newtons. A force of $-\mathbf{i} + \mathbf{j} + \mathbf{k}$ Newtons is being applied. What other force should be applied to achieve the desired total force?
14. The total force acting on an object is to be $\mathbf{i} + \mathbf{j} + 3\mathbf{k}$ Newtons. A force of $-\mathbf{i} - \mathbf{j} + \mathbf{k}$ Newtons is being applied. What other force should be applied to achieve the desired total force?
15. A bird flies from its nest 5 km. in the direction 60° north of east where it stops to rest on a tree. It then flies 10 km. in the direction due southeast and lands atop a telephone pole. Place an xy coordinate system so that the origin is the bird's nest, and the positive x axis points east and the positive y axis points north. Find the displacement vector from the nest to the telephone pole.

16. A car is stuck in the mud. There is a cable stretched tightly from this car to a tree which is 20 feet long. A person grasps the cable in the middle and pulls with a force of 100 pounds perpendicular to the stretched cable. The center of the cable moves two feet and remains still. What is the tension in the cable? The tension in the cable is the force exerted on this point by the part of the cable nearer the car as well as the force exerted on this point by the part of the cable nearer the tree.

Chapter 6

Vector Products

6.1 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 6.1.1 Let \mathbf{a}, \mathbf{b} be two vectors in \mathbb{R}^n define $\mathbf{a} \cdot \mathbf{b}$ as

$$\mathbf{a} \cdot \mathbf{b} \equiv \sum_{k=1}^n a_k b_k.$$

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 6.1.2 The dot product satisfies the following properties.

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} \tag{6.1}$$

$$\mathbf{a} \cdot \mathbf{a} \geq 0 \text{ and equals zero if and only if } \mathbf{a} = \mathbf{0} \tag{6.2}$$

$$(\alpha \mathbf{a} + \beta \mathbf{b}) \cdot \mathbf{c} = \alpha (\mathbf{a} \cdot \mathbf{c}) + \beta (\mathbf{b} \cdot \mathbf{c}) \tag{6.3}$$

$$\mathbf{c} \cdot (\alpha \mathbf{a} + \beta \mathbf{b}) = \alpha (\mathbf{c} \cdot \mathbf{a}) + \beta (\mathbf{c} \cdot \mathbf{b}) \tag{6.4}$$

$$|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a} \tag{6.5}$$

You should verify these properties. Also be sure you understand that 6.4 follows from the first three and is therefore redundant. It is listed here for the sake of convenience.

Example 6.1.3 Find $(1, 2, 0, -1) \cdot (0, 1, 2, 3)$.

This equals $0 + 2 + 0 + -3 = -1$.

Example 6.1.4 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 a fundamental inequality known as the **Cauchy Schwarz 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 6.1.5 *The dot product satisfies the inequality*

$$|\mathbf{a} \cdot \mathbf{b}| \leq |\mathbf{a}| |\mathbf{b}|. \quad (6.6)$$

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 6.6 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 6.2, $f(t) \geq 0$ for all $t \in \mathbb{R}$. Also from 6.3, 6.4, 6.1, and 6.5

$$\begin{aligned} f(t) &= \mathbf{a} \cdot (\mathbf{a} + t\mathbf{b}) + t\mathbf{b} \cdot (\mathbf{a} + t\mathbf{b}) \\ &= \mathbf{a} \cdot \mathbf{a} + t(\mathbf{a} \cdot \mathbf{b}) + t\mathbf{b} \cdot \mathbf{a} + t^2\mathbf{b} \cdot \mathbf{b} \\ &= |\mathbf{a}|^2 + 2t(\mathbf{a} \cdot \mathbf{b}) + |\mathbf{b}|^2 t^2. \end{aligned}$$

Now

$$\begin{aligned} f(t) &= |\mathbf{b}|^2 \left(t^2 + 2t \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} + \frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} \right) \\ &= |\mathbf{b}|^2 \left(t^2 + 2t \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} + \left(\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 - \left(\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 + \frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} \right) \\ &= |\mathbf{b}|^2 \left(\left(t + \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 + \left(\frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} - \left(\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 \right) \right) \geq 0 \end{aligned}$$

for all $t \in \mathbb{R}$. In particular $f(t) \geq 0$ when $t = -(\mathbf{a} \cdot \mathbf{b} / |\mathbf{b}|^2)$ which implies

$$\frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} - \left(\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|^2} \right)^2 \geq 0. \quad (6.7)$$

Multiplying both sides by $|\mathbf{b}|^4$,

$$|\mathbf{a}|^2 |\mathbf{b}|^2 \geq (\mathbf{a} \cdot \mathbf{b})^2$$

which yields 6.6.

From Theorem 6.1.5, equality holds in 6.6 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 equality is obtained in 6.6 so it can be assumed both vectors are non zero and that equality is obtained in 6.7. This implies that $f(t) = 0$ when $t = -(\mathbf{a} \cdot \mathbf{b} / |\mathbf{b}|^2)$ and so from 6.2, it follows that for this value of t , $\mathbf{a} + t\mathbf{b} = \mathbf{0}$ showing $\mathbf{a} = -t\mathbf{b}$. ■

You should note that the entire argument was based only on the properties of the dot product listed in 6.1 - 6.5. This means that whenever something satisfies these properties, the Cauchy Schwartz inequality holds. There are many other instances of these properties besides vectors in \mathbb{R}^n .

The Cauchy Schwartz inequality allows a proof of the **triangle inequality** for distances in \mathbb{R}^n in much the same way as the triangle inequality for the absolute value.

Theorem 6.1.6 (Triangle inequality) For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$

$$|\mathbf{a} + \mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}| \quad (6.8)$$

and equality holds if and only if one of the vectors is a nonnegative scalar multiple of the other. Also

$$||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}| \quad (6.9)$$

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 6.8.

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 6.1.5 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,

$$|\mathbf{a}| - |\mathbf{b}| \leq |\mathbf{a} - \mathbf{b}| \quad (6.10)$$

Similarly,

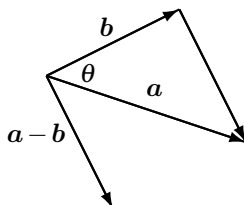
$$|\mathbf{b}| - |\mathbf{a}| \leq |\mathbf{b} - \mathbf{a}| = |\mathbf{a} - \mathbf{b}|. \quad (6.11)$$

It follows from 6.10 and 6.11 that 6.9 holds. This is because $||\mathbf{a}| - |\mathbf{b}||$ equals the left side of either 6.10 or 6.11 and either way, $||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}|$. ■

6.2 The Geometric Significance Of The Dot Product

6.2.1 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,

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta. \quad (6.12)$$

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 6.2.1 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 6.12 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 6.2.2 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 6.12

$$\mathbf{u} \cdot \mathbf{v} = 3 \times 4 \times \cos(60^\circ) = 3 \times 4 \times 1/2 = 6.$$

Observation 6.2.3 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 6.2.4 Determine whether the two vectors $2\mathbf{i} + \mathbf{j} - \mathbf{k}$ and $\mathbf{i} + 3\mathbf{j} + 5\mathbf{k}$ are perpendicular.

When you take this dot product you get $2 + 3 - 5 = 0$ and so these two are indeed perpendicular.

Definition 6.2.5 When two lines intersect, the angle between the two lines is the smaller of the two angles determined.

Example 6.2.6 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

$$\cos \theta = \frac{-6}{14} = \frac{-3}{7}. \quad (6.13)$$

We don't want this angle because it is obtuse. The angle desired is the acute angle given by

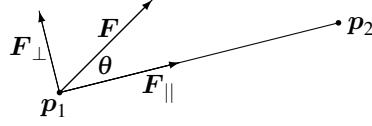
$$\cos \theta = \frac{3}{7}.$$

It is obtained by using replacing one of the direction vectors with -1 times it.

6.2.2 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 6.2.7 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 6.2.8 Let \mathbf{F} and \mathbf{D} be nonzero vectors. Then there exist unique vectors \mathbf{F}_{\parallel} and \mathbf{F}_{\perp} such that

$$\mathbf{F} = \mathbf{F}_{\parallel} + \mathbf{F}_{\perp} \tag{6.14}$$

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 6.14 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

$$\mathbf{F}_{\perp} \cdot \mathbf{D} = \mathbf{F} \cdot \mathbf{D} - \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D} \cdot \mathbf{D} = \mathbf{F} \cdot \mathbf{D} - \mathbf{F} \cdot \mathbf{D} = 0.$$

■

Example 6.2.9 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 6.2.10 Find $\text{proj}_{\mathbf{u}}(\mathbf{v})$ 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 6.2.8, 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 6.2.11 Suppose \mathbf{a} , and \mathbf{b} are vectors and $\mathbf{b}_{\perp} = \mathbf{b} - \text{proj}_{\mathbf{a}}(\mathbf{b})$. What is the magnitude of \mathbf{b}_{\perp} in terms of the included angle?

$$\begin{aligned}
|\mathbf{b}_\perp|^2 &= (\mathbf{b} - \text{proj}_\mathbf{a}(\mathbf{b})) \cdot (\mathbf{b} - \text{proj}_\mathbf{a}(\mathbf{b})) = \left(\mathbf{b} - \frac{\mathbf{b} \cdot \mathbf{a}}{|\mathbf{a}|^2} \mathbf{a} \right) \cdot \left(\mathbf{b} - \frac{\mathbf{b} \cdot \mathbf{a}}{|\mathbf{a}|^2} \mathbf{a} \right) \\
&= |\mathbf{b}|^2 - 2 \frac{(\mathbf{b} \cdot \mathbf{a})^2}{|\mathbf{a}|^2} + \left(\frac{\mathbf{b} \cdot \mathbf{a}}{|\mathbf{a}|^2} \right)^2 |\mathbf{a}|^2 = |\mathbf{b}|^2 \left(1 - \frac{(\mathbf{b} \cdot \mathbf{a})^2}{|\mathbf{a}|^2 |\mathbf{b}|^2} \right) \\
&= |\mathbf{b}|^2 (1 - \cos^2 \theta) = |\mathbf{b}|^2 \sin^2(\theta)
\end{aligned}$$

where θ is the included angle between \mathbf{a} and \mathbf{b} which is less than π radians. Therefore, taking square roots, $|\mathbf{b}_\perp| = |\mathbf{b}| \sin \theta$.

6.2.3 The Dot Product And Distance In \mathbb{C}^n

It is necessary to give a generalization of the dot product for vectors in \mathbb{C}^n . This definition reduces to the usual one in the case the components of the vector are real.

Definition 6.2.12 Let $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$. Thus $\mathbf{x} = (x_1, \dots, x_n)$ where each $x_k \in \mathbb{C}$ and a similar formula holding for \mathbf{y} . Then the dot product of these two vectors is defined to be

$$\mathbf{x} \cdot \mathbf{y} \equiv \sum_j x_j \overline{y_j} \equiv x_1 \overline{y_1} + \dots + x_n \overline{y_n}.$$

Notice how you put the conjugate on the entries of the vector \mathbf{y} . It makes no difference if the vectors happen to be real vectors but with complex vectors you must do it this way. The reason for this is that when you take the dot product of a vector with itself, you want to get the square of the length of the vector, a positive number. Placing the conjugate on the components of \mathbf{y} in the above definition assures this will take place. Thus

$$\mathbf{x} \cdot \mathbf{x} = \sum_j x_j \overline{x_j} = \sum_j |x_j|^2 \geq 0.$$

If you didn't place a conjugate as in the above definition, things wouldn't work out correctly. For example,

$$(1+i)^2 + 2^2 = 4 + 2i$$

and this is not a positive number.

The following properties of the dot product follow immediately from the definition and you should verify each of them.

Properties of the dot product:

1. $\mathbf{u} \cdot \mathbf{v} = \overline{\mathbf{v} \cdot \mathbf{u}}$.
2. If a, b are numbers and $\mathbf{u}, \mathbf{v}, \mathbf{z}$ are vectors then $(a\mathbf{u} + b\mathbf{v}) \cdot \mathbf{z} = a(\mathbf{u} \cdot \mathbf{z}) + b(\mathbf{v} \cdot \mathbf{z})$.
3. $\mathbf{u} \cdot \mathbf{u} \geq 0$ and it equals 0 if and only if $\mathbf{u} = \mathbf{0}$.

The norm is defined in the usual way.

Definition 6.2.13 For $\mathbf{x} \in \mathbb{C}^n$,

$$|\mathbf{x}| \equiv \left(\sum_{k=1}^n |x_k|^2 \right)^{1/2} = (\mathbf{x} \cdot \mathbf{x})^{1/2}$$

As in the case of \mathbb{R}^n , the **Cauchy Schwarz inequality** is of fundamental importance. First here is a simple lemma.

Lemma 6.2.14 *If $z \in \mathbb{C}$ there exists $\theta \in \mathbb{C}$ such that $\theta z = |z|$ and $|\theta| = 1$.*

Proof: Let $\theta = 1$ if $z = 0$ and otherwise, let $\theta = \frac{\bar{z}}{|z|}$. Recall that for $z = x + iy$, $\bar{z} = x - iy$ and $\bar{z}z = |z|^2$. ■

Theorem 6.2.15 (Cauchy Schwarz) *The following inequality holds for x_i and $y_i \in \mathbb{C}$.*

$$|(x \cdot y)| = \left| \sum_{i=1}^n x_i \bar{y}_i \right| \leq \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2} \left(\sum_{i=1}^n |y_i|^2 \right)^{1/2} = |x| |y| \quad (6.15)$$

Proof: Let $\theta \in \mathbb{C}$ such that $|\theta| = 1$ and $\theta(x \cdot y) = |x \cdot y|$. Then from the properties of the dot product,

$$\begin{aligned} (x + t\bar{\theta}y) \cdot (x + t\bar{\theta}y) &= (x \cdot x) + t^2 \bar{\theta}\theta(y \cdot y) + t(x \cdot \bar{\theta}y) + t(\bar{\theta}y \cdot x) \\ &= (x \cdot x) + t^2(y \cdot y) + t\theta(x \cdot y) + t\overline{(x \cdot \bar{\theta}y)} \\ &= (x \cdot x) + t^2(y \cdot y) + t\theta(x \cdot y) + t\theta(x \cdot y) \\ &= |x|^2 + 2t|x \cdot y| + t^2|y|^2 \geq 0 \end{aligned}$$

If $|y| = 0$, this can only happen if $x \cdot y = 0$ and so the inequality holds. If $|y| \neq 0$, then you have a parabola which opens up and has at most one real zero. Therefore, by the quadratic formula,

$$4|x \cdot y|^2 - 4|x|^2|y|^2 \leq 0$$

which yields the Cauchy Schwarz inequality. ■

By analogy to the case of \mathbb{R}^n , length or magnitude of vectors in \mathbb{C}^n can be defined.

Definition 6.2.16 *Let $z \in \mathbb{C}^n$. Then $|z| \equiv (z \cdot z)^{1/2}$.*

Theorem 6.2.17 *For length defined in Definition 6.2.16, the following hold.*

$$|z| \geq 0 \text{ and } |z| = 0 \text{ if and only if } z = 0 \quad (6.16)$$

$$\text{If } \alpha \text{ is a scalar, } |\alpha z| = |\alpha| |z| \quad (6.17)$$

$$|z + w| \leq |z| + |w|. \quad (6.18)$$

Proof: The first two claims are left as exercises. To establish the third, you use the same argument which was used in \mathbb{R}^n .

$$\begin{aligned} |z + w|^2 &= (z + w, z + w) = z \cdot z + w \cdot w + w \cdot z + z \cdot w \\ &= |z|^2 + |w|^2 + 2\operatorname{Re} w \cdot z \leq |z|^2 + |w|^2 + 2|w \cdot z| \\ &\leq |z|^2 + |w|^2 + 2|w||z| = (|z| + |w|)^2. \end{aligned}$$

All other considerations such as open and closed sets and the like are identical in this more general context with the corresponding definition in \mathbb{R}^n . The main difference is that here the scalars are complex numbers. ■

Definition 6.2.18 Suppose you have a vector space, V and for $\mathbf{z}, \mathbf{w} \in V$ and α a scalar a norm is a way of measuring distance or magnitude which satisfies the properties 6.16 - 6.18. Thus a norm is something which does the following.

$$\|\mathbf{z}\| \geq 0 \text{ and } \|\mathbf{z}\| = 0 \text{ if and only if } \mathbf{z} = \mathbf{0} \quad (6.19)$$

$$\text{If } \alpha \text{ is a scalar, } \|\alpha\mathbf{z}\| = |\alpha| \|\mathbf{z}\| \quad (6.20)$$

$$\|\mathbf{z} + \mathbf{w}\| \leq \|\mathbf{z}\| + \|\mathbf{w}\|. \quad (6.21)$$

Here is understood that for all $\mathbf{z} \in V, \|\mathbf{z}\| \in [0, \infty)$.

6.3 Exercises

- Find $(1, 2, 3, 4) \cdot (2, 0, 1, 3)$.
- Use formula 6.12 to verify the Cauchy Schwartz 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 Cauchy Schwarz 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 the angle between the vectors $\mathbf{i} - 2\mathbf{j} + \mathbf{k}$ and $\mathbf{i} + 2\mathbf{j} - 7\mathbf{k}$.
- 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 $\mathbf{v} = (1, 2, -2)$ and $\mathbf{u} = (1, 0, 3)$.
- 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?
- A large dog drags a sled for 300 feet along the ground by pulling on a rope which is 45 degrees from the horizontal with a force of 20 pounds. How much work does this force do?

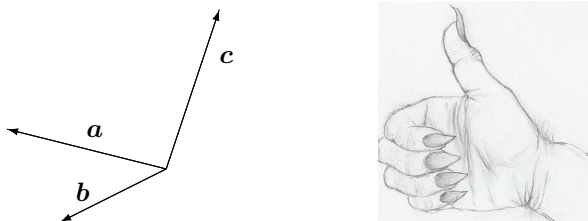
14. How much work in Newton meters does it take to slide a crate 20 meters along a loading dock by pulling on it with a 200 Newton force at an angle of 30° from the horizontal?
15. An object moves 10 meters in the direction of \mathbf{j} . There are two forces acting on this object $\mathbf{F}_1 = \mathbf{i} + \mathbf{j} + 2\mathbf{k}$, 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?
16. An object moves 10 meters in the direction of $\mathbf{j} + \mathbf{i}$. There are two forces acting on this object $\mathbf{F}_1 = \mathbf{i} + 2\mathbf{j} + 2\mathbf{k}$, 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?
17. An object moves 20 meters in the direction of $\mathbf{k} + \mathbf{j}$. There are two forces acting on this object $\mathbf{F}_1 = \mathbf{i} + \mathbf{j} + 2\mathbf{k}$, and $\mathbf{F}_2 = \mathbf{i} + 2\mathbf{j} - 6\mathbf{k}$. 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.
18. If \mathbf{a}, \mathbf{b} , and \mathbf{c} are vectors. Show that $(\mathbf{b} + \mathbf{c})_\perp = \mathbf{b}_\perp + \mathbf{c}_\perp$ where $\mathbf{b}_\perp = \mathbf{b} - \text{proj}_{\mathbf{a}}(\mathbf{b})$.
19. In the discussion of the reflecting mirror which directs all rays to a particular point $(0, p)$. Show that for any choice of positive C this point is the focus of the parabola and the directrix is $y = p - \frac{1}{C}$.
20. Suppose you wanted to make a solar powered oven to cook food. Are there reasons for using a mirror which is not parabolic? Also describe how you would design a good flash light with a beam which does not spread out too quickly.
21. Show that $(\mathbf{a} \cdot \mathbf{b}) = \frac{1}{4} [|\mathbf{a} + \mathbf{b}|^2 - |\mathbf{a} - \mathbf{b}|^2]$.
22. Prove from the axioms of the dot product the parallelogram identity which is the following: $|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2$.
23. Suppose f, g are two continuous functions defined on $[0, 1]$. Define the inner product $(f \cdot g) = \int_0^1 f(x)g(x)dx$. Show this dot product satisfies conditions 6.1 - 6.5. Explain why the Cauchy Schwarz inequality continues to hold in this context and state the Cauchy Schwarz inequality in terms of integrals.

6.4 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.

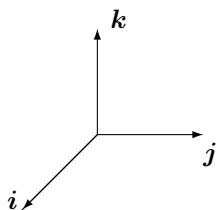
Definition 6.4.1 *Three vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ form a right handed system if when you extend the fingers of your right hand along the vector \mathbf{a} and close them in the direction of \mathbf{b} , the thumb points roughly in the direction of \mathbf{c} .*

For an example of a right handed system of vectors, see the following picture.



In this picture the vector 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 c would need to point in the opposite direction as it would for a right hand system.

From now on, the vectors i, j, k will always form a right handed system. To repeat, if you extend the fingers of our right hand along i and close them in the direction j , the thumb points in the direction of k .

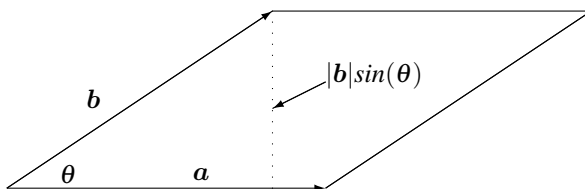


The following is the geometric description of the cross product. It gives both the direction and the magnitude and therefore specifies the vector.

Definition 6.4.2 Let a and b be two vectors in \mathbb{R}^3 . Then $a \times b$ is defined by the following two rules.

1. $|a \times b| = |a||b|\sin\theta$ where θ is the included angle.
2. $a \times b \cdot a = 0$, $a \times b \cdot b = 0$, and $a, b, a \times b$ forms a right hand system.

Note that $|a \times b|$ is the area of the parallelogram spanned by a and b .



The cross product satisfies the following properties.

$$a \times b = -(b \times a), \quad a \times a = 0, \quad (6.22)$$

For α a scalar,

$$(\alpha \mathbf{a}) \times \mathbf{b} = \alpha (\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (\alpha \mathbf{b}), \quad (6.23)$$

For \mathbf{a}, \mathbf{b} , and \mathbf{c} vectors, one obtains the distributive laws,

$$\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}, \quad (6.24)$$

$$(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}. \quad (6.25)$$

Formula 6.22 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 6.23 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 6.23. 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 6.22,

$$(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = -\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = -(\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}) = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}.$$

A proof of the distributive law is given in a later section for those who are interested. 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 6.4.3 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

$$\mathbf{a} \times \mathbf{b} = (a_2 b_3 - a_3 b_2) \mathbf{i} + (a_3 b_1 - a_1 b_3) \mathbf{j} + (a_1 b_2 - a_2 b_1) \mathbf{k}. \quad (6.26)$$

Proof: From the above table and the properties of the cross product listed,

$$\begin{aligned} & (a_1 \mathbf{i} + a_2 \mathbf{j} + a_3 \mathbf{k}) \times (b_1 \mathbf{i} + b_2 \mathbf{j} + b_3 \mathbf{k}) = \\ & a_1 b_2 \mathbf{i} \times \mathbf{j} + a_1 b_3 \mathbf{i} \times \mathbf{k} + a_2 b_1 \mathbf{j} \times \mathbf{i} + a_2 b_3 \mathbf{j} \times \mathbf{k} + \\ & \quad + a_3 b_1 \mathbf{k} \times \mathbf{i} + a_3 b_2 \mathbf{k} \times \mathbf{j} \\ & = a_1 b_2 \mathbf{k} - a_1 b_3 \mathbf{j} - a_2 b_1 \mathbf{k} + a_2 b_3 \mathbf{i} + a_3 b_1 \mathbf{j} - a_3 b_2 \mathbf{i} \\ & = (a_2 b_3 - a_3 b_2) \mathbf{i} + (a_3 b_1 - a_1 b_3) \mathbf{j} + (a_1 b_2 - a_2 b_1) \mathbf{k} \end{aligned} \quad (6.27)$$

■

It is probably impossible for most people to remember 6.26. Fortunately, there is a somewhat easier way to remember it.

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \quad (6.28)$$

where you expand the determinant along the top row. This yields

$$(a_2b_3 - a_3b_2)\mathbf{i} - (a_1b_3 - a_3b_1)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k} \quad (6.29)$$

which is the same as 6.27. If you have not seen determinants, it doesn't matter all you need here is how to evaluate 2×2 and 3×3 determinants. First consider 2×2 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 will be discussed more later.

Example 6.4.4 Find $(\mathbf{i} - \mathbf{j} + 2\mathbf{k}) \times (3\mathbf{i} - 2\mathbf{j} + \mathbf{k})$.

Use 6.28 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 6.4.5 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 6.4.4.

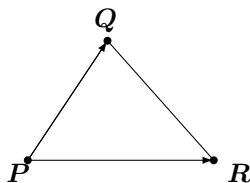
From Example 6.4.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example 6.4.4. Thus the area is $\sqrt{9 + 25 + 1} = \sqrt{35}$.

Example 6.4.6 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 6.4.7 In general, if you have three points (vectors) in \mathbb{R}^3 , $\mathbf{P}, \mathbf{Q}, \mathbf{R}$ the area of the triangle is given by

$$\frac{1}{2} |(\mathbf{Q} - \mathbf{P}) \times (\mathbf{R} - \mathbf{P})|.$$



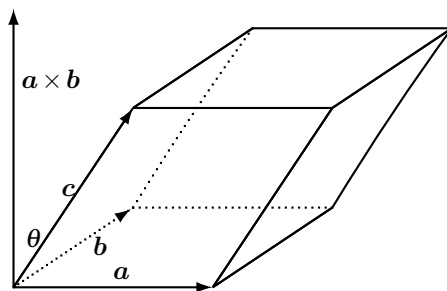
6.4.1 The Box Product

Definition 6.4.8 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 6.4.9 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 6.4.10 *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. ■

6.5 Proof of the distributive law

Here is another proof of the distributive law for the cross product. 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 $\mathbf{x} = \mathbf{a} \times (\mathbf{b} + \mathbf{c}) - (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c})$ and this shows that the following holds: $\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}$ and this proves the distributive law for the cross product another way.

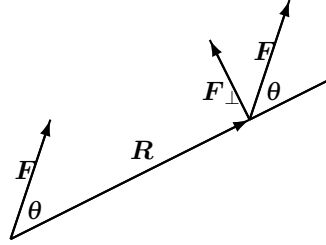
Observation 6.5.1 *Suppose you have three vectors, $\mathbf{u} = (a, b, c)$, $\mathbf{v} = (d, e, f)$, and $\mathbf{w} = (g, h, i)$. Then $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}$ is given by the following.*

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} \times \mathbf{w} &= (a, b, c) \cdot \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ d & e & f \\ g & h & i \end{vmatrix} \\ &= a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= \det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}. \end{aligned}$$

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.

6.5.1 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 6.5.2 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 6.5.3 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 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.

6.5.2 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 6.5.4 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

$$\sum_{k=1}^p (\mathbf{R}_k - \mathbf{R}_0) \times gm_k \mathbf{u} = \mathbf{0}$$

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

$$\left(\sum_{k=1}^p \mathbf{R}_k gm_k - \mathbf{R}_0 \sum_{k=1}^p gm_k \right) \times \mathbf{u} = \mathbf{0} \quad (6.30)$$

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

$$\sum_{k=1}^p \mathbf{R}_k gm_k - \mathbf{R}_0 \sum_{k=1}^p gm_k = \mathbf{0}.$$

dividing by g , and then by $\sum_{k=1}^p m_k$,

$$\mathbf{R}_0 = \frac{\sum_{k=1}^p \mathbf{R}_k m_k}{\sum_{k=1}^p m_k}. \quad (6.31)$$

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 calculus.

Example 6.5.5 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 6.31

$$\mathbf{R}_0 = \frac{5(2\mathbf{i} + 3\mathbf{j} + \mathbf{k}) + 6(\mathbf{i} - 3\mathbf{j} + 2\mathbf{k}) + 3(2\mathbf{i} - \mathbf{j} + 3\mathbf{k})}{5 + 6 + 3} = \frac{11}{7}\mathbf{i} - \frac{3}{7}\mathbf{j} + \frac{13}{7}\mathbf{k}$$

6.5.3 Angular Velocity

Definition 6.5.6 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 upper hemisphere but clockwise in the lower hemisphere? To quantify these things, you will need the concept of an angular velocity vector. Details are presented later for interesting examples. Here is a simple example. In the above 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 but it is not fixed.

Example 6.5.7 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. A complete analysis of this is given later.

Example 6.5.8 A wheel rotates counter clockwise about the vector $\mathbf{i} + \mathbf{j} + \mathbf{k}$ at 60 revolutions per minute exactly as in Example 6.5.7. 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. 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. 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.

6.6 Vector Identities And Notation

To begin with consider $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ and it is desired to simplify this quantity. It turns out this is an important quantity which 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

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = \mathbf{v}(\mathbf{u} \cdot \mathbf{w}) - \mathbf{w}(\mathbf{u} \cdot \mathbf{v}). \quad (6.32)$$

A related formula is

$$\begin{aligned}
 (\mathbf{u} \times \mathbf{v}) \times \mathbf{w} &= -[\mathbf{w} \times (\mathbf{u} \times \mathbf{v})] \\
 &= -[\mathbf{u}(\mathbf{w} \cdot \mathbf{v}) - \mathbf{v}(\mathbf{w} \cdot \mathbf{u})] \\
 &= \mathbf{v}(\mathbf{w} \cdot \mathbf{u}) - \mathbf{u}(\mathbf{w} \cdot \mathbf{v}).
 \end{aligned} \tag{6.33}$$

This derivation is simply wretched and it does nothing for other identities which may arise in applications. Actually, the above two formulas, 6.32 and 6.33 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 computation. It is a vector identity discovering machine and it is this which is the main topic in what follows.

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 6.6.1 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 6.6.2 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 6.6.3 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 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 6.6.4 *Let \mathbf{u}, \mathbf{v} be vectors in \mathbb{R}^n where the Cartesian coordinates of \mathbf{u} are (u_1, \dots, u_n) and the Cartesian coordinates of \mathbf{v} are (v_1, \dots, v_n) . 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 that 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 6.6.5 *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

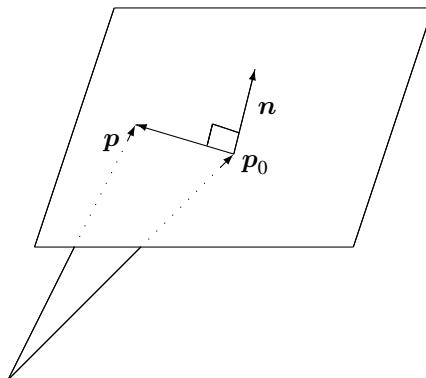
$$(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} = (\mathbf{u} \cdot \mathbf{w}) \mathbf{v} - (\mathbf{v} \cdot \mathbf{w}) \mathbf{u}.$$

6.7 Planes

You have an idea of what a plane is already. It is the span of some vectors. However, it can also be considered geometrically in terms of a 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 6.7.1 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 6.7.2 $2x + 4y - 5z = 11$ 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. The thing which makes perfect sense is the angle between two vectors. The angle between two planes requires some definition. If you think about it geometrically, you could imagine infinitely many angles between two lines both of which lie in one of the planes and which intersect at a point on a line of intersection of two planes.

Definition 6.7.3 Suppose two planes intersect in a line. The angle between the planes is defined to be the angle which is no more than $\pi/2$ between normal vectors to the respective planes.

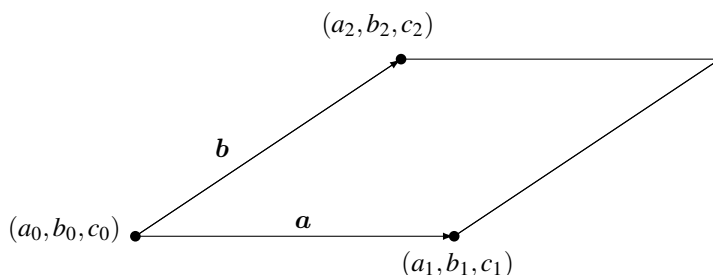
Example 6.7.4 Find the angle between the two planes $x + 2y - z = 6$ and $3x + 2y - z = 7$.

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 6.7.5 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.

Example 6.7.6 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 $y + 3z = 5$.

Proposition 6.7.7 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 6.7.8 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

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

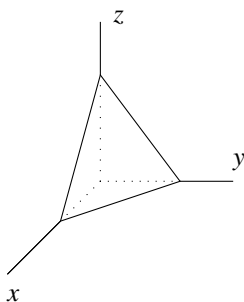
Example 6.7.9 Find the equation of the plane which contains the two lines, given by the following parametric expressions in which $t \in \mathbb{R}$.

$$(2t, 1 + t, 1 + 2t) = (x, y, z), \quad (2t + 2, 1, 3 + 2t) = (x, y, z)$$

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 6.7.10 Sketch the plane which has 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 6.7.11 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.

6.8 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 6.30 holds for $\mathbf{u} = \mathbf{i}, \mathbf{j}, \mathbf{k}$, show that this implies 6.30 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. 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 $(1, 0, 3)$, $(4, 1, 0)$ and $(-3, 1, 1)$.

9. 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?
10. Find the area of the parallelogram determined by the vectors $(1, 2, 3)$ and $(3, -2, 1)$.
11. Find the area of the parallelogram determined by the vectors $(1, 0, 3)$ and $(4, -2, 1)$.
12. Find the area of the parallelogram determined by the vectors $(1, -2, 2)$ and $(3, 1, 1)$.
13. Find the volume of the parallelepiped determined by the vectors $i - 7j - 5k$, $i - 2j - 6k$, $3i + 2j + 3k$.
14. Find the volume of the parallelepiped determined by the vectors $i + j - 5k$, $i + 5j - 6k$, $3i + j + 3k$.
15. Find the volume of the parallelepiped determined by the vectors $i + 6j + 5k$, $i + 5j - 6k$, $3i + j + k$.
16. Suppose a , b , and 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?
17. What does it mean geometrically if the box product of three vectors gives zero?
18. Find the equation of the plane through the three points $(1, 2, 3)$, $(2, -3, 1)$, $(1, 1, 7)$.
19. It is desired to find an equation of a plane containing the two vectors a and b and the point 0 . Using Problem 17, 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

$$x \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} - y \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} + z \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} = 0$$

20. 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

$$(a \times b) \cdot c = a \cdot (b \times c)$$

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.

21. Is $a \times (b \times c) = (a \times b) \times c$? What is the meaning of $a \times b \times c$? Explain. **Hint:** Try $(i \times j) \times j$.

22. 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 from the coordinate description of the cross product 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.

23. Discover a vector identity for $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$.
24. Discover a vector identity for $(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{z} \times \mathbf{w})$.
25. Discover a vector identity for $(\mathbf{u} \times \mathbf{v}) \times (\mathbf{z} \times \mathbf{w})$ in terms of box products.
26. Simplify $(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{v} \times \mathbf{w}) \times (\mathbf{w} \times \mathbf{z})$.
27. Simplify $|\mathbf{u} \times \mathbf{v}|^2 + (\mathbf{u} \cdot \mathbf{v})^2 - |\mathbf{u}|^2 |\mathbf{v}|^2$.
28. Prove that $\epsilon_{ijk} \epsilon_{ijr} = 2\delta_{kr}$.
29. 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 $\det(A) = \epsilon_{ijk} u_i v_j w_k$.
30. If A is a 3×3 matrix, show $\epsilon_{rps} \det(A) = \epsilon_{ijk} A_{ri} A_{pj} A_{sk}$.
31. Suppose A is a 3×3 matrix and $\det(A) \neq 0$. Show using 30 and 28 that

$$(A^{-1})_{ks} = \frac{1}{2\det(A)} \epsilon_{rps} \epsilon_{ijk} A_{pj} A_{ri}.$$

32. When you have a rotating rigid body with angular velocity vector $\boldsymbol{\Omega}$ then the velocity, \mathbf{u}' is given by $\mathbf{u}' = \boldsymbol{\Omega} \times \mathbf{u}$. 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 $\boldsymbol{\Omega}$ a constant vector

$$\mathbf{u}'' = \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{u}).$$

It turns out this is the centripetal acceleration. Note how it involves cross products.

33. 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) $(1, 0, 1), (2, 0, 1), (0, 1, 1)$

34. 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 7

Systems Of Equations

7.1 Systems Of Equations, Algebraic Procedures

7.1.1 Elementary Operations

Consider the following example.

Example 7.1.1 Find x and y such that

$$x + y = 7 \text{ and } 2x - y = 8. \quad (7.1)$$

The set of ordered pairs, (x, y) which solve both equations is called the **solution set**.

You can verify that $(x, y) = (5, 2)$ is a solution to the above system. The interesting question is this: If you were not given this information to verify, how could you determine the solution? You can do this by using the following basic operations on the equations, none of which change the set of solutions of the system of equations.

Definition 7.1.2 *Elementary operations* are those operations consisting of the following.

1. Interchange the order in which the equations are listed.
2. Multiply any equation by a nonzero number.
3. Replace any equation with itself added to a multiple of another equation.

Example 7.1.3 To illustrate the third of these operations on this particular system, consider the following.

$$\begin{aligned} x + y &= 7 \\ 2x - y &= 8 \end{aligned}$$

The system has the same solution set as the system

$$\begin{aligned} x + y &= 7 \\ -3y &= -6 \end{aligned}.$$

To obtain the second system, take the second equation of the first system and add -2 times the first equation to obtain

$$-3y = -6.$$

Now, this clearly shows that $y = 2$ and so it follows from the other equation that $x + 2 = 7$ and so $x = 5$.

Of course a linear system may involve many equations and many variables. The solution set is still the collection of solutions to the equations. In every case, the above operations of Definition 7.1.2 do not change the set of solutions to the system of linear equations.

Theorem 7.1.4 Suppose you have two equations, involving the variables,

$$(x_1, \dots, x_n)$$

$$E_1 = f_1, E_2 = f_2 \tag{7.2}$$

where E_1 and E_2 are expressions involving the variables and f_1 and f_2 are constants. (In the above example there are only two variables, x and y and $E_1 = x + y$ while $E_2 = 2x - y$.) Then the system $E_1 = f_1, E_2 = f_2$ has the same solution set as

$$E_1 = f_1, E_2 + aE_1 = f_2 + af_1. \tag{7.3}$$

Also the system $E_1 = f_1, E_2 = f_2$ has the same solutions as the system, $E_2 = f_2, E_1 = f_1$. The system $E_1 = f_1, E_2 = f_2$ has the same solution as the system $E_1 = f_1, aE_2 = af_2$ provided $a \neq 0$.

Proof: If (x_1, \dots, x_n) solves $E_1 = f_1, E_2 = f_2$ then it solves the first equation in $E_1 = f_1, E_2 + aE_1 = f_2 + af_1$. Also, it satisfies $aE_1 = af_1$ and so, since it also solves $E_2 = f_2$ it must solve $E_2 + aE_1 = f_2 + af_1$. Therefore, if (x_1, \dots, x_n) solves $E_1 = f_1, E_2 = f_2$ it must also solve $E_2 + aE_1 = f_2 + af_1$. On the other hand, if it solves the system $E_1 = f_1$ and $E_2 + aE_1 = f_2 + af_1$, then $aE_1 = af_1$ and so you can subtract these equal quantities from both sides of $E_2 + aE_1 = f_2 + af_1$ to obtain $E_2 = f_2$ showing that it satisfies $E_1 = f_1, E_2 = f_2$.

The second assertion of the theorem which says that the system $E_1 = f_1, E_2 = f_2$ has the same solution as the system, $E_2 = f_2, E_1 = f_1$ is seen to be true because it involves nothing more than listing the two equations in a different order. They are the same equations.

The third assertion of the theorem which says $E_1 = f_1, E_2 = f_2$ has the same solution as the system $E_1 = f_1, aE_2 = af_2$ provided $a \neq 0$ is verified as follows: If (x_1, \dots, x_n) is a solution of $E_1 = f_1, E_2 = f_2$, then it is a solution to $E_1 = f_1, aE_2 = af_2$ because the second system only involves multiplying the equation, $E_2 = f_2$ by a . If (x_1, \dots, x_n) is a solution of $E_1 = f_1, aE_2 = af_2$, then upon multiplying $aE_2 = af_2$ by the number $1/a$, you find that $E_2 = f_2$. ■

Stated simply, the above theorem shows that the elementary operations do not change the solution set of a system of equations.

Here is an example in which there are three equations and three variables. You want to find values for x, y, z such that each of the given equations are satisfied when these values are plugged in to the equations.

Example 7.1.5 Find the solutions to the system,

$$\begin{aligned} x + 3y + 6z &= 25 \\ 2x + 7y + 14z &= 58 \\ 2y + 5z &= 19 \end{aligned} \tag{7.4}$$

To solve this system replace the second equation by (-2) times the first equation added to the second. This yields the system

$$\begin{aligned}x + 3y + 6z &= 25 \\y + 2z &= 8 \\2y + 5z &= 19\end{aligned}\tag{7.5}$$

Now take (-2) times the second and add to the third. More precisely, replace the third equation with (-2) times the second added to the third. This yields the system

$$\begin{aligned}x + 3y + 6z &= 25 \\y + 2z &= 8 \\z &= 3\end{aligned}\tag{7.6}$$

At this point, you can tell what the solution is. This system has the same solution as the original system and in the above, $z = 3$. Then using this in the second equation, it follows $y + 6 = 8$ and so $y = 2$. Now using this in the top equation yields $x + 6 + 18 = 25$ and so $x = 1$. This process is called **back substitution**.

Alternatively, in 7.6 you could have continued as follows. Add (-2) times the bottom equation to the middle and then add (-6) times the bottom to the top. This yields

$$x + 3y = 7, y = 2, z = 3$$

Now add (-3) times the second to the top. This yields

$$x = 1, y = 2, z = 3,$$

a system which has the same solution set as the original system. This avoided back substitution and led to the same solution set.

7.1.2 Gauss Elimination

A less cumbersome way to represent a linear system is to write it as an **augmented matrix**. For example the linear system, 7.4 can be written as

$$\left(\begin{array}{ccc|c} 1 & 3 & 6 & 25 \\ 2 & 7 & 14 & 58 \\ 0 & 2 & 5 & 19 \end{array} \right).$$

It has exactly the same information as the original system but here it is understood there is an x column, $\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$, a y column, $\begin{pmatrix} 3 \\ 7 \\ 2 \end{pmatrix}$ and a z column, $\begin{pmatrix} 6 \\ 14 \\ 5 \end{pmatrix}$. The rows correspond to the equations in the system. Thus the top row in the augmented matrix corresponds to the equation,

$$x + 3y + 6z = 25.$$

Now when you replace an equation with a multiple of another equation added to itself, you are just taking a row of this augmented matrix and replacing it with a multiple of another

row added to it. Thus the first step in solving 7.4 would be to take (-2) times the first row of the augmented matrix above and add it to the second row,

$$\left(\begin{array}{ccc|c} 1 & 3 & 6 & 25 \\ 0 & 1 & 2 & 8 \\ 0 & 2 & 5 & 19 \end{array} \right).$$

Note how this corresponds to 7.5. Next take (-2) times the second row and add to the third,

$$\left(\begin{array}{ccc|c} 1 & 3 & 6 & 25 \\ 0 & 1 & 2 & 8 \\ 0 & 0 & 1 & 3 \end{array} \right)$$

This augmented matrix corresponds to the system

$$\begin{aligned} x + 3y + 6z &= 25 \\ y + 2z &= 8 \\ z &= 3 \end{aligned}$$

which is the same as 7.6. By back substitution you obtain the solution $x = 1, y = 6$, and $z = 3$.

In general a linear system is of the form

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned} \quad (7.7)$$

where the x_i are variables and the a_{ij} and b_i are constants. This system can be represented by the augmented matrix

$$\left(\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{array} \right). \quad (7.8)$$

Changes to the system of equations in 7.7 as a result of an elementary operations translate into changes of the augmented matrix resulting from a row operation. Note that Theorem 7.1.4 implies that the row operations deliver an augmented matrix for a system of equations which has the same solution set as the original system.

Definition 7.1.6 *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 it.*

Gauss elimination is a systematic procedure to simplify an augmented matrix to a reduced form. In the following definition, the term “**leading entry**” refers to the first nonzero entry of a row when scanning the row from left to right.

Definition 7.1.7 An augmented matrix is in **echelon form** if

1. All nonzero rows are above any rows of zeros.
2. Each leading entry of a row is in a column to the right of the leading entries of any rows above it.

How do you know when to stop doing row operations? You might stop when you have obtained an echelon form as described above, but you certainly should stop doing row operations if you have gotten a matrix in row reduced echelon form described next.

Definition 7.1.8 An augmented matrix is in **row reduced echelon form** if

1. All nonzero rows are above any rows of zeros.
2. Each leading entry of a row is in a column to the right of the leading entries of any rows above it.
3. All entries in a column above and below a leading entry are zero.
4. Each leading entry is a 1, the only nonzero entry in its column.

Example 7.1.9 Here are some matrices which are in row reduced echelon form.

$$\begin{pmatrix} 1 & 0 & 0 & 5 & 8 & 0 \\ 0 & 0 & 1 & 2 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Example 7.1.10 Here are matrices in echelon form which are not in row reduced echelon form but which are in echelon form.

$$\begin{pmatrix} 1 & 0 & 6 & 5 & 8 & 2 \\ 0 & 0 & 2 & 2 & 7 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 3 & 5 & 4 \\ 0 & 2 & 0 & 7 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Example 7.1.11 Here are some matrices which are not in echelon form.

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & -6 \\ 4 & 0 & 7 \end{pmatrix}, \begin{pmatrix} 0 & 2 & 3 & 3 \\ 1 & 5 & 0 & 2 \\ 7 & 5 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Definition 7.1.12 A **pivot position** in a matrix is the location of a leading entry in an echelon form resulting from the application of row operations to the matrix. A **pivot column** is a column that contains a pivot position.

For example consider the following.

Example 7.1.13 Suppose

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 6 \\ 4 & 4 & 4 & 10 \end{pmatrix}$$

Where are the pivot positions and pivot columns?

Replace the second row by -3 times the first added to the second. This yields

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -6 \\ 4 & 4 & 4 & 10 \end{pmatrix}.$$

This is not in reduced echelon form so replace the bottom row by -4 times the top row added to the bottom. This yields

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -6 \\ 0 & -4 & -8 & -6 \end{pmatrix}.$$

This is still not in reduced echelon form. Replace the bottom row by -1 times the middle row added to the bottom. This yields

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -6 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

which is in echelon form, although not in reduced echelon form. Therefore, the pivot positions in the original matrix are the locations corresponding to the first row and first column and the second row and second columns as shown in the following:

$$\begin{pmatrix} \boxed{1} & 2 & 3 & 4 \\ 3 & \boxed{2} & 1 & 6 \\ 4 & 4 & 4 & 10 \end{pmatrix}$$

Thus the pivot columns in the matrix are the first two columns.

The following is the algorithm for obtaining a matrix which is in row reduced echelon form.

Algorithm 7.1.14

This algorithm tells how to start with a matrix and do row operations on it in such a way as to end up with a matrix in row reduced echelon form.

1. Find the first nonzero column from the left. This is the first pivot column. The position at the top of the first pivot column is the first pivot position. Switch rows if necessary to place a nonzero number in the first pivot position.

2. Use row operations to zero out the entries below the first pivot position.
3. Ignore the row containing the most recent pivot position identified and the rows above it. Repeat steps 1 and 2 to the remaining sub-matrix, the rectangular array of numbers obtained from the original matrix by deleting the rows you just ignored. Repeat the process until there are no more rows to modify. The matrix will then be in echelon form.
4. Moving from right to left, use the nonzero elements in the pivot positions to zero out the elements in the pivot columns which are above the pivots.
5. Divide each nonzero row by the value of the leading entry. The result will be a matrix in row reduced echelon form.

This row reduction procedure applies to both augmented matrices and non augmented matrices. There is nothing special about the augmented column with respect to the row reduction procedure.

Example 7.1.15 Here is a matrix.

$$\begin{pmatrix} 0 & 0 & 2 & 3 & 2 \\ 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \end{pmatrix}$$

Do row reductions till you obtain a matrix in echelon form. Then complete the process by producing one in row reduced echelon form.

The pivot column is the second. Hence the pivot position is the one in the first row and second column. Switch the first two rows to obtain a nonzero entry in this pivot position.

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \end{pmatrix}$$

Step two is not necessary because all the entries below the first pivot position in the resulting matrix are zero. Now ignore the top row and the columns to the left of this first pivot position. Thus you apply the same operations to the smaller matrix

$$\begin{pmatrix} 2 & 3 & 2 \\ 1 & 2 & 2 \\ 0 & 0 & 0 \\ 0 & 2 & 1 \end{pmatrix}.$$

The next pivot column is the third corresponding to the first in this smaller matrix and the second pivot position is therefore, the one which is in the second row and third column.

In this case it is not necessary to switch any rows to place a nonzero entry in this position because there is already a nonzero entry there. Multiply the third row of the original matrix by -2 and then add the second row to it. This yields

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \end{pmatrix}.$$

The next matrix the steps in the algorithm are applied to is

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

The first pivot column is the first column in this case and no switching of rows is necessary because there is a nonzero entry in the first pivot position. Therefore, the algorithm yields for the next step

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -3 \end{pmatrix}.$$

Now the algorithm will be applied to the matrix

$$\begin{pmatrix} 0 \\ -3 \end{pmatrix}$$

There is only one column and it is nonzero so this single column is the pivot column. Therefore, the algorithm yields the following matrix for the echelon form.

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

To complete placing the matrix in reduced echelon form, multiply the third row by 3 and add -2 times the fourth row to it. This yields

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Next multiply the second row by 3 and take 2 times the fourth row and add to it. Then add the fourth row to the first.

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 0 \\ 0 & 0 & 6 & 9 & 0 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Next work on the fourth column in the same way.

$$\begin{pmatrix} 0 & 3 & 3 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Take $-1/2$ times the second row and add to the first.

$$\begin{pmatrix} 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Finally, divide by the value of the leading entries in the nonzero rows.

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

The above algorithm is the way a computer would obtain a reduced echelon form for a given matrix. It is not necessary for you to pretend you are a computer but if you like to do so, the algorithm described above will work. The main idea is to do row operations in such a way as to end up with a matrix in echelon form or row reduced echelon form because when this has been done, the resulting augmented matrix will allow you to describe the solutions to the linear system of equations in a meaningful way. When you do row operations until you obtain row reduced echelon form, the process is called the Gauss Jordan method. Otherwise, it is called Gauss elimination.

Example 7.1.16 Give the complete solution to the system of equations, $5x + 10y - 7z = -2$, $2x + 4y - 3z = -1$, and $3x + 6y + 5z = 9$.

The augmented matrix for this system is

$$\begin{pmatrix} 2 & 4 & -3 & -1 \\ 5 & 10 & -7 & -2 \\ 3 & 6 & 5 & 9 \end{pmatrix}$$

Multiply the second row by 2, the first row by 5, and then take (-1) times the first row and add to the second. Then multiply the first row by $1/5$. This yields

$$\begin{pmatrix} 2 & 4 & -3 & -1 \\ 0 & 0 & 1 & 1 \\ 3 & 6 & 5 & 9 \end{pmatrix}$$

Now, combining some row operations, take (-3) times the first row and add this to 2 times the last row and replace the last row with this. This yields.

$$\begin{pmatrix} 2 & 4 & -3 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 21 \end{pmatrix}.$$

One more row operation, taking (-1) times the second row and adding to the bottom yields.

$$\begin{pmatrix} 2 & 4 & -3 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 20 \end{pmatrix}.$$

This is impossible because the last row indicates the need for a solution to the equation

$$0x + 0y + 0z = 20$$

and there is no such thing because $0 \neq 20$. This shows there is no solution to the three given equations. When this happens, the system is called **inconsistent**. In this case it is very easy to describe the solution set. The system has no solution.

Here is another example based on the use of row operations.

Example 7.1.17 Give the complete solution to the system of equations, $3x - y - 5z = 9$, $y - 10z = 0$, and $-2x + y = -6$.

The augmented matrix of this system is

$$\begin{pmatrix} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ -2 & 1 & 0 & -6 \end{pmatrix}$$

Replace the last row with 2 times the top row added to 3 times the bottom row combining two row operations. This gives

$$\begin{pmatrix} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ 0 & 1 & -10 & 0 \end{pmatrix}.$$

The entry, 3 in this sequence of row operations is called the **pivot**. It is used to create zeros in the other places of the column. Next take -1 times the middle row and add to the bottom. Here the 1 in the second row is the pivot.

$$\begin{pmatrix} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Take the middle row and add to the top and then divide the top row which results by 3.

$$\begin{pmatrix} 1 & 0 & -5 & 3 \\ 0 & 1 & -10 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

This is in reduced echelon form. The equations corresponding to this reduced echelon form are $y = 10z$ and $x = 3 + 5z$. Apparently z can equal any number. Let's call this number t .¹ Therefore, the solution set of this system is $x = 3 + 5t$, $y = 10t$, and $z = t$ where t is completely arbitrary. The system has an infinite set of solutions which are given in the above simple way. This is what it is all about, finding the solutions to the system.

There is some terminology connected to this which is useful. Recall how each column corresponds to a variable in the original system of equations. The variables corresponding to a pivot column are called **basic variables**. The other variables are called **free variables**. In Example 7.1.17 there was one free variable, z , and two basic variables, x and y . In describing the solution to the system of equations, the free variables are assigned a parameter. In Example 7.1.17 this parameter was t . Sometimes there are many free variables and in these cases, you need to use many parameters. Here is another example.

Example 7.1.18 Find the solution to the system

$$x + 2y - z + w = 3$$

$$x + y - z + w = 1$$

$$x + 3y - z + w = 5$$

The augmented matrix is

$$\begin{pmatrix} 1 & 2 & -1 & 1 & 3 \\ 1 & 1 & -1 & 1 & 1 \\ 1 & 3 & -1 & 1 & 5 \end{pmatrix}.$$

Take -1 times the first row and add to the second. Then take -1 times the first row and add to the third. This yields

$$\begin{pmatrix} 1 & 2 & -1 & 1 & 3 \\ 0 & -1 & 0 & 0 & -2 \\ 0 & 1 & 0 & 0 & 2 \end{pmatrix}$$

Now add the second row to the bottom row

$$\begin{pmatrix} 1 & 2 & -1 & 1 & 3 \\ 0 & -1 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{7.9}$$

This matrix is in echelon form and you see the basic variables are x and y while the free variables are z and w . Assign s to z and t to w . Then the second row yields the equation, $y = 2$ while the top equation yields the equation, $x + 2y - s + t = 3$ and so since $y = 2$, this

¹In this context t is called a **parameter**.

gives $x + 4 - s + t = 3$ showing that $x = -1 + s - t$, $y = 2$, $z = s$, and $w = t$. It is customary to write this in the form

$$\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} -1 + s - t \\ 2 \\ s \\ t \end{pmatrix}. \quad (7.10)$$

This is another example of a system which has an infinite solution set but this time the solution set depends on two parameters, not one. Most people find it less confusing in the case of an infinite solution set to first place the augmented matrix in row reduced echelon form rather than just echelon form before seeking to write down the description of the solution. In the above, this means we don't stop with the echelon form 7.9. Instead we first place it in reduced echelon form as follows.

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

Then the solution is $y = 2$ from the second row and $x = -1 + z - w$ from the first. Thus letting $z = s$ and $w = t$, the solution is given in 7.10.

The number of free variables is always equal to the number of **different** parameters used to describe the solution. If there are no free variables, then either there is no solution as in the case where row operations yield an echelon form like

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & -2 \\ 0 & 0 & 1 \end{pmatrix}$$

or there is a unique solution as in the case where row operations yield an echelon form like

$$\begin{pmatrix} 1 & 2 & 2 & 3 \\ 0 & 4 & 3 & -2 \\ 0 & 0 & 4 & 1 \end{pmatrix}.$$

Also, sometimes there are free variables and no solution as in the following:

$$\begin{pmatrix} 1 & 2 & 2 & 3 \\ 0 & 4 & 3 & -2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

There are a lot of cases to consider but it is not necessary to make a major production of this. Do row operations till you obtain a matrix in echelon form or reduced echelon form and determine whether there is a solution. If there is, see if there are free variables. In this case, there will be infinitely many solutions. Find them by assigning different parameters to the free variables and obtain the solution. If there are no free variables, then there will be a unique solution which is easily determined once the augmented matrix is in echelon or row reduced echelon form. In every case, the process yields a straightforward way to describe the solutions to the linear system. As indicated above, you are probably less likely

to become confused if you place the augmented matrix in row reduced echelon form rather than just echelon form.

In summary,

Definition 7.1.19 A system of linear equations is a list of equations,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

where a_{ij} are numbers, and b_j is a number. The above is a system of m equations in the n variables, x_1, x_2, \dots, x_n . Nothing is said about the relative size of m and n . Written more simply in terms of summation notation, the above can be written in the form

$$\sum_{j=1}^n a_{ij}x_j = f_i, \quad i = 1, 2, 3, \dots, m$$

It is desired to find (x_1, \dots, x_n) solving each of the equations listed.

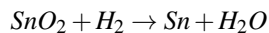
As illustrated above, such a system of linear equations may have a unique solution, no solution, or infinitely many solutions and these are the only three cases which can occur for any linear system. Furthermore, you do exactly the same things to solve any linear system. You write the augmented matrix and do row operations until you get a simpler system in which it is possible to see the solution, usually obtaining a matrix in echelon or reduced echelon form. All is based on the observation that the row operations do not change the solution set. You can have more equations than variables, fewer equations than variables, etc. It doesn't matter. You always set up the augmented matrix and go to work on it.

Definition 7.1.20 A system of linear equations is called **consistent** if there exists a solution. It is called **inconsistent** if there is no solution.

These are reasonable words to describe the situations of having or not having a solution. If you think of each equation as a condition which must be satisfied by the variables, consistent would mean there is some choice of variables which can satisfy all the conditions. Inconsistent would mean there is no choice of the variables which can satisfy each of the conditions.

7.1.3 Balancing Chemical Reactions

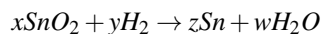
Consider the chemical reaction



Here the elements involved are tin Sn oxygen O and Hydrogen H . Some chemical reaction happens and you end up with some tin and some water. The question is, how much do you start with and how much do you end up with.

The balance of mass requires that you have the same number of oxygen, tin, and hydrogen on both sides of the reaction. However, this does not happen in the above. For

example, there are two oxygen atoms on the left and only one on the right. The problem is to find numbers x, y, z, w such that



and both sides have the same number of atoms of the various substances. You can do this in a systematic way by setting up a system of equations which will require that this take place. Thus you need

$$\text{Sn} : \quad x = z$$

$$\text{O} : \quad 2x = w$$

$$\text{H} : \quad 2y = 2w$$

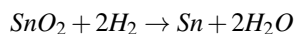
The augmented matrix for this system of equations is then

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

Row reducing this yields

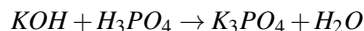
$$\left(\begin{array}{ccccc} 1 & 0 & 0 & -\frac{1}{2} & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 0 \end{array} \right)$$

Thus you could let $w = 2$ and this would yield $x = 1, y = 2$, and $z = 1$. Hence, the description of the reaction which has the same numbers of atoms on both sides would be



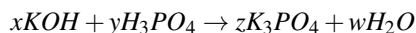
You see that this preserves the total number of atoms and so the chemical equation is balanced. Here is another example

Example 7.1.21 Potassium is denoted by K , oxygen by O , phosphorus by P and hydrogen by H . The reaction is



balance this equation.

You need to have



Equations which preserve the total number of atoms of each element on both sides of the equation are

$$K : \quad x = 3z$$

$$O : \quad x + 4y = 4z + w$$

$$H : \quad x + 3y = 2w$$

$$P : \quad y = z$$

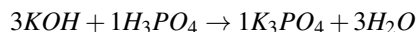
The augmented matrix for this system is

$$\left(\begin{array}{ccccc} 1 & 0 & -3 & 0 & 0 \\ 1 & 4 & -4 & -1 & 0 \\ 1 & 3 & 0 & -2 & 0 \\ 0 & 1 & -1 & 0 & 0 \end{array} \right)$$

Then the row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -\frac{1}{3} & 0 \\ 0 & 0 & 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

You could let $w = 3$ and this yields $x = 3, y = 1, z = 1$. Then the balanced equation is



Note that this results in the same number of atoms on both sides.

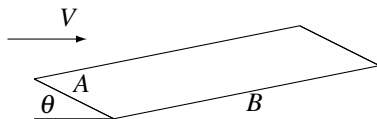
Of course these numbers you are finding would typically be the number of moles of the molecules on each side. Thus three moles of KOH added to one mole of H_3PO_4 yields one mole of K_3PO_4 and three moles of H_2O , water.

Note that in this example, you have a row of zeros. This means that some of the information in computing the appropriate numbers was redundant. If this can happen with a single reaction, think how much more it could happen if you were dealing with hundreds of reactions. This aspect of the problem can be understood later in terms of the rank of a matrix.

For an introduction to the chemical considerations mentioned here, there is a nice site on the web <http://chemistry.about.com/od/chemicalreactions/a/reactiontypes.htm> where there is a sample test and examples of chemical reactions. For names of the various elements symbolized by the various letters, you can go to the site <http://chemistry.about.com/od/elementfacts/a/elementlist.htm>. [Chemical elements](#) Of course these things are in standard chemistry books, but if you have not seen much chemistry, these sites give a nice introduction to these concepts.

7.1.4 Dimensionless Variables*

This section shows how solving systems of equations can be used to determine appropriate dimensionless variables. It is only an introduction to this topic. I got this example from [18]. This considers a specific example of a simple airplane wing shown below. We assume for simplicity that it is just a flat plane at an angle to the wind which is blowing against it with speed V as shown.



The angle is called the angle of incidence, B is the span of the wing and A is called the chord. Denote by l the lift. Then this should depend on various quantities like θ, V, B, A and so forth. Here is a table which indicates various quantities on which it is reasonable to

expect l to depend.

Variable	Symbol	Units
chord	A	m
span	B	m
angle incidence	θ	$m^0 kg^0 sec^0$
speed of wind	V	$m sec^{-1}$
speed of sound	V_0	$m sec^{-1}$
density of air	ρ	$kg m^{-3}$
viscosity	μ	$kg sec^{-1} m^{-1}$
lift	l	$kg sec^{-2} m$

Here m denotes meters, sec refers to seconds and kg refers to kilograms. All of these are likely familiar except for μ . One can simply decree that these are the dimensions of something called viscosity but it might be better to consider this a little more.

Viscosity is a measure of how much internal friction is experienced when the fluid moves. It is roughly a measure of how “sticky” the fluid is. Consider a piece of area parallel to the direction of motion of the fluid. To say that the viscosity is large is to say that the tangential force applied to this area must be large in order to achieve a given change in speed of the fluid in a direction normal to the tangential force. Thus

$$\mu (\text{area}) (\text{velocity gradient}) = \text{tangential force.}$$

Hence

$$(\text{units on } \mu) m^2 \left(\frac{m}{sec m} \right) = kg sec^{-2} m$$

Thus the units on μ are $kg sec^{-1} m^{-1}$ as claimed above.

Then one would think that you would want

$$l = f(A, B, \theta, V, V_0, \rho, \mu)$$

However, this is very cumbersome because it depends on seven variables. Also, it doesn't make very good sense. It is likely that without much care, a change in the units such as going from meters to feet would result in an incorrect value for l . The way to get around this problem is to look for l as a function of dimensionless variables multiplied by something which has units of force. It is helpful because first of all, you will likely have fewer independent variables and secondly, you could expect the formula to hold independent of the way of specifying length, mass and so forth. One looks for

$$l = f(g_1, \dots, g_k) \rho V^2 AB$$

where the units on $\rho V^2 AB$ are

$$\frac{kg}{m^3} \left(\frac{m}{sec} \right)^2 m^2 = \frac{kg \times m}{sec^2}$$

which are the units of force. Each of these g_i is of the form

$$A^{x_1} B^{x_2} \theta^{x_3} V^{x_4} V_0^{x_5} \rho^{x_6} \mu^{x_7} \quad (7.11)$$

and each g_i is independent of the dimensions. That is, this expression must not depend on meters, kilograms, seconds, etc. Thus, placing in the units for each of these quantities, one needs

$$m^{x_1} m^{x_2} (m^{x_4} \sec^{-x_4}) (m^{x_5} \sec^{-x_5}) (kg m^{-3})^{x_6} (kg \sec^{-1} m^{-1})^{x_7} = m^0 kg^0 \sec^0$$

Notice that there are no units on θ because it is just the radian measure of an angle. Hence its dimensions consist of length divided by length, thus it is dimensionless. Then this leads to the following equations for the x_i .

$$\begin{aligned} m : \quad & x_1 + x_2 + x_4 + x_5 - 3x_6 - x_7 = 0 \\ \sec : \quad & -x_4 - x_5 - x_7 = 0 \\ kg : \quad & x_6 + x_7 = 0 \end{aligned}$$

Then the augmented matrix for this system of equations is

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

The row reduced echelon form is then

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

and so the solutions are of the form

$$x_1 = -x_2 - x_7, \quad x_3 = x_3, \quad x_4 = -x_5 - x_7, \quad x_6 = -x_7$$

Thus, in terms of vectors, the solution is

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} -x_2 - x_7 \\ x_2 \\ x_3 \\ -x_5 - x_7 \\ x_5 \\ -x_7 \\ x_7 \end{pmatrix}$$

Thus the free variables are x_2, x_3, x_5, x_7 . By assigning values to these, we can obtain dimensionless variables by placing the values obtained for the x_i in the formula 7.11. For example, let $x_2 = 1$ and all the rest of the free variables are 0. This yields

$$x_1 = -1, x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 0, x_6 = 0, x_7 = 0.$$

The dimensionless variable is then $A^{-1}B^1$. This is the ratio between the span and the chord. It is called the aspect ratio, denoted as AR . Next let $x_3 = 1$ and all others equal zero. This

gives for a dimensionless quantity the angle θ . Next let $x_5 = 1$ and all others equal zero. This gives

$$x_1 = 0, x_2 = 0, x_3 = 0, x_4 = -1, x_5 = 1, x_6 = 0, x_7 = 0.$$

Then the dimensionless variable is $V^{-1}V_0^1$. However, it is written as V/V_0 . This is called the Mach number \mathcal{M} . Finally, let $x_7 = 1$ and all the other free variables equal 0. Then

$$x_1 = -1, x_2 = 0, x_3 = 0, x_4 = -1, x_5 = 0, x_6 = -1, x_7 = 1$$

then the dimensionless variable which results from this is $A^{-1}V^{-1}\rho^{-1}\mu$. It is customary to write it as $\text{Re} = (AV\rho)/\mu$. This one is called the Reynolds number. It is the one which involves viscosity. Thus we would look for

$$l = f(\text{Re}, AR, \theta, \mathcal{M}) \text{ kg} \times \text{m} / \text{sec}^2$$

This is quite interesting because it is easy to vary Re by simply adjusting the velocity or A but it is hard to vary things like μ or ρ . Note that all the quantities are easy to adjust. Now this could be used, along with wind tunnel experiments to get a formula for the lift which would be reasonable. Obviously, you could consider more variables and more complicated situations in the same way.

7.2 MATLAB And Row Reduced Echelon Form

MATLAB will find the row reduced echelon form of a matrix and save you the trouble of tedious computations. You open matlab. You will see `>>`. Then next to it you type the following:

```
rref([1,2,3,4;2,5,6,10;3,2,0,-5])
```

Then press enter on your keyboard. It will give the following.

```
ans =
```

```
1 0 0 -3
```

```
0 1 0 2
```

```
0 0 1 1
```

In usual notation, this is the row reduced echelon form of the matrix

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 5 & 6 & 10 \\ 3 & 2 & 0 & -5 \end{pmatrix}$$
 Notice how you enter a row by placing commas between entries and then when you start a new row, you put a ; to indicate it is a new row. You do something similar for another matrix. You can also simply leave a space between the entries of a row and it will know what to do, but you indicate a new row by using ;. The semicolon ; is also used to defer an operation. MATLAB will know about it but won't do anything.

In using MATLAB, you press shift enter to go to a new line. One thing might be helpful to mention about MATLAB. It is very good at manipulating matrices and vectors and there is distinctive notation used to accomplish this. For example say you type

$$x=[1,2,3]; y=[2,3,4]; x.*y$$

and then press "enter". You will get 2,6,12. You would get an error if you wrote $x*y$. Similarly, type

$$[2,4,6,8]./[1,2,3,4]$$

and press "enter". This yields 2,2,2,2. The expression $[2,4,6,8]/[1,2,3,4]$ doesn't make any sense.

7.3 Exercises

- Find the point (x_1, y_1) which lies on both lines, $x + 3y = 1$ and $4x - y = 3$.
- Solve Problem 1 graphically. That is, graph each line and see where they intersect.
- Find the point of intersection of the two lines $3x + y = 3$ and $x + 2y = 1$.
- Solve Problem 3 graphically. That is, graph each line and see where they intersect.
- Do the three lines, $x + 2y = 1$, $2x - y = 1$, and $4x + 3y = 3$ have a common point of intersection? If so, find the point and if not, tell why they don't have such a common point of intersection.
- Do the three planes, $x + y - 3z = 2$, $2x + y + z = 1$, and $3x + 2y - 2z = 0$ have a common point of intersection? If so, find one and if not, tell why there is no such point.
- You have a system of k equations in two variables, $k \geq 2$. Explain the geometric significance of
 - No solution.
 - A unique solution.
 - An infinite number of solutions.
- Here is an augmented matrix in which $*$ denotes an arbitrary number and \blacksquare denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & * & * \\ 0 & \blacksquare & * & * & 0 & * \\ 0 & 0 & \blacksquare & * & * & * \\ 0 & 0 & 0 & 0 & \blacksquare & * \end{array} \right)$$

- Here is an augmented matrix in which $*$ denotes an arbitrary number and \blacksquare denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$\left(\begin{array}{ccc|c} \blacksquare & * & * & * \\ 0 & \blacksquare & * & * \\ 0 & 0 & \blacksquare & * \end{array} \right)$$

- Here is an augmented matrix in which $*$ denotes an arbitrary number and \blacksquare denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & * & * \\ 0 & \blacksquare & 0 & * & 0 & * \\ 0 & 0 & 0 & \blacksquare & * & * \\ 0 & 0 & 0 & 0 & \blacksquare & * \end{array} \right)$$

11. Here is an augmented matrix in which $*$ denotes an arbitrary number and \blacksquare denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & * & * \\ 0 & \blacksquare & * & * & 0 & * \\ 0 & 0 & 0 & 0 & \blacksquare & 0 \\ 0 & 0 & 0 & 0 & * & \blacksquare \end{array} \right)$$

12. Suppose a system of equations has fewer equations than variables. Must such a system be consistent? If so, explain why and if not, give an example which is not consistent.
13. If a system of equations has more equations than variables, can it have a solution? If so, give an example and if not, tell why not.
14. Find h such that

$$\left(\begin{array}{cc|c} 2 & h & 4 \\ 3 & 6 & 7 \end{array} \right)$$

is the augmented matrix of an inconsistent matrix.

15. Find h such that

$$\left(\begin{array}{cc|c} 1 & h & 3 \\ 2 & 4 & 6 \end{array} \right)$$

is the augmented matrix of a consistent matrix.

16. Find h such that

$$\left(\begin{array}{cc|c} 1 & 1 & 4 \\ 3 & h & 12 \end{array} \right)$$

is the augmented matrix of a consistent matrix.

17. Choose h and k such that the augmented matrix shown has one solution. Then choose h and k such that the system has no solutions. Finally, choose h and k such that the system has infinitely many solutions.

$$\left(\begin{array}{cc|c} 1 & h & 2 \\ 2 & 4 & k \end{array} \right).$$

18. Choose h and k such that the augmented matrix shown has one solution. Then choose h and k such that the system has no solutions. Finally, choose h and k such that the system has infinitely many solutions.

$$\left(\begin{array}{cc|c} 1 & 2 & 2 \\ 2 & h & k \end{array} \right).$$

19. Determine if the system is consistent. If so, is the solution unique?

$$x + 2y + z - w = 2$$

$$x - y + z + w = 1$$

$$2x + y - z = 1$$

$$4x + 2y + z = 5$$

20. Determine if the system is consistent. If so, is the solution unique?

$$x + 2y + z - w = 2$$

$$x - y + z + w = 0$$

$$2x + y - z = 1$$

$$4x + 2y + z = 3$$

21. Find the general solution of the system whose augmented matrix is

$$\left(\begin{array}{ccc|c} 1 & 2 & 0 & 2 \\ 1 & 3 & 4 & 2 \\ 1 & 0 & 2 & 1 \end{array} \right).$$

22. Find the general solution of the system whose augmented matrix is

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

23. Find the general solution of the system whose augmented matrix is

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & 1 \\ 1 & 0 & 4 & 2 \end{array} \right).$$

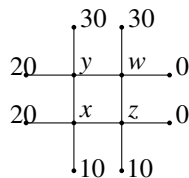
24. Find the general solution of the system whose augmented matrix is

$$\left(\begin{array}{ccccc|c} 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 2 & 1 \\ 1 & 2 & 0 & 0 & 1 & 3 \\ 1 & 0 & 1 & 0 & 2 & 2 \end{array} \right).$$

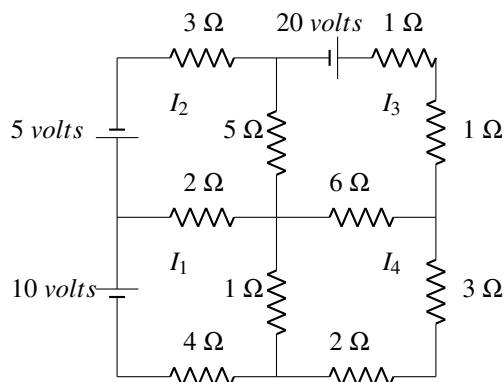
25. Find the general solution of the system whose augmented matrix is

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

26. Give the complete solution to the system of equations, $7x + 14y + 15z = 22$, $2x + 4y + 3z = 5$, and $3x + 6y + 10z = 13$.
27. Give the complete solution to the system of equations, $3x - y + 4z = 6$, $y + 8z = 0$, and $-2x + y = -4$.
28. Give the complete solution to the system of equations, $9x - 2y + 4z = -17$, $13x - 3y + 6z = -25$, and $-2x - z = 3$.
29. Give the complete solution to the system of equations, $65x + 84y + 16z = 546$, $81x + 105y + 20z = 682$, and $84x + 110y + 21z = 713$.
30. Give the complete solution to the system of equations, $8x + 2y + 3z = -3$, $8x + 3y + 3z = -1$, and $4x + y + 3z = -9$.
31. Give the complete solution to the system of equations, $-8x + 2y + 5z = 18$, $-8x + 3y + 5z = 13$, and $-4x + y + 5z = 19$.
32. Give the complete solution to the system of equations, $3x - y - 2z = 3$, $y - 4z = 0$, and $-2x + y = -2$.
33. Give the complete solution to the system of equations, $-9x + 15y = 66$, $-11x + 18y = 79$, $-x + y = 4$, and $z = 3$.
34. Give the complete solution to the system of equations, $-19x + 8y = -108$, $-71x + 30y = -404$, $-2x + y = -12$, $4x + z = 14$.
35. Consider the system $-5x + 2y - z = 0$ and $-5x - 2y - z = 0$. Both equations equal zero and so $-5x + 2y - z = -5x - 2y - z$ which is equivalent to $y = 0$. Thus x and z can equal anything. But when $x = 1$, $z = -4$, and $y = 0$ are plugged in to the equations, it doesn't work. Why?
36. Four times the weight of Gaston is 150 pounds more than the weight of Ichabod. Four times the weight of Ichabod is 660 pounds less than seventeen times the weight of Gaston. Four times the weight of Gaston plus the weight of Siegfried equals 290 pounds. Brunhilde would balance all three of the others. Find the weights of the four sisters.
37. The steady state temperature, u in a plate solves Laplace's equation, $\Delta u = 0$. One way to approximate the solution which is often used is to divide the plate into a square mesh and require the temperature at each node to equal the average of the temperature at the four adjacent nodes. This procedure is justified by the mean value property of harmonic functions. In the following picture, the numbers represent the observed temperature at the indicated nodes. Your task is to find the temperature at the interior nodes, indicated by x, y, z , and w . One of the equations is $z = \frac{1}{4}(10 + 0 + w + x)$.



38. Consider the following diagram of four circuits.



Those jagged places denote resistors and the numbers next to them give their resistance in ohms, written as Ω . The breaks in the lines having one short line and one long line denote a voltage source which causes the current to flow in the direction which goes from the longer of the two lines toward the shorter along the unbroken part of the circuit. The current in amps in the four circuits is denoted by I_1, I_2, I_3, I_4 and it is understood that the motion is in the counter clockwise direction. If I_k ends up being negative, then it just means the current flows in the clockwise direction. Then Kirchhoff's law states that

The sum of the resistance times the amps in the counter clockwise direction around a loop equals the sum of the voltage sources in the same direction around the loop.

In the above diagram, the top left circuit should give the equation

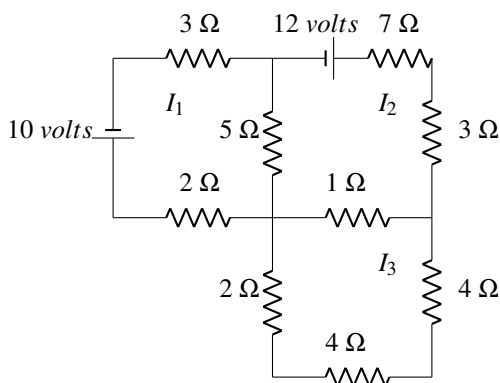
$$2I_2 - 2I_1 + 5I_2 - 5I_3 + 3I_2 = 5$$

For the circuit on the lower left, you should have

$$4I_1 + I_1 - I_4 + 2I_1 - 2I_2 = -10$$

Write equations for each of the other two circuits and then give a solution to the resulting system of equations. You might use a computer algebra system to find the solution. It might be more convenient than doing it by hand.

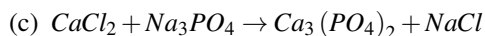
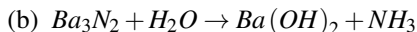
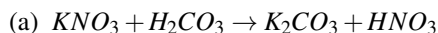
39. Consider the following diagram of three circuits.



Those jagged places denote resistors and the numbers next to them give their resistance in ohms, written as Ω . The breaks in the lines having one short line and one long line denote a voltage source which causes the current to flow in the direction which goes from the longer of the two lines toward the shorter along the unbroken part of the circuit. The current in amps in the four circuits is denoted by I_1, I_2, I_3 and it is understood that the motion is in the counter clockwise direction. If I_k ends up being negative, then it just means the current flows in the clockwise direction. Then Kirchhoff's law states that

The sum of the resistance times the amps in the counter clockwise direction around a loop equals the sum of the voltage sources in the same direction around the loop. Find I_1, I_2, I_3 .

40. Here are some chemical reactions. Balance them.



41. In the section on dimensionless variables 119 it was observed that $\rho V^2 AB$ has the units of force. Describe a systematic way to obtain such combinations of the variables which will yield something which has the units of force.

Chapter 8

Matrices

8.1 Addition And Scalar Multiplication Of Matrices

You have now solved systems of equations by writing them in terms of an augmented matrix and then doing row operations on this augmented matrix. It turns out such rectangular arrays of numbers are important from many other different points of view. Numbers are also called **scalars**. In this book, numbers will generally be either real or complex numbers. I will refer to the set of numbers as \mathbb{F} sometimes when it is not important to worry about whether the number is real or complex. Thus \mathbb{F} can be either the real numbers, \mathbb{R} or the complex numbers \mathbb{C} . However, most of the algebraic considerations hold for more general fields of scalars.

A **matrix** is a rectangular array of numbers. Several of them are referred to as **matrices**. For example, here is a matrix.

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 2 & 8 & 7 \\ 6 & -9 & 1 & 2 \end{pmatrix}$$

The size or dimension of a matrix is defined as $m \times n$ where m is the number of rows and n is the number of columns. The above matrix is a 3×4 matrix because there are three rows and four columns. The first row is $(1 \ 2 \ 3 \ 4)$, the second row is $(5 \ 2 \ 8 \ 7)$ and so forth.

The first column is $\begin{pmatrix} 1 \\ 5 \\ 6 \end{pmatrix}$. When specifying the size of a matrix, you always list the

number of rows before the number of columns. Also, you can remember the columns are like columns in a Greek temple. They stand upright while the rows just lie there like rows made by a tractor in a plowed field. Elements of the matrix are identified according to position in the matrix. For example, 8 is in position 2, 3 because it is in the second row and the third column. You might remember that you always list the rows before the columns by using the phrase **Rowman Catholic**. The symbol, (a_{ij}) refers to a matrix. The entry in the i^{th} row and the j^{th} column of this matrix is denoted by a_{ij} . Using this notation on the above matrix, $a_{23} = 8, a_{32} = -9, a_{12} = 2$, etc.

There are various operations which are done on matrices. Matrices can be added multiplied by a scalar, and multiplied by other matrices. To illustrate scalar multiplication,

consider the following example in which a matrix is being multiplied by the scalar 3.

$$3 \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 2 & 8 & 7 \\ 6 & -9 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 6 & 9 & 12 \\ 15 & 6 & 24 & 21 \\ 18 & -27 & 3 & 6 \end{pmatrix}.$$

The new matrix is obtained by multiplying every entry of the original matrix by the given scalar. If A is an $m \times n$ matrix, $-A$ is defined to equal $(-1)A$.

Two matrices must be the same size to be added. The sum of two matrices is a matrix which is obtained by adding the corresponding entries. Thus

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 2 \end{pmatrix} + \begin{pmatrix} -1 & 4 \\ 2 & 8 \\ 6 & -4 \end{pmatrix} = \begin{pmatrix} 0 & 6 \\ 5 & 12 \\ 11 & -2 \end{pmatrix}.$$

Two matrices are equal exactly when they are the same size and the corresponding entries are identical. Thus

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

because they are different sizes. As noted above, you write (c_{ij}) for the matrix C whose i^{th} entry is c_{ij} . In doing arithmetic with matrices you must define what happens in terms of the c_{ij} sometimes called the **entries** of the matrix or the **components** of the matrix.

The above discussion stated for general matrices is given in the following definition.

Definition 8.1.1 (Scalar Multiplication) If $A = (a_{ij})$ and k is a scalar, then $kA = (ka_{ij})$.

Example 8.1.2 $7 \begin{pmatrix} 2 & 0 \\ 1 & -4 \end{pmatrix} = \begin{pmatrix} 14 & 0 \\ 7 & -28 \end{pmatrix}.$

Definition 8.1.3 (Addition) If $A = (a_{ij})$ and $B = (b_{ij})$ are two $m \times n$ matrices. Then $A + B = C$ where

$$C = (c_{ij})$$

for $c_{ij} = a_{ij} + b_{ij}$.

Example 8.1.4

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 2 & 3 \\ -6 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 4 & 6 \\ -5 & 2 & 5 \end{pmatrix}$$

To save on notation, we will often use A_{ij} to refer to the i^{th} entry of the matrix A .

Definition 8.1.5 (The zero matrix) The $m \times n$ zero matrix is the $m \times n$ matrix having every entry equal to zero. It is denoted by 0 .

Example 8.1.6 The 2×3 zero matrix is $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$

Note there are 2×3 zero matrices, 3×4 zero matrices, etc. In fact there is a zero matrix for every size.

Definition 8.1.7 (*Equality of matrices*) Let A and B be two matrices. Then $A = B$ means that the two matrices are of the same size and for $A = (a_{ij})$ and $B = (b_{ij})$, $a_{ij} = b_{ij}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$.

The following properties of matrices can be easily verified. You should do so. These properties are called the vector space axioms.

- Commutative Law Of Addition.

$$A + B = B + A, \quad (8.1)$$

- Associative Law for Addition.

$$(A + B) + C = A + (B + C), \quad (8.2)$$

- Existence of an Additive Identity

$$A + 0 = A, \quad (8.3)$$

- Existence of an Additive Inverse

$$A + (-A) = 0, \quad (8.4)$$

Also for α, β scalars, the following additional properties hold.

- Distributive law over Matrix Addition.

$$\alpha(A + B) = \alpha A + \alpha B, \quad (8.5)$$

- Distributive law over Scalar Addition

$$(\alpha + \beta)A = \alpha A + \beta A, \quad (8.6)$$

- Associative law for Scalar Multiplication

$$\alpha(\beta A) = \alpha\beta(A), \quad (8.7)$$

- Rule for Multiplication by 1.

$$1A = A. \quad (8.8)$$

8.2 Multiplication of Matrices

As an example, consider the Commutative Law of Addition. Let $A + B = C$ and $B + A = D$. Why is $D = C$?

$$C_{ij} = A_{ij} + B_{ij} = B_{ij} + A_{ij} = D_{ij}.$$

Therefore, $C = D$ because the ij^{th} entries are the same. Note that the conclusion follows from the commutative law of addition of numbers.

From now on, we will typically write vectors as columns. Thus, when we write $\mathbf{x} \in \mathbb{F}^n$ we typically mean

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

We will also use the following convention.

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}^T = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}, \quad \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

The rules for adding and multiplying by a constant remain the same. To add, you add corresponding entries and to multiply by a scalar, you multiply every entry by the scalar. Consider the following system of equations:

$$\begin{aligned} x + y &= 1 \\ 2x - y + z &= 2 \\ x + y &= 1 \end{aligned}$$

Another way to write this is

$$x \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + y \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + z \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

That expression on the left is called a linear combination of the three vectors listed there whenever x, y, z are numbers. Another way to write it is

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

The rows of the above matrix are $\begin{pmatrix} 1 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 2 & -1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}$. The columns of this matrix are $\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$. It is called a 3×3 matrix because it has three rows and three columns. More generally, we have the following definition.

Definition 8.2.1 An $m \times n$ matrix is a rectangular array of numbers which has m rows and n columns. We write this as

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{n2} & \cdots & A_{mn} \end{pmatrix}$$

Thus the entry in the i^{th} row and the j^{th} column is denoted as A_{ij} . As suggested above,

$$\begin{aligned} A\mathbf{x} &= \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{n2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\ &= x_1 \begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{pmatrix} + x_2 \begin{pmatrix} A_{12} \\ A_{22} \\ \vdots \\ A_{n2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{mn} \end{pmatrix} \end{aligned}$$

Note that $A\mathbf{x}$ is in \mathbb{F}^m and the i^{th} entry of this vector $A\mathbf{x}$ is

$$A_{i1}x_1 + A_{i2}x_2 + \cdots + A_{in}x_n = \sum_{j=1}^n A_{ij}x_j.$$

In other words, the i^{th} entry of $A\mathbf{x}$ is the dot product of the i^{th} row of A with the vector \mathbf{x} . Symbolically,

$$(A\mathbf{x})_i = \sum_j A_{ij}x_j \quad (8.9)$$

We like to write \mathbf{x} to denote an $n \times 1$ matrix which is often called a vector. Then \mathbf{x}^T will denote a $1 \times n$ matrix or row vector.

Example 8.2.2

$$\begin{aligned} &\begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= x_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} 2 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} x_1 + 2x_2 + x_3 \\ x_1 + 2x_3 \end{pmatrix} \end{aligned}$$

Note that if A is $m \times n$ then $A\mathbf{x}$ is an $m \times 1$ matrix provided \mathbf{x} is $n \times 1$. Thus A makes a vector in \mathbb{F}^n into a vector in \mathbb{F}^m .

Example 8.2.3 Show the following:

$$\begin{pmatrix} 1 & -1 & 2 \\ 3 & 2 & 1 \\ 2 & 3 & -3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 5 \\ 10 \\ -1 \end{pmatrix}$$

Example 8.2.4 Write the system of equations

$$\begin{aligned}x + 2y - z &= 2 \\ x - 3y + z &= 1\end{aligned}$$

in the form $A\mathbf{x} = \mathbf{b}$.

According to the above, this system can be written as

$$\begin{pmatrix} 1 & 2 & -1 \\ 1 & -3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

The following is the most fundamental observation about multiplying a matrix times a vector.

Theorem 8.2.5 Let A be an $m \times n$ matrix and let \mathbf{x}, \mathbf{y} be two vectors in \mathbb{F}^n with a, b two scalars. Then

$$A(a\mathbf{x} + b\mathbf{y}) = aA\mathbf{x} + bA\mathbf{y}$$

Proof: By the above definition and the way we add vectors,

$$\begin{aligned}(A(a\mathbf{x} + b\mathbf{y}))_i &= \sum_j A_{ij}(ax_j + by_j) = a \sum_j A_{ij}x_j + b \sum_j A_{ij}y_j \\ &= a(A\mathbf{x})_i + b(A\mathbf{y})_i = (aA\mathbf{x} + bA\mathbf{y})_i\end{aligned}$$

Since the i^{th} entries coincide, it follows that $A(a\mathbf{x} + b\mathbf{y}) = aA\mathbf{x} + bA\mathbf{y}$ as claimed. ■

Definition 8.2.6 Define some special vectors \mathbf{e}_i as follows:

$$\mathbf{e}_i \equiv \overbrace{\begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}}^{1 \text{ in the } i^{\text{th}} \text{ position}}^T$$

Thus in \mathbb{F}^3 , we would have

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Observation 8.2.7 Let A be an $m \times n$ matrix. Then for $\mathbf{e}_i \in \mathbb{F}^n$, $A\mathbf{e}_i$ delivers the i^{th} column of A . To see this,

$$(A\mathbf{e}_i)_k = \sum_j A_{kj}(\mathbf{e}_i)_j = A_{ki}$$

because $(\mathbf{e}_i)_j = 0$ unless $j = i$ when it is 1. Thus, for k arbitrary, the k^{th} entry of $A\mathbf{e}_i$ is A_{ki} . Thus the result of multiplying by \mathbf{e}_i is

$$\begin{pmatrix} A_{1i} & A_{2i} & \cdots & A_{ni} \end{pmatrix}^T$$

which is indeed the i^{th} column. Another way to see this is to let

$$A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_i & \cdots & \mathbf{a}_n \end{pmatrix},$$

$$A\mathbf{e}_i = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_i & \cdots & \mathbf{a}_n \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = 1\mathbf{a}_i = \mathbf{a}_i$$

the i^{th} column of A .

8.3 Linear Transformations and Matrices

We can also refer to a linear transformation as a linear function. These are defined as follows.

Definition 8.3.1 Let T be a function defined on \mathbb{F}^n which takes vectors in \mathbb{F}^n to vectors in \mathbb{F}^m . This is written as $T : \mathbb{F}^n \rightarrow \mathbb{F}^m$. It is a linear function or equivalently linear transformation if it satisfies the following: For a, b scalars and \mathbf{x}, \mathbf{y} vectors in \mathbb{F}^n it follows that

$$T(a\mathbf{x} + b\mathbf{y}) = aT\mathbf{x} + bT\mathbf{y}$$

In words: It goes across addition and you can factor out scalars. Then notice that an $m \times n$ matrix A has the property that if \mathbf{x} is in \mathbb{F}^n then $A\mathbf{x}$ is in \mathbb{F}^m and by Theorem 8.2.5, if $T\mathbf{x} \equiv A\mathbf{x}$ for A an $m \times n$ matrix, then it follows that T is a linear function.

The following definition defines a linear function and notes that matrix multiplication gives an example of such a thing. The next theorem shows that this is the only way it can happen.

Theorem 8.3.2 Let T be a linear transformation, $T : \mathbb{F}^n \rightarrow \mathbb{F}^m$. Then there exists an $m \times n$ matrix A such that for all $\mathbf{x} \in \mathbb{F}^n$, you have $T\mathbf{x} = A\mathbf{x}$. This matrix is given by

$$\begin{pmatrix} T\mathbf{e}_1 & \cdots & T\mathbf{e}_n \end{pmatrix}$$

Proof: Let \mathbf{x} be arbitrary and $\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T$. Then

$$\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$$

It follows that, since T is linear,

$$T\mathbf{x} = T(x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n) = x_1T\mathbf{e}_1 + \cdots + x_nT\mathbf{e}_n = \begin{pmatrix} T\mathbf{e}_1 & \cdots & T\mathbf{e}_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

and so the matrix which does what is claimed is the one whose i^{th} column is $T\mathbf{e}_i$. That is

$$A\mathbf{x} = \begin{pmatrix} T\mathbf{e}_1 & \cdots & T\mathbf{e}_n \end{pmatrix} \mathbf{x} \blacksquare$$

8.4 Multiplication of Matrices

Say you have A an $m \times n$ matrix and B an $n \times p$ matrix. We want to define an $m \times p$ matrix called AB such that

$$(AB)\mathbf{x} = A(B\mathbf{x}) \quad (8.10)$$

In other words, we want the linear transformation determined by AB to be the same as first doing a linear transformation determined by B and then when this has been done, do the linear transformation determined by A to what you got. In short, we want matrix multiplication to correspond to composition of linear functions. Then 8.10 is satisfied if and only if

$$\begin{aligned} ((AB)\mathbf{x})_i &\equiv \sum_l (AB)_{il} x_l = (A(B\mathbf{x}))_i = \sum_k A_{ik} (B\mathbf{x})_k \\ &= \sum_k A_{ik} \sum_l B_{kl} x_l = \sum_k \sum_l A_{ik} B_{kl} x_l = \sum_l \left(\sum_k A_{ik} B_{kl} \right) x_l \end{aligned}$$

If this is to hold for all choices of \mathbf{x} , then it must also hold for $\mathbf{x} = \mathbf{e}_r$. Thus the i^{th} entry of the r^{th} column is $(AB)\mathbf{e}_r$ and using this, we obtain

$$(AB)_{ir} = \sum_l \left(\sum_k A_{ik} B_{kl} \right) (\mathbf{e}_r)_l = \sum_k A_{ik} B_{kr}$$

Thus if we have the requirement that matrix multiplication corresponds to composition of the corresponding linear transformations, we are forced to conclude the following definition for matrix multiplication.

Definition 8.4.1 Let A be $m \times n$ and B be $n \times p$. Then AB is $m \times p$ and the ir^{th} entry of (AB) is given by

$$(AB)_{ir} \equiv \sum_k A_{ik} B_{kr}$$

That is, the ir^{th} entry is the dot product of the i^{th} row of A with the r^{th} column of B .

Note that from this definition, you must have the number of columns of A equal to the number of rows of B in order to make any sense of the product. Indeed, this must be so when you consider matrix multiplication in terms of linear transformations. A linear transformation $T : \mathbb{F}^n \rightarrow \mathbb{F}^m$ is only defined on vectors in \mathbb{F}^n .

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 .

$$\begin{array}{c} \text{these must match!} \\ (m \times \widehat{n}) (\widehat{n} \times p) = m \times p \end{array}$$

Note the two outside numbers give the size of the product. Remember:

If the two middle numbers don't match, you can't multiply the matrices!

Example 8.4.2 Let

$$A = \begin{pmatrix} 1 & -1 & 2 \\ 3 & -2 & 1 \end{pmatrix}, B = \begin{pmatrix} 2 & 3 \\ -1 & 1 \\ 0 & 3 \end{pmatrix}$$

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.

$$AB = \begin{pmatrix} 3 & 8 \\ 8 & 10 \end{pmatrix}, BA = \begin{pmatrix} 11 & -8 & 7 \\ 2 & -1 & -1 \\ 9 & -6 & 3 \end{pmatrix}$$

Note that 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 juicy 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.

Although matrix multiplication is not commutative, it does have several very important properties.

Proposition 8.4.3 *If all multiplications and additions make sense, the following hold for matrices A, B, C and a, b scalars.*

$$A(aB + bC) = a(AB) + b(AC) \quad (8.11)$$

$$(B + C)A = BA + CA \quad (8.12)$$

$$A(BC) = (AB)C \quad (8.13)$$

Proof: Using Definition 8.4.1,

$$\begin{aligned} (A(aB + bC))_{ij} &= \sum_k A_{ik} (aB + bC)_{kj} \\ &= \sum_k A_{ik} (aB_{kj} + bC_{kj}) \\ &= a \sum_k A_{ik} B_{kj} + b \sum_k A_{ik} C_{kj} \\ &= a(AB)_{ij} + b(AC)_{ij} \\ &= (a(AB) + b(AC))_{ij}. \end{aligned}$$

Thus $A(B + C) = AB + AC$ as claimed. Formula 8.12 is entirely similar.

Formula 8.13 is the associative law of multiplication. Using Definition 8.4.1,

$$\begin{aligned} (A(BC))_{ij} &= \sum_k A_{ik} (BC)_{kj} \\ &= \sum_k A_{ik} \sum_l B_{kl} C_{lj} \\ &= \sum_l (AB)_{il} C_{lj} \\ &= ((AB)C)_{ij}. \end{aligned}$$

This proves 8.13. ■

Note that the claim about the associative law happens because when you have functions f, g, h such that it makes sense to take their composition in that order, we have $f \circ (g \circ h) = (f \circ g) \circ h$ and matrix multiplication corresponds to composition of the corresponding linear functions. This is the real reason for the associative law.

8.4.1 The Transpose

Another important operation on matrices is that of taking the **transpose**. The following example shows what is meant by this operation, denoted by placing a T as an exponent on the matrix.

$$\begin{pmatrix} 1 & 4 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 & 2 \\ 4 & 1 & 6 \end{pmatrix}$$

What happened? The first column became the first row and the second column became the second row. Thus the 3×2 matrix became a 2×3 matrix. The number 3 was in the second row and the first column and it ended up in the first row and second column. Here is the definition.

Definition 8.4.4 Let A be an $m \times n$ matrix. Then A^T denotes the $n \times m$ matrix which is defined as follows.

$$(A^T)_{ij} = A_{ji}$$

In words, the i^{th} row becomes the i^{th} column.

Example 8.4.5

$$\begin{pmatrix} 1 & 2 & -6 \\ 3 & 5 & 4 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 \\ 2 & 5 \\ -6 & 4 \end{pmatrix}.$$

The transpose of a matrix has the following important properties.

Lemma 8.4.6 Let A be an $m \times n$ matrix and let B be a $n \times p$ matrix. Then

$$(AB)^T = B^T A^T \quad (8.14)$$

and if α and β are scalars,

$$(\alpha A + \beta B)^T = \alpha A^T + \beta B^T \quad (8.15)$$

Proof: From the definition,

$$\begin{aligned} ((AB)^T)_{ij} &= (AB)_{ji} \\ &= \sum_k A_{jk} B_{ki} \\ &= \sum_k (B^T)_{ik} (A^T)_{kj} \\ &= (B^T A^T)_{ij} \end{aligned}$$

The proof of Formula 8.15 is left as an exercise and this proves the lemma. ■

Definition 8.4.7 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 8.4.8 *Let*

$$A = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 5 & -3 \\ 3 & -3 & 7 \end{pmatrix}.$$

Then A is symmetric.

Example 8.4.9 *Let*

$$A = \begin{pmatrix} 0 & 1 & 3 \\ -1 & 0 & 2 \\ -3 & -2 & 0 \end{pmatrix}$$

Then A is skew symmetric.

8.5 Some Examples of Linear Functions on \mathbb{R}^n

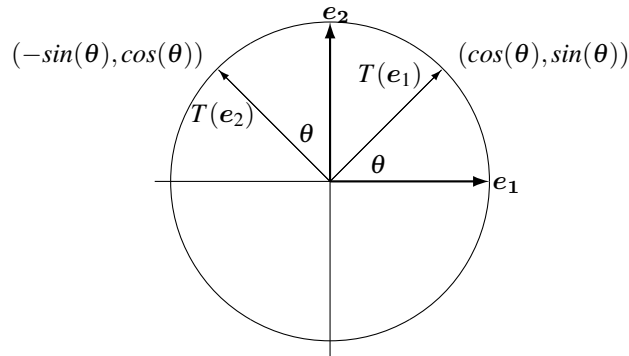
There are many examples of linear functions and we give a couple next.

8.5.1 Rotations in \mathbb{R}^2

Sometimes you need to find a matrix which represents a given linear transformation which is described in geometrical terms. The idea is to produce a matrix which you can multiply a vector by to get the same thing as some geometrical description. A good example of this is the problem of rotation of vectors discussed above. Consider the problem of rotating through an angle of θ .

Example 8.5.1 *Determine the matrix which represents the linear transformation defined by rotating every vector through an angle of θ .*

Let $e_1 \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $e_2 \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. These identify the geometric vectors which point along the positive x axis and positive y axis as shown.



From the above, you only need to find $T e_1$ and $T e_2$, the first being the first column of the desired matrix, A and the second being the second column. From the definition of the

\cos, \sin the coordinates of $T(e_1)$ are as shown in the picture. The coordinates of $T(e_2)$ also follow from simple trigonometry. Thus

$$Te_1 = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}, Te_2 = \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix}.$$

Therefore, from Theorem 8.3.2,

$$A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

For those who prefer a more algebraic approach, the definition of $(\cos(\theta), \sin(\theta))$ is as the x and y coordinates of the point $(1, 0)$. Now the point of the vector from $(0, 0)$ to $(0, 1)$, e_2 is exactly $\pi/2$ further along along the unit circle. Therefore, when it is rotated through an angle of θ the x and y coordinates are given by

$$(x, y) = (\cos(\theta + \pi/2), \sin(\theta + \pi/2)) = (-\sin \theta, \cos \theta).$$

Example 8.5.2 Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of ϕ and then through an angle θ . Thus you want the linear transformation which rotates all angles through an angle of $\theta + \phi$.

Let $T_{\theta+\phi}$ denote the linear transformation which rotates every vector through an angle of $\theta + \phi$. Then to get $T_{\theta+\phi}$, you could first do T_ϕ and then do T_θ where T_ϕ is the linear transformation which rotates through an angle of ϕ and T_θ is the linear transformation which rotates through an angle of θ . Denoting the corresponding matrices by $A_{\theta+\phi}$, A_ϕ , and A_θ , you must have for every x

$$A_{\theta+\phi}x = T_{\theta+\phi}x = T_\theta T_\phi x = A_\theta A_\phi x.$$

Consequently, you must have

$$\begin{aligned} A_{\theta+\phi} &= \begin{pmatrix} \cos(\theta + \phi) & -\sin(\theta + \phi) \\ \sin(\theta + \phi) & \cos(\theta + \phi) \end{pmatrix} = A_\theta A_\phi \\ &= \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}. \end{aligned}$$

You know how to multiply matrices. Do so to the pair on the right. This yields

$$\begin{aligned} &\begin{pmatrix} \cos(\theta + \phi) & -\sin(\theta + \phi) \\ \sin(\theta + \phi) & \cos(\theta + \phi) \end{pmatrix} \\ &= \begin{pmatrix} \cos \theta \cos \phi - \sin \theta \sin \phi & -\cos \theta \sin \phi - \sin \theta \cos \phi \\ \sin \theta \cos \phi + \cos \theta \sin \phi & \cos \theta \cos \phi - \sin \theta \sin \phi \end{pmatrix}. \end{aligned}$$

Don't these look familiar? They are the usual trig. identities for the sum of two angles derived here using linear algebra concepts.

You do not have to stop with two dimensions. You can consider rotations and other geometric concepts in any number of dimensions. This is one of the major advantages

of linear algebra. You can break down a difficult geometrical procedure into small steps, each corresponding to multiplication by an appropriate matrix. Then by multiplying the matrices, you can obtain a single matrix which can give you numerical information on the results of applying the given sequence of simple procedures. That which you could never visualize can still be understood to the extent of finding exact numerical answers. Another example follows.

Example 8.5.3 Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of $\pi/6$ and then reflecting through the x axis.

As shown in Example 8.5.2, the matrix of the transformation which involves rotating through an angle of $\pi/6$ is

$$\begin{pmatrix} \cos(\pi/6) & -\sin(\pi/6) \\ \sin(\pi/6) & \cos(\pi/6) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\sqrt{3} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}\sqrt{3} \end{pmatrix}$$

The matrix for the transformation which reflects all vectors through the x axis is

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

Therefore, the matrix of the linear transformation which first rotates through $\pi/6$ and then reflects through the x axis is

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{2}\sqrt{3} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}\sqrt{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\sqrt{3} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2}\sqrt{3} \end{pmatrix}.$$

8.5.2 Projections

In Physics it is important to consider the work done by a force field on an object. This involves the concept of projection onto a vector. Suppose you want to find the projection of a vector, v onto the given vector, u , denoted by $P_u(v)$. This is done using the dot product as follows.

$$P_u(v) = \left(\frac{v \cdot u}{u \cdot u} \right) u$$

Because of properties of the dot product, the map $v \rightarrow P_u(v)$ is linear,

$$\begin{aligned} P_u(\alpha v + \beta w) &= \left(\frac{\alpha v + \beta w \cdot u}{u \cdot u} \right) u = \alpha \left(\frac{v \cdot u}{u \cdot u} \right) u + \beta \left(\frac{w \cdot u}{u \cdot u} \right) u \\ &= \alpha P_u(v) + \beta P_u(w). \end{aligned}$$

Example 8.5.4 Let the projection map be defined above and let $u = (1, 2, 3)^T$. Does this linear transformation come from multiplication by a matrix? If so, what is the matrix?

You can find this matrix in the same way as in the previous example. Let e_i denote the vector in \mathbb{R}^n which has a 1 in the i^{th} position and a zero everywhere else. Thus a typical vector, $x = (x_1, \dots, x_n)^T$ can be written in a unique way as

$$x = \sum_{j=1}^n x_j e_j.$$

From the way you multiply a matrix by a vector, it follows that $P_u(e_i)$ gives the i^{th} column of the desired matrix. Therefore, it is only necessary to find

$$P_u(e_i) \equiv \left(\frac{e_i \cdot u}{u \cdot u} \right) u$$

For the given vector in the example, this implies the columns of the desired matrix are

$$\frac{1}{14} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \frac{2}{14} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \frac{3}{14} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Hence the matrix is

$$\frac{1}{14} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix}.$$

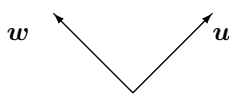
8.5.3 Rotations About A Particular Vector

The problem is to find the matrix of the linear transformation which rotates all vectors about a given unit vector u which is possibly not one of the coordinate vectors i, j , or k . Suppose for $|c| \neq 1$

$$u = (a, b, c), \quad \sqrt{a^2 + b^2 + c^2} = 1.$$

First I will produce a matrix which maps u to k such that the right handed rotation about k corresponds to the right handed rotation about u . Then I will rotate about k and finally, I will multiply by the inverse of the first matrix to get the desired result.

To begin, find vectors w, v such that $w \times v = u$. Let

$$w = \left(-\frac{b}{\sqrt{a^2 + b^2}}, \frac{a}{\sqrt{a^2 + b^2}}, 0 \right).$$


This vector is clearly perpendicular to u . Then $v = (a, b, c) \times w \equiv u \times w$. Thus from the geometric description of the cross product, $w \times v = u$. Computing the cross product gives

$$\begin{aligned} v &= (a, b, c) \times \left(-\frac{b}{\sqrt{a^2 + b^2}}, \frac{a}{\sqrt{a^2 + b^2}}, 0 \right) \\ &= \left(-c \frac{a}{\sqrt{(a^2 + b^2)}}, -c \frac{b}{\sqrt{(a^2 + b^2)}}, \frac{a^2}{\sqrt{(a^2 + b^2)}} + \frac{b^2}{\sqrt{(a^2 + b^2)}} \right) \end{aligned}$$

Now I want to have $Tw = i, Tv = j, Tu = k$. What does this? It is the inverse of the matrix which takes i to w, j to v , and k to u . This matrix is

$$\begin{pmatrix} -\frac{b}{\sqrt{a^2+b^2}} & -\frac{c}{\sqrt{a^2+b^2}}a & a \\ \frac{a}{\sqrt{a^2+b^2}} & -\frac{c}{\sqrt{a^2+b^2}}b & b \\ 0 & \frac{a^2+b^2}{\sqrt{a^2+b^2}} & c \end{pmatrix}.$$

Its inverse is

$$\begin{pmatrix} -\frac{1}{\sqrt{a^2+b^2}}b & \frac{1}{\sqrt{a^2+b^2}}a & 0 \\ -\frac{c}{\sqrt{a^2+b^2}}a & -\frac{c}{\sqrt{a^2+b^2}}b & \sqrt{a^2+b^2} \\ a & b & c \end{pmatrix}$$

Therefore, the matrix which does the rotating is

$$\begin{pmatrix} -\frac{b}{\sqrt{a^2+b^2}} & -\frac{c}{\sqrt{a^2+b^2}}a & a \\ \frac{a}{\sqrt{a^2+b^2}} & -\frac{c}{\sqrt{a^2+b^2}}b & b \\ 0 & \frac{a^2+b^2}{\sqrt{a^2+b^2}} & c \end{pmatrix} \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

$$\begin{pmatrix} -\frac{1}{\sqrt{a^2+b^2}}b & \frac{1}{\sqrt{a^2+b^2}}a & 0 \\ -\frac{c}{\sqrt{a^2+b^2}}a & -\frac{c}{\sqrt{a^2+b^2}}b & \sqrt{a^2+b^2} \\ a & b & c \end{pmatrix}$$

This yields a matrix whose columns are

$$\begin{pmatrix} \frac{b^2 \cos \theta + c^2 a^2 \cos \theta + a^4 + a^2 b^2}{a^2 + b^2} \\ \frac{-ba \cos \theta + cb^2 \sin \theta + ca^2 \sin \theta + c^2 ab \cos \theta + ba^3 + b^3 a}{a^2 + b^2} \\ -(\sin \theta)b - (\cos \theta)ca + ca \end{pmatrix},$$

$$\begin{pmatrix} \frac{-ba \cos \theta - ca^2 \sin \theta - cb^2 \sin \theta + c^2 ab \cos \theta + ba^3 + b^3 a}{a^2 + b^2} \\ \frac{a^2 \cos \theta + c^2 b^2 \cos \theta + a^2 b^2 + b^4}{a^2 + b^2} \\ (\sin \theta)a - (\cos \theta)cb + cb \end{pmatrix},$$

$$\begin{pmatrix} (\sin \theta)b - (\cos \theta)ca + ca \\ -(\sin \theta)a - (\cos \theta)cb + cb \\ (a^2 + b^2) \cos \theta + c^2 \end{pmatrix}$$

Using the assumption that \mathbf{u} is a unit vector so that $a^2 + b^2 + c^2 = 1$, it follows the desired matrix has the following as columns.

$$\begin{pmatrix} \cos \theta - a^2 \cos \theta + a^2 \\ -ba \cos \theta + ba + c \sin \theta \\ -(\sin \theta)b - (\cos \theta)ca + ca \end{pmatrix}, \begin{pmatrix} -ba \cos \theta + ba - c \sin \theta \\ -b^2 \cos \theta + b^2 + \cos \theta \\ (\sin \theta)a - (\cos \theta)cb + cb \end{pmatrix}$$

$$\begin{pmatrix} (\sin \theta)b - (\cos \theta)ca + ca \\ -(\sin \theta)a - (\cos \theta)cb + cb \\ (1 - c^2)\cos \theta + c^2 \end{pmatrix}$$

This was done under the assumption that $|c| \neq 1$. However, if this condition does not hold, you can verify directly that the above still gives the correct answer.

8.6 The Inverse of a Matrix

8.6.1 The Identity And Inverses

There is a special matrix called I and referred to as the identity matrix. It is always a square matrix, meaning the number of rows equals the number of columns and it has the property that there are ones down the main diagonal and zeroes elsewhere. Here are some identity matrices of various sizes.

$$(1), \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The first is the 1×1 identity matrix, the second is the 2×2 identity matrix, the third is the 3×3 identity matrix, and the fourth is the 4×4 identity matrix. By extension, you can likely see what the $n \times n$ identity matrix would be. It is so important that there is a special symbol to denote the ij^{th} entry of the identity matrix $I_{ij} = \delta_{ij}$ where δ_{ij} is the **Kronecker symbol** defined by

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

It is called the **identity matrix** because it is a **multiplicative identity** in the following sense.

Lemma 8.6.1 Suppose A is an $m \times n$ matrix and I_n is the $n \times n$ identity matrix. Then $AI_n = A$. If I_m is the $m \times m$ identity matrix, it also follows that $I_mA = A$.

Proof:

$$(AI_n)_{ij} = \sum_k A_{ik} \delta_{kj} = A_{ij}$$

and so $AI_n = A$. The other case is left as an exercise for you. ■

Definition 8.6.2 An $n \times n$ matrix A has an **inverse**, A^{-1} if and only if $AA^{-1} = A^{-1}A = I$. Such a matrix is called **invertible**.

It is very important to observe that the inverse of a matrix, if it exists, is unique. Another way to think of this is that if it acts like the inverse, then it is the inverse.

Theorem 8.6.3 Suppose A^{-1} exists and $AB = BA = I$. Then $B = A^{-1}$.

Proof:

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

Unlike ordinary multiplication of numbers, it can happen that $A \neq 0$ but A may fail to have an inverse. This is illustrated in the following example.

Example 8.6.4 Let $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Does A have an inverse?

One might think A would have an inverse because it does not equal zero. However,

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

and if A^{-1} existed, this could not happen because you could write

$$\begin{aligned} \begin{pmatrix} 0 \\ 0 \end{pmatrix} &= A^{-1} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) = A^{-1} \left(A \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right) = \\ &= (A^{-1}A) \begin{pmatrix} -1 \\ 1 \end{pmatrix} = I \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \end{aligned}$$

a contradiction. Thus the answer is that A does not have an inverse.

Example 8.6.5 Let $A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. Show $\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ is the inverse of A .

To check this, multiply

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

and

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

showing that this matrix is indeed the inverse of A .

8.6.2 Finding The Inverse Of A Matrix

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

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

This requires the solution of the systems of equations,

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

and

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

Writing the augmented matrix for these two systems gives

$$\left(\begin{array}{cc|c} 1 & 1 & 1 \\ 1 & 2 & 0 \end{array} \right) \quad (8.16)$$

for the first system and

$$\left(\begin{array}{cc|c} 1 & 1 & 0 \\ 1 & 2 & 1 \end{array} \right) \quad (8.17)$$

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

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

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

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

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

Now solve the second system, 8.17 to find z and w . Take (-1) times the first row and add to the second to get

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

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

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

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

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

Didn't the above seem rather repetitive? Note that exactly the same row operations were used in both systems. In each case, the end result was something of the form $(I|v)$

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

To simplify this procedure, you could have written

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

and row reduced till you obtained

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

and read off the inverse as the 2×2 matrix on the right side.

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

PROCEDURE 8.6.6 Suppose A is an $n \times n$ matrix. To find A^{-1} if it exists, form the augmented $n \times 2n$ matrix

$$(A|I)$$

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

$$(I|B). \quad (8.18)$$

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

Actually, all this shows is how to find a right inverse if it exists. What has been shown from the above discussion is that $AB = I$. Later, I will show that this right inverse is **the** inverse. See Corollary 28.1.15 presented later. However, it is not hard to see that this should be the case as follows.

The row operations are all reversible. If the row operation involves switching two rows, the reverse row operation involves switching them again to get back to where you started. If the row operation involves multiplying a row by $a \neq 0$, then you would get back to where you began by multiplying the row by $1/a$. The third row operation involving addition of c times row i to row j can be reversed by adding $-c$ times row i to row j .

In the above procedure, a sequence of row operations applied to I yields B while the same sequence of operations applied to A yields I . Therefore, the sequence of reverse row operations in the opposite order applied to B will yield I and applied to I will yield A . That is, there are row operations which provide

$$(B|I) \rightarrow (I|A)$$

and as just explained, A must be a right inverse for B . Therefore, $BA = I$. Hence B is both a right and a left inverse for A because $AB = BA = I$.

If it is impossible to row reduce $(A|I)$ to get $(I|B)$, then in particular, it is impossible to row reduce A to I and consequently impossible to do a sequence of row operations to I and get A . Later it will be made clear that the only way this can happen is that it is possible to row reduce A to a matrix of the form $\begin{pmatrix} C \\ \mathbf{0} \end{pmatrix}$ where $\mathbf{0}$ is a row of zeros. Then there will be no solution to the system of equations represented by the augmented matrix

$$\left(\begin{array}{c|c} C & \mathbf{0} \\ \hline \mathbf{0} & 1 \end{array} \right)$$

Using the reverse row operations in the opposite order on both matrices in the above, it follows that there must exist \mathbf{a} such that there is no solution to the system of equations represented by $(A|\mathbf{a})$. Hence A fails to have an inverse, because if it did, then there would be a solution \mathbf{x} to the equation $A\mathbf{x} = \mathbf{a}$ given by $A^{-1}\mathbf{a}$.

Example 8.6.7 Let $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{pmatrix}$. Find A^{-1} if it exists.

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

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

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

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

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

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

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

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

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

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

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

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

Therefore, the inverse is

$$\begin{pmatrix} -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{pmatrix}$$

Example 8.6.8 Let $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4 \end{pmatrix}$. Find A^{-1} if it exists.

Write the augmented matrix $(A|I)$

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 2 & 2 & 4 & 0 & 0 & 1 \end{array} \right)$$

and proceed to do row operations attempting to obtain $(I|A^{-1})$. Take (-1) times the top row and add to the second. Then take (-2) times the top row and add to the bottom.

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

Next add (-1) times the second row to the bottom row.

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

At this point, you can see there will be no inverse because you have obtained a row of zeros in the left half of the augmented matrix $(A|I)$. Thus there will be no way to obtain I on the left.

Example 8.6.9 Let $A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$. Find A^{-1} if it exists.



Form the augmented matrix

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

Now do row operations until the $n \times n$ matrix on the left becomes the identity matrix. This yields after some computations,

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

and so the inverse of A is the matrix on the right,

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

Checking the answer is easy. Just multiply the matrices and see if it works.

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

Always check your answer because if you are like some of us, you will usually have made a mistake.

Example 8.6.10 *In this example, it is shown how to use the inverse of a matrix to find the solution to a system of equations. Consider the following system of equations. Use the inverse of a suitable matrix to give the solutions to this system.*

$$\begin{pmatrix} x + z = 1 \\ x - y + z = 3 \\ x + y - z = 2 \end{pmatrix}.$$

The system of equations can be written in terms of matrices as

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}. \quad (8.19)$$

More simply, this is of the form $A\mathbf{x} = \mathbf{b}$. Suppose you find the inverse of the matrix A^{-1} . Then you could multiply both sides of this equation by A^{-1} to obtain

$$\mathbf{x} = (A^{-1}A)\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{b}.$$

This gives the solution as $\mathbf{x} = A^{-1}\mathbf{b}$. Note that once you have found the inverse, you can easily get the solution for different right hand sides without any effort. It is always just $A^{-1}\mathbf{b}$. In the given example, the inverse of the matrix is

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

This was shown in Example 8.6.9. Therefore, from what was just explained, the solution to the given system is

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & -1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ -2 \\ -\frac{3}{2} \end{pmatrix}.$$

What if the right side of 8.19 had been $\begin{pmatrix} 0 & 1 & 3 \end{pmatrix}^T$? What would be the solution to

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}?$$

By the above discussion, it is just

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & -1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix}.$$

This illustrates why once you have found the inverse of a given matrix, you can use it to solve many different systems easily.

8.7 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.

`>>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 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.

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.

8.8 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:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 \\ 4 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$

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. Let $A = \begin{pmatrix} 1 & 1 \\ -2 & -1 \\ 1 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 1 & -1 & -2 \\ 2 & 1 & -2 \end{pmatrix}$, and $C = \begin{pmatrix} 1 & 1 & -3 \\ -1 & 2 & 0 \\ -3 & -1 & 0 \end{pmatrix}$.

Find if possible.

- (a) AB
- (b) BA
- (c) AC
- (d) CA
- (e) CB
- (f) BC

6. 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) $(AB)^2 = A^2B^2$
- (c) $(A + B)^2 = A^2 + 2AB + B^2$
- (d) $(A + B)^2 = A^2 + AB + BA + B^2$
- (e) $A^2B^2 = A(AB)B$
- (f) $(A + B)^3 = A^3 + 3A^2B + 3AB^2 + B^3$
- (g) $(A + B)(A - B) = A^2 - B^2$

7. Let $A = \begin{pmatrix} -1 & -1 \\ 3 & 3 \end{pmatrix}$. Find all 2×2 matrices, B such that $AB = 0$.

8. 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.

9. 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?

10. 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?

11. In 8.1 - 8.8 describe $-A$ and 0 .

12. 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.

13. 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.

14. Suppose M is a 3×3 skew symmetric matrix. Show there exists a vector Ω such that for all $u \in \mathbb{R}^3$

$$Mu = \Omega \times u$$

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 $\omega_1 i + \omega_2 j + \omega_3 k$.

15. Using only the properties 8.1 - 8.8 show $-A$ is unique.
16. Using only the properties 8.1 - 8.8 show 0 is unique.
17. Using only the properties 8.1 - 8.8 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.
18. Using only the properties 8.1 - 8.8 and previous problems show $(-1)A = -A$.
19. Prove 8.15.
20. Prove that $I_m A = A$ where A is an $m \times n$ matrix.
21. Give an example of matrices, A, B, C such that $B \neq C$, $A \neq 0$, and yet $AB = AC$.
22. 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?
23. 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 $AC = AD$.
24. Explain why if $AB = AC$ and A^{-1} exists, then $B = C$.
25. Give an example of a matrix A such that $A^2 = I$ and yet $A \neq I$ and $A \neq -I$.
26. Give an example of matrices, A, B such that neither A nor B equals zero and yet $AB = 0$.
27. Give another example other than the one given in this section of two square matrices, A and B such that $AB \neq BA$.
28. Let

$$A = \begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix}.$$

Find A^{-1} if possible. If A^{-1} does not exist, determine why.

29. Let

$$A = \begin{pmatrix} 0 & 1 \\ 5 & 3 \end{pmatrix}.$$

Find A^{-1} if possible. If A^{-1} does not exist, determine why.

30. Let

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 0 \end{pmatrix}.$$

Find A^{-1} if possible. If A^{-1} does not exist, determine why.

31. Let

$$A = \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}.$$

Find A^{-1} if possible. If A^{-1} does not exist, determine why.

32. 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 .

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. Using the inverse of the matrix, find the solution to the systems

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \\ -2 \end{pmatrix}.$$

Now give the solution in terms of a, b , and c to

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

41. Using the inverse of the matrix, find the solution to the systems

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \\ -2 \end{pmatrix}.$$

Now give the solution in terms of a, b , and c to

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

42. Using the inverse of the matrix, find the solution to the system

$$\begin{pmatrix} -1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 3 & \frac{1}{2} & -\frac{1}{2} & -\frac{5}{2} \\ -1 & 0 & 0 & 1 \\ -2 & -\frac{3}{4} & \frac{1}{4} & \frac{9}{4} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

43. 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 $\mathbf{x} = A^{-1}\mathbf{b}$.
44. Prove that if A^{-1} exists and $A\mathbf{x} = \mathbf{0}$ then $\mathbf{x} = \mathbf{0}$.
45. 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 $B = A^{-1}$.
46. Show that if A is an invertible $n \times n$ matrix, then so is A^T and $(A^T)^{-1} = (A^{-1})^T$.
47. 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 45.

48. 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 45.

49. If A is invertible, show $(A^2)^{-1} = (A^{-1})^2$. **Hint:** Use Problem 45.

50. If A is invertible, show $(A^{-1})^{-1} = A$. **Hint:** Use Problem 45.

51. 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, $A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A^T\mathbf{y}$. Use the definition of matrix multiplication to do this.

52. Use the result of Problem 51 to verify directly that $(AB)^T = B^TA^T$ without making any reference to subscripts.

53. 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 $|(A^2)_{ij}| \leq R^2$. Then generalize to show that $|(A^m)_{ij}| \leq R^m$. 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 x of the form

$$x = Ax + b, \text{ or } (I - A)x = b$$

The vector Ax is called the intermediate demand and the vectors $A^k x$ have economic meaning. From the above,

$$x = Ib + Ab + A^2b + \cdots$$

The series is also called the Neuman series. It is important in functional analysis.

54. An elementary matrix is one which results from doing a row operation to the identity matrix. Thus the elementary matrix E which results from adding a times the i^{th} row to the j^{th} row would have $a\delta_{ik} + \delta_{jk}$ as the jk^{th} entry and all other rows would be unchanged. That is δ_{rs} provided $r \neq j$. Show that multiplying this matrix on the left of an appropriate sized matrix A results in doing the row operation to the matrix A . You might also want to verify that the other elementary matrices have the same effect, doing the row operation which resulted in the elementary matrix to A .
55. Let a be a fixed vector. The function T_a defined by $T_a v = a + v$ has the effect of translating all vectors by adding a . Show this is not a linear transformation. Explain why it is not possible to realize T_a in \mathbb{R}^3 by multiplying by a 3×3 matrix.
56. In spite of Problem 55 we can represent both translations and rotations 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 $v = (v_1, v_2, v_3)^T$, 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 rotations and translations all at once by forming appropriate 4×4 matrices.

Chapter 9

Subspaces Spans and Bases

The span of some vectors consists of all linear combinations of these vectors. A linear combination of vectors is just a finite sum of scalars times vectors.

Definition 9.0.1 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 9.0.2 Let $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ be vectors in \mathbb{F}^n . Form the $n \times p$ matrix

$$A \equiv \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix}$$

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

$$\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = x_1 \mathbf{u}_1 + \cdots + x_p \mathbf{u}_p$$

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 9.0.3 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 $A(\mathbb{F}^p)$. Thus column space equals $\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_p)$ where the \mathbf{u}_i are the columns of A .

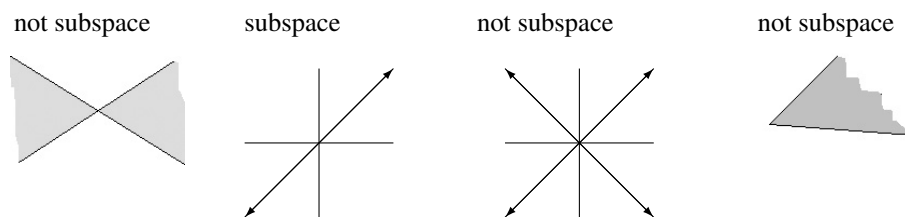
What do you really mean when you say there is a solution \mathbf{x} to a linear system of equations $A\mathbf{x} = \mathbf{b}$? You mean that \mathbf{b} is in the span of the columns of A . After all, if $A = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix}$, you are looking for $\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_p \end{pmatrix}^T$ such that $x_1 \mathbf{u}_1 + x_2 \mathbf{u}_2 + \cdots + x_p \mathbf{u}_p = A\mathbf{x} = \mathbf{b}$.

9.1 Subspaces

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 9.1.1 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 9.1.2 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 55.

Proposition 9.1.3 *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. ■

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 9.1.4 *Let A be an $m \times n$ matrix where $m < n$. Then $N(A)$ contains nonzero vectors.*

Proof: First consider the case where A is a $1 \times n$ matrix for $n > 1$. Say

$$A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}$$

If $a_1 = 0$, consider the vector $\mathbf{x} = \mathbf{e}_1$. If $a_1 \neq 0$, let

$$\mathbf{x} = \begin{pmatrix} b \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

where b is chosen to satisfy the equation

$$a_1 b + \sum_{k=2}^n a_k = 0$$

Suppose now that the theorem is true for any $m \times n$ matrix with $n > m$ and consider an $(m \times 1) \times n$ matrix A where $n > m + 1$. If the first column of A is $\mathbf{0}$, then you could let $\mathbf{x} = \mathbf{e}_1$ as above. If the first column is not the zero vector, then by doing row operations, the equation $A\mathbf{x} = \mathbf{0}$ can be reduced to the equivalent system

$$A_1 \mathbf{x} = \mathbf{0}$$

where A_1 is of the form

$$A_1 = \begin{pmatrix} 1 & \mathbf{a}^T \\ \mathbf{0} & B \end{pmatrix}$$

where B is an $m \times (n-1)$ matrix. Since $n > m+1$, it follows that $(n-1) > m$ and so by induction, there exists a nonzero vector $\mathbf{y} \in \mathbb{F}^{n-1}$ such that $B\mathbf{y} = \mathbf{0}$. Then consider the vector

$$\mathbf{x} = \begin{pmatrix} b \\ \mathbf{y} \end{pmatrix}$$

$A_1\mathbf{x}$ has for its top entry the expression $b + \mathbf{a}^T\mathbf{y}$. Letting $B = \begin{pmatrix} b_1^T \\ \vdots \\ b_m^T \end{pmatrix}$, the i^{th} entry of $A_1\mathbf{x}$

for $i > 1$ is of the form $b_i^T\mathbf{y} = 0$. Thus if b is chosen to satisfy the equation $b + \mathbf{a}^T\mathbf{y} = 0$, then $A_1\mathbf{x} = \mathbf{0}$. ■

Now here is a very fundamental definition.

Definition 9.1.5 Let $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ be vectors in \mathbb{F}^n . 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. Thus Theorem 9.1.4 says that if you have more than n vectors in \mathbb{F}^n this set of vectors will be dependent.

With this preparation, here is a major theorem.

Theorem 9.1.6 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: Let $\mathbf{u}_i = \sum_{j=1}^s a_{ji}\mathbf{v}_j$. This is merely giving names to the scalars in the linear combination which yields \mathbf{u}_i . Now suppose that $s < r$. Then if A is the matrix which has a_{ji} in the j^{th} row and the i^{th} column, it follows from Theorem 9.1.4 that there exists a vector in \mathbb{F}^r such that $A\mathbf{x} = \mathbf{0}$ but $\mathbf{x} \neq \mathbf{0}$. However, then

$$\begin{aligned} \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_r \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_r \end{pmatrix} &= \sum_{i=1}^r x_i \mathbf{u}_i = \sum_{i=1}^r x_i \sum_{j=1}^s a_{ji} \mathbf{v}_j \\ &= \sum_{j=1}^s \sum_{i=1}^r a_{ji} x_i \mathbf{v}_j = \sum_{j=1}^s (A\mathbf{x})_j \mathbf{v}_j \\ &= \sum_{j=1}^s 0 \mathbf{v}_j = \mathbf{0} \end{aligned}$$

Which contradicts the assertion that the set of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is linearly independent. Indeed, there is a nonzero vector \mathbf{x} for which $\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_r \end{pmatrix} \mathbf{x} = \mathbf{0}$. Thus we cannot have $s < r$ and so the only other possibility is that $s \geq r$. ■

Definition 9.1.7 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 9.1.8 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 9.1.6, $r \leq s$ because each \mathbf{u}_i is in the span of $\{\mathbf{v}_1, \dots, \mathbf{v}_s\}$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$ is independent. Then also $r \geq s$ by the same reasoning. ■

Definition 9.1.9 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 9.1.8.

Observation 9.1.10 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

$$\mathbf{0} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T$$

and so each $x_i = 0$. Thus $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ 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 9.1.11 Suppose $\mathbf{v} \notin \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is linearly independent. Then $\{\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{v}\}$ 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 \mathbf{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 subspace equals the span of some vectors. This is the content of the next theorem.

Theorem 9.1.12 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 9.1.11, $\{\mathbf{u}_1, \mathbf{u}_2\}$ 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.

Theorem 9.1.13 *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 $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ contrary to Theorem 9.1.6.

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 9.1.11, 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 9.1.11. 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 to the list and get an independent set which is longer than a spanning set contrary to Theorem 9.1.6. 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 9.1.11, $\{\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{v}_{k+1}\}$ 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 9.1.6. ■

Definition 9.1.14 *The rank of a matrix A is the dimension of $\text{Im}(A)$ which is the same as the column space of A .*

Observation 9.1.15 *When you have a matrix A and you do row operations to it. The solutions to the system of equations having augmented matrix $(A|\mathbf{0})$ are unchanged. This was the entire basis for using row operations to solve systems of equations which was presented earlier. Thus, if you can row reduce a matrix and obtain one for which it is clear that all columns are in the span of certain columns and that these certain columns form an independent set, then you have found the rank. You just need to count the number of these special columns. Actually, the row reduced echelon form is designed to do this very thing.*

Example 9.1.16 *Let $A =$*

$$\begin{pmatrix} 1 & 0 & -5 & -7 & -3 \\ 1 & 1 & 1 & -1 & 1 \\ 0 & 1 & 6 & 6 & 4 \end{pmatrix} \quad (9.1)$$

Determine its rank.

The row reduced echelon form for the above matrix is

$$\begin{pmatrix} 1 & 0 & -5 & -7 & -3 \\ 0 & 1 & 6 & 6 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (9.2)$$

and so its rank is 2 because every column is in the span of the first two columns. You can think of the above as the row reduced version of several systems of equations, those which have the following augmented matrices.

$$\begin{pmatrix} 1 & 0 & -5 \\ 1 & 1 & 1 \\ 0 & 1 & 6 \end{pmatrix}, \begin{pmatrix} 1 & 0 & -7 \\ 1 & 1 & -1 \\ 0 & 1 & 6 \end{pmatrix}, \begin{pmatrix} 1 & 0 & -3 \\ 1 & 1 & 1 \\ 0 & 1 & 4 \end{pmatrix},$$

In each case, you can obtain the third column as a linear combination of the first two. Thus the last three columns in 9.1 are linear combinations of the first two columns in 9.1. Therefore, any linear combination of the columns of 9.1 can also be written as a linear combination of the first two columns of 9.1. In other words, the span of the columns of 9.1 equals the span of the first two columns of 9.1. Also, from 9.2, we can see that the first two columns of 9.1 are independent. Therefore, these columns are a basis for $\text{Im}(A)$.

Similar considerations apply to determining whether some vectors are independent. Remember the definition. To determine whether some vectors are independent, make them the columns of a matrix A and determine the solution set to $A\mathbf{x} = \mathbf{0}$. If there is only the zero solution, then the vectors are independent. If there are more solutions then these vectors are not independent.

Theorem 9.1.17 *Let A be an $n \times n$ matrix. Then A^{-1} exists if and only if the rank of A equals n .*

Proof: If the rank of A is n , then no column is a linear combination of the others because, by definition, the columns are independent. In particular, one can row reduce A to obtain I . Hence row reduction will do $(A|I) \rightarrow (I|B)$ and B is the inverse. Perhaps a better way to see this is to note that the columns are independent because the span of the columns has dimension n . Hence no column can be deleted and have the shorter list still span the column space. Thus every vector in \mathbb{R}^n is in the column space. It follows for each $\mathbf{b} \in \mathbb{R}^n$, there exists \mathbf{x} such that $A\mathbf{x} = \mathbf{b}$. Thus A maps onto \mathbb{R}^n . Considered as a linear transformation, A is onto and it is also one to one because if $A\mathbf{x} = \mathbf{b}$, $A\hat{\mathbf{x}} = \mathbf{b}$, then $A(\mathbf{x} - \hat{\mathbf{x}}) = \mathbf{0}$ and so $\mathbf{x} - \hat{\mathbf{x}} = \mathbf{0}$ since otherwise, there would be a non trivial linear combination of the columns of A which is $\mathbf{0}$ contrary to the observation that the columns are independent. Thus you can define a linear transformation, denoted as A^{-1} by $A^{-1}(A\mathbf{x}) \equiv \mathbf{x}$ and $A(A^{-1}\mathbf{x}) = \mathbf{x}$. Then the matrix of A^{-1} , still denoted as A^{-1} is the desired inverse.

If A^{-1} exists, then it is obvious the columns of A are independent because a typical linear combination of these columns is of the form $A\mathbf{x}$. If it equals $\mathbf{0}$ then $A^{-1}(A\mathbf{x}) = (A^{-1}A)\mathbf{x} = \mathbf{x} = \mathbf{0}$. Thus the rank is n . ■

9.2 Exercises

1. 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 that

$$\{Ax : x \in P(u_1, \dots, u_n)\}$$

is also a parallelepiped.

2. In the context of Problem 1, draw $P(e_1, e_2)$ where e_1, e_2 are the standard basis vectors for \mathbb{R}^2 . Thus $e_1 = (1, 0)$, $e_2 = (0, 1)$. Now suppose

$$E = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

where E is the elementary matrix which takes the third row and adds to the first. Draw

$$\{Ex : x \in P(e_1, e_2)\}.$$

In other words, draw the result of doing E to the vectors in $P(e_1, e_2)$. Next draw the results of doing the other elementary matrices to $P(e_1, e_2)$. An elementary matrix is one which is obtained from doing one of the row operations to the identity matrix.

3. Determine which matrices are in row reduced echelon form.

$$\begin{array}{ll} \text{(a)} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 7 \end{pmatrix} & \text{(c)} \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 5 \\ 0 & 0 & 1 & 2 & 0 & 4 \\ 0 & 0 & 0 & 0 & 1 & 3 \end{pmatrix} \\ \text{(b)} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix} & \end{array}$$

4. Row reduce the following matrices to obtain the row reduced echelon form. List the pivot columns in the original matrix.

$$\begin{array}{ll} \text{(a)} \begin{pmatrix} 1 & 2 & 0 & 3 \\ 2 & 1 & 2 & 2 \\ 1 & 1 & 0 & 3 \end{pmatrix} & \text{(c)} \begin{pmatrix} 1 & 2 & 1 & 3 \\ -3 & 2 & 1 & 0 \\ 3 & 2 & 1 & 1 \end{pmatrix} \\ \text{(b)} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -2 \\ 3 & 0 & 0 \\ 3 & 2 & 1 \end{pmatrix} & \end{array}$$

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.

$$\text{(a)} \begin{pmatrix} 1 & 2 & 0 \\ 3 & 2 & 1 \\ 2 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}$$

$$(b) \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 1 \\ 2 & 1 & 0 \\ 0 & 2 & 0 \end{pmatrix}$$

$$(d) \begin{pmatrix} 0 & 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 3 & 2 & 6 & 0 & 5 & 4 \\ 0 & 1 & 1 & 2 & 0 & 2 & 2 \\ 0 & 2 & 1 & 4 & 0 & 3 & 2 \end{pmatrix}$$

$$(c) \begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 2 & 2 \\ 0 & 3 & 2 & 12 & 1 & 6 & 8 \\ 0 & 1 & 1 & 5 & 0 & 2 & 3 \\ 0 & 2 & 1 & 7 & 0 & 3 & 4 \end{pmatrix}$$

$$(e) \begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 3 & 2 & 6 & 1 & 5 & 1 \\ 0 & 1 & 1 & 2 & 0 & 2 & 1 \\ 0 & 2 & 1 & 4 & 0 & 3 & 1 \end{pmatrix}$$

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 $A(BA - I) = \mathbf{0}$ 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.

$$\begin{pmatrix} 1 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 0 & 1 \end{pmatrix}^T, \begin{pmatrix} 3 & 0 & 0 \end{pmatrix}^T$$

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 whether 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 whether 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 whether 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. Determine whether 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 .

$$\begin{pmatrix} 1 & 0 & 3 \end{pmatrix}^T, \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T, \begin{pmatrix} 1 & 2 & 0 \end{pmatrix}^T$$

27. Determine whether 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 .

$$\begin{pmatrix} 1 & 0 & 3 \end{pmatrix}^T, \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T, \begin{pmatrix} 1 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^T$$

28. Determine whether 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 .

$$\begin{pmatrix} 1 & 0 & 3 \end{pmatrix}^T, \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T, \begin{pmatrix} 1 & 1 & 3 \end{pmatrix}^T, \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^T$$

29. 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.

30. 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.

31. 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.

32. 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.
33. If you have 6 vectors in \mathbb{F}^5 , is it possible they are linearly independent? Explain.
34. Suppose A is an $m \times n$ matrix and $\{\mathbf{w}_1, \dots, \mathbf{w}_k\}$ is a linearly independent set of vectors in $A(\mathbb{F}^n) \subseteq \mathbb{F}^m$. Now suppose $A(\mathbf{z}_i) = \mathbf{w}_i$. Show $\{\mathbf{z}_1, \dots, \mathbf{z}_k\}$ is also independent.

35. 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.
36. 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.
37. 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.
38. If $\mathbf{b} \neq \mathbf{0}$, can the solution set of $A\mathbf{x} = \mathbf{b}$ be a plane through the origin? Explain.
39. 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.
40. 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.
41. 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.
42. Suppose there is a unique solution to a system of linear equations. What must be true of the pivot columns in the augmented matrix.
43. State whether each of the following sets of data are possible for the matrix equation $A\mathbf{x} = \mathbf{b}$. 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 $\text{rank}(A|\mathbf{b}) = 4$. **Hint:** This says \mathbf{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 $\text{rank}(A|\mathbf{b}) = 2$.
 - (c) A is a 4×2 matrix, $\text{rank}(A) = 4$ and $\text{rank}(A|\mathbf{b}) = 4$. **Hint:** This says \mathbf{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 $\text{rank}(A|\mathbf{b}) = 5$. **Hint:** This says \mathbf{b} is not in the span of the columns.
 - (e) A is a 4×2 matrix, $\text{rank}(A) = 2$ and $\text{rank}(A|\mathbf{b}) = 2$.
44. 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 $\mathbf{e}_1, \dots, \mathbf{e}_m$ occur as columns in the row reduced echelon form for A .
45. 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 \mathbf{x} such that $A\mathbf{x} = \mathbf{0}$, 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 .
46. Explain why an $n \times n$ matrix A is both one to one and onto if and only if its rank is n .

47. 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

$$x - \sum_{i=1}^k z_i \in \ker(B).$$

48. 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?
49. 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.
50. 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?

Chapter 10

Eigenvalues and Eigenvectors

10.1 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.

Definition 10.1.1 *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 it can be an eigenvalue.*

Now here is an important observation which really is just a re statement of the above definition.

Theorem 10.1.2 *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 9.1.17 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. Also, it gives an introduction to the important topic of determinants which will be presented in more detail later.

10.2 An Introduction to Determinants

Here in this section, I will summarize the main properties of determinants without detailed proofs. Proofs are presented later in the book. The idea is that you get used to using them first.

10.2.1 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 10.2.1 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 10.2.2 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 10.2.3 Suppose A is a 3×3 matrix. The ij^{th} **minor**, denoted 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 10.2.4 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 10.2.5 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 10.2.6 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 10.2.7 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 10.2.8 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”.

$$\overbrace{1(-1)^{1+1} \det \begin{pmatrix} 3 & 2 \\ 2 & 1 \end{pmatrix}}^{\text{cof}(A)_{11}} + \overbrace{4(-1)^{2+1} \det \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}}^{\text{cof}(A)_{21}} + \overbrace{3(-1)^{3+1} \det \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}}^{\text{cof}(A)_{31}} = 0.$$

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.

$$\overbrace{4(-1)^{2+1} \det \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}}^{\text{cof}(A)_{21}} + \overbrace{3(-1)^{2+2} \det \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}}^{\text{cof}(A)_{22}} + \overbrace{2(-1)^{2+3} \det \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}}^{\text{cof}(A)_{23}} = 0.$$

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 10.2.9 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 $(-1)^{i+j} \times$ (the ij^{th} 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 28 which is presented much later. This is summarized in the following theorem whose conclusion is incredible.

Theorem 10.2.10 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 10.2.11 Expand $\det \begin{pmatrix} 1 & 2 & -1 & 1 \\ 2 & 3 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 3 & 1 \end{pmatrix}$ along first column.

It is

$$1 \det \begin{pmatrix} 3 & 1 & 1 \\ 1 & 0 & 0 \\ 2 & 3 & 1 \end{pmatrix} - 2 \det \begin{pmatrix} 2 & -1 & 1 \\ 1 & 0 & 0 \\ 2 & 3 & 1 \end{pmatrix}$$

$$+1 \det \begin{pmatrix} 2 & -1 & 1 \\ 3 & 1 & 1 \\ 2 & 3 & 1 \end{pmatrix} - 1 \det \begin{pmatrix} 2 & -1 & 1 \\ 3 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} = 0$$

10.2.2 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 10.2.12 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 10.2.13 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 10.2.14 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

$$\begin{aligned} & 1 \det \begin{pmatrix} 2 & 6 & 7 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{pmatrix} + 0(-1)^{2+1} \det \begin{pmatrix} 2 & 3 & 77 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{pmatrix} + \\ & 0(-1)^{3+1} \det \begin{pmatrix} 2 & 3 & 77 \\ 2 & 6 & 7 \\ 0 & 0 & -1 \end{pmatrix} + 0(-1)^{4+1} \det \begin{pmatrix} 2 & 3 & 77 \\ 2 & 6 & 7 \\ 0 & 3 & 33.7 \end{pmatrix} \end{aligned}$$

and the only nonzero term in the expansion is

$$1 \det \begin{pmatrix} 2 & 6 & 7 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{pmatrix}.$$

Now expand this along the first column to obtain

$$\begin{aligned}
 & 1 \times \left(2 \times \det \begin{pmatrix} 3 & 33.7 \\ 0 & -1 \end{pmatrix} + 0(-1)^{2+1} \det \begin{pmatrix} 6 & 7 \\ 0 & -1 \end{pmatrix} \right. \\
 & \quad \left. + 0(-1)^{3+1} \det \begin{pmatrix} 6 & 7 \\ 3 & 33.7 \end{pmatrix} \right) \\
 &= 1 \times 2 \times \det \begin{pmatrix} 3 & 33.7 \\ 0 & -1 \end{pmatrix}
 \end{aligned}$$

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.

10.2.3 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 10.2.15 *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 10.2.16 *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 10.2.17 *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 10.2.18 *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 10.2.19 *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 10.2.20 *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 10.2.21 *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 10.2.22 *In Theorems 10.2.16 - 10.2.20 you can replace the word, “row” with the word “column”.*

There are two other major properties of determinants which do not involve row operations.

Theorem 10.2.23 *Let A and B be two $n \times n$ matrices. Then*

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

Also,

$$\det(A) = \det(A^T).$$

Example 10.2.24 *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 $\det(A) \det(B) = 8 \times (-5) = -40$.

10.2.4 Finding Determinants Using Row Operations

Theorems 10.2.20 - 10.2.22 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 10.2.25 *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 10.2.20, 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 10.2.20 this didn't change the value of the determinant. Then the last row was multiplied by (-3) . By Theorem 10.2.16 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 10.2.20. 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

$$\det(D) = 1(-3) \det \begin{pmatrix} 11 & 22 \\ 14 & -17 \end{pmatrix} = 1485$$

and so $\det(C) = -1485$ and $\det(A) = \det(B) = \left(\frac{-1}{3}\right)(-1485) = 495$.

Example 10.2.26 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 10.2.20 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 10.2.22 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 10.2.20 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.

$$\det \begin{pmatrix} 0 & -1 & -1 \\ -8 & -4 & 1 \\ 10 & -8 & -4 \end{pmatrix}$$

$$\text{which equals } 8 \det \begin{pmatrix} -1 & -1 \\ -8 & -4 \end{pmatrix} + 10 \det \begin{pmatrix} -1 & -1 \\ -4 & 1 \end{pmatrix} = -82$$

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.

10.3 Applications

10.3.1 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 8.6.2 on Page 144. Also recall the definition of the cofactor matrix given in Definition 10.2.9 on Page 176. 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 Section 27.2.1.

Theorem 10.3.1 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 10.3.2 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 10.2.20 - 10.2.22 on Page 178, the determinant of this matrix equals the determinant of the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -8 \\ 0 & 0 & -2 \end{pmatrix}$$

which equals 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 it is correct.

Example 10.3.3 Find the inverse of the matrix

$$A = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{6} & \frac{1}{3} & -\frac{1}{2} \\ -\frac{5}{6} & \frac{2}{3} & -\frac{1}{2} \end{pmatrix}$$

First find its determinant. This determinant is $\frac{1}{6}$. I will replace each entry in the above matrix with the cofactor corresponding to the position of that entry. Then I will take the transpose and multiply by 6. You should check that the result is as follows.

$$6 \begin{pmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{6} & -\frac{1}{3} \\ -\frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}^T.$$

This yields

$$6 \begin{pmatrix} 1/6 & 1/3 & 1/6 \\ 1/3 & 1/6 & -1/3 \\ -1/6 & 1/6 & 1/6 \end{pmatrix}^T = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix}$$

Always check your work.

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1/2 & 0 & 1/2 \\ -1/6 & 1/3 & -1/2 \\ -5/6 & 2/3 & -1/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

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.

10.3.2 Finding Eigenvalues Using Determinants

Theorem 10.3.1 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 \mathbf{x} such that $A\mathbf{x} = \lambda\mathbf{x}$ where λ is called an eigenvalue. Thus you have $(A - \lambda I)\mathbf{x} = \mathbf{0}$ for $\mathbf{x} \neq \mathbf{0}$. If $(A - \lambda I)^{-1}$ were to exist, then you could multiply by it on the left and obtain $\mathbf{x} = \mathbf{0}$ 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 11.5.2 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 10.3.4 *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.*

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

$$\begin{aligned} \det(A - \lambda I) &= \det(S^{-1}BS - \lambda S^{-1}IS) = \det(S^{-1}(B - \lambda I)S) \\ &= \det(S^{-1}) \det(B - \lambda I) \det(S) = \det(S^{-1}S) \det(B - \lambda I) \\ &= \det(I) \det(B - \lambda I) = \det(B - \lambda I) \end{aligned} \tag{10.1}$$

Chapter 11

Matrices and The Inner Product

Recall the inner product or dot product.

$$\mathbf{a} \cdot \mathbf{b} \equiv \sum_k a_k \overline{b_k}$$

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”. Thus, in this chapter, we will adopt the notation (\mathbf{a}, \mathbf{b}) for the dot product. The first thing to consider is the notion of the adjoint of a matrix.

Definition 11.0.1 Let A be an $m \times n$ matrix. Then its adjoint, denoted as A^* is the transpose of the conjugate of A . That is, you replace each entry of A with its complex conjugate and take the transpose of what you got. Thus

$$\begin{pmatrix} i & 2 & 1+i \\ 3 & 1-i & 1 \end{pmatrix}^* = \begin{pmatrix} -i & 3 \\ 2 & 1+i \\ 1-i & 1 \end{pmatrix}$$

In symbols, $(A^*)_{rs} = \overline{A_{sr}}$. **Note that** $(A^*)^* = A$.

The reason the adjoint is so important is the following proposition which in fact can be used as a definition of the adjoint instead of the above explicit description in terms of entries.

Proposition 11.0.2 Let A be an $m \times n$ matrix and let $\mathbf{x} \in \mathbb{F}^m$ and $\mathbf{y} \in \mathbb{F}^n$. Then

$$(\mathbf{x}, A\mathbf{y}) = (A^*\mathbf{x}, \mathbf{y})$$

Also, if B is an $n \times m$ matrix such that the above holds for all $\mathbf{x} \in \mathbb{F}^m$ and $\mathbf{y} \in \mathbb{F}^n$, then $B = A^*$.

Proof: This follows directly from the definition of the inner product and the properties of the complex conjugate which were reviewed in Section 2.3.

$$(\mathbf{x}, A\mathbf{y}) = \sum_k x_k \overline{(A\mathbf{y})_k} = \sum_k x_k \overline{\sum_j A_{kj} y_j} = \sum_k x_k \sum_j \overline{A_{kj}} \overline{y_j}$$

$$= \sum_j \sum_k A_{jk}^* x_k \overline{y_j} = \sum_j (A^* x)_j \overline{y_j} = (A^* x, y)$$

Now suppose for all x, y $(x, Ay) = (Bx, y)$. Then you have $(A^* x, y) = (Bx, y)$ for all x, y and so $(A^* x - Bx, y) = 0$ for all x, y . In particular this holds for $y = A^* x - Bx$. Thus $A^* x - Bx = 0$ for each x . Hence $A^* = B$. To see this, note that $(A^* - B)e_j$ says that the j^{th} column of $A^* - B$ is zero. ■

The last part of this argument deserves a little more emphasis. If you have an $m \times n$ matrix M , then M is the zero matrix if and only if $Mx = 0$ for all $x \in \mathbb{F}^n$. In other words, to show something is zero, you show it sends every vector to 0 . Equivalently, a matrix M is **not zero** if and only if there is some vector x for which $Mx \neq 0$.

11.1 Eigenvalues and Eigenvectors

Here I will consider eigenvectors and eigenvalues from a different point of view. Rather than determinants, this approach depends on what is really fundamental about linear algebra, linear independence. Consider the collection of $n \times n$ matrices consisting of complex numbers $M_{n \times n}$. Now consider the special matrices E_{ij} which has a 1 in the i^{th} row and j^{th} column and zeros in all other positions. Then according to the way we add and multiply matrices by scalars, $M_{n \times n}$ can be considered as \mathbb{C}^{n^2} . For example, consider the case where

$n = 2$. Instead of writing $\begin{pmatrix} 1 & 2 & 3 & 4 \end{pmatrix}$, we bend it and write it as $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. Thus,

if a_{ij} is the entry in $i j^{\text{th}}$ position of one of these matrices, then we can recover A by forming the sum

$$\sum_i \sum_j a_{ij} E_{ij} \quad (11.1)$$

Thus these E_{ij} span the $n \times n$ matrices which, as just noted, can be considered as vectors in \mathbb{C}^{n^2} . Furthermore, these matrices E_{ij} are independent because if the above sum in 11.1 equals 0, then since a_{ij} is the entry in the $i j^{\text{th}}$ position, it follows that $a_{ij} = 0$. Thus these matrices are a basis for $M_{n \times n}$ and the dimension of $M_{n \times n} = n^2$ which we already knew from the above identification of $M_{n \times n}$ with \mathbb{C}^{n^2} .

Define $A^0 \equiv I$ for any matrix $A \in M_{n \times n}$. Consider the matrices

$$I, A, A^2, \dots, A^{n^2}$$

There are $n^2 + 1$ of these matrices and so they can't be independent. Hence there are scalars c_0, \dots, c_{n^2} not all zero such that

$$c_0 I + c_1 A + \dots + c_{n^2} A^{n^2} = 0$$

In other words, there is a polynomial

$$q(\lambda) = c_m \lambda^m + \dots + c_1 \lambda + c_0$$

such that

$$q(A) \equiv c_m A^m + \dots + c_1 A + c_0 I = 0 \text{ in } M_{n \times n}$$

Out of all such polynomials, let $\hat{p}(\lambda)$ be one which has the smallest degree. Denote by $p(\lambda)$ the polynomial which results by dividing by the leading coefficient. Thus $p(\lambda)$ is the monic polynomial of smallest degree which has $p(A) = 0$.

Definition 11.1.1 *The minimum polynomial for an $n \times n$ matrix A is the polynomial which has smallest degree and is monic such that $p(A) = 0$. In fact, it is unique and you might think about why this is so using the division algorithm.*

Now we can give the definition of eigenvalues and eigenvectors. Recall it is as follows.

Definition 11.1.2 *Let A be an $n \times n$ matrix. A **NONZERO VECTOR** x is said to be an eigenvector for A if there is some number λ such that*

$$Ax = \lambda x$$

This number is called an eigenvalue.

It turns out that every $n \times n$ matrix has an eigenvalue and that in fact every root of the minimum polynomial is an eigenvalue. Recall that by the fundamental theorem of algebra, (See Section 2.9), the minimum polynomial has a root. In fact, we can completely factor the minimum polynomial.

Proposition 11.1.3 *Let $p(\lambda)$ be a monic polynomial of degree $m \geq 1$ having complex coefficients. Then there are complex numbers μ_1, \dots, μ_m , possibly not all distinct such that*

$$p(\lambda) = (\lambda - \mu_1)(\lambda - \mu_2) \cdots (\lambda - \mu_m)$$

Proof: If $m = 1$, there is nothing to show. Suppose then that the Proposition is true for some $m \geq 1$ and suppose $p(\lambda)$ is a monic polynomial of degree $m + 1$.

From the fundamental theorem of algebra, there is a root to $p(\lambda)$. Denote this root as μ_1 . From Lemma 2.8.2, the division algorithm,

$$p(\lambda) = (\lambda - \mu_1)k(\lambda) + r(\lambda)$$

where the degree of $r(\lambda)$ is less than 1 or else $r(\lambda) = 0$. Thus $r(\lambda) = r$. However, if we evaluate both sides at $\lambda = \mu_1$ we get $p(\mu_1) = 0 = r$ and so

$$p(\lambda) = (\lambda - \mu_1)k(\lambda)$$

Now $k(\lambda)$ is also a monic polynomial which can be seen by comparing the leading coefficient of both sides and it has degree m . Therefore, by induction, there are m complex numbers $\mu_2, \mu_3, \dots, \mu_{m+1}$ such that

$$p(\lambda) = (\lambda - \mu_1)k(\lambda) = p(\lambda) = (\lambda - \mu_1) \cdots (\lambda - \mu_{m+1})$$

Thus the proposition holds for $m = 1$ and if it holds for m , then it also holds for $m + 1$. Therefore, it is valid for any positive integer m . ■

Theorem 11.1.4 *Let A be an $n \times n$ matrix. Let its minimum polynomial be $p(\lambda)$. Let*

$$p(\lambda) = (\lambda - \mu_1) \cdots (\lambda - \mu_m)$$

Then for each μ_j , there is an eigenvector x_j such that $Ax_j = \mu_j x_j$.

Proof: First note that $IB = BI$ for any square matrix B . Next note that

$$0 = p(A) = (A - \mu_1 I) \cdots (A - \mu_m I) \quad (11.2)$$

Also note that

$$(A - \mu I)(A - \lambda I) = A^2 - (\lambda + \mu)A + \mu\lambda I = (A - \lambda I)(A - \mu I)$$

Thus all the factors in the above product 11.2 can be interchanged and thereby placed in any order in the product. We know that for any y ,

$$(A - \mu_j I) \left[(A - \mu_1 I) \cdots (A - \mu_{j-1} I) (A - \mu_{j+1} I) \cdots (A - \mu_m I) \right] y = 0$$

However, there is some y_j such that

$$(A - \mu_1 I) \cdots (A - \mu_{j-1} I) (A - \mu_{j+1} I) \cdots (A - \mu_m I) y_j \neq 0$$

since otherwise, $p(\lambda)$ didn't really have smallest degree. Then let

$$x_j = (A - \mu_1 I) \cdots (A - \mu_{j-1} I) (A - \mu_{j+1} I) \cdots (A - \mu_m I) y_j \blacksquare$$

The minimum polynomial can be computed although it might seem a little tedious. In the above discussion, the minimum polynomial is only known to have degree no more than n^2 . Actually it can be shown that the degree of the minimum polynomial is never more than n although it might be less than n . We will show this later as part of the theory of the determinant but in the meantime, one should go ahead and use it. Here is an example.

Example 11.1.5 *Let*

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$$

Find its minimum polynomial.

The matrices are $I, \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}^2$. These will end up being linearly dependent. They are $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 5 & 5 \\ 5 & 10 \end{pmatrix}$. The polynomial is obtained by finding a linear combination of these equal to 0. Lets make these into column vectors and use row operations.

$$\begin{pmatrix} 1 & 2 & 5 \\ 0 & 1 & 5 \\ 0 & 1 & 5 \\ 1 & 3 & 10 \end{pmatrix}$$

Now we row reduce this to get

$$\begin{pmatrix} 1 & 0 & -5 \\ 0 & 1 & 5 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Thus, as explained earlier, the last column is -5 times the first added to 5 times the second. Thus

$$\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}^2 = -5 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 5 \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$$

You can see from the row reduced echelon form that no smaller linear combination relating the matrices I, A, A^2 is possible. Hence the minimal polynomial is

$$\lambda^2 - 5\lambda + 5$$

The eigenvalues are therefore, the roots of this polynomial. They are

$$\frac{5}{2} + \frac{1}{2}\sqrt{5}, \frac{5}{2} - \frac{1}{2}\sqrt{5}$$

Now one can find eigenvectors associated with these. Consider the first of them. We want a nonzero vector x such that $(A - (\frac{5}{2} + \frac{1}{2}\sqrt{5})I)x = 0$. Thus we need consider the augmented matrix

$$\begin{pmatrix} 2 - (\frac{5}{2} + \frac{1}{2}\sqrt{5}) & 1 & 0 \\ 1 & 3 - (\frac{5}{2} + \frac{1}{2}\sqrt{5}) & 0 \end{pmatrix}$$

We row reduce this to obtain

$$\begin{pmatrix} 1 & \frac{1}{2} - \frac{1}{2}\sqrt{5} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Thus the eigenvectors are of the form

$$y \begin{pmatrix} \frac{1}{2}\sqrt{5} - \frac{1}{2} \\ 1 \end{pmatrix}, y \in \mathbb{C}$$

Example 11.1.6 Find the minimum polynomial for

$$A = \begin{pmatrix} 0 & -2 & -2 \\ 2 & 5 & 4 \\ -1 & -2 & -1 \end{pmatrix}$$

We look for linear combinations for A^0, A, A^2, A^3 . These are the matrices, listed in order of decreasing powers are

$$\begin{pmatrix} -6 & -14 & -14 \\ 14 & 29 & 28 \\ -7 & -14 & -13 \end{pmatrix}, \begin{pmatrix} -2 & -6 & -6 \\ 6 & 13 & 12 \\ -3 & -6 & -5 \end{pmatrix}, \begin{pmatrix} 0 & -2 & -2 \\ 2 & 5 & 4 \\ -1 & -2 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

We can arrange them as column vectors in \mathbb{C}^9 as done earlier, but it might be easier to simply look at the entries in a single row or column. Lets pick the first column of each. Thus the augmented matrix to solve would be

$$\begin{pmatrix} 1 & 0 & -2 & -6 & 0 \\ 0 & 2 & 6 & 14 & 0 \\ 0 & -1 & -3 & -7 & 0 \end{pmatrix}$$

Row reduce this.

$$\begin{pmatrix} 1 & 0 & -2 & -6 & 0 \\ 0 & 1 & 3 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Thus column one of A^2 equals -2 times column one of I added to three times column one of A . Thus consider the polynomial $\lambda^2 - 3\lambda + 2$. This seems to work in so far as the first column of A is concerned and there is no polynomial of smaller degree which will work. Therefore, let's check to see if this sends A to 0. If it does, then it must be the minimal polynomial. When you do the computations, you find that this indeed does send A to 0 and so it is the minimum polynomial. The eigenvalues are 1 and 2. You can now find the eigenvectors for these using row operations.

11.2 Using Matlab

It is routine to find this polynomial and so it is not surprising that MATLAB is able to do it for you. I recommend doing this rather than all the trouble just described. The syntax to use is this:

```
>> A=[1,2,3;3,-3,1;2,7,1]; minpoly(A) Here you press enter. It gives:
1 1 -24 -69
```

These are the coefficients of the minimum polynomial which is

$$\lambda^3 + \lambda^2 - 24\lambda - 69$$

The matrix you entered was

$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & -3 & 1 \\ 2 & 7 & 1 \end{pmatrix}$$

You open MATLAB and you see `>>`. Then type in just what is above. The `;` at the end after entering the matrix says for MATLAB to know the matrix but not to rewrite it. You can of course follow the same pattern to enter any square matrix you like. Then of course you are faced with the problem of finding the roots of the polynomial. Sometimes you can't do this exactly. Neither can MATLAB. However, when the polynomial can be factored, MATLAB can do it for you. Here is the syntax.

```
>> syms x
```

```
factor(x^2-3*x+2) (here you press enter and what results is:)
```

```
[x-1, x-2]
```

To get to a new line in MATLAB you press shift enter. You factored $x^2 - 3x + 2$. You can enter any polynomial you like, but sometimes they can't be factored exactly. When this happens, MATLAB will just return the original polynomial. This is its way of saying that it has no idea how to do it.

11.3 Distance and Unitary Matrices

Some matrices preserve lengths of vectors. That is $|Ux| = |x|$ for any x in \mathbb{C}^n . Such a matrix is called unitary. 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 Problem 19 on Page 168. The traditional definition of unitary is as follows.

Definition 11.3.1 Let $U \in M_{n \times n}$. Then U is called unitary if $U^*U = UU^* = I$. When U consists entirely of real entries, a unitary matrix is called an orthogonal matrix.

Then the following proposition relates this to preservation of lengths of vectors.

Proposition 11.3.2 An $n \times n$ matrix U is unitary 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{C}^n and let $\theta \in \mathbb{C}$, $|\theta| = 1$, and $\theta(U^*Uu - u, v) = |(U^*Uu - u, v)|$. Therefore from the axioms of the inner product,

$$\begin{aligned} |u|^2 + |v|^2 + 2\operatorname{Re} \theta(u, v) &= |\theta u|^2 + |v|^2 + \theta(u, v) + \bar{\theta}(v, u) \\ &= |\theta u + v|^2 = (U(\theta u + v), U(\theta u + v)) \\ &= (U\theta u, U\theta u) + (Uv, Uv) + (U\theta u, Uv) + (Uv, U\theta u) \\ &= |\theta u|^2 + |v|^2 + \theta(U^*Uu, v) + \bar{\theta}(v, U^*Uu) \\ &= |u|^2 + |v|^2 + 2\operatorname{Re} \theta(U^*Uu, v) \end{aligned}$$

and so, subtracting the ends, it follows that for all u, v ,

$$0 = 2\operatorname{Re} \theta(U^*Uu - u, v) = 2|(U^*Uu - u, v)|$$

from the above choice of θ . Now let $v = U^*Uu - u$. It follows that

$$U^*Uu - u = (U^*U - I)u = 0.$$

This is true for all u and so $U^*U = I$. Thus it is also true that $UU^* = I$. One can use the fact shown in Problem 19 on Page 168.

Conversely, if $U^*U = I$, then

$$|Uu|^2 = (Uu, Uu) = (U^*Uu, u) = (u, u) = |u|^2$$

Thus U preserves distance. ■

11.4 Schur's Theorem

The most significant theorem about eigenvalues and eigenvectors in the space of $n \times n$ complex matrices is Schur's theorem. First is a simple version of the Gram Schmidt theorem.

Definition 11.4.1 A set of vectors in \mathbb{F}^n , $\mathbb{F} = \mathbb{R}$ or \mathbb{C} , $\{x_1, \dots, x_k\}$ is called an **orthonormal** set of vectors if

$$\overline{x_i}^T x_j = x_i^* x_j = \delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Note this is the same as saying that $(x_i, x_j) = \delta_{ij}$ although here it will be slightly more convenient to define the inner product differently. Indeed, we are really working with the inner product $\langle x, y \rangle = x^* y$ whereas the usual inner product is $(x, y) = x^T \bar{y}$. This alternate version of the inner product is actually more convenient in matrix theory so we use it here. The difference is that with this new version, the complex conjugate comes out of the first entry rather than the second.

What does it mean to say that $U^* U = I$ which is the definition for U to be unitary? This says that for $U = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}$, $U^* = \begin{pmatrix} \overline{u_1}^T \\ \vdots \\ \overline{u_n}^T \end{pmatrix}$ and so from the way we multiply matrices in which the i^{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

$$u_i^* u_j = \delta_{ij}$$

in other words, the columns of U are orthonormal. From this simple observation, we get the following important theorem.

Theorem 11.4.2 *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 unitary. Thus if $Ux = 0$, you can multiply on the left on both sides with U^* and obtain $x = U^* Ux = U^* 0 = 0$. Thus, from the definition of linear independence, Definition 9.1.5, it follows that the columns of U comprise an independent set of vectors. ■

Theorem 11.4.3 *Let v_1 be a unit vector ($|v_1| = 1$) in \mathbb{R}^n , $n > 1$. Then there exist vectors*

$$\{v_2, \dots, v_n\}$$

such that this set of vectors is an orthonormal set of vectors.

Proof: The equation for x , $\overline{v_1}^T x = 0$ has a nonzero solution x by Theorem 9.1.4. Pick such a solution and divide by its magnitude to get v_2 a unit vector such that $\overline{v_1}^T \cdot v_2 = 0$. Now suppose v_1, \dots, v_k have been chosen such that $\{v_1, \dots, v_k\}$ is an orthonormal set of vectors. Then consider the equations

$$\overline{v_j}^T x = 0 \quad j = 1, 2, \dots, k$$

This amounts to the situation of Theorem 9.1.4 in which there are more variables than equations. Therefore, by this theorem, there exists a nonzero x solving all these equations. Divide by its magnitude and this gives v_{k+1} . Continue this way. At the last step, you obtain v_n and the resulting set is an orthonormal set. ■

Thus, as observed above, the matrix $\begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$ is a unitary matrix. With this preparation, here is Schur's theorem. First is some terminology. An $n \times n$ matrix T is called

upper triangular if it is of the form

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

meaning that all entries are zero below the main diagonal, consisting of those entries of the form T_{ii} .

Theorem 11.4.4 *Let A be a real or complex $n \times n$ matrix. Then there exists a unitary matrix U such that*

$$U^*AU = T, \quad (11.3)$$

where T is an upper triangular matrix. If A has all real entries and eigenvalues, then U can be chosen to be orthogonal.

Proof: The theorem is clearly true if A is a 1×1 matrix. Just let $U = 1$ the 1×1 matrix which has 1 down the main diagonal and zeros elsewhere. Suppose it is true for $(n-1) \times (n-1)$ matrices and let A be an $n \times n$ matrix. Then let v_1 be a unit eigenvector for A . That is, there exists λ_1 such that

$$Av_1 = \lambda_1 v_1, \quad |v_1| = 1.$$

By Theorem 11.4.3 there exists $\{v_1, \dots, v_n\}$, an orthonormal set in \mathbb{C}^n . Let U_0 be a matrix whose i^{th} column is v_i . Then from the above, it follows U_0 is unitary. Then from the way you multiply matrices $U_0^*AU_0$ is of the form

$$\begin{pmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_n^* \end{pmatrix} \begin{pmatrix} \lambda_1 v_1 & Av_2 & \cdots & Av_n \end{pmatrix} = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

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

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

an upper triangular matrix. Consider

$$U_1 \equiv \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \tilde{U}_1 \end{pmatrix}$$

From the way we multiply matrices, this is a unitary matrix and

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

where T is upper triangular. Then let $U = U_0 U_1$. Both of the U_i are unitary and so U must also be unitary. Indeed

$$U^* U = (U_0 U_1)^* U_0 U_1 = U_1^* U_0^* U_0 U_1 = U_1^* U_1 = I.$$

Then $U^* A U = T$.

If A is real having real eigenvalues, all of the above can be accomplished using the real dot product and using real eigenvectors. Thus the unitary matrix can be assumed real. ■

The diagonal entries of T are each eigenvalues of A . This will become clear later when we discuss the determinant and the characteristic polynomial. However, it is clear right now that T and A have the same eigenvalues. If $Tx = \lambda x$ for nonzero x , then

$$\begin{aligned} U^* A U x &= \lambda U^* U x \\ U^* (A U x - \lambda U x) &= 0 \end{aligned}$$

Now multiply both sides by U and obtain that Ux is an eigenvector for A . It is nonzero because U preserves lengths. Similar reasoning shows that every eigenvalue of A is an eigenvalue of T . Thus one obtains the following important corollary.

Corollary 11.4.5 *Let A be an $n \times n$ matrix. Then $\det(A)$ equals the product of the eigenvalues of A .*

Proof: Let $U^* A U = T$ where T is upper triangular. Then

$$\text{product of eigenvalues of } A = \text{product of eigenvalues of } T = \det(T) = \det(A)$$

The reason for the last equality is that from Theorem 10.2.23,

$$\det(T) = \det(U^T A U) = \det(U^T) \det(A) \det(U) = \det(U^T) \det(U) \det(A)$$

Now $U^T U = I$ and so $\det(U^T) \det(U) = \det(I) = 1$. ■

The following result is about Hermitian matrices. These are those matrices for which the upper triangular matrix in Schur's theorem is actually a real diagonal matrix.

Definition 11.4.6 *An $n \times n$ matrix A is Hermitian if $A = A^*$. Thus a real symmetric matrix is Hermitian but so is*

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

In this book, we are mainly interested in real symmetric matrices.

The next theorem is the main result.

Theorem 11.4.7 *If A is an $n \times n$ Hermitian matrix, there exists a unitary matrix U such that*

$$U^* A U = D \tag{11.4}$$

where D is a real diagonal matrix. That is, D has nonzero entries only on the main diagonal and these are real. Furthermore, the columns of U are an orthonormal basis of eigenvectors for \mathbb{C}^n . If A is real and symmetric, then U can be assumed to be a real orthogonal matrix and the columns of U form an orthonormal basis for \mathbb{R}^n . Furthermore, if A is an $n \times n$ matrix and there is a unitary matrix U such that $U^* A U = D$ where D is real and diagonal, then A is Hermitian.

Proof: From Schur's theorem above, there exists U unitary (real and orthogonal if A is real) such that

$$U^*AU = T$$

where T is an upper triangular matrix. Then from the rules for the transpose,

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

Thus $T = T^*$ and T is upper triangular. This can only happen if T is really a diagonal matrix having real entries on the main diagonal. (If $i \neq j$, one of T_{ij} or T_{ji} equals zero. But $T_{ij} = \overline{T_{ji}}$ and so they are both zero. Also $T_{ii} = \overline{T_{ii}}$.)

Finally, let

$$U = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{pmatrix}$$

where the \mathbf{u}_i denote the columns of U and

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

The equation, $U^*AU = D$ implies

$$\begin{aligned} AU &= \begin{pmatrix} A\mathbf{u}_1 & A\mathbf{u}_2 & \cdots & A\mathbf{u}_n \end{pmatrix} \\ &= UD = \begin{pmatrix} \lambda_1\mathbf{u}_1 & \lambda_2\mathbf{u}_2 & \cdots & \lambda_n\mathbf{u}_n \end{pmatrix} \end{aligned}$$

where the entries denote the columns of AU and UD respectively. Therefore, $A\mathbf{u}_i = \lambda_i\mathbf{u}_i$ and since the matrix is unitary, the ij^{th} entry of U^*U equals δ_{ij} and so

$$\delta_{ij} = \overline{\mathbf{u}_i}^T \mathbf{u}_j = \overline{\mathbf{u}_i^T \mathbf{u}_j} = \overline{(\mathbf{u}_i, \mathbf{u}_j)}$$

This proves the corollary because it shows the vectors $\{\mathbf{u}_i\}$ form an orthonormal basis. In case A is real and symmetric, simply ignore all complex conjugations in the above argument.

Finally suppose that $U^*AU = D$ where D is real and diagonal. Thus $D^* = D$. Then

$$A = UDU^*$$

Thus $A^* = UD^*U^* = UDU^* = A$. This last uses the fact that $(AB)^* = B^*A^*$. ■

Example 11.4.8 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 and then normalizing the vectors, 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

$$\begin{pmatrix} \frac{1}{6}\sqrt{6} & -\frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \\ \frac{1}{3}\sqrt{6} & 0 & -\frac{1}{3}\sqrt{3} \\ \frac{1}{6}\sqrt{6} & \frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \end{pmatrix}^T \begin{pmatrix} 1 & -4 & 13 \\ -4 & 10 & -4 \\ 13 & -4 & 1 \end{pmatrix}.$$

$$\begin{pmatrix} \frac{1}{6}\sqrt{6} & -\frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \\ \frac{1}{3}\sqrt{6} & 0 & -\frac{1}{3}\sqrt{3} \\ \frac{1}{6}\sqrt{6} & \frac{1}{2}\sqrt{2} & \frac{1}{3}\sqrt{3} \end{pmatrix} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & -12 & 0 \\ 0 & 0 & 18 \end{pmatrix}$$

11.5 Diagonalization

Theorem 11.4.7 is a special case of something known as diagonalization.

Definition 11.5.1 An $n \times n$ matrix A is diagonalizable if there exists an invertible matrix S such that

$$S^{-1}AS = D$$

where D is a diagonal matrix.

The following theorem gives the condition under which a matrix is diagonalizable.

Theorem 11.5.2 An $n \times n$ matrix S is invertible if and only if its columns are linearly independent.

Proof: First note that if S is $n \times n$ and its columns are linearly independent, then these columns must also span all of \mathbb{F}^n since otherwise, there would be a vector \mathbf{v} not in the span and you could add it in to the list and get $n + 1$ vectors in an independent set. This is contrary to Theorem 9.1.6. Thus $\text{Im}(S) = \mathbb{F}^n$. Also, if $S\mathbf{x} = S\mathbf{y}$, then $S(\mathbf{x} - \mathbf{y}) = \mathbf{0}$ and so $\mathbf{x} = \mathbf{y}$ (S is one to one.). Thus we can define $S^{-1}\mathbf{y}$ to be that vector such that $S(S^{-1}\mathbf{y}) = \mathbf{y}$. Then S^{-1} is a linear transformation because of the following reasoning.

$$S(S^{-1}(a\mathbf{x} + b\mathbf{y})) = a\mathbf{x} + b\mathbf{y}$$

$$\begin{aligned} S(aS^{-1}\mathbf{x} + bS^{-1}\mathbf{y}) &= aS(S^{-1}\mathbf{x}) + bS(S^{-1}\mathbf{y}) \\ &= a\mathbf{x} + b\mathbf{y} \end{aligned}$$

Thus, since S is one to one, as explained above, it follows that

$$S^{-1}(a\mathbf{x} + b\mathbf{y}) = aS^{-1}\mathbf{x} + bS^{-1}\mathbf{y}$$

Therefore, there is a matrix, still denoted as S^{-1} such that for any $\mathbf{x} \in \mathbb{F}^n$, $S(S^{-1}\mathbf{x}) = (SS^{-1})\mathbf{x} = \mathbf{x}$. Hence $SS^{-1} = I$. By Problem 19 on Page 168, $S^{-1}S = I$ also. Alternatively, $S(S^{-1}S) = (SS^{-1})S = S$. Hence for all \mathbf{x} , $S(S^{-1}S)\mathbf{x} = S\mathbf{x}$ and so $S(S^{-1}S\mathbf{x} - I\mathbf{x}) = \mathbf{0}$ and so, since S is one to one, $S^{-1}S\mathbf{x} = I\mathbf{x}$ for all \mathbf{x} showing that $S^{-1}S = I$ also. ■

Thus if the columns of a matrix are linearly independent, then the matrix is invertible. On the other hand, if the matrix S is invertible, then if $S\mathbf{x} = \mathbf{0}$ one could multiply both sides by S^{-1} and obtain $\mathbf{x} = \mathbf{0}$ and so the columns of S are linearly independent.

Theorem 11.5.3 *An $n \times n$ matrix is diagonalizable if and only if \mathbb{F}^n has a basis of eigenvectors of A . Furthermore, you can take the matrix S described above, to be given as*

$$S = \begin{pmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_n \end{pmatrix}$$

where here the \mathbf{s}_k are the eigenvectors in the basis for \mathbb{F}^n . If A is diagonalizable, the eigenvalues of A are the diagonal entries of the diagonal matrix.

Proof: To say that A is diagonalizable, is to say that

$$S^{-1}AS = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

the λ_i being elements of \mathbb{F} . This is to say that for $S = \begin{pmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{pmatrix}$, \mathbf{s}_k being the k^{th} column,

$$A \begin{pmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{pmatrix} = \begin{pmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

which is equivalent, from the way we multiply matrices, that

$$\begin{pmatrix} A\mathbf{s}_1 & \cdots & A\mathbf{s}_n \end{pmatrix} = \begin{pmatrix} \lambda_1\mathbf{s}_1 & \cdots & \lambda_n\mathbf{s}_n \end{pmatrix}$$

which is equivalent to saying that the columns of S are eigenvectors and the diagonal matrix has the eigenvalues down the main diagonal. Since S^{-1} is invertible, these eigenvectors are a basis. Similarly, if there is a basis of eigenvectors, one can take them as the columns of S and reverse the above steps, finally concluding that A is diagonalizable. ■

11.6 Approximations

11.6.1 Fredholm Alternative

First is a useful proposition which tells when there is a solution to a system of equations.

$$Ax = b \quad (11.5)$$

It is based on the simple observation that the equation has a solution if and only if the row reduced echelon form of $\begin{pmatrix} A & | & b \end{pmatrix}$ has no row of the form $\begin{pmatrix} 0 & \cdots & 0 & \blacktriangledown \end{pmatrix}$.

Proposition 11.6.1 *Let A be an $m \times n$ matrix and let b be an $m \times 1$ column vector. Then there exists a solution to 11.5 if and only if*

$$\text{rank} \begin{pmatrix} A & | & b \end{pmatrix} = \text{rank}(A). \quad (11.6)$$

Proof: Place $\begin{pmatrix} A & | & b \end{pmatrix}$ and A in row reduced echelon form, respectively B and C . If the above condition on rank is true, then both B and C have the same number of nonzero rows. In particular, you cannot have in B a row of the form

$$\begin{pmatrix} 0 & \cdots & 0 & \blacktriangledown \end{pmatrix}$$

where $\blacktriangledown \neq 0$. Therefore, there will exist a solution to the system 11.5.

Conversely, suppose there exists a solution. This means there cannot be such a row in B described above. Therefore, B and C must have the same number of zero rows and so they have the same number of nonzero rows. Therefore, the rank of the two matrices in 11.6 is the same.

Another way to see this is as follows. To say there is a solution to 11.5 is to say that b is in the span of the columns of A which is to say that the rank of A is the rank of $\begin{pmatrix} A & | & b \end{pmatrix}$ because to delete the vector b from the list of column vectors of $\begin{pmatrix} A & | & b \end{pmatrix}$ does not change the rank. ■

There is a very useful version of Proposition 11.6.1 known as the **Fredholm alternative**.

The following definition is used to state the Fredholm alternative.

Definition 11.6.2 *Let $S \subseteq \mathbb{R}^m$. Then $S^\perp \equiv \{z \in \mathbb{R}^m : z \cdot s = 0 \text{ for every } s \in S\}$. The funny exponent, \perp is called “perp”.*

Now note

$$N(A^T) \equiv \{z : A^T z = 0\} = \left\{ z : \sum_{k=1}^m z_k a_k = 0 \right\}$$

Here the a_k are the rows of A because they are the columns of A^T .

Lemma 11.6.3 *Let A be a real $m \times n$ matrix, let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Then*

$$(Ax \cdot y) = (x \cdot A^T y)$$

Proof: This follows right away from the definition of the dot product and matrix multiplication.

$$(A\mathbf{x} \cdot \mathbf{y}) = \sum_{k,l} A_{kl} x_l y_k = \sum_{k,l} (A^T)_{lk} x_l y_k = (\mathbf{x} \cdot A^T \mathbf{y}). \blacksquare$$

Now it is time to state the Fredholm alternative. The first version of this is the following theorem.

Theorem 11.6.4 *Let A be a real $m \times n$ matrix and let $\mathbf{b} \in \mathbb{R}^m$. There exists a solution \mathbf{x} to the equation $A\mathbf{x} = \mathbf{b}$ if and only if $\mathbf{b} \in (N(A^T))^\perp$.*

Proof: First suppose $\mathbf{b} \in (N(A^T))^\perp$. Then this says that if $A^T \mathbf{x} = \mathbf{0}$, it follows that

$$\mathbf{b} \cdot \mathbf{x} = \mathbf{x}^T \mathbf{b} = 0.$$

In other words, on taking the transpose, if

$$\mathbf{x}^T A = \mathbf{0}^T \text{ then } \mathbf{x}^T \mathbf{b} = 0.$$

Thus, if P is a product of elementary matrices such that PA is in row reduced echelon form, then if PA has a row of zeros, in the k^{th} position, obtained from the k^{th} row of P times A , then there is also a zero in the k^{th} position of $P\mathbf{b}$. This is because the k^{th} position in $P\mathbf{b}$ is just the k^{th} row of P times \mathbf{b} . Thus the row reduced echelon forms of A and $\left(A \mid \mathbf{b} \right)$ have the same number of zero rows. Thus $\text{rank} \left(A \mid \mathbf{b} \right) = \text{rank}(A)$. By Proposition 11.6.1, there exists a solution \mathbf{x} to the system $A\mathbf{x} = \mathbf{b}$. It remains to prove the converse.

Let $\mathbf{z} \in N(A^T)$ and suppose $A\mathbf{x} = \mathbf{b}$. I need to verify $\mathbf{b} \cdot \mathbf{z} = 0$. By Lemma 11.6.3,

$$\mathbf{b} \cdot \mathbf{z} = A\mathbf{x} \cdot \mathbf{z} = \mathbf{x} \cdot A^T \mathbf{z} = \mathbf{x} \cdot \mathbf{0} = 0 \blacksquare$$

This implies the following corollary which is also called the Fredholm alternative. The “alternative” becomes more clear in this corollary.

Corollary 11.6.5 *Let A be an $m \times n$ matrix. Then A maps \mathbb{R}^n onto \mathbb{R}^m if and only if the only solution to $A^T \mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$.*

Proof: If the only solution to $A^T \mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$, then $N(A^T) = \{\mathbf{0}\}$ and so $N(A^T)^\perp = \mathbb{R}^m$ because every $\mathbf{b} \in \mathbb{R}^m$ has the property that $\mathbf{b} \cdot \mathbf{0} = 0$. Therefore, $A\mathbf{x} = \mathbf{b}$ has a solution for any $\mathbf{b} \in \mathbb{R}^m$ because the \mathbf{b} for which there is a solution are those in $N(A^T)^\perp$ by Theorem 11.6.4. In other words, A maps \mathbb{R}^n onto \mathbb{R}^m .

Conversely if A is onto, then if $A^T \mathbf{x} = \mathbf{0}$, there exists \mathbf{y} such that $\mathbf{x} = A\mathbf{y}$ and then $A^T A\mathbf{y} = \mathbf{0}$ and so $|A\mathbf{y}|^2 = A\mathbf{y} \cdot A\mathbf{y} = A^T A\mathbf{y} \cdot \mathbf{y} = \mathbf{0} \cdot \mathbf{y} = 0$ and so $\mathbf{x} = A\mathbf{y} = \mathbf{0}$. \blacksquare

Here is an amusing example.

Example 11.6.6 *Let A be an $m \times n$ matrix in which $m > n$. Then A cannot map onto \mathbb{R}^m .*

The reason for this is that A^T is an $n \times m$ where $m > n$ and so in the augmented matrix

$$(A^T | \mathbf{0})$$

there must be some free variables. Thus there exists a nonzero vector \mathbf{x} such that $A^T \mathbf{x} = \mathbf{0}$. Hence A^T is not one to one and so A is not onto.

11.6.2 Least Squares

Suppose there is no solution to the system $Ax = b$. This happens often in applications where you want to find the best solution. In other words, you want to find x such that for all \hat{x} ,

$$|Ax - b| \leq |A\hat{x} - b|$$

It turns out that the solution to this problem is any solution x to

$$A^T Ax = A^T b$$

So this raises the question whether there is a solution to this last equation.

In order to present this material using notation which is common in more general situations, we begin to denote the dot product $x \cdot y$ as (x, y) . Thus, the property of the transpose mentioned above about how it interacts with the dot product is written as $(Ax, y) = (x, A^T y)$.

Theorem 11.6.7 *Let A be a real $m \times n$ matrix and let $b \in \mathbb{R}^m$. Then there exists a solution x to the system*

$$A^T Ax = A^T b$$

Proof: First note that $(A^T A)^T = A^T A$. Thus, by the Fredholm alternative, it suffices to verify that $A^T b$ is in $\left(N(A^T A)^T\right)^\perp$. So suppose $A^T Az = 0$. Does it follow that $(z, A^T b) = 0$? First note that $N(A^T A) = N(A)$. To see this note that since any matrix times the zero vector is zero, the left side is at least as large as the right. But if $A^T Ax = 0$, then $0 = (A^T Ax, x) = (Ax, Ax) = |Ax|^2$ so $Ax = 0$. Hence the two sets are the same. Thus $(z, A^T b) = (Az, b) = (0, b) = 0$. By Fredholm alternative, it follows there exists a solution to the above equation. ■

Next we verify that any solution to this equation is a solution to the least squares problem of finding x such that Ax is as close as possible to b .

Theorem 11.6.8 $|Ax - b| \leq |A\hat{x} - b|$ for all \hat{x} in \mathbb{R}^n if and only if $A^T Ax = A^T b$.

Proof: x is such that Ax is as close as possible to b if and only if $|A(x + tz) - b|^2$ is minimized when $t = 0$ for any choice of z . This equals

$$(Ax - b + tAz, Ax - b + tAz)$$

Now, expanding this yields

$$|Ax - b|^2 + 2t(Ax - b, Az) + t^2|Az|^2$$

If x solves the minimization problem, then taking a derivative and setting equal to 0 gives

$$0 = (Ax - b, Az) = (A^T(Ax - b), z) = (A^T Ax - A^T b, z)$$

for all z . In particular this holds for $z = A^T Ax - A^T b$. Hence $A^T Ax = A^T b$.

Conversely, if the equation holds, then $0 = (Ax - b, Az) = (A^T(Ax - b), z)$ and so for any z ,

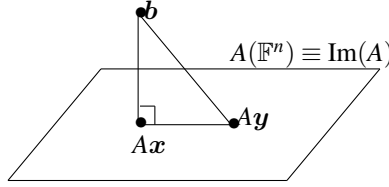
$$|A(x + tz) - b|^2 = |Ax - b|^2 + t^2|Az|^2$$

which shows that the minimization property holds since you could let $t = 1$ in the above. ■

Corollary 11.6.9 $|Ax - b| \leq |A\hat{x} - b|$ for all $\hat{x} \in \mathbb{R}^n$ if and only if $0 = (Ax - b, Az)$ for every $z \in \mathbb{R}^n$.

Proof: This is the content of the above theorem because $0 = (Ax - b, Az)$ for every $z \in \mathbb{R}^n$ if and only if $A^T Ax = A^T b$. ■

The corollary says that the vector $Ax - b$ is perpendicular to the subspace $\text{Im}(A)$ is the same as saying that Ax is as close as possible to b . Note that this orthogonality condition $0 = (Ax - b, Az)$ is equivalent to saying $0 = (Ax - b, Ax - Az) = (b - Ax, Ax - Az)$. Here is a picture which illustrates the conclusion of this important theorem.



Next consider the problem of projection onto a subspace. Letting V be a subspace of \mathbb{R}^n , and letting $b \in \mathbb{R}^n$, how do we find $x \in V$ such that $|b - x| \leq |b - y|$ for every $y \in V$?

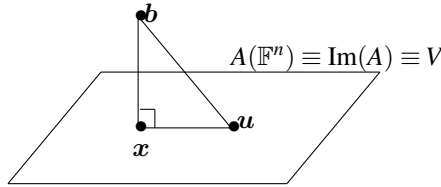
The subspace has a basis, $\{v_1, \dots, v_m\}$, $m \leq n$. Let

$$A = \begin{pmatrix} v_1 & \cdots & v_m & 0 & \cdots & 0 \end{pmatrix}$$

Thus V is the column space of A , the span of the columns of A which is also $\text{Im}(A)$. Thus the question is to find Ay which is closer to b than any Az . Isn't this just what was solved above? Then the closest point to b in V will be $x \equiv Ay$. From the above explanation, y must satisfy

$$A^T Ay = A^T b \text{ so } \begin{pmatrix} b - Ay, Az \end{pmatrix} = 0$$

for all $z \in \mathbb{R}^n$. In other words, you need x to be the point in V which satisfies $(b - x, u) = 0$ for all u in V . Note that since x is in V , a generic point of V is of the form $u - x$. Thus it makes no difference whether we write $(b - x, u) = 0$ for all u in V or $(b - x, u - x) = 0$ for all u in V . The following picture illustrates what was just shown.



Theorem 11.6.10 Let V be a finite dimensional subspace of \mathbb{R}^n and let $b \in \mathbb{R}^n$. Then there exists a unique point of V which is closest to b out of all points of V . This point x is characterized by the equation

$$(b - x, z) = 0 \quad (11.7)$$

for all $z \in V$.

Proof: The existence of this point follows from Theorem 11.6.8. However, to emphasize the uniqueness, suppose x, \hat{x} both are closest points. Then from the characterization of these points in 11.7,

$$(b - x, x - \hat{x}) = 0, (b - \hat{x}, x - \hat{x}) = 0$$

Then subtracting these yields $(x - \hat{x}, x - \hat{x}) = 0$ thus showing uniqueness. ■

11.6.3 Regression lines

In experimental work, one often wants to determine relationships which are not exactly determined. For example, you might want to find whether a vaccine is effective. In terms of linear equations, this amounts to there being no solution to a system of equations but still needing to answer a question about the data.

Example 11.6.11 *The least squares regression line is the line $y = mx + b$ which approximates data points (x_i, y_i) which typically come from some sort of experiment. It is desired to choose m, b in such a way that the sum of the squares of the errors between the value predicted by the line and the observed values is as small as possible. In other words, you want to minimize $\sum_i (y_i - (mx_i + b))^2$. Ideally, the sum would be zero and this would correspond to the data points being on a straight line. This will never occur in any realistic situation in which the data points come from experiments.*

Suppose you are given points in xy plane

$$\{(x_i, y_i)\}_{i=1}^n$$

and you would like to find constants m and b such that the line $y = mx + b$ goes through all these points. Of course this will be impossible in general. Therefore, try to find m, b to get as close as possible. The desired system is

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} \equiv A \begin{pmatrix} m \\ b \end{pmatrix}$$

which is of the form $\mathbf{y} = A\mathbf{x}$ and it is desired to choose m and b to make

$$\left| A \begin{pmatrix} m \\ b \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right|^2$$

as small as possible. According to Theorem 11.6.8, the best values for m and b occur as the solution to

$$A^T A \begin{pmatrix} m \\ b \end{pmatrix} = A^T \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}.$$

Thus, after computing $A^T A, A^T \mathbf{y}$

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}$$

Solving this system of equations for m and b ,

$$m = \frac{-(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i) + (\sum_{i=1}^n x_i y_i)n}{(\sum_{i=1}^n x_i^2)n - (\sum_{i=1}^n x_i)^2}$$

and

$$b = \frac{-(\sum_{i=1}^n x_i)\sum_{i=1}^n x_i y_i + (\sum_{i=1}^n y_i)\sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i^2)n - (\sum_{i=1}^n x_i)^2}.$$

One could clearly do a least squares fit for curves of the form $y = ax^2 + bx + c$ in the same way. In this case you want to solve as well as possible for a, b , and c the system

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and one would use the same technique as above. Many other similar problems are important, including many in higher dimensions and they are all solved the same way.

Example 11.6.12 Find the least squares regression line for the data

$$(0, 1), (2, 3), (2, 4), (3, 4), (3, 5), (4, 6), (4, 5)$$

You would ideally want to solve the following system of equations

$$\begin{pmatrix} 0 & 1 \\ 2 & 1 \\ 2 & 1 \\ 3 & 1 \\ 3 & 1 \\ 4 & 1 \\ 4 & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 4 \\ 4 \\ 5 \\ 6 \\ 5 \end{pmatrix}$$

Of course there is no solution so you look for a least squares solution. You have $A^T A$ equals

$$\begin{pmatrix} 58 & 18 \\ 18 & 7 \end{pmatrix}$$

and $A^T \mathbf{b}$ is

$$\begin{pmatrix} 85 \\ 28 \end{pmatrix}$$

and so you need to solve

$$\begin{pmatrix} 58 & 18 \\ 18 & 7 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 85 \\ 28 \end{pmatrix}$$

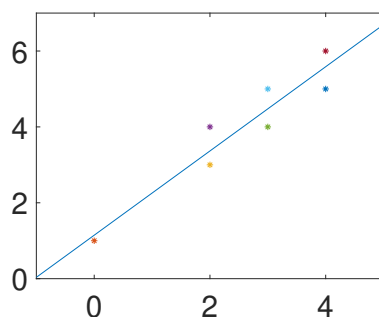
The solution is:

$$\begin{pmatrix} \frac{91}{82} \\ \frac{47}{41} \end{pmatrix} = \begin{pmatrix} 1.1098 \\ 1.1463 \end{pmatrix}$$

Thus the least squares line is

$$y = 1.1098x + 1.1463$$

If you graph these data points and the line, you will see how the line tries to do the impossible by picking a route through the data points which minimizes the error which results.



11.6.4 Identifying the Closest Point

For V a finite dimensional subspace as in the above theorem, how can we identify the closest point to \mathbf{b} in Theorem 11.6.10? Suppose a basis for V is $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$. Then we would have unique scalars c_k (The c_k are unique because the \mathbf{y} is unique.)

$$\mathbf{y} = \sum_{k=1}^m c_k \mathbf{v}_k$$

Therefore, taking inner products, it follows that for each j ,

$$(\mathbf{y}, \mathbf{v}_j) = \sum_{k=1}^m c_k (\mathbf{v}_k, \mathbf{v}_j) \quad (11.8)$$

Wouldn't it be nice if $(\mathbf{v}_k, \mathbf{v}_j) = \delta_{kj}$? Recall that δ_{ij} equals 1 if $i = j$ and 0 if $i \neq j$. If this happens, the vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ are called an orthonormal set of vectors. In this case, you would have $c_j = (\mathbf{y}, \mathbf{v}_j)$ because on the right, the sum would reduce to c_j . In fact, if you have a basis of vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, there always exists another basis $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ which is orthonormal. Furthermore, there is an algorithm for finding this improved basis. It is called the Gram Schmidt process.

Lemma 11.6.13 *Let $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be a linearly independent subset of \mathbb{R}^p , $p \geq n$. Then there exists orthonormal vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ which have the property that for each $k \leq n$, $\text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$.*

Proof: Let $\mathbf{u}_1 \equiv \mathbf{v}_1/|\mathbf{v}_1|$. Thus for $k = 1$, $\text{span}(\mathbf{u}_1) = \text{span}(\mathbf{v}_1)$ 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 $\text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$. Then define

$$\mathbf{u}_{k+1} \equiv \frac{\mathbf{v}_{k+1} - \sum_{j=1}^k (\mathbf{v}_{k+1}, \mathbf{u}_j) \mathbf{u}_j}{\left| \mathbf{v}_{k+1} - \sum_{j=1}^k (\mathbf{v}_{k+1}, \mathbf{u}_j) \mathbf{u}_j \right|}, \quad (11.9)$$

where the denominator is not equal to zero because the \mathbf{v}_j form a basis, and so

$$\mathbf{v}_{k+1} \notin \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$$

Thus by induction,

$$\mathbf{u}_{k+1} \in \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{v}_{k+1}) = \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k, \mathbf{v}_{k+1}).$$

Also, $\mathbf{v}_{k+1} \in \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{u}_{k+1})$ which is seen easily by solving 11.9 for \mathbf{v}_{k+1} , and it follows

$$\text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k, \mathbf{v}_{k+1}) = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{u}_{k+1}).$$

If $l \leq k$,

$$\begin{aligned} (\mathbf{u}_{k+1}, \mathbf{u}_l) &= C \left((\mathbf{v}_{k+1}, \mathbf{u}_l) - \sum_{j=1}^k (\mathbf{v}_{k+1}, \mathbf{u}_j) (\mathbf{u}_j, \mathbf{u}_l) \right) = \\ C \left((\mathbf{v}_{k+1}, \mathbf{u}_l) - \sum_{j=1}^k (\mathbf{v}_{k+1}, \mathbf{u}_j) \delta_{lj} \right) &= C((\mathbf{v}_{k+1}, \mathbf{u}_l) - (\mathbf{v}_{k+1}, \mathbf{u}_l)) = 0. \end{aligned}$$

The vectors, $\{\mathbf{u}_j\}_{j=1}^n$, generated in this way are therefore orthonormal because each vector has unit length. ■

Corollary 11.6.14 *If you have a basis for \mathbb{R}^p ,*

$$\{\mathbf{u}_1, \dots, \mathbf{u}_m, \mathbf{u}_{m+1}, \dots, \mathbf{u}_p\}$$

and $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ is orthonormal, then when the Gram Schmidt process is used on this basis, it returns $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$. Thus it is always possible to extend an orthonormal set of vectors to an orthonormal basis.

Proof: This follows right away from the algorithm. ■

Did we ever use the fact that all of this is taking place in \mathbb{R}^p ? No, this was never used at all! In fact everything in the Gram Schmidt process holds if V is a subspace of an arbitrary inner product space. You just need something which is a vector space which has an inner product to have it all work out exactly the same. A vector space is something in which you can add the “vectors” and multiply them by scalars in the usual way which we do for vectors in \mathbb{R}^n .

Now return to the stated problem which was to compute the closest point in V . This is the content of the next theorem.

Theorem 11.6.15 *Let V be an m dimensional subspace of \mathbb{R}^p having orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$. Let $\mathbf{b} \in \mathbb{R}^p$ and let \mathbf{y} be the point of V closest to \mathbf{b} . Then*

$$\mathbf{y} = \sum_{k=1}^m (\mathbf{b}, \mathbf{u}_k) \mathbf{u}_k \quad (11.10)$$

Proof: We only need to show that this satisfies the orthogonality condition 11.7. But this is fairly obvious because, from properties of the inner product and \mathbf{y} given above,

$$(\mathbf{y}, \mathbf{u}_j) = \left(\sum_{k=1}^m (\mathbf{b}, \mathbf{u}_k) \mathbf{u}_k, \mathbf{u}_j \right) = \sum_{k=1}^m (\mathbf{b}, \mathbf{u}_k) (\mathbf{u}_k, \mathbf{u}_j) = (\mathbf{b}, \mathbf{u}_j)$$

Thus $(\mathbf{b} - \mathbf{y}, \mathbf{u}_j) = 0$. Since this holds for every basis vector, it holds for every $\mathbf{z} \in V$ also.

$$(\mathbf{z}, \mathbf{b} - \mathbf{y}) = \left(\sum_{j=1}^m (\mathbf{z}, \mathbf{u}_j) \mathbf{u}_j, \mathbf{b} - \mathbf{y} \right) = \sum_{j=1}^m (\mathbf{z}, \mathbf{u}_j) (\mathbf{u}_j, \mathbf{b} - \mathbf{y}) = 0$$

Therefore, the orthogonality condition holds for \mathbf{y} given by the above formula and so \mathbf{y} equals the above sum in 11.10. ■

The sum in 11.10 is the Fourier series approximation to \mathbf{b} . The scalars $(\mathbf{b}, \mathbf{u}_k)$ are the Fourier coefficients. Note that all this works any time you have a norm which comes from an inner product, something which satisfies the same axioms as the dot product. That is, $|\mathbf{x}| = (\mathbf{x}, \mathbf{x})$ where (\cdot, \cdot) satisfies the inner product axioms:

1. $(f, g) = \overline{(g, f)}$
2. $(af + bg, h) = a(f, h) + b(g, h)$
3. $(f, f) \geq 0$ and equals 0 only if $f = 0$

The conjugate is placed on the (g, f) to include the case of a complex inner product. Just ignore it in the case where the scalars are real numbers.

Now we generalize these ideas more.

Theorem 11.6.16 *Let V be a finite dimensional subspace of an inner product space X , something with an inner product. (X is a nonempty set which satisfies the vector space axioms. In addition it has an inner product satisfying the inner product axioms.) If $\mathbf{b} \in X$, there exists a unique $\mathbf{y} \in V$ such that $|\mathbf{b} - \mathbf{y}| \leq |\mathbf{b} - \mathbf{z}|$ for all $\mathbf{z} \in V$. This point is characterized by $(\mathbf{b} - \mathbf{y}, \mathbf{z}) = 0$ for all $\mathbf{z} \in V$.*

Proof: Letting $t \in \mathbb{R}$,

$$|\mathbf{b} - (\mathbf{y} + t\mathbf{z})|^2 = |\mathbf{b} - \mathbf{y}|^2 - 2t(\mathbf{b} - \mathbf{y}, \mathbf{z}) + t^2|\mathbf{z}|^2$$

If \mathbf{y} is closest to \mathbf{b} then taking the derivative and setting $t = 0$, we must have $(\mathbf{b} - \mathbf{y}, \mathbf{z}) = 0$. Conversely, if this equals zero, let $t = 1$ and you have

$$|\mathbf{b} - (\mathbf{y} + \mathbf{z})|^2 = |\mathbf{b} - \mathbf{y}|^2 + |\mathbf{z}|^2$$

and so \mathbf{y} solves the minimization property. It only remains to show the existence of such \mathbf{y} satisfying $(\mathbf{b} - \mathbf{y}, \mathbf{z}) = 0$. However, using the Gram Schmidt process, there is an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ whose span is V . Then all that remains is to verify that $\sum_{i=1}^n (\mathbf{b}, \mathbf{u}_i) \mathbf{u}_i$ satisfies the orthogonality condition. Indeed,

$$\left(\mathbf{b} - \sum_{i=1}^n (\mathbf{b}, \mathbf{u}_i) \mathbf{u}_i, \mathbf{u}_j \right) = (\mathbf{b}, \mathbf{u}_j) - \sum_{i=1}^n (\mathbf{b}, \mathbf{u}_i) \delta_{ij} = 0$$

Since the \mathbf{u}_i are a basis, it follows that $(\mathbf{b} - \mathbf{y}, \mathbf{z}) = 0$ for all $\mathbf{z} \in V$. ■

Note that any time the norm comes from an inner product, something which satisfies the properties of the dot product, all of this holds. You don't need to be considering vectors in \mathbb{R}^n . It was only the axioms of the inner product which were used.

11.6.5 Using MATLAB

You may notice that the Gram Schmidt process is pretty tedious but routine. Therefore, it is not surprising that it can be automated using MATLAB. One way to do it would be to use what is known as the QR factorization. Given a real $m \times n$ matrix A which has linearly independent columns, you can always write it in the form $A = QR$ where Q is an orthogonal matrix and R is an upper triangular matrix in the sense that all entries are 0 below the main diagonal. Actually, the computer algebra system doesn't use the Gram Schmidt process. It uses something called Householder reflections, but one obtains the same essentials although sometimes a different set of vectors, the columns of Q being an orthonormal set. The span of these columns will coincide with the span of the columns of A and in fact, the span of the first k columns of A will be the span of the first k columns of Q just as in the Gram Schmidt process. Here is the syntax:

```
>>A=[1,2,3;4,2,1;2,6,7;1,-4,2];[Q,R]=qr(A)
```

Then press enter and you get the following.

Q=	R=
-0.2132 0.1756 -0.2593 0.9255	-4.6904 -3.8376 -4.9036
-0.8528 -0.1892 0.4862 -0.0244	0 6.7285 3.4453
-0.4264 0.6485 -0.5139 -0.3653	0 0 -5.2043
-0.2132 -0.7161 -0.6575 -0.0974	0 0 0

If you want to see something horrible, replace `qr(A)` with `qr(sym(A))`. This way it gives the exact values. You can check your work by `>>Q*Q'` and press enter. The Q' means the conjugate transpose in MATLAB. Since everything is real here, this is just the transpose.

There is so much more that could be discussed about the QR factorization, but this will suffice here.

As to plotting data with a curve as in the least squares example, use the following syntax.

```
x=-1:1:5;
y=1.11*x+1.146;
plot(x,y,0,1,'*',2,3,'*',2,4,'*',3,4,'*',3,5,'*',4,6,'*',4,5,'*')
```

In MATLAB, you press shift enter to get to a new line and you press enter to get it to do something.

11.7 The Singular Value Decomposition*

In this section, A will be an $m \times n$ matrix. To begin with, here is a simple lemma.

Lemma 11.7.1 *Let A be an $m \times n$ matrix. Then A^*A is Hermitian and all its eigenvalues are nonnegative.*

Proof: It is obvious that A^*A is Hermitian because $(A^*A)^* = A^*(A^*)^* = A^*A$. Suppose $A^*Ax = \lambda x$. Then

$$\lambda |x|^2 = (\lambda x, x) = (A^*Ax, x) = (Ax, Ax) \geq 0. \blacksquare$$

Definition 11.7.2 *Let A be an $m \times n$ matrix. The singular values of A are the square roots of the positive eigenvalues of A^*A .*

With this definition and lemma here is the main theorem on the singular value decomposition.

Theorem 11.7.3 *Let A be an $m \times n$ matrix. Then there exist unitary matrices, U and V of the appropriate size such that*

$$U^*AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

where σ is of the form

$$\sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{pmatrix}$$

for the σ_i the singular values of A .

Proof: By the above lemma and Theorem 11.4.7 there exists an orthonormal basis, $\{v_i\}_{i=1}^n$ such that $A^*Av_i = \sigma_i^2v_i$ where $\sigma_i^2 > 0$ for $i = 1, \dots, k$, ($\sigma_i > 0$) and equals zero if $i > k$. Thus for $i > k$, $Av_i = 0$ because

$$(Av_i, Av_i) = (A^*Av_i, v_i) = (0, v_i) = 0.$$

For $i = 1, \dots, k$, define $u_i \in \mathbb{F}^m$ by

$$u_i \equiv \sigma_i^{-1}Av_i.$$

Thus $Av_i = \sigma_i u_i$. Now

$$\begin{aligned} (u_i, u_j) &= (\sigma_i^{-1}Av_i, \sigma_j^{-1}Av_j) = (\sigma_i^{-1}v_i, \sigma_j^{-1}A^*Av_j) \\ &= (\sigma_i^{-1}v_i, \sigma_j^{-1}\sigma_j^2v_j) = \frac{\sigma_j}{\sigma_i}(v_i, v_j) = \delta_{ij}. \end{aligned}$$

Thus $\{u_i\}_{i=1}^k$ is an orthonormal set of vectors in \mathbb{F}^m . Also,

$$AA^*u_i = AA^*\sigma_i^{-1}Av_i = \sigma_i^{-1}AA^*Av_i = \sigma_i^{-1}A\sigma_i^2v_i = \sigma_i^2u_i.$$

Now extend $\{u_i\}_{i=1}^k$ to an orthonormal basis for all of \mathbb{F}^m , $\{u_i\}_{i=1}^m$ and let

$$U \equiv (u_1 \cdots u_m)$$

while $V \equiv (v_1 \cdots v_n)$. Thus U is the matrix which has the u_i as columns and V is defined as the matrix which has the v_i as columns. Then

$$U^*AV = \begin{pmatrix} u_1^* \\ \vdots \\ u_k^* \\ \vdots \\ u_m^* \end{pmatrix} A(v_1 \cdots v_n) = \begin{pmatrix} u_1^* \\ \vdots \\ u_k^* \\ \vdots \\ u_m^* \end{pmatrix} (\sigma_1 u_1 \cdots \sigma_k u_k, 0 \cdots 0) = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

where σ is given in the statement of the theorem. ■

The singular value decomposition has as an immediate corollary the following interesting result.

Corollary 11.7.4 *Let A be an $m \times n$ matrix. Then the rank of A and A^* equals the number of singular values.*

Proof: Since V and U are unitary, it follows that

$$\begin{aligned} \text{rank}(A) &= \text{rank}(U^*AV) = \text{rank}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \\ &= \text{number of singular values.} \end{aligned}$$

Also since U, V are unitary,

$$\begin{aligned} \text{rank}(A^*) &= \text{rank}(V^*A^*U) = \text{rank}((U^*AV)^*) \\ &= \text{rank}\left(\left(\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\right)^*\right) = \text{number of singular values.} \blacksquare \end{aligned}$$

This is based on the simple observation that for A an $m \times n$ matrix, the dimension of $\text{Im}(A)$ is the same as the dimension of $\text{Im}(UAV)$ if U, V are invertible matrices of the right size. Indeed, $\text{Im}(UAV) = \text{Im}(UA)$ because V being invertible maps \mathbb{F}^n onto \mathbb{F}^n . The dimension of $\text{Im}(UA)$ and the dimension of $\text{Im}(A)$ must be the same because U is one to one. Thus if a basis for $\text{Im}(A)$ is $\{\mathbf{a}_1, \dots, \mathbf{a}_k\}$, columns of A , then a basis for UA will be $\{U\mathbf{a}_1, \dots, U\mathbf{a}_k\}$.

11.8 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 that $T\mathbf{v} = \mathbf{v}$ if $\mathbf{v} \in V \equiv \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_r)$. Now show that $T(T(\mathbf{x})) = T(\mathbf{x})$ and that $|T\mathbf{x} - T\mathbf{y}| \leq |\mathbf{x} - \mathbf{y}|$.
3. Find the minimum polynomials for the following matrices and use to obtain the eigenvalues of the matrix. The set of all eigenvectors associated with an eigenvalue λ is called the eigenspace. Determine the eigenspaces for each of these matrices.

(a) $\begin{pmatrix} 9 & 4 \\ -20 & -9 \end{pmatrix}$

(b) $\begin{pmatrix} -3 & -2 \\ 10 & 6 \end{pmatrix}$

(c) $\begin{pmatrix} 5 & -2 & 2 \\ 2 & 0 & 1 \\ -4 & 2 & -1 \end{pmatrix}$

$$(d) \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 2 \end{pmatrix}$$

$$(e) \begin{pmatrix} 2 & 0 & 1 \\ 1 & 1 & 1 \\ -1 & 0 & 0 \end{pmatrix}$$

$$(f) \begin{pmatrix} 6 & -2 & 3 \\ 3 & 0 & 2 \\ -5 & 2 & -2 \end{pmatrix}$$

4. Suppose you have $p(\lambda)$ is the minimum polynomial for a square $n \times n$ matrix A . Show that this matrix is invertible if and only if the constant term of the minimum polynomial is non zero. In this case, give a formula for A^{-1} in terms of powers of A . Say

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

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.

5. Find least squares solutions to the following systems of equations.

$$(a) \begin{pmatrix} 1 & 2 \\ -1 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$(b) \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$(c) \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$$

6. 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.

$$(a) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \quad (b) \begin{pmatrix} 1 & 2 & -3 \\ 2 & 1 & 4 \\ -3 & 4 & 7 \end{pmatrix} \quad (c) \begin{pmatrix} 0 & -2 & -3 \\ 2 & 0 & -4 \\ 3 & 4 & 0 \end{pmatrix}$$

7. 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.
8. 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.

9. 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 unitary matrices. Do it here for the special case that the matrix is orthogonal. It will be simpler.
10. 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.

11. 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 10.

12. 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.
13. If A, B are symmetric matrices, does it follow that AB is also symmetric?
14. Suppose A, B are symmetric and $AB = BA$. Does it follow that AB is symmetric?
15. 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) $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix}$

16. 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.

17. 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.

18. 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.

19. 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.

20. 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

$$\lambda \mathbf{x}^T \bar{\mathbf{x}} = (A\mathbf{x})^T \bar{\mathbf{x}} = \mathbf{x}^T A\bar{\mathbf{x}} = \mathbf{x}^T \overline{A\mathbf{x}} = \mathbf{x}^T \bar{\lambda} \bar{\mathbf{x}} = \bar{\lambda} \mathbf{x}^T \bar{\mathbf{x}}$$

and so $\lambda = \bar{\lambda}$. 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.

$$\lambda (\mathbf{x} \cdot \mathbf{y}) = \lambda \mathbf{x} \cdot \mathbf{y} = A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A\mathbf{y} = \mathbf{x} \cdot \mu \mathbf{y} = \mu (\mathbf{x} \cdot \mathbf{y}) = \mu (\mathbf{x} \cdot \mathbf{y})$$

and so

$$(\lambda - \mu) (\mathbf{x} \cdot \mathbf{y}) = 0.$$

Since $\lambda \neq \mu$, it follows $\mathbf{x} \cdot \mathbf{y} = 0$.

21. Suppose U is an orthogonal $n \times n$ matrix. Explain why $\text{rank}(U) = n$.
22. Show that if A is an Hermitian 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 $(\mathbf{x}, \mathbf{y}) = 0$. Also all eigenvalues are real. Supply reasons for each step in the following argument. First

$$\lambda (\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, \mathbf{x}) = (\mathbf{x}, A\mathbf{x}) = (\mathbf{x}, \lambda \mathbf{x}) = \bar{\lambda} (\mathbf{x}, \mathbf{x})$$

and so $\lambda = \bar{\lambda}$. This shows that all eigenvalues are real. Now let $\mathbf{x}, \mathbf{y}, \mu$ and λ be given as above.

$$\lambda (\mathbf{x}, \mathbf{y}) = (\lambda \mathbf{x}, \mathbf{y}) = (A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A\mathbf{y}) = (\mathbf{x}, \mu \mathbf{y}) = \bar{\mu} (\mathbf{x}, \mathbf{y}) = \mu (\mathbf{x}, \mathbf{y})$$

and so $(\lambda - \mu) (\mathbf{x}, \mathbf{y}) = 0$. Since $\lambda \neq \mu$, it follows $(\mathbf{x}, \mathbf{y}) = 0$.

23. Show that the eigenvalues and eigenvectors of a real matrix occur in conjugate pairs.
24. 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.

25. 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.
26. 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) $\left\{\left(\frac{1}{3}, \frac{2}{3}, \frac{2}{3}\right), \left(\frac{-2}{3}, \frac{-1}{3}, \frac{2}{3}\right), \left(\frac{2}{3}, \frac{-2}{3}, \frac{1}{3}\right)\right\}$
27. Show that if $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ is an orthonormal set of vectors in \mathbb{F}^n , then it is a basis.
Hint: It was shown earlier that this is a linearly independent set.
28. Fill in the missing entries to make the matrix orthogonal.

$$\begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & - & - \\ - & \frac{\sqrt{6}}{3} & - \end{pmatrix}.$$

29. Fill in the missing entries to make the matrix orthogonal.

$$\begin{pmatrix} \frac{2}{3} & \frac{\sqrt{2}}{2} & \frac{1}{6}\sqrt{2} \\ \frac{2}{3} & - & - \\ - & 0 & - \end{pmatrix}$$

30. Fill in the missing entries to make the matrix orthogonal.

$$\begin{pmatrix} \frac{1}{3} & -\frac{2}{\sqrt{5}} & - \\ \frac{2}{3} & 0 & - \\ - & - & \frac{4}{15}\sqrt{5} \end{pmatrix}$$

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 $U^T A U = D$.

$$A = \begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

Hint: One eigenvalue is -2.

32. 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.

33. 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.

34. 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 $U^T A U = D$.

$$A = \begin{pmatrix} -\frac{5}{3} & \frac{1}{15}\sqrt{6}\sqrt{5} & \frac{8}{15}\sqrt{5} \\ \frac{1}{15}\sqrt{6}\sqrt{5} & -\frac{14}{5} & -\frac{1}{15}\sqrt{6} \\ \frac{8}{15}\sqrt{5} & -\frac{1}{15}\sqrt{6} & \frac{7}{15} \end{pmatrix}$$

Hint: The eigenvalues are $-3, -2, 1$.

35. 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 $U^T A U = D$.

$$A = \begin{pmatrix} 3 & 0 & 0 \\ 0 & \frac{3}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{3}{2} \end{pmatrix}.$$

36. 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 $U^T A U = D$.

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 5 & 1 \\ 0 & 1 & 5 \end{pmatrix}.$$

37. 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 $U^T A U = D$.

$$A = \begin{pmatrix} \frac{4}{3} & \frac{1}{3}\sqrt{3}\sqrt{2} & \frac{1}{3}\sqrt{2} \\ \frac{1}{3}\sqrt{3}\sqrt{2} & 1 & -\frac{1}{3}\sqrt{3} \\ \frac{1}{3}\sqrt{2} & -\frac{1}{3}\sqrt{3} & \frac{5}{3} \end{pmatrix}$$

Hint: The eigenvalues are $0, 2, 2$ where 2 is listed twice because it is a root of multiplicity 2.

38. 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 $U^T A U = D$.

$$A = \begin{pmatrix} 1 & \frac{1}{6}\sqrt{3}\sqrt{2} & \frac{1}{6}\sqrt{3}\sqrt{6} \\ \frac{1}{6}\sqrt{3}\sqrt{2} & \frac{3}{2} & \frac{1}{12}\sqrt{2}\sqrt{6} \\ \frac{1}{6}\sqrt{3}\sqrt{6} & \frac{1}{12}\sqrt{2}\sqrt{6} & \frac{1}{2} \end{pmatrix}$$

Hint: The eigenvalues are 2, 1, 0.

39. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{6}\sqrt{3}\sqrt{2} & -\frac{7}{18}\sqrt{3}\sqrt{6} \\ \frac{1}{6}\sqrt{3}\sqrt{2} & \frac{3}{2} & -\frac{1}{12}\sqrt{2}\sqrt{6} \\ -\frac{7}{18}\sqrt{3}\sqrt{6} & -\frac{1}{12}\sqrt{2}\sqrt{6} & -\frac{5}{6} \end{pmatrix}$$

Hint: The eigenvalues are 1, 2, -2 .

40. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix

$$\begin{pmatrix} -\frac{1}{2} & -\frac{1}{5}\sqrt{6}\sqrt{5} & \frac{1}{10}\sqrt{5} \\ -\frac{1}{5}\sqrt{6}\sqrt{5} & \frac{7}{5} & -\frac{1}{5}\sqrt{6} \\ \frac{1}{10}\sqrt{5} & -\frac{1}{5}\sqrt{6} & -\frac{9}{10} \end{pmatrix}$$

Hint: The eigenvalues are $-1, 2, -1$ where -1 is listed twice because it has multiplicity 2 as a zero of the characteristic equation.

41. 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.
42. You are doing experiments and have obtained the ordered pairs,

$$(0, 1), (1, 2), (2, 3.5), (3, 4)$$

Find m and b such that $y = mx + b$ approximates these four points as well as possible. Now do the same thing for $y = ax^2 + bx + c$, finding a, b , and c to give the best approximation.

43. Suppose you have several ordered triples, (x_i, y_i, z_i) . Describe how to find a polynomial,

$$z = a + bx + cy + dxy + ex^2 + fy^2$$

for example giving the best fit to the given ordered triples. Is there any reason you have to use a polynomial? Would similar approaches work for other combinations of functions just as well?

44. 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) $(3, 0, -4), (5, 0, 10), (-7, 1, 1)$

45. Using the Gram Schmidt process or the QR factorization, find an orthonormal basis for the span of the vectors, $(1, 2, 1), (2, -1, 3)$, and $(1, 0, 0)$.
46. Using the Gram Schmidt process or the QR factorization, find an orthonormal basis for the span of the vectors, $(1, 2, 1, 0), (2, -1, 3, 1)$, and $(1, 0, 0, 1)$.
47. 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.
48. 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.
49. Let A, B be a $m \times n$ matrices. Define an inner product on the set of $m \times n$ matrices by

$$(A, B)_F \equiv \text{trace}(AB^*).$$

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.

50. Let A be an $m \times n$ matrix. Show $\|A\|_F^2 \equiv (A, A)_F = \sum_j \sigma_j^2$ where the σ_j are the singular values of A .
51. 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.
52. Using Problem 51 and Schur's theorem, show that the trace of an $n \times n$ matrix equals the sum of the eigenvalues.
53. If A is a general $n \times n$ matrix having possibly repeated eigenvalues, show there is a sequence $\{A_k\}$ of $n \times n$ matrices having distinct eigenvalues which has the property that the ij^{th} entry of A_k converges to the ij^{th} entry of A for all ij . **Hint:** Use Schur's theorem.

Chapter 12

Vector Valued Functions

12.1 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 12.1.1 *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 12.1.2 *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 12.1.3 *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 12.1.4 *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*

$$(a\mathbf{f} + b\mathbf{g})(x) = a\mathbf{f}(x) + b\mathbf{g}(x).$$

$\mathbf{f} \cdot \mathbf{g}$ 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

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

which is defined as

$$\mathbf{g} \circ \mathbf{f}(\mathbf{x}) \equiv \mathbf{g}(\mathbf{f}(\mathbf{x})).$$

This is called the composition of the two functions.

You should note that $\mathbf{f}(\mathbf{x})$ is not a function. It is the value of the function at the point \mathbf{x} . The name of the function is \mathbf{f} . Nevertheless, people often write $\mathbf{f}(\mathbf{x})$ to denote a function and it does not cause too many problems in beginning courses. When this is done, the variable, \mathbf{x} should be considered as a generic variable free to be anything in $D(\mathbf{f})$. I will use this slightly sloppy abuse of notation whenever convenient.

Example 12.1.5 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} .

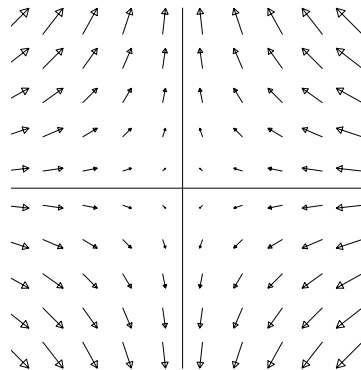
Example 12.1.6 Suppose $\mathbf{f}(t) = (2t, 1+t^2)$ and $\mathbf{g} : \mathbb{R}^2 \rightarrow \mathbb{R}$ is given by $\mathbf{g}(x, y) \equiv x + y$. Then $\mathbf{g} \circ \mathbf{f} : \mathbb{R} \rightarrow \mathbb{R}$ and

$$\mathbf{g} \circ \mathbf{f}(t) = \mathbf{g}(\mathbf{f}(t)) = \mathbf{g}(2t, 1+t^2) = 1 + 2t + t^2.$$

12.2 Vector Fields

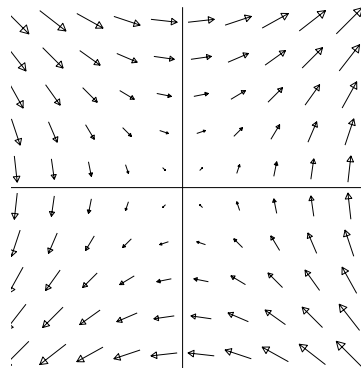
Some people find it useful to try and draw pictures to illustrate a vector valued function. This can be a very useful idea in the case where the function takes points in $D \subseteq \mathbb{R}^2$ and delivers a vector in \mathbb{R}^2 . For many points $(x, y) \in D$, you draw an arrow of the appropriate 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 12.2.1 Draw a picture of the vector field $(-x, y)$ which gives the velocity of a fluid flowing in two dimensions.



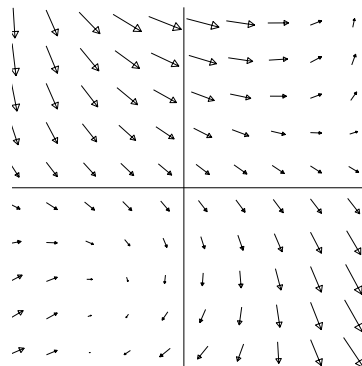
You can see how the arrows indicate the motion of this fluid.

Example 12.2.2 Draw a picture of the vector field (y, x) for the velocity of a fluid flowing in two dimensions.



Here is another such example.

Example 12.2.3 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.



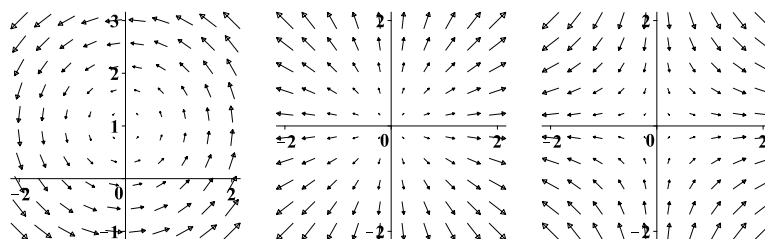
These pictures were drawn by maple. Note how they reveal both the direction and the magnitude of the vectors. However, if you try to draw these by hand, you will mainly waste time.

12.3 Exercises

1. 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.



2. 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) $\mathbf{f} \times \mathbf{g}$
 - (b) $\mathbf{g} \times \mathbf{f}$
 - (c) $\mathbf{f} \cdot \mathbf{g}$
 - (d) $\mathbf{f} \times \mathbf{g} \cdot \mathbf{h}$
 - (e) $\mathbf{f} \times (\mathbf{g} \times \mathbf{h})$
 - (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) $\mathbf{f} \times \mathbf{g} \times \mathbf{h}$
 - (b) $(k \times \mathbf{g}) \times \mathbf{h}$
 - (c) $(\mathbf{f} \cdot \mathbf{g}) \times \mathbf{h}$
 - (d) $(\mathbf{f} \times \mathbf{g}) \cdot \mathbf{h}$

(e) $lg \cdot k$ (f) $f \times (g + h)$

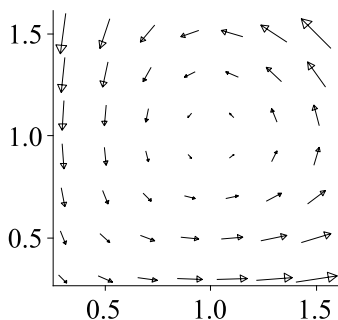
8. The Lotka Volterra system of differential equations, proposed in 1925 and 1926 by Lotka and Volterra respectively, is intended to model the interaction of predators and prey. An example of this situation is that of wolves and moose living on Isle Royal in the middle of Lake Superior. In these equations 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 $(1, 1)$?

12.4 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.

Definition 12.4.1 A function $f : D(f) \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ 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

$$|f(x) - f(y)| < \varepsilon.$$

f is continuous if it is continuous at every point of $D(f)$.

Note the total similarity to the scalar valued case.

12.4.1 Sufficient Conditions For Continuity

The next theorem is a fundamental result which allows less worry about the $\varepsilon \delta$ definition of continuity.

Theorem 12.4.2 *The following assertions are valid.*

1. *The function $a\mathbf{f} + b\mathbf{g}$ is continuous at \mathbf{x} whenever \mathbf{f}, \mathbf{g} are continuous at $\mathbf{x} \in D(\mathbf{f}) \cap D(\mathbf{g})$ and $a, b \in \mathbb{R}$.*
2. *If \mathbf{f} is continuous at \mathbf{x} , $\mathbf{f}(\mathbf{x}) \in D(\mathbf{g}) \subseteq \mathbb{R}^p$, and \mathbf{g} is continuous at $\mathbf{f}(\mathbf{x})$, then $\mathbf{g} \circ \mathbf{f}$ is continuous at \mathbf{x} .*
3. *If $\mathbf{f} = (f_1, \dots, f_q) : D(\mathbf{f}) \rightarrow \mathbb{R}^q$, then \mathbf{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 $f(\mathbf{x}) = |\mathbf{x}|$ is continuous.*

The proof of this theorem is in the last section of this chapter. Its conclusions are not surprising. For example the first claim says that $(a\mathbf{f} + b\mathbf{g})(\mathbf{y})$ is close to $(a\mathbf{f} + b\mathbf{g})(\mathbf{x})$ when \mathbf{y} is close to \mathbf{x} provided the same can be said about \mathbf{f} and \mathbf{g} . For the second claim, if \mathbf{y} is close to \mathbf{x} , $\mathbf{f}(\mathbf{x})$ is close to $\mathbf{f}(\mathbf{y})$ and so by continuity of \mathbf{g} at $\mathbf{f}(\mathbf{x})$, $\mathbf{g}(\mathbf{f}(\mathbf{y}))$ is close to $\mathbf{g}(\mathbf{f}(\mathbf{x}))$. 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.

For functions defined on \mathbb{R}^n , there is a notion of polynomial just as there is for functions defined on \mathbb{R} .

Definition 12.4.3 *Let α be an n dimensional multi-index. This means*

$$\alpha = (\alpha_1, \dots, \alpha_n)$$

where each α_i is a natural number or zero. Also, let

$$|\alpha| \equiv \sum_{i=1}^n |\alpha_i|$$

The symbol \mathbf{x}^α means

$$\mathbf{x}^\alpha \equiv x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n}.$$

An n dimensional polynomial of degree m is a function of the form

$$p(\mathbf{x}) = \sum_{|\alpha| \leq m} d_\alpha \mathbf{x}^\alpha.$$

where the d_α are real numbers.

The above theorem implies that polynomials are all continuous.

12.5 Limits Of A Function

As in the case of scalar valued functions of one variable, a concept closely related to continuity is that of the **limit of a function**. The notion of limit of a function makes sense at points \mathbf{x} , which are limit points of $D(\mathbf{f})$ and this concept is defined next.

Definition 12.5.1 Let $A \subseteq \mathbb{R}^m$ be a set. A point \mathbf{x} , is a *limit point* of A if $B(\mathbf{x}, r)$ contains infinitely many points of A for every $r > 0$.

Definition 12.5.2 Let $\mathbf{f} : D(\mathbf{f}) \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ 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,

$$|\mathbf{L} - \mathbf{f}(\mathbf{y})| < \varepsilon.$$

Theorem 12.5.3 If $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$ and $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}_1$, then $\mathbf{L} = \mathbf{L}_1$.

Proof: Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $0 < |\mathbf{y} - \mathbf{x}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < \varepsilon, |\mathbf{f}(\mathbf{y}) - \mathbf{L}_1| < \varepsilon.$$

Pick such a \mathbf{y} . There exists one because \mathbf{x} is a limit point of $D(\mathbf{f})$. Then

$$|\mathbf{L} - \mathbf{L}_1| \leq |\mathbf{L} - \mathbf{f}(\mathbf{y})| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}_1| < \varepsilon + \varepsilon = 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows $\mathbf{L} = \mathbf{L}_1$. ■

As in the case of functions of one variable, one can define what it means for $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \pm\infty$.

Definition 12.5.4 If $\mathbf{f}(\mathbf{x}) \in \mathbb{R}$, $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \infty$ if for every number l , there exists $\delta > 0$ such that whenever $|\mathbf{y} - \mathbf{x}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then $\mathbf{f}(\mathbf{y}) > l$. $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = -\infty$ if for every number l , there exists $\delta > 0$ such that whenever $|\mathbf{y} - \mathbf{x}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then $\mathbf{f}(\mathbf{y}) < l$.

The following theorem is just like the one variable version of calculus.

Theorem 12.5.5 Suppose $\mathbf{f} : D(\mathbf{f}) \rightarrow \mathbb{R}^q$. Then for \mathbf{x} a limit point of $D(\mathbf{f})$,

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L} \tag{12.1}$$

if and only if

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} f_k(\mathbf{y}) = L_k \tag{12.2}$$

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

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}, \lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{g}(\mathbf{y}) = \mathbf{K}$$

where $\mathbf{K}, \mathbf{L} \in \mathbb{R}^q$. Then if $a, b \in \mathbb{R}$,

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} (a\mathbf{f}(\mathbf{y}) + b\mathbf{g}(\mathbf{y})) = a\mathbf{L} + b\mathbf{K}, \quad (12.3)$$

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f} \cdot \mathbf{g}(\mathbf{y}) = \mathbf{L} \cdot \mathbf{K} \quad (12.4)$$

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

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) \times \mathbf{g}(\mathbf{y}) = \mathbf{L} \times \mathbf{K}. \quad (12.5)$$

If g is scalar valued with $\lim_{\mathbf{y} \rightarrow \mathbf{x}} g(\mathbf{y}) = K \neq 0$,

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) g(\mathbf{y}) = \mathbf{L}K. \quad (12.6)$$

Also, if h is a continuous function defined near \mathbf{L} , then

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} h \circ \mathbf{f}(\mathbf{y}) = h(\mathbf{L}). \quad (12.7)$$

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: Suppose (12.1). Then letting $\varepsilon > 0$ be given there exists $\delta > 0$ such that if $0 < |\mathbf{y} - \mathbf{x}| < \delta$, it follows

$$|f_k(\mathbf{y}) - L_k| \leq |\mathbf{f}(\mathbf{y}) - \mathbf{L}| < \varepsilon$$

which verifies (12.2).

Now suppose (12.2) holds. Then letting $\varepsilon > 0$ be given, there exists δ_k such that if $0 < |\mathbf{y} - \mathbf{x}| < \delta_k$, then

$$|f_k(\mathbf{y}) - L_k| < \frac{\varepsilon}{\sqrt{p}}.$$

Let $0 < \delta < \min(\delta_1, \dots, \delta_p)$. Then if $0 < |\mathbf{y} - \mathbf{x}| < \delta$, it follows

$$\begin{aligned} |\mathbf{f}(\mathbf{y}) - \mathbf{L}| &= \left(\sum_{k=1}^p |f_k(\mathbf{y}) - L_k|^2 \right)^{1/2} \\ &< \left(\sum_{k=1}^p \frac{\varepsilon^2}{p} \right)^{1/2} = \varepsilon. \end{aligned}$$

Each of the remaining assertions follows immediately from the coordinate descriptions of the various expressions and the first part. However, I will give a different argument for these.

The proof of (12.3) is left for you. It is like a corresponding theorem for continuous functions. Now (12.4) is to be verified. Let $\varepsilon > 0$ be given. Then by the triangle inequality,

$$\begin{aligned} |\mathbf{f} \cdot \mathbf{g}(\mathbf{y}) - \mathbf{L} \cdot \mathbf{K}| &\leq |\mathbf{f} \cdot \mathbf{g}(\mathbf{y}) - \mathbf{f}(\mathbf{y}) \cdot \mathbf{K}| + |\mathbf{f}(\mathbf{y}) \cdot \mathbf{K} - \mathbf{L} \cdot \mathbf{K}| \\ &\leq |\mathbf{f}(\mathbf{y})| |\mathbf{g}(\mathbf{y}) - \mathbf{K}| + |\mathbf{K}| |\mathbf{f}(\mathbf{y}) - \mathbf{L}|. \end{aligned}$$

There exists δ_1 such that if $0 < |\mathbf{y} - \mathbf{x}| < \delta_1$ and $\mathbf{y} \in D(\mathbf{f})$, then

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < 1,$$

and so for such \mathbf{y} , the triangle inequality implies, $|\mathbf{f}(\mathbf{y})| < 1 + |\mathbf{L}|$. Therefore, for $0 < |\mathbf{y} - \mathbf{x}| < \delta_1$,

$$|\mathbf{f} \cdot \mathbf{g}(\mathbf{y}) - \mathbf{L} \cdot \mathbf{K}| \leq (1 + |\mathbf{K}| + |\mathbf{L}|)[|\mathbf{g}(\mathbf{y}) - \mathbf{K}| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}|]. \quad (12.8)$$

Now let $0 < \delta_2$ be such that if $\mathbf{y} \in D(\mathbf{f})$ and $0 < |\mathbf{x} - \mathbf{y}| < \delta_2$,

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < \frac{\varepsilon}{2(1 + |\mathbf{K}| + |\mathbf{L}|)}, \quad |\mathbf{g}(\mathbf{y}) - \mathbf{K}| < \frac{\varepsilon}{2(1 + |\mathbf{K}| + |\mathbf{L}|)}.$$

Then letting $0 < \delta \leq \min(\delta_1, \delta_2)$, it follows from (12.8) that

$$|\mathbf{f} \cdot \mathbf{g}(\mathbf{y}) - \mathbf{L} \cdot \mathbf{K}| < \varepsilon$$

and this proves (12.4).

Consider (12.5). Let δ_1 be as above. From the properties of the cross product,

$$\begin{aligned} |\mathbf{f}(\mathbf{y}) \times \mathbf{g}(\mathbf{y}) - \mathbf{L} \times \mathbf{K}| &\leq |\mathbf{f}(\mathbf{y}) \times \mathbf{g}(\mathbf{y}) - \mathbf{f}(\mathbf{y}) \times \mathbf{K}| + |\mathbf{f}(\mathbf{y}) \times \mathbf{K} - \mathbf{L} \times \mathbf{K}| \\ &= |\mathbf{f}(\mathbf{y}) \times (\mathbf{g}(\mathbf{y}) - \mathbf{K})| + |(\mathbf{f}(\mathbf{y}) - \mathbf{L}) \times \mathbf{K}| \end{aligned}$$

Now from the geometric description of the cross product,

$$\leq |\mathbf{f}(\mathbf{y})| |\mathbf{g}(\mathbf{y}) - \mathbf{K}| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}| |\mathbf{K}|$$

Then if $0 < |\mathbf{y} - \mathbf{x}| < \delta_1$, this is no larger than

$$(1 + |\mathbf{L}|) |\mathbf{g}(\mathbf{y}) - \mathbf{K}| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}| |\mathbf{K}| \leq (1 + |\mathbf{K}| + |\mathbf{L}|) [|\mathbf{g}(\mathbf{y}) - \mathbf{K}| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}|]$$

and now the conclusion follows as before in the case of the dot product.

The proof of (12.6) is left to you.

Consider (12.7). Since \mathbf{h} is continuous near \mathbf{L} , it follows that for $\varepsilon > 0$ given, there exists $\eta > 0$ such that if $|\mathbf{y} - \mathbf{L}| < \eta$, then

$$|\mathbf{h}(\mathbf{y}) - \mathbf{h}(\mathbf{L})| < \varepsilon$$

Now since $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$, there exists $\delta > 0$ such that if $0 < |\mathbf{y} - \mathbf{x}| < \delta$, then

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < \eta.$$

Therefore, if $0 < |\mathbf{y} - \mathbf{x}| < \delta$,

$$|\mathbf{h}(\mathbf{f}(\mathbf{y})) - \mathbf{h}(\mathbf{L})| < \varepsilon.$$

It only remains to verify the last assertion. Assume $|\mathbf{f}(\mathbf{y}) - \mathbf{b}| \leq r$. It is required to show that $|\mathbf{L} - \mathbf{b}| \leq r$. If this is not true, then $|\mathbf{L} - \mathbf{b}| > r$. Consider $B(\mathbf{L}, |\mathbf{L} - \mathbf{b}| - r)$. Since \mathbf{L} is the limit of \mathbf{f} , it follows $\mathbf{f}(\mathbf{y}) \in B(\mathbf{L}, |\mathbf{L} - \mathbf{b}| - r)$ whenever $\mathbf{y} \in D(\mathbf{f})$ is close enough to \mathbf{x} . Thus, by the triangle inequality,

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < |\mathbf{L} - \mathbf{b}| - r$$

and so

$$\begin{aligned} r &< |L - \mathbf{b}| - |\mathbf{f}(\mathbf{y}) - L| \leq ||\mathbf{b} - L| - |\mathbf{f}(\mathbf{y}) - L|| \\ &\leq |\mathbf{b} - \mathbf{f}(\mathbf{y})|, \end{aligned}$$

a contradiction to the assumption that $|\mathbf{b} - \mathbf{f}(\mathbf{y})| \leq r$. ■

The relation between continuity and limits is as follows.

Theorem 12.5.6 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 12.5.7 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 12.5.8 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.

12.6 Properties Of Continuous Functions

Functions of p variables have many of the same properties as functions of one variable. First there is a version of the extreme value theorem generalizing the one dimensional case.

Theorem 12.6.1 *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, $x_1, x_2 \in C$ such that for all $x \in C$,*

$$f(x_1) \leq f(x) \leq f(x_2).$$

There is also the long technical theorem about sums and products of continuous functions. These theorems are proved later in this chapter.

Theorem 12.6.2 *The following assertions are valid.*

1. *The function $a\mathbf{f} + b\mathbf{g}$ is continuous at \mathbf{x} when \mathbf{f}, \mathbf{g} are continuous at $\mathbf{x} \in D(\mathbf{f}) \cap D(\mathbf{g})$ and $a, b \in \mathbb{R}$.*
2. *If \mathbf{f} and \mathbf{g} are each real valued functions continuous at \mathbf{x} , then $\mathbf{f}\mathbf{g}$ is continuous at \mathbf{x} . If, in addition to this, $g(\mathbf{x}) \neq 0$, then \mathbf{f}/\mathbf{g} is continuous at \mathbf{x} .*
3. *If \mathbf{f} is continuous at \mathbf{x} , $\mathbf{f}(\mathbf{x}) \in D(\mathbf{g}) \subseteq \mathbb{R}^p$, and \mathbf{g} is continuous at $\mathbf{f}(\mathbf{x})$, then $\mathbf{g} \circ \mathbf{f}$ is continuous at \mathbf{x} .*
4. *If $\mathbf{f} = (f_1, \dots, f_q) : D(\mathbf{f}) \rightarrow \mathbb{R}^q$, then \mathbf{f} is continuous if and only if each f_k is a continuous real valued function.*
5. *The function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, given by $f(\mathbf{x}) = |\mathbf{x}|$ is continuous.*

12.7 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 $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K|\mathbf{x} - \mathbf{y}|$ 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 Holder continuous.
7. Suppose $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ is given by $f(\mathbf{x}) = 3x_1x_2 + 2x_3^2$. Use Theorem 12.4.2 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 which involves quotients of functions encountered in the previous problem.

10. Let

$$f(x, y) \equiv \begin{cases} \frac{2x^2 - y^2}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}.$$

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) $\lim_{(x,y) \rightarrow (0,0)} \frac{x(x^2 - y^2)}{(x^2 + y^2)} = 0.$

(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:** It might help to write this in terms of 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. $f : D \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ is Lipschitz continuous or just Lipschitz for short if there exists a constant K such that

$$|f(x) - f(y)| \leq K|x - y|$$

for all $x, y \in D$. Show every Lipschitz function is uniformly continuous which means that given $\varepsilon > 0$ there exists $\delta > 0$ independent of x such that if $|x - y| < \delta$, then $|f(x) - f(y)| < \varepsilon$.

14. 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.

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 \mathbf{x} is defined to be a limit point of a set A if and only if for all $r > 0$, $B(\mathbf{x}, r)$ contains a point of A different than \mathbf{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 $D(\mathbf{f})$ equals this set, then \mathbf{f} is continuous. Show that more generally, if \mathbf{f} is any function for which $D(\mathbf{f})$ has no limit points, then \mathbf{f} is continuous.
20. Let $\{\mathbf{x}_k\}_{k=1}^n$ 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 $\lim_{\mathbf{x} \rightarrow \mathbf{0}} \frac{\sin(|\mathbf{x}|)}{|\mathbf{x}|}$ and prove your answer from the definition of limit.
23. Suppose \mathbf{g} is a continuous vector valued function of one variable defined on $[0, \infty)$. Prove

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \mathbf{g}(|\mathbf{x}|) = \mathbf{g}(|\mathbf{x}_0|).$$

12.8 Open And Closed Sets

Eventually, one must consider functions which are defined on subsets of \mathbb{R}^n and their properties. The next definition will end up being quite important. It describes a type of subset of \mathbb{R}^n with the property that if \mathbf{x} is in this set, then so is \mathbf{y} whenever \mathbf{y} is close enough to \mathbf{x} .

Definition 12.8.1 Recall that for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$|\mathbf{x} - \mathbf{y}| = \left(\sum_{i=1}^n |x_i - y_i|^2 \right)^{1/2}.$$

Also let

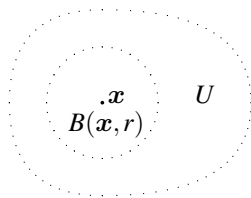
$$B(\mathbf{x}, r) \equiv \{\mathbf{y} \in \mathbb{R}^n : |\mathbf{x} - \mathbf{y}| < r\}$$

Let $U \subseteq \mathbb{R}^n$. U is an **open set** if whenever $\mathbf{x} \in U$, there exists $r > 0$ such that $B(\mathbf{x}, r) \subseteq U$. More generally, if U is any subset of \mathbb{R}^n , $\mathbf{x} \in U$ is an **interior point** of U if there exists $r > 0$ such that $B(\mathbf{x}, r) \subseteq U$. 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.

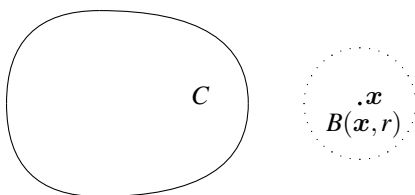
Definition 12.8.2 A subset, C , of \mathbb{R}^n is called a **closed set** if $\mathbb{R}^n \setminus C$ is an open set. The symbol $\mathbb{R}^n \setminus C$ denotes everything in \mathbb{R}^n which is not in C . It is also called the **complement** of C . The symbol S^C is a short way of writing $\mathbb{R}^n \setminus S$.

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.

It is roughly the case that 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 $\mathbb{R}^n \setminus C$ is open, there exists a ball, $B(x, r)$ contained entirely in $\mathbb{R}^n \setminus C$. If you look at $\mathbb{R}^n \setminus C$, what would be its skin? It can't be in $\mathbb{R}^n \setminus C$ 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}^n 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 $\mathbb{R}^n \setminus \emptyset = \mathbb{R}^n$. Therefore, \emptyset is both open and closed. From this, it follows \mathbb{R}^n is also both open and closed.

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.



Theorem 12.8.3 Let $x \in \mathbb{R}^n$ and let $r \geq 0$. Then $B(x, r)$ is an open set. Also,

$$D(x, r) \equiv \{y \in \mathbb{R}^n : |y - x| \leq r\}$$

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

$$\begin{aligned} |z - x| &= |z - y + y - x| \\ &\leq |z - y| + |y - x| \\ &< r_1 + |y - x| = r - |x - y| + |y - x| = r. \end{aligned}$$

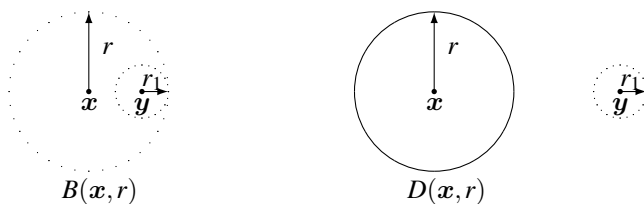
Note that if $r = 0$ then $B(x, r) = \emptyset$, the empty set. This is because if $y \in \mathbb{R}^n$, $|x - y| \geq 0$ and so $y \notin B(x, 0)$. Since \emptyset has no points in it, it must be open because every point in it, (There are none.) satisfies the desired property of being an interior point.

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 $B(y, \delta) \subseteq \mathbb{R}^n \setminus D(x, r)$. Since y was an arbitrary point in $\mathbb{R}^n \setminus D(x, r)$, it follows $\mathbb{R}^n \setminus D(x, r)$ is an open set which shows, from the definition, that $D(x, r)$ is a closed set as claimed. ■

A picture which is descriptive of the conclusion of the above theorem which also implies the manner of proof is the following.



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 12.8.4 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 $B \subseteq \mathbb{R}^n$. Then if

$$(\mathbf{x}, \mathbf{y}) \in A \times B, \mathbf{x} = (x_1, \dots, x_m) \text{ and } \mathbf{y} = (y_1, \dots, y_n),$$

the following identification will be made.

$$(\mathbf{x}, \mathbf{y}) = (x_1, \dots, x_m, y_1, \dots, y_n) \in \mathbb{R}^{n+m}.$$

Similarly, starting with something in \mathbb{R}^{n+m} , you can write it in the form (\mathbf{x}, \mathbf{y}) where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$. 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 12.8.5 A set, $A \subseteq \mathbb{R}^n$ is said to be **bounded** if there exist finite intervals, $[a_i, b_i]$ such that

$$A \subseteq \prod_{i=1}^n [a_i, b_i].$$

Theorem 12.8.6 Let U be an open set in \mathbb{R}^m and let V be an open set in \mathbb{R}^n . Then $U \times V$ is an open set in \mathbb{R}^{n+m} . If C is a closed set in \mathbb{R}^m and H is a closed set in \mathbb{R}^n , then $C \times H$ is a closed set in \mathbb{R}^{n+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

$$B((\mathbf{x}, \mathbf{y}), \delta) \equiv \left\{ (\mathbf{s}, \mathbf{t}) \in \mathbb{R}^{n+m} : \sum_{k=1}^m |x_k - s_k|^2 + \sum_{j=1}^n |y_j - t_j|^2 < \delta^2 \right\}$$

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 $B((\mathbf{x}, \mathbf{y}), \delta) \subseteq \mathbb{R}^{n+m} \setminus (C \times H)$. 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, $B((\mathbf{x}, \mathbf{y}), r) \subseteq \mathbb{R}^{n+m} \setminus (C \times H)$ 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, $H \subseteq \prod_{i=m+1}^{m+n} [a_i, b_i]$ for intervals $[a_{m+1}, b_{m+1}], \dots, [a_{m+n}, b_{m+n}]$. Therefore, $C \times H \subseteq \prod_{i=1}^{m+n} [a_i, b_i]$.

■

12.9 Exercises

1. Let $U = \{(x, y, z) \text{ such that } z > 0\}$. Determine whether U is open, closed or neither.
2. Let $U = \{(x, y, z) \text{ such that } z \geq 0\}$. Determine whether U is open, closed or neither.
3. Let $U = \{(x, y, z) \text{ such that } \sqrt{x^2 + y^2 + z^2} < 1\}$. Determine whether U is open, closed or neither.
4. Let $U = \{(x, y, z) \text{ such that } \sqrt{x^2 + y^2 + z^2} \leq 1\}$. Determine whether U is open, closed or neither.
5. Show carefully that \mathbb{R}^n is both open and closed.
6. Show that every open set in \mathbb{R}^n is the union of open balls contained in it.
7. Show the intersection of any two open sets is an open set.
8. If S is a nonempty subset of \mathbb{R}^p , a point x is said to be a **limit point** of S if $B(x, r)$ contains infinitely many points of S for each $r > 0$. Show this is equivalent to saying that $B(x, r)$ contains a point of S different than x for each $r > 0$.
9. 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.

Chapter 13

Some Fundamentals*



This section contains the proofs of the theorems which were stated without proof along with some other significant topics which will be useful later. These topics are of fundamental significance but are difficult.

13.1 Combinations Of Continuous Functions

Theorem 13.1.1 *The following assertions are valid.*

1. *The function $a\mathbf{f} + b\mathbf{g}$ is continuous at \mathbf{x} when \mathbf{f}, \mathbf{g} are continuous at $\mathbf{x} \in D(\mathbf{f}) \cap D(\mathbf{g})$ and $a, b \in \mathbb{R}$.*
2. *If \mathbf{f} and \mathbf{g} are each real valued functions continuous at \mathbf{x} , then $\mathbf{f}\mathbf{g}$ is continuous at \mathbf{x} . If, in addition to this, $\mathbf{g}(\mathbf{x}) \neq 0$, then \mathbf{f}/\mathbf{g} is continuous at \mathbf{x} .*
3. *If \mathbf{f} is continuous at \mathbf{x} , $\mathbf{f}(\mathbf{x}) \in D(\mathbf{g}) \subseteq \mathbb{R}^p$, and \mathbf{g} is continuous at $\mathbf{f}(\mathbf{x})$, then $\mathbf{g} \circ \mathbf{f}$ is continuous at \mathbf{x} .*
4. *If $\mathbf{f} = (f_1, \dots, f_q) : D(\mathbf{f}) \rightarrow \mathbb{R}^q$, then \mathbf{f} is continuous if and only if each f_k is a continuous real valued function.*
5. *The function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, given by $f(\mathbf{x}) = |\mathbf{x}|$ is continuous.*

Proof: Begin with (1). Let $\varepsilon > 0$ be given. By assumption, there exist $\delta_1 > 0$ such that whenever $|\mathbf{x} - \mathbf{y}| < \delta_1$, it follows $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \frac{\varepsilon}{2(|a|+|b|+1)}$ and there exists $\delta_2 > 0$ such that whenever $|\mathbf{x} - \mathbf{y}| < \delta_2$, it follows that $|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| < \frac{\varepsilon}{2(|a|+|b|+1)}$. Then let $0 < \delta \leq \min(\delta_1, \delta_2)$. If $|\mathbf{x} - \mathbf{y}| < \delta$, then everything happens at once. Therefore, using the triangle inequality

$$|a\mathbf{f}(\mathbf{x}) + b\mathbf{f}(\mathbf{x}) - (a\mathbf{g}(\mathbf{y}) + b\mathbf{g}(\mathbf{y}))|$$

$$\begin{aligned}
&\leq |a| |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + |b| |\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| \\
&< |a| \left(\frac{\varepsilon}{2(|a| + |b| + 1)} \right) + |b| \left(\frac{\varepsilon}{2(|a| + |b| + 1)} \right) < \varepsilon.
\end{aligned}$$

Now begin on (2). There exists $\delta_1 > 0$ such that if $|\mathbf{y} - \mathbf{x}| < \delta_1$, then

$$|f(\mathbf{x}) - f(\mathbf{y})| < 1$$

Therefore, for such \mathbf{y} ,

$$|f(\mathbf{y})| < 1 + |f(\mathbf{x})|.$$

It follows that for such \mathbf{y} ,

$$\begin{aligned}
|fg(\mathbf{x}) - fg(\mathbf{y})| &\leq |f(\mathbf{x})g(\mathbf{x}) - g(\mathbf{x})f(\mathbf{y})| + |g(\mathbf{x})f(\mathbf{y}) - f(\mathbf{y})g(\mathbf{y})| \\
&\leq |g(\mathbf{x})| |f(\mathbf{x}) - f(\mathbf{y})| + |f(\mathbf{y})| |g(\mathbf{x}) - g(\mathbf{y})| \\
&\leq (1 + |g(\mathbf{x})| + |f(\mathbf{y})|) [|g(\mathbf{x}) - g(\mathbf{y})| + |f(\mathbf{x}) - f(\mathbf{y})|].
\end{aligned}$$

Now let $\varepsilon > 0$ be given. There exists δ_2 such that if $|\mathbf{x} - \mathbf{y}| < \delta_2$, then

$$|g(\mathbf{x}) - g(\mathbf{y})| < \frac{\varepsilon}{2(1 + |g(\mathbf{x})| + |f(\mathbf{y})|)},$$

and there exists δ_3 such that if $|\mathbf{x} - \mathbf{y}| < \delta_3$, then

$$|f(\mathbf{x}) - f(\mathbf{y})| < \frac{\varepsilon}{2(1 + |g(\mathbf{x})| + |f(\mathbf{y})|)}$$

Now let $0 < \delta \leq \min(\delta_1, \delta_2, \delta_3)$. Then if $|\mathbf{x} - \mathbf{y}| < \delta$, all the above hold at once and

$$\begin{aligned}
&|fg(\mathbf{x}) - fg(\mathbf{y})| \leq \\
&(1 + |g(\mathbf{x})| + |f(\mathbf{y})|) [|g(\mathbf{x}) - g(\mathbf{y})| + |f(\mathbf{x}) - f(\mathbf{y})|] \\
&< (1 + |g(\mathbf{x})| + |f(\mathbf{y})|) \left(\frac{\varepsilon}{2(1 + |g(\mathbf{x})| + |f(\mathbf{y})|)} + \frac{\varepsilon}{2(1 + |g(\mathbf{x})| + |f(\mathbf{y})|)} \right) = \varepsilon.
\end{aligned}$$

This proves the first part of (2). To obtain the second part, let δ_1 be as described above and let $\delta_0 > 0$ be such that for $|\mathbf{x} - \mathbf{y}| < \delta_0$,

$$|g(\mathbf{x}) - g(\mathbf{y})| < |g(\mathbf{x})|/2$$

and so by the triangle inequality,

$$-|g(\mathbf{x})|/2 \leq |g(\mathbf{y})| - |g(\mathbf{x})| \leq |g(\mathbf{x})|/2$$

which implies $|g(\mathbf{y})| \geq |g(\mathbf{x})|/2$, and $|g(\mathbf{y})| < 3|g(\mathbf{x})|/2$.

Then if $|\mathbf{x} - \mathbf{y}| < \min(\delta_0, \delta_1)$,

$$\begin{aligned}
\left| \frac{f(\mathbf{x})}{g(\mathbf{x})} - \frac{f(\mathbf{y})}{g(\mathbf{y})} \right| &= \left| \frac{f(\mathbf{x})g(\mathbf{y}) - f(\mathbf{y})g(\mathbf{x})}{g(\mathbf{x})g(\mathbf{y})} \right| \\
&\leq \frac{|f(\mathbf{x})g(\mathbf{y}) - f(\mathbf{y})g(\mathbf{x})|}{\left(\frac{|g(\mathbf{x})|^2}{2} \right)} \\
&= \frac{2|f(\mathbf{x})g(\mathbf{y}) - f(\mathbf{y})g(\mathbf{x})|}{|g(\mathbf{x})|^2}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{2}{|g(x)|^2} [|f(x)g(y) - f(y)g(y) + f(y)g(y) - f(y)g(x)|] \\
&\leq \frac{2}{|g(x)|^2} [|g(y)||f(x) - f(y)| + |f(y)||g(y) - g(x)|] \\
&\leq \frac{2}{|g(x)|^2} \left[\frac{3}{2} |g(x)||f(x) - f(y)| + (1 + |f(x)|)|g(y) - g(x)| \right] \\
&\leq \frac{2}{|g(x)|^2} (1 + 2|f(x)| + 2|g(x)|) [|f(x) - f(y)| + |g(y) - g(x)|] \\
&\equiv M [|f(x) - f(y)| + |g(y) - g(x)|]
\end{aligned}$$

where

$$M \equiv \frac{2}{|g(x)|^2} (1 + 2|f(x)| + 2|g(x)|)$$

Now let δ_2 be such that if $|x - y| < \delta_2$, then

$$|f(x) - f(y)| < \frac{\varepsilon}{2} M^{-1}$$

and let δ_3 be such that if $|x - y| < \delta_3$, then

$$|g(y) - g(x)| < \frac{\varepsilon}{2} M^{-1}.$$

Then if $0 < \delta \leq \min(\delta_0, \delta_1, \delta_2, \delta_3)$, and $|x - y| < \delta$, everything holds and

$$\begin{aligned}
\left| \frac{f(x)}{g(x)} - \frac{f(y)}{g(y)} \right| &\leq M [|f(x) - f(y)| + |g(y) - g(x)|] \\
&< M \left[\frac{\varepsilon}{2} M^{-1} + \frac{\varepsilon}{2} M^{-1} \right] = \varepsilon.
\end{aligned}$$

This completes the proof of the second part of (2). Note that in these proofs no effort is made to find some sort of “best” δ . The problem is one which has a yes or a no answer. Either it is or it is not continuous.

Now begin on (3). 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 . Let $\varepsilon > 0$ be given. Then there exists $\eta > 0$ such that if $|y - f(x)| < \eta$ and $y \in D(g)$, it follows that $|g(y) - g(f(x))| < \varepsilon$. It follows from continuity of f at x that there exists $\delta > 0$ such that if $|x - z| < \delta$ and $z \in D(f)$, then $|f(z) - f(x)| < \eta$. Then if $|x - z| < \delta$ and $z \in D(g \circ f) \subseteq D(f)$, all the above hold and so

$$|g(f(z)) - g(f(x))| < \varepsilon.$$

This proves part (3).

Part (4) says: 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. Then

$$\begin{aligned}
|f_k(x) - f_k(y)| &\leq |f(x) - f(y)| \equiv \left(\sum_{i=1}^q |f_i(x) - f_i(y)|^2 \right)^{1/2} \\
&\leq \sum_{i=1}^q |f_i(x) - f_i(y)|.
\end{aligned} \tag{13.1}$$

Suppose first that \mathbf{f} is continuous at \mathbf{x} . Then there exists $\delta > 0$ such that if $|\mathbf{x} - \mathbf{y}| < \delta$, then $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon$. The first part of the above inequality then shows that for each $k = 1, \dots, q$, $|f_k(\mathbf{x}) - f_k(\mathbf{y})| < \varepsilon$. This shows the only if part. Now suppose each function f_k is continuous. Then if $\varepsilon > 0$ is given, there exists $\delta_k > 0$ such that whenever $|\mathbf{x} - \mathbf{y}| < \delta_k$

$$|f_k(\mathbf{x}) - f_k(\mathbf{y})| < \varepsilon/q.$$

Now let $0 < \delta \leq \min(\delta_1, \dots, \delta_q)$. For $|\mathbf{x} - \mathbf{y}| < \delta$, the above inequality holds for all k and so the last part of (13.1) implies

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq \sum_{i=1}^q |f_i(\mathbf{x}) - f_i(\mathbf{y})| < \sum_{i=1}^q \frac{\varepsilon}{q} = \varepsilon.$$

This proves part (4).

To verify part (5), let $\varepsilon > 0$ be given and let $\delta = \varepsilon$. Then if $|\mathbf{x} - \mathbf{y}| < \delta$, the triangle inequality implies

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| = ||\mathbf{x}| - |\mathbf{y}|| \leq |\mathbf{x} - \mathbf{y}| < \delta = \varepsilon.$$

This proves part (5) and completes the proof of the theorem. ■

13.2 The Nested Interval Lemma

First, here is the one dimensional nested interval lemma.

Lemma 13.2.1 *Let $I_k = [a_k, b_k]$ be closed intervals, $a_k \leq b_k$, such that $I_k \supseteq I_{k+1}$ for all k . Then there exists a point c which is contained in all these intervals. If $\lim_{k \rightarrow \infty} (b_k - a_k) = 0$, then there is exactly one such point.*

Proof: Note that the $\{a_k\}$ are an increasing sequence and that $\{b_k\}$ is a decreasing sequence. Now note that if $m < n$, then

$$a_m \leq a_n \leq b_n$$

while if $m > n$,

$$b_n \geq b_m \geq a_m.$$

It follows that $a_m \leq b_n$ for any pair m, n . Therefore, each b_n is an upper bound for all the a_m and so if $c \equiv \sup \{a_k\}$, then for each n , it follows that $c \leq b_n$ and so for all, $a_n \leq c \leq b_n$ which shows that c is in all of these intervals.

If the condition on the lengths of the intervals holds, then if c, c' are in all the intervals, then if they are not equal, then eventually, for large enough k , they cannot both be contained in $[a_k, b_k]$ since eventually $b_k - a_k < |c - c'|$. This would be a contradiction. Hence $c = c'$. ■

Definition 13.2.2 *The **diameter** of a set S , is defined as*

$$\text{diam}(S) \equiv \sup \{|\mathbf{x} - \mathbf{y}| : \mathbf{x}, \mathbf{y} \in S\}.$$

Thus $\text{diam}(S)$ is just a careful description of what you would think of as the diameter. It measures how stretched out the set is.

Here is a multidimensional version of the nested interval lemma.

Lemma 13.2.3 Let $I_k = \prod_{i=1}^p [a_i^k, b_i^k] \equiv \{x \in \mathbb{R}^p : x_i \in [a_i^k, b_i^k]\}$ and suppose that for all $k = 1, 2, \dots$,

$$I_k \supseteq I_{k+1}.$$

Then there exists a point $c \in \mathbb{R}^p$ which is an element of every I_k . If $\lim_{k \rightarrow \infty} \text{diam}(I_k) = 0$, then the point c is unique.

Proof: For each $i = 1, \dots, p$, $[a_i^k, b_i^k] \supseteq [a_i^{k+1}, b_i^{k+1}]$ and so, by Lemma 13.2.1, there exists a point $c_i \in [a_i^k, b_i^k]$ for all k . Then letting $c \equiv (c_1, \dots, c_p)$ it follows $c \in I_k$ for all k . If the condition on the diameters holds, then the lengths of the intervals $\lim_{k \rightarrow \infty} [a_i^k, b_i^k] = 0$ and so by the same lemma, each c_i is unique. Hence c is unique. ■

I will sometimes refer to the above Cartesian product of closed intervals as an interval to emphasize the analogy with one dimensions, and sometimes as a box.

13.3 Convergent Sequences, Sequential Compactness

A mapping $f : \{k, k+1, k+2, \dots\} \rightarrow \mathbb{R}^p$ is called a sequence. We usually write it in the form $\{a_j\}$ where it is understood that $a_j \equiv f(j)$. In the same way as for sequences of real numbers, one can define what it means for convergence to take place.

Definition 13.3.1 A sequence, $\{a_k\}$ is said to **converge** to a if for every $\varepsilon > 0$ there exists n_ε such that if $n > n_\varepsilon$, then $|a - a_n| < \varepsilon$. The usual notation for this is $\lim_{n \rightarrow \infty} a_n = a$ although it is often written as $a_n \rightarrow a$.

One can also define a subsequence in the same way as in the case of real valued sequences.

Definition 13.3.2 $\{a_{n_k}\}$ is a **subsequence** of $\{a_n\}$ if $n_1 < n_2 < \dots$.

The following theorem says the limit, if it exists, is unique.

Theorem 13.3.3 If a sequence, $\{a_n\}$ converges to a and to b then $a = b$.

Proof: There exists n_ε such that if $n > n_\varepsilon$ then $|a_n - a| < \frac{\varepsilon}{2}$ and if $n > n_\varepsilon$, then $|a_n - b| < \frac{\varepsilon}{2}$. Then pick such an n .

$$|a - b| < |a - a_n| + |a_n - b| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since ε is arbitrary, this proves the theorem. ■

The following is the definition of a Cauchy sequence in \mathbb{R}^p .

Definition 13.3.4 $\{a_n\}$ is a **Cauchy sequence** if for all $\varepsilon > 0$, there exists n_ε such that whenever $n, m \geq n_\varepsilon$,

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

A sequence is Cauchy, means the terms are “bunching up to each other” as m, n get large.

Theorem 13.3.5 The set of terms in a Cauchy sequence in \mathbb{R}^p is bounded in the sense that for all n , $|a_n| < M$ for some $M < \infty$.

Proof: Let $\varepsilon = 1$ in the definition of a Cauchy sequence and let $n > n_1$. Then from the definition,

$$|a_n - a_{n_1}| < 1.$$

It follows that for all $n > n_1$,

$$|a_n| < 1 + |a_{n_1}|.$$

Therefore, for all n ,

$$|a_n| \leq 1 + |a_{n_1}| + \sum_{k=1}^{n_1} |a_k|. \quad \blacksquare$$

Theorem 13.3.6 *If a sequence $\{a_n\}$ in \mathbb{R}^p converges, then the sequence is a Cauchy sequence. Also, if some subsequence of a Cauchy sequence converges, then the original sequence converges.*

Proof: Let $\varepsilon > 0$ be given and suppose $a_n \rightarrow a$. Then from the definition of convergence, there exists n_ε such that if $n > n_\varepsilon$, it follows that

$$|a_n - a| < \frac{\varepsilon}{2}$$

Therefore, if $m, n \geq n_\varepsilon + 1$, it follows that

$$|a_n - a_m| \leq |a_n - a| + |a - a_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

showing that, since $\varepsilon > 0$ is arbitrary, $\{a_n\}$ is a Cauchy sequence. It remains to that the last claim.

Suppose then that $\{a_n\}$ is a Cauchy sequence and $a = \lim_{k \rightarrow \infty} a_{n_k}$ where $\{a_{n_k}\}_{k=1}^\infty$ is a subsequence. Let $\varepsilon > 0$ be given. Then there exists K such that if $k, l \geq K$, then $|a_k - a_l| < \frac{\varepsilon}{2}$. Then if $k > K$, it follows $n_k > K$ because n_1, n_2, n_3, \dots is strictly increasing as the subscript increases. Also, there exists K_1 such that if $k > K_1$, $|a_{n_k} - a| < \frac{\varepsilon}{2}$. Then letting $n > \max(K, K_1)$, pick $k > \max(K, K_1)$. Then

$$|a - a_n| \leq |a - a_{n_k}| + |a_{n_k} - a_n| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Therefore, the sequence converges. \blacksquare

Definition 13.3.7 *A set K in \mathbb{R}^p is said to be **sequentially compact** if every sequence in K has a subsequence which converges to a point of K .*

Theorem 13.3.8 *If $I_0 = \prod_{i=1}^p [a_i, b_i]$ where $a_i \leq b_i$, then I_0 is sequentially compact.*

Proof: Let $\{a_k\}_{k=1}^\infty \subseteq I_0$ and consider all sets of the form $\prod_{i=1}^p [c_i, d_i]$ where $[c_i, d_i]$ equals either $\left[a_i, \frac{a_i + b_i}{2}\right]$ or $[c_i, d_i] = \left[\frac{a_i + b_i}{2}, b_i\right]$. Thus there are 2^p of these sets because there are two choices for the i^{th} slot for $i = 1, \dots, p$. Also, if x and y are two points in one of these sets,

$$|x_i - y_i| \leq 2^{-1} |b_i - a_i|.$$

$$\text{diam}(I_0) = \left(\sum_{i=1}^p |b_i - a_i|^2 \right)^{1/2},$$

$$\begin{aligned} |\mathbf{x} - \mathbf{y}| &= \left(\sum_{i=1}^p |x_i - y_i|^2 \right)^{1/2} \\ &\leq 2^{-1} \left(\sum_{i=1}^p |b_i - a_i|^2 \right)^{1/2} \equiv 2^{-1} \text{diam}(I_0). \end{aligned}$$

In particular, since $\mathbf{d} \equiv (d_1, \dots, d_p)$ and $\mathbf{c} \equiv (c_1, \dots, c_p)$ are two such points,

$$D_1 \equiv \left(\sum_{i=1}^p |d_i - c_i|^2 \right)^{1/2} \leq 2^{-1} \text{diam}(I_0)$$

Denote by $\{J_1, \dots, J_{2^p}\}$ these sets determined above. Since the union of these sets equals all of $I_0 \equiv I$, it follows that for some J_k , the sequence, $\{\mathbf{a}_i\}$ is contained in J_k for infinitely many k . Let that one be called I_1 . Next do for I_1 what was done for I_0 to get $I_2 \subseteq I_1$ such that the diameter is half that of I_1 and I_2 contains $\{\mathbf{a}_k\}$ for infinitely many values of k . Continue in this way obtaining a nested sequence $\{I_k\}$ such that $I_k \supseteq I_{k+1}$, and if $\mathbf{x}, \mathbf{y} \in I_k$, then $|\mathbf{x} - \mathbf{y}| \leq 2^{-k} \text{diam}(I_0)$, and I_n contains $\{\mathbf{a}_k\}$ for infinitely many values of k for each n . Then by the nested interval lemma, there exists \mathbf{c} such that \mathbf{c} is contained in each I_k . Pick $\mathbf{a}_{n_1} \in I_1$. Next pick $n_2 > n_1$ such that $\mathbf{a}_{n_2} \in I_2$. If $\mathbf{a}_{n_1}, \dots, \mathbf{a}_{n_k}$ have been chosen, let $\mathbf{a}_{n_{k+1}} \in I_{k+1}$ and $n_{k+1} > n_k$. This can be done because in the construction, I_n contains $\{\mathbf{a}_k\}$ for infinitely many k . Thus the distance between \mathbf{a}_{n_k} and \mathbf{c} is no larger than $2^{-k} \text{diam}(I_0)$, and so $\lim_{k \rightarrow \infty} \mathbf{a}_{n_k} = \mathbf{c} \in I_0$. ■

Corollary 13.3.9 *Let K be a closed and bounded set of points in \mathbb{R}^p . Then K is sequentially compact.*

Proof: Since K is closed and bounded, there exists a closed rectangle, $\prod_{k=1}^p [a_k, b_k]$ which contains K . Now let $\{\mathbf{x}_k\}$ be a sequence of points in K . By Theorem 13.3.8, there exists a subsequence $\{\mathbf{x}_{n_k}\}$ such that $\mathbf{x}_{n_k} \rightarrow \mathbf{x} \in \prod_{k=1}^p [a_k, b_k]$. However, K is closed and each of the points of the sequence is in K so $\mathbf{x} \in K$. If not, then since K^C is open, it would follow that eventually $\mathbf{x}_{n_k} \in K^C$ which is impossible. ■

Theorem 13.3.10 *Every Cauchy sequence in \mathbb{R}^p converges.*

Proof: Let $\{\mathbf{a}_k\}$ be a Cauchy sequence. By Theorem 13.3.5, there exists some large enough box $\prod_{i=1}^p [a_i, b_i]$ containing all the terms of $\{\mathbf{a}_k\}$. Therefore, by Theorem 13.3.8, a subsequence converges to a point of $\prod_{i=1}^p [a_i, b_i]$. By Theorem 13.3.6, the original sequence converges. ■

13.4 Continuity And The Limit Of A Sequence

Just as in the case of a function of one variable, there is a very useful way of thinking of continuity in terms of limits of sequences found in the following theorem. In words, it says a function is continuous if it takes convergent sequences to convergent sequences whenever possible.

Theorem 13.4.1 *A function $f : D(f) \rightarrow \mathbb{R}^q$ is continuous at $x \in D(f)$ if and only if, whenever $x_n \rightarrow x$ with $x_n \in D(f)$, it follows $f(x_n) \rightarrow f(x)$.*

Proof: Suppose first that f is continuous at x and let $x_n \rightarrow x$. Let $\varepsilon > 0$ be given. By continuity, there exists $\delta > 0$ such that if $|y - x| < \delta$, then $|f(y) - f(x)| < \varepsilon$. However, there exists n_δ such that if $n \geq n_\delta$, then $|x_n - x| < \delta$, and so for all n this large,

$$|f(x) - f(x_n)| < \varepsilon$$

which shows $f(x_n) \rightarrow f(x)$.

Now suppose the condition about taking convergent sequences to convergent sequences holds at x . Suppose f fails to be continuous at x . Then there exists $\varepsilon > 0$ and $x_n \in D(f)$ such that $|x - x_n| < \frac{1}{n}$, yet

$$|f(x) - f(x_n)| \geq \varepsilon.$$

But this is clearly a contradiction because, although $x_n \rightarrow x$, $f(x_n)$ fails to converge to $f(x)$. It follows f must be continuous after all. ■

13.5 The Extreme Value Theorem And Uniform Continuity

Definition 13.5.1 *A function f having values in \mathbb{R}^p is said to be bounded if the set of values of f is a bounded set.*

Lemma 13.5.2 *Let $C \subseteq \mathbb{R}^p$ be closed and bounded and let $f : C \rightarrow \mathbb{R}^s$ be continuous. Then f is bounded.*

Proof: Suppose not. Then since f is not bounded, there exists x_n such that

$$f(x_n) \notin \prod_{i=1}^s (-n, n) \equiv R_n.$$

By Corollary 13.3.9, C is sequentially compact, and so there exists a subsequence $\{x_{n_k}\}$ which converges to $x \in C$. Now $f(x) \in R_m$ for large enough m . Hence, by continuity of f , it follows $f(x_{n_k}) \in R_m$ for all k large enough, contradicting the construction. ■

Here is a proof of the extreme value theorem.

Theorem 13.5.3 *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 $x_1, x_2 \in C$ such that for all $x \in C$,*

$$f(x_1) \leq f(x) \leq f(x_2).$$

Proof: Let $M = \sup \{f(x) : x \in C\}$. Then by Lemma 13.5.2, M is a finite number. Is $f(x_2) = M$ for some x_2 ? If not, you could consider the function

$$g(x) \equiv \frac{1}{M - f(x)}$$

and g would be a continuous and unbounded function defined on C , contrary to Lemma 13.5.2. Therefore, there exists $x_2 \in C$ such that $f(x_2) = M$. A similar argument applies to show the existence of $x_1 \in C$ such that

$$f(x_1) = \inf \{f(x) : x \in C\}.$$

■

As in the case of a function of one variable, there is a concept of uniform continuity.

Definition 13.5.4 A function $f : D(f) \rightarrow \mathbb{R}^q$ is uniformly continuous if for every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever x, y are points of $D(f)$ such that $|x - y| < \delta$, it follows $|f(x) - f(y)| < \varepsilon$.

Theorem 13.5.5 Let $f : K \rightarrow \mathbb{R}^q$ be continuous at every point of K where K is a closed and bounded set in \mathbb{R}^p . Then f is uniformly continuous.

Proof: Suppose not. Then there exists $\varepsilon > 0$ and sequences $\{x_j\}$ and $\{y_j\}$ of points in K such that

$$|x_j - y_j| < \frac{1}{j}$$

but $|f(x_j) - f(y_j)| \geq \varepsilon$. Then by Corollary 13.3.9 on Page 241 which says K is sequentially compact, there is a subsequence $\{x_{n_k}\}$ of $\{x_j\}$ which converges to a point $x \in K$. Then since $|x_{n_k} - y_{n_k}| < \frac{1}{k}$, it follows that $\{y_{n_k}\}$ also converges to x . Therefore,

$$\varepsilon \leq \lim_{k \rightarrow \infty} |f(x_{n_k}) - f(y_{n_k})| = |f(x) - f(x)| = 0,$$

a contradiction. Therefore, f is uniformly continuous as claimed. ■

13.6 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 convergence most encountered in complex analysis. Pointwise convergence is more often encountered in real analysis and necessitates much more difficult theorems.

Definition 13.6.1 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.

Example 13.6.2 *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 13.6.3 *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. ■

13.7 Root Test

The root test has to do with when a series of real or complex numbers converges. I am assuming the reader has been exposed to infinite series. However, this that I am about to explain is a little more general than what is usually seen in calculus. If you have a sequence of real numbers $\{a_k\}_{k=1}^\infty$, if

$$A_n \equiv \sup_{k \geq n} a_k$$

then the sequence $\{A_n\}$ is decreasing. In the above, $\sup_{k \geq n} a_k$ means the least upper bound of all a_k for $k \geq n$ or if there is no upper bound, A_n is simply said to equal ∞ . This is just a formality to make it easy to give an easy discussion. Then, since $\{A_n\}$ is a decreasing sequence, there are two cases. One is that it is bounded below and the other case is that it isn't. In the first case, the sequence must converge to the greatest lower bound of the A_n and in the second case, we say that the sequence converges to $-\infty$. Then

$$\limsup_{n \rightarrow \infty} a_n \equiv \lim_{n \rightarrow \infty} \left(\sup_{k \geq n} a_k \right)$$

Thus, if $\limsup_{n \rightarrow \infty} a_n < r$, it follows that for all n large enough every $a_k < r$. If $\limsup_{n \rightarrow \infty} a_k > r$, it means there are infinitely many k such that $a_k > r$.

Theorem 13.7.1 Let $\mathbf{a}_k \in \mathbb{F}^p$, \mathbb{F} is either \mathbb{R} or \mathbb{C} and consider $\sum_{k=1}^{\infty} \mathbf{a}_k$. Then this series converges absolutely if

$$\limsup_{k \rightarrow \infty} |\mathbf{a}_k|^{1/k} = r < 1.$$

The series diverges spectacularly if $\limsup_{k \rightarrow \infty} |\mathbf{a}_k|^{1/k} > 1$ and if

$$\limsup_{k \rightarrow \infty} |\mathbf{a}_k|^{1/k} = 1,$$

the test fails.

Proof: Suppose first that $\limsup_{k \rightarrow \infty} |\mathbf{a}_k|^{1/k} = r < 1$. Then letting $R \in (r, 1)$, it follows from the definition of \limsup that for all k large enough,

$$|\mathbf{a}_k|^{1/k} \leq R$$

Hence there exists N such that if $k \geq N$, then $|\mathbf{a}_k| \leq R^k$. Let $M_k = |\mathbf{a}_k|$ for $k < N$ and let $M_k = R^k$ for $k \geq N$. Then

$$\sum_{k=1}^{\infty} M_k \leq \sum_{k=1}^{N-1} |\mathbf{a}_k| + \frac{R^N}{1-R} < \infty$$

and so, by the Weierstrass M test applied to the series of constants, the series converges and also converges absolutely. If

$$\limsup_{k \rightarrow \infty} |\mathbf{a}_k|^{1/k} = r > 1,$$

then letting $r > R > 1$, it follows that for infinitely many k ,

$$|\mathbf{a}_k| > R^k$$

and so there is a subsequence which is unbounded. In particular, the series cannot converge and in fact diverges spectacularly. In case that the $\limsup = 1$, you can consider $\sum_{n=1}^{\infty} \frac{1}{n}$ which diverges by calculus and $\sum_{n=1}^{\infty} \frac{1}{n^2}$ which converges, also from calculus. However, the \limsup equals 1 for both of these. ■

This is a major theorem because the \limsup always exists. As an important application, here is a corollary.

Corollary 13.7.2 If $\sum_k \mathbf{a}_k$ converges, then $\limsup_{k \rightarrow \infty} |\mathbf{a}_k|^{1/k} \leq 1$.

If the sequence has values in X a complete normed linear space, there is no change in the conclusion or proof of the above theorem. You just replace $|\cdot|$ with $\|\cdot\|$ the symbol for the norm.

13.8 Convergence of Sums

One can consider convergence of infinite series the same way as done in calculus.

Definition 13.8.1 The symbol $\sum_{k=1}^{\infty} f_k(x)$ means $\lim_{n \rightarrow \infty} \sum_{k=1}^n f_k(x)$ provided this limit exists. This is called pointwise convergence of the infinite sum. Thus the infinite sum means the limit of the sequence of partial sums. The infinite sum is said to converge uniformly if the sequence of partial sums converges uniformly.

Note how this theorem includes the case of $\sum_{k=1}^{\infty} a_k$ as a special case. Here the a_k don't depend on x .

The following theorem is very useful. It tells how to recognize that an infinite sum is converging or converging uniformly. First is a little lemma which reviews standard calculus.

Lemma 13.8.2 Suppose $M_k \geq 0$ and $\sum_{k=1}^{\infty} M_k$ converges. Then

$$\lim_{m \rightarrow \infty} \sum_{k=m}^{\infty} M_k = 0$$

Proof: By assumption, there is N such that if $m \geq N$, then if $n > m$,

$$\left| \sum_{k=1}^n M_k - \sum_{k=1}^m M_k \right| = \sum_{k=m+1}^n M_k < \varepsilon/2$$

Then letting $n \rightarrow \infty$, one can pass to a limit and conclude that

$$\sum_{k=m+1}^{\infty} M_k < \varepsilon$$

It follows that for $m > N$, $\sum_{k=m}^{\infty} M_k < \varepsilon$. The part about passing to a limit follows from the fact that $n \rightarrow \sum_{k=m+1}^n M_k$ is an increasing sequence which is bounded above by $\sum_{k=1}^{\infty} M_k$. Therefore, it converges by completeness of \mathbb{R} . ■

Theorem 13.8.3 For $x \in S$, if $\sum_{k=1}^{\infty} |f_k(x)| < \infty$, then $\sum_{k=1}^{\infty} f_k(x)$ converges pointwise. If there exists M_k such that $M_k \geq |f_k(x)|$ for all $x \in S$, then $\sum_{k=1}^{\infty} f_k(x)$ converges uniformly.

Proof: Let $m < n$. Then

$$\left| \sum_{k=1}^n f_k(x) - \sum_{k=1}^m f_k(x) \right| \leq \sum_{k=m}^{\infty} |f_k(x)| < \varepsilon/2$$

whenever m is large enough due to the assumption that $\sum_{k=1}^{\infty} |f_k(x)| < \infty$. Thus the partial sums are a Cauchy sequence and so the series converges pointwise.

If $M_k \geq |f_k(x)|$ for all $x \in S$, then for M large enough,

$$\left| \sum_{k=1}^n f_k(x) - \sum_{k=1}^m f_k(x) \right| \leq \sum_{k=m}^{\infty} |f_k(x)| \leq \sum_{k=m}^{\infty} M_k < \varepsilon/2$$

Thus, taking sup

$$\left\| \sum_{k=1}^n f_k(\cdot) - \sum_{k=1}^m f_k(\cdot) \right\| \leq \varepsilon/2 < \varepsilon$$

and so the partial sums are uniformly Cauchy sequence. Hence they converge uniformly to what is defined as $\sum_{k=1}^{\infty} f_k(x)$ for $x \in S$. ■

Some of the following exercises have been essentially done in the above discussion. Try doing them yourself. There are also some new topics.

13.9 Connected Sets

Stated informally, connected sets are those which are in one piece. In order to define what is meant by this, I will first consider what it means for a set to **not** be in one piece. This is called **separated**. Connected sets are defined in terms of **not** being separated. This is why theorems about connected sets sometimes seem a little tricky.

Definition 13.9.1 Let A be a nonempty subset \mathbb{R}^n . Then \bar{A} is defined to be the intersection of all closed sets which contain A . This is called the closure of A . Note the whole space, \mathbb{R}^n is one such closed set which contains A .

Lemma 13.9.2 Let A be a nonempty set in \mathbb{R}^n . Then \bar{A} is a closed set and

$$\bar{A} = A \cup A'$$

where A' denotes the set of limit points of A .

Proof: First of all, denote by \mathcal{C} the set of closed sets which contain A . Then

$$\bar{A} = \cap \mathcal{C}$$

and this will be closed if its complement is open. However,

$$\bar{A}^C = \cup \{H^C : H \in \mathcal{C}\}.$$

Each H^C is open and so the union of all these open sets must also be open. This is because if x is in this union, then it is in at least one of them. Hence it is an interior point of that one. But this implies it is an interior point of the union of them all which is an even larger set. Thus \bar{A} is closed.

The interesting part is the next claim. First note that from the definition, $A \subseteq \bar{A}$ so if $x \in A$, then $x \in \bar{A}$. Now consider $y \in A'$ but $y \notin A$. If $y \notin \bar{A}$, a closed set, then there exists $B(y, r) \subseteq \bar{A}^C$. Thus y cannot be a limit point of A , a contradiction. Therefore,

$$A \cup A' \subseteq \bar{A}$$

Next suppose $x \in \bar{A}$ and suppose $x \notin A$. Then if $B(x, r)$ contains no points of A different than x , since x itself is not in A , it would follow that $B(x, r) \cap A = \emptyset$ and so recalling that open balls are open, $B(x, r)^C$ is a closed set containing A so from the definition, it also contains \bar{A} which is contrary to the assertion that $x \in \bar{A}$. Hence if $x \notin A$, then $x \in A'$ and so

$$A \cup A' \supseteq \bar{A} \quad \blacksquare$$

Now is a definition about what it means to not be connected. This is called separated.

Definition 13.9.3 A set, S in \mathbb{R}^n , is separated if there exist sets A, B such that

$$S = A \cup B, A, B \neq \emptyset, \text{ and } \bar{A} \cap B = \bar{B} \cap A = \emptyset.$$

In this case, the sets A and B are said to separate S . A set is connected if it is not separated. Remember \bar{A} denotes the closure of the set A .

Note that the concept of connected sets is defined in terms of what it is not. This makes it somewhat difficult to understand. One of the most important theorems about connected sets is the following.

Theorem 13.9.4 *Suppose \mathcal{U} is a set of connected sets and that there exists a point p which is in all of these connected sets. Then $K \equiv \cup \mathcal{U}$ is connected.*

Proof: Suppose

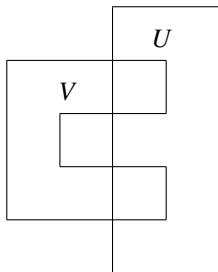
$$K = A \cup B$$

where $\bar{A} \cap B = \bar{B} \cap A = \emptyset, A \neq \emptyset, B \neq \emptyset$. Let $U \in \mathcal{U}$. Then

$$U = (U \cap A) \cup (U \cap B)$$

and this would separate U if both sets in the union are nonempty since the limit points of $U \cap B$ are contained in the limit points of B . It follows that every set of \mathcal{U} is contained in one of A or B . Suppose then that some $U \subseteq A$. Then all $U \in \mathcal{U}$ must be contained in A because if one is contained in B , this would violate the assumption that they all have a point p in common. Thus K is connected after all because this requires $B = \emptyset$. Alternatively, p is in one of these sets. Say $p \in A$. Then by the above argument every U must be in A because if not, the above would be a separation of U . Thus $B = \emptyset$. ■

The intersection of connected sets is not necessarily connected as is shown by the following picture.



Theorem 13.9.5 *Let $f : X \rightarrow \mathbb{R}^m$ be continuous where X is connected. Then $f(X)$ is also connected.*

Proof: To do this you show $f(X)$ is not separated. Suppose to the contrary that $f(X) = A \cup B$ where A and B separate $f(X)$. Then consider the sets $f^{-1}(A)$ and $f^{-1}(B)$. If $z \in f^{-1}(B)$, then $f(z) \in B$ and so $f(z)$ is not a limit point of A . Therefore, there exists an open set, U containing $f(z)$ such that $U \cap A = \emptyset$. But then, the continuity of f implies that $f^{-1}(U)$ is an open set containing z such that $f^{-1}(U) \cap f^{-1}(A) = \emptyset$. Therefore, $f^{-1}(B)$ contains no limit points of $f^{-1}(A)$. Similar reasoning implies $f^{-1}(A)$ contains no limit points of $f^{-1}(B)$. It follows that X is separated by $f^{-1}(A)$ and $f^{-1}(B)$, contradicting the assumption that X was connected. ■

An arbitrary set can be written as a union of maximal connected sets called connected components. This is the concept of the next definition.

Definition 13.9.6 *Let S be a set and let $p \in S$. Denote by C_p the union of all connected subsets of S which contain p . This is called the connected component determined by p .*

Theorem 13.9.7 *Let C_p be a connected component of a set S . Then C_p is a connected set and if $C_p \cap C_q \neq \emptyset$, then $C_p = C_q$.*

Proof: Let \mathcal{C} denote the connected subsets of S which contain p . By Theorem 13.9.4, $\bigcup \mathcal{C} = C_p$ is connected. If $x \in C_p \cap C_q$, then from Theorem 13.9.4, $C_p \supseteq C_p \cup C_q$ and so $C_p \supseteq C_q$. The inclusion goes the other way by the same reason. ■

This shows the connected components of a set are equivalence classes and partition the set.

A set, I is an interval in \mathbb{R} if and only if whenever $x, y \in I$ then $(x, y) \subseteq I$. The following theorem is about the connected sets in \mathbb{R} .

Theorem 13.9.8 *A set C in \mathbb{R} is connected if and only if C is an interval.*

Proof: Let C be connected. If C consists of a single point, p , there is nothing to prove. The interval is just $[p, p]$. Suppose $p < q$ and $p, q \in C$. You need to show $(p, q) \subseteq C$. If

$$x \in (p, q) \setminus C$$

let $C \cap (-\infty, x) \equiv A$, and $C \cap (x, \infty) \equiv B$. Then $C = A \cup B$ and the sets A and B separate C contrary to the assumption that C is connected.

Conversely, let I be an interval. Suppose I is separated by A and B . Pick $x \in A$ and $y \in B$. Suppose without loss of generality that $x < y$. Now define the set,

$$S \equiv \{t \in [x, y] : [x, t] \subseteq A\}$$

and let l be the least upper bound of S . Then $l \in \bar{A}$ so $l \notin B$ which implies $l \in A$. But if $l \notin \bar{B}$, then for some $\delta > 0$,

$$(l, l + \delta) \cap B = \emptyset$$

contradicting the definition of l as an upper bound for S . Therefore, $l \in \bar{B}$ which implies $l \notin A$ after all, a contradiction. It follows I must be connected. ■

This yields a generalization of the intermediate value theorem from one variable calculus.

Corollary 13.9.9 *Let E be a connected set in \mathbb{R}^n and suppose $f : E \rightarrow \mathbb{R}$ and that $y \in (f(e_1), f(e_2))$ where $e_i \in E$. Then there exists $e \in E$ such that $f(e) = y$.*

Proof: From Theorem 13.9.5, $f(E)$ is a connected subset of \mathbb{R} . By Theorem 13.9.8 $f(E)$ must be an interval. In particular, it must contain y . This proves the corollary. ■

The following theorem is a very useful description of the open sets in \mathbb{R} .

Theorem 13.9.10 *Let U be an open set in \mathbb{R} . Then there exist countably many disjoint open sets $\{(a_i, b_i)\}_{i=1}^{\infty}$ such that $U = \bigcup_{i=1}^{\infty} (a_i, b_i)$.*

Proof: Let $p \in U$ and let $z \in C_p$, the connected component determined by p . Since U is open, there exists, $\delta > 0$ such that $(z - \delta, z + \delta) \subseteq U$. It follows from Theorem 13.9.4 that

$$(z - \delta, z + \delta) \subseteq C_p.$$

This shows C_p is open. By Theorem 13.9.8, this shows C_p is an open interval, (a, b) where $a, b \in [-\infty, \infty]$. There are therefore at most countably many of these connected components because each must contain a rational number and the rational numbers are countable. Denote by $\{(a_i, b_i)\}_{i=1}^{\infty}$ the set of these connected components. ■

Definition 13.9.11 A set E in \mathbb{R}^n is *arcwise connected* if for any two points, $\mathbf{p}, \mathbf{q} \in E$, there exists a closed interval, $[a, b]$ and a continuous function, $\gamma: [a, b] \rightarrow E$ such that $\gamma(a) = \mathbf{p}$ and $\gamma(b) = \mathbf{q}$.

An example of an arcwise connected space would be any subset of \mathbb{R}^n which is the continuous image of an interval. Arcwise connected is not the same as connected. A well known example is the following.

$$\left\{ \left(x, \sin \frac{1}{x} \right) : x \in (0, 1] \right\} \cup \{ (0, y) : y \in [-1, 1] \} \quad (13.2)$$

You can verify that this set of points in \mathbb{R}^2 is not arcwise connected but is connected.

Lemma 13.9.12 In \mathbb{R}^n , $B(\mathbf{z}, r)$ is arcwise connected.

Proof: This is easy from the convexity of the set. If $\mathbf{x}, \mathbf{y} \in B(\mathbf{z}, r)$, then let $\gamma(t) = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$ for $t \in [0, 1]$.

$$\begin{aligned} \|\mathbf{x} + t(\mathbf{y} - \mathbf{x}) - \mathbf{z}\| &= \|(1-t)(\mathbf{x} - \mathbf{z}) + t(\mathbf{y} - \mathbf{z})\| \\ &\leq (1-t)\|\mathbf{x} - \mathbf{z}\| + t\|\mathbf{y} - \mathbf{z}\| \\ &< (1-t)r + tr = r \end{aligned}$$

showing $\gamma(t)$ stays in $B(\mathbf{z}, r)$. ■

Proposition 13.9.13 If $X \neq \emptyset$ is arcwise connected, then it is connected.

Proof: Let $p \in X$. Then by assumption, for any $x \in X$, there is an arc joining p and x . This arc is connected because it is the continuous image of an interval which is connected. Since x is arbitrary, every x is in a connected subset of X which contains p . Hence $C_p = X$ and so X is connected. ■

Theorem 13.9.14 Let U be an open subset of a \mathbb{R}^n . Then U is arcwise connected if and only if U is connected. Also the connected components of an open set are open sets.

Proof: By Proposition 13.9.13 it is only necessary to verify that if U is connected and open in the context of this theorem, then U is arcwise connected. Pick $\mathbf{p} \in U$. Say $\mathbf{x} \in U$ satisfies \mathcal{P} if there exists a continuous function, $\gamma: [a, b] \rightarrow U$ such that $\gamma(a) = \mathbf{p}$ and $\gamma(b) = \mathbf{x}$.

$$A \equiv \{ \mathbf{x} \in U \text{ such that } \mathbf{x} \text{ satisfies } \mathcal{P} \}$$

If $\mathbf{x} \in A$, then Lemma 13.9.12 implies $B(\mathbf{x}, r) \subseteq U$ is arcwise connected for small enough r . Thus letting $\mathbf{y} \in B(\mathbf{x}, r)$, there exist intervals, $[a, b]$ and $[c, d]$ and continuous functions having values in U , γ, η such that $\gamma(a) = \mathbf{p}, \gamma(b) = \mathbf{x}, \eta(c) = \mathbf{x}$, and $\eta(d) = \mathbf{y}$. Then let $\gamma_1: [a, b+d-c] \rightarrow U$ be defined as

$$\gamma_1(t) \equiv \begin{cases} \gamma(t) & \text{if } t \in [a, b] \\ \eta(t+c-b) & \text{if } t \in [b, b+d-c] \end{cases}$$

Then it is clear that γ_1 is a continuous function mapping \mathbf{p} to \mathbf{y} and showing that $B(\mathbf{x}, r) \subseteq A$. Therefore, A is open. $A \neq \emptyset$ because since U is open there is an open set, $B(\mathbf{p}, \delta)$ containing \mathbf{p} which is contained in U and is arcwise connected.

Now consider $B \equiv U \setminus A$. I claim this is also open. If B is not open, there exists a point $z \in B$ such that every open set containing z is not contained in B . Therefore, letting $B(z, \delta)$ be such that $z \in B(z, \delta) \subseteq U$, there exist points of A contained in $B(z, \delta)$. But then, a repeat of the above argument shows $z \in A$ also. Hence B is open and so if $B \neq \emptyset$, then $U = B \cup A$ and so U is separated by the two sets B and A contradicting the assumption that U is connected.

It remains to verify the connected components are open. Let $z \in C_p$ where C_p is the connected component determined by p . Then picking $B(z, \delta) \subseteq U$, $C_p \cup B(z, \delta)$ is connected and contained in U and so it must also be contained in C_p . Thus z is an interior point of C_p . ■

As an application, consider the following corollary.

Corollary 13.9.15 *Let $f : \Omega \rightarrow \mathbb{Z}$ be continuous where Ω is a connected open set in \mathbb{R}^n . Then f must be a constant.*

Proof: Suppose not. Then it achieves two different values, k and $l \neq k$. Then $\Omega = f^{-1}(l) \cup f^{-1}(\{m \in \mathbb{Z} : m \neq l\})$ and these are disjoint nonempty open sets which separate Ω . To see they are open, note

$$f^{-1}(\{m \in \mathbb{Z} : m \neq l\}) = f^{-1}\left(\bigcup_{m \neq l} \left(m - \frac{1}{6}, m + \frac{1}{6}\right)\right)$$

which is the inverse image of an open set while $f^{-1}(l) = f^{-1}\left((l - \frac{1}{6}, l + \frac{1}{6})\right)$ also an open set. ■

13.10 Exercises

1. Suppose $\{x_n\}$ is a sequence contained in a closed set C such that $\lim_{n \rightarrow \infty} x_n = x$. Show that $x \in C$. **Hint:** Recall that a set is closed if and only if the complement of the set is open. That is if and only if $\mathbb{R}^n \setminus C$ is open.
2. Show using Problem 1 and Theorem 13.3.8 that every closed and bounded set is sequentially compact. **Hint:** If C is such a set, then $C \subseteq I_0 \equiv \prod_{i=1}^n [a_i, b_i]$. Now if $\{x_n\}$ is a sequence in C , it must also be a sequence in I_0 . Apply Problem 1 and Theorem 13.3.8.
3. Prove the extreme value theorem, a continuous function achieves its maximum and minimum on any closed and bounded set C , using the result of Problem 2. **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 using Problem 2. Do the same for the minimum.
4. 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 $|f(x) - f(y)| < \varepsilon$. 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 $|f(x_k) - f(y_k)| \geq \varepsilon$. Now use Problem 2 to obtain a convergent subsequence.

5. From Problem 2 every closed and bounded set is sequentially compact. Are these the only sets which are sequentially compact? Explain.
6. A set whose elements are open sets \mathcal{C} is called an **open cover** of H if $\cup \mathcal{C} \supseteq H$. In other words, \mathcal{C} is an open cover of H if every point of H is in at least one set of \mathcal{C} . Show that if \mathcal{C} is an open cover of a closed and bounded set H then there exists $\delta > 0$ such that whenever $x \in H$, $B(x, \delta)$ is contained in some set of \mathcal{C} . This number δ is called a **Lebesgue number**. **Hint:** If there is no Lebesgue number for H , let $H \subseteq I = \prod_{i=1}^n [a_i, b_i]$. Use the process of chopping the intervals in half to get a sequence of nested intervals, I_k contained in I where $\text{diam}(I_k) \leq 2^{-k} \text{diam}(I)$ and there is no Lebesgue number for the open cover on $H_k \equiv H \cap I_k$. Now use the nested interval theorem to get c in all these H_k . For some $r > 0$ it follows $B(c, r)$ is contained in some open set of \mathcal{C} . But for large k , it must be that $H_k \subseteq B(c, r)$ which contradicts the construction. You fill in the details.
7. A set is **compact** if for every open cover of the set, there exists a finite subset of the open cover which also covers the set. Show every closed and bounded set in \mathbb{R}^p is compact. 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.
8. 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.

9. 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 8. This is a special case of a major theorem called Urysohn's lemma.

Chapter 14

Vector Valued Functions Of One Variable

14.1 Limits Of A Vector Valued Function Of One Real Variable

As in the case of a scalar valued function of one variable, the derivative is defined as

$$\lim_{h \rightarrow 0} \frac{\mathbf{f}(t_0 + h) - \mathbf{f}(t_0)}{h}.$$

Thus the derivative of a function of one variable involves a limit. The following is the definition of what is meant by a limit. The new topic is the case of one sided limits although there is really nothing essentially new from what was done earlier. Here is the definition.

Definition 14.1.1 *In the case where $D(\mathbf{f})$ is only assumed to satisfy $D(\mathbf{f}) \supseteq (t, t + r)$,*

$$\lim_{s \rightarrow t+} \mathbf{f}(s) = \mathbf{L}$$

if and only if for all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < s - t < \delta,$$

then

$$|\mathbf{f}(s) - \mathbf{L}| < \varepsilon.$$

In the case where $D(\mathbf{f})$ is only assumed to satisfy $D(\mathbf{f}) \supseteq (t - r, t)$,

$$\lim_{s \rightarrow t-} \mathbf{f}(s) = \mathbf{L}$$

if and only if for all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < t - s < \delta,$$

then

$$|\mathbf{f}(s) - \mathbf{L}| < \varepsilon.$$

One can also consider limits as a variable “approaches” infinity. Of course nothing is “close” to infinity and so this requires a slightly different definition.

$$\lim_{t \rightarrow \infty} \mathbf{f}(t) = \mathbf{L}$$

if for every $\varepsilon > 0$ there exists l such that whenever $t > l$,

$$|\mathbf{f}(t) - \mathbf{L}| < \varepsilon \quad (14.1)$$

and

$$\lim_{t \rightarrow -\infty} \mathbf{f}(t) = \mathbf{L}$$

if for every $\varepsilon > 0$ there exists l such that whenever $t < l$, (14.1) holds.

Note that in all of this the definitions are identical to the case of scalar valued functions. The only difference is that here $|\cdot|$ refers to the norm or length in \mathbb{R}^p where maybe $p > 1$.

Example 14.1.2 Let $\mathbf{f}(t) = (\cos t, \sin t, t^2 + 1, \ln(t))$. Find $\lim_{t \rightarrow \pi/2} \mathbf{f}(t)$.

Use Theorem 12.5.5 on Page 223 and the continuity of the functions to write this limit equals

$$\begin{aligned} & \left(\lim_{t \rightarrow \pi/2} \cos t, \lim_{t \rightarrow \pi/2} \sin t, \lim_{t \rightarrow \pi/2} (t^2 + 1), \lim_{t \rightarrow \pi/2} \ln(t) \right) \\ &= \left(0, 1, \ln\left(\frac{\pi^2}{4} + 1\right), \ln\left(\frac{\pi}{2}\right) \right). \end{aligned}$$

Example 14.1.3 Let $\mathbf{f}(t) = \left(\frac{\sin t}{t}, t^2, t + 1\right)$. Find $\lim_{t \rightarrow 0} \mathbf{f}(t)$.

Recall that $\lim_{t \rightarrow 0} \frac{\sin t}{t} = 1$. Then from Theorem 12.5.5 on Page 223, $\lim_{t \rightarrow 0} \mathbf{f}(t) = (1, 0, 1)$.

14.2 The Derivative And Integral

The following definition is on the derivative and integral of a vector valued function of one variable.

Definition 14.2.1 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 $\mathbf{f} \in R([a, b])$.

Here is a simple proposition which is useful to have.

Proposition 14.2.2 *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.

$$\begin{aligned} \left| \int_a^b \mathbf{f}(t) dt \right| &= \left| \left(\int_a^b f_1(t) dt, \dots, \int_a^b f_n(t) dt \right) \right| \\ &= \left(\sum_{i=1}^n \left| \int_a^b f_i(t) dt \right|^2 \right)^{1/2} \leq \left(\sum_{i=1}^n \left(\int_a^b |f_i(t)| dt \right)^2 \right)^{1/2} \\ &\leq \left(n \max_i \left(\int_a^b |f_i(t)| dt \right)^2 \right)^{1/2} = \sqrt{n} \max_i \left(\int_a^b |f_i(t)| dt \right) \\ &\leq \sqrt{n} \int_a^b |\mathbf{f}(t)| dt \blacksquare \end{aligned}$$

As in the case of a scalar valued function differentiability implies continuity but not the other way around.

Theorem 14.2.3 *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 14.2.4 *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 14.2.5 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,

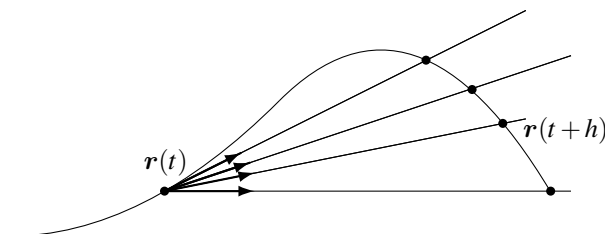
$$\lim_{h \rightarrow 0} \frac{\mathbf{f}(x+h) - \mathbf{f}(x)}{h} = \lim_{h \rightarrow 0} \mathbf{0} = \mathbf{0}$$

Example 14.2.6 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 $\mathbf{f}'(t) = (a, b)$.

14.2.1 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} .



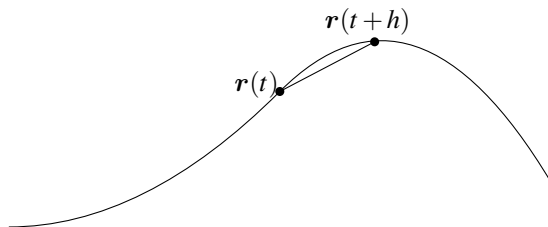
In this picture there are unit vectors in the direction of the vector from $\mathbf{r}(t)$ to $\mathbf{r}(t+h)$. You can see that it is reasonable to suppose these unit vectors, if they converge, converge to a unit vector \mathbf{T} which is tangent to the curve at the point $\mathbf{r}(t)$. Now each 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

$$\frac{\mathbf{r}(t_0 + h) - \mathbf{r}(t_0)}{h} \quad (14.2)$$

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, (14.2) converges to $\mathbf{v} \equiv (v_1, \dots, v_q)$ where $v_k = r'_k(t)$. This is because of Theorem 12.5.5 on Page 223, which says that the term in (14.2) gets close to a vector \mathbf{v} if and only if all the coordinate functions of the term in (14.2) 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 14.2.7 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 14.2.8 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 $\mathbf{r}'(1) = (\cos 1, 2, 1)$.

14.2.2 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 14.2.9 Let $a, b \in \mathbb{R}$ and suppose $\mathbf{f}'(t)$ and $\mathbf{g}'(t)$ exist. Then the following formulas are obtained.

$$(a\mathbf{f} + b\mathbf{g})'(t) = a\mathbf{f}'(t) + b\mathbf{g}'(t). \quad (14.3)$$

$$(\mathbf{f} \cdot \mathbf{g})'(t) = \mathbf{f}'(t) \cdot \mathbf{g}(t) + \mathbf{f}(t) \cdot \mathbf{g}'(t) \quad (14.4)$$

If \mathbf{f}, \mathbf{g} have values in \mathbb{R}^3 , then

$$(\mathbf{f} \times \mathbf{g})'(t) = \mathbf{f}(t) \times \mathbf{g}'(t) + \mathbf{f}'(t) \times \mathbf{g}(t) \quad (14.5)$$

The formulas, (14.4), and (14.5) are referred to as the product rule.

Proof: The first formula is left for you to prove. Consider the second, (14.4).

$$\begin{aligned} & \lim_{h \rightarrow 0} \frac{\mathbf{f} \cdot \mathbf{g}(t+h) - \mathbf{f} \cdot \mathbf{g}(t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{\mathbf{f}(t+h) \cdot \mathbf{g}(t+h) - \mathbf{f}(t+h) \cdot \mathbf{g}(t)}{h} + \frac{\mathbf{f}(t+h) \cdot \mathbf{g}(t) - \mathbf{f}(t) \cdot \mathbf{g}(t)}{h} \\ &= \lim_{h \rightarrow 0} \left(\mathbf{f}(t+h) \cdot \frac{(\mathbf{g}(t+h) - \mathbf{g}(t))}{h} + \frac{(\mathbf{f}(t+h) - \mathbf{f}(t))}{h} \cdot \mathbf{g}(t) \right) \\ &= \lim_{h \rightarrow 0} \sum_{k=1}^n f_k(t+h) \frac{(g_k(t+h) - g_k(t))}{h} + \sum_{k=1}^n \frac{(f_k(t+h) - f_k(t))}{h} g_k(t) \\ &= \sum_{k=1}^n f_k(t) g'_k(t) + \sum_{k=1}^n f'_k(t) g_k(t) = \mathbf{f}'(t) \cdot \mathbf{g}(t) + \mathbf{f}(t) \cdot \mathbf{g}'(t). \end{aligned}$$

Formula (14.5) is left as an exercise which follows from the product rule and the definition of the cross product. ■

Example 14.2.10 Let $\mathbf{r}(t) = (t^2, \sin t, \cos t)$ and let $\mathbf{p}(t) = (t, \ln(t+1), 2t)$. Simplify the expression $(\mathbf{r}(t) \times \mathbf{p}(t))'$.

From (14.5) this equals $(2t, \cos t, -\sin t) \times (t, \ln(t+1), 2t) + (t^2, \sin t, \cos t) \times (1, \frac{1}{t+1}, 2)$.

Example 14.2.11 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 14.2.12 An object has position $\mathbf{r}(t) = (t^3, \frac{t}{1+t}, \sqrt{t^2+2})$ 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 14.2.13 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 14.2.14 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

$$\frac{(\mathbf{r}(t) - P_0) \cdot (-\mathbf{r}'(t))}{|\mathbf{r}(t) - P_0|} = \frac{(\mathbf{r}(t) - P_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - P_1|} \quad (14.6)$$

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 (14.6),

$$\frac{d}{dt}(|\mathbf{r}(t) - \mathbf{P}_1|) + \frac{d}{dt}(|\mathbf{r}(t) - \mathbf{P}_0|) = 0$$

showing that $|\mathbf{r}(t) - \mathbf{P}_1| + |\mathbf{r}(t) - \mathbf{P}_0| = C$ for some constant C . This implies the curve of intersection of the plane with the room is an ellipse having \mathbf{P}_0 and \mathbf{P}_1 as the foci.

14.2.3 Leibniz's Notation

Leibniz's notation also generalizes routinely. For example, $\frac{dy}{dt} = \mathbf{y}'(t)$ with other similar notations holding.

14.3 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 (14.5) from the component description of the cross product.
13. Prove (14.5) from the formula $(\mathbf{f} \times \mathbf{g})_i = \varepsilon_{ijk} f_j g_k$.
14. Prove (14.5) 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 $\mathbf{r}'(t) = \mathbf{0}$ for all $t \in (a, b)$, show that there exists a constant vector \mathbf{c} such that $\mathbf{r}(t) = \mathbf{c}$ for all $t \in (a, b)$.
 18. If $\mathbf{F}'(t) = \mathbf{f}(t)$ for all $t \in (a, b)$ and \mathbf{F} is continuous on $[a, b]$, show that $\int_a^b \mathbf{f}(t) dt = \mathbf{F}(b) - \mathbf{F}(a)$.
 19. Verify that if $\boldsymbol{\Omega} \times \mathbf{u} = \mathbf{0}$ for all \mathbf{u} , then $\boldsymbol{\Omega} = \mathbf{0}$.

14.4 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}^n . This is done by defining things in such a way that the more general concept reduces to the earlier notion. First it is necessary to consider what is meant by arc length.

14.4.1 Arc Length And Orientations

The application of the integral considered here is the concept of the **length of a curve**.

Definition 14.4.1 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,

$$\lim_{h \rightarrow 0+} \frac{x_i(a+h) - x_i(a)}{h},$$

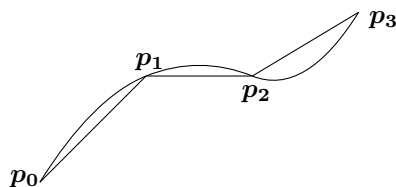
and $x'_i(b)$ defined similarly as the derivative from the left.

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) .

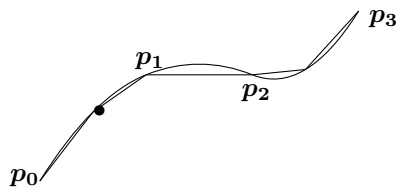
4. $|\mathbf{p}'(t)| \equiv \left(\sum_{i=1}^n |x'_i(t)|^2 \right)^{1/2} \neq 0$ for all $t \in [a, b]$.

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 $[a, b]$, $a = t_0 < \dots < t_n = b$ 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 approximated by

$$\sum_{k=1}^n |\mathbf{p}(t_k) - \mathbf{p}(t_{k-1})|.$$

Since the functions in the parametrization are differentiable, it is reasonable to expect this to be close to

$$\sum_{k=1}^n |\mathbf{p}'(t_{k-1})| (t_k - t_{k-1})$$

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 14.4.2 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 in the last section of this chapter.

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

$$f_i(t) \equiv t p_i + (1-t) q_i, t \in [0, 1].$$

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 14.6.6 on Page 272 but the proof is mathematically fairly advanced so it is presented later.

Corollary 14.4.3 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 14.4.4 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 $\mathbf{f} \sim \mathbf{g}$.

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 $\mathbf{g}^{-1} \circ \mathbf{f}$ 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 14.4.5 The following hold for \sim .

$$\mathbf{f} \sim \mathbf{f}; \tag{14.7}$$

$$\text{If } \mathbf{f} \sim \mathbf{g} \text{ then } \mathbf{g} \sim \mathbf{f}; \tag{14.8}$$

$$\text{If } \mathbf{f} \sim \mathbf{g} \text{ and } \mathbf{g} \sim \mathbf{h}, \text{ then } \mathbf{f} \sim \mathbf{h}. \tag{14.9}$$

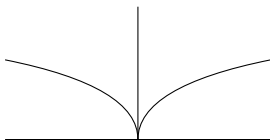
Proof: Formula (14.7) is obvious because $f^{-1} \circ f(t) = t$ so it is clearly an increasing function. If $f \sim g$ then $f^{-1} \circ g$ is increasing. Now $g^{-1} \circ f$ must also be increasing because it is the inverse of $f^{-1} \circ g$. This verifies (14.8). To see (14.9), $f^{-1} \circ h = (f^{-1} \circ g) \circ (g^{-1} \circ h)$ 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 14.4.3, 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 parametrizations, h , which satisfy $h \sim f$ are called the equivalence class determined by f and those $h \sim g$ 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.

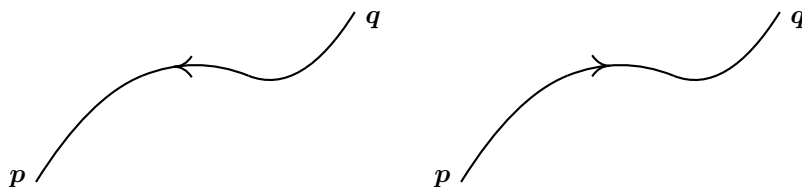
Sometimes people wonder why it is required, in the definition of a smooth curve that $p'(t) \neq 0$. Imagine t is time and $p(t)$ gives the location of a point in space. If $p'(t)$ is allowed to equal zero, the point can stop and change directions abruptly, producing a pointy place in C . Here is an example.

Example 14.4.6 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 14.4.7 A curve C is *piecewise smooth* if there exist points on this curve, denoted by 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 $\cup_{k=1}^n C_{p_{k-1}p_k} = C$. In other words, it is piecewise smooth if it consists of a finite number of smooth curves linked together.

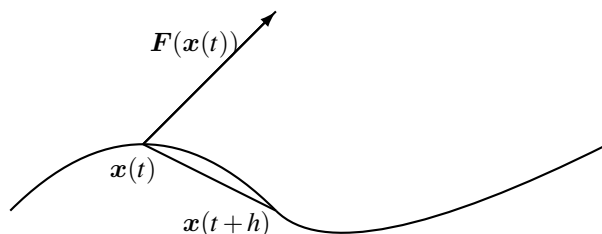
Note that Example 14.4.6 is an example of a piecewise smooth curve although it is not smooth.

14.4.2 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 14.4.8 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 wiggly 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 14.4.9 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 14.4.10 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

$$\begin{aligned} \int_c^d \mathbf{F}(\mathbf{g}(s)) \cdot \mathbf{g}'(s) ds &= \int_a^b \mathbf{F}(\mathbf{g}(\phi(t))) \cdot \mathbf{g}'(\phi(t)) \phi'(t) dt \\ &= \int_a^b \mathbf{F}(\mathbf{f}(t)) \cdot \frac{d}{dt} (\mathbf{g}(\mathbf{g}^{-1} \circ \mathbf{f}(t))) dt = \int_a^b \mathbf{F}(\mathbf{f}(t)) \cdot \mathbf{f}'(t) dt. \quad \blacksquare \end{aligned}$$

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 religion which often abounds in apparent 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 14.4.11 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 14.4.12 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 14.4.13 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. $\mathbf{R}(l) = \mathbf{r}(\phi^{-1}(l)) = \mathbf{r}\left(\phi^{-1}\left(\int_a^b |\mathbf{r}'(\tau)| d\tau\right)\right) = \mathbf{r}(b)$ 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 14.4.14 As to piecewise smooth curves, recall these 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

$$\int_C \mathbf{F} \cdot d\mathbf{R} \equiv \int_{C_{\mathbf{p}_1 \mathbf{p}_2}} \mathbf{F} \cdot d\mathbf{R} + \int_{C_{\mathbf{p}_2 \mathbf{p}_3}} \mathbf{F} \cdot d\mathbf{R} + \dots + \int_{C_{\mathbf{p}_{(n-1)} \mathbf{p}_n}} \mathbf{F} \cdot d\mathbf{R}$$

14.4.3 Another Notation For Line Integrals

Definition 14.4.15 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.

14.5 Exercises

- 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.
- 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.
- Let $\mathbf{r}(t) = (t, \cos(3t), \sin(3t))$ for $t \in [0, 1]$. Find the length of this curve.
- 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 , which is given by the formula $\mathbf{F}(x, y, z) \equiv 2xy\mathbf{i} + (x^2 + 2zy)\mathbf{j} + y^2\mathbf{k}$. 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 .
- In the following, a force field is specified followed by the parametrization of a curve. Find the work.
 - $\mathbf{F} = (x, y, z), \mathbf{r}(t) = (t, t^2, t + 1), t \in [0, 1]$
 - $\mathbf{F} = (x - y, y + z, z), \mathbf{r}(t) = (\cos(t), t, \sin(t)), t \in [0, \pi]$
 - $\mathbf{F} = (x^2, y^2, z + x), \mathbf{r}(t) = (t, 2t, t + t^2), t \in [0, 1]$
 - $\mathbf{F} = (z, y, x), \mathbf{r}(t) = (t^2, 2t, t), t \in [0, 1]$
- 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
 - $\mathbf{F} = (2xy, x^2 + 2y, 1)$
 - $\mathbf{F} = (yz^2, xz^2, 2xyz + 1)$
 - $\mathbf{F} = (\cos x, -\sin y, 1)$
 - $\mathbf{F} = (2x \sin y, x^2 \cos y, 1)$
- *Read ahead about the gradient in Definition 16.3.5 on Page 290. 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. For each f in the above, compute $f(1, 2, 3) - f(0, 0, 0)$ and compare with your solutions to the above line integrals. You should get the same thing from $f(1, 2, 3) - f(0, 0, 0)$. This is not a coincidence and will be fully discussed later. Such vector fields are called **conservative**.

8. 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 $\int_C \mathbf{F} \cdot d\mathbf{R}$.
9. 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.
10. 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 $\mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + t\mathbf{k}$.
11. 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 $\mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + t\mathbf{k}$.
12. 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:**

$$\int_a^b \mathbf{F}(x(t), y(t), z(t)) \cdot (x'(t), y'(t), z'(t)) dt =$$

$$\int_a^b m(x''(t), y''(t), z''(t)) \cdot (x'(t), y'(t), z'(t)) dt,$$

etc.

13. 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 .

14. 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 $\int_C \mathbf{F} \cdot d\mathbf{R}$.
15. 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 $\mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + t\mathbf{k}$. You should get the same answer in this case. This is because the vector field happens to be conservative. (More on this later.)

14.6 Independence Of Parametrization*



Recall that if $\mathbf{p}(t) : t \in [a, b]$ was a parametrization of a smooth curve C , the length of C is defined as $\int_a^b |\mathbf{p}'(t)| dt$. 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.

14.6.1 Hard Calculus

Recall Theorem 13.4.1 about continuity and convergent sequences. It said roughly that a function \mathbf{f} is continuous at \mathbf{x} if and only if whenever $\mathbf{x}_k \rightarrow \mathbf{x}$, then $\mathbf{f}(\mathbf{x}_k) \rightarrow \mathbf{f}(\mathbf{x})$. Also recall the following Lemma from Volume 1, whose proof is summarized below for convenience.

Lemma 14.6.1 *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.*

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 $y_t \equiv ty_1 + (1-t)y_2$. It follows that $x_t < y_t$ for all $t \in [0, 1]$. Now define

$$h(t) \equiv (\phi(y_t) - \phi(x_t))(y_t - x_t).$$

Then $h(0) < 0$, $h(1) > 0$ but by assumption, $h(t) \neq 0$ for any $t \in (0, 1)$, a contradiction.

This property of being either strictly increasing or strictly decreasing on (a, b) carries over to $[a, b]$ by the continuity of ϕ .

It only remains to verify ϕ^{-1} is continuous. If not, there exists $s_n \rightarrow s$ where s_n and s are points of $\phi([a, b])$ but $|\phi^{-1}(s_n) - \phi^{-1}(s)| \geq \varepsilon$. By sequential compactness of $[a, b]$, there is a subsequence, still denoted by n , such that $|\phi^{-1}(s_n) - t_1| \rightarrow 0$. Thus $s_n \rightarrow \phi(t_1)$, so $s = \phi(t_1)$, and $t_1 = \phi^{-1}(s)$, a contradiction. ■

Corollary 14.6.2 *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} continuous at $f(x)$? Since f is decreasing, if $f(x) < f(y)$, then $y \equiv f^{-1}(f(y)) < x \equiv f^{-1}(f(x))$ and so f^{-1} is also decreasing. 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 14.6.3 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 14.6.1 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 14.6.4 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} ,

$$\lim_{f(x) \rightarrow f(x_1)} \frac{f^{-1}(f(x)) - f^{-1}(f(x_1))}{f(x) - f(x_1)} = \lim_{x \rightarrow x_1} \frac{x - x_1}{f(x) - f(x_1)} = \frac{1}{f'(x_1)}. \quad \blacksquare$$

14.6.2 Independence Of Parametrization

Theorem 14.6.5 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 $[c, d]$ which is Riemann integrable¹,

$$\int_c^d f(s) ds = \int_a^b f(\phi(t)) |\phi'(t)| dt$$

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 14.6.1 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 14.6.6 Let $\mathbf{f} : [a, b] \rightarrow C$, $\mathbf{g} : [c, d] \rightarrow C$ be parameterizations of a smooth curve which satisfy conditions (1) - (5). Then $\phi(t) \equiv \mathbf{g}^{-1} \circ \mathbf{f}(t)$ 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 \mathbf{f} and \mathbf{g} satisfy. It only remains to verify continuity on $[a, b]$ because then the final claim follows from Lemma 14.6.1. 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 $[c, d]$, (See Theorem 13.3.8 on Page 240.) 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 $\mathbf{g}^{-1} \circ \mathbf{f}(t_n) \rightarrow s$ while $t_n \rightarrow t$. Therefore, the continuity of \mathbf{f} and \mathbf{g} imply $\mathbf{f}(t_n) \rightarrow \mathbf{f}(t)$ and $\mathbf{f}(t_n) \rightarrow \mathbf{f}(t)$. Thus, $\mathbf{g}(s) = \mathbf{f}(t)$, so $s = \mathbf{g}^{-1} \circ \mathbf{f}(t) = \phi(t)$, a contradiction. Therefore, ϕ is continuous as claimed. ■

Theorem 14.6.7 The length of a smooth curve is not dependent on which parametrization is used.

Proof: Let C be the curve and suppose $\mathbf{f} : [a, b] \rightarrow C$ and $\mathbf{g} : [c, d] \rightarrow C$ both satisfy conditions (1) - (5). Is it true that $\int_a^b |\mathbf{f}'(t)| dt = \int_c^d |\mathbf{g}'(s)| ds$?

Let $\phi(t) \equiv \mathbf{g}^{-1} \circ \mathbf{f}(t)$ 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

$$h_i(g_i(s)) = s.$$

¹Recall that all continuous functions of this sort are Riemann integrable.

This implies that for $s \in I$,

$$h'_i(g_i(s)) = \frac{1}{g'_i(s)}. \quad (14.10)$$

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 $t \in J_1$,

$$\phi'(t) = h'_i(f_i(t)) f'_i(t) = h'_i(g_i(s)) f'_i(t) = \frac{f'_i(t)}{g'_i(\phi(t))} \quad (14.11)$$

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, $\mathbf{f}'(t) = \mathbf{g}'(\phi(t)) \phi'(t)$ and so, by Theorem 14.6.5,

$$\int_{\phi(a+\delta)}^{\phi(b-\delta)} |\mathbf{g}'(s)| ds = \int_{a+\delta}^{b-\delta} |\mathbf{g}'(\phi(t))| |\phi'(t)| dt = \int_{a+\delta}^{b-\delta} |\mathbf{f}'(t)| dt.$$

Now using the continuity of ϕ , \mathbf{g}' , and \mathbf{f}' on $[a, b]$ and letting $\delta \rightarrow 0+$ in the above, yields

$$\int_c^d |\mathbf{g}'(s)| ds = \int_a^b |\mathbf{f}'(t)| dt. \quad \blacksquare$$

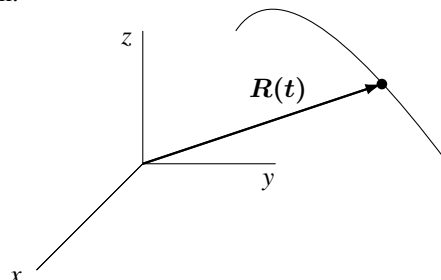
Chapter 15

Motion On A Space Curve

15.1 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 15.1.1 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 $|\mathbf{T}'| \equiv \kappa(t) |\mathbf{v}(t)|$, it follows

$$\mathbf{T}'(t) = |\mathbf{T}'(t)| \mathbf{N}(t) = \kappa(t) |\mathbf{v}(t)| \mathbf{N}(t). \blacksquare$$

Definition 15.1.2 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 N in the case where $T' \neq 0$ 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 15.1.3 For $R(t)$ the position vector of a space curve, the acceleration is given by the formula

$$a = \frac{d|v|}{dt}T + \kappa|v|^2N \equiv a_T T + a_N N. \quad (15.1)$$

Furthermore, $a_T^2 + a_N^2 = |a|^2$.

Proof:

$$a = \frac{dv}{dt} = \frac{d}{dt}(R') = \frac{d}{dt}(|v|T) = \frac{d|v|}{dt}T + |v|T' = \frac{d|v|}{dt}T + |v|^2\kappa N.$$

This proves the first part.

For the second part,

$$\begin{aligned} |a|^2 &= (a_T T + a_N N) \cdot (a_T T + a_N N) \\ &= a_T^2 T \cdot T + 2a_N a_T T \cdot N + a_N^2 N \cdot N = a_T^2 + a_N^2 \end{aligned}$$

because $T \cdot N = 0$. ■

From 15.1 and the geometric properties of the cross product,

$$a \times v = \kappa|v|^2 N \times v$$

Hence, using the geometric description of the cross product again using that the angle between N and T is 90° ,

$$|a \times v| = \kappa|v|^2|v|, \quad \kappa = \frac{|a \times v|}{|v|^3} = \frac{|v \times a|}{|v|^3} \quad (15.2)$$

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 $R(t)$ and $R(\tau)$, then from the formula above for the curvature,

$$\kappa(t) = \frac{|T'(t)|}{|v(t)|} = \frac{\left| \frac{dT}{d\tau} \frac{d\tau}{dt} \right|}{\left| \frac{dR}{d\tau} \frac{d\tau}{dt} \right|} = \frac{\left| \frac{dT}{d\tau} \right|}{\left| \frac{dR}{d\tau} \right|} \equiv \kappa(\tau).$$

From this, it is possible to give an important formula from physics. Suppose an object orbits a point at constant speed v . In the above notation, $|v| = v$. 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

$$R(t) = (r \cos t, r \sin t)$$

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.

Thus using 15.2 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 15.1 that

$$\mathbf{a} = \frac{1}{r} |\mathbf{v}|^2 \mathbf{N} = \frac{1}{r} v^2 \mathbf{N}$$

Example 15.1.4 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.

First of all, $\mathbf{v}(t) = (-\sin t, 1, 2t)$ and so the speed is given by

$$|\mathbf{v}| = \sqrt{\sin^2(t) + 1 + 4t^2}.$$

Therefore,

$$a_T = \frac{d}{dt} \left(\sqrt{\sin^2(t) + 1 + 4t^2} \right) = \frac{\sin(t) \cos(t) + 4t}{\sqrt{(2 + 4t^2 - \cos^2 t)}}.$$

It remains to find a_N . To do this, you can find the curvature first if you like.

$$\mathbf{a}(t) = \mathbf{R}''(t) = (-\cos t, 0, 2).$$

Then

$$\kappa = \frac{|(-\cos t, 0, 2) \times (-\sin t, 1, 2t)|}{\left(\sqrt{\sin^2(t) + 1 + 4t^2} \right)^3} = \frac{\sqrt{4 + (-2 \sin(t) + 2(\cos(t))t)^2 + \cos^2(t)}}{\left(\sqrt{\sin^2(t) + 1 + 4t^2} \right)^3}$$

Then $a_N = \kappa |\mathbf{v}|^2$

$$\begin{aligned} &= \frac{\sqrt{4 + (-2 \sin(t) + 2(\cos(t))t)^2 + \cos^2(t)}}{\left(\sqrt{\sin^2(t) + 1 + 4t^2} \right)^3} (\sin^2(t) + 1 + 4t^2) \\ &= \frac{\sqrt{4 + (-2 \sin(t) + 2(\cos(t))t)^2 + \cos^2(t)}}{\sqrt{\sin^2(t) + 1 + 4t^2}}. \end{aligned}$$

You can observe the formula $a_N^2 + a_T^2 = |\mathbf{a}|^2$ holds. Indeed $a_N^2 + a_T^2 =$

$$\begin{aligned} &\left(\frac{\sqrt{4 + (-2 \sin(t) + 2(\cos(t))t)^2 + \cos^2(t)}}{\sqrt{\sin^2(t) + 1 + 4t^2}} \right)^2 + \left(\frac{\sin(t) \cos(t) + 4t}{\sqrt{(2 + 4t^2 - \cos^2 t)}} \right)^2 \\ &= \frac{4 + (-2 \sin t + 2(\cos t)t)^2 + \cos^2 t}{\sin^2 t + 1 + 4t^2} + \frac{(\sin t \cos t + 4t)^2}{2 + 4t^2 - \cos^2 t} = \cos^2 t + 4 = |\mathbf{a}|^2 \end{aligned}$$

15.1.1 Some Simple Techniques

Recall the formula for acceleration is

$$\mathbf{a} = a_T \mathbf{T} + a_N \mathbf{N} \quad (15.3)$$

where $a_T = \frac{d|v|}{dt}$ and $a_N = \kappa |v|^2$. Of course one way to find a_T and a_N is to just find $|v|$, $\frac{d|v|}{dt}$ and κ and plug in. However, there is another way which might be easier. Take the dot product of both sides with \mathbf{T} . 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

$$\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T} = a_N \mathbf{N} \quad (15.4)$$

and taking norms of both sides,

$$|\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}| = a_N.$$

Also from (15.4),

$$\frac{\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}}{|\mathbf{a} - (\mathbf{a} \cdot \mathbf{T}) \mathbf{T}|} = \frac{a_N \mathbf{N}}{a_N |\mathbf{N}|} = \mathbf{N}.$$

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, consider the example worked above which was fairly messy. I will make it easier by selecting a value of t and by using the above simplifying techniques.

Example 15.1.5 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 15.1.6 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 15.1.7 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 15.1.8 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.

15.2 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 15.2.1 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 15.2.2 The vector $\mathbf{T}(s)$ is called the unit tangent vector and the vector $\mathbf{N}(s)$ is called the **principal normal**. The function $\kappa(s)$ in the above lemma is called the **curvature**. When $\mathbf{T}'(s) \neq 0$ so the principal normal is defined, the vector $\mathbf{B}(s) \equiv \mathbf{T}(s) \times \mathbf{N}(s)$ is called the **binormal**.

The binormal is normal to the osculating plane and \mathbf{B}' tells how fast this vector changes. Thus it measures the rate at which the curve twists.

Lemma 15.2.3 Let $\mathbf{R}(s)$ be a parametrization of a space curve with respect to arc length and let the vectors \mathbf{T} , \mathbf{N} , and \mathbf{B} be as defined above. Then $\mathbf{B}' = \mathbf{T} \times \mathbf{N}'$ and there exists a scalar function $\tau(s)$ such that $\mathbf{B}' = \tau\mathbf{N}$.

Proof: From the definition of $\mathbf{B} = \mathbf{T} \times \mathbf{N}$, and you can differentiate both sides and get $\mathbf{B}' = \mathbf{T}' \times \mathbf{N} + \mathbf{T} \times \mathbf{N}'$. Now recall that \mathbf{T}' is a multiple called curvature multiplied by \mathbf{N} so the vectors \mathbf{T}' and \mathbf{N} have the same direction, so $\mathbf{B}' = \mathbf{T} \times \mathbf{N}'$. Therefore, \mathbf{B}' is either zero or is perpendicular to \mathbf{T} . But also, from the definition of \mathbf{B} , \mathbf{B} is a unit vector and so $\mathbf{B}(s) \cdot \mathbf{B}(s) = 1$. Differentiating this, $\mathbf{B}'(s) \cdot \mathbf{B}(s) + \mathbf{B}(s) \cdot \mathbf{B}'(s) = 0$ showing that \mathbf{B}' is perpendicular to \mathbf{B} also. Therefore, \mathbf{B}' is a vector which is perpendicular to both vectors \mathbf{T} and \mathbf{B} and since this is in three dimensions, \mathbf{B}' must be some scalar multiple of \mathbf{N} , and this multiple is called τ . Thus $\mathbf{B}' = \tau\mathbf{N}$ as claimed. ■

Lets go over this last claim a little more. The following situation is obtained. There are two vectors \mathbf{T} and \mathbf{B} which are perpendicular to each other and both \mathbf{B}' and \mathbf{N} are perpendicular to these two vectors, hence perpendicular to the plane determined by them. Therefore, \mathbf{B}' must be a multiple of \mathbf{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 $\mathbf{T}' = 0$, 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 15.2.4 (Serret Frenet) Let $\mathbf{R}(s)$ be the parametrization with respect to arc length of a space curve and $\mathbf{T}(s) = \mathbf{R}'(s)$ is the unit tangent vector. Suppose $|\mathbf{T}'(s)| \neq 0$ so the principal normal $\mathbf{N}(s) = \frac{\mathbf{T}'(s)}{|\mathbf{T}'(s)|}$ is defined. The binormal is the vector $\mathbf{B} \equiv \mathbf{T} \times \mathbf{N}$ so $\mathbf{T}, \mathbf{N}, \mathbf{B}$ 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 .

$$\mathbf{B}' = \tau \mathbf{N}, \mathbf{T}' = \kappa \mathbf{N}, \mathbf{N}' = -\kappa \mathbf{T} - \tau \mathbf{B}$$

where κ is the curvature and is nonnegative and τ is the torsion.

Proof: $\kappa \geq 0$ because $\kappa = |\mathbf{T}'(s)|$. The first two equations are already established. To get the third, note that $\mathbf{B} \times \mathbf{T} = \mathbf{N}$ which follows because $\mathbf{T}, \mathbf{N}, \mathbf{B}$ 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

$$\mathbf{N}' = \mathbf{B}' \times \mathbf{T} + \mathbf{B} \times \mathbf{T}' = \tau \mathbf{N} \times \mathbf{T} + \kappa \mathbf{B} \times \mathbf{N}.$$

Now recall again that $\mathbf{T}, \mathbf{N}, \mathbf{B}$ is a right hand system. Thus

$$\mathbf{N} \times \mathbf{T} = -\mathbf{B}, \mathbf{B} \times \mathbf{N} = -\mathbf{T}.$$

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 $\mathbf{R}(s) = \mathbf{R}_0 + \int_0^s \mathbf{T}(r) dr$.

The vectors \mathbf{B}, \mathbf{T} , and \mathbf{N} are vectors which are functions of position on the space curve. Often, especially in applications, you deal with a space curve which is parameterized by a function of t where t is time. Thus a value of t would correspond to a point on this curve and you could let $\mathbf{B}(t), \mathbf{T}(t)$, and $\mathbf{N}(t)$ be the binormal, unit tangent, and principal normal at this point of the curve. The following example is typical.

Example 15.2.5 Given the circular helix, $\mathbf{R}(t) = (a \cos t)\mathbf{i} + (a \sin t)\mathbf{j} + (bt)\mathbf{k}$, find the arc length $s(t)$, the unit tangent vector $\mathbf{T}(t)$, the principal normal $\mathbf{N}(t)$, the binormal $\mathbf{B}(t)$, the curvature $\kappa(t)$, and the torsion, $\tau(t)$. Here $t \in [0, T]$.

The arc length is $s(t) = \int_0^t \left(\sqrt{a^2 + b^2} \right) dr = \left(\sqrt{a^2 + b^2} \right) t$. Now the tangent vector is obtained using the chain rule as

$$\mathbf{T} = \frac{d\mathbf{R}}{ds} = \frac{d\mathbf{R}}{dt} \frac{dt}{ds} = \frac{1}{\sqrt{a^2 + b^2}} \mathbf{R}'(t) = \frac{1}{\sqrt{a^2 + b^2}} ((-a \sin t)\mathbf{i} + (a \cos t)\mathbf{j} + b\mathbf{k})$$

The principal normal:

$$\frac{d\mathbf{T}}{ds} = \frac{d\mathbf{T}}{dt} \frac{dt}{ds} = \frac{1}{a^2 + b^2} ((-a \cos t)\mathbf{i} + (-a \sin t)\mathbf{j} + 0\mathbf{k})$$

and so

$$\mathbf{N} = \frac{d\mathbf{T}}{ds} / \left| \frac{d\mathbf{T}}{ds} \right| = -((\cos t)\mathbf{i} + (\sin t)\mathbf{j})$$

The binormal:

$$\mathbf{B} = \frac{1}{\sqrt{a^2 + b^2}} \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -a \sin t & a \cos t & b \\ -\cos t & -\sin t & 0 \end{vmatrix} = \frac{1}{\sqrt{a^2 + b^2}} ((b \sin t)\mathbf{i} - b \cos t \mathbf{j} + a \mathbf{k})$$

Now the curvature $\kappa(t) = \left| \frac{d\mathbf{T}}{ds} \right| = \sqrt{\left(\frac{a \cos t}{a^2 + b^2} \right)^2 + \left(\frac{a \sin t}{a^2 + b^2} \right)^2} = \frac{a}{a^2 + b^2}$. Note the curvature is constant in this example. The final task is to find the torsion. Recall that $\mathbf{B}' = \tau \mathbf{N}$ where the derivative on \mathbf{B} is taken with respect to arc length. Therefore, remembering that t is a function of s ,

$$\begin{aligned} \mathbf{B}'(s) &= \frac{1}{\sqrt{a^2 + b^2}} ((b \cos t)\mathbf{i} + (b \sin t)\mathbf{j}) \frac{dt}{ds} = \frac{1}{a^2 + b^2} ((b \cos t)\mathbf{i} + (b \sin t)\mathbf{j}) \\ &= \tau (-(\cos t)\mathbf{i} - (\sin t)\mathbf{j}) = \tau \mathbf{N} \end{aligned}$$

and it follows $-b/(a^2 + b^2) = \tau$.

An important application of the usefulness of these ideas involves the decomposition of the acceleration in terms of these vectors of an object moving over a space curve.

Corollary 15.2.6 *Let $\mathbf{R}(t)$ be a space curve and denote by $\mathbf{v}(t)$ the velocity, $\mathbf{v}(t) = \mathbf{R}'(t)$, let $v(t) \equiv |\mathbf{v}(t)|$ denote the speed, and let $\mathbf{a}(t)$ denote the acceleration. Then $\mathbf{v} = v\mathbf{T}$ and $\mathbf{a} = \frac{dv}{dt}\mathbf{T} + \kappa v^2 \mathbf{N}$.*

Proof: $\mathbf{T} = \frac{d\mathbf{R}}{ds} = \frac{d\mathbf{R}}{dt} \frac{dt}{ds} = \mathbf{v} \frac{dt}{ds}$. Also, $s = \int_0^t v(r) dr$ and so $\frac{ds}{dt} = v$ which implies $\frac{dt}{ds} = \frac{1}{v}$. Therefore, $\mathbf{T} = \mathbf{v}/v$ which implies $\mathbf{v} = v\mathbf{T}$ as claimed.

Now the acceleration is just the derivative of the velocity and so by the Serrat Frenet formulas,

$$\mathbf{a} = \frac{dv}{dt}\mathbf{T} + v \frac{d\mathbf{T}}{dt} = \frac{dv}{dt}\mathbf{T} + v \frac{d\mathbf{T}}{ds} v = \frac{dv}{dt}\mathbf{T} + v^2 \kappa \mathbf{N}$$

Note how this decomposes the acceleration into a component tangent to the curve and one which is normal to it. Also note that from the above, $v|\mathbf{T}'| \frac{\mathbf{T}'(t)}{|\mathbf{T}'|} = v^2 \kappa \mathbf{N}$ and so $\frac{|\mathbf{T}'|}{v} = \kappa$ and $\mathbf{N} = \frac{\mathbf{T}'(t)}{|\mathbf{T}'|}$. ■

15.3 Exercises

1. Find a parametrization for the intersection of the planes $2x + y + 3z = -2$ and $3x - 2y + z = -4$.
2. Find a parametrization for the intersection of the plane $3x + y + z = -3$ and the circular cylinder $x^2 + y^2 = 1$.
3. Find a parametrization for the intersection of the plane $4x + 2y + 3z = 2$ and the elliptic cylinder $x^2 + 4z^2 = 9$.

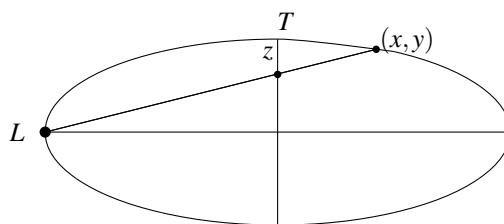
4. Find a parametrization for the straight line joining $(1, 2, 1)$ and $(-1, 4, 4)$.
5. Find a parametrization for the intersection of the surfaces $3y + 3z = 3x^2 + 2$ and $3y + 2z = 3$.
6. Find a formula for the curvature of the curve $y = \sin x$ in the xy plane.
7. An object moves over the curve (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.
8. Find a formula for the curvature of the space curve in \mathbb{R}^2 , $(x(t), y(t))$.
9. 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 $a_T \mathbf{T} + a_N \mathbf{N}$.
10. An object moves over the helix, $(\cos t, \sin t, t)$. Find the normal and tangential components of the acceleration of this object as a function of t and write the acceleration in the form $a_T \mathbf{T} + a_N \mathbf{N}$.
11. 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 $a_T \mathbf{T} + a_N \mathbf{N}$.
12. An object moves over the helix, $(\cos t, \sin t, 2t)$. Find the osculating plane at the point of the curve corresponding to $t = \pi/4$.
13. 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} .

14. Suppose $|\mathbf{R}(t)| = c$ where c is a constant $\mathbf{R}(t)$. Show the velocity, $\mathbf{R}'(t)$ is always perpendicular to $\mathbf{R}(t)$.
15. 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?
16. 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 $s = 0$.
17. 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 $s = 0$.
18. 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 $s = 0$.

19. Consider the curve obtained from the graph of $y = f(x)$. Find a formula for the curvature.
20. Consider the curve in the plane $y = e^x$. Find the point on this curve at which the curvature is a maximum.
21. 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 $x = 0$?
22. An object of mass m moving over a space curve is acted on by a force \mathbf{F} . 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.
23. 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.

Chapter 16

Functions Of Many Variables

16.1 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 16.1.1 Let A denote a nonempty subset of \mathbb{R}^p . A point x is said to be a **limit point** of the set A if for every $r > 0$, $B(x, r)$ contains infinitely many points of A .

Example 16.1.2 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 x is any point of \mathbb{R}^3 , $B(x, 1/4)$ contains at most one point.

Example 16.1.3 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 $x \in U$, There exists $B(x, r) \subseteq U$ for some $r > 0$. Now consider the line segment $x + tr e_1$ where $t \in [0, 1/2]$. This describes infinitely many points and they are all in $B(x, r)$ because $|x + tr e_1 - x| = tr < r$. 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 16.1.4 Let $f : D(f) \subseteq \mathbb{R}^p \rightarrow \mathbb{R}^q$ where $q, p \geq 1$ 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,

$$|L - f(y)| < \varepsilon.$$

The condition that x must be a limit point of $D(f)$ if you are to take a limit at x is what makes the limit well defined.

Proposition 16.1.5 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 12.5.5 on Page 223.

Theorem 16.1.6 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 $a, b \in \mathbb{R}$,

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} a\mathbf{f}(\mathbf{y}) + b\mathbf{g}(\mathbf{y}) = a\mathbf{L} + b\mathbf{K}, \quad (16.1)$$

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f} \cdot \mathbf{g}(\mathbf{y}) = \mathbf{L} \cdot \mathbf{K} \quad (16.2)$$

Also, if \mathbf{h} is a continuous function defined near \mathbf{L} , then

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{h} \circ \mathbf{f}(\mathbf{y}) = \mathbf{h}(\mathbf{L}). \quad (16.3)$$

For a vector valued function

$$\mathbf{f}(\mathbf{y}) = (f_1(\mathbf{y}), \dots, f_q(\mathbf{y}))^T,$$

$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L} = (L_1, \dots, L_k)^T$ if and only if

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} f_k(\mathbf{y}) = L_k \quad (16.4)$$

for each $k = 1, \dots, p$.

In the case where \mathbf{f} and \mathbf{g} have values in \mathbb{R}^3

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) \times \mathbf{g}(\mathbf{y}) = \mathbf{L} \times \mathbf{K}. \quad (16.5)$$

Also recall Theorem 12.5.6 on Page 226.

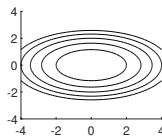
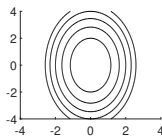
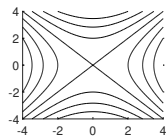
Theorem 16.1.7 For $\mathbf{f} : D(\mathbf{f}) \rightarrow \mathbb{R}^q$ and $\mathbf{x} \in D(\mathbf{f})$ such that \mathbf{x} is a limit point of $D(\mathbf{f})$, it follows \mathbf{f} is continuous at \mathbf{x} if and only if $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x})$.

16.2 Exercises

1. Sketch the contour graph of the function of two variables $f(x, y) = (x - 1)^2 + (y - 2)^2$.

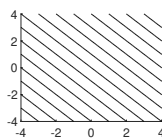
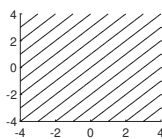
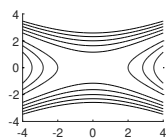
2. Which of the following functions could correspond to the following contour graphs?

$$z = x^2 + 3y^2, z = 3x^2 + y^2, z = x^2 - y^2, z = x + y.$$



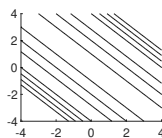
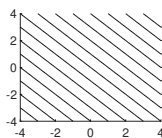
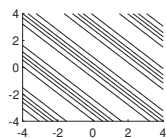
3. Which of the following functions could correspond to the following contour graphs?

$$z = x^2 - 3y^2, z = y^2 + 3x^2, z = x - y, z = x + y.$$



4. Which of the following functions could correspond to the following contour graphs?

$$z = \sin(x + y), z = x + y, z = (x + y)^2, z = x^2 - y.$$



5. Find the following limits if they exist. If they do not exist, explain why.

(a) $\lim_{(x,y) \rightarrow (0,0)} \frac{x^2 - y^2}{x^2 + y^2}$

(b) $\lim_{(x,y) \rightarrow (0,0)} \frac{2x^3 + xy^2 - x^2 - 2y^2}{x^2 + 2y^2}$

(c) $\lim_{(x,y) \rightarrow (0,0)} \frac{\sin(x^2 + y^2)}{x^2 + y^2}$

(d) $\lim_{(x,y) \rightarrow (0,0)} \frac{\sin(x^2 + 2y^2)}{x^2 + 2y^2}$

(e) $\lim_{(x,y) \rightarrow (0,0)} \frac{\sin(x^2 + 2y^2)}{2x^2 + y^2}$

(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) $\lim_{(x,y) \rightarrow (0,0)} \frac{xy}{x^2 + y^2}$

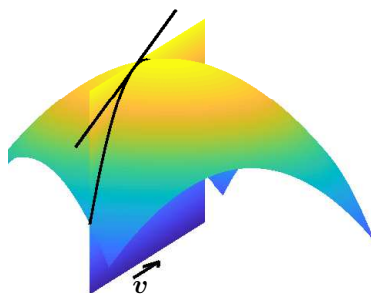
$$(d) \lim_{(x,y) \rightarrow (1,0)} \frac{x^3 - 3x^2 + 3x - 1 - y^2x + y^2}{x^2 - 2x + 1 + y^2}$$

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.

16.3 The Directional Derivative And Partial Derivatives

16.3.1 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 in the xy plane and $x_0 \equiv (x_0, y_0)$ is a point in the xy plane. When (x, y) moves in the direction of v , this results in a change in $z = f(x, y)$ as shown in the picture. The directional derivative in this direction is 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. If you looked at it from the side, you would be getting the slope of the indicated tangent line. 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.

Definition 16.3.1 Let $f : U \rightarrow \mathbb{R}$ where U is an open set in \mathbb{R}^n and let v be a unit vector. For $x \in U$, define the **directional derivative** of f in the direction v , at the point x as

$$D_v f(x) \equiv \lim_{t \rightarrow 0} \frac{f(x + tv) - f(x)}{t}.$$

Example 16.3.2 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.

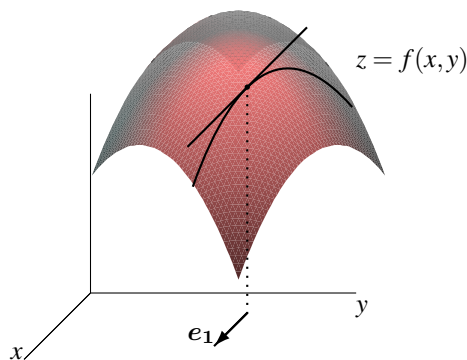
16.3.2 Partial Derivatives

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.

Thus in case of a function of two variables, the directional derivative in the direction $\mathbf{i} = \mathbf{e}_1$ is the slope of the indicated straight line in the following picture.



As in the case of a general directional derivative, you fix y and take the derivative of the function $x \rightarrow f(x, y)$. More generally, even in situations which cannot be drawn, the definition of a partial derivative is as follows.

Definition 16.3.3 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 ,

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) \equiv D_{\mathbf{e}_i}f(\mathbf{x}).$$

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 called the **partial 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} .

Example 16.3.4 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, provided the partial derivatives exist and are continuous.

Definition 16.3.5 Suppose $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ where U is an open set and the partial derivatives of f all exist and are continuous on U . Under these conditions, 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 16.3.6 In the situation of Definition 16.3.5 and 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 16.3.7 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 16.3.8 *In the situation of Definition 16.3.5, 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 16.3.9 *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 16.3.10 *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 16.3.11 *Find the partial derivative with respect to x of the function $\mathbf{f}(x, y, z, w) = (xy^2, z \sin(xy), z^3 x)^T$.*

From the above definition, $\mathbf{f}_x(x, y, z) = D_1 \mathbf{f}(x, y, z) = (y^2, zy \cos(xy), z^3)^T$.

16.4 Exercises

- Find the directional derivative of $f(x, y, z) = x^2y + z^4$ in the direction of the vector $(1, 3, -1)$ when $(x, y, z) = (1, 1, 1)$.
- Find the directional derivative of $f(x, y, z) = \sin(x + y^2) + z$ in the direction of the vector $(1, 2, -1)$ when $(x, y, z) = (1, 1, 1)$.
- 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 $(x, y, z) = (1, 1, 1)$.
- Using the conclusion of Proposition 16.3.6, prove Proposition 16.3.8 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 .
- Find the largest value of the directional derivative of $f(x, y, z) = \ln(x + y^2) + z^2$ at the point $(1, 1, 1)$.
- Find the smallest value of the directional derivative of $f(x, y, z) = x \sin(4xy^2) + z^2$ at the point $(1, 1, 1)$.
- 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?
- Find the partial derivative with respect to y of the function $\mathbf{f}(x, y, z, w) = (y^2, z^2 \sin(xy), z^3 x)^T$.
- Find the partial derivative with respect to x of the function $\mathbf{f}(x, y, z, w) = (wx, zx \sin(xy), z^3 x)^T$.
- Find $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$ for $f =$
 - $x^2y^2z + w$
 - $e^2 + xy + z^2$
 - $\sin(z^2) + \cos(xy)$
 - $\ln(x^2 + y^2 + 1) + e^z$
 - $\sin(xyz) + \cos(xy)$
- Find $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$ for $f =$
 - $x^2y + \cos(xy) + z^3y$
 - $e^{x^2+y^2}z \sin(x + y)$
 - $z^2 \sin^3(e^{x^2+y^3})$
 - $x^2 \cos(\sin(\tan(z^2 + y^2)))$
 - 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?

16.5 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.

Theorem 16.5.1 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

$$\Delta(s, t) \equiv \frac{1}{st} \left\{ \overbrace{f(x+t, y+s) - f(x+t, y)}^{h(t)} - \overbrace{(f(x, y+s) - f(x, y))}^{h(0)} \right\}. \quad (16.6)$$

Note that $(x+t, y+s) \in U$ because

$$\begin{aligned} |(x+t, y+s) - (x, y)| &= |(t, s)| = (t^2 + s^2)^{1/2} \\ &\leq \left(\frac{r^2}{4} + \frac{r^2}{4} \right)^{1/2} = \frac{r}{\sqrt{2}} < r. \end{aligned}$$

As implied above, $h(t) \equiv f(x+t, y+s) - f(x, y)$. Therefore, by the mean value theorem from calculus and the (one variable) chain rule,

$$\begin{aligned} \Delta(s, t) &= \frac{1}{st} (h(t) - h(0)) = \frac{1}{st} h'(\alpha t) t \\ &= \frac{1}{s} (f_x(x + \alpha t, y+s) - f_x(x, y)) \end{aligned}$$

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 (16.6), $\Delta(s, t)$ 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) ,

$$\lim_{(s,t) \rightarrow (0,0)} \Delta(s, t) = f_{xy}(x, y) = f_{yx}(x, y). \quad \blacksquare$$

The following is obtained from the above by simply fixing all the variables except for the two of interest.

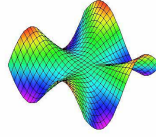
Corollary 16.5.2 *Suppose U is an open subset of \mathbb{R}^n and $f : U \rightarrow \mathbb{R}$ has the property that for two indices k, l , f_{x_k} , f_{x_l} , $f_{x_l x_k}$, and $f_{x_k x_l}$ exist on U and $f_{x_k x_l}$ and $f_{x_l x_k}$ are both continuous at $\mathbf{x} \in U$. Then $f_{x_k x_l}(\mathbf{x}) = f_{x_l x_k}(\mathbf{x})$.*

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 16.5.3 *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.

16.6 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 Laplace².

²Laplace was a great physicist and mathematician of the 1700's. He made fundamental contributions to mechanics and astronomy.

Definition 16.6.1 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 $u_{tt} - \Delta u = 0$.

Example 16.6.2 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 16.6.3 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 16.6.4 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 $u_{tt} - \Delta u = 0$.

16.7 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.

5. Let f be a function which has continuous derivatives. Show that $u(t, x) = f(x - ct)$ solves the wave equation $u_{tt} - c^2 \Delta u = 0$. What about $u(x, t) = f(x + ct)$?
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) $x^3 - 3xy^2$
 - (b) $3x^2y - y^3$
 - (c) $x^3 - 3xy^2 + 2x^2 - 2y^2$
 - (d) $3x^2y - y^3 + 4xy$
 - (e) $3x^2 - y^3 + 4xy$
 - (f) $3x^2y - y^3 + 4y$
 - (g) $x^3 - 3x^2y^2 + 2x^2 - 2y^2$
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 $u(x, t) = \frac{1}{\sqrt{t}} e^{-x^2/4c^2t}$ solves the heat equation $u_t = c^2 u_{xx}$.

Chapter 17

The Derivative Of A Function Of Many Variables

17.1 The Derivative Of Functions Of One Variable

First consider the notion of the derivative of a function of one variable.

Observation 17.1.1 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 17.1.2 A vector valued function of a vector \mathbf{v} is called $\mathbf{o}(\mathbf{v})$ (referred to as “little \mathbf{o} of \mathbf{v} ”) if

$$\lim_{|\mathbf{v}| \rightarrow 0} \frac{\mathbf{o}(\mathbf{v})}{|\mathbf{v}|} = \mathbf{0}. \quad (17.1)$$

Thus for a function of one variable, the function $f(x+h) - f(x) - f'(x)h$ is $\mathbf{o}(h)$. When we say a function is $\mathbf{o}(h)$, 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,

$$\mathbf{o}(\mathbf{v}) = \mathbf{o}(\mathbf{v}) + \mathbf{o}(\mathbf{v}), \mathbf{o}(\mathbf{v}) = 45 \times \mathbf{o}(\mathbf{v}), \mathbf{o}(\mathbf{v}) = \mathbf{o}(\mathbf{v}) - \mathbf{o}(\mathbf{v}), \text{etc.}$$

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 17.1.3 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

$$f(x+h) - f(x) = ph + o(h) \quad (17.2)$$

In this case, $p = f'(x)$.

Proof: From the above observation it follows that if $f'(x)$ does exist, then (17.2) holds. Suppose then that (17.2) 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 17.1.4 Let $f(x) = x^3$. Find $f'(x)$.

$$\begin{aligned} f(x+h) &= (x+h)^3 = x^3 + 3x^2h + 3xh^2 + h^3 \\ &= f(x) + 3x^2h + (3xh + h^2)h. \end{aligned}$$

Since $(3xh + h^2)h = o(h)$, it follows $f'(x) = 3x^2$.

Example 17.1.5 Let $f(x) = \sin(x)$. Find $f'(x)$.

$$\begin{aligned} f(x+h) - f(x) &= \sin(x+h) - \sin(x) = \sin(x)\cos(h) + \cos(x)\sin(h) - \sin(x) \\ &= \cos(x)\sin(h) + \sin(x)\frac{(\cos(h)-1)}{h}h \\ &= \cos(x)h + \cos(x)\frac{(\sin(h)-h)}{h}h + \sin(x)\frac{(\cos(h)-1)}{h}h. \end{aligned}$$

Now

$$\cos(x)\frac{(\sin(h)-h)}{h}h + \sin(x)\frac{(\cos(h)-1)}{h}h = o(h). \quad (17.3)$$

Remember the fundamental limits which allowed you to find the derivative of $\sin(x)$ were

$$\lim_{h \rightarrow 0} \frac{\sin(h)}{h} = 1, \quad \lim_{h \rightarrow 0} \frac{\cos(h)-1}{h} = 0. \quad (17.4)$$

These same limits are what is needed to verify (17.3).

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 .

17.2 Exercises

1. Determine which of the following functions are $o(h)$.

- (a) h^2
- (b) $h \sin(h)$
- (c) $|h|^{3/2} \ln(|h|)$
- (d) $h^2x + yh^3$
- (e) $\sin(h^2)$
- (f) $\sin(h)$
- (g) $xh \sin(\sqrt{|h|}) + x^5h^2$
- (h) $\exp(-1/|h|^2)$

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_1v_2
- (b) $v_2 \sin(v_1)$
- (c) $v_1^2 + v_2$
- (d) $v_2 \sin(v_1 + v_2)$
- (e) $v_1(v_1 + v_2 + xv_3)$
- (f) $(e^{v_1} - 1 - v_1)$
- (g) $(\mathbf{x} \cdot \mathbf{v})|\mathbf{v}|$

3. Here are some vector valued functions of $\mathbf{v} \in \mathbb{R}^n$. Determine which ones are $\mathbf{o}(\mathbf{v})$.

- (a) $(\mathbf{x} \cdot \mathbf{v}) \mathbf{v}$
- (b) $\sin(v_1) \mathbf{v}$
- (c) $\sqrt{|\mathbf{x} \cdot \mathbf{v}|} |\mathbf{v}|^{2/3}$
- (d) $\sqrt{|\mathbf{x} \cdot \mathbf{v}|} |\mathbf{v}|^{1/2}$
- (e) $\left(\sin \left(\sqrt{|\mathbf{x} \cdot \mathbf{v}|} \right) - \sqrt{|\mathbf{x} \cdot \mathbf{v}|} \right) \cdot |\mathbf{v}|^{-1/4}$
- (f) $\exp \left(-1/|\mathbf{v}|^2 \right)$
- (g) $\mathbf{v}^T A \mathbf{v}$ 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 $p > 1$.

17.3 The Derivative Of Functions Of Many Variables

The way of thinking about the derivative in Theorem 17.1.3 is exactly what is needed to define the derivative of a function of n variables. Recall the following definition.

Definition 17.3.1 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 $T(a\mathbf{x} + b\mathbf{y}) = aT(\mathbf{x}) + bT(\mathbf{y})$.

Recall that from the properties of matrix multiplication, if A is an $p \times n$ 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. Of course it might not be worth the fuss to think of it this way for a function of one variable but this is the way you must think of it for a function of n variables.

Definition 17.3.2 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,

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + T\mathbf{h} + \mathbf{o}(\mathbf{h}). \quad (17.5)$$

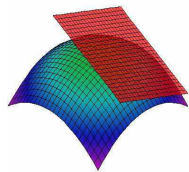
The derivative of the function \mathbf{f} , denoted by $D\mathbf{f}(\mathbf{x})$, is this linear transformation. Thus

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + D\mathbf{f}(\mathbf{x})\mathbf{h} + \mathbf{o}(\mathbf{h})$$

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 $\mathbf{x} \in \mathbb{R}^2$ and f has values 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 $Df(\mathbf{x})$ would not be well defined.

Theorem 17.3.3 Suppose f is differentiable, as given above in (17.5). Then T is uniquely determined. Furthermore, the matrix of T is the following $p \times n$ matrix

$$\left(\begin{array}{ccc} \frac{\partial f(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{array} \right)$$

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,

$$\begin{aligned} f(\mathbf{x} + t\mathbf{h}) &= f(\mathbf{x}) + Tt\mathbf{h} + \mathbf{o}(t\mathbf{h}) \\ f(\mathbf{x} + t\mathbf{h}) &= f(\mathbf{x}) + T_1t\mathbf{h} + \mathbf{o}(t\mathbf{h}) \end{aligned}$$

Now $\mathbf{o}(t\mathbf{h}) = \mathbf{o}(t)$ and so, subtracting these yields

$$Tt\mathbf{h} - T_1t\mathbf{h} = \mathbf{o}(t)$$

Divide both sides by t to obtain

$$T\mathbf{h} - T_1\mathbf{h} = \frac{\mathbf{o}(t)}{t}$$

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 8.3.2, this is the matrix whose i^{th} column is $T\mathbf{e}_i$. However, from the definition of T , letting $t \neq 0$,

$$\begin{aligned} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t} &= \frac{1}{t} (T(t\mathbf{e}_i) + \mathbf{o}(t\mathbf{e}_i)) \\ &= T(\mathbf{e}_i) + \frac{\mathbf{o}(t\mathbf{e}_i)}{t} = T(\mathbf{e}_i) + \frac{\mathbf{o}(t)}{t} \end{aligned}$$

Then letting $t \rightarrow 0$, it follows that

$$T\mathbf{e}_i = \frac{\partial f}{\partial x_i}(\mathbf{x})$$

Recall from theorem 8.3.2 this shows the matrix of the linear transformation is as claimed.

■

Other notations which are often used for this matrix or the linear transformation are $\mathbf{f}'(\mathbf{x})$, $J(\mathbf{x})$, and even $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$ or $\frac{d\mathbf{f}}{d\mathbf{x}}$. Also, the above definition can now be written in the form

$$\mathbf{f}(\mathbf{x} + \mathbf{v}) = \mathbf{f}(\mathbf{x}) + \sum_{j=1}^p \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_j} v_j + \mathbf{o}(\mathbf{v})$$

or

$$\mathbf{f}(\mathbf{x} + \mathbf{v}) - \mathbf{f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \end{pmatrix} \mathbf{v} + \mathbf{o}(\mathbf{v})$$

Here is an example of a scalar valued nonlinear function.

Example 17.3.4 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 17.3.5 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

$$\mathbf{f}(\mathbf{x} + \mathbf{v}) = \mathbf{f}(\mathbf{x}) + \begin{pmatrix} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_p} \end{pmatrix} \mathbf{v} + \mathbf{o}(\mathbf{v}).$$

Taking the i^{th} coordinate of the above equation yields

$$f_i(\mathbf{x} + \mathbf{v}) = f_i(\mathbf{x}) + \sum_{j=1}^p \frac{\partial f_i(\mathbf{x})}{\partial x_j} v_j + o(\mathbf{v}),$$

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

$$\mathbf{f}(\mathbf{x} + \mathbf{v}) = \mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v}), \quad (17.6)$$

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 17.3.6 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 17.3.7 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 $\mathbf{o}(h, k, l)$ 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 $\mathbf{o}(\mathbf{v})$. Sometimes you have to go directly to the definition and verify it is differentiable from the definition. For example, you may have seen the following important example in one variable calculus.

Example 17.3.8 Let $f(x) = \begin{cases} x^2 \sin\left(\frac{1}{x}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$. 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 17.3.9 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

$$\begin{aligned} & f(1 + h_1, 2 + h_2) - f(1, 2) \\ &= \begin{pmatrix} (1 + h_1)^2(2 + h_2) + (2 + h_2)^2 \\ (2 + h_2)^3(1 + h_1) \end{pmatrix} - \begin{pmatrix} 6 \\ 8 \end{pmatrix} \\ &= \begin{pmatrix} 5h_2 + 4h_1 + 2h_1h_2 + 2h_1^2 + h_1^2h_2 + h_2^2 \\ 8h_1 + 12h_2 + 12h_1h_2 + 6h_2^2 + 6h_2^2h_1 + h_2^3 + h_2^3h_1 \end{pmatrix} \\ &= \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \begin{pmatrix} 2h_1h_2 + 2h_1^2 + h_1^2h_2 + h_2^2 \\ 12h_1h_2 + 6h_2^2 + 6h_2^2h_1 + h_2^3 + h_2^3h_1 \end{pmatrix} \\ &= \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + o(h). \end{aligned}$$

Therefore, the matrix of the derivative is $\begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix}$.

Example 17.3.10 Let $f(x, y) = \begin{pmatrix} x^3y + y^2 \\ xy^2 + 1 \end{pmatrix}$. Find $Df(x, y)$.

You know that if there is a derivative, its standard matrix is of the form

$$\begin{pmatrix} f_{1x}(x, y) & f_{1y}(x, y) \\ f_{2x}(x, y) & f_{2y}(x, y) \end{pmatrix} = \begin{pmatrix} 3x^2y & x^3 + 2y \\ y^2 & 2xy \end{pmatrix}$$

Does it work? Is

$$\begin{pmatrix} (x + u)^3(y + v) + (y + v)^2 \\ (x + u)(y + v)^2 + 1 \end{pmatrix} - \begin{pmatrix} x^3y + y^2 \\ xy^2 + 1 \end{pmatrix}$$

$$-\begin{pmatrix} 3x^2y & x^3+2y \\ y^2 & 2xy \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = o \begin{pmatrix} u \\ v \end{pmatrix}?$$

Doing the computations, it follows the left side of the equal sign is of the form

$$\begin{pmatrix} 3x^2uv + 3xu^2y + 3xu^2v + u^3y + u^3v + v^2 \\ xv^2 + 2uyv + uv^2 \end{pmatrix}$$

This is $o \begin{pmatrix} u \\ v \end{pmatrix}$ because it involves terms like uv, u^2v , etc. Each term being of degree 2 or more.

17.4 Exercises

- Use the definition of the derivative to find the 1×1 matrix which is the derivative of the following functions.
 - $f(t) = t^2 + t$.
 - $f(t) = t^3$.
 - $f(t) = t \sin(t)$.
 - $f(t) = \ln(t^2 + 1)$.
 - $f(t) = t|t|$.
- 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 $Df(x) = 0$.
- Use the above definition of the derivative to prove the product rule for functions of 1 variable.
- Let $f(x, y) = x \sin(y)$. Compute the derivative directly from the definition.
- Let $f(x, y) = x^2 \sin(y)$. Compute the derivative directly from the definition.
- Let $\mathbf{f}(x, y) = \begin{pmatrix} x^2 + y \\ y^2 \end{pmatrix}$. Compute the derivative directly from the definition.
- Let $\mathbf{f}(x, y) = \begin{pmatrix} x^2y \\ x + y^2 \end{pmatrix}$. Compute the derivative directly from the definition.
- Let $f(x, y) = x^\alpha y^\beta$. Show $Df(x, y) = \begin{pmatrix} \alpha x^{\alpha-1} y^\beta & x^\alpha \beta y^{\beta-1} \end{pmatrix}$.
- Let $\mathbf{f}(x, y) = \begin{pmatrix} x^2 \sin(y) \\ x^2 + y \end{pmatrix}$. Find $D\mathbf{f}(x, y)$.
- Let $f(x, y) = \sqrt{x} \sqrt[3]{y}$. Find the approximate change in f when (x, y) goes from $(4, 8)$ to $(4.01, 7.99)$.

11. 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 .
12. 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 $nf(t)^{n-1}f'(t)$.
13. 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

$$nf(x, y)^{n-1}Df(x, y)$$

14. 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.

17.5 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 17.5.1 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 17.3.3 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 17.5.2 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.

$$\mathbf{f}(\mathbf{x} + v_k \mathbf{e}_k) - \mathbf{f}(\mathbf{x}) = \int_0^1 \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x} + tv_k \mathbf{e}_k) v_k dt. \quad (17.7)$$

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

$$\mathbf{h}'(t) = \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x} + t v_k \mathbf{e}_k) v_k$$

Therefore,

$$\mathbf{f}(\mathbf{x} + v_k \mathbf{e}_k) - \mathbf{f}(\mathbf{x}) = \mathbf{h}(1) - \mathbf{h}(0) = \int_0^1 \mathbf{h}'(t) dt = \int_0^1 \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x} + t v_k \mathbf{e}_k) v_k dt.$$

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,

$$\begin{aligned} \mathbf{f}(\mathbf{x} + \mathbf{v}) - \mathbf{f}(\mathbf{x}) &= \sum_{i=1}^n \left(\mathbf{f}\left(\mathbf{x} + \sum_{k=i}^n v_k \mathbf{e}_k\right) - \mathbf{f}\left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k\right) \right) \\ &= \sum_{i=1}^n \int_0^1 \frac{\partial \mathbf{f}}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + t v_i \mathbf{e}_i \right) v_i dt \\ &= \sum_{i=1}^n \int_0^1 \left(\frac{\partial \mathbf{f}}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + t v_i \mathbf{e}_i \right) v_i - \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}) v_i \right) dt \\ &\quad + \sum_{i=1}^n \int_0^1 \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}) v_i dt = \sum_{i=1}^n \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}) v_i \\ &\quad + \sum_{i=1}^n \int_0^1 \left(\frac{\partial \mathbf{f}}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + t v_i \mathbf{e}_i \right) - \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}) \right) v_i dt \\ &= \sum_{i=1}^n \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}) v_i + o(\mathbf{v}) \end{aligned}$$

and this shows \mathbf{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 14.2.2,

$$\begin{aligned} &\left| \int_0^1 \left(\frac{\partial \mathbf{f}}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + t v_i \mathbf{e}_i \right) - \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}) \right) v_i dt \right| \\ &\leq \sqrt{p} \int_0^1 \left| \frac{\partial \mathbf{f}}{\partial x_i} \left(\mathbf{x} + \sum_{k=i+1}^n v_k \mathbf{e}_k + t v_i \mathbf{e}_i \right) - \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}) \right| dt |\mathbf{v}| \end{aligned}$$

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 \mathbf{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 17.5.3 Let $f(x, y) = \begin{pmatrix} x^2y + y^2 \\ y^3x \end{pmatrix}$. Find $Df(x, y)$.

From Theorem 17.5.2 this function is differentiable because all possible partial derivatives are continuous. Thus

$$Df(x, y) = \begin{pmatrix} 2xy & x^2 + 2y \\ y^3 & 3y^2x \end{pmatrix}.$$

In particular,

$$Df(1, 2) = \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix}.$$

Here is another example.

Example 17.5.4 Let $f(x_1, x_2, x_3) = \begin{pmatrix} x_1^2x_2 + x_2^2 \\ x_2x_1 + x_3 \\ \sin(x_1x_2x_3) \end{pmatrix}$. 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 17.5.5 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 17.5.6 Let $f : U \rightarrow \mathbb{R}^q$ where U is an open subset of \mathbb{R}^p . If f is differentiable at \mathbf{x} , then f is continuous at \mathbf{x} .

Proof: From the definition of what it means to be differentiable,

$$\begin{aligned} |f(\mathbf{y}) - f(\mathbf{x})| &= |Df(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \mathbf{o}(\mathbf{y} - \mathbf{x})| \\ &\leq |Df(\mathbf{x})(\mathbf{y} - \mathbf{x})| + |\mathbf{y} - \mathbf{x}| \end{aligned}$$

provided $\|\mathbf{y} - \mathbf{x}\|$ is sufficiently small. Letting M denote the matrix of $D\mathbf{f}(\mathbf{x})$,

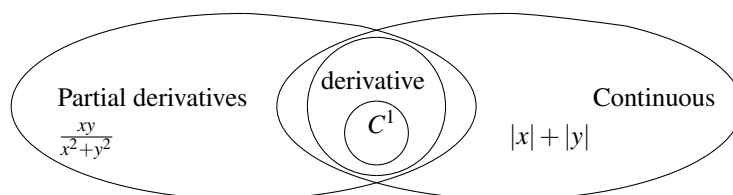
$$\|D\mathbf{f}(\mathbf{x})(\mathbf{y} - \mathbf{x})\|^2 = \sum_i \left| \sum_j M_{ij}(y_j - x_j) \right|^2 \leq \|\mathbf{y} - \mathbf{x}\| \sum_i \left(\sum_j |M_{ij}| \right)^2$$

and so, for \mathbf{y} close enough to \mathbf{x} , there exists a constant C such that

$$\|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})\| \leq C \|\mathbf{y} - \mathbf{x}\|$$

which shows that \mathbf{f} is continuous at \mathbf{x} . ■

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 17.3.8 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 17.5.7 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 $(0,0)$.

$$\frac{f(x,0) - f(0,0)}{x} = \frac{0-0}{x} = 0$$

and so $f_x(0,0) = 0$. Similarly,

$$\frac{f(0,y) - f(0,0)}{y} = \frac{0-0}{y} = 0$$

and so $f_y(0,0) = 0$. 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 $x \rightarrow 0$.

17.6 The Chain Rule

17.6.1 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

$$Df(g(x))Dg(x) = D(f \circ g)(x). \quad (17.8)$$

17.6.2 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 (17.8) both exist then the same formula holds in this more general case. Thus

$$D\mathbf{f}(\mathbf{g}(x))D\mathbf{g}(x) = D(\mathbf{f} \circ \mathbf{g})(x)$$

Note this all makes sense because $D\mathbf{f}(\mathbf{g}(x))$ is a $q \times p$ matrix and $D\mathbf{g}(x)$ is a $p \times n$ matrix. Remember it is all right to do $(q \times p)(p \times n)$. The middle numbers match. More precisely,

Theorem 17.6.1 (Chain rule) *Let U be an open set in \mathbb{R}^n , let V be an open set in \mathbb{R}^p , let $\mathbf{g} : U \rightarrow \mathbb{R}^p$ be such that $\mathbf{g}(U) \subseteq V$, and let $\mathbf{f} : V \rightarrow \mathbb{R}^q$. Suppose $D\mathbf{g}(x)$ exists for some $x \in U$ and that $D\mathbf{f}(\mathbf{g}(x))$ exists. Then $D(\mathbf{f} \circ \mathbf{g})(x)$ exists and furthermore,*

$$D(\mathbf{f} \circ \mathbf{g})(x) = D\mathbf{f}(\mathbf{g}(x))D\mathbf{g}(x). \quad (17.9)$$

In particular,

$$\frac{\partial (\mathbf{f} \circ \mathbf{g})(x)}{\partial x_j} = \sum_{i=1}^p \frac{\partial \mathbf{f}(\mathbf{g}(x))}{\partial y_i} \frac{\partial g_i(x)}{\partial x_j}. \quad (17.10)$$

There is an easy way to remember this in terms of the repeated index summation convention presented earlier. Let $\mathbf{y} = \mathbf{g}(\mathbf{x})$ and $z = f(\mathbf{y})$. Then the above says

$$\frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x_k} = \frac{\partial z}{\partial x_k}. \quad (17.11)$$

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 17.6.2 Let $f(u, v) = \sin(uv)$ and let $u(x, y, t) = t \sin x + \cos y$ and $v(x, y, t, s) = \tan x + y^2 + ts$. 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 (17.11), $\frac{\partial z}{\partial t} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial t} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial t} = v \cos(uv) \sin(x) + u s \cos(uv)$. Here $y_1 = u, y_2 = v, t = x_k$. Also,

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial x} = v \cos(uv) t \cos(x) + u s \sec^2(x) \cos(uv).$$

Clearly you can continue in this way, taking partial derivatives with respect to any of the other variables.

Example 17.6.3 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

$$\mathbf{g}(x, y) = \begin{pmatrix} x^2 y + z \\ \sin(xy) \end{pmatrix}.$$

Then

$$D\mathbf{g}(x, y, z) = \begin{pmatrix} 2xy & x^2 & 1 \\ y \cos(xy) & x \cos(xy) & 0 \end{pmatrix}.$$

Also $Df(u_1, u_2) = (u_2 \cos(u_1), \sin(u_1))$. Therefore, the derivative is

$$\begin{aligned} & Df(u_1, u_2) D\mathbf{g}(x, y, z) \\ &= (u_2 \cos(u_1), \sin(u_1)) \begin{pmatrix} 2xy & x^2 & 1 \\ y \cos(xy) & x \cos(xy) & 0 \end{pmatrix} \\ &= (2u_2 (\cos u_1) xy + (\sin u_1) y \cos xy, u_2 (\cos u_1) x^2 \\ &\quad + (\sin u_1) x \cos xy, u_2 \cos u_1) \\ &= (w_x, w_y, w_z) \end{aligned}$$

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

$$\mathbf{g}(x, y, z) = \begin{pmatrix} u_1(x, y, z) \\ u_2(x, y, z) \end{pmatrix}$$

then $D(f \circ \mathbf{g})(x, y, z)$ 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 17.6.4 Let $w = f(u_1, u_2, u_3) = u_1^2 + u_3 + u_2$ and

$$\mathbf{g}(x, y, z) = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} x + 2yz \\ x^2 + y \\ z^2 + x \end{pmatrix}$$

Find $\frac{\partial w}{\partial x}$ and $\frac{\partial w}{\partial z}$.

By the chain rule,

$$\begin{aligned}(w_x, w_y, w_z) &= \begin{pmatrix} w_{u_1} & w_{u_2} & w_{u_3} \end{pmatrix} \begin{pmatrix} u_{1x} & u_{1y} & u_{1z} \\ u_{2x} & u_{2y} & u_{2z} \\ u_{3x} & u_{3y} & u_{3z} \end{pmatrix} = \\ & (w_{u_1}u_{1x} + w_{u_2}u_{2x} + w_{u_3}u_{3x}, w_{u_1}u_{1y} + w_{u_2}u_{2y} + w_{u_3}u_{3y}, \\ & w_{u_1}u_{1z} + w_{u_2}u_{2z} + w_{u_3}u_{3z})\end{aligned}$$

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 17.6.5 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,

$$\begin{aligned} D\mathbf{f}(u_1, u_2) &= \begin{pmatrix} 2u_1 & 1 \\ 1 & \cos u_2 \end{pmatrix}, \\ D\mathbf{g}(x_1, x_2, x_3) &= \begin{pmatrix} x_2 & x_1 & 1 \\ 1 & 2x_2 & 0 \end{pmatrix}. \end{aligned}$$

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,

$$\frac{\partial f_1 \circ g}{\partial x_1}(x_1, x_2, x_3) = (2x_1x_2 + 2x_3)x_2 + 1,$$

$$\frac{\partial f_2 \circ g}{\partial x_3}(x_1, x_2, x_3) = 1, \quad \frac{\partial f_2 \circ g}{\partial x_2}(x_1, x_2, x_3) = x_1 + 2x_2(\cos(x_2^2 + x_1)).$$

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

$$\begin{aligned} \frac{\partial z_1}{\partial x_1} &= \frac{\partial z_1}{\partial u_1} \frac{\partial u_1}{\partial x_1} + \frac{\partial z_1}{\partial u_2} \frac{\partial u_2}{\partial x_1} \\ &= 2u_1x_2 + 1 = 2(x_1x_2 + x_3)x_2 + 1. \end{aligned}$$

Example 17.6.6 Let

$$\mathbf{f}(u_1, u_2, u_3) = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} u_1^2 + u_2u_3 \\ u_1^2 + u_2^3 \\ \ln(1 + u_3^2) \end{pmatrix}$$

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 $D(\mathbf{f} \circ \mathbf{g})(x) =$

$$\begin{aligned} & \begin{pmatrix} 2u_1 & u_3 & u_2 \\ 2u_1 & 3u_2^2 & 0 \\ 0 & 0 & \frac{2u_3}{(1+u_3^2)} \end{pmatrix} \begin{pmatrix} 1 & 2x_2 & \cos(x_3) & -\sin(x_4) \\ -1 & 0 & 0 & 2x_4 \\ 0 & 0 & 2x_3 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 2u_1 - u_3 & 4u_1x_2 & 2u_1\cos x_3 + 2u_2x_3 & -2u_1\sin x_4 + 2u_3x_4 + u_2 \\ 2u_1 - 3u_2^2 & 4u_1x_2 & 2u_1\cos x_3 & -2u_1\sin x_4 + 6u_2^2x_4 \\ 0 & 0 & 4\frac{u_3}{1+u_3^2}x_3 & 2\frac{u_3}{1+u_3^2} \end{pmatrix} \quad (17.12) \end{aligned}$$

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 (17.12). Thus $\frac{\partial z}{\partial x_2}$ equals

$$\begin{pmatrix} \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_2}{\partial x_2} \\ \frac{\partial z_3}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 4u_1 x_2 \\ 4u_1 x_2 \\ 0 \end{pmatrix} = \begin{pmatrix} 4(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)) x_2 \\ 4(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)) x_2 \\ 0 \end{pmatrix}$$

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.

17.7 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 $z = f(\mathbf{y}) = (y_1^2 + y_2^2 + \sin y_3 + y_4)$ and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_4 + x_3 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_4 \\ x_4 + x_2 \end{pmatrix}$. Find

$D(f \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2, 3, 4$.

5. Let

$$z = \mathbf{f}(\mathbf{y}) = \begin{pmatrix} y_1^2 + \sin y_2 + \tan y_3 \\ y_1^2 y_2 + y_3 \end{pmatrix}$$

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

$$\det(D\mathbf{f}(\mathbf{f}^{-1}(\mathbf{y}))) \det(D\mathbf{f}^{-1}(\mathbf{y})) = 1$$

17.7.1 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 17.7.1 *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 $v = 7$ and $z = 8$ in terms of g and γ .

This is just an exercise in using the chain rule. Differentiate the two sides with respect to t .

$$\frac{1}{g}v\frac{dv}{dt} + \frac{dz}{dt} + \frac{1}{\gamma}\frac{dP}{dt} = 0.$$

Then when $v = 7$ and $z = 8$, finding $\frac{dP}{dt}$ involves nothing more than solving the following for $\frac{dP}{dt}$.

$$\frac{7}{g}(-3) + 2 + \frac{1}{\gamma}\frac{dP}{dt} = 0$$

Thus

$$\frac{dP}{dt} = \gamma\left(\frac{21}{g} - 2\right)$$

at this instant in time.

Example 17.7.2 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

$$P_x = -\left(\frac{vv_x + z_x g}{g}\right)\gamma$$

Example 17.7.3 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 17.7.4 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 17.7.5 Polar coordinates are

$$x = r \cos \theta, y = r \sin \theta. \quad (17.13)$$

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 (17.13),

$$1 = r_x \cos \theta - (r \sin \theta) \theta_x, \quad 0 = r_x \sin \theta + (r \cos \theta) \theta_x$$

Solving these using Cramer's rule,

$$r_x = \cos(\theta), \quad \theta_x = \frac{-\sin(\theta)}{r}$$

Hence f_x in polar coordinates is

$$f_x = f_r(r, \theta) \cos(\theta) - f_\theta(r, \theta) \left(\frac{\sin(\theta)}{r} \right)$$

17.7.2 The Derivative Of The Inverse Function

Example 17.7.6 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 17.3.6 on Page 303 $DI = I$. 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

$$Df(x)^{-1} = Df^{-1}(y), y = f(x).$$

This is just like a similar situation for functions of one variable. Remember

$$(f^{-1})'(f(x)) = 1/f'(x).$$

In terms of the repeated index summation convention, suppose $y = f(x)$ so that $x = f^{-1}(y)$. Then the above can be written as

$$\delta_{ij} = \frac{\partial x_i}{\partial y_k}(f(x)) \frac{\partial y_k}{\partial x_j}(x).$$

17.7.3 Proof Of The Chain Rule

As in the case of a function of one variable, it is important to consider the derivative of a composition of two functions. As in the case of a function of one variable, this rule is called the chain rule. Its proof depends on the following fundamental lemma. This proof will include the one dimensional case. First let M be a matrix and v a vector of length 1. Then

$$|Mv|^2 = \sum_i \left(\sum_j M_{ij} v_j \right)^2 \leq \sum_i \left(\sum_j |M_{ij}| \right)^2 < \infty$$

Here is the rough idea of the following lemma.

$$\frac{|\mathcal{O}(g(x+v) - g(x))|}{|v|} = \overbrace{\frac{|\mathcal{O}(g(x+v) - g(x))|}{|g(x+v) - g(x)|}}^{\rightarrow 0 \text{ as } v \rightarrow 0} \overbrace{\frac{|g(x+v) - g(x)|}{|v|}}^{\text{bounded}}$$

Lemma 17.7.7 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 $\mathcal{O}(g(x+v) - g(x)) = \mathcal{O}(v)$.

Proof: Let

$$H(v) \equiv \begin{cases} \frac{|\mathcal{O}(g(x+v) - g(x))|}{|g(x+v) - g(x)|} & \text{if } g(x+v) - g(x) \neq 0 \\ 0 & \text{if } g(x+v) - g(x) = 0 \end{cases}$$

Then $\lim_{v \rightarrow 0} H(v) = 0$ because of continuity of g at x and

$$\frac{|\mathcal{O}(g(x+v) - g(x))|}{|v|} = H(v) \frac{|g(x+v) - g(x)|}{|v|}$$

Also

$$\frac{|g(x+v) - g(x)|}{|v|} \leq \frac{|Dg(x)v|}{|v|} + \frac{|\mathcal{O}(v)|}{|v|} = \left| Dg(x) \left(\frac{v}{|v|} \right) \right| + \frac{|\mathcal{O}(v)|}{|v|}$$

which is bounded for small v . Therefore,

$$\lim_{v \rightarrow 0} \frac{|\mathcal{O}(g(x+v) - g(x))|}{|v|} = 0. \blacksquare$$

Recall the notation $f \circ g(x) \equiv f(g(x))$. Thus $f \circ g$ is the name of a function, and this function is defined by what was just written. The following theorem is known as the **chain rule**.

Theorem 17.7.8 (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,

$$D(f \circ g)(x) = Df(g(x)) Dg(x). \quad (17.14)$$

In particular, If $y = g(x)$ so $y_i = g_i(x)$,

$$\frac{\partial (f \circ g)(x)}{\partial x_j} = \sum_{i=1}^p \frac{\partial f(g(x))}{\partial y_i} \frac{\partial g_i(x)}{\partial x_j}. \quad (17.15)$$

Proof: From the assumption that $D\mathbf{f}(\mathbf{g}(\mathbf{x}))$ exists,

$$\begin{aligned}\mathbf{f}(\mathbf{g}(\mathbf{x} + \mathbf{v})) &= \mathbf{f}(\mathbf{g}(\mathbf{x})) + D\mathbf{f}(\mathbf{g}(\mathbf{x}))(\mathbf{g}(\mathbf{x} + \mathbf{v}) - \mathbf{g}(\mathbf{x})) + \mathbf{o}(\mathbf{g}(\mathbf{x} + \mathbf{v}) - \mathbf{g}(\mathbf{x})) \\ &= \mathbf{f}(\mathbf{g}(\mathbf{x})) + D\mathbf{f}(\mathbf{g}(\mathbf{x}))(D\mathbf{g}(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v})) + \mathbf{o}(\mathbf{g}(\mathbf{x} + \mathbf{v}) - \mathbf{g}(\mathbf{x}))\end{aligned}$$

which by Lemma 17.7.7 equals

$$\begin{aligned}&= \mathbf{f}(\mathbf{g}(\mathbf{x})) + D\mathbf{f}(\mathbf{g}(\mathbf{x}))D\mathbf{g}(\mathbf{x})\mathbf{v} + D\mathbf{f}(\mathbf{g}(\mathbf{x}))\mathbf{o}(\mathbf{v}) + \mathbf{o}(\mathbf{v}) \\ &= \mathbf{f}(\mathbf{g}(\mathbf{x})) + D\mathbf{f}(\mathbf{g}(\mathbf{x}))D\mathbf{g}(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v})\end{aligned}$$

and this shows

$$D(\mathbf{f} \circ \mathbf{g})(\mathbf{x}) = D\mathbf{f}(\mathbf{g}(\mathbf{x}))D\mathbf{g}(\mathbf{x})$$

from the definition of the derivative and its uniqueness established in Theorem 17.3.3 on Page 301. ■

17.8 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 17.7.6 on Page 318. 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 17.7.5 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 17.7.5, 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.

17.9 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 $\mathbf{x} \in U$,

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \sum_{j=1}^n \frac{\partial f(\mathbf{x})}{\partial x_j} v_j + o(\mathbf{v}). \quad (17.16)$$

Now we can prove the formula for the directional derivative in terms of the gradient.

Proposition 17.9.1 *If f is differentiable at \mathbf{x} and for \mathbf{v} a unit vector*

$$D_{\mathbf{v}}f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}. \quad (17.17)$$

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 17.9.2 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 (17.17) 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 17.9.3 *Let $f : U \rightarrow \mathbb{R}$ be a differentiable function and let $\mathbf{x} \in U$. Then*

$$\max \{D_{\mathbf{v}}f(\mathbf{x}) : |\mathbf{v}| = 1\} = |\nabla f(\mathbf{x})| \quad (17.18)$$

and

$$\min \{D_{\mathbf{v}}f(\mathbf{x}) : |\mathbf{v}| = 1\} = -|\nabla f(\mathbf{x})|. \quad (17.19)$$

Furthermore, the maximum in (17.18) occurs when $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$ and the minimum in (17.19) occurs when $\mathbf{v} = -\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$.

Proof: From (17.17) 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

$$\begin{aligned} D_{\mathbf{v}}f(\mathbf{x}) &= \nabla f(\mathbf{x}) \cdot (-\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|) \\ &= -|\nabla f(\mathbf{x})|^2 / |\nabla f(\mathbf{x})| = -|\nabla f(\mathbf{x})| \end{aligned}$$

while if $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$, then

$$\begin{aligned} D_{\mathbf{v}}f(\mathbf{x}) &= \nabla f(\mathbf{x}) \cdot (\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|) \\ &= |\nabla f(\mathbf{x})|^2 / |\nabla f(\mathbf{x})| = |\nabla f(\mathbf{x})|. \quad \blacksquare \end{aligned}$$

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

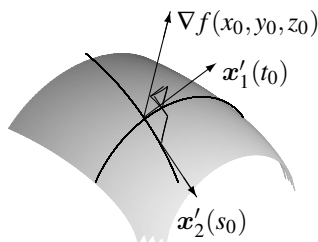
direction which involves the maximum rate of decrease in temperature. This expectation will be realized by taking $\mathbf{J} = -K\nabla u$ where K is a positive scalar function which can depend on a variety of things. The above relation between the heat flux and ∇u is usually called the Fourier heat 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.

17.10 The Gradient And Tangent Planes

The gradient has fundamental geometric significance illustrated by the following picture.



In this picture, the surface is a piece of a level surface of a function of three variables $f(x, y, z)$. Thus the surface is defined by $f(x, y, z) = c$ or more completely as

$$\{(x, y, z) : f(x, y, z) = c\}$$

For example, if $f(x, y, z) = x^2 + y^2 + z^2$, this would be a piece of a sphere. There are two smooth curves in this picture which lie in the surface having parameterizations, $\mathbf{x}_1(t) = (x_1(t), y_1(t), z_1(t))$ and $\mathbf{x}_2(s) = (x_2(s), y_2(s), z_2(s))$ which intersect at the point (x_0, y_0, z_0) on this surface.¹ This intersection occurs when $t = t_0$ and $s = s_0$. Since the points $\mathbf{x}_1(t)$ for t in an interval lie in the level surface, it follows

$$f(x_1(t), y_1(t), z_1(t)) = c$$

¹Do there exist any smooth curves which lie in the level surface of f and pass through the point (x_0, y_0, z_0) ? It turns out there do if $\nabla f(x_0, y_0, z_0) \neq \mathbf{0}$ and if the function f , is C^1 . However, this is a consequence of the implicit function theorem, one of the greatest theorems in all mathematics and a topic for an advanced calculus class. An elementary presentation is presented later.

for all t in some interval. Therefore, taking the derivative of both sides and using the chain rule on the left,

$$\frac{\partial f}{\partial x}(x_1(t), y_1(t), z_1(t))x_1'(t) + \frac{\partial f}{\partial y}(x_1(t), y_1(t), z_1(t))y_1'(t) + \frac{\partial f}{\partial z}(x_1(t), y_1(t), z_1(t))z_1'(t) = 0.$$

In terms of the gradient, this merely states

$$\nabla f(x_1(t), y_1(t), z_1(t)) \cdot \mathbf{x}_1'(t) = 0.$$

Similarly,

$$\nabla f(x_2(s), y_2(s), z_2(s)) \cdot \mathbf{x}_2'(s) = 0.$$

Letting $s = s_0$ and $t = t_0$, it follows

$$\nabla f(x_0, y_0, z_0) \cdot \mathbf{x}_1'(t_0) = 0, \quad \nabla f(x_0, y_0, z_0) \cdot \mathbf{x}_2'(s_0) = 0.$$

It follows $\nabla f(x_0, y_0, z_0)$ is perpendicular to both the direction vectors of the two indicated curves shown. Surely if things are as they should be, these two direction vectors would determine a plane which deserves to be called the tangent plane to the level surface of f at the point (x_0, y_0, z_0) and that $\nabla f(x_0, y_0, z_0)$ is perpendicular to this tangent plane at the point (x_0, y_0, z_0) .

Example 17.10.1 Find the equation of the tangent plane to the level surface

$$f(x, y, z) = 6$$

of the function $f(x, y, z) = x^2 + 2y^2 + 3z^2$ at the point $(1, 1, 1)$.

First note that $(1, 1, 1)$ is a point on this level surface. To find the desired plane it suffices to find the normal vector to the proposed plane. But $\nabla f(x, y, z) = (2x, 4y, 6z)$ and so $\nabla f(1, 1, 1) = (2, 4, 6)$. Therefore, from this problem, the equation of the plane is $(2, 4, 6) \cdot (x - 1, y - 1, z - 1) = 0$ or in other words, $2x - 12 + 4y + 6z = 0$.

Example 17.10.2 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 17.10.3 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 the direction vector of 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

$$(1, \sqrt{3}, 4) + t(2, 2\sqrt{3}, -1).$$

17.11 Exercises

- Find the gradient of $f =$
 - $x^2y + z^3$ at $(1, 1, 2)$
 - $z \sin(x^2y) + 2^{x+y}$ at $(1, 1, 0)$
 - $u \ln(x + y + z^2 + w)$ at $(x, y, z, w, u) = (1, 1, 1, 1, 2)$
 - $\sin(xy) + z^3$ at $(1, \pi, 1)$
 - $\ln(x + y^2)z$
 - $z \ln(4 + \sin(xy))$ at the point $(0, \pi, 1)$
- Find the directional derivatives of f at the indicated point in the direction $\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}}\right)$.
 - $x^2y + z^3$ at $(1, 1, 1)$
 - $z \sin(x^2y) + 2^{x+y}$ at $(1, 1, 0)$
 - $xy + z^2 + 1$ at $(1, 2, 3)$
 - $\sin(xy) + z$ at $(0, 1, 1)$
 - $x^y + z$ at $(1, 1, 1)$.
 - $\sin(\sin(x + y)) + z$ at the point $(1, 0, 1)$.
- Find the directional derivatives of the given function at the indicated point in the indicated direction.
 - $\sin(x^2 + y) + z^2$ at $(0, \pi/2, 1)$ in direction of $(1, 1, 2)$.
 - $x^{(x+y)} + \sin(zx)$ at $(1, 0, 0)$ in the direction of $(2, -1, 0)$.
 - $z^{\sin(x)} + y$ at $(0, 1, 1)$ in the direction of $(1, 1, 3)$.
- Find the tangent plane to the indicated level surface at the indicated point.
 - $x^2y + z^3 = 2$ at $(1, 1, 1)$
 - $z \sin(x^2y) + 2^{x+y} = 2 \sin 1 + 4$ at $(1, 1, 2)$
 - $\cos(x) + z \sin(x + y) = 1$ at $(-\pi, \frac{3\pi}{2}, 2)$
- 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.
- The level surfaces $x^2 + y^2 + z^2 = 4$ and $z + x^2 + y^2 = 4$ have the point $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 1)$ 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. The point $(1, 1, \sqrt{2})$ is on the level surface $x^2 + y^2 + z^2 = 4$ and the level surface $y^2 + 2z^2 = 5$. Find an equation for the line tangent to the curve of intersection of these two surfaces at this point.
9. *In a slightly more general setting, 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 as t varies, the point determined by $(x(t), y(t), z(t))$ moves over the curve.

Chapter 18

Optimization

18.1 Local Extrema

The following definition describes what is meant by a local maximum or local minimum.

Definition 18.1.1 Suppose $f : D(f) \rightarrow \mathbb{R}$ where $D(f) \subseteq \mathbb{R}^n$. A point $\mathbf{x} \in D(f) \subseteq \mathbb{R}^n$ 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 18.1.2 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}^n and suppose \mathbf{x} is a local maximum (minimum) for f . Then consider the real valued function of one variable, $h(t) \equiv f(\mathbf{x} + t\mathbf{v})$ for small $|t|$. Since 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,

$$Df(\mathbf{x}) = \begin{pmatrix} f_{x_1}(\mathbf{x}) & \cdots & f_{x_n}(\mathbf{x}) \end{pmatrix}$$

and so $\nabla f(\mathbf{x}) = 0$. This proves the following theorem.

Theorem 18.1.3 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 18.1.4 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 18.1.5 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 18.1.6 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) is $(x_0, y_0, \sqrt{4 - x_0^2 - y_0^2})$ 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 18.1.7 An open box is to contain 32 cubic feet. Find the dimensions which will result in the least surface area.

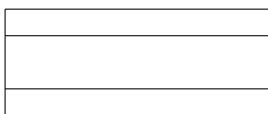
Let the height of the box be z and the length and width be x and y respectively. Then $xyz = 32$ and so $z = 32/xy$. The total area is $xy + 2xz + 2yz$ and so in terms of the two variables x and y , the area is $A = xy + \frac{64}{y} + \frac{64}{x}$. To find best dimensions you note these must result in a local minimum.

$$A_x = \frac{yx^2 - 64}{x^2} = 0, \quad A_y = \frac{xy^2 - 64}{y^2}.$$

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?

18.2 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.
- The function $f(x, y) = 2x^2 + y^2$ is defined on the disk $x^2 + y^2 \leq 1$. Find its maximum value.
- Find the point on the surface $z = x^2 + y + 1$ which is closest to $(0, 0, 0)$.
- 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

$$f(x, y) = \begin{cases} -1 & \text{if } y = x^2, x \neq 0 \\ (y - x^2)^2 & \text{if } y \neq x^2 \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

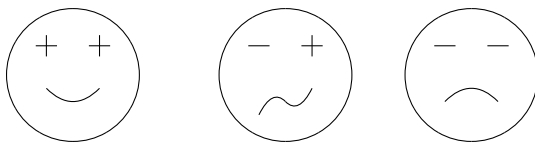
Show that $\nabla f(0, 0) = \mathbf{0}$. Now show that if (a, b) is any nonzero unit vector, the function $t \rightarrow f(ta, tb)$ has a local minimum of 0 when $t = 0$. Thus in every direction, this function has a local minimum at $(0, 0)$ but the function f does not have a local minimum at $(0, 0)$.

18.3 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 18.3.1 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 18.3.2 Let $f : U \rightarrow \mathbb{R}$ for U an open set in \mathbb{R}^n 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**.

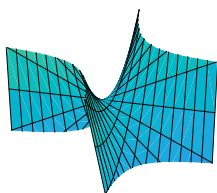
Here is an example.

Example 18.3.3 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 18.3.4 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.

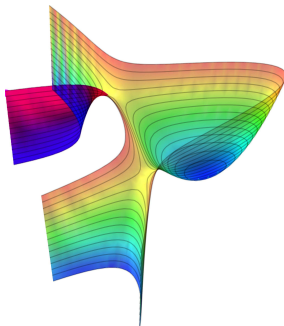
Next consider $(0, -1)$ at this point the Hessian matrix is $\begin{pmatrix} 4 & -12 \\ -12 & 4 \end{pmatrix}$ and the eigenvalues are 16, -8 . Therefore, this point is a saddle point. To determine this, find the eigenvalues.

$$\det\left(\lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 4 & -12 \\ -12 & 4 \end{pmatrix}\right) = \lambda^2 - 8\lambda - 128 = (\lambda + 8)(\lambda - 16)$$

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.

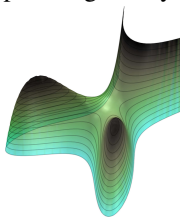
Below is a graph of this function which illustrates the behavior near saddle points.



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 18.3.5 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 18.3.6 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.

18.4 Exercises

1. Use the second derivative test on the critical points $(1, 1)$, and $(1, -1)$ for Example 18.3.5. The function is $6xy^2 - 2x^3 - 3x^4$.
2. If $H = H^T$ and $Hx = \lambda x$ while $Hx = \mu x$ for $\lambda \neq \mu$, show that $x \cdot y = 0$.
3. 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.

$$f(x, y) = -x^4 + 2x^3 + 39x^2 + 10yx^2 - 10yx - 40x - y^2 - 8y - 16.$$
4. 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.

$$f(x, y) = -3x^4 + 6x^3 + 37x^2 + 10yx^2 - 10yx - 40x - 3y^2 - 24y - 48.$$
5. 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.

$$f(x, y) = 5x^4 - 10x^3 + 17x^2 - 6yx^2 + 6yx - 12x + 5y^2 - 20y + 20.$$
6. 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.

$$f(x, y) = 4x^4 - 8x^3 - 4yx^2 + 4yx + 8x - 4x^2 + 4y^2 + 16y + 16.$$
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.

$$f(x, y, z) = \frac{1}{3}x^2 + \frac{32}{3}x + \frac{4}{3} - \frac{16}{3}yx - \frac{58}{3}y - \frac{4}{3}zx - \frac{46}{3}z + \frac{1}{3}y^2 - \frac{4}{3}zy - \frac{5}{3}z^2.$$
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.

$$f(x, y, z) = -\frac{5}{3}x^2 + \frac{2}{3}x - \frac{2}{3} + \frac{8}{3}yx + \frac{2}{3}y + \frac{14}{3}zx - \frac{28}{3}z - \frac{5}{3}y^2 + \frac{14}{3}zy - \frac{8}{3}z^2.$$
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.

$$f(x, y, z) = -\frac{11}{3}x^2 + \frac{40}{3}x - \frac{56}{3} + \frac{8}{3}yx + \frac{10}{3}y - \frac{4}{3}zx + \frac{22}{3}z - \frac{11}{3}y^2 - \frac{4}{3}zy - \frac{5}{3}z^2.$$

10. 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.

$$f(x, y, z) = -\frac{2}{3}x^2 + \frac{28}{3}x + \frac{37}{3} + \frac{14}{3}yx + \frac{10}{3}y - \frac{4}{3}zx - \frac{26}{3}z - \frac{2}{3}y^2 - \frac{4}{3}zy + \frac{7}{3}z^2.$$

11. *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.

12. 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.

13. 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.

$$f(x, y) = 2x^4 - 4x^3 + 42x^2 + 8yx^2 - 8yx - 40x + 2y^2 + 20y + 50.$$

14. 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.

$$f(x, y) = 4x^4 - 16x^3 - 4x^2 - 4yx^2 + 8yx + 40x + 4y^2 + 40y + 100.$$

15. 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.

$$f(x, y) = 5x^4 - 30x^3 + 45x^2 + 6yx^2 - 18yx + 5y^2.$$

16. Find the critical points of the following function of three variables and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = \frac{10}{3}x^2 - \frac{44}{3}x + \frac{64}{3} - \frac{10}{3}yx + \frac{16}{3}y + \frac{2}{3}zx - \frac{20}{3}z + \frac{10}{3}y^2 + \frac{2}{3}zy + \frac{4}{3}z^2.$$

17. Find the critical points of the following function of three variables and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = -\frac{7}{3}x^2 - \frac{146}{3}x + \frac{83}{3} + \frac{16}{3}yx + \frac{4}{3}y - \frac{14}{3}zx + \frac{94}{3}z - \frac{7}{3}y^2 - \frac{14}{3}zy + \frac{8}{3}z^2.$$

18. Find the critical points of the following function of three variables and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = \frac{2}{3}x^2 + 4x + 75 - \frac{14}{3}yx - 38y - \frac{8}{3}zx - 2z + \frac{2}{3}y^2 - \frac{8}{3}zy - \frac{1}{3}z^2.$$

19. Find the critical points of the following function of three variables and classify them as local minima, local maxima or saddle points.

$$f(x, y, z) = 4x^2 - 30x + 510 - 2yx + 60y - 2zx - 70z + 4y^2 - 2zy + 4z^2.$$

20. 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.

$$f(x, y, z) = -\frac{1}{6}x^4 + \frac{5}{3}x^3 - \frac{25}{6}x^2 + \frac{10}{3}yx^2 - \frac{50}{3}yx + \frac{19}{3}zx^2 - \frac{95}{3}zx - \frac{5}{3}y^2 - \frac{10}{3}zy - \frac{1}{6}z^2.$$

21. 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.

$$f(x, y, z) = -\frac{3}{2}x^4 + 6x^3 - 6x^2 + zx^2 - 2zx - 2y^2 - 12y - 18 - \frac{3}{2}z^2.$$

22. 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.

23. 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.

24. Suppose $f(x, y)$, a function of two variables defined on all \mathbb{R}^n 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 $f = -1$.

18.5 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 18.5.1 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.

Example 18.5.2 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,

$$(y, x + 1) - \lambda (2x, 2y) = 0.$$

¹There exists such a point because the sphere is closed and bounded.

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 $(-1, 0), (\frac{1}{2}, \frac{\sqrt{3}}{2}), (\frac{1}{2}, -\frac{\sqrt{3}}{2})$. 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 $(\frac{1}{2}, \frac{\sqrt{3}}{2})$. The minimum value is $-\frac{3\sqrt{3}}{4}$ and it occurs at $(\frac{1}{2}, -\frac{\sqrt{3}}{2})$.

Example 18.5.3 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

$$\left(\frac{1}{2}, \frac{3}{2}\right), \text{ and } \left(-\frac{1}{2}, -\frac{3}{2}\right). \quad (18.1)$$

$$f\left(\frac{1}{2}, \frac{3}{2}\right) = \left(\frac{1}{2}\right)\left(\frac{3}{2}\right) - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$f\left(-\frac{1}{2}, -\frac{3}{2}\right) = \left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right) - \left(-\frac{1}{2}\right)^2 = \frac{1}{2}$$

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 18.5.4 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 18.5.5 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 18.5.6 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

$$g_i(\mathbf{x}_0) = 0, \quad i = 1, \dots, m, \quad g_i \text{ a } C^1 \text{ function}, \quad (18.2)$$

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

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \dots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \quad (18.3)$$

holds.

To help remember how to use 18.3, 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 18.3 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 18.5.7 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 $-\frac{2}{3}\sqrt{30} - \frac{2}{3}\sqrt{3}$.

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 .

18.6 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 $(0, 1)$. Therefore, you need to consider these points also.

18.7 Proof Of The Second Derivative Test*

A version of the following theorem is due to Lagrange, about 1790.

Theorem 18.7.1 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.)

Proof: There exists K such that

$$f(x) - \left(f(c) + \sum_{k=1}^n \frac{f^{(k)}(c)}{k!} (x-c)^k + K(x-c)^{n+1} \right) = 0 \quad (18.4)$$

In fact,

$$K = \frac{-f(x) + \left(f(c) + \sum_{k=1}^n \frac{f^{(k)}(c)}{k!} (x-c)^k \right)}{(x-c)^{n+1}}.$$

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 18.4 got replaced by t .

Therefore, $F(c) = 0$ by the way K was chosen and also $F(x) = 0$. Then $F'(t) =$

$$\begin{aligned}
 & - \left(f'(t) - \left(\sum_{k=1}^n \frac{f^{(k)}(t)}{k!} k (x-t)^{k-1} - \sum_{k=1}^n \frac{f^{(k+1)}(t)}{k!} (x-t)^k \right) \right) \\
 & = - \left(f'(t) - \left(\sum_{k=0}^{n-1} \frac{f^{(k+1)}(t)}{k!} (x-t)^k - \sum_{k=1}^n \frac{f^{(k+1)}(t)}{k!} (x-t)^k \right) \right) \\
 & = - \left(f'(t) - \left(f^{(n+1)}(t) (x-t)^n + K(n+1) (x-t)^n \right) \right) \\
 & = -f'(t) + f'(t) - f^{(n+1)}(t) (x-t)^n + K(n+1) (x-t)^n \\
 & = -f^{(n+1)}(t) \frac{1}{n!} (x-t)^n + K(n+1) (x-t)^n
 \end{aligned}$$

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

$$\begin{aligned}
 K(n+1) &= f^{(n+1)}(\xi) \frac{1}{n!} \\
 K &= \frac{f^{(n+1)}(\xi)}{(n+1)!} \blacksquare
 \end{aligned}$$

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.

Definition 18.7.2 The matrix $\left(\frac{\partial^2 f}{\partial x_i \partial x_j} (x) \right)$ is called the Hessian matrix, denoted by $H(x)$.

Now recall the Taylor formula with the Lagrange form of the remainder.

Theorem 18.7.3 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}^n . Suppose $f \in C^2(U)$. Let $x \in U$ and let $r > 0$ be such that

$$B(x, r) \subseteq U.$$

Then for $\|v\| < r$ consider

$$f(x+tv) - f(x) \equiv h(t)$$

for $t \in [0, 1]$. Then from Taylor's theorem for the case where $m = 2$ and the chain rule, using the repeated index summation convention and the chain rule,

$$h'(t) = \frac{\partial f}{\partial x_i} (x+tv) v_i, \quad h''(t) = \frac{\partial^2 f}{\partial x_j \partial x_i} (x+tv) v_i v_j.$$

Thus

$$h''(t) = \mathbf{v}^T H(\mathbf{x} + t\mathbf{v}) \mathbf{v}.$$

From Theorem 18.7.3 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

$$\begin{aligned} f(\mathbf{x} + \mathbf{v}) &= f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{v} + \frac{1}{2} \mathbf{v}^T H(\mathbf{x}) \mathbf{v} + \\ &\quad \frac{1}{2} (\mathbf{v}^T (H(\mathbf{x} + t\mathbf{v}) - H(\mathbf{x})) \mathbf{v}) \end{aligned} \quad (18.5)$$

where the last term satisfies

$$\lim_{|\mathbf{v}| \rightarrow 0} \frac{1}{2} \frac{(\mathbf{v}^T (H(\mathbf{x} + t\mathbf{v}) - H(\mathbf{x})) \mathbf{v})}{|\mathbf{v}|^2} = 0 \quad (18.6)$$

because of the continuity of the entries of $H(\mathbf{x})$.

Theorem 18.7.4 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 (18.5) implies

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \frac{1}{2} \mathbf{v}^T H(\mathbf{x}) \mathbf{v} + \frac{1}{2} (\mathbf{v}^T (H(\mathbf{x} + t\mathbf{v}) - H(\mathbf{x})) \mathbf{v}) \quad (18.7)$$

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 11.4.7, $H(\mathbf{x})$ 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 (18.7), replacing \mathbf{v} with $s\mathbf{v}$ and letting t depend on s ,

$$f(\mathbf{x} + s\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2} s^2 \mathbf{v}^T H(\mathbf{x}) \mathbf{v} +$$

$$\frac{1}{2}s^2 (\mathbf{v}^T (H(\mathbf{x}+t\mathbf{s}\mathbf{v}) - H(\mathbf{x})) \mathbf{v})$$

which implies

$$\begin{aligned} f(\mathbf{x}+s\mathbf{v}) &= f(\mathbf{x}) + \frac{1}{2}s^2\lambda^2|\mathbf{v}|^2 + \frac{1}{2}s^2 (\mathbf{v}^T (H(\mathbf{x}+t\mathbf{s}\mathbf{v}) - H(\mathbf{x})) \mathbf{v}) \\ &\geq f(\mathbf{x}) + \frac{1}{4}s^2\lambda^2|\mathbf{v}|^2 \end{aligned}$$

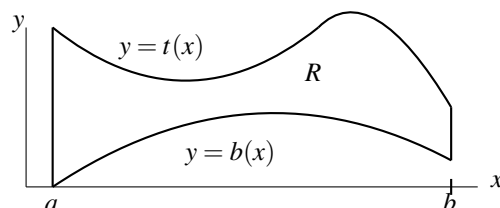
whenever s is small enough. Thus in the direction \mathbf{v} the function has a local minimum at \mathbf{x} . The assertion about the local maximum in some direction follows similarly. ■

Chapter 19

The Riemannnn Integral On \mathbb{R}^n

19.1 Methods For Double Integrals

This chapter is on the Riemannnn integral for a function of n 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 19.1.1 Suppose $t(x) = 4 - x^2$, $b(x) = 0$ and $a = -2, b = 2$. Compute the iterated integral described above for $f(x, y) = xy + y$.

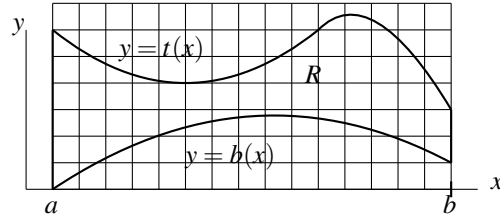
You should sketch the graphs of these functions. Filling in the limits as above, we obtain

$$\int_{-2}^2 \int_0^{4-x^2} (xy+y) dy dx = \int_{-2}^2 \frac{1}{2} (x^2-4)^2 (x+1) dx = \frac{256}{15}$$

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 $\int_a^b \int_{b(x)}^{t(x)} f(x, y) dy dx =$

$$\begin{aligned} \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \int_{b(x)}^{t(x)} f(x, y) dy dx &\approx \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \int_{b(x_i)}^{t(x_i)} f(x_i, y) dy dx \\ &\approx \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \sum_{j=m(i)}^{M(i)} f(x_i, y_{ij}) \Delta y dx \\ &= \sum_{i=1}^n \sum_{j=m(i)}^{M(i)} f(x_i, y_{ij}) \Delta y \Delta x \end{aligned}$$

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 “Riemannn 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 Riemannnn integral.

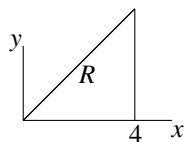
Definition 19.1.2 Let R be a bounded region in the xy plane and let f be a bounded function defined on R . We say f is Riemannnn integrable if there exists a number, denoted by $\int_R f dA$ and called the Riemannnn integral such that if $\epsilon > 0$ is given, then whenever one imposes a sufficiently 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

$$\left| \int_R f dA - \sum_i f(x_i, y_i) \text{area}(Q_i) \right| < \epsilon$$

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 Riemannnn integral. It is not for the faint of heart. It is only there for those who have a compelling need to understand all the details.

Example 19.1.3 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.

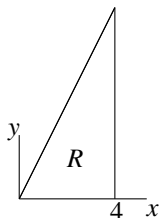
$$\int_y^4 (x^2y + yx) dx = \frac{88}{3}y - \frac{1}{3}y^4 - \frac{1}{2}y^3$$

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 19.1.4 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

$$\int_{y/2}^4 (x^2y) dx = \frac{64}{3}y - \frac{1}{24}y^4$$

and so

$$\int_R f dA = \int_0^8 \left(\frac{64}{3}y - \frac{1}{24}y^4 \right) dy = \frac{2048}{5}$$

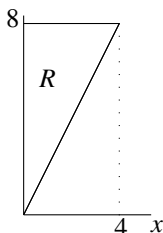
the same answer.

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 19.1.5 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.

19.1.1 Density And Mass

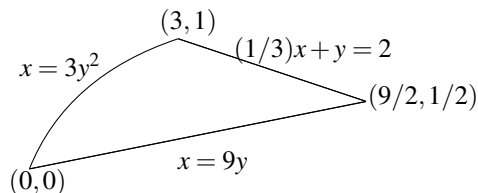
Consider a two dimensional material. Of course there is no such thing but a flat plate might be modeled as one. The density ρ is a function of position and is defined as follows. Consider a small chunk of area dA located at the point whose Cartesian coordinates are (x,y) . Then the mass of this small 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 19.1.6 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.

19.2 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. $\int_{-3}^3 \int_{-x}^x x^2 dy dx$
 Your answer for the iterated integral should be $\int_3^0 \int_{-3}^{-y} x^2 dx dy + \int_0^{-3} \int_{-3}^y x^2 dx dy + \int_0^3 \int_y^3 x^2 dx dy + \int_{-3}^0 \int_{-y}^3 x^2 dx dy$. 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. $\int_{-2}^2 \int_{-x}^x x^2 dy dx$.

19.3 Methods For Triple Integrals

19.3.1 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 .

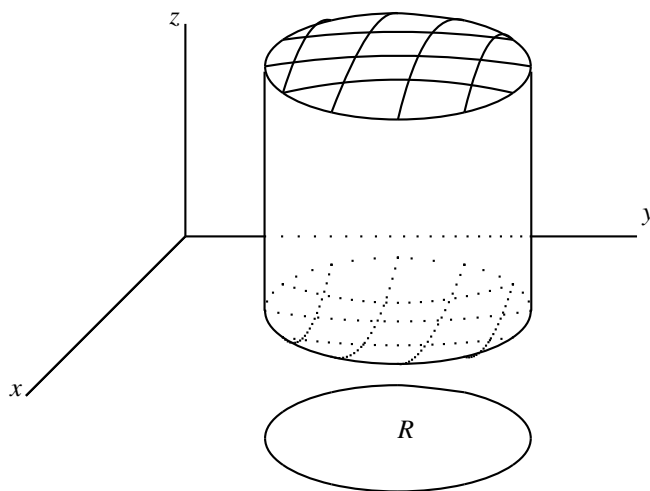
Definition 19.3.1 Let R be a bounded region in the \mathbb{R}^3 and let f be a bounded function defined on R . We say f is Riemannnn integrable if there exists a number, denoted by $\int_R f dV$ and called the Riemannnn integral such that if $\varepsilon > 0$ is given, then whenever one imposes a sufficiently 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

$$\left| \int_R f dV - \sum_i f(x_i, y_i, z_i) \text{volume}(Q_i) \right| < \varepsilon$$

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 Riemannnn integral of a function of three or more variables.

19.3.2 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

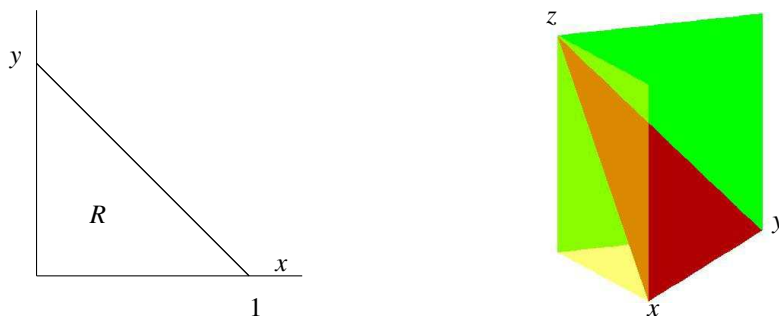
$$\int_E f dV = \int_R \int_{a(x,y)}^{b(x,y)} f(x, y, z) dz dA$$

It might be helpful to think of $dV = dz dA$. 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 func-

tion 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 19.3.2 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 19.3.3 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 19.3.4 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

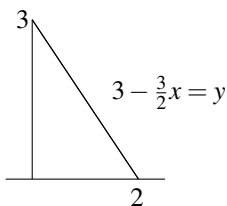
$$\begin{aligned} & \int_0^\pi \int_0^{3y} \int_0^{y+z} \cos(x + y) dx dz dy \\ &= \int_0^\pi (-1 - 16 \cos^5 y + 20 \cos^3 y - 5 \cos y - 3 (\sin y) y + 2 \cos^2 y) dy = -3\pi \end{aligned}$$

Example 19.3.5 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

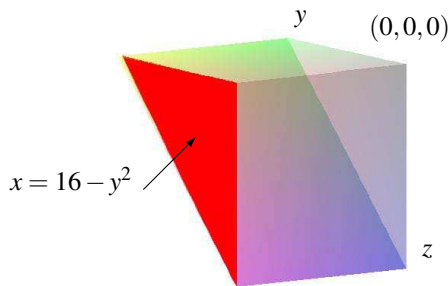
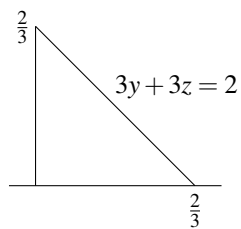
$$\int_0^3 \int_0^{\frac{2}{3}(3-y)} f(x,y) dx dy.$$

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 19.3.6 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

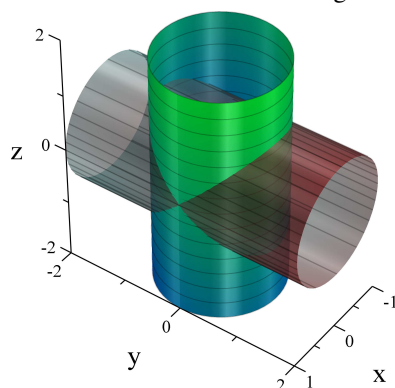
$$\int_0^{\frac{2}{3}} \int_0^{\frac{2}{3}-y} \int_0^{16-y^2} dx dz dy = \frac{860}{243}$$

Example 19.3.7 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.

19.4 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.

$$\int_0^1 \int_0^z \int_0^z f(x,y,z) dx dy dz = \int_?^? \int_?^? \int_?^? f(x,y,z) dx dz dy,$$

$$\begin{aligned}\int_0^1 \int_0^z \int_0^{2z} f(x, y, z) \, dx dy dz &= \int_0^2 \int_0^2 \int_0^2 f(x, y, z) \, dy dz dx, \\ \int_0^1 \int_0^z \int_0^z f(x, y, z) \, dx dy dz &= \int_0^2 \int_0^2 \int_0^2 f(x, y, z) \, dz dy dx, \\ \int_0^1 \int_{z/2}^{\sqrt{z}} \int_0^{y+z} f(x, y, z) \, dx dy dz &= \int_0^2 \int_0^2 \int_0^2 f(x, y, z) \, dx dz dy, \\ \int_4^6 \int_2^6 \int_0^4 f(x, y, z) \, dx dy dz &= \int_0^2 \int_0^2 \int_0^2 f(x, y, z) \, dz dy dx.\end{aligned}$$

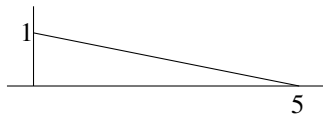
3. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x + y + \frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$.
4. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x + \frac{1}{2}y + \frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$.
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 $x \geq 0$.

19.4.1 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 19.4.1 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 19.4.2 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 19.4.3 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 $\rho(x, y, z) = z$

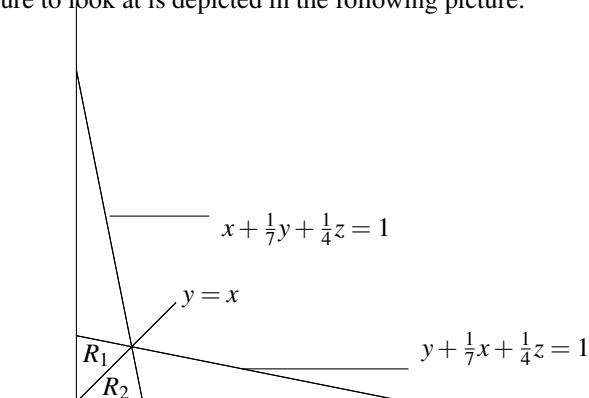
When $z = 0$ the surface $z = 9 - x^2 - y^2$ intersects the xy plane in a circle of radius 3 centered at $(0, 0)$. Since $x, y \geq 0$, it is only a quarter of a circle of interest, the part where both these variables are nonnegative. For each (x, y) inside this quarter circle, z goes from 0 to $9 - x^2 - y^2$. 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 19.4.4 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

$$\int_0^{7/8} \int_x^{1-\frac{x}{7}} \int_0^{4(1-\frac{1}{7}x-y)} dz dy dx + \int_0^{7/8} \int_y^{1-\frac{y}{7}} \int_0^{4(1-\frac{1}{7}y-x)} dz dx dy = \frac{7}{6}$$

19.5 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 20

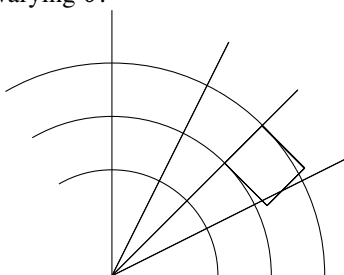
The Integral In Other Coordinates

20.1 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 20.1.1 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 20.1.2 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

$$\int_0^{2\pi} \int_0^a \overbrace{r dr d\theta}^{dA} = \frac{2}{3} \pi a^3$$

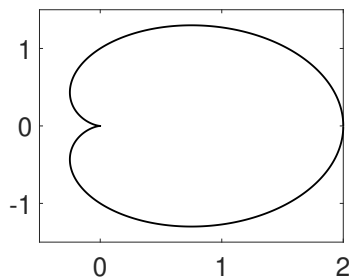
Notice how in these examples the circular disk is really a rectangle $[0, 2\pi] \times [0, a]$. This is why polar coordinates are so useful. The next example was worked earlier from a different point of view.

Example 20.1.3 Find the area of the inside of the cardioid $r = 1 + \cos \theta$, $\theta \in [0, 2\pi]$.

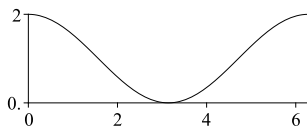
Here the integral is

$$\int_0^{2\pi} \int_0^{1+\cos(\theta)} r dr d\theta = \frac{3}{2} \pi$$

To see how impossible this problem is in rectangular coordinates, draw the graph of the cardioid.



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.



Example 20.1.4 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 $r dr d\theta$ 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 20.1.5 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 20.1.6 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

$$\int_0^{2\pi} \int_0^{2+\sin(\theta)} (3 + r \cos(\theta)) r dr d\theta = \frac{27}{2}\pi$$

20.2 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.

20.3 Cylindrical And Spherical Coordinates

Cylindrical coordinates are defined as follows.

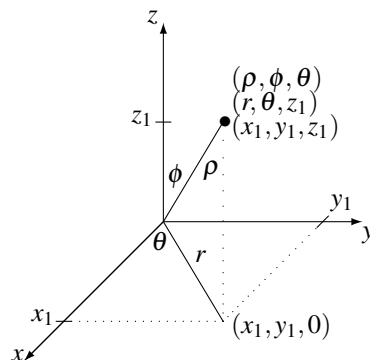
$$\begin{aligned} \mathbf{x}(r, \theta, z) &\equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r \cos(\theta) \\ r \sin(\theta) \\ z \end{pmatrix}, \\ r &\geq 0, \theta \in [0, 2\pi), z \in \mathbb{R} \end{aligned}$$

Spherical coordinates are a little harder. These are given by

$$\mathbf{x}(\rho, \theta, \phi) \equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix},$$

$$\rho \geq 0, \theta \in [0, 2\pi), \phi \in [0, \pi]$$

The following picture relates the various coordinates.

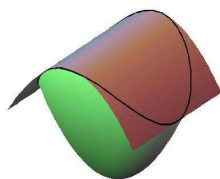


In this picture, ρ is the distance between the origin, the point whose Cartesian coordinates are $(0,0,0)$ and the point indicated by a dot and labelled as (x_1, y_1, z_1) , (r, θ, z_1) , and (ρ, ϕ, θ) . The angle between the positive z axis and the line between the origin and the point indicated by 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 $\rho \sin \phi = r$.

20.3.1 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 20.3.1 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

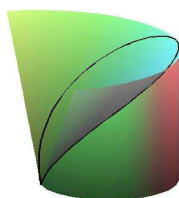
$$\int_0^{2\pi} \int_0^1 \int_{4(r^2 \cos^2(\theta)) + 2(r^2 \sin^2(\theta))}^{4-2(r^2 \sin^2(\theta))} dz r dr d\theta = 2\pi$$

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 20.3.2 Find the volume of the three dimensional region between the graphs of $z = 0$, $z = \sqrt{x^2 + y^2}$, and the cylinder $(x - 1)^2 + y^2 = 1$.

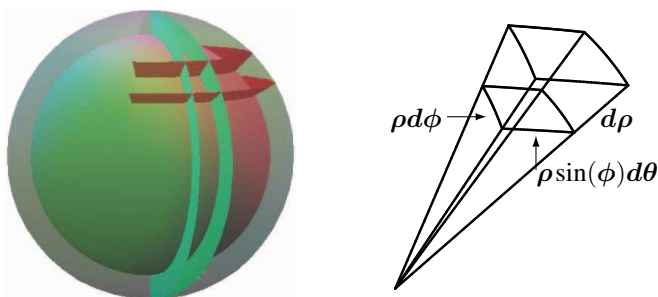


Consider the cylinder. It reduces to $r^2 = 2r \cos \theta$ or more simply $r = 2 \cos \theta$. This is the graph of a circle having radius 1 and centered at $(1, 0)$. Therefore, $\theta \in [-\pi/2, \pi/2]$. It follows that the cylindrical coordinate description of this volume is

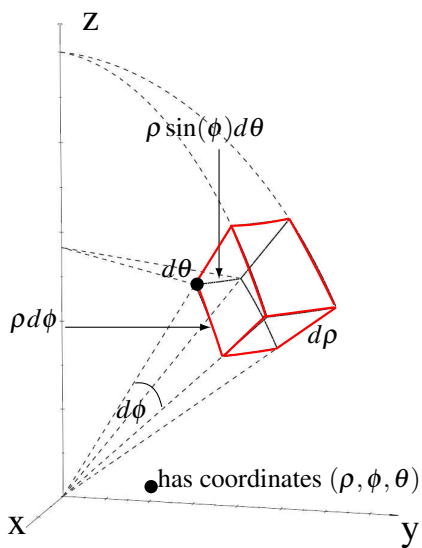
$$\int_{-\pi/2}^{\pi/2} \int_0^{2 \cos \theta} \int_0^r dz r dr d\theta = \frac{32}{9}$$

20.3.2 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 20.3.3 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

$$\begin{aligned} \int_{B_R} 1 dV &= \int_U \rho^2 \sin \phi dV \\ &= \int_0^R \int_0^\pi \int_0^{2\pi} \rho^2 \sin \phi d\theta d\phi d\rho = \frac{4}{3} R^3 \pi. \end{aligned}$$

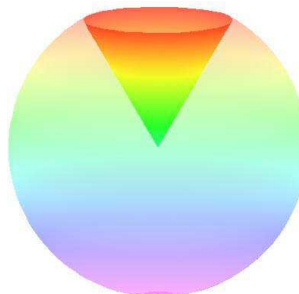
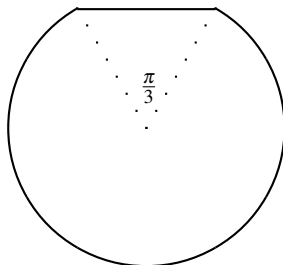
The reason this was effortless, is that the ball, B_R is realized as a box in terms of the spherical coordinates. Remember what was pointed out earlier about setting up iterated integrals over boxes.

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.

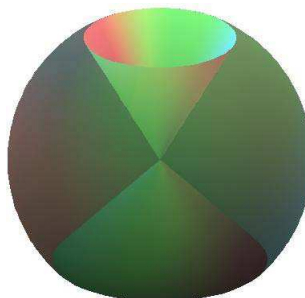
Example 20.3.4 A cone is cut out of a ball of radius R as shown in the following picture, the diagram on the left being a side view. The angle of the cone is $\pi/3$. Find the volume of what is left.



Use spherical coordinates. This volume is then

$$\int_{\pi/6}^{\pi} \int_0^{2\pi} \int_0^R \rho^2 \sin(\phi) d\rho d\theta d\phi = \frac{2}{3}\pi R^3 + \frac{1}{3}\sqrt{3}\pi R^3$$

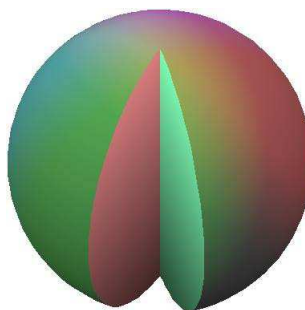
Now change the example a little by cutting out a cone at the bottom which has an angle of $\pi/2$ as shown. What is the volume of what is left?



This time you would have the volume equals

$$\int_{\pi/6}^{3\pi/4} \int_0^{2\pi} \int_0^R \rho^2 \sin(\phi) d\rho d\theta d\phi = \frac{1}{3}\sqrt{2}\pi R^3 + \frac{1}{3}\sqrt{3}\pi R^3$$

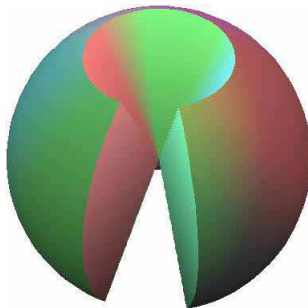
Example 20.3.5 Next suppose the ball of radius R is a sort of an orange and you remove a slice as shown in the picture. What is the volume of what is left? Assume the slice is formed by the two half planes $\theta = 0$ and $\theta = \pi/4$.



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 20.3.6 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

$$\int_{\pi/6}^{3\pi/4} \int_{\pi/4}^{2\pi} \int_0^R \rho^2 \sin(\phi) d\rho d\theta d\phi = \frac{7}{24}\sqrt{2}\pi R^3 + \frac{7}{24}\sqrt{3}\pi R^3$$

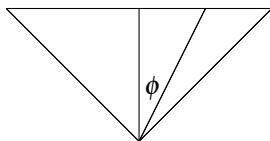
As to the integral, it equals

$$\int_{\pi/6}^{3\pi/4} \int_{\pi/4}^{2\pi} \int_0^R (\rho \sin(\phi) \cos(\theta)) \rho^2 \sin(\phi) d\rho d\theta d\phi = -\frac{1}{192}\sqrt{2}R^4 (7\pi + 3\sqrt{3} + 6)$$

This is because, in terms of spherical coordinates, $x = \rho \sin(\phi) \cos(\theta)$.

Example 20.3.7 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 20.3.8 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

$$\det \begin{pmatrix} \cos \theta & -r \sin \theta & 0 \\ \sin \theta & r \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = r.$$

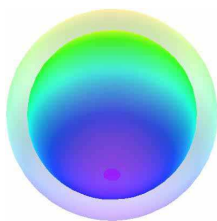
It follows the volume element in cylindrical coordinates is $r d\theta dr dz$.

Example 20.3.9 In the cone of Example 20.3.7 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 20.3.10 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

$$\frac{x\mathbf{i} + y\mathbf{j} + (z - z_0)\mathbf{k}}{|x\mathbf{i} + y\mathbf{j} + (z - z_0)\mathbf{k}|} \frac{1}{(x^2 + y^2 + (z - z_0)^2)} G\alpha dV =$$

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.

$$(x\mathbf{i} + y\mathbf{j} + (z - z_0)\mathbf{k}) \frac{1}{(x^2 + y^2 + (z - z_0)^2)^{3/2}} G\alpha dV$$

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.

$$\int_0^{2\pi} \int_a^b \int_0^\pi \frac{(\rho \cos \phi - z_0) \rho^2 \sin \phi}{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{3/2}} d\phi d\rho d\theta. \quad (20.1)$$

Rewrite the inside integral and use integration by parts to obtain this inside integral equals

$$\begin{aligned} & \frac{1}{2z_0} \int_0^\pi (\rho^2 \cos \phi - \rho z_0) \frac{(2z_0 \rho \sin \phi)}{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{3/2}} d\phi = \\ & \frac{1}{2z_0} \left(-2 \frac{-\rho^2 - \rho z_0}{\sqrt{(\rho^2 + z_0^2 + 2\rho z_0)}} + 2 \frac{\rho^2 - \rho z_0}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0)}} \right. \\ & \quad \left. - \int_0^\pi 2\rho^2 \frac{\sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} d\phi \right). \end{aligned} \quad (20.2)$$

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 20.2 equals

$$\begin{aligned} & \frac{1}{2z_0} \left(4\rho - \int_0^\pi 2\rho^2 \frac{\sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} d\phi \right) \\ & = \frac{1}{2z_0} \left(4\rho - \frac{1}{z_0} \int_0^\pi \rho \frac{2\rho z_0 \sin \phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)}} d\phi \right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2z_0} \left(4\rho - \frac{2\rho}{z_0} (\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{1/2} \Big|_0^\pi \right) \\
&= \frac{1}{2z_0} \left(4\rho - \frac{2\rho}{z_0} [(\rho + z_0) - (\rho - z_0)] \right) = 0.
\end{aligned}$$

Therefore, in this case the inner integral of 20.1 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 20.2 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, 20.2 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

$$\begin{aligned}
&\frac{-\rho}{z_0^2} \left((\rho^2 + z_0^2 - 2\rho z_0 \cos \phi)^{1/2} \Big|_0^\pi \right) \\
&= \frac{-\rho}{z_0^2} [(\rho + z_0) - (z_0 - \rho)] = -\frac{2\rho^2}{z_0^2}.
\end{aligned}$$

Thus the inner integral of 20.1 reduces to the above simple expression. Therefore, 20.1 equals

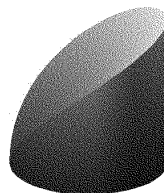
$$\int_0^{2\pi} \int_a^b \left(-\frac{2}{z_0^2} \rho^2 \right) d\rho d\theta = -\frac{4}{3} \pi \frac{b^3 - a^3}{z_0^2}$$

and so

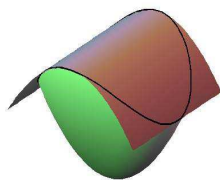
$$\begin{aligned}
&\alpha G k \int_H \frac{(z - z_0)}{[x^2 + y^2 + (z - z_0)^2]^{3/2}} dV \\
&= \alpha G k \left(-\frac{4}{3} \pi \frac{b^3 - a^3}{z_0^2} \right) = -k G \frac{\text{total mass}}{z_0^2}.
\end{aligned}$$

20.4 Exercises

1. Find the volume of the region bounded by $z = 0$, $x^2 + (y - 2)^2 = 4$, and $z = \sqrt{x^2 + y^2}$.



2. Find the volume of the region $z \geq 0, x^2 + y^2 \leq 4$, and $z \leq 4 - \sqrt{x^2 + y^2}$.
3. Find the volume of the region which is between the surfaces $z = 5y^2 + 9x^2$ and $z = 9 - 4y^2$.



4. 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.
5. 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.
6. Find the volume between $z = 3 - x^2 - y^2$ and $z = 2\sqrt{x^2 + y^2}$.
7. 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?
8. Find the volume of the cone defined by $z \in [0, 4]$ having angle $\pi/2$. Use spherical coordinates.
9. 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 20.3.10 on Page 375 check out all the details by working the integrals to be sure the steps are right.
12. What if the hollow sphere in Example 20.3.10 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.

$$\begin{aligned}
 (a) \quad & \int_{-2}^2 \int_0^{\sqrt{4-x^2}} \int_0^x xy dz dy dx \\
 (b) \quad & \int_{-1}^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \int_0^{x+y} dz dx dy \\
 (c) \quad & \int_0^1 \int_0^{\sqrt{1-x^2}} \int_x^1 dz dy dx \\
 (d) \quad & \text{For } a > 0, \int_{-a}^a \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} \int_{-\sqrt{a^2-x^2-y^2}}^{\sqrt{a^2-x^2-y^2}} dz dy dx \\
 (e) \quad & \int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{-\sqrt{4-x^2-y^2}}^{\sqrt{4-x^2-y^2}} dz dy dx
 \end{aligned}$$

14. Convert the following integrals into integrals involving spherical coordinates and then evaluate them.

$$\begin{aligned}
 (a) \quad & \int_{-a}^a \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} \int_{-\sqrt{a^2-x^2-y^2}}^{\sqrt{a^2-x^2-y^2}} dz dy dx \\
 (b) \quad & \int_{-1}^1 \int_0^{\sqrt{1-x^2}} \int_{-\sqrt{1-x^2-y^2}}^{\sqrt{1-x^2-y^2}} dz dy dx \\
 (c) \quad & \int_{-\sqrt{2}}^{\sqrt{2}} \int_{-\sqrt{2-x^2}}^{\sqrt{2-x^2}} \int_{\sqrt{x^2+y^2}}^{\sqrt{4-x^2-y^2}} dz dy dx \\
 (d) \quad & \int_{-\sqrt{3}}^{\sqrt{3}} \int_{-\sqrt{3-x^2}}^{\sqrt{3-x^2}} \int_1^{\sqrt{4-x^2-y^2}} dz dy dx \\
 (e) \quad & \int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{-\sqrt{4-x^2-y^2}}^{\sqrt{4-x^2-y^2}} dz dy dx
 \end{aligned}$$

20.5 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_n so $dV = dx_1 dx_2 \cdots dx_n$ 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 20.5.1 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.

$$\text{volume of } P(\mathbf{u}_1, \dots, \mathbf{u}_p) \equiv (\det(\mathbf{u}_i \cdot \mathbf{u}_j))^{1/2}.$$

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

$$x = f_1(u_1, u_2, u_3), y = f_2(u_1, u_2, u_3), z = f_3(u_1, u_2, u_3) \quad (20.3)$$

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

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} f_1(u_1, u_2, u_3) \\ f_2(u_1, u_2, u_3) \\ f_3(u_1, u_2, u_3) \end{pmatrix} = \mathbf{f}(\mathbf{u}) \quad (20.4)$$

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,

$$\begin{aligned} \mathbf{f}(u_{10} + s_1 du_1, u_{20} + s_2 du_2, u_{30} + s_3 du_3) - \mathbf{f}(u_{01}, u_{02}, u_{03}) = \\ D\mathbf{f}(u_{10}, u_{20}, u_{30}) \begin{pmatrix} s_1 du_1 \\ s_2 du_2 \\ s_3 du_3 \end{pmatrix} + \mathbf{o} \begin{pmatrix} s_1 du_1 \\ s_2 du_2 \\ s_3 du_3 \end{pmatrix} \end{aligned}$$

where the last term may be taken equal to $\mathbf{0}$ because 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}^n and \mathbf{x} is a point in V , a subset of n dimensional space. Thus, letting the Cartesian coordinates of \mathbf{x} be given by $\mathbf{x} = (x_1, \dots, x_n)^T$, 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 n vectors $\left\{ \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i \right\}_{i=1}^n$.

From Definition 20.5.1, the volume of this infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ is given by

$$\left(\det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i \cdot \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_j} du_j \right) \right)^{1/2} \quad (20.5)$$

in which there is no sum on the repeated index. Now in general, if there are n vectors in \mathbb{R}^n , $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$,

$$\det(\mathbf{v}_i \cdot \mathbf{v}_j)^{1/2} = |\det(\mathbf{v}_1, \dots, \mathbf{v}_n)| \quad (20.6)$$

where this last matrix is the $n \times n$ matrix which has the i^{th} column equal to \mathbf{v}_i . The reason for this is that the matrix whose ij^{th} entry is $\mathbf{v}_i \cdot \mathbf{v}_j$ is just the product of the two matrices,

$$\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{pmatrix} (\mathbf{v}_1, \dots, \mathbf{v}_n)$$

where the first on the left is the matrix having the i^{th} row equal to \mathbf{v}_i^T while the matrix on the right is just the matrix having the i^{th} column equal to \mathbf{v}_i . Therefore, since the determinant of a matrix equals the determinant of its transpose,

$$\det(\mathbf{v}_i \cdot \mathbf{v}_j) = \det \left(\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{pmatrix} (\mathbf{v}_1, \dots, \mathbf{v}_n) \right) = \det(\mathbf{v}_1, \dots, \mathbf{v}_n)^2$$

and so taking square roots yields (20.6). Therefore, from the properties of determinants, (20.5) equals

$$\left| \det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1} du_1, \dots, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_n} du_n \right) \right| = \left| \det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1}, \dots, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_n} \right) \right| du_1 \cdots du_n$$

This is the infinitesimal chunk of volume corresponding to the point $\mathbf{f}(\mathbf{u}_0)$ in V .

Definition 20.5.2 Let $\mathbf{x} = \mathbf{f}(\mathbf{u})$ be as described above. Then the symbol

$$\frac{\partial(x_1, \dots, x_n)}{\partial(u_1, \dots, u_n)},$$

called the *Jacobian determinant*, is defined by

$$\det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1}, \dots, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_n} \right) \equiv \frac{\partial(x_1, \dots, x_n)}{\partial(u_1, \dots, u_n)}.$$

Also, the symbol $\left| \frac{\partial(x_1, \dots, x_n)}{\partial(u_1, \dots, u_n)} \right| du_1 \cdots du_n$ 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 20.5.3 Suppose U is an open subset of \mathbb{R}^n for $n > 0$ 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 $h : \mathbf{f}(U) \rightarrow \mathbb{R}$,

$$\int_U h(\mathbf{f}(\mathbf{u})) \left| \frac{\partial(x_1, \dots, x_n)}{\partial(u_1, \dots, u_n)} \right| dV = \int_{\mathbf{f}(U)} h(\mathbf{x}) dV.$$

Example 20.5.4 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

$$\int_{1/2}^2 \int_1^3 \frac{u}{(v+1)^2} du dv = \frac{4}{3}.$$

20.6 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 $y = 4x$.
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 .

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.

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.

20.7 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 20.7.1 *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. See Volume 1 to see this explained with point masses. The only difference is that here the sums need to be replaced with integrals.

Example 20.7.2 *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.

$$\int_R (x^2 + y^2) \rho(\mathbf{x}) dV. \quad (20.7)$$

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 20.7.3 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 $(0, 0, \frac{18}{7})$.

20.8 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 21

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.

21.1 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.

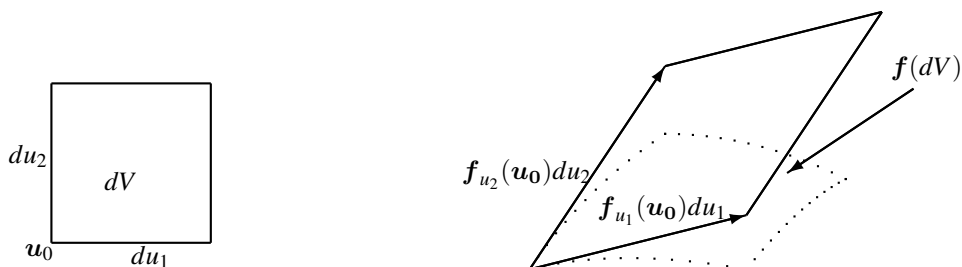
Definition 21.1.1 Let $\mathbf{u}_1, \mathbf{u}_2$ be vectors in \mathbb{R}^3 . The 2 dimensional parallelogram determined by these vectors will be denoted by $P(\mathbf{u}_1, \mathbf{u}_2)$ and it is defined as

$$P(\mathbf{u}_1, \mathbf{u}_2) \equiv \left\{ \sum_{j=1}^2 s_j \mathbf{u}_j : s_j \in [0,1] \right\}.$$

Then the area of this parallelogram is

$$\text{area } P(\mathbf{u}_1, \mathbf{u}_2) \equiv |\mathbf{u}_1 \times \mathbf{u}_2|.$$

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 21.1.1, the two dimensional volume of this infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ is given by

$$\left| \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_1} du_1 \times \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_2} du_2 \right| = \left| \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_1} \times \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_2} \right| du_1 du_2 \quad (21.1)$$

$$= |\mathbf{f}_{u_1} \times \mathbf{f}_{u_2}| du_1 du_2 \quad (21.2)$$

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 21.1.2 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.

$$\int_{\mathbf{f}(U)} h(\mathbf{x}) dA \equiv \int_U h(\mathbf{f}(\mathbf{u})) |\mathbf{f}_{u_1}(\mathbf{u}) \times \mathbf{f}_{u_2}(\mathbf{u})| du_1 du_2.$$

Definition 21.1.3 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 21.1.4 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

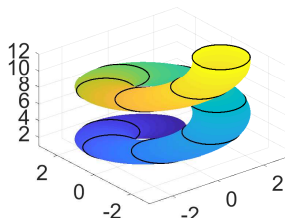
$$\sin(x^2 + \ln(7 + y^2 \sin x)) + \sin(zx)e^z = 11 \sin(xyz)?$$

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 21.1.5 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

$$\mathbf{f}_t = \begin{pmatrix} -2 \sin t & 2 \cos t & 1 \end{pmatrix}^T, \quad \mathbf{f}_s = \begin{pmatrix} -\sin s & \cos s & 0 \end{pmatrix}^T$$

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

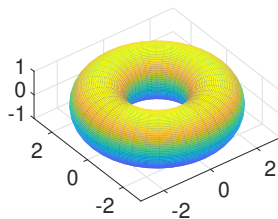
$$\int_0^{2\pi} \int_0^{12} \sqrt{5 - 4 \sin^2 t \sin^2 s - 8 \sin t \sin s \cos t \cos s - 4 \cos^2 t \cos^2 s} dt ds.$$

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 21.1.6 Let $U = [0, 2\pi] \times [0, 2\pi]$ and for $(t, s) \in U$, let

$$\mathbf{f}(t, s) = (2 \cos t + \cos t \cos s, -2 \sin t - \sin t \cos s, \sin s)^T.$$

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,

$$\mathbf{f}_t = \begin{pmatrix} -2 \sin t - \sin t \cos s \\ -2 \cos t - \cos t \cos s \\ 0 \end{pmatrix}, \quad \mathbf{f}_s = \begin{pmatrix} -\cos t \sin s \\ \sin t \sin s \\ \cos s \end{pmatrix}$$

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 21.1.7 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

$$\int_{\mathbf{f}(U)} h dA = \int_0^{2\pi} \int_0^{2\pi} \left(\overbrace{2 \cos t + \cos t \cos s}^{x \text{ on the surface}} \right)^2 (\cos s + 2) ds dt = 22\pi^2$$

21.2 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

$$\mathbf{f}_x = \begin{pmatrix} 1 \\ 0 \\ f_x \end{pmatrix}, \quad \mathbf{f}_y = \begin{pmatrix} 0 \\ 1 \\ f_y \end{pmatrix}$$

and

$$|\mathbf{f}_x \times \mathbf{f}_y| = \sqrt{1 + f_y^2 + f_x^2}$$

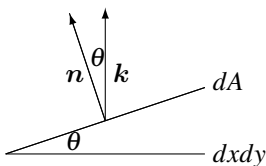
so the area element is

$$\sqrt{1 + f_y^2 + f_x^2} dx dy.$$

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 $y = f(x, z)$.

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 21.2.1 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 described as

$$(x, y, \sqrt{x^2 + y^2})$$

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.

21.3 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 to allow rotating the graph to see it from different angles.

21.4 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, they are irrelevant.

21.5 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. 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 21.5.1 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

$$x = 2 \sin \phi \cos \theta, y = 2 \sin \phi \sin \theta, z = 2 \cos \phi$$

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.

21.6 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.

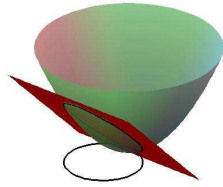
16. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let

$$\mathbf{f}(\theta, \alpha) \equiv (\cos \theta (4 + 3 \cos \alpha), -\sin \theta (4 + 3 \cos \alpha), 3 \sin \alpha)^T, \quad h(\mathbf{x}) = \cos^2 \theta,$$

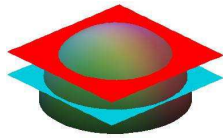
where θ is such that $\mathbf{x} = (\cos \theta (4 + 3 \cos \alpha), -\sin \theta (4 + 3 \cos \alpha), 3 \sin \alpha)^T$. Find $\int_{\mathbf{f}([0, 2\pi] \times [0, 2\pi])} h dA$. **Hint:** Check whether $\mathbf{f}_\theta \cdot \mathbf{f}_\alpha = 0$. This might make the computations reasonable.

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 22

Calculus Of Vector Fields

22.1 Divergence And Curl Of A Vector Field

Here the important concepts of divergence and curl are defined.

Definition 22.1.1 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.

$$(\text{curl}(\mathbf{f})(\mathbf{x}))_i \equiv (\nabla \times \mathbf{f}(\mathbf{x}))_i \equiv \epsilon_{ijk} \partial_j f_k(\mathbf{x})$$

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 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 22.1.2 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

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ xy & z-y & \sin(x) + z \end{vmatrix} =$$

$$\mathbf{i} \left(\frac{\partial}{\partial y} (\sin(x) + z) - \frac{\partial}{\partial z} (z - y) \right) - \mathbf{j} \left(\frac{\partial}{\partial x} (\sin(x) + z) - \frac{\partial}{\partial z} (xy) \right) +$$

$$\mathbf{k} \left(\frac{\partial}{\partial x} (z - y) - \frac{\partial}{\partial y} (xy) \right) = -\mathbf{i} - \cos(x)\mathbf{j} - x\mathbf{k}.$$

22.1.1 Vector Identities

There are many interesting identities which relate the gradient, divergence and curl.

Theorem 22.1.3 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.

$$\begin{aligned} (\nabla \times (\nabla \times \mathbf{f}))_i &= \varepsilon_{ijk} \partial_j (\nabla \times \mathbf{f})_k = \varepsilon_{ijk} \partial_j (\varepsilon_{krs} \partial_r f_s) \\ &= \varepsilon_{ijk} \varepsilon_{krs} \partial_j (\partial_r f_s) = (\delta_{ir} \delta_{js} - \delta_{is} \delta_{jr}) \partial_j (\partial_r f_s) \\ &= \partial_j (\partial_i f_j) - \partial_j (\partial_j f_i) = \partial_i (\partial_j f_j) - \partial_j (\partial_j f_i) \\ &= \left(\nabla (\nabla \cdot \mathbf{f}) - \nabla^2 \mathbf{f} \right)_i \end{aligned}$$

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. ■

22.1.2 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.

$$\mathbf{f}(x, y, z) = \left(\int_0^z g_2(x, y, t) dt, -\int_0^z g_1(x, y, t) dt + \int_0^x g_3(t, y, 0) dt, 0 \right)^T. \quad (22.1)$$

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.

$$\frac{\partial}{\partial x} \int_a^b h(x, t) dt = \int_a^b \frac{\partial h}{\partial x}(x, t) dt. \quad (22.2)$$

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 22.1.4 Suppose h and $\frac{\partial h}{\partial x}$ are continuous on the rectangle $R = [c, d] \times [a, b]$. Then (22.2) holds.

Proof: Let Δx be such that $x, x + \Delta x$ are both in $[c, d]$. By Theorem 13.5.5 on Page 243 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 22.1.3 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 Stoke's theorem. This scalar field ϕ , is called a **scalar potential** for \mathbf{f} .

22.1.3 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 22.1.5 Let U be a bounded open set in \mathbb{R}^n 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

$$u(\mathbf{x}_0) > \max \{u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2 : \mathbf{x} \in \partial U\}.$$

Therefore, $u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2$ also has its maximum in U because for ε small enough,

$$u(\mathbf{x}_0) + \varepsilon |\mathbf{x}_0|^2 > u(\mathbf{x}_0) > \max \{u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2 : \mathbf{x} \in \partial U\}$$

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, $\nabla^2(u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2) = \nabla^2 u(\mathbf{x}) + 2n\varepsilon$. (Why?) Therefore,

$$0 \geq \nabla^2 u(\mathbf{x}_1) + 2n\varepsilon \geq 2n\varepsilon,$$

a contradiction. ■

22.2 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 22.1.3.

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 (22.1) 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 22.1.5, why is $\nabla^2 (\varepsilon |x|^2) = 2n\varepsilon$?

13. Using Theorem 22.1.5, 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 $u \in C^2(U) \cap C(\bar{U})$ 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.

22.3 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 22.3.1 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

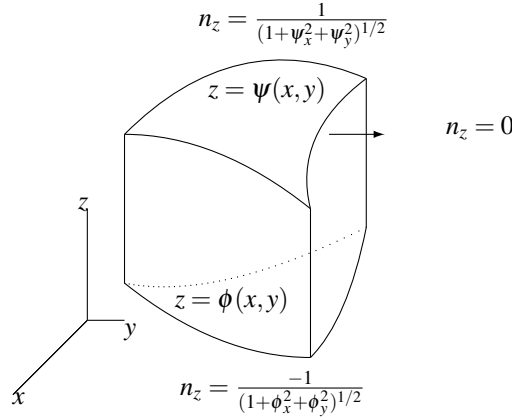
$$V = \{(x, y, z) : \phi(x, y) \leq z \leq \psi(x, y) \text{ for } (x, y) \in D\}$$

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 to be 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 22.3.2 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,

$$\begin{aligned} \int_V \frac{\partial F}{\partial z}(x, y, z) dV &= \int_D \int_{\phi(x, y)}^{\psi(x, y)} \frac{\partial F}{\partial z}(x, y, z) dz dx dy \\ &= \int_D [F(x, y, \psi(x, y)) - F(x, y, \phi(x, y))] dx dy \end{aligned} \quad (22.3)$$

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 (22.3) 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 22.3.3 *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.

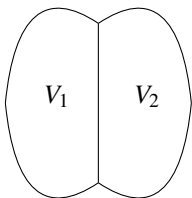
Theorem 22.3.4 *Let V be cylindrical in each of the coordinate directions and let \mathbf{F} be a C^1 vector field defined on V . Then*

$$\int_V \nabla \cdot \mathbf{F} dV = \int_{\partial V} \mathbf{F} \cdot \mathbf{n} dA.$$

Proof: From the above lemma and corollary,

$$\begin{aligned}
 \int_V \nabla \cdot \mathbf{F} dV &= \int_V \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z} dV \\
 &= \int_{\partial V} (F_1 n_x + F_2 n_y + F_3 n_z) dA \\
 &= \int_{\partial V} \mathbf{F} \cdot \mathbf{n} dA. \quad \blacksquare
 \end{aligned}$$

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. When all is said and done, one finds that the conclusion of the divergence theorem is usually true and the theorem can be used with confidence.

Example 22.3.5 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 22.3.6 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

$$x = \sin \phi \cos \theta, y = \sin \phi \sin \theta, z = \cos \phi$$

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 22.3.7 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.

22.3.1 Coordinate Free Concept Of Divergence

The divergence theorem also makes possible a coordinate free definition of the divergence.

Theorem 22.3.8 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

$$\operatorname{div} \mathbf{F}(\mathbf{x}) = \lim_{\delta \rightarrow 0^+} \frac{1}{v(B(\mathbf{x}, \delta))} \int_{\partial B(\mathbf{x}, \delta)} \mathbf{F} \cdot \mathbf{n} dA. \quad (22.4)$$

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 (22.4). ■

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.

22.4 Some Applications Of The Divergence Theorem

22.4.1 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 426) 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

$$\nabla p = \rho \mathbf{g}. \quad (22.5)$$

It turns out that the pressure in a lake at depth z is equal to $62.5z$. This is easy to see from (22.5). 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 (22.5) is more interesting because it does not require ρ to be constant.

22.4.2 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 (22.5) 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 (22.5) because in this equation $\mathbf{g} = -g\mathbf{k}$.

22.4.3 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.

$$\frac{\partial}{\partial t} (\rho(\mathbf{x}, t) c(\mathbf{x}, t) u(\mathbf{x}, t)) = \nabla \cdot (k(u, \mathbf{x}, t) \nabla u(\mathbf{x}, t)) + f(\mathbf{x}, u, t). \quad (22.6)$$

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 (22.6) 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

$$\frac{\partial}{\partial t} u(\mathbf{x}, t) = \frac{1}{\rho c} \nabla \cdot (k(\mathbf{x}) \nabla u(\mathbf{x}, t)) + f(\mathbf{x}, t). \quad (22.7)$$

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.

22.4.4 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.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = f(\mathbf{y}, t) \quad (22.8)$$

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.

22.4.5 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

$$\mathbf{y}(0, \mathbf{x}) = \mathbf{x}. \quad (22.9)$$

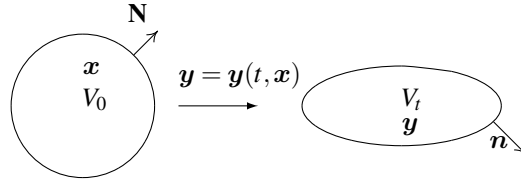
The derivative,

$$D_2 \mathbf{y}(t, \mathbf{x}) \equiv D_{\mathbf{x}} \mathbf{y}(t, \mathbf{x})$$

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 ij^{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 22.4.1 $\rho_0(\mathbf{x}) = \rho(t, \mathbf{y}(t, \mathbf{x})) \det(F)$ 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 22.9. 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 $\det(F) > 0$.

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 22.4.1 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,

$$\int_{V_0} \rho_0(\mathbf{x}) \mathbf{v}_t(t, \mathbf{x}) dV = \int_{V_0} \mathbf{b}_0(t, \mathbf{x}) dV + \int_{\partial V_t} \mathbf{e}_i T_{ij} n_j dA, \quad (22.10)$$

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 22.10 is the sum of integrals over small pieces of the form

$$\int_{P_t} T_{ij} n_j dA \quad (22.11)$$

where P_t is parameterized by $\mathbf{y}(s, t)$, $(s, t) \in D$. Thus the integral in 22.11 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

$$\int_D T_{ij}(\mathbf{x}(s, t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \left(\frac{\partial \mathbf{y}}{\partial x_\alpha} \times \frac{\partial \mathbf{y}}{\partial x_\beta} \right)_j ds dt \quad (22.12)$$

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 22.12 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 22.10 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

$$\begin{aligned}
\rho_0(\mathbf{x}) \mathbf{v}_t(t, \mathbf{x}) &= \mathbf{b}_0(t, \mathbf{x}) + \mathbf{e}_i \frac{\partial (\det(F) (TF^{-T})_{ip})}{\partial x_p} \\
&= \mathbf{b}_0(t, \mathbf{x}) + \operatorname{div} (\det(F) (TF^{-T})) \\
&= \mathbf{b}_0(t, \mathbf{x}) + \operatorname{div} S.
\end{aligned} \tag{22.13}$$

As just shown, the relation between the Cauchy stress and the Piola Kirchhoff stress is

$$S = \det(F) (TF^{-T}), \tag{22.14}$$

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.

22.4.6 The Reynolds Transport Formula

The Reynolds transport formula is another interesting application of the divergence theorem which is a generalization of the formula for taking the derivative under an integral.

$$\frac{d}{dt} \int_{a(t)}^{b(t)} f(x, t) dx = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}(x, t) dx + f(b(t), t) b'(t) - f(a(t), t) a'(t)$$

Of course there are difficult analytical questions connected with such a formal procedure, but these can be easily justified with sufficient machinery involving the Lebesgue integral. An elementary version of theorems necessary to justify this will be fussy and unpleasant so I am going to emphasize the derivation of the formula without worrying about interchange of limit considerations and whether the divergence theorem holds for the region of interest.

First is an interesting lemma about the determinant. A $p \times p$ matrix can be thought of as a vector in \mathbb{C}^{p^2} . Just imagine stringing it out into one long list of numbers. In fact, a way to give the norm of a matrix is just $\sum_i \sum_j |A_{ij}|^2 \equiv \|A\|^2$. You might check to see that this is the same as $(\text{trace}(AA^*))^{1/2} = \|A\|$. It is called the Frobenius norm for a matrix. Also recall that \det maps $p \times p$ matrices to \mathbb{C} . It makes sense to ask for the derivative of \det on the set of invertible matrices, an open subset of \mathbb{C}^{p^2} with the norm measured as just described. This is because $A \rightarrow \det(A)$ is continuous so the set where $\det(A) \neq 0$ would be an open set. Recall that $\text{trace}(AB) = \text{trace}(BA)$ whenever both products make sense. Indeed,

$$\text{trace}(AB) = \sum_i \sum_j A_{ij} B_{ji} = \text{trace}(BA)$$

This next lemma is a very interesting observation about the determinant of a matrix added to the identity.

Lemma 22.4.2 $\det(I + U) = 1 + \text{trace}(U) + o(U)$ where $o(U)$ is defined in terms of the Frobenius norm for $p \times p$ matrices.

Proof: This is obvious if $p = 1$ or 2 . Assume true for $n - 1$. Then for U an $n \times n$, $I + U =$

$$\begin{pmatrix} 1 + U_{11} & U_{12} & \cdots & U_{1n} \\ U_{21} & 1 + U_{22} & \cdots & U_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ U_{n1} & \cdots & U_{n(n-1)} & 1 + U_{nn} \end{pmatrix}$$

Expand along the last column and use induction.

$$(1 + U_{nn}) \left(1 + \sum_{k=1}^{n-1} U_{kk} + o(U) \right) + o(U)$$

That last term follows from observing that you have some U_{kn} times terms which have at least one other factor involving some U_{nj} . Simply expand the resulting cofactors along the bottom row. Therefore, multiplying this out gives $1 + \text{trace}(U) + o(U)$. ■

With this lemma, it is easy to find $D\det(F)$ whenever F is invertible.

$$\begin{aligned}\det(F + U) &= \det(F(I + F^{-1}U)) = \det(F)\det(I + F^{-1}U) \\ &= \det(F)(1 + \text{trace}(F^{-1}U) + o(U)) \\ &= \det(F) + \det(F)\text{trace}(F^{-1}U) + o(U)\end{aligned}$$

Therefore,

$$\det(F + U) - \det(F) = \det(F)\text{trace}(F^{-1}U) + o(U)$$

This proves the following.

Proposition 22.4.3 *Let F^{-1} exist. Then $D\det(F)(U) = \det(F)\text{trace}(F^{-1}U)$.*

From this, suppose $F(t)$ is a $p \times p$ matrix and all entries are differentiable. Then the following describes $\frac{d}{dt}\det(F)(t)$.

Proposition 22.4.4 *Let $F(t)$ be a $p \times p$ matrix and all entries are differentiable. Then*

$$\begin{aligned}\frac{d}{dt}\det(F)(t) &= \det(F(t))\text{trace}(F^{-1}(t)F'(t)) \\ &= \det(F(t))\text{trace}(F'(t)F^{-1}(t))\end{aligned}\tag{22.15}$$

The situation of interest is where \mathbf{x} is the material coordinates and \mathbf{y} the spacial coordinates and $\mathbf{y} = \mathbf{h}(t, \mathbf{x})$ with $F = F(t, \mathbf{x}) = D_2\mathbf{h}(t, \mathbf{x})$. I will write $\nabla_{\mathbf{y}}$ to indicate the gradient with respect to the \mathbf{y} variables and F' to indicate $\frac{\partial}{\partial t}F(t, \mathbf{x})$. Note that $\mathbf{h}(t, \mathbf{x}) = \mathbf{y}$ and so by the inverse function theorem, this defines \mathbf{x} as a function of \mathbf{y} , also as smooth as \mathbf{h} because it is always assumed $\det F > 0$.

Now let V_t be $\mathbf{h}(t, V_0)$ where V_0 is an open set whose boundary is sufficient for using the divergence theorem. Let $\mathbf{f}(\mathbf{y}, t)$ be differentiable with as many derivatives as needed to make the computations valid. The idea is to simplify

$$\frac{d}{dt} \int_{V_t} \mathbf{f}(t, \mathbf{y}) dV(\mathbf{y})$$

This will involve the change of variables in which the Jacobian will be $\det(F)$. It will not be necessary to take the absolute value because $\det(F) \leq 0$ is not physically possible. Then, it is fairly routine to justify the interchange of the derivative and the integral under suitable assumptions. The best would be to use the dominated convergence theorem, but formally, it is like saying the derivative of a sum is the sum of the derivatives. There is of course the question whether the divergence theorem will continue to hold for V_t . This will end up holding under typical assumptions normally used for assumptions that the divergence theorem will hold for V_0 . For example, if $\mathbf{h}(t, \cdot)$ is smooth and the boundary of V_0 is Lipschitz, all will be well, but this is an application of things like Rademacher's theorem and the area formula.

$$\frac{d}{dt} \int_{V_t} \mathbf{f}(t, \mathbf{y}) dV(\mathbf{y}) = \frac{d}{dt} \int_{V_0} \mathbf{f}(t, \mathbf{h}(t, \mathbf{x})) \det(F) dV(\mathbf{x})\tag{22.16}$$

$$\begin{aligned}
&= \int_{V_0} \frac{\partial}{\partial t} \mathbf{f}(\cdot, \mathbf{h}(\cdot, \mathbf{x})) \det(F) dV(\mathbf{x}) + \int_{V_0} \mathbf{f}(t, \mathbf{h}(t, \mathbf{x})) \frac{\partial}{\partial t} (\det(F)) dV(\mathbf{x}) \\
&= \int_{V_0} \frac{\partial}{\partial t} \mathbf{f}(t, \mathbf{h}(t, \mathbf{x})) \det(F) dV(\mathbf{x}) \\
&\quad + \int_{V_0} \mathbf{f}(t, \mathbf{h}(t, \mathbf{x})) \operatorname{trace}(F'F^{-1}) \det(F) dV(\mathbf{x}) \\
&= \int_{V_0} \left(\frac{\partial}{\partial t} \mathbf{f}(t, \mathbf{h}(t, \mathbf{x})) + \frac{\partial \mathbf{f}}{\partial y_i} \frac{\partial y_i}{\partial t} \right) \det(F) dV(\mathbf{x}) \\
&\quad + \int_{V_0} \mathbf{f}(t, \mathbf{h}(t, \mathbf{x})) \operatorname{trace}(F'F^{-1}) \det(F) dV(\mathbf{x}) \\
&= \int_{V_t} \frac{\partial}{\partial t} \mathbf{f}(t, \mathbf{y}) dV(\mathbf{y}) + \int_{V_t} \frac{\partial \mathbf{f}}{\partial y_i} \frac{\partial y_i}{\partial t} + \mathbf{f}(t, \mathbf{y}) \operatorname{trace}(F'F^{-1}) dV(\mathbf{y})
\end{aligned}$$

Now $\mathbf{v} = \frac{\partial}{\partial t} \mathbf{h}(t, \mathbf{x})$ and also, as noted above, $\mathbf{y} = \mathbf{h}(t, \mathbf{x})$ defines \mathbf{y} as a function of \mathbf{x} and so $\operatorname{trace}(F'F^{-1}) = \frac{\partial v_i}{\partial x_\alpha} \frac{\partial x_\alpha}{\partial y_i}$. Hence the double sum $\frac{\partial v_i}{\partial x_\alpha} \frac{\partial x_\alpha}{\partial y_i}$ is $\frac{\partial v_i}{\partial y_i} = \nabla_{\mathbf{y}} \cdot \mathbf{v}$. The above then gives

$$\begin{aligned}
&\int_{V_t} \frac{\partial}{\partial t} \mathbf{f}(t, \mathbf{y}) dV(\mathbf{y}) + \int_{V_t} \left(\frac{\partial \mathbf{f}}{\partial y_i} \frac{\partial y_i}{\partial t} + \mathbf{f}(t, \mathbf{y}) \nabla_{\mathbf{y}} \cdot \mathbf{v} \right) dV(\mathbf{y}) \\
&= \int_{V_t} \frac{\partial}{\partial t} \mathbf{f}(t, \mathbf{y}) dV(\mathbf{y}) + \int_{V_t} (D_1 \mathbf{f}(\mathbf{y}, t) \mathbf{v} + \mathbf{f}(t, \mathbf{y}) \nabla_{\mathbf{y}} \cdot \mathbf{v}) dV(\mathbf{y}) \quad (22.17)
\end{aligned}$$

Now consider the i^{th} component of the second integral in the above. It is

$$\begin{aligned}
&\int_{V_t} \nabla_{\mathbf{y}} f_i(t, \mathbf{y}) \cdot \mathbf{v} + \mathbf{f}(t, \mathbf{y}) \nabla_{\mathbf{y}} \cdot \mathbf{v} dV(\mathbf{y}) \\
&= \int_{V_t} \nabla_{\mathbf{y}} \cdot (f_i(t, \mathbf{y}) \mathbf{v}) dV(\mathbf{y})
\end{aligned}$$

At this point, use the divergence theorem to get

$$= \int_{\partial V_t} f_i(t, \mathbf{y}) \mathbf{v} \cdot \mathbf{n} dA$$

Therefore, from 22.17 and 22.16,

$$\frac{d}{dt} \int_{V_t} \mathbf{f}(t, \mathbf{y}) dV(\mathbf{y}) = \int_{V_t} \frac{\partial}{\partial t} \mathbf{f}(t, \mathbf{y}) dV(\mathbf{y}) + \int_{\partial V_t} \mathbf{f}(t, \mathbf{y}) \mathbf{v} \cdot \mathbf{n} dA$$

this is the Reynolds transport formula.

22.4.7 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, (see Problem 21 on Page 518) and

$$\mathbf{y}' = \mathbf{q}(t) + Q(t)\mathbf{y}, \mathbf{n}' = Q(t)\mathbf{n}$$

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

$$T' \mathbf{n}' = QT \mathbf{n}$$

Thus

$$T' Q \mathbf{n} = QT \mathbf{n}$$

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

$$T' Q = QT$$

and so

$$T' = QTQ^T.$$

This is called frame indifference.

By 22.14, the Piola Kirchhoff stress S is related to T by

$$S = \det(F) T F^{-T}, \quad F \equiv D_x \mathbf{y}.$$

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

$$F' = D_x \mathbf{y}' = Q(t) D_x \mathbf{y} = QF$$

and so frame indifference in terms of S is

$$S' = \det(F) QTQ^T (QF)^{-T} = \det(F) QTQ^T QF^{-T} = QS$$

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.

22.4.8 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

$$\rho \dot{\mathbf{v}} = -\rho g \mathbf{k} - \nabla P(\mathbf{x}). \quad (22.18)$$

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 22.18 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.

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.

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.

22.4.9 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 $u \equiv y - x$. Thus u is the displacement. Then from (22.13) you can verify the following holds.

$$\rho_0(x) u_{tt}(t, x) = b_0(t, x) + k\Delta u(t, x) \quad (22.19)$$

In the case where ρ_0 is a constant and $b_0 = 0$, this yields

$$u_{tt} - c\Delta u = 0.$$

The wave equation is $u_{tt} - c\Delta u = 0$ and so the above gives three wave equations, one for each component.

22.4.10 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.

22.4.11 Volumes Of Balls In \mathbb{R}^n

Recall, $B(x, r)$ denotes the set of all $y \in \mathbb{R}^n$ such that $|y - x| < r$. 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,

$$\int_{B(0, r)} \operatorname{div} x dx = \int_{\partial B(0, r)} x \cdot \frac{x}{|x|} dA$$

because the unit outward normal on $\partial B(0, r)$ is $\frac{x}{|x|}$. Therefore, denoting $A(\partial B)$ as the area of ∂B ,

$$n\alpha_n r^n = rA(\partial B(0, r))$$

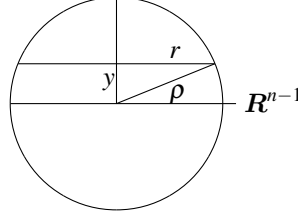
and so

$$A(\partial B(0, r)) = n\alpha_n r^{n-1}.$$

You recall the surface area of $S^2 \equiv \{x \in \mathbb{R}^3 : |x| = r\}$ is given by $4\pi r^2$ while the volume of the ball, $B(0, r)$ 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 $S^{n-1} = \{x \in \mathbb{R}^n : |x| = 1\}$. 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

$$\alpha_n = 2 \alpha_{n-1} \int_0^{\pi/2} \sin^n(\theta) d\theta. \quad (22.20)$$

Consequently,

$$\omega_n = \frac{2n\omega_{n-1}}{n-1} \int_0^{\pi/2} \sin^n(\theta) d\theta. \quad (22.21)$$

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

$$n \int_0^{\pi/2} \sin^n(\theta) d\theta = (n-1) \int_0^{\pi/2} \sin^{n-2}(\theta) d\theta \quad (22.22)$$

and so (22.21) is of the form

$$\omega_n = 2 \omega_{n-1} \int_0^{\pi/2} \sin^{n-2}(\theta) d\theta. \quad (22.23)$$

So what is α_n explicitly? Clearly $\alpha_1 = 2$ and $\alpha_2 = \pi$.

Theorem 22.4.5 $\alpha_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$ 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 $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. This is also outlined in an exercise in Volume 1.) 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, (22.23), (22.22), (22.20) and (22.21),

$$\begin{aligned} \alpha_{n+1} &= 2\alpha_n \int_0^{\pi/2} \sin^{n+1}(\theta) d\theta = 2\alpha_n \frac{n}{n+1} \int_0^{\pi/2} \sin^{n-1}(\theta) d\theta \\ &= 2\alpha_n \frac{n}{n+1} \frac{\alpha_{n-1}}{2\alpha_{n-2}} = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} \frac{n}{n+1} \pi^{1/2} \frac{\Gamma(\frac{n-2}{2}+1)}{\Gamma(\frac{n-1}{2}+1)} \\ &= \frac{\pi^{n/2}}{\Gamma(\frac{n-2}{2}+1)} \frac{n}{(\frac{n}{2})} \frac{1}{n+1} \pi^{1/2} \frac{\Gamma(\frac{n-2}{2}+1)}{\Gamma(\frac{n-1}{2}+1)} \\ &= 2\pi^{(n+1)/2} \frac{1}{n+1} \frac{1}{\Gamma(\frac{n-1}{2}+1)} = \pi^{(n+1)/2} \frac{1}{(\frac{n+1}{2})} \frac{1}{\Gamma(\frac{n-1}{2}+1)} \\ &= \pi^{(n+1)/2} \frac{1}{(\frac{n+1}{2})\Gamma(\frac{n+1}{2})} = \frac{\pi^{(n+1)/2}}{\Gamma(\frac{n+1}{2}+1)}. \quad \blacksquare \end{aligned}$$

22.4.12 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 $\mathbf{x} \neq \mathbf{x}_0$,

$$\nabla \cdot \nabla \left(qk \frac{1}{|\mathbf{x} - \mathbf{x}_0|} \right) = \nabla \cdot \mathbf{E} = 0. \quad (22.24)$$

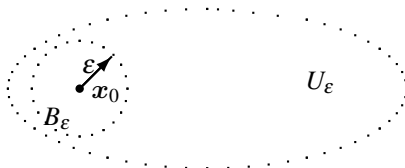
This is left as an exercise for you to verify.

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 (22.24),

$$-4\pi kq + \int_{\partial U} \mathbf{E} \cdot \mathbf{n} dS = \int_{\partial U_\varepsilon} \mathbf{E} \cdot \mathbf{n} dS = \int_{U_\varepsilon} \nabla \cdot \mathbf{E} dV = 0.$$

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.

22.5 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 $\int_V \nabla \cdot (u\nabla v) dV = \int_{\partial V} u \frac{\partial v}{\partial n} dA$ where \mathbf{n} is the exterior normal and $\frac{\partial v}{\partial n}$ denotes the directional derivative of v in the direction \mathbf{n} .
5. Let V be such that the divergence theorem holds. Show that

$$\int_V (v\nabla^2 u - u\nabla^2 v) dV = \int_{\partial V} \left(v \frac{\partial u}{\partial n} - u \frac{\partial v}{\partial n} \right) dA$$

where \mathbf{n} is the exterior normal and $\frac{\partial u}{\partial n}$ 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 $\frac{1}{3}x + \frac{1}{3}y + \frac{1}{3}z = 1$. 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 (22.19) from (22.13) and the assumption that $S = kF$.
15. Show that if $u_k, k = 1, 2, \dots, n$ each satisfies (22.7) 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 solutions of this form. Actually, it was a little more complicated because cannons are not one dimensional but it was the beginning of the study of Fourier series, a topic which fascinated mathematicians for the next 150 years and motivated the development of analysis.

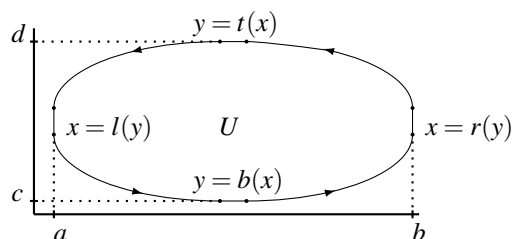
Chapter 23

Stokes And Green's Theorems

23.1 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. I will first establish Green's theorem for regions of a particular sort and then show that the theorem holds for many other regions also. Suppose a region is of the form indicated in the following picture in which

$$\begin{aligned} U &= \{(x, y) : x \in (a, b) \text{ and } y \in (b(x), t(x))\} \\ &= \{(x, y) : y \in (c, d) \text{ and } x \in (l(y), r(y))\}. \end{aligned}$$



I will refer to such a region as being convex in both the x and y directions.

Lemma 23.1.1 Let $\mathbf{F}(x, y) \equiv (P(x, y), Q(x, y))$ be a C^1 vector field defined near U where U is a region of the sort indicated in the above picture which is convex in both the x and y directions. Suppose also that the functions r, l, t , and b in the above picture are all C^1 functions and denote by ∂U the boundary of U oriented such that the direction of motion is counter clockwise. (As you walk around U on ∂U , the points of U are on your left.) Then

$$\int_{\partial U} Pdx + Qdy \equiv \int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA. \quad (23.1)$$

Proof: First consider the right side of (23.1).

$$\int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \int_c^d \int_{l(y)}^{r(y)} \frac{\partial Q}{\partial x} dx dy - \int_a^b \int_{b(x)}^{t(x)} \frac{\partial P}{\partial y} dy dx$$

$$\begin{aligned}
&= \int_c^d (Q(r(y), y) - Q(l(y), y)) dy \\
&\quad + \int_a^b (P(x, b(x)) - P(x, t(x))) dx. \tag{23.2}
\end{aligned}$$

Now consider the left side of (23.1). Denote by V the vertical parts of ∂U and by H the horizontal parts.

$$\begin{aligned}
\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} &= \int_{\partial U} ((0, Q) + (P, 0)) \cdot d\mathbf{R} \\
&= \int_c^d (0, Q(r(s), s)) \cdot (r'(s), 1) ds + \int_H (0, Q(r(s), s)) \cdot (\pm 1, 0) ds \\
&\quad - \int_c^d (0, Q(l(s), s)) \cdot (l'(s), 1) ds + \int_a^b (P(s, b(s)), 0) \cdot (1, b'(s)) ds \\
&\quad + \int_V (P(s, b(s)), 0) \cdot (0, \pm 1) ds - \int_a^b (P(s, t(s)), 0) \cdot (1, t'(s)) ds \\
&= \int_c^d Q(r(s), s) ds - \int_c^d Q(l(s), s) ds + \int_a^b P(s, b(s)) ds - \int_a^b P(s, t(s)) ds
\end{aligned}$$

which coincides with (23.2). ■

Corollary 23.1.2 *Let everything be the same as in Lemma 23.1.1 but only assume the functions r, l, t , and b are continuous and piecewise C^1 functions. Then the conclusion this lemma is still valid.*

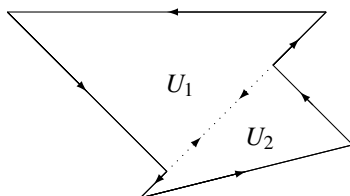
Proof: The details are left for you. All you have to do is to break up the various line integrals into the sum of integrals over sub intervals on which the function of interest is C^1 . ■

From this corollary, it follows (23.1) is valid for any triangle for example.

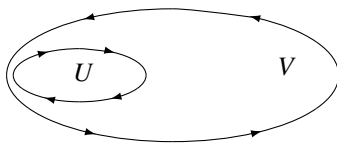
Now suppose (23.1) holds for U_1, U_2, \dots, U_m and the open sets U_k have the property that no two have nonempty intersection and their boundaries intersect only in a finite number of piecewise smooth curves. Then (23.1) must hold for $U \equiv \cup_{i=1}^m U_i$, the union of these sets. This is because

$$\begin{aligned}
&\int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \\
&= \sum_{k=1}^m \int_{U_k} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA \\
&= \sum_{k=1}^m \int_{\partial U_k} \mathbf{F} \cdot d\mathbf{R} = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}
\end{aligned}$$

because if $\Gamma = \partial U_k \cap \partial U_j$, then its orientation as a part of ∂U_k is opposite to its orientation as a part of ∂U_j and consequently the line integrals over Γ will cancel, points of Γ also not being in ∂U . As an illustration, consider the following picture for two such U_k .



Similarly, if $U \subseteq V$ and if also $\partial U \subseteq V$ and both U and V are open sets for which (23.1) holds, then the open set $V \setminus (U \cup \partial U)$ consisting of what is left in V after deleting U along with its boundary also satisfies (23.1). Roughly speaking, you can drill holes in a region for which (23.1) holds and get another region for which this continues to hold provided (23.1) holds for the holes. To see why this is so, consider the following picture which typifies the situation just described.



Then

$$\begin{aligned} \int_{\partial V} \mathbf{F} \cdot d\mathbf{R} &= \int_V \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA \\ &= \int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA + \int_{V \setminus U} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA \\ &= \int_{\partial U} \mathbf{F} \cdot d\mathbf{R} + \int_{V \setminus U} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA \end{aligned}$$

and so

$$\int_{V \setminus U} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \int_{\partial V} \mathbf{F} \cdot d\mathbf{R} - \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$$

which equals

$$\int_{\partial(V \setminus U)} \mathbf{F} \cdot d\mathbf{R}$$

where ∂V is oriented as shown in the picture. (If you walk around the region $V \setminus U$ with the area on the left, you get the indicated orientation for this curve.)

You can see that (23.1) is valid quite generally. This verifies the following theorem.

Theorem 23.1.3 (Green's Theorem) Let U be an open set in the plane and 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 it is often¹ the case that

$$\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left(\frac{\partial Q}{\partial x}(x, y) - \frac{\partial P}{\partial y}(x, y) \right) dA.$$

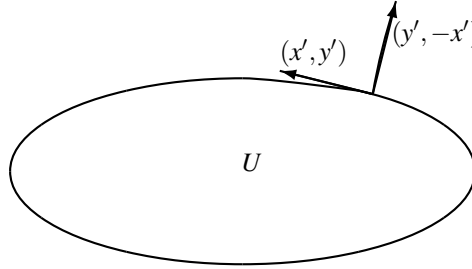
¹For a general version see the advanced calculus book by Apostol. The general versions involve the concept of a rectifiable (finite length) Jordan curve.

Here is an alternate proof of Green's theorem from the divergence theorem.

Theorem 23.1.4 (Green's Theorem) *Let U be an open set in the plane and 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 it is often the case that*

$$\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left(\frac{\partial Q}{\partial x}(x, y) - \frac{\partial P}{\partial y}(x, y) \right) dA.$$

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 23.1.5 Let U be an open set in \mathbb{R}^2 for which Green's theorem holds. Then

$$\text{Area of } U = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$$

where $\mathbf{F}(x, y) = \frac{1}{2}(-y, x)$, $(0, x)$, or $(-y, 0)$.

Proof: This follows immediately from Green's theorem. ■

Example 23.1.6 Use Proposition 23.1.5 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 23.1.5,

$$\begin{aligned} \text{Area equals} &= \frac{1}{2} \int_0^{2\pi} (-b \sin t, a \cos t) \cdot (-a \sin t, b \cos t) dt \\ &= \frac{1}{2} \int_0^{2\pi} (ab) dt = \pi ab. \end{aligned}$$

Example 23.1.7 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 23.1.8 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 23.1.9 Find $\int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$ where U is the set

$$\{(x, y) : 2 \leq x \leq 4, x \leq y \leq 4\}$$

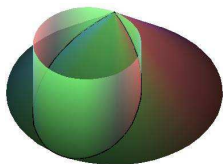
and $\mathbf{F}(x, y) = (x \sin y, y \sin x)$.

From Green's theorem, this line integral equals

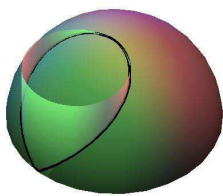
$$\int_2^4 \int_x^4 (y \cos x - x \cos y) dy dx = 4 \cos 2 - 8 \cos 4 - 8 \sin 2 - 4 \sin 4.$$

23.2 Exercises

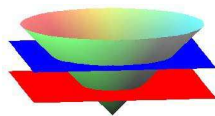
1. Find $\int_S x dS$ where S is the surface which results from the intersection of the cone $z = 2 - \sqrt{x^2 + y^2}$ with the cylinder $x^2 + y^2 - 2x = 0$.



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 $z = 2$.



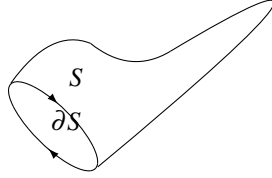
Find $\int_S \mathbf{F} \cdot \mathbf{n} dS$.

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, $(0, 0) \rightarrow (1, 0) \rightarrow (1, 1) \rightarrow (0, 0)$.

23.3 Stokes'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.

Definition 23.3.1 *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 23.3.2 *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 .*

$$(\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})(\mathbf{R}(u, v)) = ((\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u)(u, v). \quad (23.3)$$

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 (23.3). ■

The proof of Stoke's theorem given next follows [11]. First, it is convenient to give a definition.

Definition 23.3.3 *A vector valued function $\mathbf{R} : U \subseteq \mathbb{R}^m \rightarrow \mathbb{R}^n$ is said to be in $C^k(\bar{U}, \mathbb{R}^n)$ if it is the restriction to \bar{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 23.3.4 (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

$$\begin{aligned} S &\equiv \{\mathbf{R}(u, v) : (u, v) \in U\}, \\ \partial S &\equiv \{\mathbf{R}(u, v) : (u, v) \in \partial U\} \end{aligned}$$

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 ,

$$\begin{aligned} &\int_{\mathbf{R}(C)} \mathbf{F} \cdot d\mathbf{R} \\ &= \int_{\alpha}^{\beta} \mathbf{F}(\mathbf{R}(u(t), v(t))) \cdot (\mathbf{R}_u u'(t) + \mathbf{R}_v v'(t)) dt \\ &= \int_{\alpha}^{\beta} \mathbf{F}(\mathbf{R}(u(t), v(t))) \mathbf{R}_u(u(t), v(t)) u'(t) dt \\ &\quad + \int_{\alpha}^{\beta} \mathbf{F}(\mathbf{R}(u(t), v(t))) \mathbf{R}_v(u(t), v(t)) v'(t) dt \\ &= \int_C ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v) \cdot d\mathbf{r}. \end{aligned}$$

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 23.3.2, 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

$$\begin{aligned} (\mathbf{R}_u \times \mathbf{R}_v) dA &= \frac{(\mathbf{R}_u \times \mathbf{R}_v)}{|(\mathbf{R}_u \times \mathbf{R}_v)|} |(\mathbf{R}_u \times \mathbf{R}_v)| dA \\ &= \mathbf{n} dS. \end{aligned}$$

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.

23.3.1 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 $(u, v) \in Q$,

$$\mathbf{R}(u, v) \equiv u\mathbf{a} + v\mathbf{b} + \mathbf{p},$$

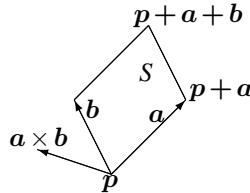
the point \mathbf{p} being a corner of the parallelogram S . Then orient ∂S consistent with the counter clockwise orientation on ∂Q . Thus, following this orientation on S you go from \mathbf{p} to $\mathbf{p} + \mathbf{a}$ to $\mathbf{p} + \mathbf{a} + \mathbf{b}$ to $\mathbf{p} + \mathbf{b}$ to \mathbf{p} . 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

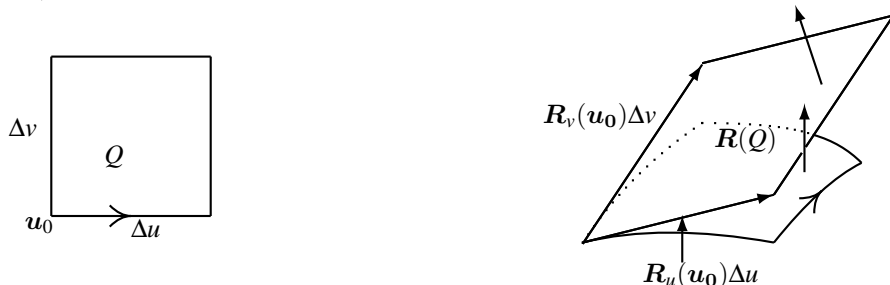
$$\mathbf{n} = \mathbf{R}_u \times \mathbf{R}_v / |\mathbf{R}_u \times \mathbf{R}_v| = \mathbf{a} \times \mathbf{b} / |\mathbf{a} \times \mathbf{b}|.$$

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.

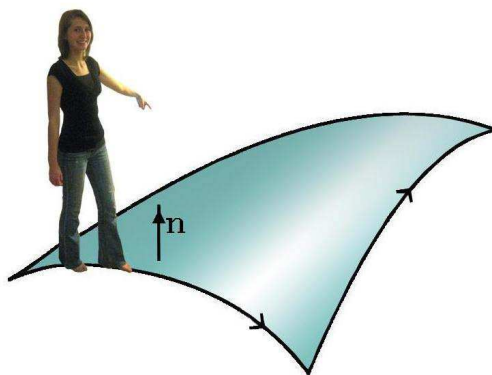
More generally, if S is a surface which is not necessarily a parallelogram but is instead as described in Theorem 23.3.4, you could consider a **small** rectangle Q contained in U and orient the boundary of $\mathbf{R}(Q)$ consistent with the counter clockwise orientation on ∂Q . Then if Q is small enough, 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.



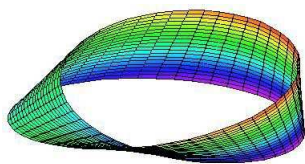
This also illustrates that you can **define** an orientation for ∂S by specifying a field of unit normal vectors for the surface, which varies continuously over the surface, and require that the motion over the boundary of the surface is such that your head points roughly in the direction of nearby normal 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.

23.3.2 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

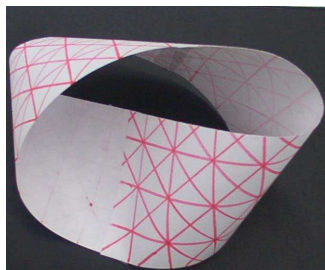
$$\begin{aligned} D = & \left(16 \sin^2 \left(\frac{\theta}{2} \right) + \frac{v^2}{2} + 4 \sin \left(\frac{\theta}{2} \right) v (\sin \theta - \cos \theta) \right. \\ & \left. + 4^3 \cos^2 \left(\frac{\theta}{2} \right) \left(\cos \left(\frac{1}{2} \theta \right) \sin \left(\frac{1}{2} \theta \right) + \cos^2 \left(\frac{1}{2} \theta \right) - \frac{1}{2} \right)^2 \right) \end{aligned}$$

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 Mobius 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.

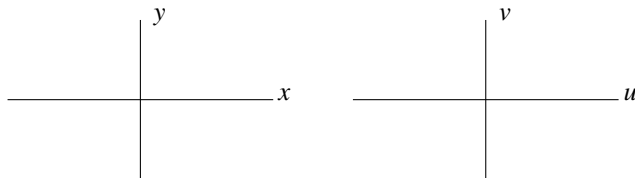


23.4 A General Green's Theorem

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 $\mathbf{R}_u \times \mathbf{R}_v$, the normal vector to V is

$$\left\| \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} \right\| \mathbf{k}$$

Suppose

$$\begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} > 0$$

so the unit normal is then just \mathbf{k} . Then Stoke's theorem applied to this special case yields

$$\int_{\partial V} \mathbf{F} \cdot d\mathbf{R} = \int_U (Q_x(x(u,v), y(u,v)) - P_x(x(u,v), y(u,v))) \mathbf{k} \cdot \mathbf{k} \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} dA$$

Now by the change of variables formula, this equals

$$= \int_V (Q_x(x,y) - P_x(x,y)) dA$$

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 23.4.1 (Green's Theorem) *Let V be an open set in the plane and let ∂V be piecewise smooth and let $\mathbf{F}(x,y) = (P(x,y), Q(x,y))$ be a C^1 vector field defined near V . Then if V is oriented counter clockwise, it is often² the case that*

$$\int_{\partial V} \mathbf{F} \cdot d\mathbf{R} = \int_V \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y) \right) dA. \quad (23.4)$$

In particular, if there exists U such as the simple convex in both directions case considered earlier for which Green's theorem holds, and $V = \mathbf{R}(U)$ where $\mathbf{R}: U \rightarrow V$ is $C^2(\overline{U}, \mathbb{R}^2)$ 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 23.4 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 of what will be of interest.

23.4.1 Conservative Vector Fields

Definition 23.4.2 *A vector field \mathbf{F} defined in a three dimensional region is said to be **conservative**³ if for every piecewise smooth closed curve C , it follows $\int_C \mathbf{F} \cdot d\mathbf{R} = 0$.*

Definition 23.4.3 *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.*

²For a general version see the advanced calculus book by Apostol. This is presented in the next section also. 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.

³There is no such thing as a liberal vector field.

Conservative vector fields are important because of the following theorem, sometimes called the fundamental theorem for line integrals.

Theorem 23.4.4 *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*

$$\int_C \mathbf{F} \cdot d\mathbf{R} = \phi(\mathbf{y}) - \phi(\mathbf{x}). \quad (23.5)$$

*Thus the line integral is path independent in this case. This function ϕ is called a **scalar potential** for \mathbf{F} .*

Proof: To save space and fussing over things which are unimportant, denote by $\mathbf{p}(x_0, x)$ a polygonal curve from x_0 to x . Thus the orientation is such that it goes from x_0 to x . The curve $\mathbf{q}(x, x_0)$ denotes the same set of points but in the opposite order. Suppose first \mathbf{F} is conservative. Fix $x_0 \in U$ and let

$$\phi(x) \equiv \int_{\mathbf{p}(x_0, x)} \mathbf{F} \cdot d\mathbf{R}.$$

This is well defined because if $\mathbf{q}(x_0, x)$ is another polygonal curve joining x_0 to x , Then the curve obtained by following $\mathbf{p}(x_0, x)$ from x_0 to x and then from x to x_0 along $\mathbf{q}(x, x_0)$ is a closed piecewise smooth curve and so by assumption, the line integral along this closed curve equals 0. However, this integral is just

$$\int_{\mathbf{p}(x_0, x)} \mathbf{F} \cdot d\mathbf{R} + \int_{\mathbf{q}(x, x_0)} \mathbf{F} \cdot d\mathbf{R} = \int_{\mathbf{p}(x_0, x)} \mathbf{F} \cdot d\mathbf{R} - \int_{\mathbf{q}(x_0, x)} \mathbf{F} \cdot d\mathbf{R}$$

which shows

$$\int_{\mathbf{p}(x_0, x)} \mathbf{F} \cdot d\mathbf{R} = \int_{\mathbf{q}(x_0, x)} \mathbf{F} \cdot d\mathbf{R}$$

and that ϕ is well defined. For small t ,

$$\begin{aligned} \frac{\phi(\mathbf{x} + t\mathbf{e}_i) - \phi(\mathbf{x})}{t} &= \frac{\int_{\mathbf{p}(x_0, \mathbf{x} + t\mathbf{e}_i)} \mathbf{F} \cdot d\mathbf{R} - \int_{\mathbf{p}(x_0, \mathbf{x})} \mathbf{F} \cdot d\mathbf{R}}{t} \\ &= \frac{\int_{\mathbf{p}(x_0, \mathbf{x})} \mathbf{F} \cdot d\mathbf{R} + \int_{\mathbf{p}(\mathbf{x}, \mathbf{x} + t\mathbf{e}_i)} \mathbf{F} \cdot d\mathbf{R} - \int_{\mathbf{p}(x_0, \mathbf{x})} \mathbf{F} \cdot d\mathbf{R}}{t}. \end{aligned}$$

Since U is open, for small t , the ball of radius $|t|$ centered at \mathbf{x} is contained in U . Therefore, the line segment from \mathbf{x} to $\mathbf{x} + t\mathbf{e}_i$ is also contained in U and so one can take $\mathbf{p}(\mathbf{x}, \mathbf{x} + t\mathbf{e}_i)(s) = \mathbf{x} + s(t\mathbf{e}_i)$ for $s \in [0, 1]$. Therefore, the above difference quotient reduces to

$$\begin{aligned} \frac{1}{t} \int_0^1 \mathbf{F}(\mathbf{x} + s(t\mathbf{e}_i)) \cdot t\mathbf{e}_i ds &= \int_0^1 F_i(\mathbf{x} + s(t\mathbf{e}_i)) ds \\ &= F_i(\mathbf{x} + s_t(t\mathbf{e}_i)) \end{aligned}$$

by the mean value theorem for integrals. Here s_t is some number between 0 and 1. By continuity of \mathbf{F} , this converges to $F_i(\mathbf{x})$ as $t \rightarrow 0$. Therefore, $\nabla\phi = \mathbf{F}$ as claimed.

Conversely, if $\nabla\phi = \mathbf{F}$, then if $\mathbf{R}: [a, b] \rightarrow \mathbb{R}^p$ is any C^1 curve joining \mathbf{x} to \mathbf{y} ,

$$\begin{aligned} \int_a^b \mathbf{F}(\mathbf{R}(t)) \cdot \mathbf{R}'(t) dt &= \int_a^b \nabla\phi(\mathbf{R}(t)) \cdot \mathbf{R}'(t) dt \\ &= \int_a^b \frac{d}{dt} (\phi(\mathbf{R}(t))) dt \\ &= \phi(\mathbf{R}(b)) - \phi(\mathbf{R}(a)) \\ &= \phi(\mathbf{y}) - \phi(\mathbf{x}) \end{aligned}$$

and this verifies (23.5) 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 \mathbf{x} to \mathbf{p} and then from \mathbf{p} to \mathbf{y} , the above would imply the integral over the curve from \mathbf{x} to \mathbf{p} is $\phi(\mathbf{p}) - \phi(\mathbf{x})$ while from \mathbf{p} to \mathbf{y} the integral would yield $\phi(\mathbf{y}) - \phi(\mathbf{p})$. Adding these gives $\phi(\mathbf{y}) - \phi(\mathbf{x})$. The formula (23.5) implies the line integral over any closed curve equals zero because the starting and ending points of such a curve are the same. ■

Example 23.4.5 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

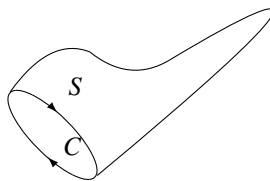
$$\sin\left(\frac{\pi}{2}\right) + 3 \cos(\pi) - (\sin(\pi) + \cos(\pi)) = -1.$$

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 23.4.6 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.

Definition 23.4.7 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 23.4.8 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 23.4.9 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 $\phi(x, y, z) = y + x^4 + \sin(x^2 + z^2)$.

23.4.2 Some Terminology

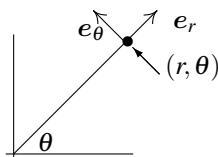
If $\mathbf{F} = (P, Q, R)$ is a vector field. Then the statement that \mathbf{F} is conservative is the same as saying the differential form $Pdx + Qdy + Rdz$ is exact. Some people like to say things in terms of vector fields and some say it in terms of differential forms. In Example [23.4.9](#), the differential form $(4x^3 + 2(\cos(x^2 + z^2))x)dx + dy + (2(\cos(x^2 + z^2))z)dz$ is exact.

Chapter 24

Moving Coordinate Systems

24.1 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 $e_r \times e_\theta = k$ 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_θ, 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

$$e_\theta = \frac{de_r}{d\theta}, \quad e_r = -\frac{de_\theta}{d\theta}. \quad (24.1)$$

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 24.1,

$$\frac{de_r}{dt} = \theta'(t) e_\theta, \quad \frac{de_\theta}{dt} = -\theta'(t) e_r \quad (24.2)$$

Using 24.2 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.

$$\begin{aligned} \mathbf{r}''(t) &= r''(t) e_r + r'(t) \frac{de_r}{dt} + r'(t) \theta'(t) e_\theta + r(t) \theta''(t) e_\theta + r(t) \theta'(t) \frac{de_\theta}{dt} \\ &= r''(t) e_r + 2r'(t) \theta'(t) e_\theta + r(t) \theta''(t) e_\theta + r(t) \theta'(t) (-e_r) \theta'(t) \\ &= \left(r''(t) - r(t) \theta'(t)^2 \right) e_r + \left(2r'(t) \theta'(t) + r(t) \theta''(t) \right) e_\theta. \end{aligned} \quad (24.3)$$

This is a very profound formula. Consider the following examples.

Example 24.1.1 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 24.3 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 24.3.

Example 24.1.2 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 24.3

$$m\mathbf{r}''(t) = -mr(t) \omega^2 e_r + m2v\omega e_\theta.$$

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.

24.2 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 24.2.1 *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.*

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.)

24.2.1 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 24.2.1 says the object moves in a plane. From the assumption that the force field is a central force field, it follows from 24.3 that

$$2r'(t)\theta'(t) + r(t)\theta''(t) = 0$$

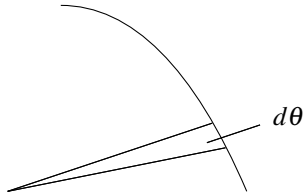
Multiply both sides of this equation by r . This yields

$$2rr'\theta' + r^2\theta'' = (r^2\theta')' = 0. \quad (24.4)$$

Consequently,

$$r^2\theta' = c \quad (24.5)$$

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,

$$\frac{dA}{dt} = \frac{1}{2} r^2 \frac{d\theta}{dt} = \frac{c}{2}. \quad (24.6)$$

24.2.2 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 24.2.1, the motion is in a plane. Now from 24.3 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

$$r''(t) - r(t)\theta'(t)^2 = -k \left(\frac{1}{r^2} \right), \quad 2r'(t)\theta'(t) + r(t)\theta''(t) = 0. \quad (24.7)$$

As in 24.4, $(r^2\theta')' = 0$ and so there exists a constant c , such that

$$r^2\theta' = c. \quad (24.8)$$

Now the other part of 24.7 and 24.8 implies

$$r''(t) - r(t) \theta'(t)^2 = r''(t) - r(t) \left(\frac{c^2}{r^4} \right) = -k \left(\frac{1}{r^2} \right). \quad (24.9)$$

It is only r as a function of θ which is of interest. Using the chain rule,

$$r' = \frac{dr}{d\theta} \frac{d\theta}{dt} = \frac{dr}{d\theta} \left(\frac{c}{r^2} \right) \quad (24.10)$$

and so also

$$\begin{aligned} r'' &= \frac{d^2 r}{d\theta^2} \left(\frac{d\theta}{dt} \right) \left(\frac{c}{r^2} \right) + \frac{dr}{d\theta} (-2) (c) (r^{-3}) \frac{dr}{d\theta} \frac{d\theta}{dt} \\ &= \frac{d^2 r}{d\theta^2} \left(\frac{c}{r^2} \right)^2 - 2 \left(\frac{dr}{d\theta} \right)^2 \left(\frac{c^2}{r^5} \right) \end{aligned} \quad (24.11)$$

Using 24.11 and 24.10 in 24.9 yields

$$\frac{d^2 r}{d\theta^2} \left(\frac{c}{r^2} \right)^2 - 2 \left(\frac{dr}{d\theta} \right)^2 \left(\frac{c^2}{r^5} \right) - r(t) \left(\frac{c^2}{r^4} \right) = -k \left(\frac{1}{r^2} \right).$$

Now multiply both sides of this equation by r^4/c^2 to obtain

$$\frac{d^2 r}{d\theta^2} - 2 \left(\frac{dr}{d\theta} \right)^2 \frac{1}{r} - r = \frac{-kr^2}{c^2}. \quad (24.12)$$

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

$$\frac{dr}{d\theta} = (-1) \rho^{-2} \frac{d\rho}{d\theta}, \quad \frac{d^2 r}{d\theta^2} = 2\rho^{-3} \left(\frac{d\rho}{d\theta} \right)^2 + (-1) \rho^{-2} \frac{d^2 \rho}{d\theta^2}.$$

Substituting this in to 24.12 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

$$\frac{d^2 \rho}{d\theta^2} + \rho = \frac{k}{c^2}, \quad (24.13)$$

which is a much nicer differential equation. Let $R = \rho - \frac{k}{c^2}$. Then in terms of R , this differential equation is

$$\frac{d^2 R}{d\theta^2} + R = 0.$$

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

$$\left(\frac{dR}{d\theta} \right)^2 + R^2 = \delta^2 \quad (24.14)$$

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 24.14 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

$$\varepsilon = (c^2/k) \delta \text{ and } p = c^2/k\varepsilon. \quad (24.15)$$

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

$$x^2 + (1 - \varepsilon^2) y^2 = \varepsilon^2 p^2 - 2p\varepsilon^2 y. \quad (24.16)$$

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.

24.2.3 Kepler's Third Law

Kepler's third law involves the time it takes for the planet to orbit the sun. From 24.16 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

$$x^2 / \left(\frac{\varepsilon^2 p^2}{1 - \varepsilon^2} \right) + \left(y + \frac{p\varepsilon^2}{1 - \varepsilon^2} \right)^2 / \left(\frac{\varepsilon^2 p^2}{(1 - \varepsilon^2)^2} \right) = 1. \quad (24.17)$$

Now note this is the equation of an ellipse and that the diameter of this ellipse is

$$\frac{2\varepsilon p}{(1 - \varepsilon^2)} \equiv 2a. \quad (24.18)$$

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 24.6 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 24.15, recalling that $k = GM$, and 24.18,

$$T^2 = \frac{4\pi^2 \varepsilon^4 p^4}{k \varepsilon p (1 - \varepsilon^2)^3} = \frac{4\pi^2 (\varepsilon p)^3}{k (1 - \varepsilon^2)^3} = \frac{4\pi^2 a^3}{k} = \frac{4\pi^2 a^3}{GM}.$$

Written more memorably, this has shown

$$T^2 = \frac{4\pi^2}{GM} \left(\frac{\text{diameter of ellipse}}{2} \right)^3. \quad (24.19)$$

This relationship is known as Kepler's third law.

24.3 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 24.3.1 *The following hold. Whenever $\mathbf{r}(t), \mathbf{s}(t)$ are two vectors from $\{\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)\}$,*

$$\mathbf{r}(t) \cdot \mathbf{s}'(t) = -\mathbf{r}'(t) \cdot \mathbf{s}(t)$$

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 24.3.2 *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

$$\mathbf{i}'(t) \cdot \mathbf{j}(t) = \boldsymbol{\Omega}(t) \cdot \mathbf{i}(t) \times \mathbf{j}(t) = \boldsymbol{\Omega}(t) \cdot \mathbf{k}(t) = \Omega_k(t)$$

Thus the component of $\boldsymbol{\Omega}(t)$ in the direction $\mathbf{k}(t)$ is determined. Next,

$$\mathbf{i}'(t) \cdot \mathbf{k}(t) = \boldsymbol{\Omega}(t) \cdot \mathbf{i}(t) \times \mathbf{k}(t) = -\Omega_j(t)$$

and so the component in the direction $\mathbf{j}(t)$ is also determined. Next,

$$\mathbf{j}'(t) \cdot \mathbf{k}(t) = \boldsymbol{\Omega}(t) \cdot \mathbf{j}(t) \times \mathbf{k}(t) = \boldsymbol{\Omega}(t) \cdot (\mathbf{j}(t) \times \mathbf{k}(t)) = \Omega_i(t)$$

so the component of $\Omega(t)$ in direction $i(t)$ is determined. Thus, if there is such an angular velocity vector, it must be of the form

$$\Omega(t) \equiv (j'(t) \cdot k(t)) i(t) - (i'(t) \cdot k(t)) j(t) + (i'(t) \cdot j(t)) k(t)$$

It only remains to verify that this vector works. Recall Lemma 24.3.1 which will be used without comment in what follows. Does the above $\Omega(t)$ work?

$$\begin{aligned} \Omega(t) \times i(t) &= (i'(t) \cdot k(t)) k(t) \\ &\quad + (i'(t) \cdot j(t)) j(t) + \left(\overbrace{i'(t) \cdot i(t)}^{=0} \right) i(t) \\ &= i'(t) \end{aligned}$$

$$\begin{aligned} \Omega(t) \times j(t) &= (j'(t) \cdot k(t)) k(t) + (i'(t) \cdot j(t)) (-i(t)) \\ &= (j'(t) \cdot k(t)) k(t) + (i(t) \cdot j'(t)) (i(t)) \\ &= j'(t) \end{aligned}$$

and finally,

$$\begin{aligned} \Omega(t) \times k(t) &= (j'(t) \cdot k(t)) (-j(t)) - (i'(t) \cdot k(t)) i(t) \\ &= (j(t) \cdot k'(t)) (j(t)) + (i(t) \cdot k'(t)) i(t) \\ &= k'(t) \end{aligned}$$

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 24.3.3 *Let $u(t)$ be a vector such that its components with respect to the basis vectors $i(t), j(t), k(t)$ are constant. Then $u'(t) = \Omega(t) \times u(t)$.*

Proof: Say $u(t) = u_i i(t) + u_j j(t) + u_k k(t)$. Then

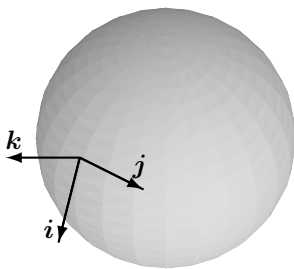
$$\begin{aligned} u'(t) &= u_i i'(t) + u_j j'(t) + u_k k'(t) \\ &= u_i \Omega(t) \times i(t) + u_j \Omega(t) \times j(t) \\ &\quad + u_k \Omega(t) \times k(t) \\ &= \Omega(t) \times (u_i i(t) + u_j j(t) + u_k k(t)) \\ &= \Omega(t) \times u(t) \quad \blacksquare \end{aligned}$$

24.4 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. In this, the motion of the earth through space around the sun is not being considered because forces resulting from this motion are negligible.

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 24.3.2 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 464 below for a more analytical explanation.

24.5 Coriolis Force and Centripetal Force

Let $\mathbf{p}(t)$ be a point which has constant components relative to the moving coordinate system described above $\{\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)\}$. For example, it could be a single point on the rotating earth. Letting $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ be a typical rectangular coordinate system fixed in space and let $\mathbf{R}(t)$ be the position vector of $\mathbf{p}(t)$ 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 $\mathbf{R}(t)$ with respect to the moving coordinate system are constants. 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

$$\mathbf{r}_B''(t) \equiv \mathbf{a}_B(t) = x''(t)\mathbf{i}(t) + y''(t)\mathbf{j}(t) + z''(t)\mathbf{k}(t)$$

and the perceived velocity would be $\mathbf{r}_B'(t) \equiv \mathbf{v}_B(t)$.

$$\mathbf{v}_B(t) \equiv x'(t)\mathbf{i}(t) + y'(t)\mathbf{j}(t) + z'(t)\mathbf{k}(t)$$

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,

$$\begin{aligned} \mathbf{v}(t) &= \mathbf{R}'(t) + \mathbf{r}_B'(t), \quad \mathbf{r}_B'(t) = \mathbf{v}_B(t) + x(t)\mathbf{i}'(t) + y(t)\mathbf{j}'(t) + z(t)\mathbf{k}'(t) \\ &= \mathbf{v}_B(t) + x(t)(\boldsymbol{\Omega}(t) \times \mathbf{i}(t)) + y(t)(\boldsymbol{\Omega}(t) \times \mathbf{j}(t)) + z(t)(\boldsymbol{\Omega}(t) \times \mathbf{k}(t)) \end{aligned}$$

and so, from the last equation for $\mathbf{r}_B'(t)$,

$$\begin{aligned} \mathbf{v}(t) &= \mathbf{v}_B(t) + \boldsymbol{\Omega}(t) \times \mathbf{r}_B(t) + \boldsymbol{\Omega}(t) \times \mathbf{R}(t) \\ &= \mathbf{v}_B(t) + \boldsymbol{\Omega}(t) \times \mathbf{r}(t) \end{aligned}$$

Now take a further derivative to find the total acceleration. Using what was just shown, it equals

$$\begin{aligned} \mathbf{a}(t) &= \mathbf{R}''(t) + \frac{d^2\mathbf{r}_B}{dt^2}(t) = \mathbf{a}_B(t) + \boldsymbol{\Omega}(t) \times \mathbf{v}_B(t) + \boldsymbol{\Omega}'(t) \times \mathbf{r}(t) + \boldsymbol{\Omega}(t) \times \mathbf{v}(t) \\ &= \mathbf{a}_B(t) + (\boldsymbol{\Omega}(t) \times \mathbf{v}_B(t)) + (\boldsymbol{\Omega}'(t) \times \mathbf{r}(t)) + \boldsymbol{\Omega}(t) \times (\mathbf{v}_B(t) + \boldsymbol{\Omega}(t) \times \mathbf{r}(t)) \\ &= \mathbf{a}_B(t) + 2(\boldsymbol{\Omega}(t) \times \mathbf{v}_B(t)) + (\boldsymbol{\Omega}'(t) \times \mathbf{r}(t)) + \boldsymbol{\Omega}(t) \times (\boldsymbol{\Omega}(t) \times \mathbf{r}(t)) \\ &= \mathbf{a}_B(t) + 2(\boldsymbol{\Omega}(t) \times \mathbf{v}_B(t)) + (\boldsymbol{\Omega}'(t) \times \mathbf{r}(t)) + \boldsymbol{\Omega}(t) \times (\boldsymbol{\Omega}(t) \times \mathbf{r}_B(t)) \\ &\quad + \boldsymbol{\Omega}(t) \times (\boldsymbol{\Omega}(t) \times \mathbf{R}(t)) \tag{24.20} \\ &= \mathbf{a}_B(t) + 2(\boldsymbol{\Omega}(t) \times \mathbf{v}_B(t)) + (\boldsymbol{\Omega}'(t) \times \mathbf{r}_B(t)) + \boldsymbol{\Omega}(t) \times (\boldsymbol{\Omega}(t) \times \mathbf{r}_B(t)) \\ &\quad + \boldsymbol{\Omega}(t) \times \mathbf{R}'(t) + \boldsymbol{\Omega}'(t) \times \mathbf{R}(t) \end{aligned}$$

$$\begin{aligned}
&= \mathbf{a}_B(t) + 2(\boldsymbol{\Omega}(t) \times \mathbf{v}_B(t)) + (\boldsymbol{\Omega}'(t) \times \mathbf{r}_B(t)) \\
&\quad + \boldsymbol{\Omega}(t) \times (\boldsymbol{\Omega}(t) \times \mathbf{r}_B(t)) + \frac{d}{dt}(\boldsymbol{\Omega}(t) \times \mathbf{R}(t)) \\
&= \mathbf{a}_B(t) + 2(\boldsymbol{\Omega}(t) \times \mathbf{v}_B(t)) + (\boldsymbol{\Omega}'(t) \times \mathbf{r}_B(t)) \\
&\quad + \boldsymbol{\Omega}(t) \times (\boldsymbol{\Omega}(t) \times \mathbf{r}_B(t)) + \mathbf{R}''(t)
\end{aligned}$$

Therefore,

$$\frac{d^2 \mathbf{r}_B}{dt^2}(t) = \mathbf{a}_B(t) + 2(\boldsymbol{\Omega}(t) \times \mathbf{v}_B(t)) + (\boldsymbol{\Omega}'(t) \times \mathbf{r}_B(t)) + \boldsymbol{\Omega}(t) \times (\boldsymbol{\Omega}(t) \times \mathbf{r}_B(t))$$

where recall that $\mathbf{a}_B(t)$ is the perceived acceleration relative to the moving coordinate system. Solving for this yields

$$\frac{d^2 \mathbf{r}_B}{dt^2}(t) - (2(\boldsymbol{\Omega}(t) \times \mathbf{v}_B(t)) + (\boldsymbol{\Omega}'(t) \times \mathbf{r}_B(t)) + \boldsymbol{\Omega}(t) \times (\boldsymbol{\Omega}(t) \times \mathbf{r}_B(t))) = \mathbf{a}_B(t)$$

The part of the acceleration on the left depending on the relative velocity is called the Coriolis acceleration. The rest of it is sometimes called centrifugal acceleration. It is felt by the observer by regarding the moving coordinates as fixed. On the earth, this force is small enough to be neglected. However, when \mathbf{v}_B is large, one can get a significant contribution from the Coriolis force.

24.6 Coriolis Force on the Rotating Earth

As shown above, on the rotating earth, $\boldsymbol{\Omega}$ is a constant and so 24.20 reduces to

$$\mathbf{a} = \mathbf{a}_B(t) + 2(\boldsymbol{\Omega}(t) \times \mathbf{v}_B(t)) + \boldsymbol{\Omega}(t) \times (\boldsymbol{\Omega}(t) \times \mathbf{r}(t)) \quad (24.21)$$

Since $\mathbf{r}_B + \mathbf{R} = \mathbf{r}$,

$$\mathbf{a}_B = \mathbf{a} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R}) - 2\boldsymbol{\Omega} \times \mathbf{v}_B - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}_B). \quad (24.22)$$

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

$\omega \overbrace{(24)(3600)}^{\text{seconds in a day}} = 2\pi$, 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

$$(7.2722 \times 10^{-5})^2 |\mathbf{r}_B|.$$

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

$$\begin{aligned}
\mathbf{a}_B &= \mathbf{a} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R}) - 2\boldsymbol{\Omega} \times \mathbf{v}_B = \\
&\quad \overbrace{-\frac{GM(\mathbf{R} + \mathbf{r}_B)}{|\mathbf{R} + \mathbf{r}_B|^3}}^{\equiv g} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R}) - 2\boldsymbol{\Omega} \times \mathbf{v}_B \equiv \mathbf{g} - 2\boldsymbol{\Omega} \times \mathbf{v}_B.
\end{aligned}$$

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 poles and at the equator, 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 $\mathbf{k}^* = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$

$$\overbrace{\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}}^{\mathbf{k}^*} = a \overbrace{\begin{pmatrix} \cos(\phi) \cos(\theta) \\ \cos(\phi) \sin(\theta) \\ -\sin(\phi) \end{pmatrix}}^{\mathbf{i}} + b \overbrace{\begin{pmatrix} -\sin(\theta) \\ \cos(\theta) \\ 0 \end{pmatrix}}^{\mathbf{j}} + c \overbrace{\begin{pmatrix} \sin(\phi) \cos(\theta) \\ \sin(\phi) \sin(\theta) \\ \cos(\phi) \end{pmatrix}}^{\mathbf{k}} \quad (24.23)$$

The solution is $a = -\sin(\phi)$, $b = 0$, and $c = \cos(\phi)$.

Now the Coriolis acceleration on the earth equals

$$2(\boldsymbol{\Omega} \times \mathbf{v}_B) = 2\omega \left(\overbrace{-\sin(\phi) \mathbf{i} + 0 \mathbf{j} + \cos(\phi) \mathbf{k}}^{\mathbf{k}^*} \right) \times (x' \mathbf{i} + y' \mathbf{j} + z' \mathbf{k}).$$

This equals

$$2\omega [(-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k}]. \quad (24.24)$$

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 pretty small and so it will be neglected. However, the Coriolis force will not be neglected.

Example 24.6.1 Suppose a rock is dropped from a tall building. Where will it strike?

Assume $\mathbf{a} = -g\mathbf{k}$ and the 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 i and 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 j component, this is

$$2gt\omega \sin \phi.$$

Two integrations give $(\omega g t^3 / 3) \sin \phi$ for the 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.

24.7 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 stopping¹. 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 observed rotation would take place on the equator. Is it possible to predict what will take place at various latitudes?

Using 24.24, in 24.22,

$$\begin{aligned} \mathbf{a}_B &= \mathbf{a} - \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R}) \\ &\quad - 2\omega [(-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k}]. \end{aligned}$$

Neglecting the small term, $\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{R})$, this becomes

$$= -g\mathbf{k} + \mathbf{T}/m - 2\omega [(-y' \cos \phi) \mathbf{i} + (x' \cos \phi + z' \sin \phi) \mathbf{j} - (y' \sin \phi) \mathbf{k}]$$

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,

$$\mathbf{T} = -T \frac{x}{l} \mathbf{i} - T \frac{y}{l} \mathbf{j} + T \frac{l - z}{l} \mathbf{k}$$

¹There 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.

and consequently, the differential equations of relative motion are

$$x'' = -T \frac{x}{ml} + 2\omega y' \cos \phi$$

$$y'' = -T \frac{y}{ml} - 2\omega (x' \cos \phi + z' \sin \phi)$$

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

$$x'' + a^2 x = by', \quad y'' + a^2 y = -bx' \quad (24.25)$$

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 ,

$$x = c \sin\left(\frac{bt}{2}\right) \sin\left(\frac{\sqrt{b^2 + 4a^2}}{2} t\right), \quad y = c \cos\left(\frac{bt}{2}\right) \sin\left(\frac{\sqrt{b^2 + 4a^2}}{2} t\right) \quad (24.26)$$

yields a solution to 24.25 along with the initial conditions,

$$x(0) = 0, y(0) = 0, x'(0) = 0, y'(0) = \frac{c\sqrt{b^2 + 4a^2}}{2}. \quad (24.27)$$

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 24.27 will hold for some value of c and so the solution to 24.25 having these initial conditions will be those of 24.26. (Some interesting mathematical details are being ignored here. Such initial value problems as 24.26 and 24.27 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 464 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 $\lim_{\phi \rightarrow \pi/2} \sec \phi = \infty$.

24.8 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 24.5. 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 24.19 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,

$$\begin{aligned} \mathbf{i}(t) &= \begin{pmatrix} \cos(\omega t) \cos \phi \\ \cos \phi \sin(\omega t) \\ -\sin \phi \end{pmatrix}, \mathbf{j}(t) = \begin{pmatrix} -\sin(\omega t) \\ \cos(\omega t) \\ 0 \end{pmatrix}, \\ \mathbf{k}(t) &= \begin{pmatrix} \sin(\phi) \cos(\omega t) \\ \sin(\phi) \sin(\omega t) \\ \cos(\phi) \end{pmatrix} \end{aligned}$$

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

$$x'' + a^2 x = by', \quad y'' + a^2 y = -bx' \quad (24.28)$$

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 24.26 and 24.27. **Hint:** If you had two solutions to 24.26 and 24.27, \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 24.28. 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.

Chapter 25

Curvilinear Coordinates

25.1 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.

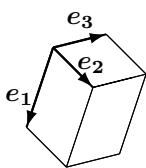


The vectors, i, j, k , are fixed. If v is a vector, there are unique scalars called components such that $v = v^1 i + v^2 j + v^3 k$. This is what it means that i, j, k is a basis. Review Section 9 at this time to see how this geometric notion relates to the general concept of a basis in a vector space.

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 25.1.1 *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 25.1.2 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

$$e^i \cdot e_j = \delta_j^i \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}. \quad (25.1)$$

Theorem 25.1.3 If $\{e_i\}_{i=1}^p$ is a basis then $\{e^i\}_{i=1}^p$ is also a basis provided 25.1 holds.

Proof: Suppose

$$v = v_i e^i. \quad (25.2)$$

Then taking the dot product of both sides of 25.2 with e_j , yields

$$v_j = v \cdot e_j. \quad (25.3)$$

Thus there is at most one choice of scalars v_j such that $v = v_j e^j$ and it is given by 25.3.

$$(v - v \cdot e_j e^j) \cdot e_k = 0$$

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 $v - v \cdot e_j e^j = 0$ 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

$$v = v^j e_j. \quad (25.4)$$

Then taking the dot product of both sides of 25.4 with e^i we see $v^i = e^i \cdot v$.

Does there exist a dual basis and is it uniquely determined?

Theorem 25.1.4 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 $e_j x^j = 0$ 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 25.1.5 $\{e_i\}_{i=1}^p$ 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 $e_i = a_i^j \mathbf{i}_j$,

$$\det(a_i^j) \neq 0. \quad (25.5)$$

If $\{e_i\}_{i=1}^p$ is a basis, then there exists a unique dual basis, $\{e^j\}_{j=1}^p$ satisfying

$$e^j \cdot e_i = \delta_i^j, \quad (25.6)$$

and that if \mathbf{v} is any vector,

$$\mathbf{v} = v_j e^j, \quad \mathbf{v} = v^j e_j. \quad (25.7)$$

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.

$$v_j = \mathbf{v} \cdot e_j, \quad v^j = \mathbf{v} \cdot e^j. \quad (25.8)$$

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 25.1.6 *The following hold.*

$$g^{ij} e_j = e^i, \quad g_{ij} e^j = e_i, \quad (25.9)$$

$$g^{ij} v_j = v^i, \quad g_{ij} v^j = v_i, \quad (25.10)$$

$$g^{ij} g_{jk} = \delta_k^i, \quad (25.11)$$

$$\det(g_{ij}) > 0, \quad \det(g^{ij}) > 0. \quad (25.12)$$

Proof: First,

$$e^i = e^i \cdot e^j e_j = g^{ij} e_j$$

by 25.7 and 25.8. Similarly, by 25.7 and 25.8,

$$e_i = e_i \cdot e_j e^j = g_{ij} e^j.$$

This verifies 25.9. To verify 25.10,

$$v^i = e^i \cdot \mathbf{v} = g^{ij} e_j \cdot \mathbf{v} = g^{ij} v_j.$$

The proof of the remaining formula in 25.10 is similar.

To verify 25.11,

$$g^{ij}g_{jk} = e^i \cdot e^j e_j \cdot e_k = ((e^i \cdot e^j) e_j) \cdot e_k = e^i \cdot e_k = \delta_k^i.$$

This shows the two determinants in 25.12 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 $e_i = a_i^j \mathbf{i}_j = b_j^i \mathbf{i}^j$, we see that since $\mathbf{i}_j = \mathbf{i}^j$, $a_i^j = b_j^i$. Therefore,

$$e_i \cdot e_j = a_i^r \mathbf{i}_r \cdot b_k^j \mathbf{i}^k = a_i^r b_k^j \delta_r^k = a_i^k b_k^j = a_i^k a_j^k.$$

It follows that for G the matrix whose ij^{th} entry is $e_i \cdot e_j$, $G = AA^T$ 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 25.11 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 25.1.7 The matrix $(g_{ij}) = G$ is called the metric tensor.

25.2 Exercises

1. Let $e_1 = \mathbf{i} + \mathbf{j}$, $e_2 = \mathbf{i} - \mathbf{j}$, $e_3 = \mathbf{j} + \mathbf{k}$. Find $e^1, e^2, e^3, (g_{ij}), (g^{ij})$. 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 $e^1 = 2\mathbf{i} + \mathbf{j}$, $e^2 = \mathbf{i} - 2\mathbf{j}$, $e^3 = \mathbf{k}$. Find $e_1, e_2, e_3, (g_{ij}), (g^{ij})$. 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 e_1, e_2, e_3 have the property that $e_i \cdot e_j = 0$ whenever $i \neq j$. Show the same is true of the dual basis and that in fact, e^i is a multiple of e_i .
4. Let e_1, \dots, e_3 be a basis for \mathbb{R}^n and let $\mathbf{v} = v^i e_i = v_i e^i$, $\mathbf{w} = w^j e_j = w_j e^j$ be two vectors. Show

$$\mathbf{v} \cdot \mathbf{w} = g_{ij} v^i w^j = g^{ij} v_i w_j.$$

5. Show if $\{e_i\}_{i=1}^3$ is a basis in \mathbb{R}^3

$$e^1 = \frac{e_2 \times e_3}{e_2 \times e_3 \cdot e_1}, e^2 = \frac{e_1 \times e_3}{e_1 \times e_3 \cdot e_2}, e^3 = \frac{e_1 \times e_2}{e_1 \times e_2 \cdot e_3}.$$

6. Let $\{e_i\}_{i=1}^n$ be a basis and define

$$e_i^* \equiv \frac{e_i}{|e_i|}, e^{*i} \equiv e^i |e_i|.$$

Show $e^{*i} \cdot e_j^* = \delta_j^i$.

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

$$v_i^* = \frac{v_i}{|\mathbf{e}_i|}, \quad v^{*i} = v^i |\mathbf{e}_i| \quad (\text{No summation on } i).$$

25.3 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

$$\mathbf{M} \text{ is } C^2, \quad (25.13)$$

$$\mathbf{M} \text{ is one to one.} \quad (25.14)$$

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.

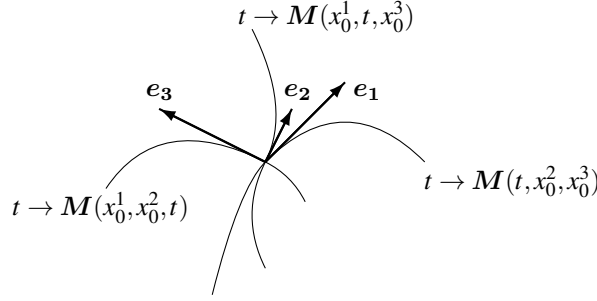
For a fixed $\mathbf{x} \in D$, we can consider the space curves,

$$t \rightarrow \mathbf{M}(\mathbf{x} + t\mathbf{i}_k) \equiv \mathbf{y}(\mathbf{x} + t\mathbf{i}_k)$$

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 25.1.5,

$$\det \left(\frac{\partial M^i}{\partial x^k} \right) \equiv \det(D\mathbf{y}(\mathbf{x})) \equiv \det(D(\mathbf{M})(\mathbf{x})) \neq 0. \quad (25.15)$$

Let

$$y^i = M^i(\mathbf{x}) \quad i = 1, \dots, n \quad (25.16)$$

so that the y^i are the usual rectangular coordinates with respect to the usual basis vectors $\{\mathbf{i}_k\}_{k=1}^n$ of the point $\mathbf{y} = \mathbf{M}(\mathbf{x})$. Letting $\mathbf{x} \equiv (x^1, \dots, x^n)$, it follows from the inverse function theorem (See Chapter 26) that $\mathbf{M}(D)$ is open, and that 25.15, 25.13, and 25.14 imply the equations 25.16 define each x^i as a C^2 function of $\mathbf{y} \equiv (y^1, \dots, y^n)$. Thus, abusing notation slightly, the equations 25.16 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,

$$\nabla x^k(\mathbf{y}) = \frac{\partial x^k(\mathbf{y})}{\partial y^j} \mathbf{i}^j.$$

Then

$$\nabla x^k(\mathbf{y}) \cdot \mathbf{e}_j = \frac{\partial x^k}{\partial y^s} \mathbf{i}^s \cdot \frac{\partial y^r}{\partial x^j} \mathbf{i}_r = \frac{\partial x^k}{\partial y^s} \frac{\partial y^s}{\partial x^j} = \delta_j^k$$

by the chain rule. Therefore, the dual basis is given by

$$\mathbf{e}^k(\mathbf{x}) = \nabla x^k(\mathbf{y}(\mathbf{x})). \quad (25.17)$$

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 \mathbf{e}^k by 25.17 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, $\mathbf{e}^k(\mathbf{x})$ may always be found by using formula 25.9 and the result is in terms of the curvilinear coordinates \mathbf{x} . Here is a familiar example.

Example 25.3.1 $D \equiv (0, \infty) \times (0, \pi) \times (0, 2\pi)$ 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

$$\mathbf{e}_1(\mathbf{x}) = \sin(x^2) \cos(x^3) \mathbf{i}_1 + \sin(x^2) \sin(x^3) \mathbf{i}_2 + \cos(x^2) \mathbf{i}_3,$$

$$\begin{aligned} \mathbf{e}_2(\mathbf{x}) &= x^1 \cos(x^2) \cos(x^3) \mathbf{i}_1 \\ &+ x^1 \cos(x^2) \sin(x^3) \mathbf{i}_2 - x^1 \sin(x^2) \mathbf{i}_3, \end{aligned}$$

$$\mathbf{e}_3(\mathbf{x}) = -x^1 \sin(x^2) \sin(x^3) \mathbf{i}_1 + x^1 \sin(x^2) \cos(x^3) \mathbf{i}_2 + 0 \mathbf{i}_3.$$

It follows the metric tensor is

$$G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (x^1)^2 & 0 \\ 0 & 0 & (x^1)^2 \sin^2(x^2) \end{pmatrix} = (g_{ij}) = (\mathbf{e}_i \cdot \mathbf{e}_j). \quad (25.18)$$

Therefore, by Theorem 25.1.6

$$\begin{aligned} G^{-1} &= (g^{ij}) \\ &= (\mathbf{e}^i, \mathbf{e}^j) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (x^1)^{-2} & 0 \\ 0 & 0 & (x^1)^{-2} \sin^{-2}(x^2) \end{pmatrix}. \end{aligned}$$

To obtain the dual basis, use Theorem 25.1.6 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, $\begin{pmatrix} y^1 & \cdots & y^n \end{pmatrix}$ are the rectangular coordinates of the point in \mathbb{R}^n .

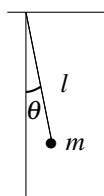
25.4 Exercises

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 $y = y(x, t)$ where t signifies time and $x \in U \subseteq \mathbb{R}^m$ for U an open set, while $y \in \mathbb{R}^n$ and suppose 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 $y = y(x(t), t)$ 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 $y = y(x, t)$. In general, the kinetic energy is defined as

$$T \equiv \frac{1}{2} m \dot{y} \cdot \dot{y} \quad (*)$$

where the dot on the top signifies differentiation with respect to t . Show

$$\frac{\partial T}{\partial \dot{x}^k} = m \dot{y} \cdot \frac{\partial y}{\partial x^k}.$$

Hint: First show

$$\dot{y} = \frac{\partial y}{\partial x^j} \dot{x}^j + \frac{\partial y}{\partial t} \quad (**)$$

and so

$$\frac{\partial \dot{y}}{\partial \dot{x}^j} = \frac{\partial y}{\partial x^j}.$$

3. \uparrow Show

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{x}^k} \right) = m \ddot{y} \cdot \frac{\partial y}{\partial x^k} + m \dot{y} \cdot \frac{\partial^2 y}{\partial x^k \partial x^r} \dot{x}^r + m \dot{y} \cdot \frac{\partial^2 y}{\partial t \partial x^k}.$$

4. \uparrow Show

$$\frac{\partial T}{\partial x^k} = m \dot{y} \cdot \left(\frac{\partial^2 y}{\partial x^r \partial x^k} \dot{x}^r + \frac{\partial^2 y}{\partial t \partial x^k} \right).$$

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 $\mathbf{y} = \mathbf{y}(\mathbf{x})$ and \mathbf{x} is a function of t . Letting G denote the metric tensor, show that the kinetic energy is of the form $\frac{1}{2}m\dot{\mathbf{x}}^T G \mathbf{x}$ 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 $\mathbf{x} = (\rho, \theta, \phi)$, 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

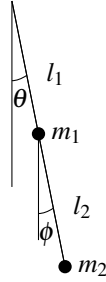
$$\mathbf{y}_\alpha = \mathbf{y}_\alpha(\mathbf{x}, t),$$

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{x}^k} \right) - \frac{\partial T}{\partial x^k} = \sum_{\alpha=1}^R \mathbf{F}_\alpha \cdot \frac{\partial \mathbf{y}_\alpha}{\partial x^k}$$

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.

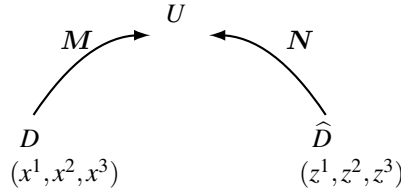


Write differential equations for θ and ϕ to describe the motion of the double pendulum.

25.5 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,

$$e_i(z) = \frac{\partial y}{\partial z^i} = \frac{\partial y}{\partial x^j} \frac{\partial x^j}{\partial z^i} = \frac{\partial x^j}{\partial z^i} e_j(x) \quad (25.19)$$

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 25.5.1 *The following transformation rules hold for pairs of curvilinear coordinates.*

$$v_i(z) = \frac{\partial x^j}{\partial z^i} v_j(x), \quad v^i(z) = \frac{\partial z^i}{\partial x^j} v^j(x), \quad (25.20)$$

$$e_i(z) = \frac{\partial x^j}{\partial z^i} e_j(x), \quad e^i(z) = \frac{\partial z^i}{\partial x^j} e^j(x), \quad (25.21)$$

$$g_{ij}(z) = \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j} g_{rs}(x), \quad g^{ij}(z) = \frac{\partial z^i}{\partial x^r} \frac{\partial z^j}{\partial x^s} g^{rs}(x). \quad (25.22)$$

Proof: We already have shown the first part of 25.21 in 25.19. Then, from 25.19,

$$\begin{aligned} e^i(z) &= e^i(z) \cdot e_j(x) e^j(x) = e^i(z) \cdot \frac{\partial z^k}{\partial x^j} e_k(z) e^j(x) \\ &= \delta_k^i \frac{\partial z^k}{\partial x^j} e^j(x) = \frac{\partial z^i}{\partial x^j} e^j(x) \end{aligned}$$

and this proves the second part of 25.21. Now to show 25.20,

$$v_i(z) = v \cdot e_i(z) = v \cdot \frac{\partial x^j}{\partial z^i} e_j(x) = \frac{\partial x^j}{\partial z^i} v \cdot e_j(x) = \frac{\partial x^j}{\partial z^i} v_j(x)$$

and

$$v^j(z) = v \cdot e^j(z) = v \cdot \frac{\partial z^i}{\partial x^j} e^i(x) = \frac{\partial z^i}{\partial x^j} v \cdot e^i(x) = \frac{\partial z^i}{\partial x^j} v^i(x).$$

To verify 25.22,

$$g_{ij}(z) = e_i(z) \cdot e_j(z) = e_r(x) \frac{\partial x^r}{\partial z^i} \cdot e_s(x) \frac{\partial x^s}{\partial z^j} = g_{rs}(x) \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j}. \blacksquare$$

25.6 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 $\frac{\partial \mathbf{F}(x)}{\partial x^j}$ is a vector and so there exist scalars, $F_{,j}^i(x)$ and $F_{i,j}(x)$ such that

$$\frac{\partial \mathbf{F}(x)}{\partial x^j} = F_{,j}^i(x) e_i(x), \quad \frac{\partial \mathbf{F}(x)}{\partial x^j} = F_{i,j}(x) e^i(x) \quad (25.23)$$

We will see how these scalars transform when the coordinates are changed.

Theorem 25.6.1 *If x and z are curvilinear coordinates,*

$$F_{,s}^r(x) = F_{,j}^i(z) \frac{\partial x^r}{\partial z^i} \frac{\partial z^j}{\partial x^s}, \quad F_{r,s}(x) \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j} = F_{i,j}(z). \quad (25.24)$$

Proof:

$$\begin{aligned} F_{,s}^r(x) e_r(x) &\equiv \frac{\partial \mathbf{F}(x)}{\partial x^s} = \frac{\partial \mathbf{F}(z)}{\partial z^j} \frac{\partial z^j}{\partial x^s} \equiv \\ F_{,j}^i(z) e_i(z) \frac{\partial z^j}{\partial x^s} &= F_{,j}^i(z) \frac{\partial z^j}{\partial x^s} \frac{\partial x^r}{\partial z^i} e_r(x) \end{aligned}$$

which shows the first formula of 25.23. To show the other formula,

$$\begin{aligned} F_{i,j}(z) e^i(z) &\equiv \frac{\partial \mathbf{F}(z)}{\partial z^j} = \frac{\partial \mathbf{F}(x)}{\partial x^s} \frac{\partial x^s}{\partial z^j} \equiv \\ F_{r,s}(x) e^r(x) \frac{\partial x^s}{\partial z^j} &= F_{r,s}(x) \frac{\partial x^s}{\partial z^j} \frac{\partial x^r}{\partial z^i} e^i(z), \end{aligned}$$

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,

$$\frac{\partial \mathbf{F}}{\partial x^j} = \frac{\partial F^i}{\partial x^j} \mathbf{e}_i(\mathbf{x}) + F^i(\mathbf{x}) \frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j}. \quad (25.25)$$

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

$$\left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} \mathbf{e}_k(\mathbf{x}) \cdot \mathbf{e}^r(\mathbf{x}) = \left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} \delta_k^r = \left\{ \begin{smallmatrix} r \\ ij \end{smallmatrix} \right\} = \frac{\partial^2 \mathbf{y}}{\partial x^j \partial x^i} \cdot \mathbf{e}^r(\mathbf{x}) \quad (25.26)$$

Therefore, from 25.25, $\frac{\partial \mathbf{F}}{\partial x^j} = \frac{\partial F^k}{\partial x^j} \mathbf{e}_k(\mathbf{x}) + F^i(\mathbf{x}) \left\{ \begin{smallmatrix} r \\ ij \end{smallmatrix} \right\} \mathbf{e}_k(\mathbf{x})$ which shows

$$F_{,j}^k(\mathbf{x}) = \frac{\partial F^k}{\partial x^j} + F^i(\mathbf{x}) \left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\}. \quad (25.27)$$

This is sometimes called the covariant derivative.

Theorem 25.6.2 *The Christoffel symbols of the second kind satisfy the following*

$$\frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j} = \left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} \mathbf{e}_k(\mathbf{x}), \quad (25.28)$$

$$\frac{\partial \mathbf{e}^i(\mathbf{x})}{\partial x^j} = - \left\{ \begin{smallmatrix} i \\ kj \end{smallmatrix} \right\} \mathbf{e}^k(\mathbf{x}), \quad (25.29)$$

$$\left\{ \begin{smallmatrix} k \\ ij \end{smallmatrix} \right\} = \left\{ \begin{smallmatrix} k \\ ji \end{smallmatrix} \right\}, \quad (25.30)$$

$$\left\{ \begin{smallmatrix} m \\ ik \end{smallmatrix} \right\} = \frac{g^{jm}}{2} \left[\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} \right]. \quad (25.31)$$

Proof: Formula 25.28 is the definition of the Christoffel symbols. We verify 25.29 next. To do so, note

$$\mathbf{e}^i(\mathbf{x}) \cdot \mathbf{e}_k(\mathbf{x}) = \delta_k^i.$$

Then from the product rule,

$$\frac{\partial e^i(x)}{\partial x^j} \cdot e_k(x) + e^i(x) \cdot \frac{\partial e_k(x)}{\partial x^j} = 0.$$

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 25.29. Formula 25.30 follows from 25.26 and equality of mixed partial derivatives.

It remains to show 25.31.

$$\frac{\partial g_{ij}}{\partial x^k} = \frac{\partial e_i}{\partial x^k} \cdot e_j + e_i \cdot \frac{\partial e_j}{\partial x^k} = \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} e_r \cdot e_j + e_i \cdot e_r \left\{ \begin{matrix} r \\ jk \end{matrix} \right\}.$$

Therefore,

$$\frac{\partial g_{ij}}{\partial x^k} = \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} g_{rj} + \left\{ \begin{matrix} r \\ jk \end{matrix} \right\} g_{ri}. \quad (25.32)$$

Switching i and k while remembering 25.30 yields

$$\frac{\partial g_{kj}}{\partial x^i} = \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} g_{rj} + \left\{ \begin{matrix} r \\ ji \end{matrix} \right\} g_{rk}. \quad (25.33)$$

Now switching j and k in 25.32,

$$\frac{\partial g_{ik}}{\partial x^j} = \left\{ \begin{matrix} r \\ ij \end{matrix} \right\} g_{rk} + \left\{ \begin{matrix} r \\ jk \end{matrix} \right\} g_{ri}. \quad (25.34)$$

Adding 25.32 to 25.33 and subtracting 25.34 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 25.1.6 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 25.31. ■

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.

25.7 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

$$\mathbf{y} \equiv y^k \mathbf{i}_k, \mathbf{e}_k(\mathbf{y}) = \mathbf{i}_k = \mathbf{e}^k(\mathbf{y}).$$

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} ,

$$\nabla \phi(\mathbf{x}) = \nabla \phi(\mathbf{y}) = \frac{\partial \phi(\mathbf{x})}{\partial x^r} \frac{\partial x^r}{\partial y^k} \mathbf{e}^k(\mathbf{y}) =$$

$$\frac{\partial \phi(\mathbf{x})}{\partial x^r} \frac{\partial x^r}{\partial y^k} \frac{\partial y^k}{\partial x^s} \mathbf{e}^s(\mathbf{x}) = \frac{\partial \phi(\mathbf{x})}{\partial x^r} \delta_s^r \mathbf{e}^s(\mathbf{x}) = \frac{\partial \phi(\mathbf{x})}{\partial x^r} \mathbf{e}^r(\mathbf{x}).$$

This shows the covariant components of $\nabla \phi(\mathbf{x})$ are

$$(\nabla \phi(\mathbf{x}))_r = \frac{\partial \phi(\mathbf{x})}{\partial x^r}, \quad (25.35)$$

Formally the same as in rectangular coordinates. To find the contravariant components, “raise the index” in the usual way. Thus

$$(\nabla \phi(\mathbf{x}))^r = g^{rk}(\mathbf{x}) (\nabla \phi(\mathbf{x}))_k = g^{rk}(\mathbf{x}) \frac{\partial \phi(\mathbf{x})}{\partial x^k}. \quad (25.36)$$

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 25.6.1,

$$F^i_{,j}(\mathbf{y}) = F^r_{,s}(\mathbf{x}) \frac{\partial x^s}{\partial y^j} \frac{\partial y^i}{\partial x^r}$$

From 25.27,

$$= \left(\frac{\partial F^r(\mathbf{x})}{\partial x^s} + F^k(\mathbf{x}) \left\{ \begin{matrix} r \\ ks \end{matrix} \right\}(\mathbf{x}) \right) \frac{\partial x^s}{\partial y^j} \frac{\partial y^i}{\partial x^r}.$$

Letting $j = i$ yields

$$\begin{aligned}
 \operatorname{div}(\mathbf{F}) &= \left(\frac{\partial F^r}{\partial x^s} + F^k(x) \left\{ \begin{matrix} r \\ ks \end{matrix} \right\} (x) \right) \frac{\partial x^s}{\partial y^i} \frac{\partial y^i}{\partial x^r} \\
 &= \left(\frac{\partial F^r}{\partial x^s} + F^k(x) \left\{ \begin{matrix} r \\ ks \end{matrix} \right\} (x) \right) \delta_r^s \\
 &= \left(\frac{\partial F^r}{\partial x^r} + F^k(x) \left\{ \begin{matrix} r \\ kr \end{matrix} \right\} (x) \right). \tag{25.37}
 \end{aligned}$$

$\left\{ \begin{matrix} r \\ kr \end{matrix} \right\}$ is simplified using the description of it in Theorem 25.6.2. Thus, from this theorem,

$$\left\{ \begin{matrix} r \\ kr \end{matrix} \right\} = \frac{g^{jr}}{2} \left[\frac{\partial g_{rj}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^r} - \frac{\partial g_{rk}}{\partial x^j} \right]$$

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,

$$\left\{ \begin{matrix} r \\ kr \end{matrix} \right\} = \frac{g^{jr}}{2} \frac{\partial g_{rj}}{\partial x^k}. \tag{25.38}$$

Now recall $g \equiv \det(g_{ij}) = \det(G) > 0$ from Theorem 25.1.6. 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 25.38,

$$\left\{ \begin{matrix} r \\ kr \end{matrix} \right\} = \frac{g^{jr}}{2} \frac{\partial g_{rj}}{\partial x^k} = \frac{1}{2g} \frac{\partial g_{rj}}{\partial x^k} A^{jr} = \frac{1}{2g} \frac{\partial g}{\partial g_{rj}} \frac{\partial g_{rj}}{\partial x^k} = \frac{1}{2g} \frac{\partial g}{\partial x^k}$$

and so from 25.37,

$$\begin{aligned}
 \operatorname{div}(\mathbf{F}) &= \frac{\partial F^k(x)}{\partial x^k} + \\
 &+ F^k(x) \frac{1}{2g(x)} \frac{\partial g(x)}{\partial x^k} = \frac{1}{\sqrt{g(x)}} \frac{\partial}{\partial x^i} \left(F^i(x) \sqrt{g(x)} \right). \tag{25.39}
 \end{aligned}$$

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 25.39 and 25.36 it follows

$$\Delta\phi(\mathbf{x}) = \frac{1}{\sqrt{g(\mathbf{x})}} \frac{\partial}{\partial x^i} \left(g^{ik}(\mathbf{x}) \frac{\partial\phi(\mathbf{x})}{\partial x^k} \sqrt{g(\mathbf{x})} \right). \quad (25.40)$$

We summarize the conclusions of this section in the following theorem.

Theorem 25.7.1 *The following formulas hold for the gradient, divergence and Laplacian in general curvilinear coordinates.*

$$(\nabla\phi(\mathbf{x}))_r = \frac{\partial\phi(\mathbf{x})}{\partial x^r}, \quad (25.41)$$

$$(\nabla\phi(\mathbf{x}))^r = g^{rk}(\mathbf{x}) \frac{\partial\phi(\mathbf{x})}{\partial x^k}, \quad (25.42)$$

$$\operatorname{div}(\mathbf{F}) = \frac{1}{\sqrt{g(\mathbf{x})}} \frac{\partial}{\partial x^i} \left(F^i(\mathbf{x}) \sqrt{g(\mathbf{x})} \right), \quad (25.43)$$

$$\Delta\phi(\mathbf{x}) = \frac{1}{\sqrt{g(\mathbf{x})}} \frac{\partial}{\partial x^i} \left(g^{ik}(\mathbf{x}) \frac{\partial\phi(\mathbf{x})}{\partial x^k} \sqrt{g(\mathbf{x})} \right). \quad (25.44)$$

Example 25.7.2 *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.

25.8 Exercises

1. Let $y^1 = x^1 + 2x^2, y^2 = x^2 + 3x^3, y^3 = x^1 + x^3$. Let

$$\mathbf{F}(\mathbf{x}) = x^1 \mathbf{e}_1(\mathbf{x}) + x^2 \mathbf{e}_2(\mathbf{x}) + (x^3)^2 \mathbf{e}_3(\mathbf{x}).$$

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

$$\left\{ l_{ij} \right\} = \frac{\partial \mathbf{e}_i}{\partial x^j} \cdot \mathbf{e}^j.$$

Find the Christoffel symbols of the second kind for spherical coordinates in which $x^1 = \phi$, $x^2 = \theta$, and $x^3 = \rho$. Do the same for cylindrical coordinates letting $x^1 = r$, $x^2 = \theta$, $x^3 = z$.

10. Show velocity can be expressed as $\mathbf{v} = v_i(\mathbf{x}) \mathbf{e}^i(\mathbf{x})$, where

$$v_i(\mathbf{x}) = \frac{\partial r_i}{\partial x^j} \frac{dx^j}{dt} - r_p(\mathbf{x}) \left\{ p \right\}_{ik} \frac{dx^k}{dt}$$

and $r_i(\mathbf{x})$ are the covariant components of the displacement vector,

$$\mathbf{r} = r_i(\mathbf{x}) \mathbf{e}^i(\mathbf{x}).$$

11. \uparrow Using problem 9 and 10, show the covariant components of velocity in spherical coordinates are

$$v_1 = \rho^2 \frac{d\phi}{dt}, v_2 = \rho^2 \sin^2(\phi) \frac{d\theta}{dt}, v_3 = \frac{d\rho}{dt}.$$

Hint: First observe that if \mathbf{r} is the position vector from the origin, then $\mathbf{r} = \rho \mathbf{e}_3$ so $r_1 = 0 = r_2$, and $r_3 = \rho$. Now use 10.

25.9 Curl and Cross Products

In this section is the curl and cross product in general curvilinear coordinates in \mathbb{R}^3 . We will always assume that for \mathbf{x} a set of curvilinear coordinates,

$$\det \left(\frac{\partial y^i}{\partial x^j} \right) > 0 \quad (25.45)$$

Where the y_i are the usual coordinates in which $\mathbf{e}_k(\mathbf{y}) = \mathbf{i}_k$.

Theorem 25.9.1 *Let 25.45 hold. Then*

$$\det \left(\frac{\partial y^i}{\partial x^j} \right) = \sqrt{g(\mathbf{x})} \quad (25.46)$$

and

$$\det \left(\frac{\partial x^i}{\partial y^j} \right) = \frac{1}{\sqrt{g(\mathbf{x})}}. \quad (25.47)$$

Proof:

$$\mathbf{e}_i(\mathbf{x}) = \frac{\partial y^k}{\partial x^i} \mathbf{i}_k$$

and so

$$g_{ij}(\mathbf{x}) = \frac{\partial y^k}{\partial x^i} \mathbf{i}_k \cdot \frac{\partial y^l}{\partial x^j} \mathbf{i}_l = \frac{\partial y^k}{\partial x^i} \frac{\partial y^k}{\partial x^j}.$$

Therefore, $g = \det(g_{ij}(\mathbf{x})) = \left(\det \left(\frac{\partial y^k}{\partial x^i} \right) \right)^2$. By 25.45, $\sqrt{g} = \det \left(\frac{\partial y^k}{\partial x^i} \right)$ as claimed. Now

$$\frac{\partial y^k}{\partial x^i} \frac{\partial x^i}{\partial y^r} = \delta_r^k$$

and so

$$\det \left(\frac{\partial x^i}{\partial y^r} \right) = \frac{1}{\sqrt{g(\mathbf{x})}}.$$

This proves the theorem.

To get the curl and cross product in curvilinear coordinates, let ϵ^{ijk} be the usual permutation symbol. Thus,

$$\epsilon^{123} = 1$$

and when any two indices in ϵ^{ijk} are switched, the sign changes. Thus

$$\epsilon^{132} = -1, \epsilon^{312} = 1, \text{ etc.}$$

Now define

$$\epsilon^{ijk}(\mathbf{x}) \equiv \epsilon^{ijk} \frac{1}{\sqrt{g(\mathbf{x})}}.$$

Then for \mathbf{x} and \mathbf{z} satisfying 25.45,

$$\epsilon^{ijk}(\mathbf{x}) \frac{\partial z^r}{\partial x^i} \frac{\partial z^s}{\partial x^j} \frac{\partial z^t}{\partial x^k} = \epsilon^{ijk} \det \left(\frac{\partial x^p}{\partial y^q} \right) \frac{\partial z^r}{\partial x^i} \frac{\partial z^s}{\partial x^j} \frac{\partial z^t}{\partial x^k}$$

$$= \varepsilon^{rst} \det \left(\frac{\partial x^p}{\partial y^q} \right) \det \left(\frac{\partial z^i}{\partial x^k} \right) = \varepsilon^{rst} \det(MN)$$

where N is the matrix whose pq^{th} entry is $\frac{\partial x^p}{\partial y^q}$ and M is the matrix whose ik^{th} entry is $\frac{\partial z^i}{\partial x^k}$. Therefore, from the definition of matrix multiplication and the chain rule, this equals

$$= \varepsilon^{rst} \det \left(\frac{\partial z^i}{\partial y^p} \right) \equiv \varepsilon^{rst} (z)$$

from the above discussion.

Now $\varepsilon^{ijk}(\mathbf{y}) = \varepsilon^{ijk}$ and for a vector field, \mathbf{F} ,

$$\text{curl}(\mathbf{F}) \equiv \varepsilon^{ijk}(\mathbf{y}) F_{k,j}(\mathbf{y}) \mathbf{e}_i(\mathbf{y}).$$

Therefore, since we know how everything transforms assuming 25.45, it is routine to write this in terms of \mathbf{x} .

$$\begin{aligned} \text{curl}(\mathbf{F}) &= \varepsilon^{rst}(\mathbf{x}) \frac{\partial y^i}{\partial x^r} \frac{\partial y^j}{\partial x^s} \frac{\partial y^k}{\partial x^t} F_{p,q}(\mathbf{x}) \frac{\partial x^p}{\partial y^k} \frac{\partial x^q}{\partial y^j} \mathbf{e}_m(\mathbf{x}) \frac{\partial x^m}{\partial y^i} \\ &= \varepsilon^{rst}(\mathbf{x}) \delta_r^m \delta_s^q \delta_t^p F_{p,q}(\mathbf{x}) \mathbf{e}_m(\mathbf{x}) = \varepsilon^{mqp}(\mathbf{x}) F_{p,q}(\mathbf{x}) \mathbf{e}_m(\mathbf{x}). \end{aligned} \quad (25.48)$$

More simplification is possible. Recalling the definition of $F_{p,q}(\mathbf{x})$,

$$\begin{aligned} \frac{\partial \mathbf{F}}{\partial x^q} &\equiv F_{p,q}(\mathbf{x}) \mathbf{e}^p(\mathbf{x}) = \frac{\partial}{\partial x^q} [F_p(\mathbf{x}) \mathbf{e}^p(\mathbf{x})] \\ &= \frac{\partial F_p(\mathbf{x})}{\partial x^q} \mathbf{e}^p(\mathbf{x}) + F_p(\mathbf{x}) \frac{\partial \mathbf{e}^p}{\partial x^q} = \frac{\partial F_p(\mathbf{x})}{\partial x^q} \mathbf{e}^p(\mathbf{x}) - F_r(\mathbf{x}) \left\{ \begin{matrix} r \\ pq \end{matrix} \right\} \mathbf{e}^p(\mathbf{x}) \end{aligned}$$

by Theorem 25.6.2. Therefore,

$$F_{p,q}(\mathbf{x}) = \frac{\partial F_p(\mathbf{x})}{\partial x^q} - F_r(\mathbf{x}) \left\{ \begin{matrix} r \\ pq \end{matrix} \right\}$$

and so

$$\text{curl}(\mathbf{F}) = \varepsilon^{mqp}(\mathbf{x}) \frac{\partial F_p(\mathbf{x})}{\partial x^q} \mathbf{e}_m(\mathbf{x}) - \varepsilon^{mqp}(\mathbf{x}) F_r(\mathbf{x}) \left\{ \begin{matrix} r \\ pq \end{matrix} \right\} \mathbf{e}_m(\mathbf{x}).$$

However, because $\left\{ \begin{matrix} r \\ pq \end{matrix} \right\} = \left\{ \begin{matrix} r \\ qp \end{matrix} \right\}$, the second term in this expression equals 0. To see this,

$$\varepsilon^{mqp}(\mathbf{x}) \left\{ \begin{matrix} r \\ pq \end{matrix} \right\} = \varepsilon^{mpq}(\mathbf{x}) \left\{ \begin{matrix} r \\ qp \end{matrix} \right\} = -\varepsilon^{mqp}(\mathbf{x}) \left\{ \begin{matrix} r \\ pq \end{matrix} \right\}.$$

Therefore, by 25.48,

$$\text{curl}(\mathbf{F}) = \varepsilon^{mqp}(\mathbf{x}) \frac{\partial F_p(\mathbf{x})}{\partial x^q} \mathbf{e}_m(\mathbf{x}). \quad (25.49)$$

What about the cross product of two vector fields? Let \mathbf{F} and \mathbf{G} be two vector fields. Then in terms of standard coordinates \mathbf{y} ,

$$\mathbf{F} \times \mathbf{G} = \varepsilon^{ijk}(\mathbf{y}) F_j(\mathbf{y}) G_k(\mathbf{y}) \mathbf{e}_i(\mathbf{y})$$

$$\begin{aligned}
&= \varepsilon^{rst}(\mathbf{x}) \frac{\partial y^i}{\partial x^r} \frac{\partial y^j}{\partial x^s} \frac{\partial y^k}{\partial x^t} F_p(\mathbf{x}) \frac{\partial x^p}{\partial y^j} G_q(\mathbf{x}) \frac{\partial x^q}{\partial y^k} e_l(\mathbf{x}) \frac{\partial x^l}{\partial y^i} \\
&= \varepsilon^{rst}(\mathbf{x}) \delta_s^p \delta_t^q \delta_r^l F_p(\mathbf{x}) G_q(\mathbf{x}) e_l(\mathbf{x}) = \varepsilon^{lpq}(\mathbf{x}) F_p(\mathbf{x}) G_q(\mathbf{x}) e_l(\mathbf{x}). \quad (25.50)
\end{aligned}$$

We summarize these results in the following theorem.

Theorem 25.9.2 Suppose \mathbf{x} is a system of curvilinear coordinates in \mathbb{R}^3 such that

$$\det \left(\frac{\partial y^i}{\partial x^j} \right) > 0.$$

Let

$$\varepsilon^{ijk}(\mathbf{x}) \equiv \varepsilon^{ijk} \frac{1}{\sqrt{g(\mathbf{x})}}.$$

Then the following formulas for curl and cross product hold in this system of coordinates.

$$\operatorname{curl}(\mathbf{F}) = \varepsilon^{mqp}(\mathbf{x}) \frac{\partial F_p(\mathbf{x})}{\partial x^q} \mathbf{e}_m(\mathbf{x}),$$

and

$$\mathbf{F} \times \mathbf{G} = \varepsilon^{lpq}(\mathbf{x}) F_p(\mathbf{x}) G_q(\mathbf{x}) \mathbf{e}_l(\mathbf{x}).$$

Chapter 26

Implicit Function Theorem*

The implicit function theorem is one of the greatest theorems in mathematics. There are many versions of this theorem which are of far greater generality than the one given here. The proof given here is like one found in one of Caratheodory's books on the calculus of variations. It 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 material on rank and determinants found in the chapter on the mathematical theory of determinants. You will also need to use the extreme value theorem for a function of n variables and the chain rule of multivariable calculus as well as everything about matrix multiplication.

Definition 26.0.1 Suppose U is an open set in $\mathbb{R}^n \times \mathbb{R}^m$ and (\mathbf{x}, \mathbf{y}) will denote a typical point of $\mathbb{R}^n \times \mathbb{R}^m$ with $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Let $\mathbf{f} : U \rightarrow \mathbb{R}^p$ be in $C^1(U)$. Then define

$$\begin{aligned} D_1 \mathbf{f}(\mathbf{x}, \mathbf{y}) &\equiv \begin{pmatrix} f_{1,x_1}(\mathbf{x}, \mathbf{y}) & \cdots & f_{1,x_n}(\mathbf{x}, \mathbf{y}) \\ \vdots & & \vdots \\ f_{p,x_1}(\mathbf{x}, \mathbf{y}) & \cdots & f_{p,x_n}(\mathbf{x}, \mathbf{y}) \end{pmatrix}, \\ D_2 \mathbf{f}(\mathbf{x}, \mathbf{y}) &\equiv \begin{pmatrix} f_{1,y_1}(\mathbf{x}, \mathbf{y}) & \cdots & f_{1,y_m}(\mathbf{x}, \mathbf{y}) \\ \vdots & & \vdots \\ f_{p,y_1}(\mathbf{x}, \mathbf{y}) & \cdots & f_{p,y_m}(\mathbf{x}, \mathbf{y}) \end{pmatrix}. \end{aligned}$$

Theorem 26.0.2 (implicit function theorem) Suppose U is an open set in $\mathbb{R}^n \times \mathbb{R}^m$. Let $\mathbf{f} : U \rightarrow \mathbb{R}^p$ be in $C^1(U)$ and suppose

$$\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}, D_1 \mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)^{-1} \text{ exists.} \quad (26.1)$$

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

$$\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y}) = \mathbf{0}. \quad (26.2)$$

Furthermore, the mapping, $\mathbf{y} \rightarrow \mathbf{x}(\mathbf{y})$ is in $C^1(B(\mathbf{y}_0, \eta))$.

Proof: Let

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} f_1(\mathbf{x}, \mathbf{y}) \\ f_2(\mathbf{x}, \mathbf{y}) \\ \vdots \\ f_n(\mathbf{x}, \mathbf{y}) \end{pmatrix}.$$

Define for $(\mathbf{x}^1, \dots, \mathbf{x}^n) \in \overline{B(\mathbf{x}_0, \delta)}^n$ and $\mathbf{y} \in B(\mathbf{y}_0, \eta)$ the following matrix.

$$J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y}) \equiv \begin{pmatrix} f_{1,x_1}(\mathbf{x}^1, \mathbf{y}) & \cdots & f_{1,x_n}(\mathbf{x}^1, \mathbf{y}) \\ \vdots & & \vdots \\ f_{n,x_1}(\mathbf{x}^n, \mathbf{y}) & \cdots & f_{n,x_n}(\mathbf{x}^n, \mathbf{y}) \end{pmatrix}. \quad (*)$$

Then by the assumption of continuity of all the partial derivatives and the extreme value theorem, 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 $(\mathbf{x}^1, \dots, \mathbf{x}^n) \in \overline{B(\mathbf{x}_0, \delta)}^n$ and $\mathbf{y} \in \overline{B(\mathbf{y}_0, \eta)}$,

$$|\det(J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y}))| > r > 0. \quad (26.3)$$

and $\overline{B(\mathbf{x}_0, \delta_0)} \times \overline{B(\mathbf{y}_0, \eta_0)} \subseteq U$. 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 $(\mathbf{x}, \mathbf{y}) \in \overline{B(\mathbf{x}_0, \delta_0)} \times \overline{B(\mathbf{y}_0, \eta_0)}$ and $i = 1, 2, \dots, n$, the i^{th} row of $D_2 \mathbf{f}(\mathbf{x}, \mathbf{y})$, given by $D_2 f_i(\mathbf{x}, \mathbf{y})$ satisfies

$$|D_2 f_i(\mathbf{x}, \mathbf{y})| < K, \quad (26.4)$$

and for all $(\mathbf{x}^1, \dots, \mathbf{x}^n) \in \overline{B(\mathbf{x}_0, \delta_0)}^n$ and $\mathbf{y} \in \overline{B(\mathbf{y}_0, \eta_0)}$ the i^{th} row of the matrix,

$$J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y})^{-1}$$

which equals $\mathbf{e}_i^T (J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y})^{-1})$ satisfies

$$|\mathbf{e}_i^T (J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y})^{-1})| < K. \quad (26.5)$$

(Recall that \mathbf{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 $\mathbf{y} \in B(\mathbf{y}_0, \eta)$ there is at most one $\mathbf{x} \in B(\mathbf{x}_0, \delta)$ 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 ,

$$h'(t_i) = \sum_{j=1}^n \frac{\partial}{\partial x_j} f_i(\mathbf{x} + t_i(\mathbf{z} - \mathbf{x}), \mathbf{y}) (z_j - x_j) = 0. \quad (26.6)$$

Then denote by \mathbf{x}^i the vector, $\mathbf{x} + t_i(\mathbf{z} - \mathbf{x})$. It follows from 26.6 that

$$J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y})(\mathbf{z} - \mathbf{x}) = \mathbf{0}$$

and so from 26.3 $z - x = 0$. (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 $y \in B(y_0, \eta)$, there exists a unique $x(y) \in B(x_0, \delta)$ such that $f(x(y), y) = 0$.

Claim: If η is small enough, then the function, $x \rightarrow h_y(x) \equiv |f(x, y)|^2$ achieves its minimum value on $\overline{B(x_0, \delta)}$ at a point of $B(x_0, \delta)$. (The existence of a point in $\overline{B(x_0, \delta)}$ at which h_y achieves its minimum follows from the extreme value theorem.)

Proof of claim: Suppose this is not the case. Then there exists a sequence $\eta_k \rightarrow 0$ and for some y_k having $|y_k - y_0| < \eta_k$, the minimum of h_{y_k} on $\overline{B(x_0, \delta)}$ occurs on a point x_k such that $|x_0 - x_k| = \delta$. Now taking a subsequence, still denoted by k , it can be assumed that $x_k \rightarrow x$ with $|x - x_0| = \delta$ and $y_k \rightarrow y_0$. This follows from the fact that $\{x \in \overline{B(x_0, \delta)} : |x - x_0| = \delta\}$ is a closed and bounded set and is therefore sequentially compact. Let $\varepsilon > 0$. Then for k large enough, the continuity of $y \rightarrow h_y(x_0)$ implies $h_{y_k}(x_0) < \varepsilon$ because $h_{y_0}(x_0) = 0$ since $f(x_0, y_0) = 0$. Therefore, from the definition of x_k , it is also the case that $h_{y_k}(x_k) < \varepsilon$. Passing to the limit yields $h_{y_0}(x) \leq \varepsilon$. Since $\varepsilon > 0$ is arbitrary, it follows that $h_{y_0}(x) = 0$ which contradicts the first part of the argument in which it was shown that for $y \in B(y_0, \eta)$ there is at most one point, x of $\overline{B(x_0, \delta)}$ where $f(x, y) = 0$. Here two have been obtained, x_0 and x . This proves the claim.

Choose $\eta < \eta_0$ and also small enough that the above claim holds and let $x(y)$ denote a point of $B(x_0, \delta)$ at which the minimum of h_y on $\overline{B(x_0, \delta)}$ is achieved. Since $x(y)$ is an interior point, you can consider $h_y(x(y) + tv)$ for $|t|$ small and conclude this function of t has a zero derivative at $t = 0$. Now

$$h_y(x(y) + tv) = \sum_{i=1}^n f_i^2(x(y) + tv, y)$$

and so from the chain rule,

$$\frac{d}{dt} h_y(x(y) + tv) = \sum_{i=1}^n \sum_{j=1}^n 2f_i(x(y) + tv, y) \frac{\partial f_i(x(y) + tv, y)}{\partial x_j} v_j.$$

Therefore, letting $t = 0$, it is required that for every v ,

$$\sum_{i=1}^n \sum_{j=1}^n 2f_i(x(y), y) \frac{\partial f_i(x(y), y)}{\partial x_j} v_j = 0.$$

In terms of matrices this reduces to

$$0 = 2f(x(y), y)^T D_1 f(x(y), y) v$$

for every vector v . Therefore,

$$0 = f(x(y), y)^T D_1 f(x(y), y)$$

From 26.3, it follows $f(x(y), y) = 0$. This proves the existence of the function $y \rightarrow x(y)$ such that $f(x(y), y) = 0$ for all $y \in B(y_0, \eta)$.

It remains to verify this function is a C^1 function. To do this, let y_1 and y_2 be points of $B(y_0, \eta)$. Then as before, consider the i^{th} component of f and consider the same argument using the mean value theorem to write

$$\begin{aligned} 0 &= f_i(x(y_1), y_1) - f_i(x(y_2), y_2) \\ &= f_i(x(y_1), y_1) - f_i(x(y_2), y_1) + f_i(x(y_2), y_1) - f_i(x(y_2), y_2) \\ &= D_1 f_i(x^i, y_1)(x(y_1) - x(y_2)) + D_2 f_i(x(y_2), y^i)(y_1 - y_2). \end{aligned} \quad (26.7)$$

where \mathbf{y}^i is a point on the line segment joining \mathbf{y}_1 and \mathbf{y}_2 . Thus from 26.4 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, defining the symbol $M(\mathbf{y}^1, \dots, \mathbf{y}^n) \equiv M$ denote the matrix having the i^{th} row equal to

$$D_2 f_i(\mathbf{x}(\mathbf{y}_2), \mathbf{y}^i),$$

it follows

$$|M(\mathbf{y}_1 - \mathbf{y}_2)| \leq \left(\sum_i K^2 |\mathbf{y}_1 - \mathbf{y}_2|^2 \right)^{1/2} = \sqrt{m}K |\mathbf{y}_1 - \mathbf{y}_2|. \quad (26.8)$$

Also, from 26.7,

$$J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y}_1)(\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)) = -M(\mathbf{y}_1 - \mathbf{y}_2) \quad (26.9)$$

and so from 26.8, 26.5, $|\mathbf{x}(\mathbf{y}_1) - \mathbf{x}(\mathbf{y}_2)| =$

$$\begin{aligned} &= |J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y}_1)^{-1} M(\mathbf{y}_1 - \mathbf{y}_2)| \\ &= \left(\sum_{i=1}^n |e_i^T J(\mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{y}_1)^{-1} M(\mathbf{y}_1 - \mathbf{y}_2)|^2 \right)^{1/2} \\ &\leq \left(\sum_{i=1}^n K^2 |M(\mathbf{y}_1 - \mathbf{y}_2)|^2 \right)^{1/2} \leq \left(\sum_{i=1}^n K^2 (\sqrt{m}K |\mathbf{y}_1 - \mathbf{y}_2|)^2 \right)^{1/2} \\ &= K^2 \sqrt{mn} |\mathbf{y}_1 - \mathbf{y}_2| \end{aligned}$$

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 26.0.3 (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

$$\mathbf{f} \text{ is } C^1(U), \text{ and } D\mathbf{f}(\mathbf{x}_0)^{-1} \text{ exists.} \quad (26.10)$$

Then there exist open sets W , and V such that

$$\mathbf{x}_0 \in W \subseteq U, \quad (26.11)$$

$$\mathbf{f} : W \rightarrow V \text{ is one to one and onto,} \quad (26.12)$$

$$\mathbf{f}^{-1} \text{ is } C^1, \quad (26.13)$$

Proof: Apply the implicit function theorem to the function

$$F(x, y) \equiv f(x) - y$$

where $y_0 \equiv f(x_0)$. Thus the function $y \rightarrow x(y)$ defined in that theorem is f^{-1} . Now let

$$W \equiv B(x_0, \delta) \cap f^{-1}(B(y_0, \eta))$$

and

$$V \equiv B(y_0, \eta).$$

This proves the theorem. ■

26.1 More Continuous Partial Derivatives

The implicit function theorem will now be improved slightly. If 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 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 $|\alpha| = \sum_i \alpha_i$,

$$D^\alpha f(x) \equiv \frac{\partial^{|\alpha|} f}{\partial \alpha_1 \partial \alpha_2 \dots \partial \alpha_n}(x), \quad D^0 f(x) \equiv f(x)$$

Theorem 26.1.1 (*implicit function theorem*) Suppose U is an open set in $\mathbb{F}^n \times \mathbb{F}^m$. Let $f : U \rightarrow \mathbb{F}^m$ be in $C^k(U)$ and suppose

$$f(x_0, y_0) = 0, \quad D_1 f(x_0, y_0)^{-1} \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m). \quad (26.14)$$

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

$$f(x(y), y) = 0. \quad (26.15)$$

Furthermore, the mapping $y \rightarrow x(y)$ is in $C^k(B(y_0, \eta))$.

Proof: From the implicit function theorem $y \rightarrow x(y)$ is C^1 . It remains to show that it is C^k for $k > 1$ assuming that f is C^k . From (26.15)

$$\frac{\partial x}{\partial y^i} = -D_1 f(x, y)^{-1} \frac{\partial f}{\partial y^i}.$$

Thus the following formula holds for $q = 1$ and $|\alpha| = q$.

$$D^\alpha x(y) = \sum_{|\beta| \leq q} M_\beta(x, y) D^\beta f(x, y) \quad (26.16)$$

where M_β is a matrix whose entries are differentiable functions of $D^\gamma x$ for $|\gamma| < q$ and $D^\tau f(x, y)$ for $|\tau| \leq q$. This follows easily from the description of $D_1 f(x, y)^{-1}$ in terms

of the cofactor matrix and the determinant of $D_1 \mathbf{f}(\mathbf{x}, \mathbf{y})$. Suppose (26.16) 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 such that its 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 (26.16) 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 26.1.2 (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,

$$\mathbf{f} \text{ is } C^k(U), \text{ and } D\mathbf{f}(\mathbf{x}_0)^{-1} \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n). \quad (26.17)$$

Then there exist open sets W , and V such that

$$\mathbf{x}_0 \in W \subseteq U, \quad (26.18)$$

$$\mathbf{f} : W \rightarrow V \text{ is one to one and onto,} \quad (26.19)$$

$$\mathbf{f}^{-1} \text{ is } C^k. \quad (26.20)$$

26.2 The Method Of Lagrange Multipliers

1. 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

$$g_i(\mathbf{x}) = 0, \quad i = 1, \dots, m \quad (26.21)$$

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 (26.21). A local minimum is defined similarly. Let $\mathbf{F} : U \times \mathbb{R} \rightarrow \mathbb{R}^{m+1}$ be defined by

$$\mathbf{F}(\mathbf{x}, a) \equiv \begin{pmatrix} f(\mathbf{x}) - a \\ g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix}. \quad (26.22)$$

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. It follows from the implicit function theorem, there exists $m+1$ variables $x_{i_1}, \dots, x_{i_{m+1}}$ such that the system

$$\mathbf{F}(\mathbf{x}, a) = \mathbf{0} \quad (26.23)$$

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 26.23 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

$$\mu \begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix}. \quad (26.24)$$

If the rank of the matrix

$$\begin{pmatrix} g_{1x_1}(\mathbf{x}_0) & \cdots & g_{mx_1}(\mathbf{x}_0) \\ \vdots & & \vdots \\ g_{1x_n}(\mathbf{x}_0) & \cdots & g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \quad (26.25)$$

is m , then we can choose $\mu = 1$ because the columns span \mathbb{R}^m . Thus there are scalars λ_i such that

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \quad (26.26)$$

at every point \mathbf{x}_0 which is either a local maximum or a local minimum. This proves the following theorem.

Theorem 26.2.1 *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 26.21, then 26.24 must hold for some scalars $\mu, \lambda_1, \dots, \lambda_m$ not all equal to zero. If the rank of the matrix in 26.25 is m , it follows 26.26 holds for some choice of the λ_i .*

26.3 The Local Structure Of C^1 Mappings*

In linear algebra it is shown that every invertible matrix can be written as a product of elementary matrices, those matrices which are obtained from doing a row operation to the identity matrix. Two of the row operations produce a matrix which will change exactly one entry of a vector when it is multiplied by the elementary matrix. The other row operation involves switching two rows and this has the effect of switching two entries in a vector when multiplied on the left by the elementary matrix. Thus, in terms of the effect on a vector, the mapping determined by the given matrix can be considered as a composition of mappings which either flip two entries of the vector or change exactly one. A similar local result is available for nonlinear mappings. I found this interesting result in the advanced calculus book by Rudin.

Definition 26.3.1 Let U be an open set in \mathbb{R}^n and let $G : U \rightarrow \mathbb{R}^n$. Then G is called primitive if it is of the form

$$G(x) = \begin{pmatrix} x_1 & \cdots & \alpha(x) & \cdots & x_n \end{pmatrix}^T.$$

Thus, G is primitive if it only changes one of the variables. A function $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is called a flip if

$$F(x_1, \dots, x_k, \dots, x_l, \dots, x_n) = (x_1, \dots, x_l, \dots, x_k, \dots, x_n)^T.$$

Thus a function is a flip if it interchanges two coordinates. Also, for $m = 1, 2, \dots, n$, define

$$P_m(x) \equiv \begin{pmatrix} x_1 & x_2 & \cdots & x_m & 0 & \cdots & 0 \end{pmatrix}^T$$

It turns out that if $h(0) = 0$, $Dh(0)^{-1}$ exists, and h is C^1 on U , then h can be written as a composition of primitive functions and flips. This is a very interesting application of the inverse function theorem.

Theorem 26.3.2 Let $h : U \rightarrow \mathbb{R}^n$ be a C^1 function with $h(0) = 0$ $Dh(0)^{-1}$ exists. Then there is an open set $V \subseteq U$ containing 0 , flips F_1, \dots, F_{n-1} , and primitive functions G_n, G_{n-1}, \dots, G_1 such that for $x \in V$,

$$h(x) = F_1 \circ \cdots \circ F_{n-1} \circ G_n \circ G_{n-1} \circ \cdots \circ G_1(x).$$

The primitive function G_j leaves x_i unchanged for $i \neq j$.

Proof: Let

$$h_1(x) \equiv h(x) = \begin{pmatrix} \alpha_1(x) & \cdots & \alpha_n(x) \end{pmatrix}^T$$

$$Dh(0)e_1 = \begin{pmatrix} \alpha_{1,1}(0) & \cdots & \alpha_{n,1}(0) \end{pmatrix}^T$$

where $\alpha_{k,1}$ denotes $\frac{\partial \alpha_k}{\partial x_1}$. Since $Dh(0)$ is one to one, the right side of this expression cannot be zero. Hence there exists some k such that $\alpha_{k,1}(0) \neq 0$. Now define

$$G_1(x) \equiv \begin{pmatrix} \alpha_k(x) & x_2 & \cdots & x_n \end{pmatrix}^T$$

Then the matrix of $DG_1(\mathbf{0})$ is of the form

$$\begin{pmatrix} \alpha_{k,1}(\mathbf{0}) & \cdots & \cdots & \alpha_{k,n}(\mathbf{0}) \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

and its determinant equals $\alpha_{k,1}(\mathbf{0}) \neq 0$. Therefore, by the inverse function theorem, there exists an open set U_1 , containing $\mathbf{0}$ and an open set V_2 containing $\mathbf{0}$ such that $G_1(U_1) = V_2$ and G_1 is one to one and onto, such that it and its inverse are both C^1 . Let F_1 denote the flip which interchanges x_k with x_1 . Now define

$$h_2(y) \equiv F_1 \circ h_1 \circ G_1^{-1}(y)$$

Thus

$$\begin{aligned} h_2(G_1(x)) &\equiv F_1 \circ h_1(x) \\ &= \left(\alpha_k(x) \quad \cdots \quad \alpha_1(x) \quad \cdots \quad \alpha_n(x) \right)^T \end{aligned} \quad (26.27)$$

Therefore,

$$P_1 h_2(G_1(x)) = \left(\alpha_k(x) \quad 0 \quad \cdots \quad 0 \right)^T.$$

Also

$$P_1(G_1(x)) = \left(\alpha_k(x) \quad 0 \quad \cdots \quad 0 \right)^T$$

so $P_1 h_2(y) = P_1(y)$ for all $y \in V_2$. Also, $h_2(\mathbf{0}) = \mathbf{0}$ and $Dh_2(\mathbf{0})^{-1}$ exists because of the definition of h_2 above and the chain rule. Since $F_1^2 = I$, the identity map, it follows from (26.27) that

$$h(x) = h_1(x) = F_1 \circ h_2 \circ G_1(x). \quad (26.28)$$

Note that on an open set $V_2 \equiv G_1(U_1)$ containing the origin, h_2 leaves the first entry unchanged. This is what $P_1 h_2(G_1(x)) = P_1(G_1(x))$ says. In contrast, $h_1 = h$ left possibly no entries unchanged.

Suppose then, that for $m \geq 2$, h_m leaves the first $m-1$ entries unchanged,

$$P_{m-1} h_m(x) = P_{m-1}(x) \quad (26.29)$$

for all $x \in U_m$, an open subset of U containing $\mathbf{0}$, and $h_m(\mathbf{0}) = \mathbf{0}$, $Dh_m(\mathbf{0})^{-1}$ exists. From (26.29), $h_m(x)$ must be of the form

$$h_m(x) = \left(x_1 \quad \cdots \quad x_{m-1} \quad \alpha_1(x) \quad \cdots \quad \alpha_n(x) \right)^T$$

where these α_k are different than the ones used earlier. Then

$$Dh_m(\mathbf{0})e_m = \left(0 \quad \cdots \quad 0 \quad \alpha_{1,m}(\mathbf{0}) \quad \cdots \quad \alpha_{n,m}(\mathbf{0}) \right)^T \neq \mathbf{0}$$

because $Dh_m(\mathbf{0})^{-1}$ exists. Therefore, there exists a $k \geq m$ such that $\alpha_{k,m}(\mathbf{0}) \neq 0$, not the same k as before. Define

$$G_m(x) \equiv \left(x_1 \quad \cdots \quad x_{m-1} \quad \alpha_k(x) \quad \cdots \quad x_n \right)^T \quad (26.30)$$

so a change in G_m occurs only in the m^{th} slot. Then $G_m(\mathbf{0}) = \mathbf{0}$ and $DG_m(\mathbf{0})^{-1}$ exists similar to the above. In fact

$$\det(DG_m(\mathbf{0})) = \alpha_{k,m}(\mathbf{0}).$$

Therefore, by the inverse function theorem, there exists an open set V_{m+1} containing $\mathbf{0}$ such that $V_{m+1} = G_m(U_m)$ with G_m and its inverse being one to one, continuous and onto. Let F_m be the flip which flips x_m and x_k . Then define h_{m+1} on V_{m+1} by

$$h_{m+1}(\mathbf{y}) = F_m \circ h_m \circ G_m^{-1}(\mathbf{y}).$$

Thus for $\mathbf{x} \in U_m$,

$$h_{m+1}(G_m(\mathbf{x})) = (F_m \circ h_m)(\mathbf{x}). \quad (26.31)$$

and consequently, since $F_m^2 = I$,

$$F_m \circ h_{m+1} \circ G_m(\mathbf{x}) = h_m(\mathbf{x}) \quad (26.32)$$

It follows

$$\begin{aligned} P_m h_{m+1}(G_m(\mathbf{x})) &= P_m(F_m \circ h_m)(\mathbf{x}) \\ &= \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(\mathbf{x}) & 0 & \cdots & 0 \end{pmatrix}^T \end{aligned}$$

and

$$P_m(G_m(\mathbf{x})) = \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(\mathbf{x}) & 0 & \cdots & 0 \end{pmatrix}^T.$$

Therefore, for $\mathbf{y} \in V_{m+1}$,

$$P_m h_{m+1}(\mathbf{y}) = P_m(\mathbf{y}).$$

As before, $h_{m+1}(\mathbf{0}) = \mathbf{0}$ and $Dh_{m+1}(\mathbf{0})^{-1}$ exists. Therefore, we can apply (26.32) repeatedly, obtaining the following:

$$\begin{aligned} h(\mathbf{x}) &= F_1 \circ h_2 \circ G_1(\mathbf{x}) \\ &= F_1 \circ F_2 \circ h_3 \circ G_2 \circ G_1(\mathbf{x}) \\ &\vdots \\ &= F_1 \circ \cdots \circ F_{n-1} \circ h_n \circ G_{n-1} \circ \cdots \circ G_1(\mathbf{x}) \end{aligned}$$

where h_n fixes the first $n-1$ entries,

$$P_{n-1} h_n(\mathbf{x}) = P_{n-1}(\mathbf{x}) = \begin{pmatrix} x_1 & \cdots & x_{n-1} & 0 \end{pmatrix}^T,$$

and so $h_n(\mathbf{x})$ is a primitive mapping of the form

$$h_n(\mathbf{x}) = \begin{pmatrix} x_1 & \cdots & x_{n-1} & \alpha(\mathbf{x}) \end{pmatrix}^T.$$

Therefore, define the primitive function $G_n(\mathbf{x})$ to equal $h_n(\mathbf{x})$. ■

Part II

Differential Equations

Chapter 27

Determinants

27.1 Basic Techniques And Properties

To begin with, the basic computations and properties of determinants are discussed. After this, complete proofs are given for those who are interested.

Actually, determinants were studied before the modern theory of linear algebra. They are very important in differential equations. Much of what was done earlier concerning eigenvalues and eigenvectors could have been presented without determinants, but things like the Wronskian are given to be certain determinants and so it is important to discuss them.

27.1.1 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 27.1.1 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 27.1.2 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 27.1.3 Suppose A is a 3×3 matrix. The ij^{th} **minor**, denoted 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 27.1.4 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 27.1.5 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 27.1.6 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 27.1.7 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 up. This process when applied to the i^{th} row (column) is known as expanding the determinant along the i^{th} row (column).

Example 27.1.8 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”.

$$\overbrace{1(-1)^{1+1}}^{\text{cof}(A)_{11}} \begin{vmatrix} 3 & 2 \\ 2 & 1 \end{vmatrix} + \overbrace{4(-1)^{2+1}}^{\text{cof}(A)_{21}} \begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} + \overbrace{3(-1)^{3+1}}^{\text{cof}(A)_{31}} \begin{vmatrix} 2 & 3 \\ 3 & 2 \end{vmatrix} = 0.$$

You see, we just followed 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.

Definition 27.1.9 Suppose A is a 4×4 matrix. The ij^{th} **minor** is the determinant of the 3×3 matrix you obtain when you delete the i^{th} row and the j^{th} column. The ij^{th} **cofactor**, $\text{cof}(A)_{ij}$ 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.

Definition 27.1.10 The determinant of a 4×4 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 together. This process when applied to the i^{th} row (column) is known as expanding the determinant along the i^{th} row (column).

Example 27.1.11 Find $\det(A)$ where

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 4 & 2 & 3 \\ 1 & 3 & 4 & 5 \\ 3 & 4 & 3 & 2 \end{pmatrix}$$

As in the case of a 3×3 matrix, you can expand this along any row or column. Lets pick the third column. $\det(A) =$

$$\begin{aligned} & 3(-1)^{1+3} \begin{vmatrix} 5 & 4 & 3 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix} + 2(-1)^{2+3} \begin{vmatrix} 1 & 2 & 4 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix} \\ & + 4(-1)^{3+3} \begin{vmatrix} 1 & 2 & 4 \\ 5 & 4 & 3 \\ 3 & 4 & 2 \end{vmatrix} + 3(-1)^{4+3} \begin{vmatrix} 1 & 2 & 4 \\ 5 & 4 & 3 \\ 1 & 3 & 5 \end{vmatrix}. \end{aligned}$$

Now you know how to expand each of these 3×3 matrices along a row or a column. If you do so, you will get -12 assuming you make no mistakes. You could expand this matrix along any row or any column and assuming you make no mistakes, you will always get the same thing which is defined to be the determinant of the matrix A . This method of evaluating a determinant by expanding along a row or a column is called the **method of Laplace expansion**.

Note that each of the four terms above involves three terms consisting of determinants of 2×2 matrices and each of these will need 2 terms. Therefore, there will be $4 \times 3 \times 2 = 24$ terms to evaluate in order to find the determinant using the method of Laplace expansion. Suppose now you have a 10×10 matrix and you follow the above pattern for evaluating determinants. By analogy to the above, there will be $10! = 3,628,800$ terms involved in the evaluation of such a determinant by Laplace expansion along a row or column. This is a lot of terms.

In addition to the difficulties just discussed, 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.

Definition 27.1.12 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 $\text{cof}(A) = (c_{ij})$ 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 $i j^{\text{th}}$ **minor** of A .) and then multiply this number by $(-1)^{i+j}$. Thus $(-1)^{i+j} \times$ (the $i j^{\text{th}}$ minor) equals the $i j^{\text{th}}$ cofactor. To make the formulas easier to remember, $\text{cof}(A)_{ij}$ will denote the $i j^{\text{th}}$ entry of the cofactor matrix.

With this definition of the cofactor matrix, here is how to define the determinant of an $n \times n$ matrix.

Definition 27.1.13 Let A be an $n \times n$ matrix where $n \geq 2$ and suppose the determinant of an $(n-1) \times (n-1)$ has been defined. Then

$$\det(A) = \sum_{j=1}^n a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^n a_{ij} \operatorname{cof}(A)_{ij}. \quad (27.1)$$

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

Theorem 27.1.14 Expanding the $n \times n$ matrix along any row or column always gives the same answer so the above definition is a good definition.

27.1.2 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 27.1.15 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 27.1.16 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 27.1.17 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

$$\begin{aligned} & 1 \times \left(2 \times \begin{vmatrix} 3 & 33.7 \\ 0 & -1 \end{vmatrix} + 0(-1)^{2+1} \begin{vmatrix} 6 & 7 \\ 0 & -1 \end{vmatrix} \right. \\ & \quad \left. + 0(-1)^{3+1} \begin{vmatrix} 6 & 7 \\ 3 & 33.7 \end{vmatrix} \right) \\ &= 1 \times 2 \times \begin{vmatrix} 3 & 33.7 \\ 0 & -1 \end{vmatrix} \end{aligned}$$

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.

27.1.3 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 27.1.18 *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 27.1.19 *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 27.1.20 *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 27.1.21 *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 27.1.22 *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 27.1.23 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 27.1.24 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 27.1.25 In Theorems 27.1.19 - 27.1.23 you can replace the word, “row” with the word “column”.

There are two other major properties of determinants which do not involve row operations.

Theorem 27.1.26 Let A and B be two $n \times n$ matrices. Then

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

Also,

$$\det(A) = \det(A^T).$$

Example 27.1.27 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 $\det(A) \det(B) = 8 \times (-5) = -40$.

27.1.4 Finding Determinants Using Row Operations

Theorems 27.1.23 - 27.1.25 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 27.1.28 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 27.1.23, 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 27.1.23 this didn't change the value of the determinant. Then the last row was multiplied by (-3) . By Theorem 27.1.19 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 27.1.23. 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 27.1.29 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 27.1.23 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 27.1.25 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 27.1.23 has the same determinant as the original matrix. Let's 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

$$8 \det \begin{pmatrix} -1 & -1 \\ -8 & -4 \end{pmatrix} + 10 \det \begin{pmatrix} -1 & -1 \\ -4 & 1 \end{pmatrix} = -82$$

We 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.

27.2 Applications

27.2.1 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 8.6.2 on Page 144. Also recall the definition of the

cofactor matrix given in Definition 27.1.12 on Page 500. 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.

Theorem 27.2.1 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 27.2.2 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 27.1.23 - 27.1.25 on Page 503, the determinant of this matrix equals the determinant of the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -8 \\ 0 & 0 & -2 \end{pmatrix}$$

which equals 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 it is correct.

Example 27.2.3 Find the inverse of the matrix

$$A = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{6} & \frac{1}{3} & -\frac{1}{2} \\ -\frac{5}{6} & \frac{2}{3} & -\frac{1}{2} \end{pmatrix}$$

First find its determinant. This determinant is $\frac{1}{6}$. The inverse is therefore equal to

$$6 \begin{pmatrix} \begin{vmatrix} 1/3 & -1/2 \\ 2/3 & -1/2 \end{vmatrix} & -\begin{vmatrix} -1/6 & -1/2 \\ -5/6 & -1/2 \end{vmatrix} & \begin{vmatrix} -1/6 & 1/3 \\ -5/6 & 2/3 \end{vmatrix} \\ -\begin{vmatrix} 0 & 1/2 \\ 2/3 & -1/2 \end{vmatrix} & \begin{vmatrix} 1/2 & 1/2 \\ -5/6 & -1/2 \end{vmatrix} & -\begin{vmatrix} 1/2 & 0 \\ -5/6 & 2/3 \end{vmatrix} \\ \begin{vmatrix} 0 & 1/2 \\ 1/3 & -1/2 \end{vmatrix} & -\begin{vmatrix} 1/2 & 1/2 \\ -1/6 & -1/2 \end{vmatrix} & \begin{vmatrix} 1/2 & 0 \\ -1/6 & 1/3 \end{vmatrix} \end{pmatrix}^T.$$

Expanding all the 2×2 determinants this yields

$$6 \begin{pmatrix} 1/6 & 1/3 & 1/6 \\ 1/3 & 1/6 & -1/3 \\ -1/6 & 1/6 & 1/6 \end{pmatrix}^T = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix}$$

Always check your work.

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1/2 & 0 & 1/2 \\ -1/6 & 1/3 & -1/2 \\ -5/6 & 2/3 & -1/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

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.

Proof of Theorem 27.2.1: From the definition of the determinant in terms of expansion along a column, 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 Theorem 27.1.21. However, expanding this matrix B_k 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,

$$\sum_{i=1}^n a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk} \equiv \begin{cases} 1 & \text{if } r = k \\ 0 & \text{if } r \neq k \end{cases}.$$

Now

$$\sum_{i=1}^n a_{ir} \operatorname{cof}(A)_{ik} = \sum_{i=1}^n a_{ir} \operatorname{cof}(A)_{ki}^T$$

which is the kr^{th} entry of $\operatorname{cof}(A)^T A$. Therefore,

$$\frac{\operatorname{cof}(A)^T}{\det(A)} A = I. \quad (27.2)$$

Using the other formula in Definition 27.1.13, and similar reasoning,

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

Now

$$\sum_{j=1}^n a_{rj} \operatorname{cof}(A)_{kj} = \sum_{j=1}^n a_{rj} \operatorname{cof}(A)_{jk}^T$$

which is the rk^{th} entry of $A \operatorname{cof}(A)^T$. Therefore,

$$A \frac{\operatorname{cof}(A)^T}{\det(A)} = I, \quad (27.3)$$

and it follows from 27.2 and 27.3 that $A^{-1} = \left(a_{ij}^{-1} \right)$, where

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

In other words,

$$A^{-1} = \frac{\operatorname{cof}(A)^T}{\det(A)}.$$

Now suppose A^{-1} exists. Then by Theorem 27.1.26,

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

so $\det(A) \neq 0$. ■

This way of finding inverses is especially useful in the case where it is desired to find the inverse of a matrix whose entries are functions.

Example 27.2.4 Suppose

$$A(t) = \begin{pmatrix} e^t & 0 & 0 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{pmatrix}$$

Show that $A(t)^{-1}$ exists and then find it.

First note $\det(A(t)) = e^t \neq 0$ so $A(t)^{-1}$ exists. The cofactor matrix is

$$C(t) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix}$$

and so the inverse is

$$\frac{1}{e^t} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix}^T = \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}.$$

27.2.2 Finding Eigenvalues Using Determinants

It was shown in Theorem 27.2.1 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 11.5.2 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 27.2.5 *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.*

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

$$\begin{aligned} \det(A - \lambda I) &= \det(S^{-1}BS - \lambda S^{-1}IS) = \det(S^{-1}(B - \lambda I)S) \\ &= \det(S^{-1}) \det(B - \lambda I) \det(S) = \det(S^{-1}S) \det(B - \lambda I) \\ &= \det(I) \det(B - \lambda I) = \det(B - \lambda I) \end{aligned} \quad (27.4)$$

Definition 27.2.6 Let A be $n \times n$. Then $\text{trace}(A) \equiv \sum_{i=1}^n A_{ii}$.

Proposition 27.2.7 Let A be $m \times n$ and let B be $n \times m$. Then $\text{trace}(AB) = \text{trace}(BA)$. Also for square matrices A, B , if $A = S^{-1}BS$, then $\text{trace}(A) = \text{trace}(B)$. Also $\det(A) = \det(B)$.

Proof: $\text{trace}(AB) \equiv \sum_i \sum_j A_{ij} B_{ji} = \sum_j \sum_i B_{ji} A_{ij} \equiv \text{trace}(BA)$. Now let A, B be as described. Then

$$\begin{aligned} \text{trace}(A) &= \text{trace}(S^{-1}BS) = \text{trace}((BS)S^{-1}) \\ &= \text{trace}(B(SS^{-1})) = \text{trace}(B) \end{aligned}$$

As to the claim about the determinant, it follows from the properties of the determinant that

$$\begin{aligned} \det(A) &= \det(S^{-1}BS) = \det(S^{-1}) \det(B) \det(S) \\ &= \det(B) \det(S^{-1}S) = \det(B) \blacksquare \end{aligned}$$

These two, the trace and the determinant are two of the so called **principal invariants** of a 3×3 matrix. The reason these are called invariants is that they are the same for A and B if these two are related as described in the above proposition. In this case, the other principal invariant is

$$\frac{1}{2} (\text{trace}(A))^2 - \frac{1}{2} \text{trace}(A^2)$$

It turns out these are related to the coefficients of the characteristic polynomial defined as

$$\det(A - \lambda I)$$

and discussed below.

To see this last is also an invariant, the above proposition implies

$$\begin{aligned} \frac{1}{2} (\text{trace}(A))^2 - \frac{1}{2} \text{trace}(A^2) &= \frac{1}{2} (\text{trace}(S^{-1}BS))^2 - \frac{1}{2} \text{trace}((S^{-1}BS)^2) \\ &= \frac{1}{2} (\text{trace}(B))^2 - \frac{1}{2} \text{trace}((S^{-1}BS)(S^{-1}BS)) \\ &= \frac{1}{2} (\text{trace}(B))^2 - \frac{1}{2} \text{trace}(S^{-1}B^2S) \\ &= \frac{1}{2} (\text{trace}(B))^2 - \frac{1}{2} \text{trace}(B^2) \end{aligned}$$

The physical reason these are important is that their invariance implies they do not change when one uses a different coordinate system to describe points. That which is physically meaningful cannot depend on coordinate system because such coordinate systems are purely artificial constructions used to identify points. Therefore, the principal invariants are good for formulating physical laws. This is as far as we go here. To see much more on these ideas, you should take a course on continuum mechanics. However, the trace and determinant also have a very interesting relation to eigenvalues.

Theorem 27.2.8 The trace of a matrix is the sum of its eigenvalues listed according to multiplicity as a root of the characteristic polynomial. Also, the determinant of the matrix equals the product of its eigenvalues.

Proof: Let A be an $n \times n$ matrix. By Schur's theorem, there is unitary U such that

$$U^*AU = T$$

where T is upper triangular. The characteristic polynomial of T is

$$(\lambda - \mu_1)(\lambda - \mu_2) \cdots (\lambda - \mu_n)$$

where μ_1, \dots, μ_n are the diagonal entries of T . From the above discussion 27.4, these must also be the eigenvalues of A listed according to multiplicity since these two matrices A, T have the same characteristic polynomial. By Proposition 27.2.7 A, T have the same determinant, but since T is upper triangular, the product of its diagonal entries is the product of the eigenvalues of A and this is the common value of the determinant of these two matrices. ■

Example 27.2.9 Find the eigenvalues of the following matrix.

$$A = \begin{pmatrix} 10 & 12 & 1 \\ -8 & -9 & 0 \\ 7 & 8 & 0 \end{pmatrix}$$

You take $\det(A - \lambda I)$ and after much fussing with details, you get the following for the characteristic polynomial.

$$-X^3 + X^2 + X - 1$$

Thus the eigenvalues are the roots of this polynomial. These roots are $1, 1, -1$ when listed according to multiplicity. You can use the above Theorem 27.2.8 as a way to check whether you likely have this right. Indeed, when you add these together, you get 1. When you take the trace of the above matrix, you get 1. This is a little reassurance that you didn't make a mistake. Note that the determinant of the above matrix is -1 which also equals the product of these eigenvalues.

27.2.3 Cramer's Rule

This formula for the inverse also implies a famous procedure known as **Cramer's rule**. Cramer's rule gives a formula for the solutions, \mathbf{x} , to a system of equations, $A\mathbf{x} = \mathbf{y}$ in the special case that A is a square matrix. Note this rule does not apply if you have a system of equations in which there is a different number of equations than variables.

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.

PROCEDURE 27.2.10 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, y if

$$\begin{pmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & -3 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

The determinant of the matrix of coefficients, $\begin{pmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & -3 & 2 \end{pmatrix}$ is -14 . From Cramer's rule, to get x , you replace the first column of A with the right side of the equation and take its determinant and divide by the determinant of A . Thus

$$x = \frac{\begin{vmatrix} 1 & 2 & 1 \\ 2 & 2 & 1 \\ 3 & -3 & 2 \end{vmatrix}}{-14} = \frac{1}{2}$$

Now to find y, z , you do something similar.

$$y = \frac{\begin{vmatrix} 1 & 1 & 1 \\ 3 & 2 & 1 \\ 2 & 3 & 2 \end{vmatrix}}{-14} = -\frac{1}{7}, \quad z = \frac{\begin{vmatrix} 1 & 2 & 1 \\ 3 & 2 & 2 \\ 2 & -3 & 3 \end{vmatrix}}{-14} = \frac{11}{14}$$

You see the pattern. For large systems Cramer's rule is less than useful if you want to find an answer. This is because to use it you must evaluate determinants. However, you have no practical way to evaluate determinants for large matrices other than row operations and if you are using row operations, you might just as well use them to solve the system to begin with. It will be a lot less trouble. Nevertheless, there are situations in which Cramer's rule is useful.

Example 27.2.11 Solve for z if

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix}$$

You could do it by row operations but it might be easier in this case to use Cramer's rule because the matrix of coefficients does not consist of numbers but of functions. Thus

$$z = \frac{\begin{vmatrix} 1 & 0 & 1 \\ 0 & e^t \cos t & t \\ 0 & -e^t \sin t & t^2 \end{vmatrix}}{\begin{vmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{vmatrix}} = t((\cos t)t + \sin t)e^{-t}.$$

You end up doing this sort of thing sometimes in ordinary differential equations in the method of variation of parameters.

27.3 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.

27.4 The Cayley Hamilton Theorem*

Definition 27.4.1 Let A be an $n \times n$ matrix. The characteristic polynomial is defined as

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

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

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

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

The Cayley Hamilton theorem states that every matrix satisfies its characteristic equation, that equation defined by $q_A(t) = 0$. It is one of the most important theorems in linear

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 27.4.2 *Suppose for all $|\lambda|$ large enough,*

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

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

Proof: Multiply by λ^{-m} to obtain

$$A_0\lambda^{-m} + A_1\lambda^{-m+1} + \cdots + A_{m-1}\lambda^{-1} + A_m = 0.$$

Now let $|\lambda| \rightarrow \infty$ to obtain $A_m = 0$. With this, multiply by λ to obtain

$$A_0\lambda^{-m+1} + A_1\lambda^{-m+2} + \cdots + A_{m-1} = 0.$$

Now let $|\lambda| \rightarrow \infty$ to obtain $A_{m-1} = 0$. Continue multiplying by λ and letting $\lambda \rightarrow \infty$ to obtain that all the $A_i = 0$. ■

With the lemma, here is a simple corollary.

Corollary 27.4.3 *Let A_i and B_i be $n \times n$ matrices and suppose*

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

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

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

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

Theorem 27.4.4 *Let A be an $n \times n$ matrix and let $q(\lambda) \equiv \det(\lambda I - A)$ be the characteristic polynomial. Then $q(A) = 0$.*

Proof: Let $C(\lambda)$ equal the transpose of the cofactor matrix of $(\lambda I - A)$ for $|\lambda|$ large. (If $|\lambda|$ is large enough, then λ cannot be in the finite list of eigenvalues of A and so for such λ , $(\lambda I - A)^{-1}$ exists.) Therefore, by Theorem 28.1.14

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

Say

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

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

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

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.

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

Therefore, collecting the terms in the general case,

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

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

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

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

It is good to keep in mind the following example when considering the above proof of the Cayley Hamilton theorem. If $p(\lambda) = q(\lambda)$ for all λ or for all λ large enough where $p(\lambda), q(\lambda)$ are polynomials having matrix coefficients, then it is not necessarily the case that $p(A) = q(A)$ for A a matrix of an appropriate size. Let

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, E_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Then a short computation shows that for all complex λ ,

$$(\lambda I + E_1)(\lambda I + E_2) = (\lambda^2 + \lambda)I = (\lambda I + E_2)(\lambda I + E_1)$$

However,

$$(NI + E_1)(NI + E_2) \neq (NI + E_2)(NI + E_1)$$

The reason this can take place is that N fails to commute with E_i . Of course a scalar commutes with any matrix so there was no difficulty in obtaining that the matrix equation held for arbitrary λ , but this factored equation does not continue to hold if λ is replaced by a matrix. In the above proof of the Cayley Hamilton theorem, this issue was avoided by considering only polynomials which are of the form $C_0 + C_1\lambda + \cdots$ in which the polynomial identity held because the corresponding matrix coefficients were equal. However, you can also argue that in the above proof, the C_i each commute with A .

Theorem 27.4.5 *Let $q(\lambda)$ be the characteristic polynomial and $p(\lambda)$ the minimal polynomial. Then there is a polynomial $l(\lambda)$ which could be a constant such that $q(\lambda) = l(\lambda)p(\lambda)$.*

Proof: By the division algorithm, $q(\lambda) = p(\lambda)l(\lambda) + r(\lambda)$ where the degree of $r(\lambda)$ is less than the degree of $p(\lambda)$ or else $r(\lambda) = 0$. But then, substituting in A , you get $r(A) = 0$ which is impossible if its degree is less than that of $p(\lambda)$. It follows that $r(\lambda) = 0$ and so the claim is established. $p(\lambda)$ “divides” $q(\lambda)$. ■

27.5 Exercises

1. Find the determinants of the following matrices.

$$(a) \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 2 \\ 0 & 9 & 8 \end{pmatrix} \text{ (The answer is 31.)} \quad 375.)$$

$$(b) \begin{pmatrix} 4 & 3 & 2 \\ 1 & 7 & 8 \\ 3 & -9 & 3 \end{pmatrix} \text{ (The answer is } -2.)$$

$$(c) \begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 3 & 2 & 3 \\ 4 & 1 & 5 & 0 \\ 1 & 2 & 1 & 2 \end{pmatrix}, \text{ (The answer is } -2.)$$

2. Find the following determinant by expanding along the first row and second column.

$$\begin{vmatrix} 1 & 2 & 1 \\ 2 & 1 & 3 \\ 2 & 1 & 1 \end{vmatrix}$$

3. Find the following determinant by expanding along the first column and third row.

$$\begin{vmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{vmatrix}$$

4. Find the following determinant by expanding along the second row and first column.

$$\begin{vmatrix} 1 & 2 & 1 \\ 2 & 1 & 3 \\ 2 & 1 & 1 \end{vmatrix}$$

5. Compute the determinant by cofactor expansion. Pick the easiest row or column to use.

$$\begin{vmatrix} 1 & 0 & 0 & 1 \\ 2 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \\ 2 & 1 & 3 & 1 \end{vmatrix}$$

6. Find the determinant using row operations.

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

7. Find the determinant using row operations.

$$\begin{vmatrix} 2 & 1 & 3 \\ 2 & 4 & 2 \\ 1 & 4 & -5 \end{vmatrix}$$

8. Find the determinant using row operations.

$$\begin{vmatrix} 1 & 2 & 1 & 2 \\ 3 & 1 & -2 & 3 \\ -1 & 0 & 3 & 1 \\ 2 & 3 & 2 & -2 \end{vmatrix}$$

9. Find the determinant using row operations.

$$\begin{vmatrix} 1 & 4 & 1 & 2 \\ 3 & 2 & -2 & 3 \\ -1 & 0 & 3 & 3 \\ 2 & 1 & 2 & -2 \end{vmatrix}$$

10. Verify an example of each property of determinants found in Theorems 27.1.23 - 27.1.25 for 2×2 matrices.
11. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

12. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} c & d \\ a & b \end{pmatrix}$$

13. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & b \\ a+c & b+d \end{pmatrix}$$

14. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & b \\ 2c & 2d \end{pmatrix}$$

15. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} b & a \\ d & c \end{pmatrix}$$

16. Let A be an $r \times r$ matrix and suppose there are $r - 1$ rows (columns) such that all rows (columns) are linear combinations of these $r - 1$ rows (columns). Show $\det(A) = 0$.
17. Show $\det(aA) = a^n \det(A)$ where here A is an $n \times n$ matrix and a is a scalar.
18. Illustrate with an example of 2×2 matrices that the determinant of a product equals the product of the determinants.
19. Is it true that $\det(A + B) = \det(A) + \det(B)$? If this is so, explain why it is so and if it is not so, give a counter example.
20. An $n \times n$ matrix is called **nilpotent** if for some positive integer, k it follows $A^k = 0$. If A is a nilpotent matrix and k is the smallest possible integer such that $A^k = 0$, what are the possible values of $\det(A)$?
21. A matrix is said to be **orthogonal** if $A^T A = I$. Thus the inverse of an orthogonal matrix is just its transpose. What are the possible values of $\det(A)$ if A is an orthogonal matrix?
22. Fill in the missing entries to make the matrix orthogonal as in Problem 21.

$$\begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{\sqrt{12}}{6} \\ \frac{1}{\sqrt{2}} & - & - \\ - & \frac{\sqrt{6}}{3} & - \end{pmatrix}.$$

23. Let A and B be two $n \times n$ matrices. $A \sim B$ (A is **similar** to B) means there exists an invertible matrix S such that $A = S^{-1}BS$. Show that if $A \sim B$, then $B \sim A$. Show also that $A \sim A$ and that if $A \sim B$ and $B \sim C$, then $A \sim C$.
24. In the context of Problem 23 show that if $A \sim B$, then $\det(A) = \det(B)$.
25. Two $n \times n$ matrices, A and B , are similar if $B = S^{-1}AS$ for some invertible $n \times n$ matrix S . Show that if two matrices are similar, they have the same characteristic polynomials. The characteristic polynomial of an $n \times n$ matrix M is the polynomial, $\det(\lambda I - M)$.
26. Tell whether the statement is true or false.
 - (a) If A is a 3×3 matrix with a zero determinant, then one column must be a multiple of some other column.
 - (b) If any two columns of a square matrix are equal, then the determinant of the matrix equals zero.
 - (c) For A and B two $n \times n$ matrices, $\det(A + B) = \det(A) + \det(B)$.
 - (d) For A an $n \times n$ matrix, $\det(3A) = 3 \det(A)$
 - (e) If A^{-1} exists then $\det(A^{-1}) = \det(A)^{-1}$.
 - (f) If B is obtained by multiplying a single row of A by 4 then $\det(B) = 4 \det(A)$.
 - (g) For A an $n \times n$ matrix, $\det(-A) = (-1)^n \det(A)$.

- (h) If A is a real $n \times n$ matrix, then $\det(A^T A) \geq 0$.
- (i) Cramer's rule is useful for finding solutions to systems of linear equations in which there is an infinite set of solutions.
- (j) If $A^k = 0$ for some positive integer, k , then $\det(A) = 0$.
- (k) If $Ax = 0$ for some $x \neq 0$, then $\det(A) = 0$.

27. Use Cramer's rule to find the solution to $x + 2y = 1, 2x - y = 2$.

28. Use Cramer's rule to find the solution to $x + 2y + z = 1, 2x - y - z = 2, x + z = 1$.

29. Here is a matrix,

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 1 \\ 3 & 1 & 0 \end{pmatrix}$$

Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

30. Here is a matrix,

$$\begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix}$$

Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

31. Here is a matrix,

$$\begin{pmatrix} 1 & 3 & 3 \\ 2 & 4 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

32. Here is a matrix,

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 1 \\ 2 & 6 & 7 \end{pmatrix}$$

Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

33. Here is a matrix,

$$\begin{pmatrix} 1 & 0 & 3 \\ 1 & 0 & 1 \\ 3 & 1 & 0 \end{pmatrix}$$

Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

34. Use the formula for the inverse in terms of the cofactor matrix to find if possible the inverses of the matrices

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 1 \\ 4 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 1 \\ 2 & 3 & 0 \\ 0 & 1 & 2 \end{pmatrix}.$$

If the inverse does not exist, explain why.

35. Here is a matrix,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}$$

Does there exist a value of t for which this matrix fails to have an inverse? Explain.

36. Here is a matrix,

$$\begin{pmatrix} 1 & t & t^2 \\ 0 & 1 & 2t \\ t & 0 & 2 \end{pmatrix}$$

Does there exist a value of t for which this matrix fails to have an inverse? Explain.

37. Here is a matrix,

$$\begin{pmatrix} e^t & \cosh t & \sinh t \\ e^t & \sinh t & \cosh t \\ e^t & \cosh t & \sinh t \end{pmatrix}$$

Does there exist a value of t for which this matrix fails to have an inverse? Explain.

38. Show that if $\det(A) \neq 0$ for A an $n \times n$ matrix, it follows that if $Ax = 0$, then $x = 0$.

39. Suppose A, B are $n \times n$ matrices and that $AB = I$. Show that then $BA = I$. **Hint:** You might do something like this: First explain why $\det(A), \det(B)$ are both nonzero. Then $(AB)A = A$ and then show $BA(BA - I) = 0$. From this use what is given to conclude $A(BA - I) = 0$. Then use Problem 38.

40. Use the formula for the inverse in terms of the cofactor matrix to find the inverse of the matrix

$$A = \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & e^t \cos t - e^t \sin t & e^t \cos t + e^t \sin t \end{pmatrix}.$$

41. Find the inverse if it exists of the matrix

$$\begin{pmatrix} e^t & \cos t & \sin t \\ e^t & -\sin t & \cos t \\ e^t & -\cos t & -\sin t \end{pmatrix}.$$

42. Here is a matrix,

$$\begin{pmatrix} e^t & e^{-t} \cos t & e^{-t} \sin t \\ e^t & -e^{-t} \cos t - e^{-t} \sin t & -e^{-t} \sin t + e^{-t} \cos t \\ e^t & 2e^{-t} \sin t & -2e^{-t} \cos t \end{pmatrix}$$

Does there exist a value of t for which this matrix fails to have an inverse? Explain.

43. Suppose A is an upper triangular matrix. Show that A^{-1} exists if and only if all elements of the main diagonal are non zero. Is it true that A^{-1} will also be upper triangular? Explain. Is everything the same for lower triangular matrices?

44. If A, B , and C are each $n \times n$ matrices and ABC is invertible, why are each of A, B , and C invertible.

45. Let $F(t) = \det \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix}$. Verify

$$F'(t) = \det \begin{pmatrix} a'(t) & b'(t) \\ c(t) & d(t) \end{pmatrix} + \det \begin{pmatrix} a(t) & b(t) \\ c'(t) & d'(t) \end{pmatrix}.$$

Now suppose

$$F(t) = \det \begin{pmatrix} a(t) & b(t) & c(t) \\ d(t) & e(t) & f(t) \\ g(t) & h(t) & i(t) \end{pmatrix}.$$

Use Laplace expansion and the first part to verify $F'(t) =$

$$\begin{aligned} & \det \begin{pmatrix} a'(t) & b'(t) & c'(t) \\ d(t) & e(t) & f(t) \\ g(t) & h(t) & i(t) \end{pmatrix} + \det \begin{pmatrix} a(t) & b(t) & c(t) \\ d'(t) & e'(t) & f'(t) \\ g(t) & h(t) & i(t) \end{pmatrix} \\ & + \det \begin{pmatrix} a(t) & b(t) & c(t) \\ d(t) & e(t) & f(t) \\ g'(t) & h'(t) & i'(t) \end{pmatrix}. \end{aligned}$$

Conjecture a general result valid for $n \times n$ matrices and explain why it will be true. Can a similar thing be done with the columns?

46. Let $Ly = y^{(n)} + a_{n-1}(x)y^{(n-1)} + \cdots + a_1(x)y' + a_0(x)y$ where the a_i are given continuous functions defined on a closed interval, (a, b) and y is some function which has n derivatives so it makes sense to write Ly . Suppose $Ly_k = 0$ for $k = 1, 2, \dots, n$. The **Wronskian** of these functions, y_i is defined as

$$W(y_1, \dots, y_n)(x) \equiv \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n-1)}(x) & \cdots & y_n^{(n-1)}(x) \end{pmatrix}$$

Show that for $W(x) = W(y_1, \dots, y_n)(x)$ to save space,

$$W'(x) = \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n)}(x) & \cdots & y_n^{(n)}(x) \end{pmatrix}.$$

Now use the differential equation, $Ly = 0$ which is satisfied by each of these functions, y_i and properties of determinants presented above to verify the differential equation $W' + a_{n-1}(x)W = 0$. Give an explicit solution of this linear differential equation, **Abel's formula**, and use your answer to verify that the Wronskian of these solutions to the equation, $Ly = 0$ either vanishes identically on (a, b) or never. **Hint:** To solve the differential equation, let $A'(x) = a_{n-1}(x)$ and multiply both sides of the differential equation by $e^{A(x)}$ and then argue the left side is the derivative of something.

47. Find the following determinants and the inverses of the given matrices. You might use MATLAB to do this with no trouble.

$$(a) \det \begin{pmatrix} 2 & 2+2i & 3-3i \\ 2-2i & 5 & 1-7i \\ 3+3i & 1+7i & 16 \end{pmatrix} \quad (b) \det \begin{pmatrix} 10 & 2+6i & 8-6i \\ 2-6i & 9 & 1-7i \\ 8+6i & 1+7i & 17 \end{pmatrix}$$

48. Find the eigenvalues and eigenvectors of the following matrices. List the eigenvalues according to multiplicity as a root of the characteristic polynomial.

$$(a) \begin{pmatrix} 4 & 7 & 5 \\ -2 & -4 & -4 \\ 1 & 3 & 4 \end{pmatrix}$$

$$(b) \begin{pmatrix} 1 & 1 & 2 \\ 0 & 0 & -2 \\ 0 & 1 & 3 \end{pmatrix}$$

$$(c) \begin{pmatrix} -3 & -7 & -2 \\ 4 & 8 & 2 \\ -2 & -3 & 1 \end{pmatrix}$$

$$(d) \begin{pmatrix} 4 & 6 & 3 \\ -2 & -3 & -2 \\ 1 & 2 & 2 \end{pmatrix}$$

49. The eigenspace for an eigenvalue λ is defined to be the span of all eigenvectors. If the dimension of the eigenspace for each λ equals the multiplicity of the eigenvalue as a root of the characteristic polynomial, then the matrix is said to be nondefective. If, for any eigenvalue, the dimension of the eigenspace called **geometric multiplicity**

is less than the algebraic multiplicity of the eigenvalue as a root of the characteristic polynomial, then the matrix is called **defective**. It can be shown that A can be diagonalized if and only if it is nondefective. See Theorem 11.5.3.

50. The typical situation is that an $n \times n$ matrix has n distinct eigenvalues. In this case, the matrix is always nondefective. This comes from the following theorem which you will show in this problem.

Theorem 27.5.1 *Let A be an $n \times n$ matrix and let $\{\mu_1, \dots, \mu_k\}$ be distinct eigenvalues corresponding to eigenvectors $\{x_1, \dots, x_k\}$. Then this set of eigenvectors is a linearly independent set.*

Do the following. If not independent, then there exist scalars a_i such that

$$\sum_{i=1}^l a_i x_i = \mathbf{0}$$

in which the a_i are not all zero and l is as small as possible for this to take place. Explain why $a_l \neq 0$ and why $l \geq 2$. Then multiply both sides on the left by A and then both sides on the left by μ_l . Subtract and obtain a contradiction of some sort, having to do with l being as small as possible and all eigenvectors being nonzero.

Chapter 28

The Mathematical Theory Of Determinants*



28.0.1 The Function sgn

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

Lemma 28.0.1 *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.*

$$\text{sgn}_n(1, \dots, n) = 1 \quad (28.1)$$

$$\text{sgn}_n(i_1, \dots, p, \dots, q, \dots, i_n) = -\text{sgn}_n(i_1, \dots, q, \dots, p, \dots, i_n) \quad (28.2)$$

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) ,

$$\begin{aligned} \text{sgn}_n(i_1, \dots, i_{\theta-1}, n, i_{\theta+1}, \dots, i_n) &\equiv \\ (-1)^{n-\theta} \text{sgn}_{n-1}(i_1, \dots, i_{\theta-1}, i_{\theta+1}, \dots, i_n) \end{aligned} \quad (28.3)$$

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

$$\begin{aligned}
 \operatorname{sgn}_n(j_1, \dots, j_n) &\equiv \operatorname{sign} \left(\prod_{r < s} (j_s - j_r) \right) \\
 &= \operatorname{sign} \left(\underbrace{(i_p - i_q)}_{\text{both } p, q} \overbrace{\prod_{p < j < q} (i_j - i_q) \prod_{p < j < q} (i_p - i_j)}^{\text{one of } p, q} \underbrace{\prod_{r < s, r, s \notin \{p, q\}} (i_s - i_r)}_{\text{neither } p \text{ nor } q} \right)
 \end{aligned}$$

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 28.0.2 *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 28.1 and 28.2, 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 28.2 gives both functions are equal to zero for that ordered list. ■

Definition 28.0.3 When you have an ordered list of distinct numbers 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 .

28.1 The Determinant

Definition 28.1.1 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,

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

28.1.1 The Definition

Definition 28.1.2 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.

28.1.2 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

$$\det(A(r_1, \dots, r_n)) = \sum_{(k_1, \dots, k_n)} \text{sgn}(k_1, \dots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} \quad (28.4)$$

and

$$A(1, \dots, n) = A.$$

Proposition 28.1.3 *Let*

$$(r_1, \dots, r_n)$$

be an ordered list of numbers from $\{1, \dots, n\}$. Then

$$\text{sgn}(r_1, \dots, r_n) \det(A)$$

$$= \sum_{(k_1, \dots, k_n)} \text{sgn}(k_1, \dots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} \quad (28.5)$$

$$= \det(A(r_1, \dots, r_n)). \quad (28.6)$$

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

$$\det(A(1, \dots, r, \dots, s, \dots, n)) = \quad (28.7)$$

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

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

$$\begin{aligned} &= \sum_{(k_1, \dots, k_n)} \text{sgn}(k_1, \dots, k_s, \dots, k_r, \dots, k_n) a_{1k_1} \cdots a_{rk_s} \cdots a_{sk_r} \cdots a_{nk_n} \\ &= \sum_{(k_1, \dots, k_n)} -\text{sgn} \left(k_1, \dots, \overbrace{k_r, \dots, k_s}^{\text{These got switched}}, \dots, k_n \right) a_{1k_1} \cdots a_{sk_r} \cdots a_{rk_s} \cdots a_{nk_n} \\ &= -\det(A(1, \dots, s, \dots, r, \dots, n)). \end{aligned} \quad (28.8)$$

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 28.0.1, this implies

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

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 28.7-28.8 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 28.1.4 *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.

28.1.3 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 28.1.5 *The following formula for $\det(A)$ is valid.*

$$\det(A) = \frac{1}{n!} \sum_{(r_1, \dots, r_n)} \sum_{(k_1, \dots, k_n)} \text{sgn}(r_1, \dots, r_n) \text{sgn}(k_1, \dots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}. \quad (28.9)$$

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 28.1.3, 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 . ■

28.1.4 The Alternating Property Of The Determinant

Corollary 28.1.6 *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 28.1.3 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 28.1.5 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.

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

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

28.1.5 Linear Combinations And Determinants

Linear combinations have been discussed already. However, here is a review and some new terminology.

Definition 28.1.7 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 28.1.8 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 $A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix}$ be the columns of A and suppose the condition that one column is a linear combination of r of the others is satisfied. Then by using Corollary

28.1.6 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 $\mathbf{a}_n = \sum_{k=1}^r c_k \mathbf{a}_k$ and so

$$\det(B) = \det \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \sum_{k=1}^r c_k \mathbf{a}_k \end{pmatrix}.$$

By Corollary 28.1.6

$$\det(B) = \sum_{k=1}^r c_k \det \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \mathbf{a}_k \end{pmatrix} = 0.$$

because there are two equal columns. The case for rows follows from the fact that $\det(A) = \det(A^T)$. ■

28.1.6 The Determinant Of A Product

Recall the following definition of matrix multiplication.

Definition 28.1.9 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 28.1.10 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 28.1.3,

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

28.1.7 Cofactor Expansions

Lemma 28.1.11 Suppose a matrix is of the form

$$M = \begin{pmatrix} A & * \\ \mathbf{0} & a \end{pmatrix} \quad (28.10)$$

or

$$M = \begin{pmatrix} A & \mathbf{0} \\ * & a \end{pmatrix} \quad (28.11)$$

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 28.0.1, $\det(M)$ 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 28.11. 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 28.10 use Corollary 28.1.5 and 28.11 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 28.1.12 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 i^{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 i^{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 28.1.13 Let A be an $n \times n$ matrix where $n \geq 2$. Then

$$\det(A) = \sum_{j=1}^n a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^n a_{ij} \operatorname{cof}(A)_{ij}. \quad (28.12)$$

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 28.1.6,

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

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 $\text{cof}(A)_{ij} \equiv (-1)^{i+j} \det(A^{ij})$. At this point, recall that from Proposition 28.1.3, 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 28.1.11,

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

Therefore,

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

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

$$\begin{aligned} \det(A) &= \det(A^T) = \sum_{j=1}^n a_{ij}^T \text{cof}(A^T)_{ij} \\ &= \sum_{j=1}^n a_{ji} \text{cof}(A)_{ji} \end{aligned}$$

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

28.1.8 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 28.1.14 A^{-1} exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = (a_{ij}^{-1})$ where

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

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

Proof: By Theorem 28.1.13 and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,

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

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 28.1.6. 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 28.1.13, 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 28.1.10,

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

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 28.1.15 *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 28.1.10 implies

$$\det B \det A = 1$$

and so $\det A \neq 0$. Therefore from Theorem 28.1.14, 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 28.1.14 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 .

28.1.9 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 $(y_1 \cdots, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer's rule.

Chapter 29

First Order Scalar ODE

29.1 First Order Linear Equations

The homogeneous first order constant coefficient linear differential equation is a differential equation of the form

$$y' + ay = 0. \quad (29.1)$$

It is arguably the most important differential equation in existence. Generalizations of it include the entire subject of linear differential equations and even many of the most important partial differential equations occurring in applications.

Here is how to find the solutions to this equation. Multiply both sides of the equation by e^{at} . Then use the product and chain rules to verify that

$$e^{at} (y' + ay) = \frac{d}{dt} (e^{at} y) = 0.$$

Therefore, since the derivative of the function $t \rightarrow e^{at} y(t)$ equals zero, it follows this function must equal some constant C . Consequently, $ye^{at} = C$ and so $y(t) = Ce^{-at}$. This shows that if there is a solution of the equation, $y' + ay = 0$, then it must be of the form Ce^{-at} for some constant, C . You should verify that every function of the form, $y(t) = Ce^{-at}$ is a solution of the above differential equation, showing that this yields all solutions. This proves the following theorem.

Theorem 29.1.1 *The solutions to the equation, $y' + ay = 0$ for a real number consist of all functions of the form, Ce^{-at} where C is some constant.*

Example 29.1.2 *Radioactive substances decay in the following way. The rate of decay is proportional to the amount present. In other words, letting $A(t)$ denote the amount of the radioactive substance at time t , $A(t)$ satisfies the following initial value problem.*

$$A'(t) = -k^2 A(t), \quad A(0) = A_0$$

where A_0 is the initial amount of the substance. What is the solution to the initial value problem?

Write the differential equation as $A'(t) + k^2 A(t) = 0$. From Theorem 29.1.1 the solution is

$$A(t) = Ce^{-k^2 t}$$

and it only remains to find C . Letting $t = 0$, it follows $A_0 = A(0) = C$. Thus $A(t) = A_0 \exp(-k^2 t)$.

Now here is another problem which is a little harder because it has something extra added in at the end.

Example 29.1.3 Find solutions to $y' = 2y + 1$.

Here is how you do it:

1. Write as $y' - 2y = 1$
2. Find an “**Integrating Factor**” $\int (-2) dt = -2t$. Note that I didn’t bother to add in the arbitrary constant. This is because it does not matter. You don’t care about finding all integrating factors. You just need one. Then an integrating factor is e^{-2t} .
3. Multiply both sides of the equation by the integrating factor.

$$e^{-2t} (y' - 2y) = \frac{d}{dt} (e^{-2t} y(t)) = e^{-2t} (1)$$

Note that the first equal sign follows from the product rule and the chain rule. **This is why we multiply by the integrating factor, to get the derivative of something equal to something known.**

4. Take antiderivatives of both sides.

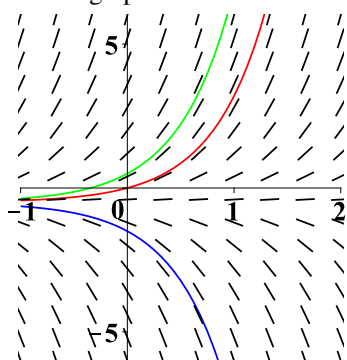
$$e^{-2t} y(t) = \int e^{-2t} dt = -\frac{1}{2} e^{-2t} + C$$

Thus

$$y(t) = -\frac{1}{2} + Ce^{2t}$$

This time you need to be sure to keep the constant of integration because it **does** matter.

Note that by varying C you get different solutions to the differential equation. Now here are graphs of a few of these solutions along with the slope field.



Note how the solutions follow the slope field. How do you determine the “right value” of C ? This involves an

INITIAL CONDITION

An initial condition involves specifying a particular point which is to lie on the graph of the solution to the differential equation. Then you can see from the picture that, having made this specification, the rest of the graph should be determined by the need to follow the slope field. When you have specified the initial condition as well as the differential equation, the problem is called an **initial value problem**.

Example 29.1.4 Find the solution to the initial value problem

$$y' = 2y + 1, y(1) = 2$$

From the above example, all solutions are of the form $y = -\frac{1}{2} + Ce^{2t}$. It is now just a matter of finding the value of C which will cause the given point $(1, 2)$, expressed by saying that $y(1) = 2$, to lie on the graph of y . Thus you need to have $2 = -\frac{1}{2} + Ce^2$. Then you just need to solve this equation for C . This yields $C = \frac{5}{2e^2}$. Therefore, **the** solution to the initial value problem is

$$y = -\frac{1}{2} + \frac{5}{2e^2}e^{2t}$$

Note the use of the definite article. There is **only one** solution to this initial value problem although there are infinitely many solutions to the differential equation, three of which were graphed above. This uniqueness property will be discussed more later, but for now, you can see roughly why this is. It comes from the need for the solution to follow the slope field, so if you specify a point on the curve, you have essentially determined it.

Example 29.1.5 Find the solution to the initial value problem

$$y' + 2ty = \sin(t)e^{-t^2}, y(0) = 3.$$

1. Find the integrating factor. $\int 2tdt = t^2$. Integrating factor: $\exp(t^2) = e^{t^2}$.
2. Multiply both sides by the integrating factor.

$$\exp(t^2)(y' + 2ty) = \frac{d}{dt}(\exp(t^2)y) = \sin(t)$$

3. Take \int of both sides.

$$\exp(t^2)y(t) = -\cos(t) + C$$

4. Solve for y

$$y = \exp(-t^2)(C - \cos(t))$$

5. Find C to satisfy the initial condition.

$$3 = C - 1, C = 4.$$

6. Place value of C you just found in the formula for y

$$y = \exp(-t^2)(4 - \cos(t))$$

Now at this point, you should check and see if it works. It needs to solve both the initial condition and the differential equation.

Example 29.1.6 Find the solutions to

$$y' + a(t)y = b(t).$$

1. Find integrating factor. $A(t) + C \equiv \int a(t)$. Integrating factor: $\exp(A(t))$

2. Multiply by $\exp(A(t))$

$$\overbrace{\exp(A(t)) (y' + a(t)y)}^{\text{chain rule and product rule}} = \frac{d}{dt} (\exp(A(t))y) = \exp(A(t))b(t)$$

3. Do \int to both sides. Pick $F(t) \in \int \exp(A(t))b(t)dt$.

$$\exp(A(t))y(t) = F(t) + C$$

$$y(t) = \exp(-A(t))F(t) + C\exp(-A(t))$$

This proves the following theorem.

Theorem 29.1.7 *The solutions to the equation, $y' + a(t)y = b(t)$ consist of all functions of the form $y(t) = e^{-A(t)}F(t) + e^{-A(t)}C$ where $F(t) \in \int e^{A(t)}b(t)dt$ and C is a constant, $A'(t) = a(t)$.*

Finally, here is a uniqueness theorem.

Theorem 29.1.8 *If $a(t)$ is a continuous function, there is at most one solution to the initial value problem, $y' + a(t)y = b(t)$, $y(r) = y_0$.*

Proof: If there were two solutions y_1 and y_2 , then letting $w = y_1 - y_2$, it follows $w' + a(t)w = 0$ and $w(r) = 0$. Then multiplying both sides of the differential equation by $e^{A(t)}$ where $A'(t) = a(t)$, it follows $(e^{A(t)}w)' = 0$ and so $e^{A(t)}w(t) = C$ for some constant, C . However, $w(r) = 0$ and so this constant can only be 0. Hence $w = 0$ and so $y_1 = y_2$. ■

Finally, consider the general linear initial value problem.

Definition 29.1.9 *A linear differential equation is one which is of the form*

$$y' + a(t)y = b(t)$$

where a, b are continuous. The corresponding initial value problem is

$$y' + a(t)y = b(t), y(t_0) = y_0.$$

Now here are the steps for solving the initial value problem.

1. Find the integrating factor $\int a(t)dt \equiv A(t) + C$. The integrating factor is $\exp(A(t)) = e^{A(t)}$.
2. Multiply both sides by the integrating factor.

$$\exp(A(t)) (y'(t) + a(t)y(t)) = \frac{d}{dt} (\exp(A(t))y(t)) = \exp(A(t))b(t)$$

Why is this so? It involves the chain rule and the product rule.

$$\begin{aligned} \frac{d}{dt} (\exp(A(t))y(t)) &= \exp(A(t))A'(t)y(t) + \exp(A(t))y'(t) \\ &= \exp(A(t))a(t)y(t) + \exp(A(t))y'(t) \\ &= \exp(A(t)) (y'(t) + a(t)y(t)) \end{aligned}$$

3. Next do $\int_{t_0}^t$ to both sides.

$$\int_{t_0}^t \frac{d}{ds} (\exp(A(s)) y(s)) ds = \int_{t_0}^t \exp(A(s)) b(s) ds$$

Then by the fundamental theorem of calculus,

$$\exp(A(t)) y(t) - \exp(A(t_0)) y(t_0) = \int_{t_0}^t \exp(A(s)) b(s) ds$$

and so, you can solve for $y(t)$ and get

$$\begin{aligned} y(t) &= \exp(-A(t)) \exp(A(t_0)) y(t_0) + \exp(-A(t)) \int_{t_0}^t \exp(A(s)) b(s) ds \\ &= \exp(A(t_0) - A(t)) y_0 + \int_{t_0}^t \exp(A(s) - A(t)) b(s) ds \end{aligned}$$

This shows that if the linear initial value problem has a solution, then it must be of the above form. Hence there is at most one solution to the initial value problem. Does the above formula actually give a solution to the initial value problem? Let $y(t)$ be given by that formula. Then

$$y(t_0) = \exp(0) y_0 + \int_{t_0}^{t_0} \exp(A(s) - A(t)) b(s) ds = y_0$$

so the initial condition holds. Does it solve the differential equation? By the chain rule and the fundamental theorem of calculus,

$$\begin{aligned} y'(t) &= (-A'(t)) \exp(A(t_0) - A(t)) y_0 + \exp(-A(t)) \exp(A(t)) b(t) \\ &\quad + (-A'(t)) \exp(-A(t)) \int_{t_0}^t \exp(A(s)) b(s) ds \\ &= (-a(t)) \exp(A(t_0) - A(t)) y_0 + \exp(-A(t)) \exp(A(t)) b(t) \\ &\quad + (-a(t)) \exp(-A(t)) \int_{t_0}^t \exp(A(s)) b(s) ds = -a(t) y(t) + b(t) \end{aligned}$$

so it also is a solution of the linear initial value problem.

Example 29.1.10 *This example illustrates a different notation for differential equations. Find the solutions to*

$$x dy + (2xy - x \sin x) dx = 0$$

The idea is you divide by dx and so the exact meaning is

$$xy' + 2xy = x \sin(x)$$

Then

$$y' + 2y = \sin x, \quad (e^{2x} y)' = e^{2x} \sin x$$

$$e^{2x} y = \int e^{2x} \sin(x) dx = \frac{1}{5} e^{2x} (2 \sin x - \cos x) + C$$

$$y = \frac{1}{5} (2 \sin x - \cos x) + C e^{-2x}$$

The reason for writing it this way is that sometimes you want to find x as a function of y and this notation is neutral in terms of which variable is the independent variable.

Example 29.1.11 A radioactive substance decays in such a way that the rate of change of the amount of the substance is a constant multiple of the amount present, the constant being negative. Thus $\frac{dA}{dt} = -kA$. There is a certain sample of decaying material. Measurements are taken after 5 years and it is found that there is about 9/10 of the original amount present. Find the half life of this material. The half life is the amount of time it takes for half of it to have decayed.

From the equation, $A = A_0 e^{-kt}$. Then $\frac{9}{10}A_0 = A_0 e^{-k(5)}$. Solving this for k yields $\frac{-\ln(.9)}{5} = k$ and so the amount of time to have half of what was started with is T given as a solution to the following equation.

$$e^{-\left(\frac{-\ln(.9)}{5}\right)(T)} = \frac{1}{2}, \text{ so } T = \frac{\ln(.5)}{\ln(.9)/5} = 32.894$$

This kind of thing is associated not just with radioactive material but with other chemicals as well. They degrade over time according to such an equation.

Example 29.1.12 The ancient Babylonians were fascinated with the idea of compound interest. They were interested in how long it would take an initial amount to double. One can understand compound interest compounded continuously using the same kind of differential equation as the above only this time the constant is positive and is the interest rate. Thus

$$\frac{dA}{dt} = kA$$

If the interest rate is 20% per year compounded continuously, how long will it take for an initial amount to double in size?

From the equation, $A = A_0 e^{.2t}$ where A_0 is the initial amount. Then you want to find T such that $2A_0 = A_0 e^{.2T}$ and so

$$T = \frac{\ln 2}{.2} = 5.0 \ln 2 = 3.4657$$

If the rate is r per year and you have n years and the interest is compounded at the end of each year rather than continuously, then the amount is given by the formula $(1+r)^n = A(n)$. Anciently, they used this kind of thing because they did not have differential equations. If the interest rate is 20% compounded monthly, then the amount after n years is $A_0 \left(1 + \frac{.2}{12}\right)^{12n}$ where A_0 is the initial amount. If $n = 3.5$, a use of a calculator shows that

$$\left(1 + \frac{.2}{12}\right)^{12(3.5)} = 2.0022$$

which is very similar to compounding the interest continuously. The rational for this formula is that if it is compounded monthly, then the interest rate per month is $.2/12$. Each successive month is called a payment period.

Example 29.1.13 A lake contains one million gallons of water. A gas tank starts to leak upstream and contaminated water mixed with gasoline starts flowing into the lake at the rate of 1000 gallons per month. This is mixed well due to large numbers of fish in the lake and water flows out at the same rate. The amount of gasoline in the contaminated water varies due to the demand for gas at the gas station and the concentration of gasoline in the contaminated water is $(1 + \sin(t))$ grams per gallon. Find a formula for the concentration of gasoline in the lake in grams per gallon as a function of time in months after a long time.

Let A be the amount of gas in the lake. Then

$$\frac{dA}{dt} = (1 + \sin(t)) \times 1000 - \frac{A}{10^6} 1000 = 1000(1 + \sin(t)) - \frac{1}{10^3} A$$

Rather than worry with the stupid numbers, write this as

$$A' + aA = b(1 + \sin(t)), \quad A(0) = 0$$

Following the procedure for finding solutions to a linear equation,

$$(e^{at} A)' = b(1 + \sin(t)) e^{at}$$

Now it follows that, taking antiderivatives of both sides,

$$e^{at} A = b \left(\frac{e^{at}}{a^3 + a} (a^2 \sin t - a \cos t + a^2 + 1) \right) + C$$

Since $A(0) = 0$, it follows that

$$0 = b \left(\frac{1}{a^3 + a} (-a + a^2 + 1) \right) + C$$

and so $C = -\frac{b}{a^3 + a} (a^2 - a + 1)$. Therefore,

$$A = be^{-at} \left(\frac{e^{at}}{a^3 + a} (a^2 \sin t - a \cos t + a^2 + 1) \right) + \frac{(a - a^2 - 1) be^{-at}}{a^3 + a}$$

Now placing in the formula the values of a and b and then simplifying the result it follows that A equals

$$10^6 e^{-0.001t} \left(1.0 e^{0.001t} - 0.001 e^{0.001t} \cos t + 1.0 \times 10^{-6} e^{0.001t} \sin t - 0.999 \right)$$

Then, dividing by the number of gallons in the lake, this yields for the number of grams per gallon

$$e^{-0.001t} \left(1.0 e^{0.001t} - 0.001 e^{0.001t} \cos t + 1.0 \times 10^{-6} e^{0.001t} \sin t - 0.999 \right)$$

After a long time, the terms having the negative exponential will disappear in the limit and this yields for the number of grams per gallon the formula

$$1 - 0.001 \cos t + 1.0 \times 10^{-6} \sin t$$

Note that this yields approximately 1 gram per gallon. Compare to the concentration of the incoming water. The concentration of the incoming water oscillates about 1 and so does the concentration of gas in the lake, although the oscillations are much much smaller. This is due to the large number of gallons in the lake. You might have expected this but you could not have predicted exact values without the differential equation.

Example 29.1.14 A pumpkin is launched 30° from the horizontal at a speed of 60 feet per second. It is acted on by the force of gravity which delivers an acceleration which is 32 feet per second squared and an acceleration due to air resistance which we assume is .2 times the speed which acts in the opposite direction to the direction of motion. Describe the position of the pumpkin as a function of time.

Let the initial position be at $(0, 0)$ and let the coordinates of the point be $(x(t), y(t))$. What is the initial velocity? It is $(30\sqrt{3}, 30)$. Then the acceleration is given by

$$(x''(t), y''(t)) = -32(0, 1) - .2(x'(t), y'(t))$$

Thus $y'' + .2y' = -32$. Let's solve for y' .

$$(e^{.2t}y')' = (-32)e^{.2t}$$

$$e^{.2t}y'(t) = \frac{-32}{.2}e^{.2t} + C, \text{ so } y'(t) = -160 + Ce^{-.2t}$$

So what is C ? When $t = 0$, we get $C - 160 = 30$ and so $C = 190$. Hence

$$y'(t) = 190e^{-.2t} - 160$$

$$y(t) = -160t - 950e^{-.2t} + D$$

What is D ? When $t = 0$ we want $y(0) = 0$ and so $D = 950$. Thus

$$y(t) = -160t - 950e^{-.2t} + 950$$

As to x ,

$$x'' + .2x' = 0 \text{ so } (x'e^{.2t})' = 0$$

and so $x'(t) = Ce^{-.2t}$. To satisfy the initial condition, $x'(t) = 30\sqrt{3}e^{-.2t}$. Then

$$x(t) = \frac{30\sqrt{3}}{-(1/5)}e^{-.2t} + D$$

What is D ? to satisfy the initial condition for the position, $D = 150\sqrt{3}$ and so

$$x(t) = -150\sqrt{3}e^{-.2t} + 150\sqrt{3}$$

The position of the pumpkin is

$$(x(t), y(t)) = \left(-150\sqrt{3}e^{-.2t} + 150\sqrt{3}, -160t - 950e^{-.2t} + 950\right)$$

The following is a summary of the above discussion.

PROCEDURE 29.1.15 *To solve the first order linear differential equation*

$$y' + a(t)y = f(t),$$

do the following:

1. Find $A(t) \in \int a(t)dt$. That is, find $A(t)$ such that $A'(t) = a(t)$.

2. Multiply both sides by the integrating factor $e^{A(t)}$.

3. The above step yields

$$(e^{A(t)}y(t))' = e^{A(t)}f(t)$$

4. Do $\int dt$ to both sides. Then choose the arbitrary constant to satisfy a given initial condition.

29.2 Bernouli Equations

Some kinds of nonlinear equations can be changed to get a linear equation. An equation of the form

$$y' + a(t)y = b(t)y^\alpha$$

is called a Bernouli equation¹. The trick is to define a new variable, $z = y^{1-\alpha}$. Then $y^\alpha z = y$ and so $z' = (1-\alpha)y^{-\alpha}y'$ which implies $\frac{1}{(1-\alpha)}y^\alpha z' = y'$. Then

$$\frac{1}{(1-\alpha)}y^\alpha z' + a(t)y^\alpha z = b(t)y^\alpha$$

and so

$$z' + (1-\alpha)a(t)z = (1-\alpha)b(t).$$

Now this is a linear equation for z . Solve it and then use the transformation to find y .

Example 29.2.1 Solve $y' + y = ty^3$.

You let $z = y^{-2}$ and make the above substitution. Thus $zy^3 = y$ and

$$z' = (-2)y^{-3}y', \quad y' = -\frac{1}{2}y^3z'$$

and so $-\frac{1}{2}y^3z' + y^3z = ty^3$. Hence, cancelling the y^3 , $z' - 2z = (-2)t$. Then

$$\frac{d}{dt}(e^{-2t}z) = -2te^{-2t}$$

and so

$$e^{-2t}z = te^{-2t} + \frac{1}{2}e^{-2t} + C$$

and so

$$y^{-2} = z = t + \frac{1}{2} + Ce^{2t}$$

and so

$$y^2 = \frac{1}{t + \frac{1}{2} + Ce^{2t}}.$$

When you get this far, it is a good idea to check and see if it works. After all, this is the point of the manipulations, to get the answer. If you get the answer, then if there is a mistake, it is no longer terribly relevant.

$$2yy' = \frac{d}{dt} \left(\frac{1}{t + \frac{1}{2} + Ce^{2t}} \right) = -\frac{8Ce^{2t} + 4}{(2t + 2Ce^{2t} + 1)^2}$$

$$y' = -\frac{8Ce^{2t} + 4}{2y(2t + 2Ce^{2t} + 1)^2}$$

¹This is named after Jacob Bernoulli (1654-1705), one of a whole family of Swiss mathematicians. Others were Johann I and II Daniel, and Nicolaus.

Then

$$\begin{aligned}
 y' + y &= -\frac{8Ce^{2t} + 4}{2y(2t + 2Ce^{2t} + 1)^2} + y \\
 &= -\frac{8Ce^{2t} + 4}{2y(2t + 2Ce^{2t} + 1)^2} + \frac{2y^2(2t + 2Ce^{2t} + 1)^2}{2y(2t + 2Ce^{2t} + 1)^2} \\
 &= -\frac{8Ce^{2t} + 4}{2y(2t + 2Ce^{2t} + 1)^2} + \frac{2\left(\frac{1}{t + \frac{1}{2} + Ce^{2t}}\right)(2t + 2Ce^{2t} + 1)^2}{2y(2t + 2Ce^{2t} + 1)^2} \\
 &= 4\frac{t}{y(2t + 2Ce^{2t} + 1)^2} = \frac{t}{y(t + Ce^{2t} + \frac{1}{2})^2} = \frac{t}{y}y^4 = ty^3
 \end{aligned}$$

so it appears to work.

The following procedure gives a summary of the above.

PROCEDURE 29.2.2 *To solve the Bernouli equation*

$$y' + a(t)y = b(t)y^\alpha, \alpha \neq 1$$

do the following:

1. Change the variable. Let $z = y^{1-\alpha}$. Then $z' = (1-\alpha)y^{-\alpha}y'$, $y^\alpha z = y$.
2. Place in the equation.

$$\frac{1}{1-\alpha}y^\alpha z' + a(t)y^\alpha z = b(t)y^\alpha$$

3. Cancel the y^α and solve the linear equation for z .

29.3 Separable Differential Equations, Stability

Separable differential equations also occur quite often in applications and they are fairly easy to deal with. This section gives a discussion of these equations.

Definition 29.3.1 *Separable differential equations are those which can be written in the form*

$$\frac{dy}{dx} = \frac{f(x)}{g(y)}.$$

The reason these are called separable is that if you formally cross multiply,

$$g(y)dy = f(x)dx$$

and the variables are “separated”. The x variables are on one side and the y variables are on the other.

Proposition 29.3.2 *If $G'(y) = g(y)$ and $F'(x) = f(x)$, then if the equation, $F(x) - G(y) = c$ specifies y as a differentiable function of x , then $x \rightarrow y(x)$ solves the separable differential equation*

$$\frac{dy}{dx} = \frac{f(x)}{g(y)}. \quad (29.2)$$

Proof: Differentiate both sides of $F(x) - G(y) = c$ with respect to x . Using the chain rule,

$$F'(x) - G'(y) \frac{dy}{dx} = 0.$$

Therefore, since $F'(x) = f(x)$ and $G'(y) = g(y)$, $f(x) = g(y) \frac{dy}{dx}$ which is equivalent to 29.2. ■

Definition 29.3.3 The curves $F(x) - G(y) = c$ for various values of c are called **integral curves or solution curves**. It makes sense to think of these as giving a solution if, near a point on the level curve, one variable is a function of the other.

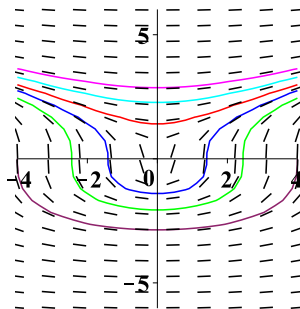
Example 29.3.4 Find the solution to the initial value problem,

$$y^2 y' = x, y(0) = 1.$$

This is a separable equation and in fact, $y^2 dy = x dx$ so the solution to the differential equation is of the form $\frac{y^3}{3} - \frac{x^2}{2} = C$ and it only remains to find the constant C . To do this, you use the initial condition. Letting $x = 0$, it follows $\frac{1}{3} = C$ and so

$$\frac{y^3}{3} - \frac{x^2}{2} = \frac{1}{3}$$

The following picture shows how the integral curves follow the tangent field.



Sometimes, you can't expect to solve for one of the variables in terms of the other. In other words, the integral curve might not be a function of one variable. Here is a nice example from [7].

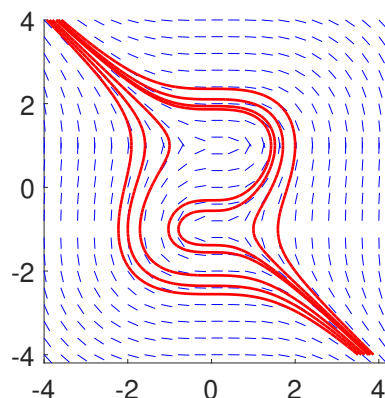
Example 29.3.5 Find integral curves for the equation

$$y' = \frac{x^2}{(1-y^2)}$$

Separating variables, you get $(1-y^2) dy = x^2 dx$ and so the integral curves are of the form

$$\left(y - \frac{y^3}{3}\right) - \frac{x^3}{3} = C$$

Here is a picture of a few of these integral curves along with the slope field.



I used MATLAB to graph the above. One thing might be helpful to mention about MATLAB. It is very good at manipulating matrices and vectors and there is distinctive notation used to accomplish this. For example say you type

```
x=[1,2,3]; y=[2,3,4]; x.*y
```

and then press “enter”. You will get 2,6,12. Of course you would get an error if you wrote $x*y$. Similarly, type

```
[2,4,6,8]/[1,2,3,4]
```

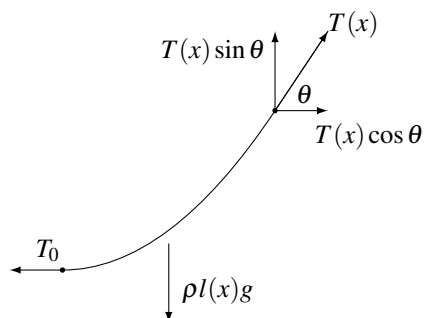
and press “enter”. This yields 2,2,2,2. Of course $[2,4,6,8]/[1,2,3,4]$ doesn’t make any sense.

You can get graphs of some integral curves in MATLAB by typing the following and then “enter”. You don’t have to type it on two lines, but if you want to do so, to get to a new line, you press “shift” and “enter”.

```
>> [x,y]=meshgrid(-4:.1:4,-4:.1:4);
z=y-(y.^3/3+x.^3/3);contour(x,y,z,[-.5,-1,-.3,1,2])
```

Example 29.3.6 *What is 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 t and at 0 respectively, as shown. 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

$$\frac{\sin \theta}{\cos \theta} = l(x) \overbrace{\rho g / T_0}^{\equiv c}.$$

Now letting $y(x)$ denote the y coordinate of the hanging chain corresponding to x ,

$$\frac{\sin \theta}{\cos \theta} = \tan \theta = y'(x).$$

Therefore, this yields

$$y'(x) = cl(x).$$

Now differentiating both sides of the differential equation,

$$y''(x) = cl'(x) = c\sqrt{1 + y'(x)^2}$$

and so

$$\frac{y''(x)}{\sqrt{1 + y'(x)^2}} = c.$$

Let $z(x) = y'(x)$ so the above differential equation becomes

$$\frac{z'(x)}{\sqrt{1 + z^2}} = c.$$

Therefore, $\int \frac{z'(x)}{\sqrt{1 + z^2}} dx = cx + d$. Change the variable in the antiderivative letting $u = z(x)$ and this yields

$$\int \frac{z'(x)}{\sqrt{1 + z^2}} dx = \int \frac{du}{\sqrt{1 + u^2}} = \sinh^{-1}(u) + C = \sinh^{-1}(z(x)) + C.$$

Therefore, combining the constants of integration,

$$\sinh^{-1}(y'(x)) = cx + d$$

and so

$$y'(x) = \sinh(cx + d).$$

Therefore,

$$y(x) = \frac{1}{c} \cosh(cx + d) + k$$

where d and k are some constants and $c = \rho g / T_0$. Curves of this sort are called catenaries. Note these curves result from an assumption that the only forces acting on the chain are as shown.

The next example has to do with population models. It was mentioned earlier. The idea is that if there were infinite resources, population growth would satisfy the differential equation

$$\frac{dy}{dt} = ky$$

where k is a constant. However, resources are not infinite and so k should be modified to be consistent with this. Instead of k , one writes $r\left(1 - \frac{y}{K}\right)$ which will cause the population growth to decrease as soon as y exceeds K . Of course the problem with this is that we are not sure whether K itself is dependent on other factors not included in the model.

Example 29.3.7 *The equation*

$$\frac{dy}{dt} = r\left(1 - \frac{y}{K}\right)y, \quad r, K > 0$$

is called the logistic equation. It models population growth. You see that the right side is equal to 0 at the two values $y = K$ and $y = 0$.

This is a separable equation. Thus

$$\frac{dy}{\left(1 - \frac{y}{K}\right)y} = rdt$$

Now you do \int to both sides. This requires partial fractions on the left.

$$\frac{1}{\left(1 - \frac{y}{K}\right)y} = \frac{1}{K-y} + \frac{1}{y}$$

Therefore,

$$\ln(y) - \ln(K-y) = rt + C$$

if $0 < y < K$. If $y > K$, you get

$$\ln(y) - \ln(y-K) = rt + C$$

Therefore, the integral curves are of the form

$$\ln\left(\frac{y}{K-y}\right) = rt + C$$

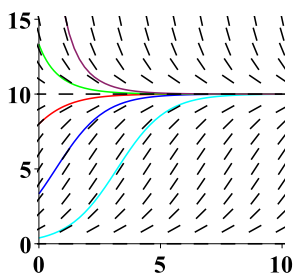
so changing the name of the constant C , it follows that for $y < K$, the integral curves are described by the following function.

$$y = K \frac{Ce^{rt}}{Ce^{rt} + 1}, \quad C > 0$$

In case $y > K$, these curves are described by

$$y = K \frac{Ce^{rt}}{Ce^{rt} - 1}, \quad C > 0$$

What follows is a picture of the slope field along with some of these integral curves in case $r = 1$ and $K = 10$.



The bottom axis is the t axis. Note how all the integral curves in the picture approach K as t increases. This is why K is called a **stable equilibrium point**.

Definition 29.3.8 Consider the equation $\frac{dy}{dt} = f(y)$. Then y_0 is called an **equilibrium point** if $f(y_0) = 0$. Note that the solution to the initial value problem $y' = f(y)$, $y(t_0) = y_0$ is just $y = y_0$. An equilibrium point is **stable** if whenever y_1 is close enough to y_0 , it follows that the solution to the initial value problem $y' = f(y)$, $y(0) = y_1$ stays close to y_0 for all $t > 0$. It is **asymptotically stable** if whenever y_1 is close enough to y_0 , it follows that for y the solution to the initial value problem, $y' = f(y)$, $y(0) = y_1$ satisfies $\lim_{t \rightarrow \infty} y(t) = y_0$. The equilibrium point y_0 is **unstable** if there are initial conditions close to y_0 but the solution does not stay close to y_0 . That is, there exists $\varepsilon > 0$ such that for any $\delta > 0$ there is y_1 with $|y_1 - y_0| < \delta$ but the solution to $y' = f(y)$, $y(0) = y_1$ has the property that for some $t > 0$, $|y(t) - y_0| \geq \varepsilon$. An equilibrium point y_0 is **semi-stable** if it is stable from one side and unstable from the other.

Now observe that $y = 0$ is the solution which results if you begin with the initial condition $y(0) = 0$. If there is nothing to start with, it can't grow. However, if you have any other positive number for $y(0)$, then you see that the solution curve approaches the stable point K . You can see this, not just by looking at the picture but also by taking the limit as $t \rightarrow \infty$ in the above formulae.

One of the interesting things about this equation is that it is possible to determine K the maximum capacity, by taking measurements at three equally spaced times. Suppose you do so at times $t, 2t, 3t$ and obtain y_1, y_2, y_3 respectively. Assume you are in the region where $y < K$. In an actual experiment, this is where you would be. Let $\lambda \equiv e^{rt}$. Then from the above formula for y , you have the equations

$$KC\lambda = y_1(C\lambda + 1), KC\lambda^2 = y_2(C\lambda^2 + 1), KC\lambda^3 = y_3(C\lambda^3 + 1)$$

Then divide the second equation by λ and compare with the first. This shows that $\lambda = y_2/y_1$. Next divide the top equation by $C\lambda$ and the last by $C\lambda^3$. This yields

$$K = y_1 \left(1 + \frac{1}{C\lambda} \right) = y_3 \left(1 + \frac{1}{C\lambda^3} \right)$$

Now it becomes possible to solve for C . This yields

$$C = \frac{(y_1^3 y_3 - y_1^2 y_2^2)}{y_1 y_2^3 - y_2^3 y_3}$$

Then substitute this in to the first equation. This obtains

$$K \left(\frac{(y_1^3 y_3 - y_1^2 y_2^2)}{y_1 y_2^3 - y_2^3 y_3} \right) \frac{y_2}{y_1} = y_1 \left(\frac{(y_1^3 y_3 - y_1^2 y_2^2)}{y_1 y_2^3 - y_2^3 y_3} \left(\frac{y_2}{y_1} \right) + 1 \right)$$

Then you can solve this for K . After some simplification, it yields

$$\frac{y_2^2 y_3 - y_1^2 y_3}{y_2^2 - y_1 y_3} = K$$

Note how the equilibrium point K was stable in the above example. There were only two equilibrium points, K and 0 . The equilibrium point 0 was unstable because if the integral curve started near 0 but slightly positive, it tended to increase to K . Here is another harder example. In this example, there are three equilibrium points.

Example 29.3.9 $\frac{dy}{dt} = -r\left(1 - \frac{y}{T}\right)\left(1 - \frac{y}{K}\right)y$, $r > 0, 0 < T < K$.

This is a separable equation.

$$\frac{dy}{\left(1 - \frac{y}{T}\right)\left(1 - \frac{y}{K}\right)y} = -r dt$$

The partial fractions expansion is

$$\frac{1}{\left(1 - \frac{y}{T}\right)\left(1 - \frac{y}{K}\right)y} = \frac{1}{K-T} \left(\frac{K}{T-y} - \frac{T}{K-y} \right) + \frac{1}{y}$$

Therefore,

$$\frac{-1}{K-T} K \ln|T-y| + \frac{1}{K-T} T \ln|K-y| + \ln|y| = -rt + C$$

Consider the case where $r = 1, T = 5, K = 10$. Then you get

$$\ln \left(\left| \frac{(10-y)y}{(5-y)^2} \right| \right) = -t + C$$

There are cases, depending on where y is. Suppose first that $y \in (0, 10)$. Then you get for a different C

$$\frac{(10-y)y}{(5-y)^2} = Ce^{-t}, \quad C > 0$$

You could solve this for y if you like and get

$$y = \frac{1}{Ce^{-t} + 1} \left(-5\sqrt{Ce^{-t} + 1} + 5Ce^{-t} + 5 \right)$$

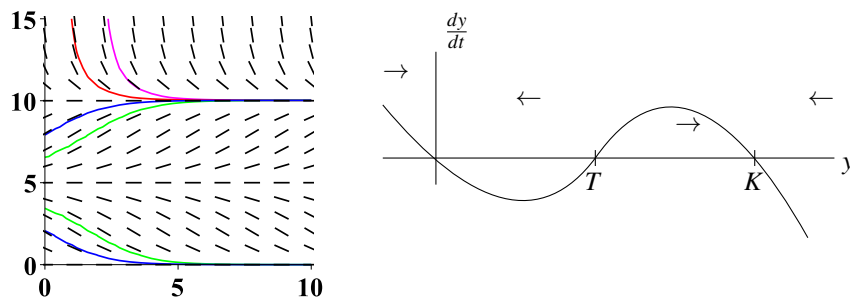
or

$$y = \frac{1}{Ce^{-t} + 1} \left(5\sqrt{Ce^{-t} + 1} + 5Ce^{-t} + 5 \right)$$

On the other hand, if $y > 10$, you get

$$\frac{(y-10)y}{(5-y)^2} = Ce^{-t}$$

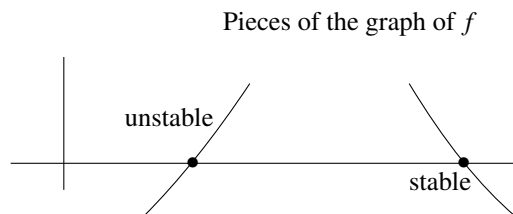
The following is a picture of some integral curves and the slope field. You see how the equilibrium points 10 and 0 are both stable but the equilibrium point 5 is not.



Is there a systematic way to figure this out without doing lots of computer generated pictures? The answer is yes! Furthermore, it is very easy to do. Consider the right side of the equation. If you graph the function $z = f(y)$, you get something which looks like the right side of the above.

Look at the graph. When $y \in (0, T)$, you have the slope of the graph is negative and so, from the equation, $\frac{dy}{dt}$ is negative and so $t \rightarrow y(t)$ is decreasing. (Remember calculus.) If $y \in (T, K)$, then the graph is positive and so $\frac{dy}{dt}$ is positive which requires that $t \rightarrow y(t)$ is increasing. When $y \in (K, \infty)$, the graph is negative and so $t \rightarrow y(t)$ is decreasing. Thus T is unstable, K is stable while 0 is also stable. I have not considered the case where $y < 0$ because this is not too interesting in the example which typically describes y as a population of something. However, you can see from the graph that if $y < 0$, then $t \rightarrow y(t)$ is increasing.

In general, you can consider $y' = f(y)$ and the equilibrium points. The following picture is descriptive of the situation. Such an equation is called **autonomous** because the function on the right depends only on y and not on t .



Proposition 29.3.10 Suppose f is continuous with continuous derivative and that $f(y_0) = 0, f'(y_0) < 0$. Then y_0 is asymptotically stable.

Proof: By continuity of f' , there is $\delta > 0$ such that for $y \in (y_0 - \delta, y_0 + \delta) = I, f'(y) \leq -2\eta, \eta > 0$. Thus

$$\begin{aligned} f(y) &= f(y_0) + f'(y_0)(y - y_0) + o(y - y_0) \\ &= f'(y_0)(y - y_0) + o(y - y_0) \end{aligned}$$

Then if $y_1 \in I$, and if $y(t)$ is the solution to the equation $y' = f(y)$ having this initial condition, then

$$\begin{aligned} y(t) - y_0 &= y_1 - y_0 + \int_0^t f(y(s)) ds \\ &= y_1 - y_0 + \int_0^t f'(y_0)(y(s) - y_0) ds + \int_0^t o(y(s) - y_0) ds \end{aligned}$$

We can also assume δ is small enough that $|o(y - y_0)| < \eta |y - y_0|$ for $y \in I$. Say $y_1 > y_0$. Then by assumption, $t \rightarrow y(t)$ is decreasing since $y' = f(y) < 0$ and so if $y(t)$ fails to converge to y_0 , there would exist $\varepsilon > 0$ which is the limit of $y(t) - y_0$. Then

$$\varepsilon + \int_0^t -f'(y_0)(y(s) - y_0) ds \leq y_1 - y_0 + \eta \int_0^t (y(s) - y_0)$$

Thus

$$\varepsilon + \int_0^t \eta \varepsilon \leq y_1 - y_0$$

which is impossible because as $t \rightarrow \infty$, the left side is unbounded. Similar reasoning shows asymptotic stability if $y_1 < y_0$. ■

PROCEDURE 29.3.11 *To solve a separable equation which can be placed in the form*

$$f(y) dy + g(x) dx = 0$$

do the following:

1. Place an \int in front each term on the left and do what it seems to say.
2. You get $F(y) + G(x) = C$. This is the general solution. Now choose C to satisfy any initial condition which may be given.
3. An equilibrium point for $y' = f(y)$ is a point y_0 where $f(y_0) = 0$. It is an asymptotically stable equilibrium if $f'(y_0) < 0$ and unstable if $f'(y_0) > 0$.

29.4 Homogeneous Equations

Sometimes equations can be made separable by changing the variables appropriately. This occurs in the case of the so called homogeneous equations, those of the form

$$y' = f\left(\frac{y}{x}\right).$$

When this sort of equation occurs, there is an easy trick which will allow you to consider a separable equation.

You define a new variable,

$$u \equiv \frac{y}{x}.$$

Thus $y = ux$ and so

$$y' = u'x + u = f(u).$$

Thus

$$\frac{du}{dx}x = f(u) - u$$

and so

$$\frac{du}{f(u) - u} = \frac{dx}{x}.$$

The variables have now been separated and you go to work on it in the usual way. This method is due to Leibniz² and dates from around 1691.

Example 29.4.1 *Find the solutions of the equation*

$$y' = \frac{y^2 + xy}{x^2}.$$

²Gottfried Wilhelm (von) Leibniz, (1646-1716) is credited with Newton as being one of the inventors of calculus. There was much controversy over who did it first. It is likely that Newton did it first, but Leibniz had superior notation. The notation $\frac{dy}{dx}$ for the derivative and the notation for integrals is due to him. Like many of these men, he was interested in many other subjects besides mathematics, such as philosophy, theology, geology, and medicine.

First note this is of the form

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

Let $u = \frac{y}{x}$ so $y = xu$. Then

$$u'x + u = u^2 + u$$

and so, separating the variables yields

$$\frac{du}{u^2} = \frac{dx}{x}$$

Hence

$$-\frac{1}{u} = \ln|x| + C$$

and so

$$\frac{y}{x} = u = \frac{1}{K - \ln|x|}$$

where $K = -C$. Hence

$$y(x) = \frac{x}{K - \ln|x|}$$

PROCEDURE 29.4.2 *To solve a homogeneous equation, one which can be placed in the form*

$$y' = f\left(\frac{y}{x}\right),$$

do the following:

1. *Define a new variable $v = y/x$. Then $y = xv$ and so $y' = v + xv'$.*
2. *Plug in to the equation.*

$$v + xv' = f(v), \quad x \frac{dv}{dx} = f(v) - v$$

$$\frac{dv}{f(v) - v} = \frac{dx}{x}$$

This is separable. Place \int before each side and do what it says. Then choose the constant of integration to satisfy any initial condition which may be present.

29.5 Exact Equations

Sometimes you have a differential equation of the form

$$M(x, y) dx + N(x, y) dy = 0$$

where $N_x = M_y$. In this happy situation, one can find a function of two variables $f(x, y)$ such that

$$f_x(x, y) = M(x, y), \quad f_y(x, y) = N(x, y) \quad (29.3)$$

and the solution to the equation is of the form

$$f(x, y) = C \quad (29.4)$$

where C is a constant. This function f is called a scalar potential or potential for short.

These equations are called **exact**. Why does * yield a solution? Say the above relation defines y as a function of x . Then using the chain rule,

$$f_x(x, y) + f_y(x, y) \frac{dy}{dx} = 0$$

and so

$$\begin{aligned} f_x(x, y) dx + f_y(x, y) dy &= 0 \\ M(x, y) dx + N(x, y) dy &= 0 \end{aligned}$$

It is easy to see that if there exists a C^2 function f with the property that $f_x = M, f_y = N$, then $N_x = M_y$. This follows because $M_y = f_{xy}$ and $N_x = f_{yx}$. By equality of mixed partial derivatives, you need to have $M_y = N_x$. In fact, if this last condition holds, then there will generally be such a potential function $f(x, y)$.

Why is it that if $N_x = M_y$ then there exists f with the properties described?

Let

$$f(x, y) \equiv \int_0^x M(t, y) dt + N(0, y).$$

Then $f_x(x, y) = M(x, y)$, and formally differentiating across the integral,

$$\begin{aligned} f_y(x, y) &= \int_0^x M_y(t, y) dt + N(0, y) = \int_0^x N_x(t, y) dt \\ &= N(x, y) - N(0, y) + N(0, y) = N(x, y) \end{aligned}$$

In general, this process of $\left(\frac{\partial}{\partial y} \int_0^x M(t, y) dt = \int_0^x M_y(t, y) dt \right)$ has not been proved, but in examples, it will be obviously true. Also, it is formally true when you think of the integral as a sort of sum and use the fact that the derivative of a sum is the sum of the derivatives.

Example 29.5.1 Find the solutions to

$$(\cos(x) + 2xy) dx + x^2 dy = 0$$

You see that this is exact ($2x = 2x$). Then the $f(x, y)$ satisfies $f_x(x, y) = \cos(x) + 2xy$ and so $f(x, y) = \sin(x) + x^2 y + g(y)$. Then taking the partial derivative with respect to y , it follows that $x^2 + g'(y) = x^2$ and so it suffices to let $g(y) = 0$. Then the solutions to this differential equation are

$$\sin(x) + x^2 y = C$$

where C is a constant which would be determined by some sort of an initial condition.

Example 29.5.2 In the above example, determine C if $(x, y) = (\frac{\pi}{2}, 0)$ is to be on the curve which yields a solution to the differential equation.

You need to have $1 = C$ because

$$\sin\left(\frac{\pi}{2}\right) + \left(\frac{\pi}{2}\right)^2 \cdot 0 = C$$

and so the solution in this case is $\sin(x) + x^2y = 1$.

All of the examples of this sort of thing are similar. Exact equations are easy to solve. Physically the solution to these equations is really a statement about the energy being constant.

PROCEDURE 29.5.3 *To solve an exact equation*

$$M(x, y) dx + N(x, y) dy = 0$$

do the following:

1. Check to see if it really is exact by seeing if $N_x = M_y$. If it is, find a scalar potential $f(x, y)$ such that $f_x = M, f_y = N$.
2. The general solution is $f(x, y) = C$. Choose C to satisfy initial conditions.

29.6 The Integrating Factor

It turns out that theoretically, this is the most general method for solving equations

$$m(x, y) dx + n(x, y) dy = 0$$

I want to stress the word “theoretically” however. If the above equation is not exact, the idea is to multiply by a function μ which will make it exact. Thus it would be sufficient to have

$$(\mu m)_y = (\mu n)_x$$

The function μ is called an integrating factor. In other words, it is required that

$$\mu_y m + \mu m_y = \mu_x n + \mu n_x \quad (29.5)$$

This is called a first order linear partial differential equation and we don't know how to solve them. However, we don't need to find all solutions, just one which works. The idea is to look for $\mu = \mu(x)$ or $\mu = \mu(y)$. For us, if there is no such easy solution, the method has failed. So what would happen if there is a solution $\mu = \mu(x)$? then you would have

$$\mu'(x) = \mu(x) \frac{m_y(x, y) - n_x(x, y)}{n(x, y)}$$

and so there will be such an integrating factor if

$$\frac{m_y(x, y) - n_x(x, y)}{n(x, y)}$$

depends only on x . Similarly, there will be an integrating factor $\mu = \mu(y)$ if

$$\frac{n_x - m_y}{m}$$

depends only on y .

Example 29.6.1 Find the solutions to

$$(2y^3 + 2y) dx + (3xy^2 + x) dy = 0$$

The equation is clearly not exact so we look for an integrating factor.

$$\mu_y (2y^3 + 2y) + \mu (6y^2 + 2) = \mu_x (3xy^2 + x) + \mu (3y^2 + 1)$$

We look for one which depends on only one variable. Let's try to find $\mu = \mu(y)$ first. If there is such a solution, then

$$\mu'(y) (2y^3 + 2y) + \mu (6y^2 + 2) = \mu (3y^2 + 1)$$

so it looks like there is such a solution.

$$\mu'(y) = \frac{-3y^2 - 1}{2y^3 + 2y} \mu$$

Thus

$$\frac{d\mu}{\mu} = \frac{-3y^2 - 1}{2y^3 + 2y} dy \quad (29.6)$$

Then

$$\ln(\mu) = \int \frac{-3y^2 - 1}{2y^3 + 2y} dy = -\frac{1}{2} \ln(y^3 + y)$$

An integrating factor would be $1/\sqrt{y^3 + y}$. This looks really ugly. Let's try and find one which depends only on x . Then

$$\mu (6y^2 + 2 - (3y^2 + 1)) = \mu'(x) (3xy^2 + x)$$

$$\mu (3y^2 + 1) = \mu'(x) (3xy^2 + x)$$

$$\mu'(x) = \mu(x) \frac{1}{x}$$

Thus $\mu = x$ is also an integrating factor. Which would you rather use? Multiply by x . The equation is now

$$(2xy^3 + 2yx) dx + (3x^2y^2 + x^2) dy = 0$$

and it is an exact equation so you are in the situation of the preceding section. You find a scalar potential. The manipulations explained in the last section yield $x^2y^3 + x^2y$ as a scalar potential. Then the solutions are

$$x^2y^3 + x^2y = C$$

All of these are the same. You begin with 29.5 and look for solutions. In particular you look for solutions that depend on only one variable. If you can find one, then the problem has been reduced to that of the preceding section. If you can't find such a solution, then you give up. Under general conditions, it can be proved that solutions exist but as usual in mathematics, there is a big gap between knowing something exists and finding it. However,

here is something nice which was discovered by Euler back in the 1700s. It is called Euler's identity along with the more famous one involving complex numbers.³

Lemma 29.6.2 *A function $M(x, y)$ is homogeneous of degree α if $M(tx, ty) = t^\alpha M(x, y)$. For such a function,*

$$\alpha M(x, y) = x \frac{\partial M}{\partial x}(x, y) + y \frac{\partial M}{\partial y}(x, y)$$

Proof: You use the chain rule to differentiate both sides of the equation

$$M(tx, ty) = t^\alpha M(x, y)$$

with respect to t . Thus

$$\alpha t^{\alpha-1} M(x, y) = x \frac{\partial M}{\partial x}(tx, ty) + y \frac{\partial M}{\partial y}(tx, ty)$$

Now let $t = 1$. ■

The reason this is pretty nice is that if you have the equation

$$M(x, y) dx + N(x, y) dy = 0$$

and both M and N are homogeneous of degree α , then

$$\frac{1}{xM + yN}$$

is an integrating factor. Here $M_x = \frac{\partial M}{\partial x}$. We verify this next. It is so if

$$\left(\frac{N}{xM + yN} \right)_x = \left(\frac{M}{xM + yN} \right)_y$$

By the quotient rule, this will be so if and only if

$$N_x(xM + yN) - N(M + xM_x + yN_x) =$$

$$M_y(xM + yN) - M(xM_y + N + yN_y)$$

In both sides of the above equation, some terms cancel and it follows that the desired result follows if and only if

$$xMN_x - (NM + xM_xN) = yNM_y - (MN + yMN_y)$$

³Euler (1707-1783) (pronounced "oiler") was a Swiss mathematician, a student of Johann Bernoulli. He is one of the most important mathematicians to ever live. He wrote more mathematics than anyone else, some 530 books and papers in all areas of the subject. His very unusual memory allowed him to continue doing mathematical research even after he went blind in 1766. Many of the ideas in this book are due to him. Like many of the other great mathematicians of his time Euler's interests were not limited to mathematics. His work is also very important in engineering and physics. A remarkable amount of notation is due to him or popularized by him. Included in this list is the summation symbol Σ , e , π , i , and $f(x)$. Like many of his time, he was a very religious man who believed the Bible was inspired. He had incredible insight but like most of us, he made mistakes because he sometimes neglected issues related to convergence. However, the need for this sort of thing was not well understood in his time. Euler died in St. Petersburg.

and this happens if and only if

$$xMN_x - xM_xN = yNM_y - yMN_y$$

which happens if and only if

$$MxN_x + MyN_y = NyM_y + NxM_x$$

if and only if

$$M(xN_x + yN_y) = N(yM_y + xM_x)$$

But this is true because by Euler's identity, $xN_x + yN_y = \alpha N$ and $yM_y + xM_x = \alpha M$ so the above is just $\alpha NM = \alpha NM$. Of course it is assumed that $xM + yN \neq 0$ in the above.

Example 29.6.3 Find the integral curves for

$$(x^2 + xy) dx + (y^2 + x^2) dy = 0$$

Of course this can be written as a homogeneous equation and the technique for solving these can be used. However, let's use this new technique which says that an integrating factor is

$$\frac{1}{x(x^2 + xy) + y(y^2 + x^2)} = \frac{1}{x^3 + 2x^2y + y^3}$$

Then multiplying by this yields an exact equation.

$$\frac{x^2 + xy}{x^3 + 2x^2y + y^3} dx + \frac{y^2 + x^2}{x^3 + 2x^2y + y^3} dy = 0$$

Unfortunately, it is too complicated for me to solve this conveniently. However, knowing that it is exact allows the use of the formula derived in showing that if $M_y = N_x$ then the equation was exact. Thus the integral curves are of the form

$$\begin{aligned} & \int_0^x M(t, y) dt + N(0, y) \\ &= \int_0^x \frac{t^2 + ty}{t^3 + 2t^2y + y^3} dt + \frac{1}{y} = C \end{aligned}$$

Now we consider an easier one.

Example 29.6.4 Find the integral curves for

$$(xy + y^2) dx + x^2 dy = 0$$

The integrating factor is

$$\frac{1}{xy(2x + y)}$$

and so the equation to solve is

$$\frac{1}{x(2x + y)} (x + y) dx + \frac{x}{y(2x + y)} dy = 0$$

Then integrating the first term with respect to x , the scalar potential is of the form

$$f(x, y) = \ln|x| - \frac{1}{2} \ln \left| x + \frac{1}{2}y \right| + g(y)$$

Then differentiating with respect to y ,

$$-\frac{1}{2(2x+y)} + g'(y) = \frac{x}{y(2x+y)}$$

$$g'(y) = \frac{1}{2y}$$

and so $g(y) = \frac{1}{2} \ln|y|$ will work. Thus the integral curves are of the form

$$\ln|x| - \frac{1}{2} \ln \left| x + \frac{1}{2}y \right| + \frac{1}{2} \ln|y| = C$$

You could simplify this if desired.

PROCEDURE 29.6.5 *To solve*

$$M(x, y) dx + N(x, y) dy = 0$$

using an integrating factor, do the following:

1. Look for an integrating factor μ which is a function of x alone. You do this if

$$\frac{M_y - N_x}{N}$$

does not depend on y . In this case, you solve

$$\mu'(x) = \mu(x) \left(\frac{M_y - N_x}{N} \right)$$

which is a separable equation. Solve and choose constant to satisfy initial condition. If this doesn't work,

2. Look for an integrating factor μ which is a function of y alone. You do this if

$$\frac{N_x - M_y}{M}$$

does not depend on x . In this case, you solve

$$\mu'(y) = \frac{N_x - M_y}{M} \mu(y)$$

which is a separable equation. Solve and choose constant to satisfy initial condition.

3. If neither of these work, check to see if M, N are both homogeneous of the same degree. If they are, you could use either the methods of homogeneous equations or Euler's formula for the integrating factor

$$\frac{1}{xM + yN}.$$

4. If none of the above works, give up. You don't know how to do it. The integrating factor exists, but you don't know how to find it.

29.7 The Case Where M, N Are Affine Linear

Something which often occurs is an equation of the form

$$(px + qy + r)dx + (\alpha x + \beta y + \gamma)dy = 0$$

It doesn't quite fit anything in the earlier discussion. It won't be exact, homogeneous, or separable or linear. However, one can massage it to get something which is homogeneous. This is illustrated in some examples.

Example 29.7.1 Find the integral curves for

$$(x + 2y + 3)dx + (2x - y + 1)dy = 0$$

Of course the problem is those constants 3, 1 so it is reasonable to change variables. Let $u = x - a, v = y - b$ where we choose a, b in an auspicious manner to get the constants to disappear. First, $dx = du, dy = dv$. Then in terms of the new variables,

$$\begin{aligned}(u + a + 2(v + b) + 3)dx + (2(u + a) - (v + b) + 1)dy &= 0 \\ (u + 2v + (a + 2b + 3))dx + (2u - v + (2a - b + 1))dy &= 0\end{aligned}$$

and we want

$$\begin{aligned}a + 2b + 3 &= 0 \\ 2a - b + 1 &= 0\end{aligned}$$

Hence we should let $a = -1$ and $b = -1$. Then with this, the equations reduce to

$$(u + 2v)du + (2u - v)dv = 0$$

This is now a homogeneous equation, or we could use the integrating factor described earlier, but, in this case, it is also an exact equation. A scalar potential is

$$\frac{u^2}{2} + 2uv - \frac{v^2}{2},$$

and so the integral curves for the original equation would be

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

The example illustrates what to do in general. You just change the variables to remove those constant terms and then obtain a homogeneous equation which can be solved by a variety of methods.

Example 29.7.2 Find the integral curves for

$$(x + y + 2)dx + (2x - y + 4)dy = 0$$

As before, let $u = x - a, v = y - b$ and write in terms of these new variables. Thus

$$(u + v + (a + b + 2))du + (2u - v + (2a - b + 4))dv = 0$$

Then you need

$$\begin{aligned}a + b + 2 &= 0 \\ 2a - b + 4 &= 0\end{aligned}$$

Thus $a = -2, b = 0$. The new equation then is

$$(u + v) du + (2u - v) dv = 0$$

This can be considered as a homogeneous equation.

$$\frac{dv}{du} = \frac{(u + v)}{v - 2u} = \frac{1 + \frac{v}{u}}{\frac{v}{u} - 2}$$

Then let $z = \frac{v}{u}$ and do the usual substitution. This yields

$$u \frac{dz}{du} = \frac{1}{z - 2} (-z^2 + 3z + 1) \quad (29.7)$$

and so, separating the variables,

$$\frac{2 - z}{z^2 - 3z - 1} dz = \frac{du}{u}$$

Then after much work one obtains integral curves of the form

$$\begin{aligned}& \ln \left(\frac{v}{u} - \frac{1}{2} \sqrt{13} - \frac{3}{2} \right)^{\frac{1}{26} \sqrt{13}} + \ln \frac{1}{\left(\frac{v}{u} + \frac{1}{2} \sqrt{13} - \frac{3}{2} \right)^{\frac{1}{26} \sqrt{13}}} \\ & + \ln \frac{1}{\sqrt{\frac{v}{u} - \frac{1}{2} \sqrt{13} - \frac{3}{2}}} + \ln \frac{1}{\sqrt{\frac{v}{u} + \frac{1}{2} \sqrt{13} - \frac{3}{2}}} - \ln |u| = C\end{aligned}$$

Then you plug in what u, v are in terms of x, y .

Actually, it was real easy to do this. The computer algebra system did it for me. Here is one which is not so ugly.

Example 29.7.3 Find the integral curve which contains the given ordered pair.

$$(6x - y - 4) dx = (y - 2x) dy, (2, 2)$$

The equation is

$$\frac{dy}{dx} = \frac{6x - y - 4}{y - 2x}$$

Now let $x = u + a, y = v + b$. Then we choose a, b such that in terms of the new variables the equation becomes homogeneous. Thus we need

$$\begin{aligned}6a - b - 4 &= 0 \\ b - 2a &= 0\end{aligned}$$

Thus we let $a = 1, b = 2$. Then the equation is

$$\frac{dv}{du} = \frac{6u - v}{v - 2u} \quad (29.8)$$

This is a homogeneous equation. Let $z = \frac{v}{u}$. Then

$$uz' = \frac{6-z}{z-2} - z = \frac{1}{z-2} (-z^2 + z + 6)$$

Separating the variables,

$$\frac{(2-z)dz}{z^2 - z - 6} = \frac{du}{u}$$

This is easily solved,

$$C - \left(\frac{4}{5} \ln |z+2| + \frac{1}{5} \ln |z-3| \right) = \ln |u|$$

The in terms of the original variables,

$$C = \left(\frac{4}{5} \ln \left| \frac{y-2}{x-1} + 2 \right| + \frac{1}{5} \ln \left| \frac{y-2}{x-1} - 3 \right| \right) + \ln |x-1|$$

Then to contain the ordered pair, you need

$$C = \frac{4}{5} \ln 2 + \frac{1}{5} \ln 3 = \frac{1}{5} \ln(48)$$

PROCEDURE 29.7.4 To solve affine linear equations of the form

$$(px + qy + r)dx + (\alpha x + \beta y + \gamma)dy = 0,$$

do the following:

1. Change the variables $u = x - a$, $v = y - b$, plug in and choose a, b to make the resulting equation homogeneous.
2. Solve the resulting homogeneous equation. Then substitute back in $x - a$ for u and $y - b$ for v . Pick the constant to satisfy initial conditions.

29.8 Linear and Nonlinear Differential Equations

Recall initial value problems for linear differential equations are those of the form

$$y' + p(t)y = q(t), \quad y(t_0) = y_0 \quad (29.9)$$

where $p(t)$ and $q(t)$ are continuous functions of t . Then if $t_0 \in [a, b]$, an interval, there exists a unique solution to the initial value problem given above which is defined for all $t \in [a, b]$. The following theorem which is really something of a review gives a proof.

Theorem 29.8.1 Let $[a, b]$ be an interval containing t_0 and let $p(t)$ and $q(t)$ be continuous functions defined on $[a, b]$. Then there exists a unique solution to 29.9 valid for all $t \in [a, b]$.

Proof: Let $P'(t) = p(t)$, $P(t_0) = 0$. For example, let $P(t) \equiv \int_{t_0}^t p(s) ds$. Then multiply both sides of the differential equation by $\exp(P(t))$. This yields

$$(y(t) \exp(P(t)))' = q(t) \exp(P(t))$$

and so, integrating both sides from t_0 to t ,

$$y(t) \exp(P(t)) - y_0 = \int_{t_0}^t q(s) \exp(P(s)) ds$$

and so

$$y(t) = \exp(-P(t)) y_0 + \exp(-P(t)) \int_{t_0}^t q(s) \exp(P(s)) ds$$

which shows that if there is a solution to 29.9, then the above formula gives that solution. Thus there is at most one solution. Also, you see the above formula makes perfect sense on the whole interval. Since the steps are reversible, this shows $y(t)$ given in the above formula is a solution. You should provide the details. Use the fundamental theorem of calculus. ■

It is not so simple for a nonlinear initial value problem of the form

$$y' = f(t, y), \quad y(t_0) = y_0.$$

Theorem 29.8.2 Let f and $\frac{\partial f}{\partial y}$ be continuous in some rectangle, $a < t < b, c < y < d$ containing the point (t_0, y_0) . Then there exists a unique local solution to the initial value problem

$$y' = f(t, y), \quad y(t_0) = y_0.$$

This means there exists an interval, I such that $t_0 \in I \subseteq (a, b)$ and a unique function, y defined on this interval which solves the above initial value problem on that interval.

A much more general theorem will be proved later. Also, in the above, it suffices to say that f is continuous on the given rectangle and that for $y, z \in [c, d], t \in [a, b]$,

$$|f(t, y) - f(t, z)| \leq K |y - z|$$

for some $K > 0$. This is called a Lipschitz condition. For now, note that it is reasonable to believe the conclusion of this theorem. Start with the point (t_0, y_0) and follow the slope field as illustrated in many of the above examples. The problem is, sometimes you can't extend the solution as far as you might like.

Example 29.8.3 Solve $y' = 1 + y^2, y(0) = 0$.

This satisfies the conditions of Theorem 29.8.2. Therefore, there is a unique solution to the above initial value problem defined on some interval containing 0. However, in this case, we can solve the initial value problem and determine exactly what happens. The equation is separable.

$$\frac{dy}{1+y^2} = dt$$

and so $\arctan(y) = t + C$. Then from the initial condition, $C = 0$. Therefore, the solution to the equation is $y = \tan(t)$. Of course this function is defined on the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$. It is impossible to extend it further because it has an asymptote at the two ends of this interval.

Theorem 29.8.2 does not say that the local solution can never be extended beyond some small interval. Sometimes it can. It depends very much on the nonlinear equation. For example, the initial value problem

$$y' = 1 + y^2 - \epsilon y^3, \quad y(0) = y_0$$

turns out to have a solution on \mathbb{R} . Here ε is a small positive number. You might think about why this is so. It is related to the fact that in this new equation, the extra term prevents y' from becoming unbounded.

Also, you don't know whether the interval of existence is symmetric about the point at which the initial condition is given.

Example 29.8.4 Solve $y' = (y - 1)^2$, $y(0) = 0$.

The equation is separable

$$\frac{dy}{(y-1)^2} = dt$$

Then integrating and using the initial condition,

$$\frac{1}{1-y} = t + 1$$

Thus

$$y = \frac{t}{t+1}$$

which makes sense on $(-1, \infty)$.

The next one looks a lot like the above, but has a solution on the whole real line.

Example 29.8.5 Consider $y' = 1 - y^2$, $y(0) = 0$.

You can verify that

$$y = \frac{e^{2t} - 1}{e^{2t} + 1}$$

is the solution and it makes sense for all t .

Hopefully, this has demonstrated that all sorts of things can happen when you are considering nonlinear equations. However, it gets even worse.

If you assume less on f in the above theorem, you sometimes can get existence but not uniqueness for the initial value problem. In the next example $\frac{\partial f}{\partial y}$ is not continuous near $(0, 0)$.

Example 29.8.6 Find the solutions to the initial value problem

$$y' = y^{1/3}, \quad y(0) = 0.$$

The equation is separable so $\frac{dy}{y^{1/3}} = dt$ and so the solutions are of the form

$$\frac{3}{2}y^{2/3} = t + C.$$

Letting $C = 0$ from the initial condition, one solution is $y = (\frac{2}{3}t)^{3/2}$ for $t > 0$. However, you can see that $y = 0$ is also a solution. Thus uniqueness is violated. Note there are two solutions to the initial value problem and both exist and solve the initial value problem on all of $[0, \infty)$.

Observation 29.8.7 *What are the main differences between linear and nonlinear equations? Linear initial value problems have an interval of existence which is the same as the interval on which the functions in the equation are continuous. Nonlinear initial value problems sometimes don't. Solutions to linear initial value problems are unique. This is not always true for nonlinear equations although if in the nonlinear equation, f and $\partial f/\partial y$ are both continuous, then you at least get uniqueness as well as existence on some possibly small interval of undetermined length.*

29.9 Computer Algebra Methods

The above methods work very well except for when they don't, which is the typical case. One can use computer algebra systems to solve such equations. In this section, the use of various systems will be discussed. The intent here is to give a reasonably simple way to obtain these solutions, not to give all possible ways to use these systems. In this book, I will be emphasizing MATLAB. However, other systems will be discussed in this section. One very easy to use system which behaves a lot like MATLAB is Scientific Notebook, which is actually based on mupad. I will mention its use also.

29.9.1 MATLAB

A frequently used computer algebra system is MATLAB. You can use this to find solutions to the initial value problem. If you want commands to appear on separate lines, you use "shift enter".

The basic version of MATLAB is sufficient to do the numerical procedures discussed. In order to do procedures which involve commands like "syms" you will need to have the symbolic math toolbox also. In particular, you need this toolbox for the first example given here in which "dsolve" is used, but not for the numerical procedures mentioned next.

Here is what you type to get MATLAB to compute the solution to

$$y' = y - .01y^2, y(0) = 2.$$

After the >> you type the following:

```
syms y(t); y(t)=dsolve(diff(y,t)==y - .01*y^2, y(0)==2)
```

After typing in the above, you press enter and here is what results.

```
>>syms y(t); y(t)=dsolve(diff(y,t)==y-.01*y^2,y(0)==2)
y(t) =
100/(exp(log(49) - t)+1)
```

If you want a graph of this solution, this is also easy to get. After doing the above, type in the following to the right of >>

```
ezplot(y(t),[0,3])
```

and then press "enter" to obtain the graph of the solution on the interval $[0, 3]$.

Similarly, you can ask for numerical solutions in case you can't find an analytical solution. MATLAB can find these also. For example, if you wanted to solve on the interval $[0, 2]$ the initial value problem

$$y' = y - .01y^5, \quad y(0) = 1,$$

You would do the following: After `>>` you type

```
f=@(x,y) y-.01*y^5;
```

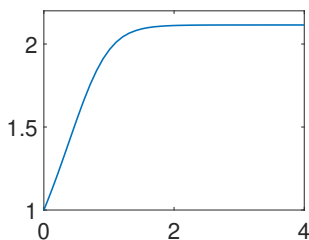
Next, type the following on a new line:

```
[x,y]=ode45(f,[0,2],1) (*)
```

and on the next new line,

```
plot(x,y)
```

and press "enter". This will give a large table of values of x followed by values of y which comes from using a suitable numerical method named `ode45` and it will also plot the solution.



If you don't want to see this large table of values, simply place a `;` at the end of `*`. This will cause MATLAB to defer displaying the table even though it knows about it.

If you placed `;` at the end of `*`, and decide you would like to see $y(.5)$ for example, you ask for the table of values. This is done by typing `[x,y]` after `>>` and then "enter" to see the whole table and simply scroll down to find an entry in the column for x which is close to $.5$. There is also another way to find the values using the `deval` function.

Another thing which is pretty easy to do in MATLAB is to change the initial conditions and graph the two solutions on the same set of axes. The above gives you a graph of $y(x)$ for $x \in [0, 2]$. It has defined the function y at least at many points. Now you can simply define another solution with a different initial condition as follows.

```
[x1,y1]=ode45(f,[0,2],2)
```

and press return. This will define the function $x1 \rightarrow y1(x1)$. Then to graph both on the same axes, you would type

```
plot(x,y,x1,y1)
```

and both will appear. You can do as many of these as you want of course. If you wanted to do a lot of graphs all at once, you can also have this done. You would do the following:

```
>> f=@(t,x) [x-x^3];
hold on
for z=-2:.5:2
[t,x]=ode45(f,[0,4],z);
plot(t,x)
end
```

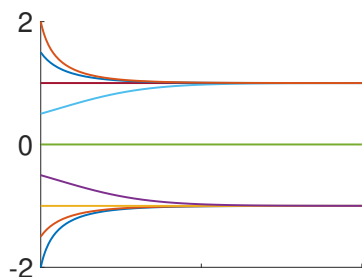

Then press “enter” and you will get graphs of solutions for initial conditions

$$-2, -1.5, -1, -.5, \dots, 2$$

With the above, which is solving

$$x' = x - x^3, \quad x(0) = z$$

for various values of z , you get the following graph.



Note how this illustrates that there are three equilibrium points $-1, 0, 1$ and that the first and third are stable but 0 is not.

You can also do the following. After defining a function, say `h=@(t,y) [y-y^3]`, you do the following:

```
sol=ode45(h,[0,7],3);
deval(sol,[1,2,3,4,5])
```

then press “enter”. You should get the values of y at the points 1, 2, 3, 4, 5. Remember that to place on a new line, you use “shift enter”. You could also use any other symbol for “sol”.

Another thing I have noticed when using MATLAB is that it sometimes puts the graph behind the command window so you don’t see it till you shrink the command window.

To adjust the appearance of the graph which results, you go to the graph and click on file and then export setup. You can make changes in a dialog box and do things like change the thickness of the lines and the size of the font very easily. Then you can save it as an eps file or several other kinds of files.

Also, when you are done, type `>> clear all` or `close all` and then “enter”. Then type `>> clf` and “enter” to get rid of any graphs it may have done and press “enter”. To clear the screen, type `>> clc` and then press “enter”. This is a very good idea because if you want to do something else, you don’t want MATLAB to be confused about what you mean and it will be confused if it can.

29.10 Exercises

Linear Equations

- Find all solutions to the following linear equations. You may need to leave answers in terms of integrals on some of them.

(a) $y' + 2ty = e^{-t^2}$

(e) $y' + \frac{1}{t-1}y = \frac{1}{(t-1)^2}$

(b) $y' - ty = e^t$

(f) $y' + \tan(t)y = \cos(t)$

(c) $y' + \cos(t)y = \cos(t)$

(g) $y' - \tan(t)y = \sec(t)$

(d) $y' + ty = \sin(t)$

(h) $y' - \tan(t)y = \sec^2(t)$

- In the above linear equations find the solution to the initial value problems when $y(0)$ equals the following numbers.

- (a) 1 (e) -2
 (b) 2 (f) 12
 (c) 3 (g) -3
 (d) 4 (h) -2
3. Solve the following initial value problem. $ty' - y = \frac{1}{t^2}$, $y(1) = 2$. Would it make any sense to give the initial condition at $t = 0$?
4. Solve the following initial value problem. $ty' + y = \frac{1}{t}$, $y(-1) = 2$. Would it make any sense to give the initial condition at $t = 0$? **Hint:** You need to remember that $\int \frac{1}{t} dt = \ln|t| + C$.
5. Solve the following initial value problems.
- (a) $\ln(t)y' + \frac{1}{t}y = \ln(t)$, $y(2) = 3$.
 (b) $\ln(t)y' - \frac{1}{t}y = \ln^2(t)$, $y(2) = 3$.
 (c) $y' + \tan(t)y = \cos^3(t)$, $y(0) = 4$.
 (d) $\cosh(t)y' + \sinh(t)y = \sinh(t)$, $y(0) = -4$
6. You have the equation $y' + p(t)y = q(t)$ where $P'(t) = p(t)$. Give a formula for all solutions to this differential equation.
7. The height of an object at time t is $y(t)$. It falls from an airplane at 30,000 feet which is traveling East at 500 miles per hour and is acted on by gravity which we will assume has acceleration equal to 32 feet per second squared and air resistance which we will suppose yields an acceleration equal to .1 times the speed of the falling object opposite to the direction of motion. If its initial velocity is in the direction of motion of the airplane, find a formula for the position of the object as a function of t in feet.
8. Solve the following differential equations. Give the general solution.
- (a) $(x^3 + y) dx - xdy = 0$
 (b) $ydx + (x - y) dy = 0$ **Hint:** You might look for x as a function of y .
 (c) $y(y^2 - x) dy = dx$
 (d) $2ydx = (x^2 - 1)(dx - dy)$
 (e) $L \frac{di}{dt} + Ri = E \sin(\omega t)$. Here L, R, E are positive constants. L symbolizes inductance and R resistance while i is the current.
9. For compounding interest n times in one year which has interest rate r per year, the amount after t years is given by $A_0 \left(1 + \frac{r}{n}\right)^{tn}$. Show that
- $$\lim_{n \rightarrow \infty} \left(1 + \frac{r}{n}\right)^{tn} = e^{rt},$$
- thus giving the same conclusion as mentioned in the chapter.
10. Consider the equation $y' + 2ty = t$, $y(0) = 32.76$. Find $\lim_{t \rightarrow \infty} y(t)$.

11. Although the gas supply was shut off, the air in the building continued to circulate. When the gas was shut off, the temperature in the building was 70 and after five hours, the temperature had fallen to a chilly 50 degrees. If the outside temperature was at 10 degrees, what is the constant in Newton's law of cooling?
12. A radioactive substance decays according to how much is present. Thus the equation is $A' = -kA$. If after 40 years, there is $5/6$ of the amount initially there still present, what is the half life of this substance?
13. You have the following initial value problem $y' + y = \sin t$, $y(0) = y_0$. Letting y be the solution to this initial value problem, find a function $u(t)$ which does not depend on y_0 and $\lim_{t \rightarrow \infty} |y(t) - u(t)| = 0$.
14. A pond which holds V cubic meters is being polluted at the rate of $10 + \sin(2\pi t)$ kg per year. The periodic source represents seasonal variability. The total volume of the lake is constant because it loses $\frac{1}{4}V$ cubic meters per year and gains the same. After a long time, what is the average amount of pollutant in this lake in a year?

Bernouli Equations

15. Solve the following initial value problems involving Bernouli equations.

(a) $y' + 2xy = xy^3$, $y(1) = 2$	(d) $y' - 2x^3y = x^3y^{-1}$, $y(1) = 1$
(b) $y' + \sin(t)y = \sin(t)y^2$, $y(1) = 1$	(e) $y' + y = x^2y^{-2}$, $y(1) = 1$
(c) $y' + 2y = x^2y^3$, $y(1) = -1$	(f) $y' + x^3y = x^3y^{-2}$, $y(1) = -1$
16. Consider $y' = py - qy^2$, $y(0) = \frac{p}{mq}$ where p, q are positive and $m > 1$. Solve this Bernouli equation and also find $\lim_{t \rightarrow \infty} y(t)$.
17. Consider $y' = 3y - y^3$, $y(0) = 1$. Solve this Bernouli equation and find $\lim_{t \rightarrow \infty} y(t)$.
18. Find the solution to the Bernouli equation $y' = (\cos t + 1)y - y^3$, $y(0) = 1$. **Hint:** You may have to leave the solution in terms of an integral.
19. Actually the drag force of a small object moving through the air is proportional not to the speed but to the square of the speed. Thus a falling object would satisfy the following equation for downward velocity. $v' = g - kv^2$. Here g is acceleration of gravity in whatever units are desired. Find $\lim_{t \rightarrow \infty} v(t)$ in terms of g, k . **Hint:** Look at the equation.
20. A Riccati equation is like a Bernouli equation except you have an extra function added in. These are of the form $y' = a(t) + b(t)y + c(t)y^2$. If you have a solution, y_1 , show that $y(t) = y_1(t) + \frac{1}{v(t)}$ will be another solution provided v satisfies a suitable first order linear equation. Thus the set of all such y will involve a constant of integration and so can be regarded as a general solution to the Riccati equation. These equations result in a very natural way when you consider $y' = f(t, y)$ and approximate $f(t, y)$ by fixing t and approximating the resulting function of y with a second order Taylor polynomial.

Separable Equations

21. Solve the following initial value problems involving separable equations. The ordered pair given is to be included in the solution curve.

- (a) $x^2 dx + (y^2 + 1) dy = 0$, $(1, 1)$ (e) $0 = \cos(y) dx + \tan(x) dy$, $(\frac{\pi}{2}, \frac{\pi}{4})$
 (b) $xy dx + (y^2 + 1) dy = 0$, $(1, 1)$ (f) $xy dx = (y^2 + 1) dy$, $(1, 1)$
 (c) $xy dx + (y^2 + 1) dy = 0$, $(1, -1)$ (g) $xy dx = (y^2 - 1) dy$, $(2, 1)$
 (d) $y dx + (y^2 + 1) x dy = 0$, $(1, 2)$
22. Find all integral curves of the equation $yx dx + e^{-x^2} dy = 0$. Graph several.
23. Find all integral curves of the equation $yx dx + \frac{1}{\ln(1+x^2)} y^3 dy = 0$. Graph several.
24. Give the integral curves to the equation $v' = g - kv^2$ mentioned above where g is acceleration of gravity and k a positive constant.
25. You have a collection of hyperbolas $x^2 - y^2 = C$ where each choice of C leads to a different hyperbola. Find another collection of curves which intersect these at a right angle. **Hint:** Say you have $f(x, y) = C$ is one of these. If you are at a point where the relation defines y as a function of x , and (x, y) is a point on one of these hyperbolas just mentioned, then $\frac{dy}{dx}$ should have a relation to the tangent line to $x^2 - y^2 = C$. Since the two curves are to be perpendicular, you should have the product of their slopes equal to -1 . Thus $\left(\frac{dy}{dx}\right)\left(\frac{x}{y}\right) = -1$.
26. Generalize the above problem. Suppose you have a family of level curves $f(x, y) = C$ and you want another family of curves which is perpendicular to this family of curves at every point of intersection. Find a differential equation which will express this condition. Recall that two curves are perpendicular if the products of the slopes of the tangent lines to the two curves equals -1 .
27. Find and determine the stability of the equilibrium points for the following separable equations.
- (a) $y' = y^2(y - 1)$ (e) $y' = \ln(1 + y^2)$
 (b) $y' = (y + 1)(y - 1)(y + 2)$ (f) $y' = e^{2y} - 1$
 (c) $y' = \sin(y)$ (g) $y' = 1 - e^{y^2}$
 (d) $y' = \cos(y)$
28. The force on an object of mass m acted on by the earth having mass M is given by Newton's formula kmM/r^2 where k is the gravitation constant first calculated by Cavendish⁴ in 1798. Letting R be the radius of the earth and letting g denote the acceleration of gravity on the earth's surface, show that $kM = R^2g$. Now suppose a large gun having its muzzle at the surface of the earth is fired away from the center of the earth such that the projectile has velocity v_0 . Explain why

$$\frac{dv}{dt} = -\frac{R^2g}{(R+r)^2}$$

⁴For about 100 years, since the time Newton claimed the existence of this gravitation constant, no one knew what it was. Henry Cavendish did an extremely sensitive experiment in 1797-1798 to determine it. It involved lead balls mirrors telescopes and a torsion balance. He was a chemist who also found ways to make hydrogen. He did many other very precise experiments in physics and chemistry.

where r is the distance to the surface of the earth and here $v = v(t)$ the speed of the projectile at time t when it is at a distance of r from the surface of the earth. Next explain why

$$v \frac{dv}{dr} = -\frac{R^2 g}{(R+r)^2}$$

The two variables are v and r . Separate the variables and find the solution to this differential equation given that the initial speed is v_0 as stated above. Show that the maximum distance from the surface of the earth is given by

$$R \left(\frac{Rg}{Rg - \frac{1}{2}v_0^2} - 1 \right)$$

provided that $Rg > \frac{1}{2}v_0^2$. What is the smallest value of v_0 such that the projectile will leave the earth and never return?

29. The Gompertz equation is $\frac{dy}{dt} = ry \ln\left(\frac{K}{y}\right)$. Find the solutions to this equation with initial condition $y(0) = y_0$. Also identify all equilibrium solutions and their stability. Also verify the inequality $ry \ln\left(\frac{K}{y}\right) \geq ry\left(1 - \frac{y}{K}\right)$ for $y \in [0, K]$. Explain why for a given initial condition $y_0 \in (0, K)$, the solution to the Gompertz equation should be at least as large as the solution to the logistic equation.
30. You have a population which satisfies the logistic equation $y' = ry\left(1 - \frac{y}{K}\right)$ and the initial condition is $y(0) = \alpha K$ where $0 < \alpha < 1/2$. How long will it take for the population to double?
31. An equilibrium point is called semi-stable if it is stable from one side and not stable from the other. Sketch the appearance of $f(y)$ near y_0 if y_0 is a semi-stable equilibrium point. Here $f(y_0) = 0$ and the differential equation is $y' = f(y)$.
32. Consider the differential equation $y' = a - y^2$ where a is a real number. Show that there are no equilibrium solutions if $a < 0$ but there are two of them if $a > 0$ and only one if $a = 0$. Discuss the stability of the two equilibrium points when $a > 0$. What about stability of equilibrium when $a = 0$?
33. Do exactly the same problem when $y' = ay - y^3$. This time show there are three equilibrium points when $a > 0$ and only one if $a < 0$. Discuss the stability of these points.
34. Do the same problem if $y' = ay - y^2$. These three problems illustrate something called bifurcation which is when the nature of the solutions changes dramatically when some parameter changes.

Homogeneous Equations

35. Find the solution curve to the following differential equations which contains the given point.
 - (a) $y' = \frac{1}{x(2x+y)}(x+y)^2$, $(1, 1)$
 - (b) $y' = -\frac{1}{x(x-2y)}(x^2 - xy + 2y^2)$, $(2, 0)$

- (c) $y' = \frac{1}{4x^2+yx} (x^2 + 4xy + y^2), (-1, 1)$
 (d) $y' = -\frac{1}{3x^2-xy} (x^2 - 3xy + y^2), (1, 1)$
 (e) $y' = \frac{1}{x(y+5x)} (x^2 + 5xy + y^2), (-1, -1)$
 (f) $y' = \frac{1}{x(3y+2x)} (x^2 + 2xy + 3y^2), (-2, 3)$
 (g) $y' = \frac{1}{x(4y-x)} (x^2 - xy + 4y^2), (3, -2)$

36. Find the solution curve to the following ODEs which contains the given point.

- (a) $y' = \frac{1}{x^2} (x^2 + y^2 + xy), (1, 1)$
 (b) $y' = \frac{1}{x^2} (4x^2 + y^2 + xy), (2, 0)$
 (c) $y' = \frac{1}{x^2} (x^2 + 9y^2 + xy), (3, 1)$
 (d) $y' = \frac{1}{x^2} (4x^2 + 2y^2 + xy), (-1, 1)$

37. Find the solution curve to the following ODEs which contains the given point.

- (a) $-(x+y)dx + (x+2y)dy = 0, (1, 1)$
 (b) $(x-y)dx + (x+3y)dy = 0, (2, 1)$
 (c) $(4x+y)dx + (x+2y)dy = 0, (-1, 2)$
 (d) $-(3x+y)dx + (x-y)dy = 0, (3, 2)$
 (e) $(3x-4y)dx + (4x-\frac{4}{3}y)dy = 0, (3, 1)$
 (f) $(-y)dx + (4y-x)dy = 0, (0, 2)$
 (g) $(-2x-\frac{31}{4}y)dx + (x-\frac{9}{4}y)dy = 0, (-1, 2).$

38. Find all solutions to $y' + \sin\left(\frac{y}{x}\right) = 1$. **Hint:** You might need to leave the answer in terms of integrals.

39. Solve: $x^2dy + (4x^2 - xy + 5y^2)dx = 0, y(3) = -1$.

40. Solve: $x^2dy + (7x^2 - xy + 4y^2)dx = 0, y(2) = -1$.

41. Solve: $x^2dy + (6x^2 - xy + 3y^2)dx = 0, y(-1) = 1$.

42. Solve: $(x^3 - 7x^2y - 5y^3)dx + (7x^3 + 5xy^2)dy = 0, y(3) = -2$.

Exact Equations and Integrating Factor

43. Find the solution curve to the following ODEs which contain the given point. First verify that the equation is exact.

- (a) $(2xy + 1)dx + x^2dy = 0, (1, 1)$
 (b) $(2x \sin y + 1)dx + (x^2 \cos y)dy = 0, (1, \frac{\pi}{2})$
 (c) $(2x \sin y - \sin x)dx + ((\cos y)x^2 + 1)dy = 0, (0, 0)$
 (d) $\left(\frac{y}{xy+1}\right)dx + \frac{1}{xy+1}(x+xy+1)dy = 0, (1, 1)$
 (e) $(y^2 \cos xy^2 + 1)dx + (2xy \cos xy^2 + 1)dy = 0, (1, 0)$

$$(f) \left(y(\tan^2 xy + 1) + y \cos xy \right) dx + \left(x(\tan^2 xy + 1) + x \cos xy + 1 \right) dy = 0, \\ (0, 1)$$

44. Find the solution curve to the following ODEs which contains the given point.

- (a) $(2y^3 + 2) dx + (3xy^2) dy = 0, (1, 1)$
- (b) $(2y^3 + 2y + 2 \cos(x^2)) dx + (3xy^2 + x) dy = 0, (1, 1)$
- (c) $(2xy^2 + y + 2xy \cos x^2) dx + (2 \sin x^2 + 3x^2 y + 2x) dy = 0, (2, 1)$
- (d) $3y^4 dx + \left(4xy^3 + \frac{5y^4}{x^2} \right) dy = 0, (1, 2)$
- (e) $(5x^4 y + 4x^3 y^3) dx + (3x^5 + 5x^4 y^2) dy = 0, (1, 1)$
- (f) $(8x^4 y^6 + 3x^3) dx + (12x^5 y^5 + 3xy^2) dy = 0, (-1, 2)$

45. Explain why every separable ODE can be considered as an exact ODE.

46. Suppose you have a family of level curves $f(x, y) = C$ where C is a constant. Also suppose that f is a harmonic function. That is $f_{xx} + f_{yy} = 0$. Consider the problem of finding another family of level curves such that each of these is perpendicular to the original level curves $f(x, y) = C$ at any point on both of them. Show that the appropriate equation to solve is $0 = f_y dx - f_x dy$. Verify that this is an exact equation. Thus there exists $g(x, y)$ such that the solutions are $g(x, y) = C$.

M, N Both Affine Linear

47. Find the integral curve for the following differential equation which contains the given point. These are also exact so you could use either method.

- (a) $(2x + y - 3) dx + (x + y - 3) dy = 0, (1, 6)$
- (b) $(y - x + 2) dx + ((x - y) - 2) dy = 0, (3, 2)$
- (c) $(x + y - 3) dx + (x + 3y - 7) dy = 0, (2, 2)$
- (d) $(2x + y - 8) dx + (x + y - 7) dy = 0, (-2, 1)$
- (e) $(x + y - 2) dx + (x + 3y - 4) dy = 0, (4, 1)$
- (f) $(y - 2x + 5) dx + (x + y + 2) dy = 0, (1, 1)$
- (g) $(y - 4x + 3) dx + (x - 5y + 4) dy = 0, (2, 1)$

48. Find the integral curves for the following differential equation.

- (a) $(2y - x) dx = (4x + y - 9) dy$
- (b) $(5x + 4y - 13) dx = (8x + y - 10) dy$
- (c) $(3x - 2y + 1) dx = (y - 4x - 3) dy$
- (d) $(4y - 4x + 4) dx = (8x + y + 11) dy$
- (e) $(2y - x - 3) dx = (4x + y + 21) dy$
- (f) $(5y - 6x + 23) dx = (10x + y - 29) dy$

An Assortment of Exercises

49. Solve: $y' + 3 \cos(t)y = 4(\cos t)e^{-3 \sin t}$, $y(0) = 1$.
50. Solve: $y' + \tan(t)y = \cos(t)$, $y(0) = -2$.
51. Solve: $x^2 dy + (4x^2 - xy + 3y^2) dx = 0$, $y(2) = -2$.
52. Solve: $(\frac{7}{2}y - 2x) dx + (x - \frac{9}{4}y) dy = 0$ which contains the point $(x, y) = (1, 2)$.
53. Solve: $x^2 dy + (3x^2 - xy + 2y^2) dx = 0$, $y(2) = -3$.
54. Solve: $(x^3 - 6x^2y - y^3) dx + (6x^3 + xy^2) dy = 0$, $y(2) = -3$. Graph the integral curve.
55. Solve: $(2y - 3x) dx + (2x - \frac{4}{3}y) dy = 0$ which contains the point $(x, y) = (1, 2)$. Graph the integral curve.
56. Solve: $y' + 5 \cos(3t)y = 2e^{-(5/3) \sin 3t} \cos 3t$, $y(0) = 2$.
57. Solve: $x^2 dy + (5x^2 - xy + 5y^2) dx = 0$, $y(-2) = -2$.
58. Solve: $(3x + \frac{19}{4}y) dx + (-4x - \frac{9}{4}y) dy = 0$ which contains the point $(x, y) = (1, 2)$.
59. Solve: $(x^3 - 3x^2y - y^3) dx + (3x^3 + xy^2) dy = 0$, $y(3) = -1$.
60. Solve: $(y) dx + (x + 4y) dy = 0$ which contains the point $(x, y) = (1, 2)$.
61. Solve: $5(t^6)y + y' = -5t^6e^{t^7}$, $y(1) = 1$.
62. Solve: $x^2 dy + (6x^2 - xy + 5y^2) dx = 0$, $y(3) = 3$.
63. Find the solutions to the equation $y' + y(3 \cos t) = 3(\cos t)e^{-3 \sin t}$.
64. Solve: $(y - 2x) dx + (\frac{9}{2}y - x) dy = 0$ which contains the point $(x, y) = (1, 2)$.
65. Solve: $x^2 dy + (2x^2 - xy + y^2) dx = 0$, $y(2) = -1$.
66. Find the solutions to the equation $y' + 2ty = te^{t^2}$.
67. Solve: $(\frac{7}{3}y - 2x) dx + (x - \frac{4}{3}y) dy = 0$ which contains the point $(x, y) = (1, 2)$.
68. Solve: $y' + \tan(2t)y = \cos 2t$, $y(0) = 2$.
69. Find the general solution to the equation

$$y' + (4x^3 + x^2 + 3x)y = \exp\left(-x^4 - \frac{1}{3}x^3 - \frac{3}{2}x^2\right) \ln(x+1)$$

70. Show that the following initial value problem fails to have a unique solution.

$$y' = y^{1/(2n+1)}, y(0) = 0, n \text{ a positive integer.}$$

71. Sometimes you have an equation of the form

$$y'' = f(y, y')$$

and you are looking for a function $t \rightarrow y(t)$ so the independent variable is missing. These can be massaged into a first order equation as follows. Let $v = y'$ and then you have

$$v' = f(y, v)$$

Now $\frac{dv}{dt} = \frac{dv}{dy} \frac{dy}{dt} = \frac{dv}{dy} v$. Thus we have

$$v \frac{dv}{dy} = f(y, v)$$

which is now a first order differential equation. Use this technique to solve the following problems. This won't always work. It is a gimmick which sometimes works.

- (a) $y'' + 2y' = 0, y(0) = 1, y'(0) = 0$
- (b) $y'' = y'(2y + 1), y(0) = 0, y'(0) = 1$
- (c) $y'' = 2yy', y(0) = 0, y'(0) = 1$
- (d) $y'' = y'(1 - 3y^2), y(0) = 1, y'(0) = 0$
- (e) $y'y'' = 2, y(0) = 1, y'(0) = 2$
- (f) $y'' = 2y, y(0) = 1, y'(0) = 2$
- (g) $y'y'' + 3y = 0, y(0) = y'(0) = 1$
- (h) $(1 + 3t^2)y'' + 6ty' - \frac{3}{t^2} = 0, y'(1) = 1, y(1) = 2$. **Hint:** This is not like the above but $\frac{d}{dt}((1 + 3t^2)y')$ gives the first two terms.
- (i) $yy'' + (y')^2 = 0$. Give a general solution involving two constants of integration.
- (j) $y'' + y(y')^2 = 0$. Give a general solution involving two constants of integration.
- (k) $y''y^2 - 2y(y')^2 = 0$, Give a general solution involving two constants of integration.
- (l) $y''y^3 - 3y'y^2 = 0$, Give a general solution involving two constants of integration.
- (m) $3(y')^2 y''y^2 + 2y(y')^4 = 0$, Give a general solution involving two constants of integration.

72. Explain how you would proceed to solve an equation of the form $y'' = f(t, y')$ where the function you are looking for is $t \rightarrow y(t)$. How many independent constants would you have in a general solution?

Computer Algebra Problems

- 73. Give a graph of the solution to the following initial value problem on the interval $[0, 5]$. $y' = -y^3 + 3y^2 + 2, y(0) = 0$.
- 74. Give a graph of the solution to the following initial value problem on the interval $[0, 5]$. $y' = -y^3 + xy^2 + 1, y(0) = 1$.

75. Solve the following initial value problems and give a graph of each on $[0, 3]$ on the same axes. $y' = \frac{1}{10}y(5-y), y(0) = .3, y' = \frac{1}{10}y(5-y), y(0) = .5, y' = \frac{1}{10}y(5-y), y(0) = -.3$.
76. Give a graph of the solutions to the differential equation $y' = ty^2 - (.1)y^3$ on the interval $[0, 5]$ which result from the initial conditions $y(0) = 1, 0, 2, -3$.
77. Give a graph of the solution to $y' = x(y^2)^{3/4} - xy^3 + 1, y(0) = 0$.
78. Use a computer algebra system to obtain a solution to the initial value problem

$$y' = \frac{y^3}{x^3 + 8y^3}, y(0) = 1$$

You may have to obtain a numerical solution in terms of a graph. It is true that the equation is homogeneous, but it might be too hard to carry out the computations. Scientific notebook has trouble with this one.

79. Use a computer algebra system to obtain the graph of the solution to the initial value problem $x^2y' = 4x^2 + xy + y^2, y(4) = 1$.
80. Find the solution to the following initial value problem, either a graph or a formula. Then graph it

$$y' = xy + \sin(x) - \frac{1}{10}y^2, y(0) = 1$$

81. When you use MATLAB or other computer algebra system to find a numerical solution to a differential equation, you are using a fairly sophisticated numerical method. The most primitive method for obtaining numerical solutions to $y' = f(t, y)$ is called Euler's method. In this method, one has a step size h and partitions the time interval into $t_0 < t_1 < \cdots < t_n = T, t_{j+1} = t_j + h$. Then letting y_0 be the initial condition, Euler's method goes like this. You iterate the following process.

$$k = f(t_i, y_i), y_{i+1} = y_i + hk, t_{i+1} = t_i + h$$

When you get to t_n , you stop. Your solution consists of a function y which interpolates the points (t_i, y_i) meaning $y_i = y(t_i)$. You can easily get MATLAB to do this for you. Here is the case of $y' = y, y(0) = 1$.

```
f=@(t,y) y; h=.01; y(1)=1; t(1)=0;
hold on; for j=1:500;
k=f(t(j),y(j)); y(j+1)=y(j)+h*k; t(j+1)=t(j)+h;
end; plot(t,y); [t(501),y(501)]
```

The first line is defining the function $f(t, y) = y$. Thus the real solution is e^t . The number $y(501)$ is the Euler solution at 5. Compare with e^5 .

82. Suppose you have the initial value problem $y' = y, y(0) = y_0$. You know the solution is $e^t y_0$. Consider the interval $[0, t]$. Consider for $k \leq n$

$$y_{k+1} = y_k + \frac{t}{n} y_k$$

Show that this is the same as finding y_1, \dots, y_n where

$$\frac{y_{k+1} - y_k}{t/n} = y_k$$

In place of $y'(s) = y(s)$, you have $\frac{y_{k+1} - y_k}{t/n} = y_k$. Now show that $y_n = \left(1 + \frac{t}{n}\right)^n y_0$. What is the limit as $n \rightarrow \infty$?

83. Suppose on an interval $[a, a+h]$, you have $y'(t) = f(t, y(t))$ and $z(t) = y(a) + (t-a)f(a, y(a))$. Suppose also the solution y has bounded continuous second derivatives. Show that $|y(h) - z(h)| < Ch^2$ for some constant C . You will need to use Taylor's theorem. This is the local error for the Euler method.

Chapter 30

Laplace Transform Methods

30.1 Linear O.D.E. With Constant Coefficients

This method of Laplace¹ transforms succeeds so well because of the algebraic technique of partial fractions and the fact that the Laplace transform is a linear mapping. It works very well to solve higher order initial value problems involving linear equations with constant coefficients and also more generally first order systems. It is all about changing a differential equation into an algebraic equation, solving that one, and then extracting the solution to the original differential equation from what was obtained.

This presentation will emphasize the algebraic procedures. The analytical questions are not trivial and are given a discussion in Section 4.2.

For an initial value problem, you can often reduce to one which has initial condition given at 0 by simply changing the independent variable. Also, this is where the initial condition is typically given anyway so in this method, I will assume all the initial conditions are given at 0.

Definition 30.1.1 Let f be a function defined on $[0, \infty)$ which has *exponential growth*, meaning that

$$|f(t)| \leq Ce^{\lambda t}$$

for some real λ . Then the **Laplace transform** of f , denoted by $\mathcal{L}(f)$ is defined as

$$F(s) \equiv \mathcal{L}f(s) = \int_0^\infty e^{-ts} f(t) dt$$

for all s sufficiently large. It is customary to write this transform as $F(s)$ or $\mathcal{L}f(s)$ and the function as $f(t)$ instead of f . In other words, t is considered a generic variable as is s and you tell the difference by whether it is t or s . It is sloppy but convenient notation.

Lemma 30.1.2 \mathcal{L} is a linear mapping in the sense that if f, g have exponential growth, then for all s large enough and a, b scalars,

$$\mathcal{L}(af(t) + bg(t))(s) = a\mathcal{L}f(s) + b\mathcal{L}g(s)$$

¹Pierre-Simon, marquis de Laplace (1749-1827) had interests in mathematics, physics, probability, and astronomy. He wrote a major book called celestial mechanics. There is also the Laplacian named after him, and Laplace's equation in potential theory. The expansion of a determinant along a row or column is called Laplace expansion. He was also involved in the development of the metric system. It is hard to overstate the importance of his contributions to mathematics and the other subjects which interested him. [14]

Proof: Let f, g be two functions having exponential growth. Then for s large enough,

$$\begin{aligned}\mathcal{L}(af(t) + bg(t)) &\equiv \int_0^\infty e^{-ts} (af(t) + bg(t)) dt \\ &= a \int_0^\infty e^{-ts} f(t) dt + b \int_0^\infty e^{-ts} g(t) dt = a\mathcal{L}f(s) + b\mathcal{L}g(s) \blacksquare\end{aligned}$$

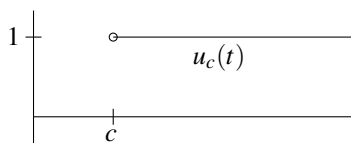
The usefulness of this method in solving differential equations, comes from the following observation.

$$\mathcal{L}(x'(t)) = \int_0^\infty x'(t) e^{-ts} dt = x(t) e^{-st} \Big|_0^\infty + \int_0^\infty s e^{-st} x(t) dt = -x(0) + s\mathcal{L}x(s).$$

In the following table, $\Gamma(p+1)$ denotes the gamma function

$$\Gamma(p+1) = \int_0^\infty e^{-t} t^p dt$$

The function $u_c(t)$ denotes the step function which equals 1 for $t > c$ and 0 for $t < c$.



The expression in Formula 20.) is defined as follows

$$\int \delta(t-c) f(t) dt = f(c)$$

It models an impulse and is sometimes called the Dirac delta function. There is no such function but it is called this anyway. In the following, n will be a positive integer and $f * g(t) \equiv \int_0^t f(t-u) g(u) du$. Also, $F(s)$ will denote $\mathcal{L}\{f(t)\}$ the Laplace transform of the function $t \rightarrow f(t)$.

Table of Laplace Transforms

$f(t)$	$F(s)$	$f(t)$	$F(s)$
1.) 1	$1/s$	12.) $e^{at} \sinh bt$	$\frac{b}{(s-a)^2 - b^2}$
2.) e^{at}	$1/(s-a)$	13.) $e^{at} \cosh bt$	$\frac{s-a}{(s-a)^2 - b^2}$
3.) t^n	$\frac{n!}{s^{n+1}}$	14.) $t^n e^{at}$	$\frac{n!}{(s-a)^{n+1}}$
4.) $t^p, p > -1$	$\frac{\Gamma(p+1)}{s^{p+1}}$	15.) $u_c(t)$	$\frac{e^{-cs}}{s}$
5.) $\sin at$	$\frac{a}{s^2 + a^2}$	16.) $u_c(t) f(t-c)$	$e^{-cs} F(s)$
6.) $\cos at$	$\frac{s}{s^2 + a^2}$	17.) $e^{ct} f(t)$	$F(s-c)$
7.) e^{ibt}	$\frac{s+ib}{s^2 + b^2}$	18.) $f(ct)$	$\frac{1}{c} F\left(\frac{s}{c}\right)$
8.) $\sinh at$	$\frac{a}{s^2 - a^2}$	19.) $f * g(t)$	$F(s) G(s)$
9.) $\cosh at$	$\frac{s}{s^2 - a^2}$	20.) $\delta(t-c)$	e^{-cs}
10.) $e^{at} \sin bt$	$\frac{b}{(s-a)^2 + b^2}$	21.) $f'(t)$	$sF(s) - f(0)$
11.) $e^{at} \cos bt$	$\frac{s-a}{(s-a)^2 + b^2}$	22.) $(-t)^n f(t)$	$\frac{d^n F}{ds^n}(s)$

You should verify the claims in this table. It is best if you do it yourself. The fundamental result in using Laplace transforms is this. If you have $F(s) = G(s)$ then aside from finitely many jumps on each bounded interval, it follows that $f(t) = g(t)$. Thus you just go backwards in the table to find the desired functions. To see this shown, see Section 4.2 on Page 47. I will illustrate with a second order differential equation having constant coefficients. Of course you can change to a first order system and this will be the emphasis next, but you can also use the method directly. Note

$$\begin{aligned}
 \int_0^\infty y''(t) e^{-st} dt &= y'(t) e^{-st} \Big|_0^\infty + s \int_0^\infty y'(t) e^{-st} dt \\
 &= -y'(0) + s \int_0^\infty y'(t) e^{-st} dt \\
 &= -y'(0) + s \left[y(t) e^{-st} \Big|_0^\infty + s \int_0^\infty y(t) e^{-st} dt \right] \\
 &= -y'(0) - sy(0) + s^2 Y(s)
 \end{aligned} \tag{30.1}$$

A similar formula holds for higher derivatives. You can also get this by iterating 21.

Example 30.1.3 Find all solutions to the equation $y'' - 2y' + y = e^{-t}$.

From the table, first go to y'' . This gives $-y'(0) - sy(0) + s^2 Y(s)$ then you go to the next term which gives $-2sY(s) + 2y(0)$ and finally, you get $Y(s)$ from the y . On the right you get from formula 2. $1/(s+1)$. Therefore, you have

$$\begin{aligned}
 s^2 Y(s) - 2sY(s) + Y(s) - y'(0) - sy(0) + 2y(0) &= \frac{1}{s+1} \\
 (s^2 - 2s + 1) Y(s) &= y'(0) + (s-2)y(0) + \frac{1}{s+1}
 \end{aligned}$$

Thus we find the Laplace transform of the function desired.

$$\begin{aligned} Y(s) &= y'(0) \frac{1}{s^2 - 2s + 1} + y(0) \frac{s-2}{s^2 - 2s + 1} + \frac{\frac{1}{s+1}}{(s^2 - 2s + 1)} \\ &= y'(0) \frac{1}{s^2 - 2s + 1} + y(0) \frac{s-2}{s^2 - 2s + 1} + \frac{\frac{1}{s+1}}{(s^2 - 2s + 1)} \end{aligned}$$

Now you go backwards in the table. This typically involves doing partial fractions to get something which is in the table. It may be tedious, but is completely routine. You can also get this from a computer algebra system. More on this later. Thus we need

$$\begin{aligned} y'(0) \mathcal{L}^{-1} \left(\frac{1}{s^2 - 2s + 1} \right) + y(0) \mathcal{L}^{-1} \left(\frac{s-2}{s^2 - 2s + 1} \right) + \mathcal{L}^{-1} \left(\frac{1}{(s+2)(s^2 - 2s + 1)} \right) \\ \frac{\frac{1}{s+1}}{(s^2 - 2s + 1)} = \frac{1}{4(s+1)} + \frac{1}{2(s-1)^2} - \frac{1}{4(s-1)} \end{aligned}$$

Now you go backwards in the table to find that this comes from

$$\frac{1}{4}e^{-t} + \frac{1}{2}te^t - \frac{1}{4}e^t.$$

Next consider the other two terms.

$$\frac{s-2}{s^2 - 2s + 1} = -\frac{1}{(s-1)^2} + \frac{1}{s-1}$$

These are in the table.

$$\begin{aligned} \mathcal{L}^{-1} \left(\frac{s-2}{s^2 - 2s + 1} \right) &= -te^t + e^t \\ \mathcal{L}^{-1} \left(\frac{1}{s^2 - 2s + 1} \right) &= te^t \end{aligned}$$

Therefore, our solution is

$$y'(0)te^t + y(0)(-te^t + e^t) + \frac{1}{4}e^{-t} + \frac{1}{2}te^t - \frac{1}{4}e^t$$

If you specify $y'(0) = y(0) = 1$, then you will find the unique solution to the differential equation with initial conditions. It is

$$y(t) = \frac{1}{2}te^t + \frac{3}{4}e^t + \frac{1}{4}e^{-t}$$

You can check that this satisfies the initial conditions and the equation.

Another important formula mentioned in the above table is Formula 19. In this formula,

$$f * g(t) \equiv \int_0^t f(t-u)g(u)du = \int_0^t g(t-u)f(u)du$$

You can use change of variables to observe that the last equation is true so $f * g = g * f$.

Why is this formula so? It follows from the definition and interchanging the order of integration.

$$\begin{aligned}
\int_0^\infty e^{-st} f * g(t) dt &= \int_0^\infty e^{-st} \int_0^t f(t-u) g(u) du dt = \int_0^\infty \int_0^t e^{-st} f(t-u) g(u) du dt \\
&= \int_0^\infty \int_0^t e^{-s(t-u)} f(t-u) e^{-su} g(u) du dt = \int_0^\infty e^{-su} g(u) \int_u^\infty e^{-s(t-u)} f(t-u) dt du \\
&= \int_0^\infty e^{-su} g(u) \int_0^\infty e^{-sv} f(v) dv du = G(s) F(s)
\end{aligned}$$

Now here is another example in which the right side of the equation is such that it will be hard to find the Laplace transform.

Example 30.1.4 Solve the initial value problem $y'' + 5y' + 6y = \sin(t^2)$, $y(0) = 1$, $y'(0) = 0$.

Using the initial conditions and taking the Laplace transform of both sides,

$$s^2 Y(s) - y'(0) - sy(0) + 5sY(s) - 5y(0) + 6Y(s) = \mathcal{L}(\sin(t^2))$$

Now solve for $Y(s)$

$$Y(s)(s^2 + 5s + 6) = 5 + s + \mathcal{L}(\sin(t^2))$$

and so

$$Y(s) = \frac{5+s}{s^2+5s+6} + \frac{1}{s^2+5s+6} \mathcal{L}(\sin(t^2)) \quad (30.2)$$

Now

$$\begin{aligned}
\frac{1}{s^2+5s+6} &= -\frac{1}{s+3} + \frac{1}{s+2} \\
\frac{5+s}{s^2+5s+6} &= -\frac{2}{s+3} + \frac{3}{s+2}
\end{aligned}$$

so going backwards in the table

$$\frac{1}{s^2+5s+6} = \mathcal{L}(-e^{-3t}) + \mathcal{L}(e^{-2t}) = \mathcal{L}(-e^{-3t} + e^{-2t})$$

$$\frac{5+s}{s^2+5s+6} = \mathcal{L}(-2e^{-3t} + 3e^{-2t})$$

Using the convolution formula, and taking inverse Laplace transforms by going backwards in the table, it follows from 30.2

$$\begin{aligned}
y(t) &= -2e^{-3t} + 3e^{-2t} + \int_0^t (-e^{-3(t-u)} + e^{-2(t-u)}) \sin(u^2) du \\
&= -2e^{-3t} + 3e^{-2t} - e^{-3t} \int_0^t e^{3u} \sin(u^2) du + 3e^{-2t} \int_0^t e^{2u} \sin(u^2) du
\end{aligned}$$

If you are interested in a finite time interval, there is no loss of generality in using this method because any continuous function on $[0, T]$ can be considered the restriction to $[0, T]$ of one having exponential growth.

30.2 First Order Systems, Constant Coefficients

You want to find a matrix valued function $\Phi(t)$ such that

$$\Phi'(t) = A\Phi(t), \Phi(0) = I, A \text{ is } p \times p \quad (30.3)$$

Such a matrix is called a fundamental matrix. It turns out that if you can find $\Phi(t)$, you can always solve the first order system

$$x' = Ax + f, x(0) = x_0 \quad (30.4)$$

I also want to have $A\Phi(t) = \Phi(t)A$.

What is meant by the above symbols? The idea is that $\Phi(t)$ is a matrix whose entries are differentiable functions of t . The meaning of $\Phi'(t)$ is the matrix whose entries are the derivatives of the entries of $\Phi(t)$. For example, abusing notation slightly,

$$\begin{pmatrix} t & t^2 \\ \sin(t) & \tan(t) \end{pmatrix}' = \begin{pmatrix} 1 & 2t \\ \cos(t) & \sec^2(t) \end{pmatrix}.$$

What are some properties of this derivative? Does the product rule hold for example?

Lemma 30.2.1 Suppose $\Phi(t)$ is $m \times n$ and $\Psi(t)$ is $n \times p$ and these are differentiable matrices. Then

$$(\Phi(t)\Psi(t))' = \Phi'(t)\Psi(t) + \Phi(t)\Psi'(t)$$

Proof: By definition,

$$\begin{aligned} ((\Phi(t)\Psi(t))')_{ij} &= ((\Phi(t)\Psi(t))_{ij})' = \left(\sum_k \Phi(t)_{ik} \Psi(t)_{kj} \right)' \\ &= \sum_k \Phi'(t)_{ik} \Psi(t)_{kj} + \sum_k \Phi(t)_{ik} \Psi'(t)_{kj} \\ &= (\Phi'(t)\Psi(t))_{ij} + (\Phi(t)\Psi'(t))_{ij} \end{aligned}$$

and so the conclusion follows. ■

Now consider how to find the fundamental matrix $\Phi(t)$ to begin with. I will illustrate with an example.

Example 30.2.2 Let $A = \begin{pmatrix} -1 & 2 \\ -3 & 4 \end{pmatrix}$. Find the fundamental matrix.

I want $\Phi'(t) = A\Phi(t)$, $\Phi(0) = I$. Take the Laplace transform of both sides. By this I mean replace each entry of the matrix with its Laplace transform. Then if $F(s)$ is the name of the Laplace transform of $\Phi(t)$,

$$sF(s) - I = AF(s) \text{ so } (sI - A)F(s) = I$$

and so $F(s) = (sI - A)^{-1}$. Now this is easy to find using the formula for the inverse presented earlier. Recall you took the transpose of the cofactor matrix and divided by the determinant to get the inverse. See Theorem 27.2.1. In this example,

$$F(s) = (sI - A)^{-1} = \left(s \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} -1 & 2 \\ -3 & 4 \end{pmatrix} \right)^{-1} = \begin{pmatrix} \frac{s-4}{s^2-3s+2} & \frac{2}{s^2-3s+2} \\ -\frac{3}{s^2-3s+2} & \frac{s+1}{s^2-3s+2} \end{pmatrix}$$

Note how the entries are all rational functions. This will ALWAYS happen no matter what matrix you use and this follows from that method for finding the inverse in terms of the transpose of the cofactor method. Since theoretically, by the fundamental theorem of algebra, one can always factor a polynomial into a product of linear and irreducible quadratics in the denominator of those rational functions, this process will ALWAYS work with the caveat that one might not be able to actually carry out the factorization of the polynomials in the denominator. However this shows that the fundamental matrix does exist and that your ability to explicitly compute it is exactly as good as your ability to factor a polynomial. In this case, I can take the inverse Laplace transform of that matrix on the right and get

$$F(s) = \begin{pmatrix} \frac{3}{s-1} - \frac{2}{s-2} & \frac{2}{s-2} - \frac{2}{s-1} \\ \frac{3}{s-1} - \frac{3}{s-2} & \frac{3}{s-2} - \frac{2}{s-1} \end{pmatrix}$$

$$\Phi(t) = \begin{pmatrix} 3e^t - 2e^{2t} & 2e^{2t} - 2e^t \\ 3e^t - 3e^{2t} & 3e^{2t} - 2e^t \end{pmatrix}$$

Does it work?

$$D_t \begin{pmatrix} 3e^t - 2e^{2t} & 2e^{2t} - 2e^t \\ 3e^t - 3e^{2t} & 3e^{2t} - 2e^t \end{pmatrix} = \begin{pmatrix} 3e^t - 4e^{2t} & 4e^{2t} - 2e^t \\ 3e^t - 6e^{2t} & 6e^{2t} - 2e^t \end{pmatrix}$$

$$\begin{pmatrix} -1 & 2 \\ -3 & 4 \end{pmatrix} \begin{pmatrix} 3e^t - 2e^{2t} & 2e^{2t} - 2e^t \\ 3e^t - 3e^{2t} & 3e^{2t} - 2e^t \end{pmatrix} = \begin{pmatrix} 3e^t - 4e^{2t} & 4e^{2t} - 2e^t \\ 3e^t - 6e^{2t} & 6e^{2t} - 2e^t \end{pmatrix}$$

so yes, it solves the equation. Also $\Phi(0) = I$. Thus this is indeed the fundamental matrix.

30.2.1 Some Technical Considerations*

Now if $F(s) = (sI - A)^{-1} = \mathcal{L}(\Phi(t))$, is $\Phi'(t) = A\Phi(t)$? is $\Phi(0) = I$? Here we are assuming that the entries of $\Phi(t)$ have exponential growth. Then multiplying through by $(sI - A)$,

$$I = (sI - A) \int_0^\infty e^{-st} \Phi(t) dt = (I - A/s) \int_0^\infty se^{-st} \Phi(t) dt \quad (30.5)$$

$$(I - A/s) \int_0^\infty se^{-st} (\Phi(t) - \Phi(0)) dt + (I - A/s) \Phi(0) \quad (30.6)$$

because $\int_0^\infty se^{-st} dt = 1$, this being true for all large enough s . Letting $s \rightarrow \infty$, the first term converges to 0. Here is roughly why this is so. Letting $\delta > 0$, be so small that all entries of $\Phi(t)$ are closer than ε to the entries of $\Phi(0)$ whenever $t < \delta$,

$$\left\| \int_\delta^\infty se^{-st} (\Phi(t) - \Phi(0)) dt \right\| \leq \int_\delta^\infty se^{-st} (Ce^{\lambda t} + C) dt$$

where $\|A\|$ will denote the maximum of the absolute values of all entries of A and it is assumed that each of these is no more than $Ce^{\lambda t}$. Now that integral on the right can be computed and it equals the following for large s

$$\left(\frac{1}{-s + \lambda} e^{-st + \lambda t} s - e^{-st} \right) \Big|_\delta^\infty = \left(\frac{1}{s - \lambda} e^{-s\delta + \lambda\delta} s + e^{-s\delta} \right)$$

Letting $s \rightarrow \infty$, this clearly converges to 0. Also,

$$\left\| \int_0^\delta se^{-st} (\Phi(t) - \Phi(0)) dt \right\| \leq \int_0^\delta se^{-st} \varepsilon dt < \varepsilon$$

and so for all s large enough, $\|(I - A/s) \int_0^\infty se^{-st} (\Phi(t) - \Phi(0)) dt\| < \varepsilon$ showing that this does indeed converge to 0. The last term in 30.6 converges to $\Phi(0)$ as $s \rightarrow \infty$ and so we do indeed have $\Phi(0) = I$.

What about $\Phi'(t) = A\Phi(t)$? For large s , integrate by parts using $\Phi(0) = I$ to obtain

$$\begin{aligned} \int_0^\infty e^{-st} \Phi'(t) dt &= -I + \int_0^\infty se^{-st} \Phi(t) dt = -I + sF(s) \\ \int_0^\infty e^{-st} A\Phi(t) dt &= AF(s) \end{aligned}$$

Is $-I + sF(s) = AF(s)$? Yes because $F(s) = (sI - A)^{-1}$ and so the Laplace transforms of $\Phi'(t)$ and $A\Phi(t)$ are the same. This means the two functions are the same because they are both continuous, something which is shown later that the Laplace transform determines the functions from which it comes. This has shown the following important theorem.

Theorem 30.2.3 *Let A be a $p \times p$ matrix and suppose $F(s) = (sI - A)^{-1} = \mathcal{L}(\Phi(t))$ for $\Phi(t)$ a matrix whose entries have exponential growth. Then $\Phi'(t) = A\Phi(t)$, $\Phi(0) = I$. Conversely, if $\Phi'(t) = A\Phi(t)$, $\Phi(0) = I$, then $\mathcal{L}(\Phi(t)) = (sI - A)^{-1}$. Thus the fundamental matrix is unique.*

As noted above, one can ALWAYS find from the table of Laplace transforms an explicit solution $\Phi(t)$ whose Laplace transform is $(sI - A)^{-1}$ provided you can factor the polynomials in the denominators of the rational functions which are the entries of $(sI - A)^{-1}$. Such factorizations always exist by the fundamental theorem of algebra and so the fundamental matrix always exists. Thus your ability to find an explicit formula for such a fundamental matrix is exactly as good as your ability to factor polynomials which occur as denominators in the formula for $(sI - A)^{-1}$. Note that $\Phi(t)$ is unique, because if you have one then its Laplace transform must be $(sI - A)^{-1}$.

One other item is of interest in these fundamental matrices and this is the group property.

Theorem 30.2.4 *The following hold:*

- Lemma 30.2.5**
1. *If $\Psi'(t) = A\Psi(t)$, $\Psi(0) = 0$, then $\Psi(t) = 0$.*
 2. *If $\Phi(t)$ is the fundamental matrix for A then $\Phi(t)A = A\Phi(t)$.*
 3. *Also if $\Phi(t)$ is the fundamental matrix, then $\Phi(t+u) = \Phi(t)\Phi(u)$. In particular $\Phi(t)^{-1} = \Phi(-t)$.*

Proof: 1. Consider the first claim. Letting $G(s)$ be the Laplace transform, it follows that

$$sG(s) - 0 = AG(s)$$

for all s large enough. This is impossible unless $G(s) = 0$. Therefore, $\mathcal{L}(0) = \mathcal{L}(\Psi(t))$ and so $0 = \Psi(t)$ from what is shown later about the Laplace transform determining the function.

2. $(A\Phi(t) - \Phi(t)A)' = A^2\Phi(t) - A\Phi(t)A = A(A\Phi(t) - \Phi(t)A)$, and also $A\Phi(0) - \Phi(0)A = A - A = 0$ for from **1.**, it follows that $A\Phi(t) - \Phi(t)A = 0$.

3. Using **2.**, and letting t be the variable of differentiation,

$$\begin{aligned} (\Phi(t+u) - \Phi(t)\Phi(u))' &= A\Phi(t+u) - A\Phi(t)\Phi(u) \\ &= A(\Phi(t+u) - \Phi(t)\Phi(u)) \end{aligned}$$

Also $\Phi(0+u) - \Phi(0)\Phi(u) = \Phi(u) - \Phi(u) = 0$ so by part **1.**, it follows that

$$\Phi(t+u) - \Phi(t)\Phi(u) = 0. \blacksquare$$

30.2.2 Solving a First Order System

If you can find the fundamental matrix, it is easy to solve a first order system.

$$x' = Ax + f, \quad x(0) = x_0 \quad (30.7)$$

Multiply on the left by $\Phi(-t)$ and permute A and $\Phi(t)$ as needed using Theorem 30.2.4.

$$\Phi(-t)x' - A\Phi(-t)x = \Phi(-t)f$$

$$(\Phi(-t)x)' = \Phi(-t)f(t)$$

Now integrate and obtain

$$\Phi(-t)x(t) - x_0 = \int_0^t \Phi(-u)f(u)du$$

Now multiply on left by $\Phi(t)$ to obtain

$$\begin{aligned} x(t) &= \Phi(t)x_0 + \int_0^t \Phi(t)\Phi(-u)f(u)du \\ &= \Phi(t)x_0 + \int_0^t \Phi(t-u)f(u)du \end{aligned}$$

Therefore, there is at most one solution to 30.7 and if there is one, then this is it.

Theorem 30.2.6 *There exists a unique solution to 30.7 and it is given by*

$$x(t) = \Phi(t)x_0 + \int_0^t \Phi(t-u)f(u)du$$

Proof: I just showed there is at most one solution. It only remains to verify that the above works. However, the formula can be written as

$$x(t) = \Phi(t)x_0 + \Phi(t) \int_0^t \Phi(-u)f(u)du$$

When $t = 0$ this yields x_0 as it should. Now differentiate. Using the product rule,

$$\begin{aligned} x'(t) &= A\Phi(t)x_0 + A\Phi(t) \int_0^t \Phi(-u)f(u)du + \overbrace{\Phi(t)\Phi(-t)}^I f(t) \\ &= Ax(t) + f(t). \blacksquare \end{aligned}$$

Example 30.2.7 Find the solution to

$$\mathbf{x}' = \begin{pmatrix} -4 & -3 \\ 6 & 5 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \cos t \\ e^t \end{pmatrix}, \mathbf{x}(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

First find the fundamental matrix. One goes backwards in the table to find the following.

$$\begin{aligned} (sI - A)^{-1} &= \left(s \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} -4 & -3 \\ 6 & 5 \end{pmatrix} \right)^{-1} \\ &= \begin{pmatrix} -\frac{s-5}{-s^2+s+2} & \frac{3}{-s^2+s+2} \\ -\frac{1}{-\frac{1}{6}s^2+\frac{1}{6}s+\frac{1}{3}} & -\frac{\frac{1}{6}s+\frac{2}{3}}{-\frac{1}{6}s^2+\frac{1}{6}s+\frac{1}{3}} \end{pmatrix} \end{aligned}$$

Then using going backwards in the table and writing in terms of cosh and sinh,

$$\Phi(t) = \begin{pmatrix} e^{\frac{1}{2}t} (\cosh \frac{3}{2}t - 3 \sinh \frac{3}{2}t) & -2 (\sinh \frac{3}{2}t) e^{\frac{1}{2}t} \\ 4 (\sinh \frac{3}{2}t) e^{\frac{1}{2}t} & e^{\frac{1}{2}t} (\cosh \frac{3}{2}t + 3 \sinh \frac{3}{2}t) \end{pmatrix}$$

Then the solution is

$$\begin{aligned} \mathbf{x}(t) &= \begin{pmatrix} e^{\frac{1}{2}t} (\cosh \frac{3}{2}t - 3 \sinh \frac{3}{2}t) & -2 (\sinh \frac{3}{2}t) e^{\frac{1}{2}t} \\ 4 (\sinh \frac{3}{2}t) e^{\frac{1}{2}t} & e^{\frac{1}{2}t} (\cosh \frac{3}{2}t + 3 \sinh \frac{3}{2}t) \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &\quad + \int_0^t \Phi(t-u) \mathbf{f}(u) du \\ \mathbf{x}(t) &= \begin{pmatrix} e^{\frac{1}{2}t} (\cosh \frac{3}{2}t - 5 \sinh \frac{3}{2}t) \\ e^{\frac{1}{2}t} (\cosh \frac{3}{2}t + 7 \sinh \frac{3}{2}t) \end{pmatrix} + \int_0^t \Phi(t-s) \begin{pmatrix} \cos s \\ e^s \end{pmatrix} ds \end{aligned}$$

Doing the integrations, one obtains

$$\begin{aligned} &\int_0^t \Phi(t-s) \begin{pmatrix} \cos s \\ e^s \end{pmatrix} ds \\ &= \begin{pmatrix} \frac{1}{10} e^{-t} (15e^{2t} - 14e^{3t} + 14(\cos t)e^t + 8e^t \sin t - 15) \\ -\frac{1}{10} e^{-t} (25e^{2t} - 28e^{3t} + 18(\cos t)e^t + 6e^t \sin t - 15) \end{pmatrix} \end{aligned}$$

It follows that $\mathbf{x}(t) =$

$$\begin{pmatrix} \frac{3}{2}e^t + \frac{4}{5}\sin t - \frac{3}{2}e^{-t} - \frac{7}{5}e^{2t} + (\cosh \frac{3}{2}t)e^{\frac{1}{2}t} \\ -5(\sinh \frac{3}{2}t)e^{\frac{1}{2}t} + \frac{7}{5}(\cos t)e^te^{-t} \\ \frac{3}{2}e^{-t} - \frac{3}{5}\sin t - \frac{5}{2}e^t + \frac{14}{5}e^{2t} + (\cosh \frac{3}{2}t)e^{\frac{1}{2}t} \\ + 7(\sinh \frac{3}{2}t)e^{\frac{1}{2}t} - \frac{9}{5}(\cos t)e^te^{-t} \end{pmatrix}$$

Using the table as just described really is a pretty good way to solve these kinds of equations, but there is a much easier way to do it. You let the computer algebra system do the tedious work for you. Here is the general idea for a first order system. Be patient. I will

consider specific examples a little later. However, if you are looking for something which will solve all first order systems in closed form using known elementary functions, then you are looking for something which is not there. You can indeed speak of it in general theoretical terms but the only problems which are completely solvable in closed form are those for which you can exactly find the eigenvalues of the matrix. Unfortunately, this involves solving polynomial equations and none of us can do these in general.

30.2.3 Using a Computer Algebra System

You want to solve

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}, \mathbf{x}(0) = \mathbf{x}_0$$

Use the Property 21. and take Laplace transforms of both sides. Thus

$$sX(s) - \mathbf{x}_0 = AX(s) + F(s)$$

where $X(s)$ is the Laplace transform of $\mathbf{x}(t)$ and $F(s)$ is the Laplace transform of $\mathbf{f}(t)$. Then you can solve for $X(s)$, at least for large enough s so that $(sI - A)^{-1}$ exists. Thus

$$(sI - A)X(s) = \mathbf{x}_0 + F(s)$$

Then

$$X(s) = (sI - A)^{-1}(\mathbf{x}_0 + F(s))$$

Note that there is even a formula for $(sI - A)^{-1}$. See Theorem 28.1.14. Thus you can **always** find $X(s)$. Then having done so, it is a matter of finding the function whose Laplace transform gives $X(s)$. By hand, you would consider each entry of $X(s)$ and by using partial fractions, you would go backwards in the table. It won't always work. Sometimes you won't be able to factor the polynomials enough to carry this out and even when it does work, it will be pretty tedious. This is why you should use Matlab or some computer algebra system. Here is an example which can be done. The reason I know it will work out is that I cooked it up to work out. I picked a matrix whose eigenvalues are known. I also picked the forcing function to be something which will tend to make things work.

Example 30.2.8 Solve the following first order system.

$$\mathbf{x}' = \begin{pmatrix} 2 & 2 & -1 \\ -1 & 0 & 1 \\ -1 & 0 & 2 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \cos t \\ \sin t \\ e^t \end{pmatrix}, \mathbf{x}(0) = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

Following the above general procedure, the Laplace transform of the forcing function is

$$\begin{pmatrix} \frac{s}{s^2+1} \\ \frac{1}{s^2+1} \\ \frac{1}{s-1} \end{pmatrix}$$

and so

$$X(s) = \left(\begin{pmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & s \end{pmatrix} - \begin{pmatrix} 2 & 2 & -1 \\ -1 & 0 & 1 \\ -1 & 0 & 2 \end{pmatrix} \right)^{-1} \left(\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} \frac{s}{s^2+1} \\ \frac{1}{s^2+1} \\ \frac{1}{s-1} \end{pmatrix} \right)$$

Now I compute this.

$$X(s) = \begin{pmatrix} -\frac{1}{(s^2+1)(s-1)^3}(-s^4+s^3+2) \\ \frac{1}{(s-1)^2} \frac{s^2-3s+4}{s^3-2s^2+s-2} \\ \frac{1}{(s-1)^3(s^3-2s^2+s-2)}(s^5-3s^4+3s^3-2s^2+s+2) \end{pmatrix}$$

At this point, I use partial fractions and go backwards in the table or I ask a computer algebra system to find the inverse Laplace transform. I recommend using the computer algebra system. Thus

$$\mathbf{x}(t) = \begin{pmatrix} \frac{1}{2}\cos t + \frac{1}{2}e^t - \frac{1}{2}t^2e^t + \frac{3}{2}te^t \\ \frac{1}{2}e^t - \frac{9}{10}\cos t - \frac{3}{10}\sin t + \frac{2}{5}e^{2t} - te^t \\ \frac{1}{5}\cos t - \frac{1}{10}\sin t + \frac{4}{5}e^{2t} - \frac{1}{2}t^2e^t + \frac{1}{2}te^t \end{pmatrix}$$

This is then the solution to the first order system. I used Scientific Notebook to do all of these computations. However, one can also use Matlab. You will need Matlab and the symbolic math toolbox installed for this to work.

```
>>syms s t; a=(enter initial vector here); b=(enter sI-A here); c=(enter f(t) here);
simplify(ilaplace(inv(b)*(a+laplace(c))))
```

I will use this to solve the above problem.

```
>> syms s t; a=[1;0;1]; b=[s-2 -2 1;1 s -1;1 0 s-2]; c=[cos(t);sin(t);exp(t)];
simplify(ilaplace(inv(b)*(a+laplace(c))))
```

Note the use of square brackets in entering the matrix. You must use these. You enter one row at a time with a space between successive entries and a semicolon to indicate the start of a new row. Then you press enter on your keyboard and it will produce the following:

```
cos(t)/2 + exp(t)/2 - (t^2*exp(t))/2 + (3*t*exp(t))/2
(2*exp(2*t))/5 - (9*cos(t))/10 + exp(t)/2 - (3*sin(t))/10 - t*exp(t)
(4*exp(2*t))/5 + cos(t)/5 - sin(t)/10 - (t^2*exp(t))/2 + (t*exp(t))/2
```

The advantage to using Scientific notebook is the result comes out looking a lot nicer but you get the same thing either way. In fact Scientific notebook is based on mupad which is part of the symbolic math toolbox in Matlab.

Example 30.2.9 Find the fundamental matrix of

$$A = \begin{pmatrix} -3 & 2 & -1 \\ 0 & -1 & 1 \\ 4 & -4 & 3 \end{pmatrix}$$

and use to solve the initial value problem

$$\mathbf{x}' = A\mathbf{x} + \begin{pmatrix} \ln(t^2+1) \\ \sin(t^2) \\ \cos(t) \end{pmatrix}, \quad \mathbf{x}(0) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

It will not be possible to give a closed form solution for this problem but we can write it in terms of an integral if the fundamental matrix is found.

$$\begin{aligned}\Psi(s) &= \left(s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} -3 & 2 & -1 \\ 0 & -1 & 1 \\ 4 & -4 & 3 \end{pmatrix} \right)^{-1} \\ &= \begin{pmatrix} \frac{s-1}{s^2+2s+1} & \frac{2}{s^2+2s+1} & -\frac{1}{s^2+2s+1} \\ -\frac{4}{-s^3-s^2+s+1} & -\frac{s^2-5}{-s^3-s^2+s+1} & -\frac{s+3}{-s^3-s^2+s+1} \\ \frac{1}{\frac{1}{4}s^2-\frac{1}{4}} & -\frac{1}{\frac{1}{4}s^2-\frac{1}{4}} & \frac{\frac{1}{4}s+\frac{3}{4}}{\frac{1}{4}s^2-\frac{1}{4}} \end{pmatrix}\end{aligned}$$

Therefore,

$$\Phi(t) = \begin{pmatrix} -e^{-t}(2t-1) & 2te^{-t} & -te^{-t} \\ e^t - e^{-t} - 2te^{-t} & 2e^{-t} - e^t + 2te^{-t} & e^t - e^{-t} - te^{-t} \\ 4\sinh t & -4\sinh t & 2e^t - e^{-t} \end{pmatrix}$$

Then the solution is

$$\begin{aligned}x(t) &= \begin{pmatrix} te^{-t} - e^{-t}(2t-1) \\ e^t - te^{-t} \\ 2e^t - e^{-t} \end{pmatrix} + \\ &+ \int_0^t \Phi(t-u) \begin{pmatrix} \ln(u^2+1) \\ \sin(u^2) \\ \cos(u) \end{pmatrix} du\end{aligned}$$

where $\Phi(t)$ is given above.

To find the fundamental matrix in Matlab, you would use the following syntax.

```
>> syms s t; b=[s+3 -2 1;0 s+1 -1;-4 4 s-3];
simplify(ilaplace(inv(b)))
```

Then you press enter and it gives the following:

```
[ -exp(-t)*(2*t - 1), 2*t*exp(-t), -t*exp(-t)]
[ -exp(-t)*(2*t - exp(2*t) + 1), exp(-t)*(2*t - exp(2*t) + 2), -exp(-t)*(t - exp(2*t) + 1)]
[ 2*exp(t) - 2*exp(-t), 2*exp(-t) - 2*exp(t), 2*exp(t) - exp(-t)]
```

which is just a messier version of what was obtained above using Scientific notebook.

Actually it might be a little easier to use the following syntax. You can adjust as needed.

```
>>syms s t; b=eye(3); c=[-3 2 -1;0 -1 1;4 -4 3];
simplify(ilaplace(inv(s*b-c)))
```

Here you just need to enter the matrix for c and the symbol `eye(3)` says it is a 3×3 identity matrix. Matlab then does the rest for you.

30.3 Homogeneous Particular and General Solutions

It is convenient to split things up into homogeneous problems and then look for particular solutions. First of all, is a definition of the general solution to a homogeneous problem.

Definition 30.3.1 Let A be an $n \times n$ matrix. The general solution to the homogeneous problem is defined to be all solutions to the equation

$$x' = Ax$$

Note how there is no initial condition. We just look for all solutions to the above differential equation. The following theorem describes all of these solutions.

Theorem 30.3.2 *The general solution to the homogeneous problem $\mathbf{x}' = A\mathbf{x}$ consists of all vectors of the form $\Phi(t)\mathbf{c}$ where \mathbf{c} is a vector in \mathbb{F}^n and $\Phi(t)$ is the fundamental matrix of A .*

Proof: Let \mathbf{x} be a solution to the equation. Then $\mathbf{x}(0) = \mathbf{c}$ for some \mathbf{c} . Consider $\Phi(t)\mathbf{c}$ and $\mathbf{x}(t)$ both solve $\mathbf{x}' = A\mathbf{x}$ the first doing so because

$$\Phi'(t)\mathbf{c} = A\Phi(t)\mathbf{c}$$

Thus $\Phi(t)\mathbf{c}$ and $\mathbf{x}(t)$ both solve the same differential equation and have the same initial condition. Therefore, these are the same and this shows that the set of solutions to $\mathbf{x}' = A\mathbf{x}$ consists of $\Phi(t)\mathbf{c}$ for $\mathbf{c} \in \mathbb{F}^n$ as claimed. ■

Example 30.3.3 *Find the general solution to $\mathbf{x}' = A\mathbf{x}$ where*

$$A = \begin{pmatrix} 4 & 1 & 2 & 1 \\ -3 & 0 & -2 & -1 \\ -7 & -3 & -4 & -3 \\ 8 & 4 & 6 & 5 \end{pmatrix}$$

According to the above theory, it suffices to find the fundamental matrix. The inverse of $sI - A$ is the matrix which has the following columns, beginning at the left and moving toward the right:

$$\begin{pmatrix} \frac{s+2}{s^2-2s+1} \\ -\frac{3}{s^2-2s+1} \\ \frac{7s-13}{-s^3+4s^2-5s+2} \\ -\frac{1}{3s-5}(4s-7)\frac{6s-10}{-s^3+4s^2-5s+2} \end{pmatrix}, \begin{pmatrix} \frac{1}{s^2-2s+1} \\ \frac{s-2}{s^2-2s+1} \\ \frac{3s-5}{-s^3+4s^2-5s+2} \\ -\frac{2s-3}{3s-5}\frac{6s-10}{-s^3+4s^2-5s+2} \end{pmatrix},$$

$$\begin{pmatrix} \frac{2}{s^2-2s+1} \\ -\frac{2}{s^2-2s+1} \\ -\frac{s^2-8s+11}{-s^3+4s^2-5s+2} \\ \frac{6s-10}{-s^3+4s^2-5s+2} \end{pmatrix}, \begin{pmatrix} \frac{1}{s^2-2s+1} \\ -\frac{1}{s^2-2s+1} \\ \frac{3s-5}{-s^3+4s^2-5s+2} \\ -\frac{s^2+s-4}{-s^3+4s^2-5s+2} \end{pmatrix}$$

This was done by a computer algebra system. Now take inverse Laplace transforms of this to get the fundamental matrix $\Phi(t) =$

$$\begin{pmatrix} e^t(3t+1) & te^t & 2te^t & te^t \\ -3te^t & -e^t(t-1) & -2te^t & -te^t \\ e^t - e^{2t} - 6te^t & e^t - e^{2t} - 2te^t & 2e^t - e^{2t} - 4te^t & e^t - e^{2t} - 2te^t \\ 2e^{2t} - 2e^t + 6te^t & 2e^{2t} - 2e^t + 2te^t & 2e^{2t} - 2e^t + 4te^t & 2e^{2t} - e^t + 2te^t \end{pmatrix}$$

therefore, the general solution is of the form $\Phi(t)\mathbf{c}$ where $\mathbf{c} \in \mathbb{F}^n$. In other words, it is the set of linear combinations of the columns of $\Phi(t)$. Since $\Phi(t)^{-1} = \Phi(-t)$, the columns

are linearly independent and this shows that the dimension of the solution space is n if A is $n \times n$. In the above example, the dimension of the general solution is 4 because A is 4×4 .

Now consider the general solution to

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}$$

There is a very easy way to describe this. It is just the general solution to $\mathbf{x}' = A\mathbf{x}$ added to \mathbf{x}_p where \mathbf{x}_p is any particular solution to the above nonhomogeneous equation.

Theorem 30.3.4 *The general solution to $\mathbf{x}' = A\mathbf{x} + \mathbf{f}$ consists of all solutions to this equation. It is of the form $\Phi(t)\mathbf{c} + \mathbf{x}_p$ where \mathbf{x}_p is a particular solution meaning $\mathbf{x}_p' = A\mathbf{x}_p + \mathbf{f}$.*

Proof: Anything of the form $\Phi(t)\mathbf{c} + \mathbf{x}_p$ is a solution to $\mathbf{x}' = A\mathbf{x} + \mathbf{f}$. It remains to verify that this is the only way it can happen. Let $\mathbf{z}' = A\mathbf{z} + \mathbf{f}$ and consider $(\mathbf{z} - \mathbf{x}_p)$. Then

$$(\mathbf{z} - \mathbf{x}_p)' = \mathbf{z}' - \mathbf{x}_p' = A\mathbf{z} + \mathbf{f} - (A\mathbf{x}_p + \mathbf{f}) = A(\mathbf{z} - \mathbf{x}_p)$$

and so $\mathbf{z} - \mathbf{x}_p$ is a solution to $\mathbf{x}' = A\mathbf{x}$. Therefore, from Theorem 30.3.2, there exists $\mathbf{c} \in \mathbb{F}^n$ such that $\mathbf{z}(t) = \Phi(t)\mathbf{c} + \mathbf{x}_p(t)$. ■

Example 30.3.5 *Find the general solution to*

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}$$

where

$$A = \begin{pmatrix} 2 & -4 & -2 \\ 3 & -4 & -2 \\ -3 & 10 & 6 \end{pmatrix}, \quad \mathbf{f}(t) = \begin{pmatrix} e^t \sin t \\ e^{-t} \cos t \\ t \end{pmatrix}$$

First I will find the fundamental matrix using the following syntax.

```
>>syms s t; b=eye(3);
c=[2 -4 -2;3 -4 -2;-3 10 6];f=[exp(t)*sin(t);exp(-t)*cos(t);t];
simplify(ilaplace(inv(s*b-c)))
simplify(ilaplace(inv(s*b-c)*laplace(f)))
```

The first line starting with “simplify” will give the fundamental matrix and the second will give a particular solution. The claim about the first was already considered. As to the second, if \mathbf{x} is a particular solution with zero initial condition,

$$sX(s) = AX(s) + F(s)$$

In the above syntax, the $F(s)$ comes from $\text{laplace}(f)$. Then

$$X(s) = (sI - A)^{-1} F(s)$$

and this involves $\text{inv}(s*b-c)*\text{laplace}(f)$ in the above syntax. Then you do ilaplace to this thing to get a particular solution. Try it. You will get a horrendous mess but Matlab has no problem in doing it.

This has shown how to solve first order systems at least up to a suitable variation of constants formula. There is one other topic which is sometimes useful and that is the convolution integral and its relation to the Laplace transform.

Theorem 30.3.6 Suppose $F(s)$ is the Laplace transform of $f(t)$ and $G(s)$ is the Laplace transform of $g(t)$. Then $F(s)G(s)$ is the Laplace transform of

$$\int_0^t f(u)g(t-u)du = \int_0^t f(t-u)g(u)du \equiv f * g(t)$$

Proof: To be rigorous, you really need to replace improper integrals with integrals over a finite interval and then take a limit, but the idea is essentially as follows:

$$\begin{aligned} \int_0^\infty e^{-st} \int_0^t f(t-u)g(u)dudt &= \int_0^\infty \int_u^\infty e^{-st} f(t-u)g(u)dtdu \\ &= \int_0^\infty \int_u^\infty e^{-s(t-u)} f(t-u) e^{-su} g(u)dtdu \\ &= \int_0^\infty \int_0^\infty e^{-sr} f(r) e^{-su} g(u)drdu \\ &= \int_0^\infty e^{-su} g(u) \left(\int_0^\infty e^{-sr} f(r)dr \right) du \\ &= \int_0^\infty e^{-sr} f(r)dr \int_0^\infty e^{-su} g(u)du \\ &= F(s)G(s) \end{aligned}$$

The other formula follows from changing the variable. ■

Note that $F(s)$ could be a matrix and $G(s)$ could be a vector. You simply need the multiplication to make sense.

Example 30.3.7 Find a particular solution to

$$\mathbf{x}'(t) = A\mathbf{x}(t) + \mathbf{f}(t)$$

where

$$\mathbf{f}(t) = \begin{pmatrix} t \\ t \\ \ln(t^2 + 1) \end{pmatrix}$$

and

$$A = \begin{pmatrix} -1 & 0 & -6 \\ -2 & 1 & -5 \\ 1 & 0 & 4 \end{pmatrix}$$

There is no way you will find a decent closed form solution to this in terms of elementary functions because of the horrible $\ln(t^2 + 1)$ but this is not really a problem because you can find a particular solution in terms of a convolution. You just need to find the fundamental matrix which is not hard. I will use the following to find the fundamental matrix.

```
syms s t; b=eye(3); c=[-1 0 -6;-2 1 -5;1 0 4];
```

```
simplify(ilaplace(inv(s*b-c)))
```

This yields for $\Phi(t)$

$$\begin{pmatrix} -e^t(2e^t - 3) & 0 & -6e^t(e^t - 1) \\ -e^t(t + e^t - 1) & e^t & -e^t(2t + 3e^t - 3) \\ e^t(e^t - 1) & 0 & e^t(3e^t - 2) \end{pmatrix}$$

Then using the above theorem, $\mathbf{x}_p(t) =$

$$\int_0^t \begin{pmatrix} -e^u(2e^u-3) & 0 & -6e^u(e^u-1) \\ -e^u(u+e^u-1) & e^u & -e^u(2u+3e^u-3) \\ e^u(e^u-1) & 0 & e^u(3e^u-2) \end{pmatrix} \begin{pmatrix} t-u \\ t-u \\ \ln((t-u)^2+1) \end{pmatrix} du$$

This gives a perfectly good description of a particular solution. Thus the general solution is of the form

$$\begin{pmatrix} -e^t(2e^t-3) & 0 & -6e^t(e^t-1) \\ -e^t(t+e^t-1) & e^t & -e^t(2t+3e^t-3) \\ e^t(e^t-1) & 0 & e^t(3e^t-2) \end{pmatrix} \mathbf{c} + \mathbf{x}_p(t)$$

Here \mathbf{c} is an arbitrary vector in \mathbb{R}^n . Note how this is essentially a return to the notion of the variation of constants formula presented earlier.

Of course all of this depends on being able to say that if two functions have the same Laplace transform, then they must in some sense be the same function. This will be discussed later when it will also be shown how to explicitly go backwards in the table and find the original function given its Laplace transform.

30.4 Higher Order Scalar Linear Equations

Recall these are differential equations which are of the form

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y = f(t)$$

The notation $y^{(k)}$ means the k^{th} derivative, and it is assumed that all given functions are continuous. There is nothing new about these equations. They can all be studied as special cases of first order systems.

The following is a procedure for changing one of these higher order linear equations into a first order system which is the right way to study differential equations.

PROCEDURE 30.4.1 Consider the equation

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y = f$$

To write as a first order system, do the following.

$$\begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(n) \end{pmatrix} = \begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{pmatrix}$$

Then, suppressing the dependence on t ,

$$\begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(n-1) \\ x(n) \end{pmatrix}' = \begin{pmatrix} x(2) \\ x(3) \\ \vdots \\ x(n) \\ y^{(n)} \end{pmatrix} = \begin{pmatrix} x(2) \\ x(3) \\ \vdots \\ x(n) \\ -(a_{n-1}x(n) + \cdots + a_1x(2) + a_0x(1)) + f \end{pmatrix}$$

In terms of matrices,

$$\begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(n-1) \\ x(n) \end{pmatrix}' = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-2} & -a_{n-1} \end{pmatrix} \begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(n-1) \\ x(n) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ f \end{pmatrix}$$

In case $f = 0$ so you have a homogeneous equation,

$$\begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(n-1) \\ x(n) \end{pmatrix}' = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-2} & -a_{n-1} \end{pmatrix} \begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(n-1) \\ x(n) \end{pmatrix} \quad (30.8)$$

it follows that in the above reduction to a first order system,

$$\begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(n-1) \\ x(n) \end{pmatrix} = \begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \\ y^{(n-1)} \end{pmatrix}$$

where y is the solution to the higher order scalar equation and y is a solution to this scalar higher order equation if and only if

$$\begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \\ y^{(n-1)} \end{pmatrix}$$

is a solution to the above first order system.

When you have a fundamental matrix for A $\Phi(t)$, recall that the determinant of $\Phi(t)$ is not zero because this matrix has an inverse, namely $\Phi(-t)$. In general, if you have $x'_k = Ax_k$ for $k \leq p$ where A is a $p \times p$ matrix, you could form

$$\Psi(t) \equiv \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix}(t) \quad (30.9)$$

and $\Psi' = \begin{pmatrix} Ax_1 & Ax_2 & \cdots & Ax_p \end{pmatrix} = A\Psi(t)$.

Theorem 30.4.2 Let $\Psi(t)$ be as in 30.9 where $x'_k = Ax_k$. Then $\Psi(t)^{-1}$ exists for all t if and only if $\Psi(0)^{-1}$ exists and if this happens, then the fundamental matrix is $\Phi(t) = \Psi(t)\Psi(0)^{-1}$.

Proof: \Leftarrow Say $\Psi(0)^{-1}$ exists. Then $(\Psi\Psi(0)^{-1})' = (A\Psi\Psi(0))$ and so $\Phi(t) \equiv \Psi(t)\Psi(0)^{-1}$ is the fundamental matrix. Recall that there is only one and that it is invertible. Thus $\Psi(t)^{-1}$ exists for all t .

\Rightarrow If $\Psi(t)^{-1}$ exists for all t , then this is true for $t = 0$. ■

The above says that if $\Psi(t)$ is given by 30.9 then $\det(\Psi(t))$ either vanishes for all t or for no t . This determinant is called the Wronskian and this little observation is known as the Wronskian alternative. Also note that the general solution is of the form $\Phi(t)c$ as explained above. If $\Psi(0)$ is invertible, this is $\Psi(t)\Psi(0)^{-1}c$ but a generic c can be written as $\Psi(0)^{-1}\Psi(0)c$ and so the general solution is of the form $\Psi(t)c$ exactly when $\Psi(0)^{-1}$ exists.

Theorem 30.4.3 Consider the equation $Ly \equiv y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y = 0$ and suppose $Ly_k = 0$ for $k = 1, 2, \dots, n$. Then every solution to $Ly = 0$ is of the form $\sum_{k=1}^n c_k y_k$ if and only if $W(y_1, \dots, y_n)(t) \neq 0$ for some t . If this Wronskian condition holds for some t , then it holds for all t . That is, the Wronskian vanishes identically or never.

In the case that $W(y_1, \dots, y_n)(t) \neq 0$ for some t , we say that the general solution to $Ly = 0$ consists of expressions of the form $\sum_{k=1}^n c_k y_k$.

A useful way to recognize that you have the general solution in the case of second order equations is as follows.

Proposition 30.4.4 Suppose y_i , $i = 1, 2$ is a solution to

$$y'' + p(t)y' + q(t)y = 0$$

then the general solution to the equation is of the form

$$\{c_1 y_1 + c_2 y_2, c_1, c_2 \in \mathbb{R}\}$$

if and only if y_1/y_2 is nonconstant.

Proof: By the quotient rule,

$$\frac{d}{dt} \left(\frac{y_2}{y_1} \right) = \frac{y_2' y_1 - y_1' y_2}{(y_1)^2} = \frac{W(y_1, y_2)(t)}{y_1^2(t)}$$

and so the Wronskian is nonzero at some point if and only if y_2/y_1 is not constant so that the derivative of the quotient is not zero. ■

Chapter 31

Numerical Solutions For Systems

You usually can't factor the characteristic polynomial and so you usually can't find explicit solutions to the system

$$\mathbf{x}'(t) = A\mathbf{x}(t) + \mathbf{f}(t), \mathbf{x}(0) = \mathbf{x}_0$$

Another serious difficulty is the case where \mathbf{f} depends not just on t but also on \mathbf{x} . This is the case of nonlinear equations.

This is really just a more complicated problem than finding the integral when you are unable to find an antiderivative in terms of known functions. In the simpler case of finding integrals, there are numerical methods for determining the integral. It is no different in the case of systems of ordinary differential equations.

31.1 A Few Numerical Methods

One way to obtain an approximate solution to a system of equations

$$\mathbf{y}' = \mathbf{F}(t, \mathbf{y}), \mathbf{y}(0) = \mathbf{y}_0$$

would be to replace the derivative with a difference quotient as follows:

$$\frac{\mathbf{y}_i - \mathbf{y}_{i-1}}{h} = \mathbf{F}(t_i, \mathbf{y}_i), i \geq 1, \mathbf{y}_0 = \mathbf{x}_0$$

where here you have a uniform partition of $[0, a]$, $t_0 < t_1 < \dots < t_n = a$ where $h = t_i - t_{i-1}$ and \mathbf{x}_0 is the given initial condition. I will leave it to you to see that there is a solution to the above discrete problem. Then if you want, you could define a function $\mathbf{y}_n(t)$ to be a piecewise linear function which equals \mathbf{y}_i at t_i . This is an example of a numerical solution. The above method is called the Euler method. It isn't as good as what a computer algebra system will use. No one who is serious about getting numerical solutions will use this method. A slightly better method is the improved Euler method.

PROCEDURE 31.1.1 *To find a numerical solution to*

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \mathbf{y}(0) = \mathbf{y}_0$$

using the improved Euler method, do the following:

$$k_1 \equiv f(t_k, y_k), k_2 \equiv f(t_k + h, y_k + k_1 h), t_k = kh, y_{k+1} = y_k + \frac{h}{2}(k_1 + k_2)$$

then there is some constant C such that $|y(kh) - y_k| < Ch^2$.

It predicts what the slope should be at a point and then averages the two values to advance another step.

This problem of getting solutions to first order systems of differential equations has been studied extensively and a book like this is not the place to see a careful description of the best methods. However, one of the very best was developed long before computers by Runge and Kutta in 1901.

PROCEDURE 31.1.2 To find a numerical solution to

$$y' = f(t, y), y(0) = y_0$$

using the Runge-Kutta algorithm, do the following: For $t_k = kh, k = 0, \dots$,

$$\begin{aligned} k_1 &\equiv f(t_k, y_k), k_2 \equiv f\left(t_k + \frac{h}{2}, y_k + k_1 \frac{h}{2}\right), \\ k_3 &\equiv f\left(t_k + \frac{h}{2}, y_k + k_2 \frac{h}{2}\right), k_4 \equiv f(t_k + h, y_k + k_3 h) \\ y_{k+1} &= y_k + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \end{aligned}$$

then there is some constant C such that $|y(kh) - y_k| < Ch^4$.

You can have MATLAB use the Runge-Kutta algorithm to numerically find a solution to a system of ordinary differential equations. I will illustrate with a first order system which comes from the Van der Pol equation. The exact equation studied is not too important at this point. My intent is to illustrate the syntax used. Here it is:

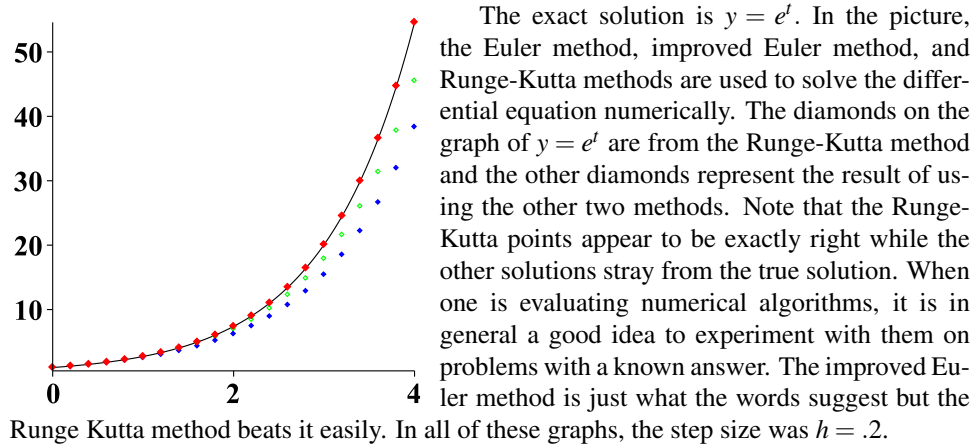
```
f=@(t,x)[x(2),-((x(1)^2-1)*x(2)+x(1))]; n=300; h=.05;
y(1,:)= [1,0]; t(1)=0;
hold on
for r=1:n
k1=f(t(r),y(r,:)); k2=f(t(r)+h/2,y(r,:)+k1*(h/2));
k3=f(t(r)+h/2,y(r,:)+k2*(h/2)); k4=f(t(r)+h,y(r,:)+k3*h);
y(r+1,:)=y(r,:)+(h/6)*(k1+2*k2+2*k3+k4); t(r+1)=t(r)+h;
end
plot(t,y(:,1),t,y(:,2))
interp1(t,[y(:,1),y(:,2)],[2,3,4,5,6])
```

Then press “enter”. This will compute the solution to the differential equation

$$\begin{pmatrix} x \\ y \end{pmatrix}' = \begin{pmatrix} y \\ -((x^2 - 1)y + x) \end{pmatrix}, \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

It will graph both components as functions of t , and it will give you a table of values at the points $t = 2, 3, 4, 5, 6$.

To illustrate how the Runge Kutta algorithm works in comparison to the other two, consider the initial value problem $y' = y$, $y(0) = 1$. Then the three methods give the following graphs.



31.2 Using MATLAB to Find Solutions

A computer algebra system will use the best algorithms whenever possible.

$$\begin{pmatrix} x \\ y \end{pmatrix}' = \begin{pmatrix} t^2 & \sin(x) \\ t+1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \quad \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Then you would type in the following:

```
>>f=@(t,y)[-sin(y(2))*y(2)-y(2);(t+1)*y(1)+2*y(2)];
[t,x]=ode45(f,[0:.05:2],[0;1]); plot(t,x)
```

The vector $\begin{pmatrix} x & y \end{pmatrix}^T$ is denoted as $y = \begin{pmatrix} y(1) & y(2) \end{pmatrix}^T$. Then press “enter” and it will graph these functions on $[0, 2]$. You should see two graphs, one for $x(t)$ and one for $y(t)$. The first is 0 when $t = 0$ and the second is 1 when $t = 0$. In the second line, .05 is the minimum step size for t . You can change this is you like.

If you want a table, you type in

```
>> s=ode45(f,[0,2],[0;1]);
deval(s,[0,.2,.4,.6,.8,1,1.2,1.4,1.6,1.8,2])
```

Then when you press “enter”, you get a table of column vectors which give the values of the vector $\begin{pmatrix} y(1) \\ y(2) \end{pmatrix}$ at the specified values of t . You can also click on the data cursor icon on the top of the graph. Then place the little cross on a point of the curve which interests you and left click. It will display the ordered pair on this point, a value for t and one for

either $y(1)$ or $y(2)$. In the line which has `ode45`, if you type `ode45(f,[0,2],[0;1])`, then MATLAB will decide on the step size for you.

Then, when you have what you want, you ought to type “clear all” and then “enter” and then type `clf` and then “enter” to get rid of any figures. This is so you can do something else without closing MATLAB and starting it over again. MATLAB remembers the functions which have been defined and so unless you do this, it may think you are referring to something other than what you want if you do another computation without closing it down.

You are not limited to systems which have two variables. For example, suppose you wanted to get a solution to

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}' = \begin{pmatrix} -x^3 + x - y \\ z - y + \sin(z) \\ x \end{pmatrix}, \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix}(0) = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

Then you would enter something like the following.

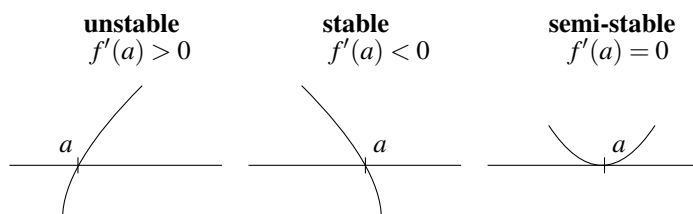
```
>> f=@(t,y)[-y(1)^3+y(1)-y(2);y(3)-y(2)+sin(y(3));y(1)];
      [t,x]=ode45(f,[0:.05:2],[0;1;1]);plot(t,x)
```

Of course the solution to the initial value problem is a space curve. Suppose you wanted to see the graph of this space curve. Try this

```
>> f=@(t,y)[-y(1)^3+y(1)-sin(y(2));y(3)-y(2)+sin(y(3));y(1)];
      [t,x]=ode45(f,[0:.05:30],[0;-1;1]);
      plot3(x(:,1),x(:,2),x(:,3),'LineWidth',2)
```

31.3 Stability of Equilibrium Points

Recall the simple case discussed earlier in which you are considering $y' = f(y)$, $y(a) = 0$. This is summarized in the following picture



The stability is determined by the sign of $f'(a)$. However, if $f'(a) = 0$, then the equilibrium point can be stable from one side and not from the other. In the example of the above, it is stable from the left and unstable from the right. This is because if y is close to a but less than a , the derivative y' is positive so the solution to the differential equation increases. If y is close to a but larger than a , then $y' < 0$ and so the solution moves away from a . Similar considerations show why the other two claims are so, stable if $f'(a) < 0$ and unstable if $f'(a) > 0$.

However, the situation is even more complicated when $f'(a) = 0$. You could have f be strictly increasing through $(a, 0)$ in which case, you would have that the equilibrium

point is unstable. Think $f(y) = (y - a)^3$. Then f has a 0 derivative at a but is increasing. Similarly, you could have f decreasing through $(a, 0)$ in which case, the equilibrium point would be stable. The point is, anything can happen when $f'(a) = 0$.

You should regard $f'(a)$ as an eigenvalue for the linear map $x \rightarrow f'(a)x$. The eigenvalue is negative implies stability. The eigenvalue is positive implies not stable. The eigenvalue is 0 means anything can happen.

The situation is completely similar for nonlinear systems of equations.

Definition 31.3.1 Consider $\mathbf{y}' = \mathbf{f}(\mathbf{y})$, where we always assume \mathbf{f} is C^1 . A point \mathbf{a} is called an equilibrium point when $\mathbf{f}(\mathbf{a}) = \mathbf{0}$.

From the notion of differentiability,

$$\mathbf{f}(\mathbf{a} + \mathbf{y}) = \mathbf{0} + D\mathbf{f}(\mathbf{a})\mathbf{y} + o(\mathbf{y})$$

The situation which generalizes what happens with functions of one variable is as follows. Let \mathbf{a} be an equilibrium point for the differential equation $\mathbf{y}' = \mathbf{f}(\mathbf{y})$. Thus $\mathbf{f}(\mathbf{a}) = \mathbf{0}$.

eigenvalues of $D\mathbf{f}(\mathbf{a})$ have negative real parts	equilibrium is stable
some eigenvalue of $D\mathbf{f}(\mathbf{a})$ has positive real part	equilibrium is unstable
some eigenvalue of $D\mathbf{f}(\mathbf{a})$ has zero real part	you have no idea

So what exactly is meant by stable? It is the same as in the case of scalar valued equations.

Definition 31.3.2 An equilibrium point \mathbf{a} for $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ is stable if whenever the initial condition \mathbf{y}_0 is sufficiently close to \mathbf{a} , it follows that the solution to the initial value problem $\mathbf{y}' = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(0) = \mathbf{y}_0$ will stay close to \mathbf{a} . Also, \mathbf{a} is asymptotically stable if whenever \mathbf{y}_0 is close enough to \mathbf{a} , then the solution of the initial value problem just described converges to \mathbf{a} as $t \rightarrow \infty$.

In fact, one has a little more in case all eigenvalues are negative.

eigenvalues of $D\mathbf{f}(\mathbf{a})$ have negative real parts	equilibrium is asymptotically stable
some eigenvalue of $D\mathbf{f}(\mathbf{a})$ has positive real part	equilibrium is unstable
some eigenvalue of $D\mathbf{f}(\mathbf{a})$ has zero real part	you have no idea

Example 31.3.3 Consider the following system of equations for which $\begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^T$ is an equilibrium point. Determine whether this is a stable equilibrium.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}' = \begin{pmatrix} xy - 12z - 5x \\ y^2 - y - 2x - 6z \\ xy^2 + x + 2z \end{pmatrix}$$

You need to get the derivative of the right side. The matrix of this is

$$\begin{pmatrix} y-5 & x & -12 \\ -2 & 2y-1 & -6 \\ y^2+1 & 2xy & 2 \end{pmatrix}$$

At the equilibrium point, you get

$$\begin{pmatrix} -5 & 0 & -12 \\ -2 & -1 & -6 \\ 1 & 0 & 2 \end{pmatrix}$$

The eigenvalues are $-1, -2, -1$ and so this equilibrium point is stable.

Example 31.3.4 The point $(1, 0, 0)^T$ is an equilibrium point of the following system

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}' = \begin{pmatrix} 11x + 11y - 12z + xy - 11 \\ 6z - 7y - 6x + yz + 6 \\ 4x + 4y - 6z + xz - 4 \end{pmatrix}$$

Determine whether this point is stable.

You need to find the derivative. It equals

$$\begin{pmatrix} y+11 & x+11 & -12 \\ -6 & z-7 & y+6 \\ z+4 & 4 & x-6 \end{pmatrix}$$

At the equilibrium point you get

$$\begin{pmatrix} 11 & 12 & -12 \\ -6 & -7 & 6 \\ 4 & 4 & -5 \end{pmatrix}$$

Now you consider the eigenvalues for this matrix. In this case, there is a positive eigenvalue and so the equilibrium point is unstable.

Of course there is a problem with this. How do you find the sign of the eigenvalues. You don't need to know the eigenvalues exactly, just their signs. However, MATLAB can tell you the approximate eigenvalues. To find them in this case, you do the following.

$$A=[11 \ 12 \ -12;-6 \ -7 \ 6;4 \ 4 \ -5];\text{eig}(A)$$

You enter the rows starting with the top row and then the next and so forth. You type the numbers from left to right leaving a space between numbers or you can put a comma between them. When you start a new row, you tell MATLAB this is the case by placing ; there. Then type eig(A) and press enter. It will give you the eigenvalues.

Example 31.3.5 $\begin{pmatrix} 1 & 1 & 0 \end{pmatrix}^T$ is an equilibrium point for the differential equation

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}' = \begin{pmatrix} 7x + 14y - 3z + xy - 22 \\ z - 9y - 5x + yz + 14 \\ z^2 - 2z + 3x + 5y - 8 \end{pmatrix}$$

Determine its stability.

Find the derivative.

$$\begin{pmatrix} y+7 & x+14 & -3 \\ -5 & z-9 & y+1 \\ 3 & 5 & 2z-2 \end{pmatrix}$$

Now find this at the equilibrium point.

$$\begin{pmatrix} 8 & 15 & -3 \\ -5 & -9 & 2 \\ 3 & 5 & -2 \end{pmatrix}$$

Next you need to consider the real parts of the eigenvalues. Use MATLAB. This gives the eigenvalues are $-1 + i$, $-1 - i$, and -1 so they have negative real parts and this shows that the equilibrium point is stable.

When the eigenvalues include one which has real part equal to 0 and none of them having positive part larger than 0, then you really don't know much. Nevertheless, there is a way to consider this case also, but I do not plan to include it in this book. It involves something called the center manifold and understanding it properly requires a little too much hard mathematics. However, it is also true that in many cases of interest, the system takes place in the plane and in this case, you can often figure out what is happening by simply having MATLAB graph the space curves resulting from various initial conditions.

For example, consider

$$\begin{pmatrix} x \\ y \end{pmatrix}' = \begin{pmatrix} -4y^2 \\ 2x \end{pmatrix}$$

If you graph the space curves which result from many different small initial conditions, you will see that $(0,0)$ is a stable point although not asymptotically stable.

```
hold on
r=.1; f=@(t,x)[-4*x(2)^3;2*x(1)];
for n=1:10
[t,x]=ode45(f,[0,40],[0,n*.1]);
plot(x(:,1),x(:,2),'LineWidth',1.3)
end
```

You could modify this just a little and find a situation where $(0,0)$ is asymptotically stable.

$$\begin{pmatrix} x \\ y \end{pmatrix}' = \begin{pmatrix} -4y^2 \\ 2x + .1y \end{pmatrix}$$

```
hold on
r=.1; f=@(t,x)[-4*x(2)^3;2*x(1)+.1*x(2)];
for n=1:3
[t,x]=ode45(f,[0,40],[0,n*.1]);
plot(x(:,1),x(:,2),'LineWidth',1.3)
end
```

If you solve numerically and graph the solution using the above syntax, you will see the solution spiral in towards $(0,0)$.

A great deal more can be said concerning stability and more generally the geometric behavior of solutions to ordinary differential equations, especially for systems which have

solutions in the plane. In the next section are some major results about these things. For much more see a text on ordinary differential equations. My book has a good deal more discussion. See [25].

31.4 Periodic Orbits, Poincare Bendixon Theorem

The fundamental result in this subject, at least in the plane, is the Poincare Bendixon theorem.¹

Definition 31.4.1 *A periodic orbit is a set of points*

$$\{\mathbf{x}(t, \mathbf{x}_0), t \geq 0\}$$

such that for some $T > 0$, $\mathbf{x}(t+T, \mathbf{x}_0) = \mathbf{x}(t, \mathbf{x}_0)$ for all $t \geq 0$. The number T is called a period. Thus the point $\mathbf{x}(t, \mathbf{x}_0)$ goes around and around always returning to the point from where it started.

Now the following is the Poincare Bendixon theorem which gives existence of periodic orbits in the plane.

Theorem 31.4.2 *Let D be the closure of a bounded region of the plane such that \mathbf{f} is a C^1 function which has no zeros in D , and suppose that $\mathbf{x}(t, \mathbf{x}_0)$ stays in D for all $t \geq 0$ if $\mathbf{x}_0 \in D$, where this is the solution to*

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}), \mathbf{x}(0) = \mathbf{x}_0$$

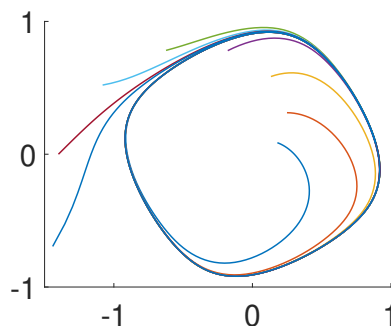
Then letting $\Lambda_+ = \cup_{t \geq 0} \mathbf{x}(t, \mathbf{x}_0)$, it follows that Λ_+ is either a periodic orbit or $t \rightarrow \mathbf{x}(t, \mathbf{x}_0)$ spirals in toward a periodic orbit.

It is a plausible result. Say you have that every initial condition which starts off in a bounded closed set stays in that set and there are no equilibrium points. Thus $t \rightarrow \mathbf{x}(t, \mathbf{x}_0)$ just keeps moving. Then from this theorem, there must be a periodic orbit somewhere such that either this function traces out a periodic orbit or it gets close to one. For example, consider the system

$$\begin{pmatrix} x \\ y \end{pmatrix}' = \begin{pmatrix} x + y - x(x^2 + 2y^2) \\ -x + y - y(2x^2 + y^2) \end{pmatrix}$$

From looking at the eigenvalues of the matrix in the almost linear system, you will see that they are both positive. Hence every solution near $(0,0)$ but not equal to $(0,0)$ must fail to remain near $(0,0)$. Also, you can see from the equations that the solutions cannot get very large because the sign of x' will change to oppose $|x|$ getting large and a similar condition happens for y' . Therefore, there should exist a periodic orbit from the above theorem. The following picture illustrates what happens for various initial conditions. Note how the solutions spiral in toward a periodic orbit.

¹Ivar Otto Bendixson (1861-1935) was a Swedish mathematician. He is most famous for the Poincare Bendixon theorem presented here. He also did work in topology.



Definition 31.4.3 A saddle point \mathbf{x}_0 for $\mathbf{x}' = \mathbf{f}(\mathbf{x})$ is an equilibrium point ($\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$) which is not stable, but which has the property that in every set of the form

$$\{\mathbf{x} : r > |\mathbf{x} - \mathbf{x}_0| > 0\}$$

there are points for which the solution having these as initial conditions converges to \mathbf{x}_0 as $t \rightarrow \infty$. These saddle points occur for example if you have a negative and a positive eigenvalue for $D\mathbf{f}(\mathbf{x}_0)$.

The following very interesting theorem can be obtained from the above.

Theorem 31.4.4 If you have a periodic orbit of a solution to an autonomous two dimensional differential equation, $\mathbf{x}' = \mathbf{f}(\mathbf{x})$, then it must go around some equilibrium point. If there is only one equilibrium point inside the periodic orbit, then it cannot be a saddle point.

31.5 Exercises

1. The Van der Pol equation describes nonlinear oscillations. It is

$$x'' + (x^2 - 1)x' + x = 0 \quad (31.1)$$

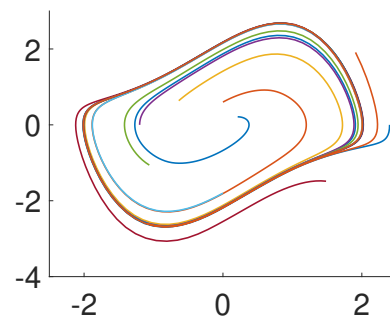
Show that it has a non constant periodic solution. Do as follows. First write as a first order system

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

Have MATLAB or some other computer algebra system give a graph of solutions for the above system corresponding to various initial conditions in a way to show the periodic solution. Try the following:

```
f=@(t,x)[x(2);-((x(1)^2-1)*x(2)+x(1))];
d=pi/4; r=.3;
hold on
for n=1:9
[t,x]=ode45(f,[0:.05:10],[n*r*cos(n*d);n*r*sin(n*d)]);
plot(x(:,1),x(:,2),'LineWidth',1.5)
end
```

Yo should get something like the following.



Chapter 32

Solutions Near a Regular Singular Point

This chapter is on solutions to an equation of the form

$$x^2 y'' + x p(x) y' + q(x) y = 0$$

where $p(x)$ and $q(x)$ can be expressed in terms of a power series centered at 0. Such equations are said to have a regular singular point. This is very different because there is generally no way to write such an equation in the form

$$y'' + p(x) y' + q(x) y = 0$$

where $p(x), q(x)$ are continuous near 0. Thus the initial value problem makes no sense. Thus, none of the above theory applies to these equations and further analysis is needed. Equations of this sort were found to be very important in the nineteenth century for various reasons. For more on these topics, you can see my book [25]. What is here is a subset of the contents of this book. Also, the most important example of this kind of equation is the Bessel equation. Whole books are available on this which will develop more of the theory than presented here or in my differential equations book. See [17].

32.1 The Euler Equations

The simplest equation to illustrate the concept of a regular singular point is the so called Euler equation, sometimes called a Cauchy Euler equation.

Definition 32.1.1 *A differential equation is called an Euler equation if it can be written in the form*

$$x^2 y'' + ax y' + by = 0.$$

Solving a Cauchy Euler equation is really easy. You look for a solution like $y = x^r$ and try to choose r in such a way that it solves the equation. Plugging this in to the above equation,

$$x^2 r(r-1)x^{r-2} + xarx^{r-1} + bx^r = 0$$

This reduces to

$$x^r(r(r-1) + ar + b) = 0$$

and so you have to solve the equation

$$r(r-1) + ar + b = 0$$

to find the values of r . If these values of r are different, say $r_1 \neq r_2$ then the general solution must be

$$C_1x^{r_1} + C_2x^{r_2}$$

because the Wronskian of the two functions will be nonzero. I know this because the ratio of the two functions is not a constant so Proposition 30.4.4 implies this gives the general solution. The reason for this is that the quotient rule gives the numerator as ± 1 times the Wronskian.

Example 32.1.2 Find the general solution to $x^2y'' - 2xy' + 2y = 0$.

You plug in x^r and look for r . Then as above this yields

$$r(r-1) - 2r + 2 = r^2 - 3r + 2 = 0$$

and so the two values of r are 1, 2. Therefore, the general solution to this equation is

$$C_1x + C_2x^2.$$

Of course there are three cases for solutions to the so called indicial equation

$$r(r-1) + ar + b = 0$$

Either the zeros are distinct and real, distinct and complex or repeated. Consider the case where they are distinct and complex next.

Example 32.1.3 Find the general solution to $x^2y'' + 3xy' + 2y = 0$.

This time you have

$$r^2 + 2r + 2 = 0$$

and the solutions are $r = -1 \pm i$. How do we interpret

$$x^{-1+i}, x^{-1-i}?$$

It is real easy. You assume always that $x > 0$ since otherwise the leading coefficient could vanish. Then

$$x^{-1+i} = e^{\ln(x)(-1+i)} = e^{-\ln(x)+i\ln(x)}$$

and by Euler's formula this equals

$$\begin{aligned} x^{-1+i} &= e^{\ln(x^{-1})} (\cos(\ln(x)) + i \sin(\ln(x))) \\ &= \frac{1}{x} (\cos(\ln(x)) + i \sin(\ln(x))) \end{aligned}$$

Corresponding to x^{-1-i} we get something similar.

$$x^{-1-i} = \frac{1}{x} ((\cos(\ln(x)) - i \sin(\ln(x))))$$

Adding these together and dividing by 2 to get the real part, the principle of superposition implies

$$\frac{1}{x} \cos(\ln(x))$$

is a solution. Then subtracting them and dividing by $2i$ you get

$$\frac{1}{x} \sin(\ln(x))$$

is a solution. Hence anything of the form

$$C_1 \frac{1}{x} \cos(\ln(x)) + C_2 \frac{1}{x} \sin(\ln(x))$$

is a solution. Is this the general solution? Of course. This follows because the ratio of the two functions is not constant and this implies their Wronskian is nonzero. See Proposition [30.4.4](#).

In the general case, suppose the solutions of the indicial equation

$$r(r-1) + ar + b = 0 \quad (32.1)$$

are $\alpha \pm i\beta$. Then the general solution for $x > 0$ is

$$C_1 x^\alpha \cos(\beta \ln(x)) + C_2 x^\alpha \sin(\beta \ln(x))$$

Finally consider the case where the zeros of the indicial equation are real and repeated. Note I have included all cases because, since the coefficients of this equation are real, the zeros come in conjugate pairs if they are not real. Suppose then that x^r is a solution of

$$x^2 y'' + axy' + by = 0$$

and that r is a repeated root. By the quadratic formula applied to the indicial equation [32.1](#),

$$r = \frac{-(a-1)}{2} \quad (32.2)$$

Then if $z(x)$ is another solution which is not a multiple of x^r , you would have

$$z(x) = x^r u(x)$$

The plug in to the equation and try to make it work.

$$x^2 (x^r u(x))'' + ax (x^r u(x))' + bx^r u(x) = 0$$

Then

$$x^2 [r(r-1)x^{r-2}u + 2rx^{r-1}u' + x^r u''] + ax(rx^{r-1}u + x^r u') + bx^r u = 0$$

Separate out the terms which multiply u

$$u \left[\overbrace{r(r-1) + ar + b}^{=0} \right] x^r = 0$$

All that is left is

$$x^{r+2}u'' + (2rx^{r+1} + ax^{r+1})u' = 0$$

Thus, using 32.2,

$$x^{r+2}u'' + ((-a+1)x^{r+1} + ax^{r+1})u' = 0$$

Therefore,

$$xu'' + u' = 0$$

and this is a first order linear equation for u' . Thus, a nonzero solution to this is

$$u' = \frac{1}{x}$$

Therefore, if $u = \ln x$, it follows that $z(x) = u(x)x^r$ is a solution to the Euler equation with the repeated roots and so another solution is

$$z = x^r \ln(x)$$

Example 32.1.4 Find the general solution of the equation

$$x^2y'' + 3xy' + y = 0.$$

In this case the indicial equation is

$$r(r-1) + 3r + 1 = r^2 + 2r + 1 = 0$$

and there is a repeated zero, $r = -1$. Therefore, the general solution is

$$y = C_1x^{-1} + C_2 \ln(x)x^{-1}.$$

This is pretty easy isn't it?

How would things be different if the equation was of the form

$$(x-a)^2y'' + a(x-a)y' + by = 0?$$

The answer is that it wouldn't be any different. You could just define a new independent variable $t \equiv (x-a)$ and then the equation in terms of t becomes

$$t^2z'' + atz + bz = 0$$

where $z(t) \equiv y(x) = y(t+a)$. You can always reduce these sorts of equations to the case where the singular point is at 0. However, you might not want to do this. If not, you look for a solution in the form $y = (x-a)^r$, plug in and determine the correct value of r . In the case of real and distinct zeros you get

$$y = C_1(x-a)^{r_1} + C_2(x-a)^{r_2}$$

In the case where $r = \alpha \pm i\beta$ you get

$$y = C_1(x-a)^\alpha \cos(\beta \ln(x-a)) + C_2(x-a)^\alpha \sin(\beta \ln(x-a))$$

for the general solution for $x > a$

In the case where r is a repeated zero, you get

$$y = C_1(x-a)^r + C_2 \ln(x-a)(x-a)^r.$$

32.2 Some Simple Observations on Power Series

This section is a review of a few facts about power series which should have been learned in calculus. If you have not seen these things, which may well be the case given the way calculus courses are systematically watered down, see my book *Calculus of one and many variables* on my web page or my differential equations book [25].

Definition 32.2.1 A function f is analytic in some open set U if for each $a \in U$, $f(x) = \sum_{k=0}^{\infty} a_k (x-a)^k$ for all x close enough to a . In other words, you can get the function near a by a power series.

Theorem 32.2.2 Suppose $f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n$ for x near a and suppose $a_0 \neq 0$. Then

$$f(x)^{-1} = \frac{1}{a_0} + h(x)$$

where $h(x) = \sum_{n=1}^{\infty} b_n (x-a)^n$ so $h(a) = 0$.

Proof: It turns out that $f(x)^{-1}$ has a power series representation near a and so $f(a)^{-1} = 1/a_0$. ■

Theorem 32.2.3 Suppose $f(x) = \sum_{n=0}^{\infty} a_n x^n$ and $g(x) = \sum_{n=0}^{\infty} b_n x^n$ for x near 0. Then $f(x)g(x)$ also has a power series near 0 and in fact,

$$f(x)g(x) = \sum_{n=0}^{\infty} \left(\sum_{k=0}^n a_{n-k} b_k \right) x^n. \quad (32.3)$$

Proof: See the material on power series in my calculus book. However, it is quite plausible.

$$\begin{aligned} & \left(\sum_{n=0}^{\infty} a_n x^n \right) \left(\sum_{n=0}^{\infty} b_n x^n \right) \\ &= (a_0 + a_1 x + a_2 x^2 + \cdots) (b_0 + b_1 x + b_2 x^2 + \cdots) \end{aligned}$$

Now formally multiply the two power series like they were polynomials and collect terms. This will yield 32.3. ■

32.3 Regular Singular Points

First of all, here is the definition of what a regular singular point is.

Definition 32.3.1 A differential equation has a regular singular point at 0 if the equation can be written in the form

$$x^2 y'' + x b(x) y' + c(x) y = 0 \quad (32.4)$$

where

$$b(x) = \sum_{n=0}^{\infty} b_n x^n, \quad c(x) = \sum_{n=0}^{\infty} c_n x^n$$

for all x near 0. Such functions are called analytic in this section. More generally, a differential equation

$$P(x)y'' + Q(x)y' + R(x)y = 0 \quad (32.5)$$

where P, Q, R are analytic near a has a regular singular point at a if it can be written in the form

$$(x-a)^2 y'' + (x-a)b(x)y' + c(x)y = 0 \quad (32.6)$$

where

$$b(x) = \sum_{n=0}^{\infty} b_n (x-a)^n, \quad \sum_{n=0}^{\infty} c_n (x-a)^n = c(x)$$

for all $|x-a|$ small enough. The equation 32.5 has a singular point at a if $P(a) = 0$.

The following table emphasizes the similarities between the Euler equations and the regular singular point equations. I have featured the point 0. If you are interested in another point a , you just replace x with $x-a$ everywhere it occurs.

	Euler equation	regular singular point
form of equation	$x^2 y'' + x b_0 y' + c_0 y = 0$	$x^2 y'' + x(b_0 + b_1 x + \cdots) y' + (c_0 + c_1 x + \cdots) y = 0$
indicial equation	$r(r-1) + b_0 r + c_0 = 0$	$r(r-1) + b_0 r + c_0 = 0$
one solution	$y = x^r$	$y = x^r \sum_{k=0}^{\infty} a_k x^k, a_0 = 1.$

Recognizing Regular Singular Points

How do you know a singular differential equation can be written a certain way? In particular, how can you recognize a regular singular point when you see one? Suppose

$$P(x)y'' + Q(x)y' + R(x)y = 0$$

where all of P, Q, R are analytic functions near a . How can you tell if it has a regular singular point at a ? Here is how. It has a regular singular point at a if

$$\lim_{x \rightarrow a} (x-a) \frac{Q(x)}{P(x)} \text{ exists}$$

$$\lim_{x \rightarrow a} (x-a)^2 \frac{R(x)}{P(x)} \text{ exists}$$

If these conditions hold, then by theorems in complex analysis it will be the case that

$$(x-a) \frac{Q(x)}{P(x)} = \sum_{n=0}^{\infty} b_n (x-a)^n,$$

and

$$(x-a)^2 \frac{R(x)}{P(x)} = \sum_{n=0}^{\infty} c_n (x-a)^n$$

for x near a . Indeed, equations of this form reduce to the form in 32.6 upon dividing by $P(x)$ and multiplying by $(x-a)^2$.

Example 32.3.2 Find the regular singular points of the equation and find the singular points.

$$x^3(x-2)^2(x-1)^2y'' + (x-2)\sin(x)y' + (1+x)y = 0$$

The singular points are 0, 2, 1. Let's consider 0 first.

$$\lim_{x \rightarrow 0} x \frac{(x-2)\sin(x)}{x^3(x-2)^2(x-1)^2}$$

does not exist. Therefore, 0 is not a regular singular point. I don't have to check any further. Now consider the singular point 2.

$$\lim_{x \rightarrow 2} (x-2) \frac{(x-2)\sin(x)}{x^3(x-2)^2(x-1)^2} = \frac{1}{8} \sin 2$$

and

$$\lim_{x \rightarrow 2} (x-2)^2 \frac{1+x}{x^3(x-2)^2(x-1)^2} = \frac{3}{8}$$

and so yes, 2 is a regular singular point. Now consider 1.

$$\lim_{x \rightarrow 1} (x-1) \frac{(x-2)\sin(x)}{x^3(x-2)^2(x-1)^2}$$

does not exist so 1 is not a regular singular point. Thus the above equation has only one regular singular point and this is where $x = 2$.

Example 32.3.3 Find the regular singular points of

$$x \sin(x)y'' + 3 \tan(x)y' + 2y = 0$$

The singular points are $0, n\pi$ where n is an integer. Let's consider a point at $n\pi$ where $n \neq 0$. To be specific, let's let $n = 3$

$$\lim_{x \rightarrow 3\pi} (x-3\pi) \frac{3 \tan(x)}{x \sin(x)} = 0$$

Similarly the limit exists for other values of n . Now consider

$$\lim_{x \rightarrow 3\pi} (x-3\pi)^2 \frac{2}{x \sin(x)} = 0$$

Similarly the limit exists for other values of n . What about 0?

$$\lim_{x \rightarrow 0} x \frac{3 \tan(x)}{x \sin(x)} = 3$$

and

$$\lim_{x \rightarrow 0} x^2 \frac{2}{x \sin(x)} = 2$$

so it appears all these singular points are regular singular points.

Example 32.3.4 Find the regular singular points of

$$x^2 \sin(x) y'' + 3 \tan(x) y' + 2y = 0$$

Let's look at $x = 0$ first. The equation has the same singular points.

$$\lim_{x \rightarrow 0} x \frac{3 \tan(x)}{x^2 \sin(x)} = \text{undefined}$$

so 0 is not a regular singular point.

$$\lim_{x \rightarrow 3\pi} (x - 3\pi) \frac{3 \tan(x)}{x^2 \sin(x)} = 0$$

and the situation is similar for other singular points $n\pi$. Also

$$\lim_{x \rightarrow 3\pi} (x - 3\pi)^2 \frac{2}{x^2 \sin(x)} = 0$$

with similar result for arbitrary $n\pi$ where $n \neq 0$. Thus in this case 0 is not a regular singular point but $n\pi$ is a regular singular point for all integers $n \neq 0$.

In general, if you have an equation which has a regular singular point at a so that the equation can be massaged to give something of the form

$$(x - a)^2 y'' + (x - a) b(x) y' + c(x) y = 0$$

you could always define a new variable $t \equiv (x - a)$ and letting $z(t) = y(x)$, you could rewrite the equation in terms of t in the form

$$t^2 z'' + t b(a + t) z' + c(a + t) z = 0$$

and thereby reduce to the case where the regular singular point is at 0. Thus there is no loss of generality in concentrating on the case where the regular singular point is at 0. In addition, the most important examples are like this. Therefore, from now on, I will consider this case. This just means you have all the series in terms of powers of x rather than the more general powers of $x - a$.

32.4 Abel's Formula

Suppose you have a differential equation

$$y'' + p(x) y' + q(x) y = 0$$

and you have two solutions to it y_1, y_2 . Abel's formula is a lovely little identity for the Wronskian of these two functions. It gives another way to show that the Wronskian either vanishes identically or not at all. From the equation,

$$\begin{aligned} y_2 y_1'' + p(x) y_2 y_1' + q(x) y_2 y_1 &= 0 \\ y_1 y_2'' + p(x) y_1 y_2' + q(x) y_1 y_2 &= 0 \end{aligned}$$

Now subtract.

$$(y_1 y_2'' - y_2 y_1'') + p(x) (y_1 y_2' - y_2 y_1') = 0 \quad (32.7)$$

From the product rule,

$$(y_1 y_2' - y_2 y_1')' = y_2'' y_1 + y_1' y_2' - (y_1'' y_2 + y_1' y_2') = (y_1 y_2'' - y_2 y_1'')$$

but $y_1 y_2' - y_2 y_1' = W(y_1, y_2)$. Hence 32.7 is of the form

$$W' + p(x)W = 0$$

and so, from the theory of linear equations,

$$W(x) = C e^{-P(x)} \text{ where } P'(x) = p(x).$$

This proves Abel's formula.

Proposition 32.4.1 *Let y_1, y_2 be two solutions to $y'' + p(x)y' + q(x)y = 0$ for x in some interval on which $p(x), q(x)$ are continuous. Then*

$$W(y_1, y_2)(x) = C e^{-P(x)}, \quad P'(x) = p(x).$$

Note how this shows directly that the Wronskian either vanishes identically or not at all. This also motivates the following procedure.

PROCEDURE 32.4.2 *Suppose y is a known solution to*

$$y'' + p(x)y' + q(x)y = 0 \tag{32.8}$$

To find another solution z which solves

$$z'' + p(x)z' + q(x)z = 0$$

do the following: For

$$W = \begin{vmatrix} y & z \\ y' & z' \end{vmatrix}$$

Find a nonzero solution $W(x)$ of

$$W'(x) + p(x)W(x) = 0$$

Then solve for z in the equation

$$z'y - zy' = W$$

BE SURE THAT THE EQUATION IS IN THE FORM DESCRIBED IN THE ABOVE PROCEDURE. THIS MEANS THE COEFFICIENT OF y'' IS 1!

32.5 Finding the Solution

Suppose you have reduced the equation to

$$x^2 y'' + x p(x) y' + q(x) y = 0 \tag{32.9}$$

where each of p, q is analytic near 0. Then letting

$$p(x) = b_0 + b_1 x + \cdots$$

$$q(x) = c_0 + c_1x + \cdots$$

you see that for small x the equation should be approximately equal to

$$x^2y'' + xb_0y' + c_0y = 0$$

which is an Euler equation. This would have a solution in the form x^r where

$$r(r-1) + b_0r + c_0 = 0,$$

the indicial equation for the Euler equation, and so it is not unreasonable to look for a solution to the equation in 32.9 which is of the form The values of r are called the exponents of the singularity.

$$x^r \sum_{k=0}^{\infty} a_k x^k, \quad a_0 \neq 0.$$

You perturb the coefficients of the Euler equation to get 32.9 and so it is not unreasonable to think you should look for a solution to 32.9 of the above form.

Example 32.5.1 Find the general solution to the equation

$$x^2y'' + x(1+x^2)y' - 2y = 0.$$

The associated Euler equation is of the form

$$x^2y'' + xy' - 2y = 0$$

and so the indicial equation is

$$r(r-1) + r - 2 = 0 \tag{32.10}$$

so $r = \sqrt{2}, r = -\sqrt{2}$. Then you would look for a solution in the form

$$y = x^r \sum_{k=0}^{\infty} a_k x^k = \sum_{k=0}^{\infty} a_k x^{k+r}$$

where $r = \pm\sqrt{2}$. Plug in to the equation.

$$\begin{aligned} & x^2 \sum_{k=0}^{\infty} a_k (k+r)(k+r-1) x^{k+r-2} \\ & + x(1+x^2) \sum_{k=0}^{\infty} a_k (k+r) x^{k+r-1} - 2 \sum_{k=0}^{\infty} a_k x^{k+r} = 0 \end{aligned}$$

This simplifies to

$$\begin{aligned} & \sum_{k=0}^{\infty} a_k (k+r)(k+r-1) x^{k+r} + \sum_{k=0}^{\infty} a_k (k+r) x^{k+r} \\ & + \sum_{k=0}^{\infty} a_k (k+r) x^{k+r+2} - 2 \sum_{k=0}^{\infty} a_k x^{k+r} = 0 \end{aligned} \tag{32.11}$$

The lowest order term is the x^r term and it yields

$$a_0(r)(r-1) + a_0(r) - 2a_0 = 0$$

but this is just $a_0(r(r-1)+r-2) = 0$. Since r is one of the zeros of 32.10, there is no restriction on the choice of a_0 . In fact, as discussed below, this lack of a requirement on a_0 is equivalent to finding the right value of r . Next consider the x^{r+1} terms. There are no such terms in the third of the above sums just as there were no x^r terms in this sum. Then

$$a_1((1+r)(r) + (1+r) - 2) = 0$$

Now if r solves 32.10 then $1+r$ does not do so because the two solutions to this equation do not differ by an integer. Therefore, the above equation requires $a_1 = 0$. At this point we can give a recurrence relation for the other a_k . To do this, change the variable of summation in the third sum of 32.11 to obtain

$$\begin{aligned} \sum_{k=0}^{\infty} a_k(k+r)(k+r-1)x^{k+r} + \sum_{k=0}^{\infty} a_k(k+r)x^{k+r} \\ + \sum_{k=2}^{\infty} a_{k-2}(k-2+r)x^{k+r} - 2 \sum_{k=0}^{\infty} a_k x^{k+r} = 0 \end{aligned}$$

Thus for $k \geq 2$,

$$a_k[(k+r)(k+r-1) + (k+r) - 2] + a_{k-2}(k-2+r) = 0$$

Hence for $k \geq 2$,

$$a_k = \frac{-a_{k-2}(k-2+r)}{[(k+r)(k+r-1) + (k+r) - 2]} = \frac{-a_{k-2}(k-2+r)}{[(k+r)(k+r-1) + (k+r) - 2]}$$

and we take $a_0 \neq 0$ while $a_1 = 0$. Now let's find the first several terms of two independent solutions, one for $r = \sqrt{2}$ and the other for $r = -\sqrt{2}$. Let $a_0 = 1$ for simplicity. Then the above recurrence relation shows that since $a_1 = 0$ all the odd terms equal 0. Also

$$a_2 = \frac{-r}{[(2+r)(2+r-1) + (2+r) - 2]} = -\frac{r}{[(2+r)(1+r) + r]}$$

while

$$a_4 = \frac{-\left(-\frac{r}{[(2+r)(1+r) + r]}\right)(4-2+r)}{[(4+r)(4+r-1) + (4+r) - 2]} = \frac{r}{[2+4r+r^2]} \frac{2+r}{[14+8r+r^2]}$$

Continuing this way, you can get as many terms as you want. Now let's put in the two values of r to obtain the beginning of the two solutions. First let $r = \sqrt{2}$

$$\begin{aligned} y_1(x) = x^{\sqrt{2}} \left(1 + \left(-\frac{\sqrt{2}}{[(2+\sqrt{2})(\sqrt{2}+1) + \sqrt{2}]} \right) x^2 + \right. \\ \left. + \left(\frac{\sqrt{2}}{[4+4\sqrt{2}]} \frac{2+\sqrt{2}}{[16+8\sqrt{2}]} \right) x^4 \dots \right) \end{aligned}$$

the solution which corresponds to $r = -\sqrt{2}$ is

$$y_2(x) = x^{-\sqrt{2}} \left(1 + \left(\frac{\sqrt{2}}{[(2-\sqrt{2})(1-\sqrt{2}) - \sqrt{2}]} \right) x^2 + \right.$$

$$\sqrt{2} \frac{-2 + \sqrt{2}}{[4 - 4\sqrt{2}][16 - 8\sqrt{2}]} x^4 + \dots$$

Then the general solution is

$$C_1 y_1 + C_2 y_2$$

and this is valid for $x > 0$. Note that the ratio of the two solutions is not a constant so this is indeed the general solution.

Generalities

For an equation

$$x^2 y'' + x p(x) y' + q(x) y = 0$$

having a regular singular point at 0, one looks for solutions in the form

$$y(x) = \sum_{n=0}^{\infty} a_n x^{r+n} \quad (32.12)$$

where r is a constant which is to be determined, in such a way that $a_0 \neq 0$. It turns out that such equations **always** have such solutions although solutions of this sort are not always enough to obtain the general solution to the equation. The constant r is called the exponent of the singularity because the solution is of the form

$$x^r a_0 + \text{higher order terms.}$$

Thus the behavior of the solution to the equation given above is like x^r for x near the singularity, 0.

If you require that 32.12 solves 32.9 and plug in, you obtain using Theorem 32.2.3

$$\begin{aligned} \sum_{n=0}^{\infty} (r+n)(r+n-1) a_n x^{n+r} + \sum_{n=0}^{\infty} \left(\sum_{k=0}^n a_k (k+r) b_{n-k} \right) x^{n+r} \\ + \sum_{n=0}^{\infty} \left(\sum_{k=0}^n c_{n-k} a_k \right) x^{n+r} = 0. \end{aligned} \quad (32.13)$$

Since $a_0 \neq 0$,

$$p(r) \equiv r(r-1) + b_0 r + c_0 = 0 \quad (32.14)$$

and this is called the indicial equation. (Note it is the indicial equation for the Euler equation which comes from deleting all the nonconstant terms in the power series for $p(x)$ and $q(x)$.) Also the following equation must hold for $n = 1, \dots$

$$p(n+r) a_n = - \sum_{k=0}^{n-1} a_k (k+r) b_{n-k} - \sum_{k=0}^{n-1} c_{n-k} a_k \equiv f_n(a_i, b_i, c_i) \quad (32.15)$$

These equations are all obtained by setting the coefficient of x^{n+r} equal to 0.

There are various cases depending on the nature of the solutions to this indicial equation. I will always assume the zeros are real, but will consider the case when the zeros are distinct and do not differ by an integer and the case when the zeros differ by a non negative integer.

It turns out that the nature of the problem changes according to which of these cases holds. You can see why this is the case by looking at the equations 32.14 and 32.15. If r_1, r_2 solve the indicial equation and $r_1 - r_2 \neq$ an integer, then with r in equation 32.15 replaced by either r_1 or r_2 , for $n = 1, \dots$, $p(n+r) \neq 0$ and so there is a unique solution to 32.15 for each $n \geq 1$ once $a_0 \neq 0$ has been chosen. Therefore, in this case that $r_1 - r_2 \neq$ an integer, equation 32.4 has a general solution in the form

$$C_1 \sum_{n=0}^{\infty} a_n x^{n+r_1} + C_2 \sum_{n=0}^{\infty} b_n x^{n+r_2}, a_0, b_0 \neq 0.$$

It is obvious this is the general solution because the ratio of the two solutions is non constant. As pointed out earlier, this requires their Wronskian to be nonzero.

On the other hand, if $r_1 - r_2 =$ an integer, then there exists a unique solution to 32.15 for each $n \geq 1$ if r is replaced by the larger of the two zeros r_1 . Therefore, in this case there is always a solution of the form

$$y_1(x) = \sum_{n=0}^{\infty} a_n x^{n+r_1}, a_0 = 1, \quad (32.16)$$

but you might very well hit a snag when you attempt to find a solution of this form with r_1 replaced with the smaller of the two zeros r_2 due to the possibility that for some $m \geq 1$, $p(m+r_2) = p(r_1) = 0$ without the right side of 32.15 vanishing. In the case when both zeros are equal, there is only one solution of the form in 32.16 since there is always a unique solution to 32.15 for $n \geq 1$. Therefore, in the case when $r_1 - r_2 =$ a non negative integer either 0 or some positive integer, you must consider other solutions. I will use Abel's formula to find the second solution. The equation solved by these two solutions is

$$x^2 y'' + x p(x) y' + q(x) y = 0$$

and dividing by x^2 to place in the right form for using Abel's formula, Proposition 32.4.1.

$$y'' + \frac{1}{x} p(x) y' + \frac{1}{x^2} q(x) y = 0$$

Thus letting y_1 be the solution of the form in 32.16, and y_2 another solution, Abel's formula gives

$$y_2' y_1 - y_2 y_1' = W \in \int e^{-P(x)} dx, P'(x) = \frac{p(x)}{x}$$

Thus, following Procedure 32.4.2

$$y_2' - \frac{y_1'}{y_1} y_2 = \frac{1}{y_1} W$$

Then using an integrating factor $e^{\ln(1/|y_1|)} = \frac{1}{|y_1|}$

$$\frac{d}{dx} \left(\frac{1}{|y_1|} y_2 \right) = \frac{1}{|y_1| y_1} W = \pm \frac{1}{y_1^2} W$$

The sign does not matter to whether you have a solution, so it suffices to let

$$\frac{d}{dx} \left(\frac{1}{y_1} y_2 \right) = \frac{1}{y_1^2} W$$

Taking antiderivatives, another solution is

$$y_2 \in y_1 \int \frac{1}{y_1^2} e^{-P} dx$$

where $P(x) \in \int x^{-1} p(x) dx$. Thus

$$P(x) \in \int \left(\frac{b_0}{x} + b_1 + b_2 x + \dots \right) dx = b_0 \ln x + b_1 x + b_2 x^2/2 + \dots$$

and so

$$-P(x) = \ln x^{-b_0} + k(x)$$

for $k(x)$ some analytic function, $k(0) = 0$. Therefore,

$$e^{-P(x)} = e^{\ln(x^{-b_0}) + k(x)} = x^{-b_0} g(x)$$

for $g(x)$ some analytic function, $g(0) = 1$. Therefore,

$$y_2 \in y_1(x) \int \frac{1}{y_1^2} \left(x^{-b_0} g(x) \right) dx, \quad g(0) = 1. \quad (32.17)$$

Next it is good to understand y_1 and r_1 in terms of b_0 . Consider the zeros to the indicial equation,

$$r(r-1) + b_0 r + c_0 = r^2 - r + b_0 r + c_0 = 0.$$

It is given that $r_1 = r_2 + m$ where m is a non negative integer. Thus the left side of the above equals

$$(r - r_2)(r - r_2 - m) = r^2 - 2rr_2 - rm + r_2^2 + r_2 m$$

and so

$$-2r_2 - m = b_0 - 1$$

which implies

$$r_2 = \frac{1 - b_0}{2} - \frac{m}{2}$$

and hence

$$\begin{aligned} r_1 = r_2 + m &= \frac{1 - b_0}{2} + \frac{m}{2} \\ y_1(x) &= x^{\frac{1-b_0+m}{2}} \sum_{n=0}^{\infty} a_n x^n, \quad a_0 = 1 \end{aligned} \quad (32.18)$$

Now from Theorem 32.2.2 and looking at 32.18 $y_1(x)^{-2}$ is of the form

$$\frac{1}{x^{1-b_0+m} (\sum_{n=0}^{\infty} a_n x^n)^2} = x^{b_0-1-m} (1 + h(x))$$

where $h(x)$ is analytic, $h(0) = 0$. Therefore, 32.17 is

$$\begin{aligned} y_2(x) &\in y_1(x) \int x^{b_0-1-m} (1 + h(x)) \left(x^{-b_0} g(x) \right) dx \\ y_2(x) &\in y_1(x) \int x^{-1-m} (1 + l(x)) dx, \quad l(0) = 0 \end{aligned} \quad (32.19)$$

Now suppose that $m > 0$. Then,

$$\frac{y_2(x)}{y_1(x)} = \frac{-x^{-m}}{m} + \sum_{n=1}^{m-1} A_n \frac{x^{n-m}}{n-m} + A_m \ln(x) + \sum_{n=m+1}^{\infty} A_n \frac{x^{n-m}}{n-m}.$$

It follows

$$y_2 = A_m \ln(x) y_1 + x^{-m} \left(\frac{-1}{m} + \sum_{n=1}^{\infty} B_n x^n \right) \overbrace{x^{r_1} \sum_{n=0}^{\infty} a_n x^n}^{y_1}.$$

Where $B_n = \frac{A_n}{n-m}$ for $n \neq m$. Therefore, y_2 has the following form.

$$y_2 = A_m \ln(x) y_1 + x^{r_2} \sum_{n=0}^{\infty} C_n x^n.$$

If $m = 0$ so there is a repeated zero to the indicial equation then 32.19 implies

$$\frac{y_2}{y_1} = \ln x + \sum_{n=1}^{\infty} \frac{A_n}{n} x^n + A_0$$

where A_0 is a constant of integration. Thus, the second solution is of the form

$$y_2 = \ln(x) y_1 + x^{r_2} \sum_{n=0}^{\infty} C_n x^n.$$

The following theorem summarizes the above discussion.

PROCEDURE 32.5.2 Let 32.4 be an equation with a regular singular point and let r_1 and r_2 be real solutions of the indicial equation, 32.14 with $r_1 \geq r_2$. Then if $r_1 - r_2$ is not equal to an integer, the general solution 32.4 may be written in the form :

$$C_1 \sum_{n=0}^{\infty} a_n x^{n+r_1} + C_2 \sum_{n=0}^{\infty} b_n x^{n+r_2}$$

where we can have $a_0 = 1$ and $b_0 = 1$. If $r_1 = r_2 = r$ then the general solution of 32.4 may be obtained in the form

$$C_1 \overbrace{\sum_{n=0}^{\infty} a_n x^{n+r}}^{y_1} + C_2 \left(\ln(x) \overbrace{\sum_{n=0}^{\infty} a_n x^{n+r}}^{y_1} + \sum_{n=0}^{\infty} C_n x^{n+r} \right)$$

where we may take $a_0 = 1$. If $r_1 - r_2 = m$, a positive integer, then the general solution to 32.4 may be written as

$$C_1 \overbrace{\left(\sum_{n=0}^{\infty} a_n x^{n+r_1} \right)}^{y_1} + C_2 \left(k \ln(x) \overbrace{\left(\sum_{n=0}^{\infty} a_n x^{n+r_1} \right)}^{y_1} + x^{r_2} \sum_{n=0}^{\infty} C_n x^n \right),$$

where k may or may not equal zero and we may take $a_0 = 1$.

This procedure indicates what one should look for in the various cases. There is more discussion in [25].

32.6 The Bessel Equations

The Bessel differential equations are

$$x^2 y'' + xy' + (x^2 - \nu^2)y = 0$$

Obviously this has a regular singular point at 0 and the indicial equation is

$$r(r-1) + r - \nu^2 = r^2 - \nu^2 = 0$$

Thus the two indices of singularity are $\pm \nu$. There are various cases according to whether ν is 0, not an integer, or an integer.

32.6.1 The Case where $\nu = 0$

First consider the case where $\nu = 0$. In this case, there exists a solution of the form $\sum_{n=0}^{\infty} a_n x^n$ and it is required to find the constants a_n . Plugging into the equation one gets

$$x^2 \sum_{n=0}^{\infty} a_n n(n-1)x^{n-2} + x \sum_{n=0}^{\infty} a_n n x^{n-1} + \sum_{n=0}^{\infty} a_n x^{n+2} = 0$$

Then change the variable of summation in the last sum. This yields

$$\sum_{n=0}^{\infty} a_n n(n-1)x^n + \sum_{n=0}^{\infty} a_n n x^n + \sum_{n=2}^{\infty} a_{n-2} x^n = 0$$

It follows that there is no restriction on a_0, a_1 but for $n \geq 2$,

$$a_n (n(n-1) + n) + a_{n-2} = a_n n^2 + a_{n-2} = 0$$

Thus $a_n = -\frac{a_{n-2}}{n^2}$.

Taking $a_0 = 1, a_1 = 0$, it follows that all odd terms equal 0 and

$$a_2 = \frac{-1}{4}, a_4 = \frac{1}{2^2} \frac{1}{4^2}, a_6 = -\frac{1}{2^2} \frac{1}{4^2} \frac{1}{6^2}, \dots$$

The pattern is now fairly clear:

$$a_{2n} = (-1)^n \frac{1}{2^n (n!)^2}$$

Then this solution is

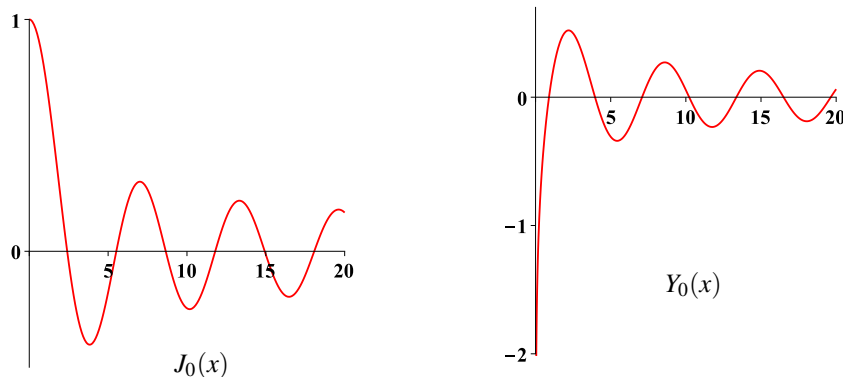
$$J_0(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{2^k (k!)^2} x^{2k} \quad (32.20)$$

Then by Theorem 32.5.2, the general solution is of the form

$$C_1 J_0(x) + C_2 \left(\ln(x) J_0(x) + \sum_{n=0}^{\infty} C_n x^n \right)$$

for suitable choice of the C_n . Thus one is bounded near $x = 0$ and the other is unbounded near $x = 0$. In fact, it is customary to let the second solution be a complicated linear

combination of these two solutions. When this is done, the function which results is known as $Y_0(x)$. Then $J_0(x)$ is the Bessel function of the first kind and the $Y_0(x)$ is called the Bessel function of the second kind. Here are graphs of these functions.



32.6.2 The Case of ν Not an Integer

Next consider the case where ν is not an integer. This time, the series is of the form

$$\sum_{n=0}^{\infty} a_n x^{n+\nu}$$

Substituting into the equation,

$$\begin{aligned} x^2 \sum_{n=0}^{\infty} a_n (n+\nu)(n+\nu-1)x^{n+\nu-2} + x \sum_{n=0}^{\infty} a_n (n+\nu)x^{n+\nu-1} \\ + \sum_{n=0}^{\infty} a_n x^{n+\nu+2} - \sum_{n=0}^{\infty} \nu^2 a_n x^{n+\nu} = 0 \end{aligned}$$

Thus a little simplification yields

$$\sum_{n=0}^{\infty} a_n (n+\nu)^2 x^{n+\nu} + \sum_{n=2}^{\infty} a_{n-2} x^{n+\nu} - \sum_{n=0}^{\infty} \nu^2 a_n x^{n+\nu} = 0$$

Then we need to have $a_1 = 0$ but let $a_0 = 1$. Then for $n \geq 2$,

$$a_n \left((n+\nu)^2 - \nu^2 \right) = -a_{n-2} \text{ so } a_n = \frac{-a_{n-2}}{(n+\nu)^2 - \nu^2} = \frac{-a_{n-2}}{n(n+2\nu)} \quad (32.21)$$

Thus all the odd terms are 0 and the first several terms are as follows.

$$a_0 = 1, a_2 = -\frac{1}{2(2+2\nu)}, a_4 = \frac{1}{2(2+2\nu)} \frac{1}{4(4+2\nu)}, \dots$$

The pattern seems clear at this point. Thus

$$a_{2n} = \frac{(-1)^n 1}{(2 \cdot 4 \cdot \dots \cdot 2n)(2+2\nu)(4+2\nu) \cdots (2n+2\nu)}$$

$$= \frac{(-1)^n 1}{2^{2n} n! (1+v)(2+v) \cdots (n+v)}$$

That product $(1+v)(2+v) \cdots (n+v)$ in the bottom will be denoted as $(n+v)_n$. Then this reduces to

$$a_{2n} = \frac{(-1)^n}{2^{2n} n! (n+v)_n}$$

Thus a solution corresponding to v is

$$x^v + \sum_{k=1}^{\infty} \frac{(-1)^k}{2^{2k} k! (k+v)_k} x^{2k+v}$$

Then this is massaged a little more. It is multiplied by the constant $\frac{1}{\Gamma(v+1)2^v}$. Recall that the Gamma function satisfies

$$\Gamma(\alpha)\alpha = \Gamma(\alpha+1)$$

Applying this rule repeatedly in the above sum yields

$$J_v(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(k+v+1)} \left(\frac{x}{2}\right)^{2k+v}$$

You can verify directly that

$$J_{-v}(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(k-v+1)} \left(\frac{x}{2}\right)^{2k-v}$$

is also a solution to the Bessel equation. The definition of $\Gamma(k-v+1)$ when the argument is negative is defined in terms of the property of the gamma function which was responsible for making $J_v(x)$ be a solution, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. Thus, for example, if $-v+1 < 0$, $\Gamma(-v+1)(-v+1) \cdots (-v+m) = \Gamma(-v+1+m)$ where m is large enough that $-v+1+m > 0$. Since v is not an integer, $-v+k$ is never zero so there is never a difficulty in encountering something which does not make sense.

The Bessel function of the first kind J_v converges to 0 as $x \rightarrow 0+$ while J_{-v} is unbounded as $x \rightarrow 0+$. Consequently, their ratio cannot be a constant and so the general solution is obtained as linear combinations of these two solutions. Of course everything changes if v is a positive integer. In this case, the second solution fails to even make sense because you could have $k-v=0$ and $\Gamma(0)$ is not even defined.

In fact, what people tabulate is a linear combination of these two solutions

$$Y_v(x) \equiv \frac{\cos(\pi v) J_v(x) - J_{-v}(x)}{\sin(\pi v)}$$

It is called the Weber function or the Neumann function. The main thing to notice here is that it is unbounded as $x \rightarrow 0$.

32.6.3 Case Where v is an Integer

Let $v = m$ a positive integer. Then you still get one solution which is of the form

$$x^m + \sum_{k=1}^{\infty} \frac{(-1)^k}{2^{2k} k! (k+m)_k} x^{2k+m}$$

and multiplying by a constant as above, you can obtain

$$J_m(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(k+m+1)} \left(\frac{x}{2}\right)^{2k+m} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! (k+m)!} \left(\frac{x}{2}\right)^{2k+m}$$

as one solution. In fact, you could consider simply replacing m with $-m$ in the above, but this will not work out. It won't work out roughly because $\Gamma(k-m+1) = \pm\infty$ for $k+1 \leq m$. Thus the sum will reduce to

$$\sum_{k=m}^{\infty} \frac{(-1)^k}{k! \Gamma(k-m+1)} \left(\frac{x}{2}\right)^{2k-m}$$

Changing the variable of summation to $k = j + m$, this becomes

$$\begin{aligned} (-1)^m \sum_{j=0}^{\infty} \frac{(-1)^j}{(j-m)! \Gamma(j+1)} \left(\frac{x}{2}\right)^{2j+m} &= (-1)^m \sum_{j=0}^{\infty} \frac{(-1)^j}{\Gamma(j-m+1) j!} \left(\frac{x}{2}\right)^{2j+m} \\ &= (-1)^m J_m(x) \end{aligned}$$

so the new solution obtained by replacing m with $-m$ is nothing more than $(-1)^m$ times the old solution. It follows that there is no way to obtain the general solution as a linear combination of these two. The second solution must involve a logarithmic term and will therefore, be unbounded near 0. However, it is convenient to define

$$J_{-m}(x) \equiv (-1)^m J_m(x)$$

The way this is dealt with is to define the second solution as

$$Y_m(x) \equiv \lim_{\nu \rightarrow m} Y_{\nu}(x)$$

because the limit does exist for all $x > 0$.

One other important consideration is easy to get which is that the solutions to Bessel's equation must oscillate about 0 like sines and cosines.

Proposition 32.6.1 *Let y be a solution of the Bessel equation*

$$x^2 y'' + xy' + (x^2 - \nu^2)y = 0$$

Then y has infinitely many zeros.

Proof: Change the independent variable to s where $x = e^s$. Thus, letting $y(s) = y(x)$, it will suffice to show that $s \rightarrow y(s)$ has infinitely many zeros. Doing the transformation yields the following differential equation for $s \rightarrow y(s)$

$$y''(s) + (e^{2s} - \nu^2)y(s) = 0$$

Obviously for all s large enough, $e^{2s} - \nu^2 > 1$. Consider now the equation

$$z'' + z = 0$$

The idea is to show that if a, b are successive zeros of z for a large enough that for $s > a$, $e^{2s} - \nu^2 > 1$ it follows that y must have a zero in $[a, b]$. Since z has infinitely many zeros, it follows that so does y .

Without loss of generality, assume z is positive on (a, b) . If it isn't, multiply by -1 to make this happen. The solution z is a linear combination of sines and cosines. It can be written in the form

$$z = A \cos(s - \phi)$$

and so $z'(a) > 0$ and $z'(b) < 0$.

If y has no zeros on $[a, b]$, then again, without loss of generality, let y be positive on $[a, b]$.

$$z''y - y''z + (1 - (e^{2s} - v^2))yz = 0$$

Thus on the open interval (a, b) ,

$$W(y, z)' = ((e^{2s} - v^2) - 1)yz > 0$$

where $W(y, z)$ is the Wronskian. It follows from the mean value theorem that $W(y, z)(a) < W(y, z)(b)$. Then

$$\begin{vmatrix} y(a) & 0 \\ y'(a) & z'(a) \end{vmatrix} < \begin{vmatrix} y(b) & 0 \\ y'(b) & z'(b) \end{vmatrix}$$

$$\text{positive} = y(a)z'(a) < y(b)z'(b) = \text{negative},$$

a contradiction. ■

For the purposes of this book, this will suffice. The main message is that there are two independent solutions, one bounded near 0 and the other unbounded as described above. Both oscillate about 0 and have infinitely many zeros. In many applications, the unbounded one is of no interest based on physical considerations.

32.7 Other Properties of Bessel Functions

Recall that for m a nonnegative integer,

$$J_m(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+m)!} \left(\frac{x}{2}\right)^{2k+m}$$

and that if $-m$ is negative,

$$J_{-m}(x) = (-1)^m J_m(x)$$

This $J_m(x)$ was the bounded solution for the Bessel equation. Note that an infinite sum of these functions is absolutely convergent. Indeed,

$$\begin{aligned} |J_m(x)| &\leq \sum_{k=0}^{\infty} \frac{1}{k!m!} \left(\left|\frac{x}{2}\right|\right)^{2k+m} \leq \frac{1}{m!} \left|\frac{x}{2}\right|^m \sum_{k=0}^{\infty} \frac{1}{k!} \left(\left(\frac{x}{2}\right)^2\right)^k \\ &= \frac{1}{m!} \left|\frac{x}{2}\right|^m \exp(x^2/4) \end{aligned} \quad (32.22)$$

Therefore, it is permissible to sum the various series which result in what follows in any order desired.

Now for $t \neq 0$,

$$e^{\frac{x}{2}} = \sum_{l=0}^{\infty} \frac{1}{l!} \left(\frac{x}{2}\right)^l t^l, \quad e^{-\frac{x}{2}} = \sum_{k=0}^{\infty} \frac{1}{k!} (-1)^k \left(\frac{x}{2}\right)^k t^{-k}$$

We multiply these two series. This will involve many terms which can be added in any order thanks to absolute convergence. To get t^m for $m \geq 0$, you need to multiply terms $l = m + k$ times the term for t^{-k} in the second sum. Thus you get for this term

$$\begin{aligned} & t^m \sum_{k=0}^{\infty} \frac{1}{(m+k)!} \left(\frac{x}{2}\right)^{m+k} \frac{1}{k!} (-1)^k \left(\frac{x}{2}\right)^k \\ &= t^m \sum_{k=0}^{\infty} (-1)^k \frac{1}{k! (m+k)!} \left(\frac{x}{2}\right)^{2k+m} = t^m J_m(x) \end{aligned}$$

This gives the terms t^m for $m \geq 0$.

What of the terms involving $m < 0$? To get these terms, you need to have $l - k = m$ so you need $k = l - m$. Thus the sum which results for these terms is

$$\begin{aligned} & t^m \sum_{l=0}^{\infty} \frac{1}{l!} \left(\frac{x}{2}\right)^l \frac{1}{(l-m)!} (-1)^{l-m} \left(\frac{x}{2}\right)^{l-m} = t^m \sum_{l=0}^{\infty} \frac{(-1)^{l-m}}{l! (l-m)!} \left(\frac{x}{2}\right)^{2l-m} \\ &= (-1)^m t^m \sum_{l=0}^{\infty} \frac{(-1)^l}{l! (l-m)!} \left(\frac{x}{2}\right)^{2l-m} = (-1)^m t^m J_{-m}(x) \end{aligned}$$

Therefore,

$$e^{\frac{x}{2}} e^{-\frac{x}{2t}} = e^{(x/2)(t-1/t)}$$

must equal the sum of t^m terms for $m \geq 0$ and the sum of t^m terms for $m < 0$. It follows that

$$\begin{aligned} e^{(x/2)(t-1/t)} &= J_0(x) + \sum_{m=1}^{\infty} t^m J_m(x) + \sum_{m=1}^{\infty} (-1)^m t^{-m} J_m(x) \\ &= J_0(x) + \sum_{m=1}^{\infty} J_m(x) (t^m + (-1)^m t^{-m}) \end{aligned}$$

Now recall that $J_{-m}(x) = (-1)^m J_m(x)$ and so

$$e^{(x/2)(t-1/t)} = J_0(x) + \sum_{m=1}^{\infty} J_m(x) t^m + \sum_{m=1}^{\infty} t^{-m} J_{-m}(x) = \sum_{m=-\infty}^{\infty} t^m J_m(x)$$

That is, $J_m(x)$ is just the m^{th} coefficient of the series for $e^{(x/2)(t-1/t)}$. This has proved the following interesting result on the generating function for Bessel equations.

Theorem 32.7.1 *For m an integer and $J_m(x) = (-1)^m J_{-m}(x)$, we have the following generating function for these Bessel functions.*

$$e^{(x/2)(t-1/t)} = \sum_{m=-\infty}^{\infty} t^m J_m(x) \quad (32.23)$$

In addition to this, there is an addition formula

$$J_m(x+y) = \sum_{k=-\infty}^{\infty} J_{m-k}(x) J_k(y) \quad (32.24)$$

Proof: It remains to obtain the above addition formula. This is remarkably easy to obtain.

$$\begin{aligned} e^{((x+y)/2)(t-1/t)} &= \sum_{m=-\infty}^{\infty} t^m J_m(x+y) \\ e^{((x+y)/2)(t-1/t)} &= e^{(x/2)(t-1/t)} e^{(y/2)(t-1/t)} \\ &= \sum_{l=-\infty}^{\infty} t^l J_l(x) \sum_{k=-\infty}^{\infty} t^k J_k(y) \end{aligned}$$

and in this product, the t^m term is the sum of products for which $l + k = m$. That is,

$$J_m(x+y) = \sum_{k=-\infty}^{\infty} J_k(y) J_{m-k}(x)$$

This shows the addition formula. ■

Of course t was completely arbitrary as long as it is not zero. Thus let it equal $e^{i\theta}$ in 32.23. Then from Euler's identity, $e^{i\theta} = (\cos(\theta) + i \sin(\theta))$,

$$e^{(x/2)(2i \sin \theta)} = \sum_{m=-\infty}^{\infty} (e^{i\theta})^m J_m(x)$$

Then using Euler's identity again,

$$\cos(x \sin(\theta)) + i \sin(x \sin(\theta)) = \sum_{m=-\infty}^{\infty} (\cos(m\theta) + i \sin(m\theta)) J_m(x)$$

Equating real and imaginary parts,

$$\begin{aligned} \cos(x \sin(\theta)) &= \sum_{m=-\infty}^{\infty} \cos(m\theta) J_m(x) \\ \sin(x \sin(\theta)) &= \sum_{m=-\infty}^{\infty} \sin(m\theta) J_m(x) \end{aligned}$$

Now recall from trig. identities,

$$\cos(a) \cos(b) + \sin(a) \sin(b) = \cos(a-b)$$

multiply the top by $\cos(n\theta)$ and the bottom by $\sin(n\theta)$ and add. Thus

$$\cos(n\theta - x \sin(\theta)) = \sum_{m=-\infty}^{\infty} \cos(n\theta - m\theta) J_m(x)$$

Because of the uniform convergence of the partial sums of the above series which follows from computations like those in 32.22, one can interchange \int_0^π with the infinite summation. This yields

$$\int_0^\pi \cos(n\theta - x \sin(\theta)) d\theta = \pi J_n(x)$$

because, unless $n = m$, $\int_0^\pi \cos(n\theta - m\theta) d\theta = 0$. Therefore, this yields the very important integral identity for $J_n(x)$,

$$J_n(x) = \frac{1}{\pi} \int_0^\pi \cos(n\theta - x \sin(\theta)) d\theta \quad (32.25)$$

The interchange of the integral with the summation follows from noting that the sums of the form $\sum_{m=-k}^k \cos(n\theta - m\theta) J_m(x)$ converge uniformly on $[0, \pi]$ to the infinite sum thanks to the $1/m!$ in the estimates of 32.22. Thus, from the fact that the integral is linear,

$$\begin{aligned} \int_0^\pi \sum_{m=-\infty}^{\infty} \cos(n\theta - m\theta) J_m(x) d\theta &= \int_0^\pi \lim_{k \rightarrow \infty} \sum_{m=-k}^k \cos(n\theta - m\theta) J_m(x) d\theta \\ &= \lim_{k \rightarrow \infty} \sum_{m=-k}^k \int_0^\pi \cos(n\theta - m\theta) J_m(x) d\theta = \sum_{m=-\infty}^{\infty} \int_0^\pi \cos(n\theta - m\theta) J_m(x) d\theta \end{aligned}$$

Theorem 32.7.2 *Let n be a positive integer. Then*

$$J_n(x) = \frac{1}{\pi} \int_0^\pi \cos(n\theta - x \sin(\theta)) d\theta$$

How do you compute $J_n(x)$? You can't get it the usual way very conveniently because the leading term vanishes at 0. This integral will give an easy way to do it. For example,

$$J_4(6) = \frac{1}{\pi} \int_0^\pi \cos(4\theta - 6 \sin(\theta)) d\theta = 0.35764$$

I just did the integral numerically in Scientific Notebook and got the answer easily. One can also produce a graph of $x \rightarrow J_4(x)$ very easily in this software by graphing the function of x given by $\frac{1}{\pi} \int_0^\pi \cos(4\theta - x \sin(\theta)) d\theta$. To do this, you simply type the expression in math mode and then select plot 2d. It has to work at it a little but will produce the graph. It knows that the variable is x and acts accordingly. In the exercises is a problem on how to do this in MATLAB. It is more elaborate.

There are whole books written on Bessel functions, [17].

32.8 Exercises

1. The Hermite equation is

$$y'' - xy' + ny = 0$$

Verify that if $n = 0$ or a positive integer, then this equation always has a polynomial solution. These are called Hermite polynomials. **Hint:** This is easier than the case of a regular singular point. Just look for a solution of the form $y = \sum_{n=0}^{\infty} a_n x^n$ and choose the a_n in such a way that the series satisfies the equation using the fact that you can differentiate a power series term by term. In this case, there should be two solutions.

2. If you have two polynomial solutions to the Hermite equation above, $p_m(x)$ corresponding to m in the equation and $p_n(x)$ corresponding to n in the equation, $n \neq m$, show that

$$\int_{-\infty}^{\infty} e^{-x^2} p_m(x) p_n(x) dx = 0$$

3. The equation

$$(1 - x^2)y'' - 2xy' + n(n+1)y = 0$$

is Legendre's equation. Note that 0 is an ordinary point for this equation. Show that for n a non-negative integer, this equation has polynomial solutions. Also explain why this equation has a regular singular point at $1, -1$.

4. In the above problem, suppose $p_k(x)$ and $p_l(x)$ are solutions, to the equations corresponding to $n = k, l$ respectively. Show that

$$\int_{-1}^1 p_k(x) p_l(x) dx = 0$$

Thus this gives an example of a collection of orthogonal polynomials.

5. The Legendre polynomials are given in the above problem but one multiplies by a constant so that the result satisfies $p_n(1) = 1$. The purpose of this problem is to find the constant. **Hint:** Use the Leibniz formula on $(x^2 - 1)^n = (x - 1)^n (x + 1)^n$.
6. The equation $(1 - x^2)y'' - xy' + n^2y = 0$ is called the Chebychev equation. Find solutions to this equation. That is, specify a recurrence relation and two solutions. Explain why there exist polynomial solutions to this equation. **Hint:** You just look for power series solutions.
7. The equation $(1 - x^2)y'' - 3xy' + n(n + 2)y = 0$ is also called the Chebychev equation. Find solutions to this equation. That is, specify a recurrence relation and two solutions. Explain why there exist polynomial solutions to this equation. **Hint:** You just look for power series solutions.
8. Specify two solutions to the following differential equation by determining a recurrence relation and then describing how to obtain two solutions. **Hint:** You just look for power series solutions.
- (a) $y''(x^2 + 1) + 5xy' + 2y = 0$. (f) $y'' - 2x^2y' - xy = 0$.
 (b) $y''(x^2 + 1) + xy' + 3y = 0$. (g) $y'' + x^2y' + 2xy = 0$.
 (c) $y''(x^2 + 1) + 7xy' + 4y = 0$. (h) $y'' - 3x^2y' - xy = 0$.
 (d) $y''(1 - 3x^2) + 6xy' + 4y = 0$. (i) $y'' + 2x^2y' - 4xy = 0$.
 (e) $y'' - 5x^2y' - 4xy = 0$.
9. Find the solution to the initial value problem $y'' + \sin(x)y' + \cos(3x)y = 0$ along with the initial conditions $y(0) = 1, y'(0) = -1$. You just need to find the first terms of the power series solution up to x^4 .
10. Find the solution to the initial value problem $y'' + \tan(2x)y' + \cos(3x)y = 0$ along with the initial conditions $y(0) = -1, y'(0) = 2$. You just need to find the first terms of the power series solution up to x^4 .
11. Find the solution to the initial value problem $y'' + \tan(5x)y' + \sec(3x)y = 0$ along with the initial conditions $y(0) = -2, y'(0) = 3$. You just need to find the first terms of the power series solution up to x^4 .
12. Find the general solution to the following Euler equations.

- (a) $y''x^2 - 3y'x + 3y = 0$. (d) $y''x^2 + 6y'x + 6y = 0$.
 (b) $y''x^2 + 4y'x - 4y = 0$. (e) $y''x^2 + 4y'x - 4y = 0$.
 (c) $y''x^2 + 2y'x - 6y = 0$. (f) $y''x^2 - 3y'x + 4y = 0$.

- (g) $y''x^2 + 5y'x + 4y = 0$. (l) $y''x^2 + 7y'x + 10y = 0$.
 (h) $y''x^2 - 5y'x + 9y = 0$. (m) $y''x^2 + 7y'x + 10y = 0$.
 (i) $y''x^2 - 3y'x + 4y = 0$. (n) $y''x^2 + 9y'x + 32y = 0$.
 (j) $y''x^2 - 3y'x + 4y = 0$. (o) $y''x^2 + 11y'x + 26y = 0$.
 (k) $y''x^2 - y'x + y = 0$. (p) $y''x^2 + 11y'x + 34y = 0$.

13. The hypergeometric equation is

$$x(1-x)y'' + (\gamma - (1+\alpha+\beta)x)y' - \alpha\beta y = 0$$

Show it has a regular singular point at 0 and that the roots of the indicial equation are 0 and $1-\gamma$.

14. In the above example, change the independent variable as follows: $t = 1/x$. Determine the equation which results in terms of t and show that the resulting equation has a regular singular point at 0 and that the roots of the indicial equation are α, β .
Hint: You need to show that $y''(x) = y''(t)t^4 + 2t^3y'(t)$, $y'(x) = -t^2y'(t)$. When you let $t = 0$, you are looking at the “point at infinity”. Thus you are showing that the “point at infinity” is a regular singular point.
15. Consider the Bessel equation in which $\nu = 1/2$. In this case, the roots of the indicial equation differ by an integer. Nevertheless, there are two solutions, neither of which involves a logarithm. Verify that for ν not an integer,

$$x^{-\nu} + \sum_{k=1}^{\infty} \frac{(-1)^k}{2^{2k}k!(k-\nu)_k} x^{2k-\nu}$$

does indeed yield a solution to the Bessel equation.

16. Show that for $\nu = 1/2$, one solution to the Bessel equation is $x^{-1/2} \sin(x)$. What is the other solution? Verify your answer. Show that one of these solutions is bounded and in fact converges to 0 as $x \rightarrow 0+$ while the other is unbounded as $x \rightarrow 0+$.
17. Explain why in every case, if you have a general solution to the Bessel equation, one of the solutions will be unbounded as $x \rightarrow 0$ and the other must converge to 0 as $x \rightarrow 0+$.
18. The Laguerre differential equation is

$$xy'' + (1-x)y' + my = 0$$

Show that when m is a nonnegative integer, there always exists a polynomial which is a solution to this differential equation. Letting $p_k(x), p_l(x)$ be polynomial solutions corresponding to $m = k, l$ respectively, show that

$$\int_0^{\infty} e^{-x} p_k(x) p_l(x) dx = 0, \quad k \neq l$$

19. Prove, Leibniz rule.

$$(fg)^{(n)} = \sum_{k=0}^n \binom{n}{k} f^{(k)} g^{(n-k)}$$

20. Suppose you have any linear second order differential equation $Ly = 0$ in which there is a general solution $C_1y + C_2z$ such that $W(y, x) \neq 0$ for $x \in [a, b]$. Show that if $y(x) = 0$, then $y'(x) \neq 0$. Why does this show that given a zero of a nonzero solution to the Bessel equation, or any other second order linear differential equation, there is a next zero?
21. Consider the equation $x^3y'' + 2xy' + y = 0$. Explain why it does not have a regular singular point at 0. Show that the only possible nonzero power series solution to this has radius of convergence equal to 0. In fact there really isn't any such series solution to this problem.
22. Consider the Bessel function $J_m(x)$ for m a positive integer. Recall the summation formula.

$$J_m(x+y) = \sum_{k=-\infty}^{\infty} J_{m-k}(x) J_k(y),$$

$$J_m(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+m)!} \left(\frac{x}{2}\right)^{2k+m}, \quad J_m(x) = (-1)^m J_{-m}(x)$$

Explain why J_m is even if m is even and J_m is odd if m is odd. Next let $m = 0$ and see what comes out of the summation formula. Then let $y = -x$ to obtain an inequality which shows that all the J_n are bounded. Show in particular that each $J_n(x)$ has the property that $|J_n(x)| \leq 1/\sqrt{2}$ if $n > 0$.

23. Use the integral formula for the Bessel function to graph $J_4(x)$ for $x \in [0, 20]$. Here is the syntax which will work for this. You put in the new lines.
- ```
hold on
for k=1:201 f=@(t,k)cos(4*t-((k-1)*.1)*sin(t));
y(k)=pi^(-1)*integral(@(t)f(t,k),0,pi);
x(k)=(k-1)*.1; plot(x,y,'linewidth',2) end.
```

## Chapter 33

# Boundary Value Problems, Fourier Series

### 33.1 Boundary Value Problems

The initial value problem can always be formulated as

$$y' = Ay + f, y(a) = y_0.$$

These are very nice problems because they always have a unique solution. A boundary value problem is different. They don't always have solutions.

**Definition 33.1.1** A two-point boundary value problem is to find a solution  $y$  to a differential equation

$$y'' + p(x)y' + q(x)y = g(x), x \in [a, b]$$

which also satisfies boundary conditions which are given at the two end points.

Examples of boundary values would be to give the value of  $y$  at the end points or the value of  $y'$  at the end points or some combination of  $y$  and  $y'$  at the end points.

**Example 33.1.2** Find the solutions to the equation  $y'' + y = \sin x$ , and boundary conditions

$$y(0) = 0, y(\pi) = 0$$

The general solution to the differential equation is easily seen to be  $A \cos(x) + B \sin(x) - \frac{1}{2}x \cos(x)$ . You have find  $A, B$  such that the boundary conditions are satisfied. Substituting  $t = 0$  yields  $A = 0$ . Then you also need  $B \sin(\pi) - \frac{1}{2}\pi \cos(\pi) = 0$  which is impossible. Therefore, there is no solution to this boundary value problem.

This is a very significant issue because there are numerical methods for solving boundary value problems. These methods will give you an answer even if there isn't one, but if it isn't there, you won't find it. Of course, just because you can't find it does not necessarily mean it isn't there. This is the interesting thing about math. In the absence of good existence and uniqueness theorems, you sometimes don't know what you are getting.

**Example 33.1.3** Find the solutions to the equation  $y'' + y = \sin x$ , and boundary conditions  $y(0) = 0, y(\frac{\pi}{2}) = 0$ .

It is the same equation, but the end points are different. As in the above example, if it has a solution, then it is of the form  $B \sin x - \frac{1}{2}x \cos(x)$ . Now let  $x = \pi/2$  and you find  $B - \frac{1}{4}\pi = 0$ . Thus a solution to this boundary value problem is  $y = -\frac{1}{2}x \cos(x)$ .

In this example, there was exactly one solution. Next consider

**Example 33.1.4** Find the solutions to the equation  $y'' + y = \sin x$ , and boundary conditions

$$y(0) = 0, y'\left(\frac{\pi}{2}\right) = \frac{\pi}{4}$$

The general solution to the differential equation is easily seen to be  $A \cos(x) + B \sin(x) - \frac{1}{2}x \cos(x)$ . You have find  $A, B$  such that the boundary conditions are satisfied. Substituting  $t = 0$  yields  $A = 0$ . Thus if there is a solution it is of the form  $y = B \sin(x) - \frac{1}{2}x \cos(x)$ . Then  $y'(x) = B \cos x - \frac{1}{2} \cos x + \frac{1}{2}x \sin x$ . Then you also need  $\frac{\pi}{4} = y'\left(\frac{\pi}{2}\right) = B \cos\left(\frac{\pi}{2}\right) + 0 + \frac{\pi}{4}$  which happens for any value of  $B$ . Therefore, for any  $B$ ,  $y = B \sin(x) - \frac{1}{2}x \cos(x)$  is a solution to this two point boundary value problem.

This is an example of a boundary value problem which has infinitely many solutions. Notice how all three examples involved the same differential equation, just different boundary conditions.

It turns out that for two point boundary value problems it is always this way. Either there are no solutions, exactly one or there are infinitely many. This may look familiar. Recall that it was this way for systems of linear equations. There are profound reasons why this similarity takes place but they are not for a book like this.

## 33.2 Eigenvalue Problems

I suppose these are best discussed through the example which will be featured most prominently.

**Example 33.2.1** Find the values of  $\lambda$  such that there exist nonzero solutions to the boundary value problem

$$\begin{aligned} y'' + \lambda y &= 0 \\ y(0) &= y(L) = 0 \end{aligned}$$

Along with any pair of boundary conditions which satisfy the conditions

$$y(0)y'(0) = 0, y(L)y'(L) = 0$$

Multiply by  $y$  and integrate from 0 to  $L$ .

$$\int_0^L y'' y dx + \lambda \int_0^L y^2 dx = 0 \quad (33.1)$$

Integrate by parts.

$$y'y|_0^L - \int_0^L (y')^2 dx + \lambda \int_0^L y^2 dx = 0$$

Consider now the boundary term. It equals 0 by assumption. Therefore,

$$-\int_0^L (y')^2 dx + \lambda \int_0^L y^2 dx = 0$$

If  $\lambda < 0$ , this equation could not be true and have  $y \neq 0$  because it would imply  $\int_0^L y^2 dx = 0$  so  $y = 0$ . Therefore, for any such example,  $\lambda \geq 0$ .

**Case 1:** Now consider some cases each of which have the property that  $yy'$  equals 0 at the end points of the interval  $[0, L]$ . First suppose  $y = 0$  at the ends of the interval. To save notation, write  $\lambda = \mu^2$ . Then you want

$$\begin{aligned} y'' + \mu^2 y &= 0 \\ y(0) &= y(L) = 0 \end{aligned}$$

The solution to the differential equation is

$$C_1 \sin \mu x + C_2 \cos \mu x$$

Insert the boundary conditions. This yields

$$C_2 = 0, C_1 \sin(\mu L) = 0$$

Therefore, for some nonnegative integer  $n$ , you must have  $\mu L = n\pi$ . You can't have  $n = 0$  since then  $y = 0$  and this is not allowed. Therefore,  $n$  is a positive integer and the eigenvalues are

$$\lambda = \frac{n^2 \pi^2}{L^2}, n = 1, 2, \dots$$

The corresponding eigenfunctions are

$$\sin\left(\frac{n\pi}{L}x\right), n = 1, 2, \dots$$

**Case 2:** Next consider the case where  $y' = 0$  at the ends. Thus you want nonzero  $y$  and  $\lambda$  such that

$$\begin{aligned} y'' + \mu^2 y &= 0 \\ y'(0) &= y'(L) = 0 \end{aligned}$$

The solution to the differential equation is

$$y = C_1 \sin \mu x + C_2 \cos \mu x$$

Then

$$y' = C_1 \mu \cos \mu x - C_2 \mu \sin \mu x$$

Insert the boundary conditions. At 0 this requires that

$$C_1 \mu = 0$$

At the right end point this requires

$$C_2 \mu \sin(\mu L) = 0$$

One case is for  $\mu = 0$ . This would result in an eigenfunction

$$y = 1$$

which is a nonzero function. Of course any nonzero multiple of this is also an eigenfunction. If  $\mu$  is not zero, then you need

$$\mu L = n\pi, \quad n = 1, 2, \dots$$

so

$$\lambda = \frac{n^2 \pi^2}{L^2}, \quad n = 0, 1, 2, \dots$$

The eigenfunctions in this case are

$$1, \cos\left(\frac{n\pi}{L}x\right), \quad n = 1, 2, \dots$$

**Case 3:** Next consider the case where  $y(0) = 0$  and  $y'(L) = 0$ . Thus you want nonzero  $y$  and  $\lambda$  such that

$$\begin{aligned} y'' + \mu^2 y &= 0 \\ y(0) &= y'(L) = 0 \end{aligned}$$

In this case, you would have

$$y = C_1 \sin \mu x + C_2 \cos \mu x$$

and on inserting the left boundary condition, this requires that  $C_2 = 0$ . Now consider the right boundary condition. You can't have  $\mu = 0$  in this case, because if you did, you would have  $y = 0$  which is not allowed. Hence you have

$$y'(L) = C_1 \mu \cos(\mu L) = 0$$

since  $\mu \neq 0$ , you must have

$$\mu L = (2n-1)\pi \text{ for } n = 1, 2, \dots$$

Therefore, in this case the eigenvalues are

$$\lambda = \frac{(2n-1)^2 \pi^2}{L^2}, \quad n = 1, 2, \dots$$

and the eigenfunctions are

$$\sin\left(\frac{(2n-1)\pi}{L}x\right), \quad n = 1, 2, \dots$$

### 33.3 Fourier Series

A Fourier series is a series which is intended to somehow approximate a given periodic function by an infinite sum of the form

$$a_0 + \sum_{k=1}^{\infty} a_k \cos\left(\frac{k\pi}{L}x\right) + \sum_{k=1}^{\infty} b_k \sin\left(\frac{k\pi}{L}x\right)$$

First of all, what is a periodic function?



**Definition 33.3.1** A function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is called **periodic** of period  $T$  if for all  $x \in \mathbb{R}$ ,

$$f(x + T) = f(x).$$

An example of a periodic function having period  $2L$  is  $x \rightarrow \sin\left(\frac{k\pi}{L}x\right)$  and  $x \rightarrow \cos\left(\frac{k\pi}{L}x\right)$ . If you want to approximate a function with these periodic functions, then it is necessary that it be periodic of period  $2L$ . Otherwise it would not be reasonable to expect to be able to approximate the function in any useful way with these periodic functions.

Before doing anything else, here are some important trig. identities.

$$\sin a \cos b = \frac{1}{2} (\sin(a + b) + \sin(a - b)) \quad (33.2)$$

$$\cos a \cos b = \frac{1}{2} (\cos(a - b) + \cos(a + b)) \quad (33.3)$$

$$\sin a \sin b = \frac{1}{2} (\cos(a - b) - \cos(a + b)) \quad (33.4)$$

These follow right away from the standard trig. identities for the sum of two angles. Here is a lemma which gives an orthogonality condition.

**Lemma 33.3.2** The following formulas hold. For  $m, n$  positive integers,

$$\int_{-L}^L \frac{1}{\sqrt{L}} \sin\left(\frac{m\pi}{L}x\right) \frac{1}{\sqrt{L}} \sin\left(\frac{n\pi}{L}x\right) dx = \begin{cases} 0 & \text{if } m \neq n \\ 1 & \text{if } m = n \end{cases}$$

$$\int_{-L}^L \frac{1}{\sqrt{L}} \cos\left(\frac{m\pi}{L}x\right) \frac{1}{\sqrt{L}} \cos\left(\frac{n\pi}{L}x\right) dx = \begin{cases} 0 & \text{if } m \neq n \\ 1 & \text{if } m = n \end{cases}$$

$$\int_{-L}^L \sin\left(\frac{m\pi}{L}x\right) \cos\left(\frac{n\pi}{L}x\right) dx = 0$$

**Proof:** Consider the first of these formulas. From one of the above trig. identities,

$$\begin{aligned} \int_{-L}^L \sin\left(\frac{m\pi}{L}x\right) \sin\left(\frac{n\pi}{L}x\right) dx &= \\ \frac{1}{2} \int_{-L}^L \cos\left(\left(\frac{m\pi}{L} - \frac{n\pi}{L}\right)x\right) - \cos\left(\left(\frac{m\pi}{L} + \frac{n\pi}{L}\right)x\right) dx \end{aligned}$$

If  $m \neq n$ , this clearly integrates to 0. If  $m = n$ , you have

$$\frac{1}{2} \int_{-L}^L \left(1 - \cos\left(\frac{2n\pi}{L}x\right)\right) dx = L$$

Thus

$$\int_{-L}^L \frac{1}{\sqrt{L}} \sin\left(\frac{n\pi}{L}x\right) \frac{1}{\sqrt{L}} \sin\left(\frac{n\pi}{L}x\right) dx = 1$$

The second formula works out the same way. Consider the third.

$$\int_{-L}^L \sin\left(\frac{m\pi}{L}x\right) \cos\left(\frac{n\pi}{L}x\right) dx =$$

$$\frac{1}{2} \int_{-L}^L \left( \sin \left( \left( \frac{m\pi}{L} + \frac{n\pi}{L} \right) x \right) + \sin \left( \left( \frac{m\pi}{L} - \frac{n\pi}{L} \right) x \right) \right) dx$$

It is easy to see that this integral is always 0 regardless the choice of  $m, n$ . ■

Now suppose you succeed in approximating  $f$  with a Fourier series in some meaningful way.

$$f(x) \approx a_0 \frac{1}{\sqrt{2L}} + \sum_{k=1}^{\infty} a_k \frac{1}{\sqrt{L}} \cos \left( \frac{k\pi}{L} x \right) + \sum_{k=1}^{\infty} b_k \frac{1}{\sqrt{L}} \sin \left( \frac{k\pi}{L} x \right) \quad (33.5)$$

What should be the formula for  $a_k$  and  $b_k$ ? Multiply both sides by  $\frac{1}{\sqrt{L}} \sin \left( \frac{m\pi}{L} x \right)$  and then integrate the resulting infinite sum by saying the integral of the sum is the sum of the integrals. Since the sum involves a limit, this is nothing but a formal and highly speculative piece of pseudo mathematical nonsense but we will not let a little thing like that get in the way. Thus

$$\begin{aligned} \int_{-L}^L f(x) \frac{1}{\sqrt{L}} \sin \left( \frac{m\pi}{L} x \right) dx &= a_0 \int_{-L}^L \frac{1}{\sqrt{2L}} \sin \left( \frac{m\pi}{L} x \right) dx + \\ &\sum_{k=1}^{\infty} a_k \int_{-L}^L \frac{1}{\sqrt{L}} \cos \left( \frac{k\pi}{L} x \right) \frac{1}{\sqrt{L}} \sin \left( \frac{m\pi}{L} x \right) dx \\ &+ \sum_{k=1}^{\infty} b_k \int_{-L}^L \frac{1}{\sqrt{L}} \sin \left( \frac{k\pi}{L} x \right) \frac{1}{\sqrt{L}} \sin \left( \frac{m\pi}{L} x \right) dx \end{aligned}$$

All these integrals equal 0 but one and that is the one involving the sine and  $k = m$ . This is by the above lemma. Therefore,

$$\int_{-L}^L f(x) \frac{1}{\sqrt{L}} \sin \left( \frac{m\pi}{L} x \right) dx = b_m \frac{1}{L} \int_{-L}^L \sin^2 \left( \frac{m\pi}{L} x \right) dx = b_m$$

It seems likely therefore, that  $b_m$  should be defined as

$$b_m = \int_{-L}^L f(x) \frac{1}{\sqrt{L}} \sin \left( \frac{m\pi}{L} x \right) dx \quad (33.6)$$

Next do the same thing after multiplying by  $\frac{1}{\sqrt{L}} \cos \left( \frac{m\pi}{L} x \right)$ . Another use of the same lemma implies that the appropriate choice for  $a_m$  is

$$a_m = \int_{-L}^L f(x) \frac{1}{\sqrt{L}} \cos \left( \frac{m\pi}{L} x \right) dx \quad (33.7)$$

Finally integrate both sides of 33.5. This yields

$$\int_{-L}^L f(x) \frac{1}{\sqrt{2L}} dx = a_0 \quad (33.8)$$

and so the appropriate description of  $a_0$  is given above. Thus the Fourier series is of the form

$$\begin{aligned} &\int_{-L}^L f(y) \frac{1}{\sqrt{2L}} dy \frac{1}{\sqrt{2L}} + \sum_{m=1}^{\infty} \left( \int_{-L}^L f(y) \frac{1}{\sqrt{L}} \cos \left( \frac{m\pi}{L} y \right) dy \right) \frac{1}{\sqrt{L}} \cos \left( \frac{m\pi}{L} x \right) \\ &+ \sum_{m=1}^{\infty} \left( \int_{-L}^L f(y) \frac{1}{\sqrt{L}} \sin \left( \frac{m\pi}{L} y \right) dy \right) \frac{1}{\sqrt{L}} \sin \left( \frac{m\pi}{L} x \right) \end{aligned}$$

Combining the  $\sqrt{L}$  terms, this yields

$$\begin{aligned} &= \frac{1}{2L} \int_{-L}^L f(y) dy + \sum_{m=1}^{\infty} \left( \frac{1}{L} \int_{-L}^L f(y) \cos\left(\frac{m\pi}{L}y\right) dy \right) \cos\left(\frac{m\pi}{L}x\right) \\ &\quad + \sum_{m=1}^{\infty} \left( \frac{1}{L} \int_{-L}^L f(y) \sin\left(\frac{m\pi}{L}y\right) dy \right) \sin\left(\frac{m\pi}{L}x\right) \end{aligned}$$

This is so far completely speculative, but this was they often did things back in the time when Fourier came up with the idea back in the early 1800s. Here is the definition of the Fourier series in which we combine the various constant terms to make it easier to remember.

**Definition 33.3.3** Let  $f$  be a function defined on  $\mathbb{R}$  which is  $2L$  periodic and Riemann integrable on every closed interval of length  $2L$ . Then the Fourier series is defined as

$$a_0 + \sum_{k=1}^{\infty} a_k \cos\left(\frac{k\pi}{L}x\right) + \sum_{k=1}^{\infty} b_k \sin\left(\frac{k\pi}{L}x\right)$$

where  $a_0, a_m, b_m$  are given as

$$\begin{aligned} a_0 &= \frac{1}{2L} \int_{-L}^L f(y) dy, a_m = \frac{1}{L} \int_{-L}^L f(y) \cos\left(\frac{m\pi}{L}y\right) dy \\ b_m &= \frac{1}{L} \int_{-L}^L f(y) \sin\left(\frac{m\pi}{L}y\right) dy \end{aligned}$$

We will refer to  $a_0, a_n, b_n$  as Fourier coefficients.

### 33.4 Mean Square Approximation

When you have two functions defined on an interval  $[a, b]$ , how do you measure the distance between them? It turns out there are infinitely many ways to do this. One way is to say the distance between  $f$  and  $g$ , denoted as  $\|f - g\|$  is defined as

$$\|f - g\| = \sup \{|f(x) - g(x)|, x \in [a, b]\}$$

To say that two functions are close in this sense is to say that for each  $x$  you have  $f(x)$  close to  $g(x)$ . The two functions are said to be uniformly close if they are close in this norm. This norm is also called the uniform norm.

This is a good way to define distance between functions, but it turns out that a more useful way in many situations is the following. You define

$$\|f - g\| \equiv \left( \int_a^b |f(x) - g(x)|^2 dx \right)^{1/2}$$

Then  $\|f - g\|$  is called the mean square norm with the above definition. You should verify that if two functions are close in the uniform norm, then they must be close in the mean square norm, but not the other way around. Often the mean square norm is denoted as  $|f - g|$ . So why is this a norm and what is meant by a norm? First here is a simple lemma.

**Lemma 33.4.1** Suppose  $f, g$  are Riemann integrable functions. Define

$$(f, g) \equiv \int_a^b f(x)g(x)dx.$$

Then the following are satisfied.

$$\begin{aligned}(f, g) &= (g, f) \\ (f, f) &\geq 0\end{aligned}$$

For  $a, b$  real numbers,

$$\begin{aligned}(af + bg, h) &= a(f, h) + b(g, h) \\ (f, ag + bh) &= a(f, g) + b(f, h)\end{aligned}$$

The following inequality called the Cauchy-Schwarz inequality holds.

$$|(f, g)| \leq |f| |g| \equiv (f, f)^{1/2} (g, g)^{1/2}$$

where  $|f|$  denotes the mean square distance defined above.

**Proof:** All of the above are completely obvious except for the last one. As to that one, note that from the first obvious properties, for  $t \in \mathbb{R}$

$$0 \leq (tf + g, tf + g) = t^2(f, f) + 2t(f, g) + (g, g)$$

If  $(f, f) = 0$  there is nothing to prove because you must have  $(f, g) = 0$  since otherwise, the above inequality would be violated for suitable choice of  $t$ . It follows that the above is a quadratic polynomial whose graph opens up and which has at most one real zero. Hence by the quadratic formula,

$$4(f, g)^2 - 4(f, f)(g, g) \leq 0$$

which reduces to the Cauchy-Schwarz inequality. ■

Now the mean square norm amounts to nothing more than  $|f| = (f, f)^{1/2}$ .

**Proposition 33.4.2** The mean square norm  $\|f\| = |f| = (f, f)^{1/2}$  satisfies the following axioms.

1.  $\|f\| \geq 0$
2. If  $a$  is a number,  $\|af\| = |a| \|f\|$
3.  $\|f + g\| \leq \|f\| + \|g\|$

**Proof:** The only one which is not completely obvious is the last. Then by the definition of the norm and the properties of  $(\cdot, \cdot)$ ,

$$\begin{aligned}\|f + g\|^2 &\equiv (f + g, f + g) = \|f\|^2 + \|g\|^2 + 2(f, g) \\ &\leq \|f\|^2 + \|g\|^2 + 2|(f, g)| \\ &\leq \|f\|^2 + \|g\|^2 + 2\|f\| \|g\| \\ &= (\|f\| + \|g\|)^2\end{aligned}$$

Now taking the square root of both sides yields the desired inequality. ■

The reason this is important is that if you have  $f$  close to  $g$  and  $h$  close to  $g$ , then you have  $f$  close to  $h$ . Indeed,

$$\|f - h\| \leq \|f - g\| + \|g - h\|$$

and if both of the terms on the right are small, then the term on the left is also.

There are  $2n + 1$  functions,  $\frac{1}{\sqrt{2L}}, \frac{1}{\sqrt{L}} \cos\left(\frac{k\pi}{L}x\right), \frac{1}{\sqrt{L}} \sin\left(\frac{j\pi}{L}x\right)$  for  $k, j \in 1, 2, \dots, n$ . Denote these functions as  $\{\phi_k\}_{k=1}^{2n+1}$  to save on notation. It was shown above that  $(\phi_k, \phi_j) = \delta_{jk}$  which is 1 if  $k = j$  and 0 if  $k \neq j$ . Then for  $f$  a Riemann integrable function on  $[-L, L]$ , our problem is to choose  $\alpha_k$  to minimize

$$\left| f - \sum_{k=1}^{2n+1} \alpha_k \phi_k \right|^2 = (A + B, A + B)$$

where  $A = f - \sum_{k=1}^{2n+1} (f, \phi_k) \phi_k$ ,  $B = \sum_{k=1}^{2n+1} ((f, \phi_k) - \alpha_k) \phi_k$ . Thus the above is

$$|A|^2 + 2(A, B) + |B|^2 \quad (*)$$

Consider the middle term.

$$\begin{aligned} \left( f - \sum_{k=1}^{2n+1} (f, \phi_k) \phi_k, \phi_j \right) &= (f, \phi_j) - \sum_k (f, \phi_k) (\phi_k, \phi_j) \\ &= (f, \phi_j) - (f, \phi_j) = 0 \end{aligned}$$

and so the middle term of  $*$  equals 0 because

$$\left( f - \sum_{k=1}^{2n+1} (f, \phi_k) \phi_k, \sum_{k=1}^{2n+1} \alpha_k \phi_k \right) = 0$$

for any choice of  $\alpha_k$  which includes the case of  $(A, B)$ . Thus  $*$  implies

$$\left| f - \sum_{k=1}^{2n+1} \alpha_k \phi_k \right|^2 = \left| f - \sum_{k=1}^{2n+1} (f, \phi_k) \phi_k \right|^2 + \left| \sum_{k=1}^{2n+1} ((f, \phi_k) - \alpha_k) \phi_k \right|^2$$

From the definition of the norm, the second term is

$$\sum_{j,k} ((f, \phi_k) - \alpha_k) ((f, \phi_j) - \alpha_j) (\phi_k, \phi_j) = \sum_k ((f, \phi_k) - \alpha_k)^2$$

Hence

$$\left| f - \sum_{k=1}^{2n+1} \alpha_k \phi_k \right|^2 = \left| f - \sum_{k=1}^{2n+1} (f, \phi_k) \phi_k \right|^2 + \sum_{k=1}^{2n+1} ((f, \phi_k) - \alpha_k)^2 \quad (**)$$

which shows that the left side is minimized exactly when  $\alpha_k = (f, \phi_k)$ . It is clear then that corresponding to  $\frac{1}{\sqrt{L}} \cos\left(\frac{k\pi x}{L}\right)$ , you would have

$$\alpha_k = \int_{-L}^L f(x) \frac{1}{\sqrt{L}} \cos\left(\frac{k\pi x}{L}\right) dx$$

and so, the term in the Fourier series which corresponds to this would be

$$\begin{aligned} & \left( \int_{-L}^L f(y) \frac{1}{\sqrt{L}} \cos\left(\frac{k\pi y}{L}\right) dy \right) \frac{1}{\sqrt{L}} \cos\left(\frac{k\pi x}{L}\right) \\ &= \left( \frac{1}{L} \int_{-L}^L f(y) \cos\left(\frac{k\pi y}{L}\right) dy \right) \cos\left(\frac{k\pi x}{L}\right) \end{aligned}$$

which is exactly what was determined earlier.

In \*\*, let each  $\alpha_k = 0$ . Then this equation implies

$$|f|^2 \geq \sum_{k=1}^{2n+1} (f, \phi_k)^2$$

which is Bessel's inequality. In particular, in the case of most interest here, this inequality is

$$\begin{aligned} |f|^2 \geq & \frac{1}{2L} \left( \int_{-L}^L f(x) dx \right)^2 + \frac{1}{L} \sum_{k=1}^n \left( \int_{-L}^L f(x) \sin\left(\frac{k\pi x}{L}\right) dx \right)^2 \\ & + \frac{1}{L} \sum_{k=1}^n \left( \int_{-L}^L f(x) \cos\left(\frac{k\pi x}{L}\right) dx \right)^2 \end{aligned} \quad (**)$$

It follows that the sequence of partial sums in the sum on the right in \*\* converges and so

$$\begin{aligned} \lim_{k \rightarrow \infty} \int_{-L}^L f(x) \cos\left(\frac{k\pi}{L}x\right) dx &= 0 \\ \lim_{k \rightarrow \infty} \int_{-L}^L f(x) \sin\left(\frac{k\pi}{L}x\right) dx &= 0 \end{aligned} \quad (33.9)$$

The two limits in 33.9 are special cases of the Riemann-Lebesgue lemma. These are the considerations which make it possible to consider the pointwise convergence properties of Fourier series. In particular, the following lemma is used.

**Lemma 33.4.3** Suppose  $f$  is a Riemann integrable function defined on  $[-L, L]$ . Then

$$\lim_{k \rightarrow \infty} \int_{-L}^L f(x) \sin\left(\left(k + \frac{1}{2}\right) \frac{\pi}{L}x\right) dx = 0$$

**Proof:** It equals

$$\lim_{k \rightarrow \infty} \left[ \int_{-L}^L f(x) \cos\left(\frac{\pi x}{2L}\right) \sin\left(\frac{k\pi}{L}x\right) dx + \int_{-L}^L f(x) \sin\left(\frac{\pi x}{2L}\right) \cos\left(\frac{k\pi}{L}x\right) dx \right]$$

and each of these converge to 0 thanks to 33.9. ■

Fourier series are really all about mean square convergence. If you are interested in pointwise approximation with trig. polynomials, there are better ways to do it than with Fourier series. However, the pointwise convergence is also very interesting and this is discussed in the next section.

### 33.5 Pointwise Convergence of Fourier Series

For each  $x$  the Fourier series yields an infinite series. One wonders whether it converges to  $f(x)$ . It is completely obvious that this is not necessarily the case. This is because the Fourier series is completely unchanged if  $f$  is changed at any finite set of points. Therefore, to obtain any sort of meaningful convergence, one must assume something about the function. The following is an elementary theorem which is a special case of a more substantial real analysis result.

**Definition 33.5.1** *The one sided limits are*

$$f(x+) \equiv \lim_{h \rightarrow 0+} f(x+h), \quad f(x-) \equiv \lim_{h \rightarrow 0+} f(x-h)$$

**Theorem 33.5.2** *Suppose  $f$  is a periodic function of period  $2L$  such that  $f$  has only finitely many jump discontinuities on the interval  $[-L, L)$ . Suppose there exists a constant  $K$  such that for all  $x$ ,*

$$|f(x+) - f(x+y)| < Ky$$

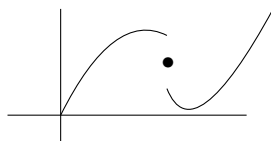
*for all sufficiently small positive  $y$ . Also*

$$|f(x-) - f(x-y)| < Ky$$

*for all  $y$  sufficiently small. Then*

$$\frac{f(x+) + f(x-)}{2} = a_0 + \sum_{k=1}^{\infty} a_k \cos\left(\frac{k\pi}{L}x\right) + \sum_{k=1}^{\infty} b_k \sin\left(\frac{k\pi}{L}x\right)$$

In words, this says that the Fourier series converges to the midpoint of the jump. A picture which represents a part of the graph of  $f$  is as follows.



You note that the dot is at the midpoint of the jump. The condition in the theorem is there to rule out excessive steepness of the graph of the function. In fact, one can do a lot better than what it says in this theorem. You should see [3] for two more general treatments of this theorem.

One way to satisfy the condition on not having excessive steepness is to have  $f$  be piecewise continuous such that if  $a, b$  are successive discontinuities, then redefining  $f$  on  $[a, b]$  to equal  $f(a+)$  at the left and  $f(b-)$  at the right, the new function has a continuous derivative on  $[a, b]$ .

Theorem 33.5.2 is the convergence theorem. I am going to give a discussion of this convergence theorem. If you are not interested in understanding why it works, ignore the proof. It is included in case someone would be interested. This important theorem or one like it was first proved in 1829 by Dirichlet.

### 33.5.1 Explanation of Pointwise Convergence Theorem

**Proof of the convergence theorem:** The convergence of sums has to do with the limit of the sequence of partial sums. Let

$$S_n f(x) = a_0 + \sum_{k=1}^n a_k \cos\left(\frac{k\pi}{L}x\right) + \sum_{k=1}^n b_k \sin\left(\frac{k\pi}{L}x\right)$$

From the definition, this equals

$$\begin{aligned} & \frac{1}{L} \int_{-L}^L \frac{f(y)}{2} dy + \sum_{k=1}^n \frac{1}{L} \int_{-L}^L f(y) \cos\left(\frac{k\pi}{L}y\right) dy \cos\left(\frac{k\pi}{L}x\right) \\ & + \sum_{k=1}^n \frac{1}{L} \int_{-L}^L f(y) \sin\left(\frac{k\pi}{L}y\right) dy \sin\left(\frac{k\pi}{L}x\right) \end{aligned}$$

This simplifies to

$$\begin{aligned} & \frac{1}{L} \int_{-L}^L \frac{f(y)}{2} dy + \\ & \sum_{k=1}^n \frac{1}{L} \int_{-L}^L f(y) \cos\left(\frac{k\pi}{L}y\right) \cos\left(\frac{k\pi}{L}x\right) + f(y) \sin\left(\frac{k\pi}{L}y\right) \sin\left(\frac{k\pi}{L}x\right) dy \end{aligned}$$

which equals

$$\frac{1}{L} \int_{-L}^L \frac{f(y)}{2} dy + \sum_{k=1}^n \frac{1}{L} \int_{-L}^L f(y) \cos\left(\frac{k\pi}{L}(x-y)\right) dy$$

Simplifying this a little more yields

$$\begin{aligned} & \int_{-L}^L \frac{1}{L} \left( \frac{1}{2} + \sum_{k=1}^n \cos\left(\frac{k\pi}{L}(x-y)\right) \right) f(y) dy \\ & \equiv \int_{-L}^L D_n(x-y) f(y) dy \end{aligned}$$

Here  $D_n(t)$  is called the Dirichlet kernel. In order to consider the convergence of the partial sums, it is necessary to study the properties of the Dirichlet kernel.

**Lemma 33.5.3** *The Dirichlet kernel is periodic of period  $2L$ .*

$$\int_{-L}^L D_n(t) dt = 2 \int_0^L D_n(t) dt = 1.$$

There is also a formula for this kernel,

$$D_n(t) = \frac{\sin\left(\left(n + \frac{1}{2}\right) \frac{\pi}{L} t\right)}{2L \sin\left(\frac{\pi}{2L} t\right)}$$

**Proof:** As indicated above,

$$D_n(t) = \frac{1}{L} \left( \frac{1}{2} + \sum_{k=1}^n \cos\left(\frac{k\pi}{L}t\right) \right)$$



and so, it is obvious that

$$\int_{-L}^L D_n(t) dt = 1. \quad (33.10)$$

From the above formula, it follows that  $D_n(t) = D_n(-t)$  and  $D_n(x+2L) = D_n(x)$ . Since  $D_n(t) = D_n(-t)$ ,

$$\int_{-L}^L D_n(t) dt = 2 \int_0^L D_n(t) dt$$

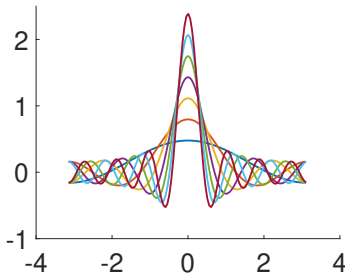
It remains to find a formula. Use 33.2

$$\begin{aligned} \sin\left(\frac{\pi}{2L}t\right) D_n(t) &= \frac{1}{L} \left( \frac{1}{2} \sin\left(\frac{\pi}{2L}t\right) + \sum_{k=1}^n \sin\left(\frac{\pi}{2L}t\right) \cos\left(\frac{k\pi}{L}t\right) \right) \\ &= \frac{1}{L} \left( \frac{1}{2} \sin\left(\frac{\pi}{2L}t\right) + \frac{1}{2} \sum_{k=1}^n \sin\left(\left(\frac{k\pi}{L} + \frac{\pi}{2L}\right)t\right) - \sin\left(\left(\frac{k\pi}{L} - \frac{\pi}{2L}\right)t\right) \right) \\ &= \frac{1}{2L} \left[ \sin\left(\frac{\pi}{2L}t\right) + \sum_{k=1}^n \sin\left(\left(k + \frac{1}{2}\right)\frac{\pi}{L}t\right) - \sum_{k=1}^n \sin\left(\left(k - \frac{1}{2}\right)\frac{\pi}{L}t\right) \right] \\ &= \frac{1}{2L} \left[ \sin\left(\frac{\pi}{2L}t\right) + \sum_{k=1}^n \sin\left(\left(k + \frac{1}{2}\right)\frac{\pi}{L}t\right) - \sum_{k=0}^{n-1} \sin\left(\left(k + \frac{1}{2}\right)\frac{\pi}{L}t\right) \right] \\ &= \frac{1}{2L} \sin\left(\left(n + \frac{1}{2}\right)\frac{\pi}{L}t\right) \end{aligned}$$

Thus the desired formula is

$$D_n(t) = \frac{\sin\left(\left(n + \frac{1}{2}\right)\frac{\pi}{L}t\right)}{2L \sin\left(\frac{\pi}{2L}t\right)} \blacksquare$$

Here is a graph of the first seven of these Dirichlet kernels,  $n \geq 1$  for  $L = \pi$ .



Next, it follows from the above that

$$S_n f(x) = \int_{-L}^L D_n(x-y) f(y) dy.$$

Change the variables. Let  $u = x - y$ . Then this reduces to

$$\int_{-L+x}^{L+x} D_n(u) f(x-u) du$$

Since  $D_n$  and  $f$  are both periodic of period  $2L$ , this equals

$$\int_{-L}^L D_n(y) f(x-y) dy$$

Therefore, since  $\int_{-L}^L D_n(y) dy = 1$ ,

$$\left| \frac{f(x+) + f(x-)}{2} - S_n f(x) \right| = \left| \frac{f(x+) + f(x-)}{2} - \int_{-L}^L D_n(y) f(x-y) dy \right|$$

$$\begin{aligned}
&= \left| \int_{-L}^L \left( \frac{f(x+) + f(x-)}{2} - f(x-y) \right) D_n(y) dy \right| \\
&= \left| \int_0^L (f(x+) + f(x-)) D_n(y) dy - \int_0^L (f(x-y) + f(x+y)) D_n(y) dy \right| \\
&\leq \left| \int_0^L \frac{f(x+) - f(x+y)}{2L \sin(\frac{\pi}{2L}y)} \sin \left( \left( n + \frac{1}{2} \right) \frac{\pi}{L} y \right) dy \right| + \\
&\quad \left| \int_0^L \frac{f(x-) - f(x-y)}{2L \sin(\frac{\pi}{2L}y)} \sin \left( \left( n + \frac{1}{2} \right) \frac{\pi}{L} y \right) dy \right|
\end{aligned}$$

Both of these converge to 0 thanks to Lemma 33.4.3. To use this lemma, it is only necessary to verify that the functions

$$y \rightarrow \frac{f(x-) - f(x-y)}{2L \sin(\frac{\pi}{2L}y)}, \quad y \rightarrow \frac{f(x+) - f(x+y)}{2L \sin(\frac{\pi}{2L}y)}$$

are each Riemann integrable on  $[-L, L]$ .

I will show this now. Each is continuous except for finitely many points of discontinuity. The only remaining issue is whether the functions are bounded as  $y \rightarrow 0$ . However, there exists a constant  $K$  such that

$$\left| \frac{f(x+) - f(x+y)}{2L \sin(\frac{\pi}{2L}y)} \right| \leq \frac{K|y|}{|2L \sin(\frac{\pi}{2L}y)|}$$

and this expression converges to  $K/\pi$ , so the function is Riemann integrable. The other function is similar. ■

**Example 33.5.4** Let  $f(x) = |x|$  for  $x \in [-1, 1]$  and let  $f$  be periodic of period 2. Find the Fourier series of  $f$ .

Here you need  $L = 1$ . Then

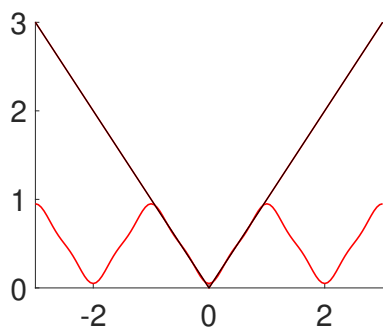
$$\begin{aligned}
a_0 &= \frac{1}{2} \int_{-1}^1 |x| dx = \frac{1}{2} \\
a_k &= \int_{-1}^1 |x| \cos(k\pi x) dx = \frac{2}{\pi^2 k^2} \left( (-1)^k - 1 \right)
\end{aligned}$$

Note that  $a_k = 0$  if  $k$  is even and it equals  $-4/(\pi^2 k^2)$  when  $k$  is odd.

Since the function is even, the  $b_k = 0$ . Therefore, the Fourier series equals

$$\frac{1}{2} - \sum_{k=1}^{\infty} \frac{4}{\pi^2 (2k-1)^2} \cos(2k-1)\pi x$$

Now here is the graph of the function between  $-1$  and  $1$  along with the sum up to 2 in the Fourier series. You will notice that after only three terms the Fourier series appears to be very close to the function on the interval  $[-1, 1]$ . This also shows how the Fourier series approximates the periodic extension of this function off this interval.



Notice that if you take  $x = 0$  in the above, the theorem on pointwise convergence of Fourier series implies that the sum converges to the value of the function which is 0. Therefore,

$$\frac{1}{2} = \sum_{k=1}^{\infty} \frac{4}{\pi^2 (2k-1)^2}$$

It follows that

$$\frac{\pi^2}{8} = \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2}.$$

This is a remarkable assertion.

Now here is another example for which the Fourier series will have to struggle harder to approximate the function.

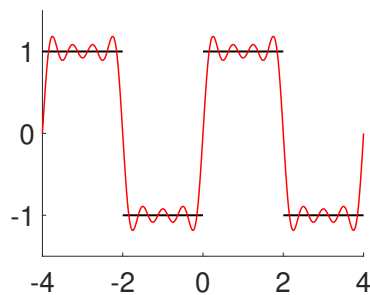
**Example 33.5.5** Let  $f(x) = 1$  on  $(0, 2]$  and  $f(x) = -1$  on  $(-2, 0]$  and  $f(x+4) = f(x)$ .

First note that  $L = 2$ . In this case, the function is odd and so all the  $a_k = 0$ .

$$b_k = \frac{1}{2} \int_{-2}^2 f(x) \sin\left(\frac{k\pi x}{2}\right) dx = \int_0^2 \sin\left(\frac{k\pi x}{2}\right) dx$$

Then  $b_k = \frac{2}{\pi k} (1 - (-1)^k)$ . Thus for  $k$  even, this is 0. For  $k$  odd, this is  $\frac{4}{\pi k}$ . It follows the Fourier series is

$$\sum_{k=1}^{\infty} \frac{4}{\pi (2k-1)} \sin\left(\frac{(2k-1)\pi x}{2}\right)$$



In the picture, is a graph of the addition of the first four terms of the Fourier series along with part of the function. Notice the way the Fourier series is struggling to do the impossible, approximate uniformly a discontinuous function with one which is very smooth. That little blip near the jump in the function will never go away by taking more terms in the sum.

Note that if you take  $x = 1$  the series must converge to 1. Therefore,

$$1 = \sum_{k=1}^{\infty} \frac{4}{\pi (2k-1)} (-1)^{k-1}$$

It follows that

$$\frac{\pi}{4} = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{2k-1}$$

This is another remarkable assertion.

### 33.5.2 Mean Square Convergence

It is the case that if  $f$  is Riemann integrable and  $2L$  periodic, then the Fourier series converges to the function  $f$  in the mean square sense. That is

$$\lim_{n \rightarrow \infty} \int_{-L}^L |f(x) - S_n f(x)|^2 dx = 0$$

I will show this now, leaving out a few details which will be reasonable to believe. Suppose that  $f$  is continuous and periodic with period  $2L$ . The Cesaro means of  $f$  are defined as follows.

$$\sigma_n f(x) \equiv \frac{1}{n+1} \sum_{k=0}^n S_k f(x), \quad S_0 f(x) = a_0 \equiv \frac{1}{2L} \int_{-L}^L f(x) dx$$

Thus, from what was shown above,

$$\begin{aligned} \sigma_n f(x) &= \frac{1}{n+1} \sum_{k=0}^n \int_{-L}^L D_k(x-y) f(y) dy \\ &= \int_{-L}^L \left( \frac{1}{n+1} \sum_{k=0}^n D_k(x-y) \right) f(y) dy \end{aligned}$$

Then the Fejer kernel is

$$F_n(t) = \frac{1}{n+1} \sum_{k=0}^n D_k(t) \quad (*)$$

We compute this now. Recall that

$$D_n(t) = \frac{\sin\left(\left(n + \frac{1}{2}\right) \frac{\pi t}{L}\right)}{2L \sin\left(\frac{\pi t}{2L}\right)}$$

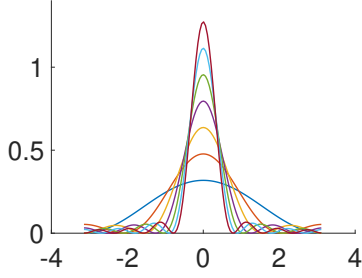
Thus

$$\begin{aligned} \sin^2\left(\frac{\pi t}{2L}\right) F_n(t) &= \frac{1}{2L} \frac{1}{n+1} \sum_{k=0}^n \sin\left(\frac{\pi t}{2L}\right) \sin\left(\left(k + \frac{1}{2}\right) \frac{\pi t}{L}\right) \\ &= \frac{1}{2L(n+1)} \frac{1}{2} \sum_{k=0}^n \left[ \cos\left(\left(k + \frac{1}{2}\right) \frac{\pi t}{L} - \left(\frac{\pi t}{2L}\right)\right) - \cos\left(\frac{\pi t}{2L} + \left(k + \frac{1}{2}\right) \frac{\pi t}{L}\right) \right] \\ &= \frac{1}{2L} \frac{1}{n+1} \frac{1}{2} \sum_{k=0}^n \left[ \cos\left(\frac{\pi k t}{L}\right) - \cos\left(\frac{\pi t}{L}(k+1)\right) \right] \\ &= \frac{1}{4L(n+1)} \left( 1 - \cos\left(\frac{\pi t}{L}(n+1)\right) \right) \end{aligned}$$

Thus

$$F_n(t) = \frac{1}{4L(n+1)} \frac{\left(1 - \cos\left(\frac{\pi t}{L}(n+1)\right)\right)}{\sin^2\left(\frac{\pi t}{2L}\right)} \quad (**)$$

Here are graphs of  $F_n(t)$  for  $n = 1, 2, \dots, 7$  for  $L = \pi$ . Notice how they are nonnegative and are large on a small interval containing 0. As you increase  $n$ , the bump in the middle gets taller.



There are certain properties which are obvious. First of all,  $\int_{-L}^L F_n(t) dt = 1$ . This follows from \* and 33.10 which pertained to the Dirichlet kernel. Another which is obvious from the above formula \*\* is that  $F_n(t) \geq 0$ . Finally, for any small  $\delta > 0$ , if  $|t| > \delta$  then

$$F_n(t) \leq \frac{1}{4L(n+1)} \frac{(1 - \cos(\frac{\pi}{L}t(n+1)))}{\sin^2(\frac{\pi}{2L}\delta)} \quad (***)$$

It follows from periodicity that for  $M \geq \max_x |f(x)|$

$$\begin{aligned} |f(x) - \sigma_n f(x)| &= \left| \int_{-L}^L (f(x) - f(t)) F_n(x-t) dt \right| \\ &= \left| \int_{-L}^L (f(x) - f(x-u)) F_n(u) du \right| \\ &\leq \int_{|u| \geq \delta} |f(x) - f(x-u)| F_n(u) du + 2M \int_{|u| < \delta} |f(x) - f(x-u)| F_n(u) du \\ &\leq 2M \frac{1}{4L(n+1)} \frac{2}{\sin^2(\frac{\pi}{2L}\delta)} + 2M \int_{|u| < \delta} |f(x) - f(x-u)| F_n(u) du \end{aligned}$$

Now if  $\varepsilon > 0$ , there is  $\delta > 0$  such that if  $|u| < \delta$ , then for all  $x$ ,  $|f(x) - f(x-u)| < \varepsilon/2$ . Thus for such a choice of  $\delta$  and \*\*\*,

$$|f(x) - \sigma_n f(x)| \leq 2M \frac{1}{4L(n+1)} \frac{2}{\sin^2(\frac{\pi}{2L}\delta)} + \frac{\varepsilon}{2}$$

and so, if  $n$  is large enough, you get

$$|f(x) - \sigma_n f(x)| < \varepsilon,$$

this for any  $x$ . Thus the convergence of  $\sigma_n f(x)$  to  $f(x)$  is uniform. It follows that

$$\lim_{n \rightarrow \infty} \int_{-L}^L |f(x) - \sigma_n f(x)|^2 dx = 0$$

This shows the following interesting result.

**Proposition 33.5.6** *If  $f$  is continuous and  $2L$  periodic, then the Cesaro means converge uniformly to  $f$  and also they converge to  $f$  in the mean square sense.*

From this, it is not hard to establish that the Cesaro means converge in mean square to any  $2L$  periodic function  $f$  which is Riemann integrable on intervals of length  $2L$ . To do this, you argue that, given a Riemann integrable function which is  $2L$  periodic, there exists a continuous function which is close to it in the mean square norm. Then apply the above proposition to this continuous function and get a Cesaro mean close to it in mean square which is close to the original function in mean square sense.

The Cesaro means are trig polynomials of the form

$$a_0 + \sum_{k=1}^n a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \sin\left(\frac{k\pi x}{L}\right).$$

One of these can be made as close as desired to  $f$  in the mean square sense. Hence the corresponding Fourier series is even closer, by the above section on mean square approximation. Thus, for every  $\varepsilon > 0$  there exists  $N$  such that if  $n > N$ , then

$$\int_{-L}^L |S_n f(x) - f(x)|^2 dx < \varepsilon$$

which says the Fourier series converge in the mean square sense to  $f$ .

Note that the above proposition also shows an improved result about pointwise convergence. The function  $f$  did not need to have any control on its derivative and yet the Cesaro means converged uniformly to the function. If the function were piecewise continuous, the Cesaro means would converge to the mid point of the jump with no condition on the derivatives from left or right. This is easy to show but is as far as this will be taken here. If you want uniform approximation using trigonometric series, you should not be using the Fourier series. You should use the Cesaro means.

### 33.6 Integrating and Differentiating Fourier Series

Suppose that  $f$  is  $2L$  periodic and piecewise continuous. This is defined next.

**Definition 33.6.1** Let  $f$  be a bounded function defined on  $[a, b]$ . It is called piecewise continuous if there is a partition of  $[a, b]$ ,  $\{x_0, \dots, x_n\}$  and for each  $k$ , a continuous function  $g_k$  such that  $f(x) = g_k(x)$  for all  $x \in (x_{k-1}, x_k)$ .

It turns out that you can integrate a Fourier series term by term. This is generally true but I will show it here for piecewise continuous  $2L$  periodic functions. Let  $f$  be such a function equal to a continuous function on  $[x_i, x_{i+1}]$  for  $i \leq n$ . Then consider

$$G(x) \equiv \int_{-L}^x (f(t) - a_0) dt$$

where  $a_0$  is the Fourier coefficient

$$a_0 = \frac{1}{2L} \int_{-L}^L f(t) dt$$

Thus  $G(-L) = G(L) = 0$  and if we continue using  $G$  to denote the  $2L$  periodic extension, it follows from Theorem 33.5.2 that the Fourier series of  $G$

$$A_0 + \sum_{n=1}^{\infty} A_n \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{L}\right)$$

converges to  $G$  at every point. This is because  $|f(t) - a_0|$  is bounded by some  $M$  due to the assumption that it is piecewise continuous and the observation that

$$|G(x) - G(\hat{x})| \leq \left| \int_{\hat{x}}^x |f(t) - a_0| \right| \leq M|x - \hat{x}|$$

Then plugging in  $\pi$  to the Fourier series for  $G$  we get

$$0 = A_0 + \sum_{n=1}^{\infty} A_n (-1)^n, \quad A_0 = - \sum_{n=1}^{\infty} A_n (-1)^n \quad (*)$$

Next consider  $A_n, n > 0$ .

$$\begin{aligned} LA_n &= \int_{-L}^L \int_{-L}^x (f(t) - a_0) dt \cos\left(\frac{n\pi x}{L}\right) dx \\ &= \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \left( \int_{-L}^{x_k} (f(t) - a_0) dt + \int_{x_k}^x (f(t) - a_0) dt \right) \cos\left(\frac{n\pi x}{L}\right) dx \\ &= \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \int_{-L}^{x_k} (f(t) - a_0) dt \cos\left(\frac{n\pi x}{L}\right) dx \\ &\quad + \sum_{k=0}^{n-1} \left( \frac{L}{n\pi} \sin\left(\frac{n\pi x}{L}\right) \int_{x_k}^x (f(t) - a_0) dt \Big|_{x_k}^{x_{k+1}} \right) \\ &= \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} G(x_k) \cos\left(\frac{n\pi x}{L}\right) dx + \sum_{k=0}^{n-1} \frac{L}{n\pi} \sin\left(\frac{n\pi x_{k+1}}{L}\right) (G(x_{k+1}) - G(x_k)) \\ &\quad - \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \frac{L}{n\pi} \sin\left(\frac{n\pi x}{L}\right) (f(x) - a_0) dx \end{aligned}$$

Now do an integration on the first sum. This yields

$$\begin{aligned} &\frac{L}{n\pi} \sum_{k=0}^{n-1} G(x_k) \left( \sin\left(\frac{n\pi x_{k+1}}{L}\right) - \sin\left(\frac{n\pi x_k}{L}\right) \right) \\ &+ \frac{L}{n\pi} \sum_{k=0}^{n-1} \sin\left(\frac{n\pi x_{k+1}}{L}\right) (G(x_{k+1}) - G(x_k)) \\ &- \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \frac{L}{n\pi} \sin\left(\frac{n\pi x}{L}\right) (f(x) - a_0) dx \end{aligned}$$

The sums simplify and the result one obtains is

$$\begin{aligned} &\frac{L}{n\pi} \sum_{k=0}^{n-1} G(x_{k+1}) \sin\left(\frac{n\pi x_{k+1}}{L}\right) - G(x_k) \sin\left(\frac{n\pi x_k}{L}\right) \\ &- \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \frac{L}{n\pi} \sin\left(\frac{n\pi x}{L}\right) (f(x) - a_0) dx \end{aligned}$$

The series telescopes and the result is 0 because  $G(L) = G(-L) = 0$ . Thus the result of it all is

$$\begin{aligned} LA_n &= - \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} \frac{L}{n\pi} \sin\left(\frac{n\pi x}{L}\right) (f(x) - a_0) dx \\ &= \int_{-L}^L - \frac{L}{n\pi} \sin\left(\frac{n\pi x}{L}\right) (f(x) - a_0) dx \end{aligned}$$

Thus

$$A_n = -\frac{L}{n\pi} \frac{1}{L} \int_{-L}^L \sin\left(\frac{n\pi x}{L}\right) (f(x) - a_0) dx = -\frac{L}{n\pi} b_n$$

Similar computations will show that for  $n > 0$ ,

$$B_n = \frac{L}{n\pi} \frac{1}{L} \int_{-L}^L \cos\left(\frac{n\pi x}{L}\right) (f(x) - a_0) dx = \frac{L}{n\pi} a_n$$

where  $a_n, b_n$  are, respectively, the cosine and sine Fourier coefficients of  $f$ . Thus we have from \*,

$$\begin{aligned} G(x) &= \int_{-L}^x (f(t) - a_0) dt = -\sum_{n=1}^{\infty} \left( -\frac{L}{n\pi} b_n \right) (-1)^n + \\ &\quad \sum_{n=1}^{\infty} -\frac{L}{n\pi} b_n \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} \frac{L}{n\pi} a_n \sin\left(\frac{n\pi x}{L}\right) \end{aligned}$$

Hence

$$\begin{aligned} \int_{-L}^x (f(t) - a_0) dt &= \sum_{n=1}^{\infty} \frac{L b_n}{n\pi} \left( \cos\left(\frac{-n\pi L}{L}\right) - \cos\left(\frac{n\pi x}{L}\right) \right) \\ &\quad + \sum_{n=1}^{\infty} \frac{L}{n\pi} a_n \sin\left(\frac{n\pi x}{L}\right) \end{aligned}$$

Thus

$$\begin{aligned} \int_{-L}^x f(t) dt &= \int_{-L}^x a_0 dt + \sum_{n=1}^{\infty} b_n \int_{-L}^x \sin\left(\frac{n\pi t}{L}\right) dt \\ &\quad + \sum_{n=1}^{\infty} a_n \int_{-L}^x \cos\left(\frac{n\pi t}{L}\right) dt \end{aligned}$$

This proves the following theorem.

**Theorem 33.6.2** *Let  $f$  be piecewise continuous and  $2L$  periodic. Then for every  $x \in [-L, L]$ ,*

$$\begin{aligned} \int_{-L}^x f(t) dt &= \int_{-L}^x a_0 dt + \sum_{n=1}^{\infty} b_n \int_{-L}^x \sin\left(\frac{n\pi t}{L}\right) dt \\ &\quad + \sum_{n=1}^{\infty} a_n \int_{-L}^x \cos\left(\frac{n\pi t}{L}\right) dt \end{aligned}$$

where  $a_0, a_k, b_k$  are the Fourier coefficients for  $f$ .

Note that there is nothing which says that the Fourier series of  $f$  converges to  $f$ ! This is a wonderful result.

You can't expect to be able to differentiate Fourier series. See the exercises. However, there is something which can be said. Suppose for  $x \in [-L, L)$

$$f(x) = f(-L) + \int_{-L}^x f'(t) dt$$



and that  $f'$  is piecewise continuous and  $2L$  periodic. Let  $f$  denote the  $2L$  periodic extension of the above  $f$ . Then let the formal Fourier series for  $f'$  be

$$a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right)$$

Then by Theorem 33.6.2,

$$\begin{aligned} \int_{-L}^x f'(t) dt &= \int_{-L}^x a_0 dt + \sum_{n=1}^{\infty} a_n \frac{L}{n\pi} \sin\left(\frac{n\pi x}{L}\right) \\ &\quad + \sum_{n=1}^{\infty} b_n \frac{L}{n\pi} \left((-1)^n - \cos\left(\frac{n\pi x}{L}\right)\right) \end{aligned}$$

Then  $a_0 \equiv \frac{1}{2L} \int_{-L}^L f'(t) dt = \frac{1}{2L} (f(L) - f(-L)) = 0$ .

$$\begin{aligned} a_n &\equiv \frac{1}{L} \int_{-L}^L f'(t) \cos\left(\frac{n\pi t}{L}\right) dt = \frac{1}{L} f(t) \cos\left(\frac{n\pi t}{L}\right) \Big|_{-L}^L \\ &\quad + \frac{1}{L} \frac{n\pi}{L} \int_{-L}^L f(t) \sin\left(\frac{n\pi t}{L}\right) dt \\ &= \frac{1}{L} \frac{n\pi}{L} \int_{-L}^L f(t) \sin\left(\frac{n\pi t}{L}\right) dt = B_n \frac{n\pi}{L} \end{aligned}$$

where  $B_n$  is the Fourier coefficient for  $f(t)$ . Similarly,

$$b_n = \frac{1}{L} \int_{-L}^L f'(t) \sin\left(\frac{n\pi t}{L}\right) dt = -\frac{1}{L} \frac{n\pi}{L} \int_{-L}^L f(t) \cos\left(\frac{n\pi t}{L}\right) dt = -\frac{n\pi}{L} A_n$$

where  $A_n$  is the  $n^{\text{th}}$  cosine Fourier coefficient for  $f$ . Thus

$$\begin{aligned} \int_{-L}^x f'(t) dt &= \sum_{n=1}^{\infty} B_n \frac{n\pi}{L} \frac{L}{n\pi} \sin\left(\frac{n\pi x}{L}\right) \\ &\quad + \sum_{n=1}^{\infty} \left(-\frac{n\pi}{L} A_n\right) \frac{L}{n\pi} \left((-1)^n - \cos\left(\frac{n\pi x}{L}\right)\right) \\ f(x) - f(-L) &= \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} A_n \cos\left(\frac{n\pi x}{L}\right) - \sum_{n=1}^{\infty} A_n (-1)^n \\ f(x) &= \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} A_n \cos\left(\frac{n\pi x}{L}\right) + \left(f(-L) - \sum_{n=1}^{\infty} A_n (-1)^n\right) \end{aligned}$$

Thus that constant on the end is  $A_0$ . It follows that

$$f(x) = A_0 + \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} A_n \cos\left(\frac{n\pi x}{L}\right)$$

and  $-\frac{n\pi}{L} A_n = b_n$ ,  $B_n \frac{n\pi}{L} = a_n$  and so

$$\begin{aligned} f'(x) &= \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) \\ &= \sum_{n=1}^{\infty} B_n \frac{n\pi}{L} \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} A_n \left(-\frac{n\pi}{L}\right) \sin\left(\frac{n\pi x}{L}\right) \\ &= \sum_{n=1}^{\infty} B_n \frac{d}{dx} \sin\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} A_n \frac{d}{dx} \cos\left(\frac{n\pi x}{L}\right) \end{aligned}$$

This proves the following.

**Theorem 33.6.3** *Let  $f$  denote the  $2L$  periodic extension of the function  $f$  given on  $[-L, L)$  by*

$$f(x) = f(-L) + \int_{-L}^x f'(t) dt$$

*and suppose  $f'$  is  $2L$  periodic and piecewise continuous. Then for each  $x \in [-L, L]$ ,*

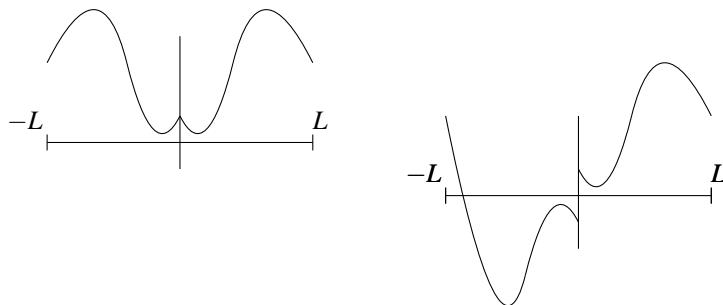
$$f(x) = A_0 + \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} A_n \cos\left(\frac{n\pi x}{L}\right)$$

*where the  $A_k, B_k$  are the Fourier coefficients of  $f$  and the Fourier series for  $f'$  is*

$$\sum_{n=1}^{\infty} B_n \frac{d}{dx} \sin\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} A_n \frac{d}{dx} \cos\left(\frac{n\pi x}{L}\right)$$

### 33.7 Odd and Even Extensions

Often, as in the above examples and in the applications which follow, the function you are finding the Fourier series for is either even or odd. One way this often occurs is when the function of interest is defined on an interval  $[0, L]$  and it is only its values on this interval which are of interest. Then you could consider either the even or the odd extension of this function to  $[-L, L]$  and then extend it to be a  $2L$  periodic function. For example, consider the following pictures.



The first of these is an even extension to  $[-L, L]$  and the second is an odd extension to  $[-L, L]$ . In the first case where there is an even extension, the Fourier coefficients are  $b_k = 0$

$$\begin{aligned} a_0 &= \frac{1}{2L} \int_{-L}^L f(x) dx = \frac{1}{L} \int_0^L f(x) dx \\ a_k &= \frac{1}{L} \int_{-L}^L f(x) \cos\left(\frac{k\pi x}{L}\right) dx = \frac{2}{L} \int_0^L f(x) \cos\left(\frac{k\pi x}{L}\right) dx \end{aligned}$$

In the second case where you are dealing with the odd extension, each  $a_k = 0$  and

$$b_k = \frac{1}{L} \int_{-L}^L f(x) \sin\left(\frac{k\pi x}{L}\right) dx = \frac{2}{L} \int_0^L f(x) \sin\left(\frac{k\pi x}{L}\right) dx$$

**Example 33.7.1** *Let  $f(x) = x$  on  $[0, 1]$ . Find the Fourier series of its even extension.*

Its even extension is nothing more than the function of Example 33.5.4. This is

$$\frac{1}{2} - \sum_{k=1}^{\infty} \frac{4}{\pi^2 (2k-1)^2} \cos(2k-1)\pi x$$

**Example 33.7.2** Let  $f(x) = x$  on  $[0, 1]$ . Find the Fourier series of its odd extension which is periodic of period 2.

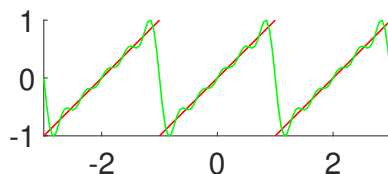
This would be the function  $f(x) = x$  on  $(-1, 1)$  extended to be periodic of period 2. Thus  $L = 1$ . Since it is an odd function, all the  $a_k = 0$  and from the above,

$$b_k = 2 \int_0^1 x \sin(k\pi x) dx = \frac{2}{\pi k} (-1)^{k+1}$$

Then the Fourier series is

$$\sum_{k=1}^{\infty} \frac{2}{\pi k} (-1)^{k+1} \sin(k\pi x)$$

The graph of the sum of the first five terms is given.



Note the difficulty in handling the jump with the little bump right before the discontinuity. This illustrates that if you are only interested in the function on  $[0, 1]$ , it would be better to use the even extension than the odd extension. However, in the applications, you don't get to choose.

Also, note that, unlike power series, Fourier series are attempting to approximate a function on a whole interval, not just near a single point. This is much more interesting.

There is a general sort of problem called a Sturm-Liouville problem discussed in Problem 13. It turns out that there are general theorems about convergence of expansions in terms of eigenfunctions to such problems [38]. However, you can often see that convergence in the mean square sense will hold from observing that the Fourier series for the eigenfunctions will converge because it is the restriction of the Fourier series of an even or odd extension as discussed in this section. For many other considerations on Sturm-Liouville problems, the old book by Ince [23] is very good. These problems have been intensively studied since around 1830.

## 33.8 Exercises

1. Let  $f(x)$  be the even extension of  $\sin x$ . Find the Fourier series and at  $x = \pi/2$  write a series which says that the Fourier series converges to the function at this point. Note that here  $L = \pi$  and so  $n\pi/L = n$ .
2. Let  $f(x)$  be the odd  $2\pi$  periodic extension of  $y = x^2$ . Find what the sum converges to at  $x = \pi/2$ . Again  $L = \pi$  and so  $n\pi/L = n$ .

3. Let  $f(x)$  be the even  $2\pi$  periodic extension of  $y = x^2$ . Find the Fourier coefficients and obtain an interesting series by letting  $x = \pi$ .
4. In Example 33.5.5 the Fourier series was found for the function  $f$  which is 1 on  $[0, 2]$  and  $-1$  on  $(-2, 0)$ .

$$\sum_{k=1}^{\infty} \frac{4}{\pi(2k-1)} \sin\left(\frac{(2k-1)\pi x}{2}\right)$$

This function has a jump so it is not differentiable at  $0, 2, 4$ , etc. However, it is differentiable at most points, other than a few jumps. Furthermore, the Fourier series converges to the function at these points. Can you differentiate the Fourier series term by term and get something which converges to the derivative of the function? What does this show about interchange of limits?

5. In one of the problems above, you found that the Fourier series for the  $2\pi$  periodic extension of  $y = x^2$  is  $\frac{\pi^2}{3} + \sum_{k=1}^{\infty} 4 \frac{(-1)^k}{k^2} \cos(kx)$ . The derivative of this function,  $y = 2x$  is sure piecewise continuous. Find the Fourier series expansion for  $y = 2x$  without any effort.
6. Find a Fourier series which converges to the  $2\pi$  periodic extension of

$$\int_{-\pi}^x \left(t^2 - \frac{\pi^2}{3}\right) dt = \frac{1}{3}x^3 - \frac{1}{3}\pi^2 x.$$

7. Suppose  $f$  is periodic with period  $2L$ . Does it follow that  $f'$  is also periodic of period  $2L$ ? Explain.
8. Here are some boundary value problems. Find nonzero solutions if there are any or determine that there are none.

$$(a) \quad y'' + \frac{1}{4}\pi^2 y = 0, \\ y(0) = 0, y(2) = 0$$

$$(e) \quad y'' + \left(\frac{1}{2}\pi\right)^2 y = 0, \\ y(0) = 0, y(1) = 0$$

$$(b) \quad y'' + \left(\frac{7\pi}{5}\right)^2 y = 0, \\ y(0) = 0, y\left(\frac{5}{2}\right) = 0$$

$$(f) \quad y'' + \pi^2 y = 0, \\ y(0) = 0, y(1) = 0$$

$$(c) \quad y'' + \left(\frac{5\pi}{3}\right)^2 y = 0, \\ y(0) = 0, y\left(\frac{3}{2}\right) = 0$$

$$(g) \quad y'' + \left(\frac{4\pi}{2}\right)^2 y = 0, \\ y(0) = 0, y\left(\frac{2}{2}\right) = 0$$

$$(d) \quad y'' + \left(\frac{2\pi}{7}\right)^2 y = 0, \\ y(0) = 0, y\left(\frac{7}{2}\right) = 0$$

$$(h) \quad y'' + \frac{9}{25}\pi^2 y = 0, \\ y(0) = 0, y\left(\frac{5}{2}\right) = 0$$

9. Here are some boundary value problems. Find nonzero solutions if there are any or determine that there are none.

$$(a) \quad y'' + \left(\frac{11\pi}{2}\right)^2 y = 0, \\ y(0) = 0, y'\left(\frac{2}{2}\right) = 0$$

$$(c) \quad y'' + \left(\frac{2\pi}{7}\right)^2 y = 0, \\ y(0) = 0, y'\left(\frac{7}{2}\right) = 0$$

$$(b) \quad y'' + \left(\frac{4\pi}{5}\right)^2 y = 0, \\ y(0) = 0, y'\left(\frac{5}{2}\right) = 0$$

$$(d) \quad y'' + \left(\frac{7\pi}{3}\right)^2 y = 0, \\ y(0) = 0, y'\left(\frac{3}{2}\right) = 0$$

$$\begin{array}{ll}
 \text{(e) } y'' + \left(\frac{\pi}{13}\right)^2 y = 0, & y(0) = 0, y'\left(\frac{5}{2}\right) = 0 \\
 y(0) = 0, y'\left(\frac{13}{2}\right) = 0 & \text{(g) } y'' + \left(\frac{5\pi}{11}\right)^2 y = 0, \\
 \text{(f) } y'' + \left(\frac{2\pi}{5}\right)^2 y = 0, & y(0) = 0, y'\left(\frac{11}{2}\right) = 0
 \end{array}$$

10. In boundary value problems like the above, why is it that there is either no nonzero solution or infinitely many?
11. In the study of buckling beams, you have an equation

$$y^{(4)}(x) + \lambda y''(x) = 0, x \in [0, L]$$

along with boundary conditions like

$$\begin{array}{ll}
 y(0) &= y'(0) = 0, \text{ clamped at left end} \\
 y(L) &= 0 = y''(L), \text{ hinged at right end}
 \end{array}$$

where  $\lambda$  increases with the axial force and depends on geometrical and physical properties of the beam. The idea is to find values of  $\lambda$  for which there is a nonzero solution to the differential equation and the boundary conditions. Assume all boundary conditions considered have  $y(0) = y(L) = 0$  and at each end, either  $y'$  or  $y''$  is equal to 0. Thus one considers beams for which each end is either clamped or hinged. Show that if  $\lambda$  is such that there exists a nonzero solution, then  $\lambda > 0$ . **Hint:** You show this by multiplying the equation by  $y$  and integrating by parts.

12. Letting  $\lambda = \delta^2$ , in the above problem, show that there exist infinitely many values for  $\delta$  and corresponding nonzero solutions to the boundary value problem for the following situation.

$$\begin{array}{ll}
 \text{(a) } y(0) = y'(0) = 0, y(L) = y'(L) = 0 \\
 \text{(b) } y(0) = y''(0) = 0, y(L) = y'(L) = 0 \\
 \text{(c) } y(0) = y''(0) = 0, y(L) = y''(L) = 0
 \end{array}$$

13. A **Sturm-Liouville problem** involves the differential equation for an unknown function of  $x$  which is denoted here by  $y$ ,

$$(p(x)y'(x))' + (\lambda q(x) + r(x))y = 0, x \in [a, b]$$

and it is assumed that  $p(t), q(t) \geq 0$  and are nonzero except for finitely many points in  $[a, b]$  for any  $t$  along with boundary conditions,

$$\begin{array}{ll}
 C_1 y(a) + C_2 y'(a) &= 0 \\
 C_3 y(b) + C_4 y'(b) &= 0
 \end{array}$$

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  $\lambda$  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

$$\int_a^b q(x) y(x) z(x) dx = 0. \quad (33.11)$$

**Hint:** Do something like this:

$$(p(x)y')'z + (\lambda_1 q(x) + r(x))yz = 0,$$

$$(p(x)z')'y + (\lambda_2 q(x) + r(x))zy = 0.$$

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 33.11 is called an orthogonality relation and it makes possible an expansion in terms of certain functions called eigenfunctions.

14. Here is a really nice result. Suppose you have  $y, z$  are both solutions of the differential equation

$$(p(x)y'(x))' + q(x)y(x) = 0$$

Show that  $p(x)W(y, z)(x) = C$  a constant. Here  $W(y, z)$  is the Wronskian.

15. In the above problem, change the variables as follows. Let  $z(x) = p(x) \frac{y'(x)}{y(x)}$  and determine the equation which results for  $z$ . This kind of equation is called a Riccati equation. In particular, show that

$$z' + \frac{1}{p(x)}z^2 + q(x) = 0$$

This kind of equation is like a Bernoulli equation with exponent 2, but with another function added in. For more on this, see [29].

16. Suppose in the equation of Problem 14 you have two solutions  $u, v$  whose Wronskian is nonzero so they are independent solutions. Suppose that  $a, b$  are consecutive zeros of  $u$  and that  $p(x) > 0$  on  $[a, b]$ . Show that  $v$  has exactly one zero in  $(a, b)$ . This is called the Sturm separation theorem. **Hint:** Use the result of the above mentioned problem and argue that  $v(a) \neq 0$  and that you can assume that  $v(a) > 0$  and that  $u$  is positive on the open interval  $(a, b)$ .
17. Letting  $[a, b] = [-\pi, \pi]$ , consider an example of a 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 13, 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.

18. Sometimes one encounters an eigenvalue problem of the form

$$x^2 y'' + xy' + (\lambda x^2 - n^2)y = 0, \quad C_1 y(L) + C_2 y'(L) = 0$$

not both  $C_i$  equal zero and  $y$  bounded near 0. Discover an orthogonality relation between this solution and one for which  $\lambda$  is changed to  $\mu$ . **Hint:** You might divide by  $x$ .

19. Let  $x \rightarrow J_n(x)$  be a solution to the Bessel equation

$$x^2 y'' + xy' + (x^2 - n^2)y = 0$$

and suppose  $\alpha$  is a positive number. Let  $z(x) \equiv J_n(\alpha x)$ . Find a differential equation satisfied by  $z$ . You should show that it satisfies  $x^2 z'' + xz + (\alpha^2 x^2 - n^2)z = 0$ .

20. Let  $\alpha, \beta$  be two zeros of the Bessel function  $J_n(x)$ . It was shown in Proposition 32.6.1 on Page 629 that there are infinitely many of these zeros. Now consider the two functions  $x \rightarrow J_n\left(\frac{\alpha}{L}x\right), x \rightarrow J_n\left(\frac{\beta}{L}x\right)$ . Show that

$$\int_0^L J_n\left(\frac{\alpha}{L}x\right) J_n\left(\frac{\beta}{L}x\right) x dx = 0$$

21. Consider

$$x^2 y'' + xy' + (\delta^2 x^2 - n^2)y = 0, \quad y(L) = 0, y \text{ bounded near } 0$$

Show that there are only certain values of  $\delta$  which work and they are of the form  $\delta^2 = (\alpha/L)^2$  where  $\alpha$  is some zero of a solution to Bessel's equation.

22. Show that the only eigenvalues  $\lambda$  for

$$x^2 y'' + xy' + (\lambda x^2 - n^2)y = 0, \quad y(L) = 0$$

are positive.

23. Recall that for  $n$  an integer, the general solution to Bessel's equation is  $C_1 J_n(x) + C_2 Y_n(x)$  where  $Y_n$  is unbounded at 0. Using the above problem, characterize all eigenvalues  $\lambda$  of the eigenvalue problem

$$x^2 y'' + xy' + (\lambda x^2 - n^2)y = 0, \quad y(L) = 0, y \text{ bounded near } 0.$$

and describe all solutions to this boundary value problem in terms of Bessel functions. **Hint:** Rule out  $Y_n$  to begin with. Then consider  $z(\sqrt{\lambda}x) = y(x)$  for  $y$  a solution to the above Sturm-Liouville equation.

24. A Sturm-Liouville eigenvalue problem involves the equation

$$(p(x)y'(x))' + (\lambda q(x) + r(x))y = 0, \quad x \in (a, b)$$

The Liouville transformation is

$$z = (p(x)q(x))^{1/4}y, \quad t = \int_c^x \left(\frac{q(s)}{p(s)}\right)^{1/2} ds, \quad c \in (a, b)$$

Then determine the equation solved by  $z$ . **Hint:** This is a little involved. First verify that the left side reduces to

$$\frac{d}{dt} \left( p \frac{d}{dx} \left( (pq)^{-1/4} \right) z + (pq)^{1/4} \frac{dz}{dt} \right) \sqrt{\frac{q}{p}} + (\lambda q + r) (pq)^{-1/4} z = 0$$

Next verify that the  $z'(t)$  terms all cancel. That way, in the above, you can neglect these terms in using the product rule. This leads to

$$\left( \frac{\frac{d}{dx} \left( -\frac{1}{4} p^{-1/4} q^{-5/4} \frac{d}{dx} (pq) \right)}{(p^{-1/4} q^{3/4})} + \frac{r(pq)^{-1/4}}{(p^{-1/4} q^{3/4})} \right) z + z'' + \lambda z = 0$$

Now argue that the equation is of the form

$$z'' + (\lambda + m(t))z = 0$$

where  $m(t)$  is a function which depends on  $p, q$ .

25. Consider the eigenvalue problem for Bessel's equation,

$$x^2 y'' + xy' + (\lambda x^2 - n^2) y = 0, \quad y(L) = 0$$

Show it can be written in self adjoint form as

$$(xy')' + \left( \lambda x - \frac{n^2}{x} \right) y = 0$$

Thus in this case,  $q(x) = x$  and  $r(x) = -n^2/x$ . What is the form of the equation if Liouville's transformation is applied to this Bessel eigenvalue problem? **Hint:** Just use the specific description of what was obtained above and that  $r(x) = -n^2/x, p(x) = q(x) = x$ , and so  $t = x$ . You should get something like

$$z'' + \lambda z + \left( \frac{1 - 4n^2}{4x^2} \right) z = 0$$

26. In the above problem, let  $\lambda = 1$  and let  $n = 1/2$  and use to find the general solution to the Bessel equation in which  $\nu = 1/2$ . Show, using the above, that this general solution is of the form

$$C_1 x^{-1/2} \cos x + C_2 x^{-1/2} \sin x.$$

27. Show that the polynomial  $q(x)$  of degree  $n$  which minimizes

$$\int_{-1}^1 |f(x) - p(x)|^2 dx$$

out of all polynomials  $p$  of degree  $n$  is the  $n^{\text{th}}$  partial sum of the Fourier series taken with respect to the Legendre polynomials  $q(x) = S_n f(x)$ , where  $S_n f(x) \equiv \sum_{k=0}^n c_k p_k(x)$ ,  $c_k = \int_{-1}^1 q_k(x) f(x) dx$ .



28. Recall the normalized Legendre polynomials

$$q_n(x) = \frac{\sqrt{2n+1}}{\sqrt{2}} p_n(x)$$

which have the property that

$$\int_{-1}^1 q_j(x) q_k(x) dx = \delta_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

If  $f$  is a Riemann integrable function, show that

$$\lim_{n \rightarrow \infty} \int_{-1}^1 f(x) q_n(x) dx = 0$$

29. Show that if  $f$  is any continuous function on  $[-1, 1]$ , then the Fourier series in terms of Legendre polynomials converges to  $f$  in the mean square sense. This means that for  $S_n f(x) \equiv \sum_{k=0}^n c_k p_k(x)$ ,  $c_k = \int_{-1}^1 q_k(x) f(x) dx$ , it follows that

$$\lim_{n \rightarrow \infty} \int_{-1}^1 |f(x) - S_n f(x)|^2 dx = 0$$

30. It can be shown that there are no continuous, nonzero solutions to Legendre's equation

$$((1-x^2)y')' + \lambda y = 0$$

defined on  $[-1, 1]$  unless  $\lambda = n(n+1)$  for  $n$  an integer. Use the above problem to show this.

31. One of the applications of Fourier series is to obtain solutions to linear differential equations which have a periodic right side. This is done by expanding the right side which is a forcing function in a Fourier series, solving the simple equation which corresponds to each term and then adding these solutions to obtain what is hoped to be a representation of the solution. Find a particular solution for each of the following. Let

$$y'' + 3y = f(t),$$

where  $f(t)$  is the step function which is periodic of period 2 and equals  $-1$  on  $[-1, 0)$  and  $1$  on  $(0, 1]$ . Here are the steps. First find a Fourier series for  $f$ . Say  $\sum_{n=1}^{\infty} b_n \sin(n\pi x)$ . Then let  $y_n$  be the solution to

$$y_n'' + 3y_n = \sin(n\pi t)$$

and then hopefully, on neglecting mathematical issues, the solution to the original problem is

$$y(t) = \sum_{n=1}^{\infty} b_n y_n(t)$$

32. Explain why the above procedure should give a particular solution if mathematical issues related to interchange of limit operations are ignored.

33. This problem is tedious but maybe it is better to do it all at once than to repeat seemingly endless virtually identical problems. In this problem,  $a$  is positive and  $b$  is a nonzero real number while  $n$  is a nonnegative integer. Find the real and imaginary parts of a solution  $y$  to

$$y'' + 2ay' + by = \exp\left(i\frac{n\pi t}{L}\right)$$

using the method of undetermined coefficients. Show that the real part is

$$\frac{2\pi L^3 a n \sin \frac{\pi}{L} n t - \pi^2 L^2 n^2 \cos \frac{\pi}{L} n t}{L^4 b^2 + 4\pi^2 L^2 a^2 n^2 - 2\pi^2 L^2 b n^2 + \pi^4 n^4}$$

and the imaginary part of the solution is

$$\frac{(L^4 b - \pi^2 L^2 n^2) \sin \frac{\pi}{L} n t - 2\pi L^3 a n \cos \frac{\pi}{L} n t}{L^4 b^2 + 4\pi^2 L^2 a^2 n^2 - 2\pi^2 L^2 b n^2 + \pi^4 n^4}$$

Explain why the real part is a particular solution to

$$y'' + 2ay' + by = \cos\left(\frac{n\pi t}{L}\right)$$

and the imaginary part is a particular solution to

$$y'' + 2ay' + by = \sin\left(\frac{n\pi t}{L}\right)$$

In case  $n$  is 0, a solution is  $1/b$ .

34. Using the above problem, describe the solution after a long time to the equation

$$y'' + 2y' + 2y = f(t)$$

where  $f(t)$  is a periodic function which has the following Fourier series. Note that the transient terms will disappear due to the fact that  $a = 1$  is positive. Note that with the above problem, you could do many other examples in which  $a$  and  $b$  are not given as here.

$$(a) \sum_{n=1}^{\infty} \frac{1}{n^2} \cos\left(\frac{n\pi t}{3}\right) + \sum_{n=1}^{\infty} \frac{1}{1+n^2} \sin\left(\frac{n\pi t}{3}\right) + 3$$

$$(b) \sum_{n=1}^{\infty} e^{-n} \cos\left(\frac{n\pi t}{2}\right) + \sum_{n=1}^{\infty} \frac{1}{n^4} \sin\left(\frac{n\pi t}{2}\right) + 1$$

$$(c) \sum_{n=1}^{\infty} \frac{1}{n^3} \cos\left(\frac{n\pi t}{4}\right) + \sum_{n=1}^{\infty} \frac{1}{n^3+1} \sin\left(\frac{n\pi t}{4}\right) - 2$$

35. Suppose you have an undamped equation

$$y'' + 4y = f(t)$$

where  $f$  is periodic. Suppose in the Fourier expansion of  $f(t)$  there is a nonzero term which is of the form  $b \sin(2t)$ . Say it describes the transverse vibrations of a bridge in the center. What will likely happen to this bridge?

36. Consider the functions  $y_n(x) = \sin(n\pi x)$  on the interval  $[0, 2]$ . Show that these functions satisfy  $\int_0^2 y_n(x) y_m(x) dx$  is 1 if  $n = m$  and zero if  $n \neq m$ . Now consider using them to expand the function  $f(x) = x$  in a Fourier series. Thus you would have

$$\sum_{n=1}^{\infty} b_n \sin(n\pi x)$$

where

$$b_n = \int_0^2 x \sin(n\pi x) dx$$

Graph the sum of the first seven terms in this Fourier series expansion along with the function it is supposedly approximating. What does this tell you about being able to approximate with orthogonal functions? Now do the same problem with the orthonormal functions  $\sin(n\frac{\pi}{2}x)$ .

37. Recall that a sequence of functions defined on  $[a, b]$   $\{f_n\}$  converges to  $f$  in the mean square sense if

$$\lim_{n \rightarrow \infty} \int_a^b |f_n(x) - f(x)|^2 dx = 0$$

consider the function  $f_n(x)$  for  $x \in [0, 1]$  defined as follows.  $f_n(x) = \sqrt{n}$  on  $(0, 1/n)$  and  $f_n(x) = 0$  for  $x$  not on this interval. Show that  $\lim_{n \rightarrow \infty} f_n(x) = 0$  for each  $x$  but  $f_n$  fails to converge to 0 in the mean square sense. Now let  $f_n(x) = 1$  for  $x \in \{1, 1/2, 1/2^2, \dots, 1/2^n\}$  but it equals zero at all other points. Show that  $f_n$  converges to 0 in the mean square sense but not at every point.

38. Using Example 33.5.5 and the convergence theorem for Fourier series, explain why

$$1 = \sum_{k=1}^{\infty} \frac{4}{\pi(2k-1)} \sin\left(\frac{(2k-1)\pi\alpha}{2}\right) \text{ for all } \alpha \in (0, 2).$$



## Chapter 34

# Some Partial Differential Equations

### 34.1 Laplacian in Orthogonal Curvilinear Coordinates

Recall the formula for the Laplacian in curvilinear coordinates

$$\Delta \phi(x) = \frac{1}{\sqrt{g(x)}} \frac{\partial}{\partial x^i} \left( g^{ik}(x) \frac{\partial \phi(x)}{\partial x^k} \sqrt{g(x)} \right)$$

where  $g(x)$  was the determinant of the metric tensor. Using this, it was shown earlier that the Laplacian in spherical coordinates can be obtained.

**Example 34.1.1** *Laplacian in spherical coordinates.*

$$\begin{aligned} \Delta f &= \frac{1}{\rho^2 \sin \phi} \left( \frac{\partial}{\partial \rho} \left( \rho^2 \sin \phi \frac{\partial f}{\partial \rho} \right) + \frac{\partial}{\partial \phi} \left( \frac{\rho \sin \phi}{\rho} \frac{\partial f}{\partial \phi} \right) + \frac{\partial}{\partial \theta} \left( \frac{\rho}{\rho \sin \phi} \frac{\partial f}{\partial \theta} \right) \right) \\ &= \frac{1}{\rho^2} \frac{\partial}{\partial \rho} \left( \rho^2 \frac{\partial f}{\partial \rho} \right) + \frac{1}{\rho^2 \sin \phi} \frac{\partial}{\partial \phi} \left( \sin(\phi) \frac{\partial f}{\partial \phi} \right) + \frac{1}{\rho^2 \sin^2 \phi} \frac{\partial^2 f}{\partial \theta^2} \end{aligned}$$

Using the same machinery, one can obtain the Laplacian in cylindrical coordinates.

**Example 34.1.2** *Laplacian in cylindrical coordinates.*

$$x = r \cos \theta$$

$$y = r \sin \theta$$

$$z = z$$

$$\Delta f = \frac{1}{r} \left( \frac{\partial}{\partial r} \left( r \frac{\partial f}{\partial r} \right) + \frac{\partial}{\partial \theta} \left( \frac{1}{r} \frac{\partial f}{\partial \theta} \right) \right) = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2}$$

## 34.2 Heat and Wave Equations

### 34.2.1 Heat Equation

Fourier's law of heat conduction is that the heat flux  $\mathbf{J}$  is proportional to the temperature gradient  $\nabla u$  where here  $u$  is the temperature. Specifically it says that

$$\mathbf{J} = -k\nabla u$$

So what is the "heat flux"? Hopefully, you saw flux integrals in calculus but here is a short review. If you have a surface  $S$  and a field of unit normals on  $S$  denoted as  $\mathbf{n}$ , then the rate at which the heat crosses  $S$  in the direction of  $\mathbf{n}$  is

$$\int_S \mathbf{J} \cdot \mathbf{n} dS$$

where this is an integral over the surface. Now consider a ball  $B$  with boundary  $S$  in a heat conducting material. Then the heat in  $B$  is given by

$$\int_B \rho c u dV$$

where  $\rho$  is the density and  $c$  the specific heat. Then if no heat is being produced by some chemical reaction for example, it follows that the time rate of change of the total heat in  $B$  is equal to the rate at which heat flows into  $B$ . Thus

$$\frac{d}{dt} \left( \int_B \rho c u dV \right) = - \int_S \mathbf{J} \cdot \mathbf{n} dS$$

where  $\mathbf{n}$  is the outer normal from  $B$ . This is why there is a minus sign on the right. You want the rate at which heat enters  $B$ . Then from the divergence theorem,

$$\frac{d}{dt} \left( \int_B \rho c u dV \right) = - \int_B \nabla \cdot \mathbf{J} dV$$

The integral is a sort of a sum, here over the spacial variables and so it makes sense to formally take the time derivative into the integral<sup>1</sup> and write, using the Fourier law of heat conduction

$$\int_B \frac{\partial (\rho c u)}{\partial t} dV = \int_B \nabla \cdot (k \nabla u) dV$$

This must hold for any ball  $B$  and so the only way this could take place is to have

$$\frac{\partial (\rho c u)}{\partial t} = \nabla \cdot (k \nabla u)$$

We now let  $k, c, \rho$  all be constants and obtain

$$\frac{\partial u}{\partial t} = \frac{k}{\rho c} \Delta u$$

Of course these things are typically not constant, especially  $k$  but if we don't assume this, we can't solve the equation.

<sup>1</sup>This is horrible mathematics because it exchanges two limit operations. However, when modeling, one doesn't worry about rigorous math.

In one dimension, this reduces to

$$u_t = \alpha^2 u_{xx}$$

and this is the equation in what follows. There are other issues besides the equation to consider.

You have a rod of length  $L$ . The heat equation for the temperature  $u$  in the rod is

$$u_t = \alpha^2 u_{xx}$$

In addition to this, there are boundary conditions given on  $u$  at the ends of the rod. For example, you could have

$$u(0, t) = u(L, t) = 0$$

and there is also an initial temperature given

$$u(x, 0) = f(x)$$

Then the idea is to find the unknown function  $u(t, x)$ . Here  $t$  is time and  $x$  is the coordinate of a point on the rod. The constant  $\alpha^2$  varies from material to material. It is different for iron than for aluminum for example. Here you have  $x \in [0, L]$  and  $t > 0$ .

This is a rectangular shape and so it is reasonable to look for a nonzero solution to the above partial differential equation and boundary condition in the form

$$u(x, t) = a(t) b(x)$$

Then

$$a'(t) b(x) = \alpha^2 a(t) b''(x)$$

One can separate the variables as follows.

$$\frac{a'(t)}{\alpha^2 a(t)} = \frac{b''(x)}{b(x)} \quad (34.1)$$

Both sides must equal to some constant  $c$  since otherwise they could not be equal. One way to see this is to differentiate both sides with respect to  $t$ . Then

$$\left( \frac{a'(t)}{\alpha^2 a(t)} \right)' = 0 \text{ and so } \frac{a'(t)}{\alpha^2 a(t)} = c,$$

a constant. Consider the side involving  $x$ .

$$b''(x) - cb(x) = 0, \quad b(0) = b(L) = 0$$

Of course you can't have  $b(x) = 0$  since if it were 0, you would have  $u(x, t) = 0$ . Therefore, from Example 33.2.1,  $-c = \frac{n^2 \pi^2}{L^2}$  where  $n$  is a positive integer and

$$b(x) = \sin\left(\frac{n\pi x}{L}\right)$$

Of course there is such a function for each  $n$  a positive integer. Having picked such a positive integer, 34.1 now forces  $a(t)$  to satisfy the equation

$$a'(t) + \frac{n^2 \pi^2 \alpha^2}{L^2} a(t) = 0$$

Therefore,

$$a(t) = a_n e^{-\frac{n^2 \pi^2 \alpha^2}{L^2} t}$$

It follows that for each  $n$ , there exists a solution to the partial differential equation along with the boundary conditions which is of the form

$$u_n(x, t) = a_n e^{-\frac{n^2 \pi^2 \alpha^2}{L^2} t} \sin\left(\frac{n\pi x}{L}\right)$$

Now if you have solutions to the differential equation along with the boundary condition and you add them together, you have another solution to these things. Therefore, it is not unreasonable to hope that this would also be true for an infinite sum of such solutions. Therefore, we look for a solution to the partial differential equation which is of the form

$$u(x, t) = \sum_{n=1}^{\infty} a_n e^{-\frac{n^2 \pi^2 \alpha^2}{L^2} t} \sin\left(\frac{n\pi x}{L}\right)$$

At least formally, such a thing would solve everything but the initial condition. Now you choose  $a_n$  in such a way that when  $t = 0$ ,

$$f(x) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{L}\right)$$

for  $x \in [0, L]$ . This is now a Fourier series problem.

Another point of view is to look for eigenfunctions.  $b$  such that

$$b''(x) + \lambda b(x) = 0, \quad b(0) = b(L) = 0$$

This is because if you had such an eigenfunction, you could replace the  $b''(x)$  with  $-\lambda b(x)$ . From Example 33.2.1 on Page 638,  $\lambda = \frac{n^2 \pi^2}{L^2}$  where  $n$  is a positive integer and

$$b(x) = \sin\left(\frac{n\pi x}{L}\right)$$

Denote by  $b_n$  this eigenfunction. Then look for a solution to the whole problem which is in the form

$$u(x, t) = \sum_{k=1}^{\infty} a_k(t) b_k(x)$$

Then proceeding formally,

$$\sum_{k=1}^{\infty} a'_k(t) b_k(x) = \alpha^2 \sum_{k=1}^{\infty} a_k(t) b''_k(x) = \sum_{k=1}^{\infty} a_k(t) \left(-\frac{k^2 \pi^2}{L^2} \alpha^2\right) b_k(x)$$

It follows that you should have

$$a'_k(t) + \frac{k^2 \pi^2 \alpha^2}{L^2} a_k(t) = 0$$

so this results in

$$u(x, t) = \sum_{k=1}^{\infty} a_k e^{-\frac{k^2 \pi^2 \alpha^2}{L^2} t} \sin\left(\frac{k\pi x}{L}\right),$$

the same as before. Then you just try and find the  $a_k$  to satisfy the initial condition.

Here is a summary of the method. This method is general and will work for all the examples discussed here.



**PROCEDURE 34.2.1** *To find the solution to an equation*

$$u_t = \alpha^2 u_{xx}, \text{ zero boundary conditions, Initial condition}$$

you do the following.

1. First find eigenfunctions, nonzero solutions to

$$y'' + \lambda^2 y = 0, \text{ boundary conditions}$$

There will typically be infinitely many of these  $\{y_n(x)\}_{n=1}^{\infty}$  corresponding to eigenvalues  $\lambda_n$  where  $\lim_{n \rightarrow \infty} \lambda_n = \infty$ .

2. Your solution will then be of the form

$$u(x, t) = \sum_{n=1}^{\infty} b_n(t) y_n(x)$$

3. Choose  $b_n(t)$  to satisfy the equation  $b'_n(t) = -\lambda_n^2 b_n(t)$  in order that the terms of the sum satisfy the partial differential equation. Thus

$$b_n(t) = b_n \exp(-t\lambda_n^2)$$

Then the solution to the problem is

$$u(x, t) = \sum_{n=1}^{\infty} b_n \exp(-t\lambda_n^2) y_n(x)$$

where  $b_n$  is chosen such that  $\sum_{n=1}^{\infty} b_n y_n(x)$  is the Fourier series expansion for the initial condition.

**Example 34.2.2** *Find the solution to the initial boundary value problem*

$$\begin{aligned} u_t &= \alpha^2 u_{xx}, \quad u(0, t) = u(2, t) = 0 \\ u(x, 0) &= 1 - (1 - x)^2 \end{aligned}$$

where

$$f(x) = \begin{cases} x & \text{if } x \in [0, 1] \\ 1 - x & \text{if } x \in [1, 2] \end{cases}$$

From the above discussion,

$$u(x, t) = \sum_{k=1}^{\infty} a_k e^{-\frac{k^2 \pi^2}{2^2} t} \sin\left(\frac{k\pi x}{2}\right)$$

the eigenfunctions being  $\sin\left(\frac{k\pi x}{2}\right)$ , and to satisfy the initial condition, you need

$$a_k = \frac{2}{2} \int_0^2 (1 - (1 - x)^2) \sin\left(\frac{k\pi x}{2}\right) dx$$

After some tedious computations, this yields

$$a_k = \frac{16}{\pi^3 k^3} \left(1 - (-1)^k\right)$$

Thus when  $k$  is even, this is 0 and when  $k$  is odd, it equals  $\frac{32}{\pi^3 k^3}$ . Thus

$$u(x, t) = \sum_{k=1}^{\infty} \frac{32}{\pi^3 (2k-1)^3} e^{-\frac{(2k-1)^2 \pi^2}{4} t} \sin\left(\frac{(2k-1) \pi x}{2}\right)$$

The next example has to do with the same equation but with one end insulated and the other held at a temperature of 0. The physical modeling of this equation shows that to consider an insulated boundary, say at  $L$ , you let  $u_x(L, t) = 0$ .

**Example 34.2.3** *Solve the problem*

$$\begin{aligned} u_t &= u_{xx}, \quad u(0, t) = u_x(2, t) = 0 \\ u(x, 0) &= 1 - (1-x)^2 \end{aligned}$$

To do this, first look for eigenfunctions. Find solutions to

$$y'' + \lambda y = 0, \quad y(0) = 0, y'(2) = 0$$

Then the eigenfunctions are in Example 33.2.1. They are

$$\sin\left(\frac{(2n-1) \pi x}{4}\right), \quad n = 1, 2, \dots$$

It follows that the solution desired is of the form

$$\sum_{n=1}^{\infty} b_n(t) \sin\left(\frac{(2n-1) \pi x}{4}\right)$$

and one needs

$$b'_n(t) = -\frac{1}{10} \left(\frac{(2n-1) \pi}{4}\right)^2 b_n(t)$$

so

$$b_n(t) = b_n \exp\left(-\frac{1}{10} \left(\frac{(2n-1) \pi}{4}\right)^2 t\right)$$

Then the Fourier series expansion of the solution is

$$\sum_{n=1}^{\infty} b_n \exp\left(-\frac{1}{10} \left(\frac{(2n-1) \pi}{4}\right)^2 t\right) \sin\left(\frac{(2n-1) \pi x}{4}\right)$$

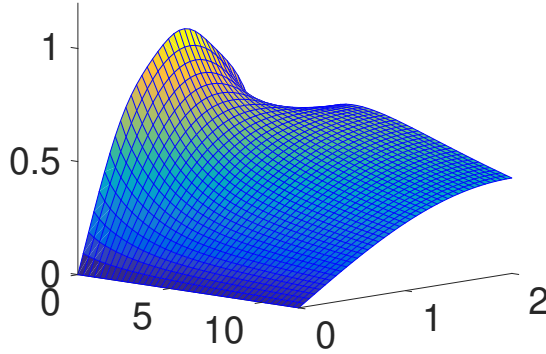
where  $b_n$  is an appropriate Fourier coefficient chosen to satisfy the initial condition. Thus

$$\begin{aligned} b_n &= \frac{2}{2} \int_0^2 \sin\left(\frac{(2n-1) \pi x}{4}\right) (1 - (1-x)^2) dx \\ &= \frac{32}{\pi^3 (2n-1)^3} (2(-1)^n \pi n - (-1)^n \pi + 4) \end{aligned}$$

Then the solution to this problem is

$$\sum_{n=1}^{\infty} \left( \frac{32}{\pi^3 (2n-1)^3} (2(-1)^n \pi n - (-1)^n \pi + 4) \right) e^{-\frac{1}{10} \left( \frac{(2n-1)\pi}{4} \right)^2 t} \sin \left( \frac{(2n-1)\pi x}{4} \right)$$

A graph of this function of two variables in which the sum is taken up to 8 for  $(t, x) \in [0, 12] \times [0, 2]$  is:

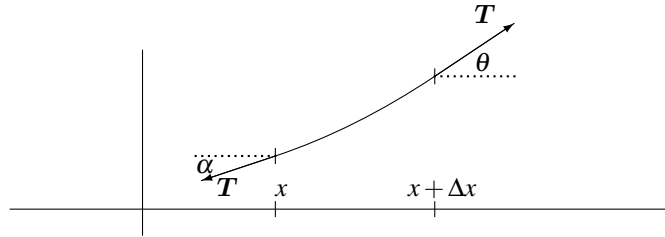


### 34.2.2 The Wave Equation

The next example is of a different sort of equation, the wave equation. This equation is of the form

$$u_{tt} = c^2 u_{xx}$$

It models the transverse displacements of a vibrating string. Here is a picture to discuss why this is an appropriate equation. It is important to note that it is a string, not a beam. This means that it cannot support itself in the sense that there is no internal stiffness. It is also very important to note that the transverse displacements are assumed to be very small. Thus the picture drawn below is blown up in the vertical direction.



Let  $\rho$  be the length density of this string which is assumed constant. This means that the mass of the segment of string shown is just  $\rho$  (length of the segment of string). Since the transverse displacements are very small, this is essentially  $\rho \Delta x$ . The force acting on the segment of string shown is  $T \sin \theta - T \sin \alpha$ , where  $T$  is the magnitude of the vector  $T$ . We assume also that the magnitude of the tension in the string is also a constant due to the assumption that the displacements are small. Let  $u(t, x)$  denote the vertical displacement from horizontal. For  $\Delta x$  small enough, the acceleration  $u_{tt}(t, x)$  should be essentially constant on the interval  $[x, x + \Delta x]$ . Then by Newton's second law,

$$\rho \Delta x u_{tt}(t, x) = T (\sin \theta - \sin \alpha)$$

Since the displacement is very small, we can assume that there is really no difference in replacing  $\sin \theta$ ,  $\sin \alpha$  with  $\tan \theta$ ,  $\tan \alpha$  respectively. But  $\tan \theta$  is just the slope of the tangent line at  $(t, x + \Delta x)$ . Thus

$$\rho \Delta x u_{tt}(t, x) = T(u_x(t, x + \Delta x) - u_x(t, x))$$

Divide by  $\Delta x$  and let  $\Delta x \rightarrow 0$  to obtain

$$\rho u_{tt} = T u_{xx}, \quad u_{tt} = \frac{T}{\rho} u_{xx}.$$

This is the wave equation for a vibrating string.

Since it is second order in  $t$  you need two initial conditions, one on the velocity and the other on the displacement in order to get a unique solution. However, other than this, the procedure is essentially the same.

**Example 34.2.4** Find the solution to the initial boundary value problem

$$\begin{aligned} u_{tt} &= \alpha^2 u_{xx}, \quad u(0, t) = u(2, t) = 0, \\ u(x, 0) &= 1 - (1 - x)^2 \\ u_t(x, 0) &= 0 \end{aligned}$$

The eigenfunctions are solutions to

$$y''(x) + \lambda y(x) = 0, \quad y(0) = 0 = y(2)$$

This is discussed in Example 33.2.1. The eigenfunctions are

$$\sin\left(\frac{n\pi x}{2}\right)$$

and the eigenvalues are  $\lambda = \frac{n^2\pi^2}{4}$ .

Then you look for a solution to the equation with boundary conditions of the form

$$a(t) \sin\left(\frac{n\pi x}{2}\right)$$

Thus you need

$$a''(t) \sin\left(\frac{n\pi x}{2}\right) = -\alpha^2 \frac{n^2\pi^2}{4} a(t) \sin\left(\frac{n\pi x}{2}\right)$$

Hence

$$a'' + \alpha^2 \frac{n^2\pi^2}{4} a = 0$$

and so, since you know the general solution to this equation, it is

$$a(t) = a_n \cos\left(\alpha \frac{n\pi}{2} t\right) + b_n \sin\left(\alpha \frac{n\pi}{2} t\right)$$

It follows that the solution to the full problem will be of the form

$$u(x, t) = \sum_{n=1}^{\infty} \left( a_n \cos\left(\alpha \frac{n\pi}{2} t\right) + b_n \sin\left(\alpha \frac{n\pi}{2} t\right) \right) \sin\left(\frac{n\pi x}{2}\right)$$

Now you need to find  $a_n$  and  $b_n$  to get the initial conditions. Letting  $t = 0$ , you need to have

$$1 - (1 - x)^2 = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{2}\right)$$

and this is something which was done earlier. You need

$$a_n = \frac{16}{\pi^3 n^3} (1 - (-1)^n)$$

Next, what about  $b_n$ ? Differentiate both sides. Thus

$$u_t(x, t) = \sum_{n=1}^{\infty} \left( a_n \left( -\alpha \frac{n\pi}{2} \right) \sin\left(\alpha \frac{n\pi}{2} t\right) + b_n \left( \alpha \frac{n\pi}{2} \right) \cos\left(\alpha \frac{n\pi}{2} t\right) \right) \sin\left(\frac{n\pi x}{2}\right)$$

Of course this operation is complete garbage because it involves the interchange of limit operations without any justification. However, we do it anyway. In fact it is all right. You can do the formal manipulations and then you can rigorously verify that what you end up with really is a solution to the problem in some sense. Now plug in  $t = 0$ . Then you need

$$0 = \sum_{n=0}^{\infty} b_n \left( \alpha \frac{n\pi}{2} \right) \sin\left(\frac{n\pi x}{2}\right)$$

Clearly you should take  $b_n = 0$ . Therefore, the desired solution is

$$u(x, t) = \sum_{n=1}^{\infty} \left( \frac{32}{\pi^3 (2n-1)^3} \cos\left(\alpha \frac{(2n-1)\pi}{2} t\right) \right) \sin\left(\frac{n\pi x}{2}\right)$$

Let's let  $\alpha^2 = .09$ . Then the specific solution is

$$u(x, t) = \sum_{n=1}^{\infty} \left( \frac{32}{\pi^3 (2n-1)^3} \cos\left(.3 \frac{(2n-1)\pi}{2} t\right) \right) \sin\left(\frac{n\pi x}{2}\right)$$

Note that from calculus, the series makes perfect sense because in fact, it converges absolutely.

**Example 34.2.5** Solve the initial boundary value problem

$$\begin{aligned} u_{tt} &= \alpha^2 u_{xx}, \quad u(0, t) = u(4, t) = 0, \\ u(x, 0) &= f(x) \\ u_t(x, 0) &= 0 \end{aligned}$$

where

$$f(x) = \begin{cases} 1 - (x-2)^2 & \text{on } [1, 3] \\ 0 & \text{on the rest of } [0, 4] \end{cases}$$

By similar reasoning to the above example,

$$u(x, t) = \sum_{n=1}^{\infty} \left( a_n \cos\left(\alpha \frac{n\pi}{4} t\right) + b_n \sin\left(\alpha \frac{n\pi}{4} t\right) \right) \sin\left(\frac{n\pi x}{4}\right)$$

Then, as above,  $b_n = 0$  and  $a_n$  must be chosen such that

$$f(x) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{4}\right)$$

Thus

$$a_n = \frac{2}{4} \int_1^3 (1 - (x-2)^2) \sin\left(\frac{n\pi x}{4}\right) dx$$

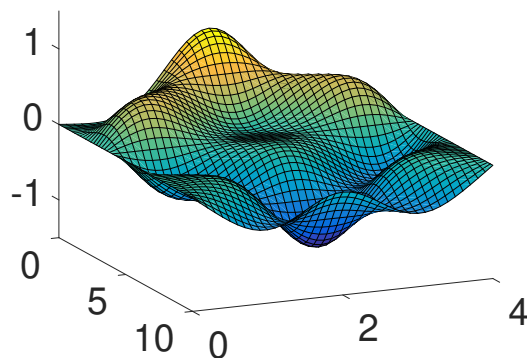
Then after doing the hard work, you end up with

$$a_n = -16 \frac{4 \cos \frac{3}{4}n\pi + n\pi \sin \frac{3}{4}n\pi - 4 \cos \frac{1}{4}n\pi + n\pi \sin \frac{1}{4}n\pi}{n^3 \pi^3}$$

Then the solution is

$$u(x, t) = \sum_{n=1}^{\infty} \left( -16 \frac{4 \cos \frac{3}{4}n\pi + n\pi \sin \frac{3}{4}n\pi - 4 \cos \frac{1}{4}n\pi + n\pi \sin \frac{1}{4}n\pi}{n^3 \pi^3} \right) \cos\left(\alpha \frac{n\pi}{4} t\right) \sin\left(\frac{n\pi x}{4}\right)$$

Let  $\alpha = .5$  to give a specific example. Here is a graph of the function of two variables in which the sum is taken up to  $n = 6$ . The  $t$  axis goes from 0 to 10 and if you fix  $t$  and imagine a cross section, it will be  $x \rightarrow u(x, t)$ .



### 34.3 Nonhomogeneous Problems

For the sake of completeness, here is a brief discussion of what can be done if you have a nonhomogeneous equation of the form  $u_t = au_{xx} + f$  along with an initial condition

$$u(x, 0) = g(x)$$

and boundary conditions. As before, there are eigenfunctions  $y_n$  satisfying the boundary conditions and

$$y_n'' = -\lambda_n^2 y_n, \quad \lim_{n \rightarrow \infty} \lambda_n = \infty$$

such that also

$$\int_0^L y_n(x) y_m(x) dx = \delta_{nm} = \begin{cases} 1 & \text{if } n = m \\ 0 & \text{if } n \neq m \end{cases}$$

and it is assumed that one can obtain a valid Fourier series expansion in terms of these eigenfunctions of all the functions of interest. Note how, for the sake of simplicity, it is assumed that

$$\int_0^L y_n^2(x) dx = 1$$

You multiply by an appropriate constant to make it this way. Thus, if the eigenfunctions are multiples of  $\sin\left(\frac{n\pi}{L}x\right)$ , you choose the multiple to satisfy the above equation. Let

$$f(x, t) = \sum_{n=0}^{\infty} f_n(t) y_n(x)$$

Thus it is desired to have

$$\sum_{n=0}^{\infty} b'_n(t) y_n(x) = -a \sum_{n=0}^{\infty} \lambda_n b_n(t) y_n(x) + \sum_{n=0}^{\infty} f_n(t) y_n(x)$$

and this is achieved if

$$b'_n(t) = -a\lambda_n b_n(t) + f_n(t)$$

which is a familiar equation, the solution being

$$b_n(t) = e^{-a\lambda_n t} b_n(0) + \int_0^t e^{-a\lambda_n(t-s)} f_n(s) ds$$

Then the solution is

$$u(x, t) = \sum_{n=0}^{\infty} \left( e^{-a\lambda_n t} b_n(0) + \int_0^t e^{-a\lambda_n(t-s)} f_n(s) ds \right) y_n(x)$$

where  $b_n(0)$  needs to be chosen to satisfy the initial condition. Thus it is required that

$$b_n(0) = \int_0^L g(u) y_n(u) du$$

In what was done earlier,  $y_n$  was typically something like  $(2/L)^{1/2} \sin\left(\frac{n\pi x}{L}\right)$ . Then the solution is

$$u(x, t) = \sum_{n=0}^{\infty} \left( e^{-a\lambda_n t} \left( \int_0^L g(u) y_n(u) du \right) + \int_0^t e^{-a\lambda_n(t-s)} f_n(s) ds \right) y_n(x)$$

**Example 34.3.1** Find the solution to

$$\begin{aligned} u_t(x, t) &= u_{xx}(x, t) + f(x, t) \\ u(0, t) &= 0 = u(2, t) \\ u(x, 0) &= x \end{aligned}$$

where  $f(x, t) = xt$ .

First find the eigenfunctions and eigenvalues for the equation

$$y'' + \lambda y = 0, y(0) = 0 = y(2)$$

You must have  $\lambda$  strictly positive. The eigenvalues are  $\lambda = \left(\frac{n\pi}{2}\right)^2$ , and the eigenfunctions are  $\sin\left(\frac{n\pi x}{2}\right)$ . Also, there is a Fourier series expansion for  $f(x, t)$  as follows.

$$f(x, t) = \sum_{n=1}^{\infty} f_n(t) \sin\left(\frac{n\pi x}{2}\right)$$

where

$$f_n(t) = \frac{2}{L} \int_0^L f(x, t) \sin\left(\frac{n\pi x}{2}\right) dx$$

Thus

$$f_n(t) = \int_0^L f(x, t) \sin\left(\frac{n\pi x}{2}\right) dx = \frac{1}{\pi^2 n^2} \left(4\pi n t (-1)^{n+1}\right)$$

Now the solution is

$$u(x, t) = \sum_{n=0}^{\infty} \left( \int_0^t e^{-\left(\frac{n\pi}{2}\right)^2(t-s)} \left( \int_0^L u \sin\left(\frac{n\pi u}{2}\right) du \right) + \frac{1}{\pi^2 n^2} \left(4\pi n s (-1)^{n+1}\right) \right) ds \sin\left(\frac{n\pi x}{2}\right)$$

Once you know how to solve this kind of problem, it becomes routine, if long, to find solutions to problems like this.

**Example 34.3.2** Find the solution to the initial-boundary value problem

$$\begin{aligned} u_t(x, t) &= u_{xx}(x, t) + f(x, t) \\ u(0, t) &= 0, u(L, t) = g(t) \\ u(x, 0) &= h(x) \end{aligned}$$

In this case, you massage the problem to get one which is like one you do know how to do which involves zero boundary conditions. Let

$$w(x, t) = u(x, t) - \frac{x}{L} g(t)$$

then

$$\begin{aligned} w_t &= u_t - \frac{x}{L} g'(t) = u_{xx} + f - \frac{x}{L} g'(t) = w_{xx} + f(x, t) - \frac{x}{L} g'(t) \\ w(0, t) &= u(0, t) = 0, \quad w(L, t) = u(L, t) - g(t) = 0 \\ w(x, 0) &= u(x, 0) - \frac{x}{L} g(0) = h(x) - \frac{x}{L} g(0) \end{aligned}$$

and now you solve for  $w$  using the above procedure. There are seemingly endless variations of this but all amount to the following.

**PROCEDURE 34.3.3** To solve

$$u_t = Au + f$$

nonzero boundary conditions

initial condition  $u(x, 0) = l(x)$



You let  $w = u - k(x, t)$  where  $k$  is a known function chosen such that the boundary conditions on  $w$  involve  $w$  or its partial  $x$  derivatives set equal to 0. Then adjust to consider the equation solved for  $w$  which is of the form

$$w_t = Aw + \hat{f}$$

zero boundary conditions

$$\text{modified initial condition } w(x, 0) = \hat{l}(x)$$

This is then of the right form which can be solved. Obtain eigenfunctions  $\{y_n\}$

$$Ay_n = -\lambda_n^2 y_n$$

Find the eigenfunction expansion for  $f$  in terms of these.

$$\sum_{n=0}^{\infty} f_n(t) y_n(x)$$

Then you need

$$w(x, t) = \sum_{n=0}^{\infty} b_n(t) y_n(x)$$

where

$$b'_n(t) = -\lambda_n^2 b_n(t) + f_n(t)$$

and  $b_n(0)$  is an appropriate Fourier coefficient chosen to satisfy the initial condition. Find  $w$  and then  $u(x, t) = w(x, t) + k(x, t)$ .

In case the problem is second order in time, you do something similar except that the differential equation for  $b_n$  will now be second order in time and you will need to adjust both  $b_n(0)$  and  $b'_n(0)$  to achieve appropriate initial conditions.

**Example 34.3.4** Solve the following

$$u_{tt} = u_{xx}, \quad u(x, 0) = 0, u_t(x, 0) = 0$$

$$u(0, t) = 0, \quad u(L, t) = \sin t$$

Initially the string is at rest and then something starts moving the right side up and down. What happens?

Following the procedure, let

$$w(x, t) = u(x, t) - \frac{x}{L} \sin(t)$$

this works because  $w$  has zero boundary conditions. Then

$$w_{tt} = u_{tt} + \frac{x}{L} \sin t = u_{xx} + \frac{x}{L} \sin t = w_{xx} + \frac{x}{L} \sin t$$

$$w(0, t) = w(L, t) = 0$$

$$w(x, 0) = 0, \quad w_t(x, 0) = -\frac{x}{L} \cos(t)$$

The eigenfunctions are  $\sin\left(\frac{n\pi}{L}x\right)$ . Then the expansion for  $(x/L)\sin(t)$  is

$$\begin{aligned} & \sum_{n=1}^{\infty} \left( \frac{2}{L} \int_0^L \left( \frac{x}{L} \sin t \right) \sin\left(\frac{n\pi}{L}x\right) dx \right) \sin\left(\frac{n\pi}{L}x\right) \\ &= \sum_{n=1}^{\infty} \left( 2 \frac{(-1)^{n+1}}{\pi n} \sin(t) \right) \sin\left(\frac{n\pi}{L}x\right) \end{aligned}$$

then the solution is

$$w(x, t) = \sum_{n=1}^{\infty} b_n(t) \sin\left(\frac{n\pi}{L}x\right)$$

where

$$b_n''(t) = -\frac{n^2\pi^2}{L^2} b_n(t) + 2 \frac{(-1)^{n+1}}{\pi n} \sin(t) \quad (*)$$

The Fourier series expansion for  $w_t(x, t) = -\frac{x}{L} \cos(t)$  in terms of these eigenfunctions is

$$\begin{aligned} & \sum_{n=1}^{\infty} \left( \frac{2}{L} \int_0^L \left( -\frac{x}{L} \cos(t) \right) \sin\left(\frac{n\pi x}{L}\right) dx \right) \sin\left(\frac{n\pi x}{L}\right) \\ &= \sum_{n=1}^{\infty} \left( 2 \frac{(-1)^n}{\pi n} \cos t \right) \sin\left(\frac{n\pi x}{L}\right) \end{aligned}$$

Now the solution to  $*$  is  $b_n(t) =$

$$\left( \cos\left(\frac{\pi}{L}nt\right) \right) b_n(0) + \frac{1}{\pi n} \left( \sin\left(\frac{\pi}{L}nt\right) \right) b_n'(0) + 2(-1)^{n+1} L^2 \frac{\sin t}{\pi^3 n^3 - \pi L^2 n}$$

Clearly the initial condition for  $w$  gives  $b_n(0) = 0$ . It remains to find  $b_n'(0)$ . The solution is

$$w(x, t) = \sum_{n=1}^{\infty} \left( \frac{1}{\pi n} \left( \sin\left(\frac{\pi}{L}nt\right) \right) b_n'(0) + 2(-1)^{n+1} L^2 \frac{\sin t}{\pi^3 n^3 - \pi L^2 n} \right) \sin\left(\frac{n\pi}{L}x\right)$$

Now

$$w_t(x, t) = \sum_{n=1}^{\infty} \left( \left( \cos\frac{\pi}{L}nt \right) b_n'(0) + 2(-1)^{n+1} L^2 \left( \frac{\cos t}{\pi^3 n^3 - \pi L^2 n} \right) \right) \sin\left(\frac{n\pi}{L}x\right)$$

and so the initial condition for  $w_t$  requires

$$\sum_{n=1}^{\infty} \left( b_n'(0) + \frac{2(-1)^{n+1} L^2}{\pi^3 n^3 - \pi L^2 n} \right) \sin\left(\frac{n\pi}{L}x\right) = \sum_{n=1}^{\infty} \left( 2 \frac{(-1)^n}{\pi n} \right) \sin\left(\frac{n\pi x}{L}\right)$$

and so

$$b_n'(0) = 2 \frac{(-1)^n}{\pi n} - \frac{2(-1)^{n+1} L^2}{\pi^3 n^3 - \pi L^2 n} = 2(-1)^n \pi \frac{n}{\pi^2 n^2 - L^2}$$

Thus

$$w(x, t) = \sum_{n=1}^{\infty} \left( \begin{aligned} & 2(-1)^n L \frac{\sin \frac{\pi}{L}nt}{\pi^2 n^2 - L^2} \\ & + 2(-1)^{n+1} L^2 \frac{\sin t}{\pi^3 n^3 - \pi L^2 n} \end{aligned} \right) \sin\left(\frac{n\pi}{L}x\right)$$

Therefore,  $u(x, t) = w(x, t) + \frac{x}{L} \sin(t)$ .

## 34.4 Laplace Equation

The Laplace equation is  $\Delta u = 0$ . In two dimensions and in rectangular coordinates,

$$\Delta u = u_{xx} + u_{yy} = 0$$

Here  $u$  is a function of the two variables  $x, y$ . Note first that  $\Delta$  is a linear operator. That is, for  $a, b$  scalars and  $u, v$  functions,

$$\Delta(au + bv) = a\Delta u + b\Delta v$$

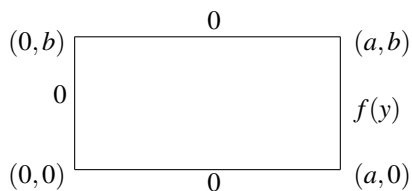
Because of this, if you have several solutions to the Laplace equation  $u_1, \dots, u_m$ , and if you have scalars  $c_i$ , then

$$\Delta\left(\sum_{i=1}^n c_i u_i\right) = \sum_{i=1}^n c_i \Delta u_i = \sum_{i=1}^n c_i 0 = 0.$$

It is understood that the point  $(x, y)$  is contained in some region in the plane. One looks for a solution to the equation which also satisfies boundary conditions on the boundary of the region. When these conditions involve given values for the function  $u$  it is called the Dirichlet problem. When it involves giving values for the normal derivative of  $u$  defined by  $\nabla u \cdot \mathbf{n}$  for  $\mathbf{n}$  the unit outer normal, it is called a Neuman problem. In this short introduction this region will be either a circular disk or a rectangle. These are called boundary value problems.

### 34.4.1 Rectangles

First consider the rectangle. Here is a typical problem. The boundary conditions are as shown in the picture, zero on the top bottom and left side and  $f(y)$  on the right.



You can solve this the usual way. Look for eigenfunctions. These need to correspond to the two opposite sides where the boundary condition is 0. Thus the eigenfunctions are the nonzero solutions to

$$f''(y) + \lambda f(y) = 0, \quad f(0) = 0 = f(b)$$

It follows the eigenfunctions are

$$\sin\left(\frac{n\pi}{b}y\right), \quad n = 1, 2, \dots$$

and the eigenvalues are  $\frac{\pi^2}{b^2}n^2, n = 1, 2, \dots$ . Next you need to find some  $g(x)$  such that  $g(x)\sin\left(\frac{n\pi}{b}y\right)$  solves the boundary conditions and the equation. The boundary conditions are automatic. Now consider the equation. You need

$$g''(x)\sin\left(\frac{n\pi}{b}y\right) + g(x)\left(-\frac{\pi^2}{b^2}n^2\sin\left(\frac{\pi}{b}ny\right)\right) = 0$$

Thus, you need

$$g''(x) - \frac{\pi^2}{b^2} n^2 g(x) = 0$$

You know the solution is

$$C_1 e^{\frac{n\pi}{b}x} + C_2 e^{-\frac{n\pi}{b}x}$$

Now it turns out that in this application, it is much more convenient to write the general solution as

$$a_n \cosh\left(\frac{n\pi}{b}x\right) + b_n \sinh\left(\frac{n\pi}{b}x\right)$$

This gives the same general solution. The above functions are linear combinations of the known solutions and so things in the above form are solutions. Furthermore, the ratio of the two solutions is not constant so their Wronskian does not vanish. Hence it is the general solution. Now you try and get the solution to the boundary value problem in the form

$$u(x, y) = \sum_{n=1}^{\infty} \left( a_n \cosh\left(\frac{n\pi}{b}x\right) + b_n \sinh\left(\frac{n\pi}{b}x\right) \right) \sin\left(\frac{n\pi}{b}y\right)$$

when  $x = 0$ , you get  $\sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi}{b}y\right) = 0$  and so each  $a_n = 0$ . When  $x = a$ , you need

$$f(y) = \sum_{n=1}^{\infty} b_n \sinh\left(\frac{n\pi}{b}a\right) \sin\left(\frac{n\pi}{b}y\right)$$

Hence you need

$$b_n \sinh\left(\frac{n\pi}{b}a\right) = \frac{2}{b} \int_0^b f(t) \sin\left(\frac{n\pi}{b}t\right) dt b_n = \frac{2}{b \sinh\left(\frac{n\pi}{b}a\right)} \int_0^b f(t) \sin\left(\frac{n\pi}{b}t\right) dt$$

Therefore, with this formula for  $b_n$

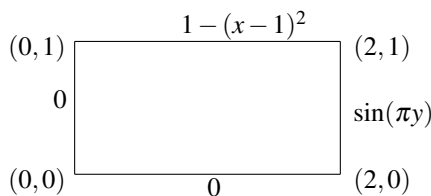
$$u(x, y) = \sum_{n=1}^{\infty} b_n \sinh\left(\frac{n\pi}{b}x\right) \sin\left(\frac{n\pi}{b}y\right)$$

This shows how to solve a more general problem in which you have functions given on the edges. You solve the problem for the situation in which there is something nonzero on exactly one edge with 0 on the others and then you add these solutions together.

**Example 34.4.1** Find the solution to the boundary value problem

$$u_{xx} + u_{yy} = 0$$

where the boundary conditions and rectangle are as expressed in the following picture.



First find the solution which has  $\sin y$  on the right and zero on the other edges. This was done in the above. It is

$$u_1(x, y) = \sum_{n=1}^{\infty} b_n \sinh(n\pi x) \sin(n\pi y)$$

where  $b_n$  is as given above. Thus

$$b_n = \frac{2}{\sinh(n\pi a)} \int_0^1 \sin(t) \sin(n\pi t) dt = \frac{2}{\sinh n\pi a} \frac{n\pi \sin 1 (-1)^n}{n^2 \pi^2 - 1}$$

Hence this partial solution is

$$u_1(x, y) = \sum_{n=1}^{\infty} \frac{2}{\sinh n\pi a} \frac{n\pi \sin 1 (-1)^n}{n^2 \pi^2 - 1} \sinh(n\pi x) \sin(n\pi y)$$

Next find the solution to the equation which has  $1 - (x - 1)^2$  on the top and zero on the other sides. This is just like what was done earlier except that you would switch  $a$  and  $b$ . You find the eigenfunctions for the two opposite zero boundary conditions. These are

$$\sin\left(\frac{n\pi}{2}x\right), n = 1, 2, \dots$$

with eigenvalues  $\frac{n^2 \pi^2}{4}$ . Next you look for solutions to the equation which involve

$$a(y) \sin\left(\frac{n\pi}{2}x\right)$$

Thus

$$a''(y) \sin\left(\frac{n\pi}{2}x\right) + a(y) \left(-\frac{\pi^2}{4} n^2 \sin \frac{\pi}{2} nx\right) = 0$$

Hence

$$a''(y) - \frac{\pi^2}{4} n^2 a(y) = 0$$

and so

$$a(y) = a_n \cosh\left(\frac{n\pi}{2}y\right) + b_n \sinh\left(\frac{n\pi}{2}y\right)$$

Then the general solution is

$$u_2(x, y) = \sum_{n=1}^{\infty} \left( a_n \cosh\left(\frac{n\pi}{2}y\right) + b_n \sinh\left(\frac{n\pi}{2}y\right) \right) \sin\left(\frac{n\pi}{2}x\right)$$

When  $y = 0$ , you are supposed to get 0 for the boundary condition. Hence  $a_n = 0$ . When  $y = b$  you need

$$1 - (1 - x)^2 = \sum_{n=1}^{\infty} \left( b_n \sinh\left(\frac{n\pi}{2}\right) \right) \sin\left(\frac{n\pi}{2}x\right)$$

Therefore, you need

$$b_n \sinh\left(\frac{n\pi}{2}\right) = \int_0^2 \left( 1 - (1 - s)^2 \right) \sin\left(\frac{n\pi}{2}s\right) ds = 8 \frac{2 - 2(-1)^n}{n^3 \pi^3}$$

Then this solution is of the form

$$\begin{aligned} u_2(x, y) &= \sum_{n=1}^{\infty} \left( 8 \frac{2 - 2(-1)^n}{n^3 \pi^3} \frac{1}{\sinh\left(\frac{n\pi}{2}\right)} \right) \sinh\left(\frac{n\pi}{2}y\right) \sin\left(\frac{n\pi}{2}x\right) \\ &= \sum_{n=1}^{\infty} \frac{32}{(2n-1)^3 \pi^3} \frac{1}{\sinh\left(\frac{(2n-1)\pi}{2}\right)} \sinh\left(\frac{(2n-1)\pi y}{2}\right) \sin\left(\frac{(2n-1)\pi}{2}x\right) \end{aligned}$$

Therefore, the solution to the boundary value problem is the sum of these two solutions.

$$\begin{aligned} u(x, y) &= \sum_{n=1}^{\infty} \frac{2}{\sinh n\pi a} \frac{n\pi \sin 1 (-1)^n}{n^2 \pi^2 - 1} \sinh(n\pi x) \sin(n\pi y) + \\ &\quad \sum_{n=1}^{\infty} \frac{32}{(2n-1)^3 \pi^3} \frac{1}{\sinh\left(\frac{(2n-1)\pi}{2}\right)} \sinh\left(\frac{(2n-1)\pi y}{2}\right) \sin\left(\frac{(2n-1)\pi}{2}x\right) \end{aligned}$$

You can probably see how to consider given functions in place of 0 on the remaining two sides.

### 34.4.2 Circular Disks

This is more interesting than the above because it is more often the case that you encounter it in real situations. Most pipes are circular for example. The Laplacian in rectangular coordinates is

$$\Delta = u_{xx} + u_{yy}$$

However, rectangular coordinates are not natural for considering circles. For example, the boundaries of a rectangle are obtained by letting one of the variables be constant. If you want something like this to happen for a circular shape, you should consider polar coordinates. For example, the boundary of a circular disk is obtained by letting  $r = c$  a constant. Recall the relation between polar and rectangular coordinates.  $\theta \in [0, 2\pi), r > 0$ ,

$$\begin{aligned} x &= r \cos \theta \\ y &= r \sin \theta \end{aligned} \tag{34.2}$$

You have a scalar field  $u$  and it is a function of a point in two dimensional space. This point can be described in terms of either polar coordinates or rectangular coordinates. Thus

$$u(x, y) = u(r, \theta)$$

there  $(x, y)$  and  $(r, \theta)$  pertain to the same point in two dimensions. As discussed above in Section 34.1, the Laplacian in polar coordinates is

$$u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2}$$

**Example 34.4.2** Find the solution to

$$\Delta u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0$$

on the disc of radius  $R$  if on the boundary of this disk,

$$u(R, \theta) = f(\theta)$$

where  $f(0) = f(2\pi)$ . This last condition is necessary because  $\theta = 0$  and  $\theta = 2\pi$  correspond to the same point on the boundary of this disk. Note how everything is in terms of the variables  $r, \theta$  and that in terms of these variables, the circular disk is actually a rectangle.

Use the method of separation of variables. Look for a solution to the equation which is of the form  $R(r)\Theta(\theta)$ .

$$r^2 R''(r)\Theta(\theta) + rR'(r)\Theta(\theta) + R(r)\Theta''(\theta) = 0$$

So divide by  $R\Theta$ . This leads to

$$r^2 \frac{R''}{R} + r \frac{R'}{R} + \frac{\Theta''}{\Theta} = 0$$

Hence

$$\frac{\Theta''}{\Theta} = -\lambda = r^2 \frac{R''}{R} + r \frac{R'}{R}$$

for some constant  $\lambda$ . First consider  $\Theta$ . You must have  $\Theta(0) = \Theta(2\pi)$ . Also

$$\Theta'' + \lambda\Theta = 0$$

Multiply both sides by  $\Theta$  and integrate. This leads to

$$\Theta'(\theta)\Theta(\theta)|_0^{2\pi} - \int_0^{2\pi} (\Theta')^2 d\theta + \lambda \int_0^{2\pi} \Theta^2 d\theta = 0$$

The boundary terms disappear because you must also have  $\Theta'(2\pi) = \Theta'(0)$ . Therefore, to have a solution, it is necessary that  $\lambda \geq 0$ . If  $\lambda = 0$ , you need to have  $\Theta' = 0$  and so  $\Theta(\theta) = C$  a constant. Otherwise, you need  $\lambda = \mu^2, \mu > 0$ . Then the solution to the equation is

$$C_1 \cos \mu\theta + C_2 \sin \mu\theta$$

and you need to have  $\Theta(0) = \Theta(2\pi)$ . Therefore, it is required that  $\mu 2\pi$  is an integer multiple of  $2\pi$  so  $\mu = n$  for  $n$  an integer. Thus the eigenvalues are the nonnegative integers and you get

$$\Theta_n(\theta) = (a_n \cos(n\theta) + b_n \sin(n\theta)), \quad n = 0, 1, 2, \dots$$

It follows that for each of these  $n$ ,

$$r^2 R_n'' + rR_n' - n^2 R_n = 0$$

This is an Euler equation and you look for solutions in the form  $R(r) = r^\alpha$ . Then to find  $\alpha$ , you insert this into the equation.

$$r^2 \alpha(\alpha-1)r^{\alpha-2} + r\alpha r^{\alpha-1} - n^2 r^\alpha = 0$$

and so you get the indicial equation

$$\alpha(\alpha-1) + \alpha - n^2 = (\alpha-n)(\alpha+n) = 0$$

Therefore, the solutions are of the form

$$c_n r^n + d_n r^{-n}$$

We can immediately conclude that  $d_n = 0$  because it makes no sense to have the solution to the differential equation be unbounded as  $r \rightarrow 0$ . Recall the theorem from calculus that on a closed and bounded set, a continuous function achieves its maximum and minimum. If  $u$  is going to be continuous, which we certainly expect it to be, then this cannot be harmonized with  $d_n \neq 0$ . Thus this has found many solutions to the partial differential equation which are of the form

$$r^n (a_n \cos(n\theta) + b_n \sin(n\theta))$$

The solution to the equation will then be an infinite sum of the functions of the above form. Thus combining the  $c_n$  with  $a_n$  and  $b_n$ ,

$$u(r, \theta) = \sum_{n=0}^{\infty} r^n (a_n \cos(n\theta) + b_n \sin(n\theta))$$

If you want to achieve the boundary condition, then you need to have  $R^n a_n =$

$$\frac{1}{\pi} \int_0^{2\pi} \cos(n\theta) f(\theta) d\theta, \quad a_0 = \frac{1}{2\pi} \int_0^{2\pi} f(\theta) d\theta, \quad R^n b_n = \frac{1}{\pi} \int_0^{2\pi} \sin(n\theta) f(\theta) d\theta$$

If you like, you can simplify this and write an interesting formula for the solution to this problem.

$$\begin{aligned} u(r, \theta) &= \frac{1}{2\pi} \int_0^{2\pi} f(\theta) d\theta + \\ &\frac{1}{\pi} \sum_{n=1}^{\infty} \frac{r^n}{R^n} \left( \left( \int_0^{2\pi} \cos(n\alpha) f(\alpha) d\alpha \right) \cos(n\theta) + \left( \int_0^{2\pi} \sin(n\alpha) f(\alpha) d\alpha \right) \sin(n\theta) \right) \\ &= \frac{1}{2\pi} \int_0^{2\pi} f(\alpha) d\alpha + \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{r^n}{R^n} \int_0^{2\pi} f(\alpha) \cos(n(\theta - \alpha)) d\alpha \end{aligned}$$

In fact, it can be proved that the infinite sum and the integral can be interchanged. This is thanks to the term  $(r/R)^n$  which yields absolute convergence. There is no problem if it were a finite sum and thanks to this term, the tail of the series is negligible. Thus one can reduce to the finite sum case and make the interchange. Thus the above implies

$$u(r, \theta) = \int_0^{2\pi} \frac{1}{\pi} \left( \frac{1}{2} + \sum_{n=1}^{\infty} \frac{r^n}{R^n} \cos(n(\theta - \alpha)) \right) f(\alpha) d\alpha$$

You can find a formula for this.

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{r^n}{R^n} \cos(nt) &= \operatorname{Re} \sum_{n=1}^{\infty} \left( \frac{r}{R} e^{it} \right)^n = \operatorname{Re} \left( \frac{\frac{r}{R} e^{it}}{1 - \frac{r}{R} e^{it}} \right) = \operatorname{Re} \left( \frac{r e^{it}}{R - r e^{it}} \right) \\ &= \frac{Rr \cos(t) - r^2}{(R - r \cos t)^2 + r^2 \sin^2(t)} = \frac{Rr \cos(t) - r^2}{R^2 - 2(\cos t)Rr + r^2} \end{aligned}$$

Then

$$\frac{1}{2} + \sum_{n=1}^{\infty} \frac{r^n}{R^n} \cos(nt) = \frac{1}{2} + \frac{Rr \cos(t) - r^2}{R^2 - 2(\cos t)Rr + r^2} = \frac{1}{2} \frac{R^2 - r^2}{R^2 - 2(\cos t)Rr + r^2}$$



Thus

$$u(r, \theta) = \frac{1}{2\pi} \int_0^{2\pi} \left( \frac{R^2 - r^2}{R^2 - 2(\cos(\theta - \alpha))Rr + r^2} \right) f(\alpha) d\alpha$$

Note that this shows that if  $r = 0$  so you are at the center, then

$$u(r, \theta) = \frac{1}{2\pi} \int_0^{2\pi} f(\alpha) d\alpha$$

so the value at the center is the average of the boundary values. This proves the following fundamental result.

**Theorem 34.4.3** *The solution to the problem*

$$\Delta u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0$$

on the disc of radius  $R$  where on the boundary of this disk,

$$u(R, \theta) = f(\theta), \quad f(0) = f(2\pi)$$

is given by the formula

$$u(r, \theta) = \frac{1}{2\pi} \int_0^{2\pi} \left( \frac{R^2 - r^2}{R^2 - 2(\cos(\theta - \alpha))Rr + r^2} \right) f(\alpha) d\alpha$$

## 34.5 Exercises

1. Solve the following initial boundary value problems.

- (a)  $u_t = u_{xx}, u(x, 0) = 1, u(0, t) = 0, u(4, t) = 0$
- (b)  $u_t = 2u_{xx}, u(x, 0) = 1, u_x(0, t) = 0, u(3, t) = 0$
- (c)  $u_t = 3u_{xx}, u(x, 0) = 1 - x, u(0, t) = 0, u(2, t) = 0$
- (d)  $u_t = 4u_{xx}, u(x, 0) = 1 - x, u(0, t) = 0, u(2, t) = 0$
- (e)  $u_t = 5u_{xx}, u(x, 0) = 1 - x, u(0, t) = 0, u_x(1, t) = 0$
- (f)  $u_t = u_{xx}, u(x, 0) = x + 1, u_x(0, t) = 0, u(2, t) = 0$
- (g)  $u_t = 3u_{xx}, u(x, 0) = x, u_x(0, t) = 0, u(1, t) = 0$
- (h)  $u_t = 3u_{xx}, u(x, 0) = x^2, u(0, t) = 0, u(5, t) = 0$
- (i)  $u_t = 4u_{xx}, u(x, 0) = 1, u(0, t) = 0, u_x(1, t) = 0$
- (j)  $u_t = u_{xx}, u(x, 0) = x, u(0, t) = 0, u_x(4, t) = 0$
- (k)  $u_t = 2u_{xx}, u(x, 0) = 1, u_x(0, t) = 0, u_x(5, t) = 0$
- (l)  $u_t = 2u_{xx}, u(x, 0) = x, u_x(0, t) = 0, u_x(4, t) = 0$
- (m)  $u_t = 2u_{xx}, u(x, 0) = 1 - x, u_x(0, t) = 0, u_x(3, t) = 0$

2. Find the solution to the initial boundary value problem

$$\begin{aligned} u_{tt} &= 3u_{xx}, u(0, t) = 0, u(5, t) = 0, \\ u(x, 0) &= 3x(x - 5), u_t(x, 0) = x^2 \end{aligned}$$

3. Find the solution to the initial boundary value problem

$$\begin{aligned}u_{tt} &= 4u_{xx}, u(0, t) = 0, u(5, t) = 0, \\u(x, 0) &= 3x(x - 5), u_t(x, 0) = x + 1\end{aligned}$$

4. Find the solution to the initial boundary value problem

$$\begin{aligned}u_{tt} &= 4u_{xx}, u(0, t) = 0, u(2, t) = 0, \\u(x, 0) &= -x(x - 2), u_t(x, 0) = x^2\end{aligned}$$

5. Describe how to solve the initial boundary value problem

$$\begin{aligned}u_{tt} + 2u_t &= 2u_{xx}, u(0, t) = 0, u(5, t) = 0, \\u(x, 0) &= -x(x - 5), u_t(x, 0) = x^2\end{aligned}$$

**Hint:** You might consider defining  $w = e^{2t}u$  and see what equation is solved by  $w$ .

6. Find the solution to the initial boundary value problem

$$\begin{aligned}u_t - 2u &= u_{xx}, u(0, t) = 0, u(5, t) = 0, \\u(x, 0) &= x\end{aligned}$$

**Hint:** It is like before. You get eigenfunctions and match coefficients.

7. Find the solution to the initial boundary value problem

$$\begin{aligned}u_t &= u_{xx} + (\cos x), u(0, t) = 0, u(2, t) = 0, \\u(x, 0) &= 1 - x\end{aligned}$$

8. Find the solution to the initial boundary value problem

$$\begin{aligned}u_t &= 2u_{xx} + (x - 1), u(0, t) = 0, u(4, t) = 0, \\u(x, 0) &= 1\end{aligned}$$

9. Find the solution to the initial boundary value problem

$$\begin{aligned}u_t &= 5u_{xx} + (x - 1), u(0, t) = 0, u(1, t) = 0, \\u(x, 0) &= x + 1\end{aligned}$$

10. Find the solution to the initial boundary value problem

$$\begin{aligned}u_t &= 2u_{xx}, u(0, t) = 0, u(5, t) = 0, \\u(x, 0) &= \begin{cases} x & \text{for } x \in [0, \frac{5}{4}] \\ 0 & \text{if } x \in (\frac{5}{4}, 5] \end{cases}\end{aligned}$$

11. Find the solution to the initial boundary value problem

$$\begin{aligned}u_t &= 3u_{xx}, u(0, t) = 0, u(3, t) = 0, \\u(x, 0) &= \begin{cases} 1 & \text{for } x \in [0, 1] \\ 0 & \text{if } x \in (1, 3] \end{cases}\end{aligned}$$

12. Find the solution to the initial boundary value problem

$$\begin{aligned} u_t &= 4u_{xx}, u(0, t) = 0, u(2, t) = 0, \\ u(x, 0) &= \begin{cases} x & \text{for } x \in [0, \frac{2}{3}] \\ 1 - \frac{1}{2}x & \text{if } x \in (\frac{2}{3}, 2] \end{cases} \end{aligned}$$

13. Find the solution to the initial boundary value problem

$$\begin{aligned} u_t &= 5u_{xx}, u(0, t) = 0, u(1, t) = 0, \\ u(x, 0) &= \begin{cases} x & \text{for } x \in [0, \frac{1}{2}] \\ 1 - x & \text{if } x \in (\frac{1}{2}, 1] \end{cases} \end{aligned}$$

14. Find the solution to the initial boundary value problem

$$\begin{aligned} u_t &= 3u_{xx}, u(0, t) = 0, u(5, t) = 0, \\ u(x, 0) &= \begin{cases} x & \text{for } x \in [0, \frac{5}{2}] \\ 5 - x & \text{if } x \in (\frac{5}{2}, 5] \end{cases} \end{aligned}$$

15. Find the solution to the initial boundary value problem

$$\begin{aligned} u_t &= 2u_{xx}, u_x(0, t) = 0, u(2, t) = 0, \\ u(x, 0) &= \begin{cases} x & \text{for } x \in [0, \frac{1}{2}] \\ \frac{2}{3} - \frac{1}{3}x & \text{if } x \in (\frac{1}{2}, 2] \end{cases} \end{aligned}$$

16. Find the solution to the initial boundary value problem

$$\begin{aligned} u_t &= 5u_{xx}, u_x(0, t) = 0, u(1, t) = 0, \\ u(x, 0) &= \begin{cases} x & \text{for } x \in [0, \frac{1}{2}] \\ 1 - x & \text{if } x \in (\frac{1}{2}, 1] \end{cases} \end{aligned}$$

17. Find the solution to the initial boundary value problem

$$\begin{aligned} u_t &= 5u_{xx}, u_x(0, t) = 0, u(2, t) = 0, \\ u(x, 0) &= \begin{cases} x & \text{for } x \in [0, \frac{2}{3}] \\ 1 - \frac{1}{2}x & \text{if } x \in (\frac{2}{3}, 2] \end{cases} \end{aligned}$$

18. Consider the following initial boundary value problem,

$$\begin{aligned} u_t &= u_{xx}, u(0, t) = 0, u(2, t) + u_x(2, t) = 0, \\ u(x, 0) &= f(x) \end{aligned}$$

Determine the appropriate equation for the eigenfunctions and show that there exists a sequence of strictly positive eigenvalues converging to  $\infty$ . Also explain why the solution  $u$  if it exists, must have a limit  $\lim_{t \rightarrow \infty} u(x, t) = w(x)$  and that this limit satisfies  $w(x) = 0$ .

19. Consider the following initial boundary value problem,

$$\begin{aligned} u_t &= u_{xx}, \quad u_x(0, t) = 0, u(2, t) + u_x(2, t) = 0, \\ u(x, 0) &= f(x) \end{aligned}$$

Determine the appropriate equation for the eigenfunctions and show that there exists a sequence of strictly positive eigenvalues converging to  $\infty$ . Also explain why the solution  $u$  if it exists, must have a limit  $\lim_{t \rightarrow \infty} u(x, t) = w(x)$  and that this limit satisfies  $w''(x) = w(x) = 0$ .

20. Consider the following initial boundary value problem,

$$\begin{aligned} u_t &= u_{xx}, \quad u_x(0, t) = 0, u_x(2, t) = 0, \\ u(x, 0) &= f(x) \end{aligned}$$

Determine the appropriate equation for the eigenfunctions and show that there exists a sequence of strictly positive eigenvalues converging to  $\infty$ . Also explain why the solution  $u$  if it exists, must have a limit  $\lim_{t \rightarrow \infty} u(x, t) = \frac{1}{2} \int_0^2 f(x) dx$ .

21. Recall that on the circular disk of radius  $R$  centered at the origin, denoted here as  $D_R$

$$u(r, \theta) = \int_0^{2\pi} \frac{1}{\pi} \left( \frac{1}{2} + \sum_{n=1}^{\infty} \frac{r^n}{R^n} \cos(n(\theta - \alpha)) \right) f(\alpha) d\alpha$$

gave the solution to  $\Delta u = 0$  and  $f(\alpha)$  a given function on the boundary where  $f(0) = f(2\pi)$ . Show, using the divergence theorem from calculus that there is at most one smooth solution to this problem. Then explain why

$$\int_0^{2\pi} \frac{1}{\pi} \left( \frac{1}{2} + \sum_{n=1}^{\infty} \frac{r^n}{R^n} \cos(n(\theta - \alpha)) \right) d\alpha = 1$$

22. Recall that on a simple computation was done which showed that

$$\frac{1}{\pi} \left( \frac{1}{2} + \sum_{n=1}^{\infty} \frac{r^n}{R^n} \cos(n(\theta - \alpha)) \right) = \frac{1}{2\pi} \frac{R^2 - r^2}{R^2 - 2(\cos(\theta - \alpha))Rr + r^2}$$

Therefore,

$$\int_0^{2\pi} \frac{1}{2\pi} \frac{R^2 - r^2}{R^2 - 2(\cos(\theta - \alpha))Rr + r^2} d\alpha = 1$$

Explain why it is also the case that

$$\frac{1}{2\pi} \frac{R^2 - r^2}{R^2 - 2(\cos(\theta - \alpha))Rr + r^2} \geq 0$$

and if  $|\theta - \alpha| \geq \delta > 0$ , then

$$\lim_{r \rightarrow R^-} \frac{1}{2\pi} \frac{R^2 - r^2}{R^2 - 2(\cos(\theta - \alpha))Rr + r^2} = 0$$

uniformly for such  $\alpha$ .

23. The solution to Laplace's equation on the disk  $D_R$  which has boundary values  $f(\alpha)$  was derived and it is

$$u(r, \theta) = \int_0^{2\pi} \frac{1}{2\pi} \frac{R^2 - r^2}{R^2 - 2(\cos(\theta - \alpha))Rr + r^2} f(\alpha) d\alpha$$

Show that

$$\lim_{r \rightarrow R^-} u(r, \theta) = f(\theta)$$

This shows how the boundary values are obtained.

24. Recall that  $u(r, \theta) =$

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} f(\theta) d\theta + \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{r^n}{R^n} & \left( \left( \int_0^{2\pi} \cos(n\alpha) f(\alpha) d\alpha \right) \cos(n\theta) \right. \\ & \left. + \left( \int_0^{2\pi} \sin(n\alpha) f(\alpha) d\alpha \right) \sin(n\theta) \right) \end{aligned}$$

Explain why if  $f$  is a  $2\pi$  periodic continuous function, it follows that there is a trigonometric polynomial which is uniformly close to  $f(\theta)$  for  $\theta \in [0, 2\pi]$ . **Hint:** From the above problem, convergence to  $f(\theta)$  as  $r \rightarrow R^-$  takes place. Note that from the argument, this actually happens uniformly thanks to the uniform continuity of  $f$ . Now argue that the tail  $\sum_{n=N}^{\infty}$  of the above series is uniformly small if  $N$  is large.

25. Let

$$f(x) = \begin{cases} x & \text{if } x \in [0, 1] \\ 2-x & \text{if } x \in [1, 2] \end{cases}$$

Solve the following initial boundary value problems

- (a)  $u_t = a^2 u_{xx}, u(x, 0) = f(x), u(0, t) = 0 = u(2, t)$
- (b)  $u_t = a^2 u_{xx}, u(x, 0) = f(x), u_x(0, t) = 0 = u_x(2, t)$
- (c)  $u_t = a^2 u_{xx}, u(x, 0) = f(x), u_x(0, t) = 0 = u_x(2, t)$



**Part III**

**Fundamentals of Complex  
Analysis**





## Chapter 35

# Analytic Functions

This part of the book is on the fundamentals of complex analysis. I will not try to give theorems in greatest possible generality. My intent is to give a fairly rigorous presentation of those parts of the subject which have the most interesting applications. I think that sometimes, when one tries to give the greatest generality and precision, the fundamental ideas are obscured. These are often very simple ideas and it is too bad when they are lost. Complex analysis is quite different than real analysis. It is relatively free of pathology and often has a much more algebraic flavor than real analysis. I am trying to emphasize these things, many of which are very important in both pure and applied math.

The fundamental theorems of Chapter 13 are going to be needed here.

### 35.1 Cauchy Riemann Equations

Of interest are functions  $f : U \rightarrow \mathbb{C}$  where  $U$  is an open subset of  $\mathbb{R}^2$  and we consider  $\mathbb{R}^2$  to equal  $\mathbb{C}$  where the ordered pair  $(x, y)$  is written as  $x + iy$ . It is customary to write  $\partial U$  to denote the boundary of the open set  $U$ . This means  $\bar{U} \setminus U$  whenever  $U$  is open and it is useful to think of it as the edge of  $U$ . Thus  $\partial B$  is a circle if  $B$  is an open ball. This will be used whenever convenient.

As noted earlier in Section 2.3, the complex numbers forms a field. That is, it acts just like the real numbers. There is a multiplication and addition which satisfy the usual properties which we think numbers should satisfy. Recall from calculus the familiar formula

$$\lim_{h \rightarrow 0} \frac{f(z+h) - f(z)}{h} \equiv f'(z)$$

When functions of many variables were encountered earlier, it was necessary to present this in another way in terms of little  $o$  notation or more directly as

$$\lim_{|v| \rightarrow 0} \frac{|f(x+v) - f(x) - Df(x)v|}{|v|} = 0$$

We had to do it this way because one cannot divide by a vector. However, in the case where  $z \in \mathbb{C}$ , no such worry is necessary. The familiar calculus formula can be used because indeed, you can divide by a nonzero complex number. This leads to the concept of an analytic function which will be presented in what follows. We will see that these are just

like long polynomials. In fact, this is the correct context for the study of power series. Then from calculus, the next thing considered is the rational functions. The generalization of this simple concept will be the meromorphic functions. Remarkable things are obtained from these simple considerations. Surprising applications are available when this theory is developed. I will demonstrate that these extravagant assertions are abundantly verified.

We will be considering line integrals and it will be assumed that the curves over which the line integrals are taken are piecewise  $C^1$ . Actually, all that is needed is that these curves have finite length but this is better considered in a book devoted primarily to the mathematical theory.

## 35.2 The Cauchy Riemann Equations

These fundamental equations pertain to a complex valued function of a complex variable. Recall the complex numbers should be considered as points in the plane. Thus a complex number is of the form  $x + iy$  where  $i^2 = -1$ . The complex conjugate is defined by

$$\overline{x + iy} \equiv x - iy$$

and for  $z$  a complex number,

$$|z| \equiv (z\bar{z})^{1/2} = \sqrt{x^2 + y^2}.$$

Thus when  $x + iy$  is considered an ordered pair  $(x, y) \in \mathbb{R}^2$  the magnitude of a complex number is nothing more than the usual norm of the ordered pair. Also for  $z = x + iy, w = u + iv$ ,

$$|z - w| = \sqrt{(x - u)^2 + (y - v)^2}$$

so in terms of all topological considerations,  $\mathbb{R}^2$  is the same as  $\mathbb{C}$ . Thus to say  $z \rightarrow f(z)$  is continuous, is the same as saying

$$(x, y) \rightarrow u(x, y), (x, y) \rightarrow v(x, y)$$

are continuous where  $f(z) \equiv u(x, y) + iv(x, y)$  with  $u$  and  $v$  being called the real and imaginary parts of  $f$ . The only new thing is that writing an ordered pair  $(x, y)$  as  $x + iy$  with the convention  $i^2 = -1$  makes  $\mathbb{C}$  into a field. You should verify that for  $z, w$  two complex numbers,  $|zw| = |z| |w|$ . Also  $\overline{z + w} = \bar{z} + \bar{w}$ .

Now here is the definition of what it means for a function to be analytic.

**Definition 35.2.1** Let  $U$  be an open subset of  $\mathbb{C}$  ( $\mathbb{R}^2$ ) and let  $f : U \rightarrow \mathbb{C}$  be a function. Then  $f$  is said to be analytic on  $U$  if for every  $z \in U$ ,

$$\lim_{\Delta z \rightarrow 0} \frac{f(z + \Delta z) - f(z)}{\Delta z} \equiv f'(z)$$

exists and is a continuous function of  $z \in U$ . For a function having values in  $\mathbb{C}$  denote by  $u(x, y)$  the real part of  $f$  and  $v(x, y)$  the imaginary part. Both  $u$  and  $v$  have real values and

$$f(x + iy) \equiv f(z) \equiv u(x, y) + iv(x, y)$$

All of the usual methods and formulas for finding the derivative which were discussed in calculus hold with no change for a function of a complex variable. That is, you have the product rule, chain rule, and quotient rule with no change. Also the differentiation of polynomials is the same. The proofs of these theorems are exactly the same as in calculus. Thus I will use the standard methods with no comment whenever convenient. The new thing is a relationship between the partial derivatives of the real and imaginary parts known as the Cauchy Riemann equations.

**Proposition 35.2.2** *Let  $U$  be an open subset of  $\mathbb{C}$ . Then  $f : U \rightarrow \mathbb{C}$  is analytic if and only if for*

$$f(x + iy) \equiv u(x, y) + iv(x, y)$$

$u(x, y), v(x, y)$  being the real and imaginary parts of  $f$ , it follows

$$u_x(x, y) = v_y(x, y), \quad u_y(x, y) = -v_x(x, y)$$

and all these partial derivatives,  $u_x, u_y, v_x, v_y$  are continuous on  $U$ . (The above equations are called the Cauchy Riemann equations.)

**Proof:** First suppose  $f$  is analytic. First let  $\Delta z = ih$  and take the limit of the difference quotient as  $h \rightarrow 0$  in the definition. Thus from the definition,

$$\begin{aligned} f'(z) &\equiv \lim_{h \rightarrow 0} \frac{f(z + ih) - f(z)}{ih} \\ &= \lim_{h \rightarrow 0} \frac{u(x, y + h) + iv(x, y + h) - (u(x, y) + iv(x, y))}{ih} \\ &= \lim_{h \rightarrow 0} \frac{1}{i} (u_y(x, y) + iv_y(x, y)) = -iu_y(x, y) + v_y(x, y) \end{aligned}$$

Next let  $\Delta z = h$  and take the limit of the difference quotient as  $h \rightarrow 0$ .

$$\begin{aligned} f'(z) &\equiv \lim_{h \rightarrow 0} \frac{f(z + h) - f(z)}{h} \\ &= \lim_{h \rightarrow 0} \frac{u(x + h, y) + iv(x + h, y) - (u(x, y) + iv(x, y))}{h} \\ &= u_x(x, y) + iv_x(x, y). \end{aligned}$$

Therefore, equating real and imaginary parts,

$$u_x = v_y, \quad v_x = -u_y \tag{35.1}$$

and this yields the Cauchy Riemann equations. Since  $z \rightarrow f'(z)$  is continuous, it follows the real and imaginary parts of this function must also be continuous. Thus from the above formulas for  $f'(z)$ , it follows from the continuity of  $z \rightarrow f'(z)$  all the partial derivatives of the real and imaginary parts are continuous.

Next suppose the Cauchy Riemann equations hold and these partial derivatives are all continuous. For  $\Delta z = h + ik$ ,

$$\begin{aligned} f(z + \Delta z) - f(z) &= u(x + h, y + k) + iv(x + h, y + k) - (u(x, y) + iv(x, y)) \\ &= u_x(x, y)h + u_y(x, y)k + i(v_x(x, y)h + v_y(x, y)k) + o((h, k)) \end{aligned}$$

$$= u_x(x, y)h + u_y(x, y)k + i(v_x(x, y)h + v_y(x, y)k) + o(\Delta z)$$

This follows since  $C^1$  implies differentiable along with the definition of the norm (absolute value) in  $\mathbb{C}$ . By the Cauchy Riemann equations this equals

$$\begin{aligned} &= u_x(x, y)h - v_x(x, y)k + i(v_x(x, y)h + u_x(x, y)k) + o(\Delta z) \\ &= u_x(x, y)(h + ik) + iv_x(x, y)(h + ik) + o(\Delta z) \\ &= u_x(x, y)\Delta z + iv_x(x, y)\Delta z + o(\Delta z) \end{aligned}$$

Dividing by  $\Delta z$  and taking a limit yields  $f'(z)$  exists and equals  $u_x(x, y) + iv_x(x, y)$  which are assumed to be continuous. ■

For functions of a real variable, it is perfectly possible for the derivative to exist and not be continuous. For example, consider

$$f(x) \equiv \begin{cases} x^2 \sin\left(\frac{1}{x}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

You can verify that  $f'(x)$  exists for all  $x$  but at 0 this derivative is not continuous. This will **NEVER** happen with functions of a complex variable. This will be shown later when it is more convenient. For now make continuity of  $f'$  part of the requirement for  $f$  to be analytic.

### 35.3 Contour Integrals

In the theory of functions of a complex variable, the most important results are those involving contour integration. The most important tools in complex analysis are Cauchy's theorem in some form and Cauchy's formula for an analytic function. These are statements about certain contour integrals. Now a contour integral is just a sort of line integral. In what follows,  $\gamma^*$  or  $\Gamma$  will denote the set of points on a curve and  $\gamma$  will denote a parametrization of the given curve. Here is the definition. It should look familiar and resemble a corresponding definition for line integrals presented earlier. In fact, these contour integrals are just line integrals.

**Definition 35.3.1** Let  $\gamma: [a, b] \rightarrow \mathbb{C}$ ,  $t \in [a, b]$  be a parametrization for a smooth oriented curve  $\Gamma$ , the direction of motion being increasing  $t \in [a, b]$  and let  $f$  be a complex valued function defined on  $\Gamma$ . Then

$$\int_{\gamma} f(z) dz \equiv \int_a^b f(\gamma(t)) \gamma'(t) dt$$

For a piecewise smooth curve  $\gamma$  going from  $z_1$  to  $z_2$  to  $\cdots$  to  $z_m$ , and for  $\gamma_{z_{(k-1)k}}^*$  the curve joining  $z_{k-1}$  to  $z_k$ ,

$$\int_{\gamma} f(z) dz \equiv \sum_{k=1}^m \int_{\gamma_{z_{(k-1)k}}^*} f(z) dz$$

**Example 35.3.2** Let  $\gamma(t) = \cos(t) + i \sin(t)$ ,  $t \in [0, 1]$  and let  $f(z) = z^2$ . Find  $\int_{\gamma} f(t) dz$ .

It equals

$$\begin{aligned}
 & \int_0^1 (\cos(t) + i \sin(t))^2 (-\sin(t) + i \cos(t)) dt \\
 &= \int_0^1 (i \cos^3 t - 3 \cos^2 t \sin t - 3 i \cos t \sin^2 t + \sin^3 t) dt \\
 &= \left( \frac{1}{3} \cos 3 - \frac{1}{3} \right) + \frac{1}{3} i \sin 3
 \end{aligned}$$

As claimed above, every contour integral reduces to a line integral. Say  $z = x + iy$  and  $f(z) = u(x, y) + iv(x, y)$  as above and  $\gamma(t) = x(t) + iy(t)$ ,  $t \in [a, b]$ . Then from the above definition,

$$\begin{aligned}
 \int_{\gamma} f(z) dz &= \int_a^b (u(x(t), y(t)) + iv(x(t), y(t))) (x'(t) + iy'(t)) dt \\
 &= \int_a^b (u(x(t), y(t))x'(t) - v(x(t), y(t))y'(t) \\
 &\quad + i(v(x(t), y(t))x'(t) + u(x(t), y(t))y'(t))) dt \\
 &\equiv \int_{\Gamma} u(x, y) dx - v(x, y) dy + i \int_{\Gamma} v(x, y) dx + u(x, y) dy
 \end{aligned}$$

which is indeed, just the sum of two line integrals. Thus all the theory of line integrals applies. In particular, the contour integral is dependent only on the smooth curves and their orientation. This yields most of the following lemma.

**Lemma 35.3.3** *Let  $f$  be defined and continuous on a piecewise smooth oriented curve  $\Gamma$  contained in  $\mathbb{C}$  having parametrization  $\gamma$ . Let the real and imaginary parts of  $f$  be denoted by  $u$  and  $v$  respectively. Then*

$$\int_{\gamma} f(z) dz = \int_{\Gamma} u dx - v dy + i \int_{\Gamma} v dx + u dy$$

Also the following estimate is available.

$$\left| \int_{\gamma} f(z) dz \right| \leq \max(|f(z)| : z \in \gamma^*) (\text{length of } \gamma^*)$$

If  $f_n$  is continuous and

$$\lim_{n \rightarrow \infty} (\sup(|f_n(z) - f(z)| : z \in \Gamma)) = 0 \quad (35.2)$$

then

$$\lim_{n \rightarrow \infty} \int_{\gamma} f_n(z) dz = \int_{\gamma} f(z) dz \quad (35.3)$$

**Proof:** It only remains to verify the estimate.  $\int_{\gamma} f(z) dz$  is some complex number  $I$  so let

$$\omega = \begin{cases} \frac{I}{|I|} & \text{if } I \neq 0 \\ 1 & \text{if } I = 0 \end{cases}$$

Thus  $|\omega| = 1$  and  $\omega \int_{\gamma} f(z) dz = \left| \int_{\gamma} f(z) dz \right|$ . Then letting  $\gamma$  be a parametrization for a smooth curve,

$$\begin{aligned} \left| \int_{\gamma} f(z) dz \right| &= \omega \int_{\gamma} f(z) dz = \int_a^b \omega f(\gamma(t)) \gamma'(t) dt \leq \int_a^b |f(\gamma(t))| |\gamma'(t)| dt \\ &\leq \max(|f(z)| : z \in \gamma^*) \int_a^b |\gamma'(t)| dt \end{aligned}$$

Now recall that this last integral is the definition of the length of  $\gamma^*$ . If the curve  $\Gamma$  is piecewise  $C^1$  composed of smooth curves  $\gamma_i$ , Then

$$\begin{aligned} \left| \int_{\Gamma} f(z) dz \right| &\equiv \left| \sum_{j=1}^m \int_{\gamma_j} f(z) dz \right| \leq \sum_{j=1}^m \left| \int_{\gamma_j} f(z) dz \right| \\ &\leq \sum_{j=1}^m \max(|f(z)| : z \in \Gamma) (\text{length of } \gamma_j^*) \\ &= \max(|f(z)| : z \in \Gamma) (\text{length of } \Gamma) \end{aligned}$$

Consider the last claim. From Theorem 13.6.3,  $z \rightarrow f(z)$  is continuous. Therefore, the integral makes sense. Also from the estimate,

$$\begin{aligned} \left| \int_{\Gamma} f(z) dz - \int_{\Gamma} f_n(z) dz \right| &= \left| \int_{\Gamma} (f(z) - f_n(z)) dz \right| \\ &\leq \max(|f(z) - f_n(z)| : z \in \gamma^*) (\text{length of } \gamma^*) \end{aligned}$$

and by assumption, this last expression converges to 0 as  $n \rightarrow \infty$ . This shows 35.3. ■

**Observation 35.3.4** In the case that  $\gamma^* = [a, b]$  an interval on the real line, the above definition of the contour integral shows that if  $\gamma$  is oriented from  $a$  to  $b$ , then  $\int_{\gamma} f(z) dz = \int_a^b f(z) dz$  and if  $\gamma$  is oriented from  $b$  to  $a$ , then  $\int_{\gamma} f(z) dz = \int_b^a f(z) dz$  where the notation on the right signifies the usual Riemann integral.

**Definition 35.3.5** If one reverses the order in which points of  $\gamma^*$  are encountered, then one replaces  $\gamma$  with  $-\gamma$  in which, for  $\gamma : [a, b] \rightarrow \mathbb{C}$ ,  $-\gamma(t)$  encounters the points of  $\gamma^*$  in the opposite order, the definition of the contour integral shows that

$$-\int_{\gamma} f(z) dz = \int_{-\gamma} f(z) dz$$

You could get a parametrization for  $-\gamma$  as  $-\gamma(t) \equiv \gamma(b-t)$  for  $t \in [0, b-a]$  or if you wanted to use the same interval, define  $-\gamma : [a, b] \rightarrow \mathbb{C}$  by  $-\gamma(t) \equiv \gamma(b+a-t)$ . A simple closed piecewise  $C^1$  curve is one which has the first point encountered equal to the last point encountered by the parametrization. We will only consider closed curves for which Green's theorem applies to the curve and its inside which will be denoted as  $U_i$ .

One other technical result is often useful. It involves interchanging the order of contour integrals.

Recall the mean value theorem for integrals from calculus.

**Lemma 35.3.6** Let  $f : [a, b] \rightarrow \mathbb{R}$  be continuous. Then there exists  $c \in (a, b)$  such that

$$f(c)(b-a) = \int_a^b f(x) dx$$

**Proof:** Let  $F(x) \equiv \int_a^x f(t) dt$ . Then by the mean value theorem,

$$F(b) - F(a) = F'(c)(b-a)$$

for some  $c \in (a, b)$ . But  $F'(x) = f(x)$  and so this proves the lemma. ■

**Lemma 35.3.7** Let  $\gamma, \eta$  be parametrizations for two smooth curves,  $\gamma([a, b])$  and  $\eta([c, d])$  and let  $f : \gamma^* \times \eta^* \rightarrow \mathbb{R}$  be continuous. Then

$$\int_{\eta} \int_{\gamma} f(z, w) dz dw = \int_{\gamma} \int_{\eta} f(z, w) dw dz$$

In other words, you can switch the contour integrals.

**Proof:** That on the left is by definition,

$$\int_{\eta} \int_{\gamma} f(z, w) dz dw = \int_c^d \int_a^b f(\gamma(t), \eta(s)) \gamma'(t) \eta'(s) dt ds$$

Let  $P$  be a partition for  $[a, b]$  and  $Q$  a partition for  $[c, d]$ . Then the above is

$$\begin{aligned} & \sum_{i=1}^n \sum_{j=1}^m \int_{s_{i-1}}^{s_i} \int_{t_{j-1}}^{t_j} f(\gamma(t), \eta(s)) \gamma'(t) \eta'(s) dt ds \\ &= \sum_{i=1}^n \sum_{j=1}^m \int_{s_{i-1}}^{s_i} f(\gamma(\hat{t}_j), \eta(s)) \gamma'(\hat{t}_j) (t_j - t_{j-1}) \eta'(s) ds \\ &= \sum_{i=1}^n \sum_{j=1}^m f(\gamma(\hat{t}_j), \eta(\hat{s}_i)) \gamma'(\hat{t}_j) \eta'(\hat{s}_i) (t_j - t_{j-1}) (s_i - s_{i-1}) \end{aligned}$$

by an application of the mean value theorem for integrals from calculus. Here  $(\hat{t}_j, \hat{s}_i) \in (t_{j-1}, t_j) \times (s_{i-1}, s_i)$ . Similarly,

$$\begin{aligned} \int_{\gamma} \int_{\eta} f(z, w) dw dz &= \sum_{j=1}^m \sum_{i=1}^n f(\gamma(\tilde{t}_j), \eta(\tilde{s}_i)) \gamma'(\tilde{t}_j) \eta'(\tilde{s}_i) (t_j - t_{j-1}) (s_i - s_{i-1}) \\ &= \sum_{i=1}^n \sum_{j=1}^m f(\gamma(\tilde{t}_j), \eta(\tilde{s}_i)) \gamma'(\tilde{t}_j) \eta'(\tilde{s}_i) (t_j - t_{j-1}) (s_i - s_{i-1}) \end{aligned}$$

where  $(\tilde{t}_j, \tilde{s}_i) \in (t_{j-1}, t_j) \times (s_{i-1}, s_i)$ . By uniform continuity, if  $\|P\|, \|Q\|$  are small enough, then

$$|f(\gamma(\tilde{t}_j), \eta(\tilde{s}_i)) \gamma'(\tilde{t}_j) \eta'(\tilde{s}_i) - f(\gamma(\hat{t}_j), \eta(\hat{s}_i)) \gamma'(\hat{t}_j) \eta'(\hat{s}_i)| < \varepsilon$$

and so

$$\left| \int_{\eta} \int_{\gamma} f(z, w) dz dw - \int_{\gamma} \int_{\eta} f(z, w) dw dz \right| < \varepsilon (b-a)(d-c).$$

Since  $\varepsilon$  is arbitrary, the two contour integrals must be equal. ■

**Theorem 35.3.8** Let  $\gamma, \eta$  be two piecewise smooth oriented curves. Then if the oriented parametrizations that go with  $\gamma^*$  are respectively  $\gamma_1, \gamma_2, \dots, \gamma_n$  and the oriented parametrizations that go with  $\eta$  are respectively  $\eta_1, \eta_2, \dots, \eta_m$ , then if  $f : \gamma^* \times \eta^* \rightarrow \mathbb{C}$  is continuous,

$$\int_{\gamma} \int_{\eta} f(z, w) dw dz = \int_{\eta} \int_{\gamma} f(z, w) dz dw$$

**Proof:** First suppose  $f$  has values in  $\mathbb{R}$ . Then, starting with the left and using Lemma 35.3.7,

$$\begin{aligned} \int_{\gamma} \int_{\eta} f(z, w) dw dz &\equiv \sum_{k=1}^n \int_{\gamma_k} \sum_{l=1}^m \int_{\eta_l} f(z, w) dw dz = \sum_{k=1}^n \sum_{l=1}^m \int_{\gamma_k} \int_{\eta_l} f(z, w) dw dz \\ &= \sum_{k=1}^n \sum_{l=1}^m \int_{\eta_l} \int_{\gamma_k} f(z, w) dw dz = \sum_{l=1}^m \sum_{k=1}^n \int_{\eta_l} \int_{\gamma_k} f(z, w) dw dz \\ &= \int_{\eta} \int_{\gamma} f(z, w) dz dw \end{aligned}$$

In the general case, you simply apply this to the real and imaginary parts of  $f$ . ■

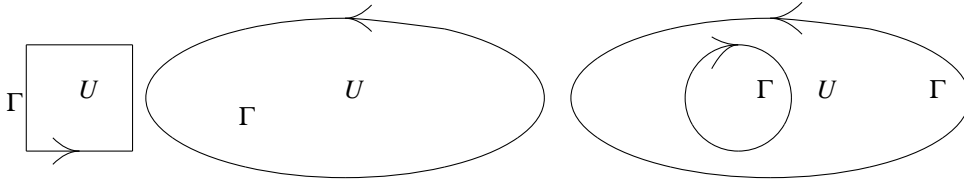
The main result is the Cauchy integral theorem which is presented next. First recall Green's theorem.

**Theorem 35.3.9** (Green's Theorem) Let  $V$  be an open set in the plane and let its boundary  $\Gamma$  be piecewise smooth and let  $\mathbf{F}(x, y) = (P(x, y), Q(x, y))$  be a  $C^1$  vector field defined near  $V$ . Then if  $\Gamma$  is oriented counter clockwise, it is often<sup>1</sup> the case that

$$\int_{\Gamma} \mathbf{F} \cdot d\mathbf{R} = \int_V \left( \frac{\partial Q}{\partial x}(x, y) - \frac{\partial P}{\partial y}(x, y) \right) dm_2. \quad (35.4)$$

In particular, if there exists  $U$  such as the simple convex in both directions case considered earlier for which Green's theorem holds, and  $V = \mathbf{R}(U)$  where  $\mathbf{R} : U \rightarrow V$  is  $C^2(\overline{U}, \mathbb{R}^2)$  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 35.4 is valid where the orientation around  $\Gamma$  is consistent with the orientation around  $U$ . Also, one can paste together regions for which Green's theorem holds to get another one for which Green's theorem holds.

Here are some examples of regions for which Green's theorem holds:



Recall that you determine the positive orientation for use with Green's theorem as follows. You regard  $\mathbf{k}$  as pointing out of the paper because the  $x$  axis points to the right and the

<sup>1</sup>For a general version see the advanced calculus book by Apostol. This is presented in the next section also. 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.



y axis points up. Then the motion is such that if your head points in the direction of  $\mathbf{k}$ , your left hand will be over the surface if you walk in the direction of the positive orientation.

These examples work for Green's theorem and if you have a  $C^2$  mapping defined near these regions, then if the resulting curves around images of  $U$  are oriented consistent with the above orientations, then you have another example of a region and its boundary for which Green's theorem holds.

## 35.4 Cauchy Integral Theorem

With the above preparation, here is the Cauchy integral theorem. It is really very simple. It involves the Cauchy Riemann equations and Green's theorem.

**Theorem 35.4.1** *Let  $U$  be an open set and suppose  $U$  and its boundary  $\Gamma$  satisfy Green's theorem where  $\Gamma$  is suitably oriented for using Green's theorem. Suppose also that  $f$  is analytic on an open set containing  $U \cup \Gamma$ . Then*

$$\int_{\Gamma} f(z) dz = 0.$$

**Proof:** From Lemma 35.3.3, the contour integral is

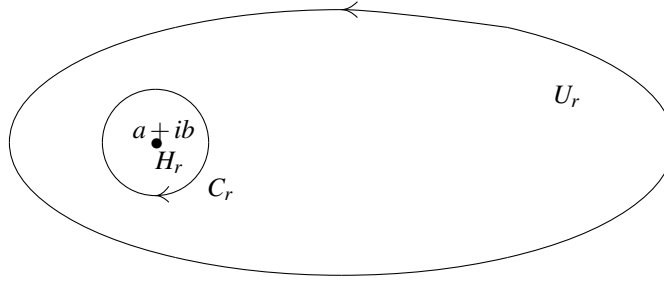
$$\int_{\Gamma} u dx - v dy + i \int_{\Gamma} v dx + u dy$$

By Green's theorem, this equals

$$\int_U (-v_x - u_y) dA + \int_U (u_x - v_y) dA = 0$$

thanks to the Cauchy Riemann equations. ■

Now this yields an easy way to check orientation of a piecewise smooth simple closed curve. Consider the following picture of a simple closed curve in which there is a hole on its inside denoted as  $H_r$ , its boundary being  $C_r$  as shown.



**Theorem 35.4.2** *Let  $\Gamma$  be a simple closed curve in  $\mathbb{C}$  and let  $z \in U$ , the inside component of  $\Gamma^c$ . Then for  $\gamma$  a parametrization of  $\Gamma$ ,*

$$n(\gamma, z) \equiv \frac{1}{2\pi i} \int_{\gamma} \frac{1}{w - z} dw = \pm 1$$

depending on the orientation of  $\Gamma$ . If  $z \notin U \cup \Gamma$ , the integral equals 0.  $n(\gamma, z)$  is called the winding number.

**Proof:** Denote by  $\Gamma_r$  the insider circle in the above picture having radius  $r$  oriented as shown. Then  $f(w) = \frac{1}{w-z}$  has a derivative which is

$$\frac{-1}{(w-z)^2}$$

a continuous function, and so its real and imaginary parts are continuous for  $w \neq z$ . Therefore, the function is analytic near  $U_r$  the open set bounded by the two curves  $\Gamma_r$  and  $\Gamma$ . It follows from the Cauchy theorem that for  $\gamma$  an orientation on  $\Gamma$  as shown and  $\hat{\gamma}_r$  an orientation as shown on  $\Gamma_r$ ,

$$\int_{\gamma} \frac{1}{w-z} dw + \int_{\hat{\gamma}_r} \frac{1}{w-z} dw = 0$$

Therefore, orienting  $\Gamma_r$  in the usual direction, a parametrization for this circle is

$$x = a + r \cos t, y = b + r \sin t, t \in [0, 2\pi]$$

Deote this parametrization by  $\gamma_r$ . Then

$$\int_{\gamma} \frac{1}{w-z} dw = \int_{\gamma_r} \frac{1}{w-z} dw$$

and using the definition of the contour integral, the right side reduces to  $2\pi i$ . Thus the winding number is 1. Therefore, if  $\Gamma$  were oriented the opposite direction, you would get  $-1$  for the winding number. If  $z \notin U \cup \Gamma$ , the function is analytic near  $U$  and so the Cauchy integral theorem implies right away that the winding number is 0. ■

The expression  $\frac{1}{2\pi i} \int_{\gamma} \frac{1}{w-z} dw \equiv n(\gamma, z)$  is called the winding number. As explained, it is either 1 or  $-1$  depending on how the curve  $\Gamma$  is oriented. The winding number can be defined with much more generality for any closed curve, simple or not. However, I will not do so, choosing instead to emphasize the most basic ideas. The greater generality is needed however, when you consider general versions of the Cauchy integral formula, a marvelous representation theorem for an analytic function.

**Definition 35.4.3** Given  $\Gamma$  a simple closed curve, the orientation is said to be positive if the winding number is 1 and negative if the winding number is  $-1$ .

## 35.5 Primitives and Cauchy Goursat Theorem

In beginning calculus, the notion of an antiderivative was very important. It is similar for functions of complex variables. The role of a primitive is also a lot like a potential in computing line integrals.

**Definition 35.5.1** A function  $F$  such that  $F' = f$  is called a **primitive** of  $f$ .

The following theorem shows that the primitive acts just like a potential, the difference being that a primitive has complex, not real values. In calculus, in the context of a function of one real variable, this is often called an antiderivative and every continuous function has one thanks to the fundamental theorem of calculus. However, it will be shown below that the situation is not at all the same for functions of a complex variable.

So what if a function has a primitive? Say  $F'(z) = f(z)$  where  $f$  is continuous.

**Theorem 35.5.2** Suppose  $\gamma$  is a piecewise  $C^1$  curve. Let its endpoints be  $p$  and  $q$  with the orientation of the curve from  $p$  to  $q$ . Suppose  $f : \gamma^* \rightarrow \mathbb{C}$  is continuous and has a primitive  $F$ . Thus  $F'(z) = f(z)$  for some open set  $\Omega \supseteq \gamma^*$ . Then

$$\int_{\gamma} f(z) dz = F(q) - F(p)$$

**Proof:** Assume first that  $\gamma$  is a  $C^1$  curve defined on an interval  $[a, b]$ . Then by definition,

$$\begin{aligned} \int_{\gamma} f(z) dz &= \int_a^b f(\gamma(t)) \gamma'(t) dt = \int_a^b \frac{d}{dt} (F(\gamma(t))) dt \\ &= F(\gamma(b)) - F(\gamma(a)) = F(q) - F(p) \end{aligned}$$

Now in the general case, you have

$$\gamma_j : [a_j, b_j] \rightarrow \mathbb{C}$$

is  $C^1$  and  $\gamma_j(b_j) = \gamma_{j+1}(a_{j+1})$ ,  $j \leq m$ . Then

$$\int_{\gamma} f(z) dz \equiv \sum_{j=1}^m \int_{\gamma_j} f(z) dz = \sum_{j=1}^m (F(\gamma_j(b_j)) - F(\gamma_j(a_j)))$$

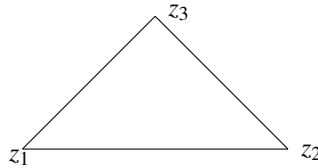
where  $\gamma_m(b_m) = q$  and  $\gamma_1(a_1) = p$ .

$$\begin{aligned} &= F(q) - F(\gamma_m(a_m)) + (F(\gamma_{m-1}(b_{m-1})) - F(\gamma_{m-1}(a_{m-1}))) \\ &\quad + (F(\gamma_{m-1}(b_{m-1})) - F(\gamma_{m-1}(a_{m-1}))) + \cdots + \\ &\quad + (F(\gamma_1(b_1)) - F(p)) \end{aligned}$$

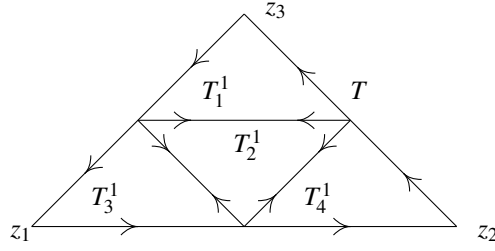
By assumption, this reduces to  $F(q) - F(p)$  because  $\gamma_j(b_j) = \gamma_{j+1}(a_{j+1})$  for each  $j$ . ■

The Cauchy Goursat theorem is the next big result. This is a major theorem which does not depend on the derivative being continuous. Thus it will also provide the needed generalization which involves not assuming that  $z \rightarrow f'(z)$  is continuous.

If you have two points in  $\mathbb{C}$ ,  $z_1$  and  $z_2$ , you can consider  $\gamma(t) \equiv z_1 + t(z_2 - z_1)$  for  $t \in [0, 1]$  to obtain a continuous bounded variation curve from  $z_1$  to  $z_2$ . More generally, if  $z_1, \dots, z_m$  are points in  $\mathbb{C}$  you can obtain a continuous bounded variation curve from  $z_1$  to  $z_m$  which consists of first going from  $z_1$  to  $z_2$  and then from  $z_2$  to  $z_3$  and so on, till in the end one goes from  $z_{m-1}$  to  $z_m$ . Denote this piecewise linear curve as  $\gamma(z_1, \dots, z_m)$ . Now let  $T$  be a triangle with vertices  $z_1, z_2$  and  $z_3$  encountered in the counter clockwise direction as shown.



Denote by  $\int_{\partial T} f(z) dz$ , the expression,  $\int_{\gamma(z_1, z_2, z_3, z_1)} f(z) dz$ . Consider the following picture.



Thus

$$\int_{\partial T} f(z) dz = \sum_{k=1}^4 \int_{\partial T_k^1} f(z) dz. \quad (35.5)$$

On the “inside lines” the integrals cancel because there are two integrals going in opposite directions for each of these inside lines. Recall the method for evaluating a line integral with a  $C^1$  parametrization.

**Theorem 35.5.3** (Cauchy Goursat) *Let  $f : \Omega \rightarrow X$ , where  $\Omega$  is an open subset of  $\mathbb{C}$  and  $X$  is a complex complete normed linear space, have the property that  $f'(z)$  exists for all  $z \in \Omega$  and let  $T$  be a triangle contained in  $\Omega$ . Then*

$$\int_{\partial T} f(w) dw = 0.$$

**Proof:** Suppose not. Then

$$\left| \int_{\partial T} f(w) dw \right| = \alpha \neq 0.$$

From 35.5 it follows

$$\alpha \leq \sum_{k=1}^4 \left| \int_{\partial T_k^1} f(w) dw \right|$$

and so for at least one of these  $T_k^1$ , denoted from now on as  $T_1$ ,

$$\left| \int_{\partial T_1} f(w) dw \right| \geq \frac{\alpha}{4}.$$

Now let  $T_1$  play the same role as  $T$ . Subdivide as in the above picture, and obtain  $T_2$  such that

$$\left| \int_{\partial T_2} f(w) dw \right| \geq \frac{\alpha}{4^2}.$$

Continue in this way, obtaining a sequence of triangles,

$$T_k \supseteq T_{k+1}, \text{diam}(T_k) \leq \text{diam}(T) 2^{-k},$$

and

$$\left| \int_{\partial T_k} f(w) dw \right| \geq \frac{\alpha}{4^k}.$$

Then let  $z \in \cap_{k=1}^{\infty} T_k$  and note that by assumption,  $f'(z)$  exists. Therefore, for all  $k$  large enough,

$$\int_{\partial T_k} f(w) dw = \int_{\partial T_k} (f(z) + f'(z)(w-z) + g(w)) dw$$

where  $|g(w)| < \varepsilon |w-z|$ . Now observe that  $w \rightarrow f(z) + f'(z)(w-z)$  has a primitive, namely,

$$F(w) = f(z)w + f'(z)(w-z)^2/2.$$

Therefore, by Theorem 35.5.2,

$$\int_{\partial T_k} f(w) dw = \int_{\partial T_k} g(w) dw.$$

From Theorem 35.3.3,

$$\begin{aligned} \frac{\alpha}{4^k} &\leq \left| \int_{\partial T_k} g(w) dw \right| \leq \varepsilon \text{diam}(T_k) (\text{length of } \partial T_k) \\ &\leq \varepsilon 2^{-k} (\text{length of } T) \text{diam}(T) 2^{-k}, \end{aligned}$$

and so

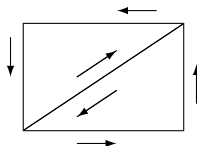
$$\alpha \leq \varepsilon (\text{length of } T) \text{diam}(T).$$

Since  $\varepsilon$  is arbitrary, this shows  $\alpha = 0$ , a contradiction. Thus  $\int_{\partial T} f(w) dw = 0$  as claimed.

■

**Note that no assumption of continuity of  $z \rightarrow f'(z)$  was needed.**

Obviously, there is a version of the above Cauchy Goursat theorem which is valid for a rectangle. Indeed, apply the Cauchy Goursat theorem for the triangles obtained from a diagonal of the rectangle. The diagonal will be oriented two different ways depending on which triangle it is a part of.



**Corollary 35.5.4** *Let  $\Omega$  be an open set on which  $f'(z)$  exists. Then if  $R$  is a rectangle contained in  $\Omega$  along with its inside, then orienting  $R$  either way results in*

$$\int_R f(z) dz = 0.$$

The following is a general version of the Cauchy integral theorem. If  $f'(z)$  exists on the inside and if  $f$  is continuous on the boundary, then the integral over the bounding curve is 0. Note how the closed curve is arbitrary, not just a triangle.

## 35.6 Functions Differentiable on a Disk, Zeros

It turns out that if a function has a derivative, then it has all of them, in contrast to functions of a real variable.

**Theorem 35.6.1** (Morera<sup>2</sup>) Let  $\Omega$  be an open set and let  $f'(z)$  exist for all  $z \in \Omega$ . Let  $D \equiv \overline{B(z_0, r)} \subseteq \Omega$ . Then there exists  $\varepsilon > 0$  such that  $f$  has a primitive on  $B(z_0, r + \varepsilon)$ .

**Proof:** Choose  $\varepsilon > 0$  small enough that  $B(z_0, r + \varepsilon) \subseteq \Omega$ . Then for  $w \in B(z_0, r + \varepsilon)$ , define

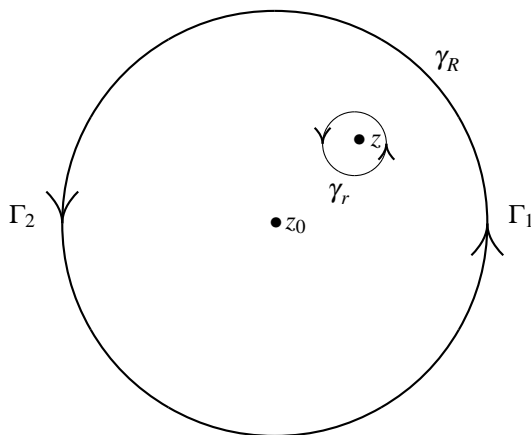
$$F(w) \equiv \int_{\gamma(z_0, w)} f(u) du.$$

Then by the Cauchy Goursat theorem, and  $w \in B(z_0, r + \varepsilon)$ , it follows that for  $|h|$  small enough,

$$\begin{aligned} \frac{F(w+h) - F(w)}{h} &= \frac{1}{h} \int_{\gamma(w, w+h)} f(u) du \\ &= \frac{1}{h} \int_0^1 f(w+th) h dt = \int_0^1 f(w+th) dt \end{aligned}$$

which converges to  $f(w)$  due to the continuity of  $f$  at  $w$ . ■

Consider the following picture where you have a large circle of radius  $R$  and a small circle of radius  $r$  centered at  $z$ , a point on the inside of  $\gamma_R$ . The Cauchy integral formula gives  $f(z)$  in terms of the values of  $f$  on the large circle.



**Theorem 35.6.2** Let  $\gamma_R$  be a positively oriented circle of radius  $R$  and let  $U$  be its inside. Suppose  $f$  has a derivative on an open set containing  $U \cup \gamma_R^*$ . Then if  $z \in U$ ,

$$f(z) = \frac{1}{2\pi i} \int_{\gamma_R} \frac{f(w)}{w-z} dw$$

**Proof:** Use  $-\gamma_r$  for the orientation of the smaller circle. Then from the Cauchy integral theorem above, if  $w \rightarrow g(w)$  is analytic,

$$0 = \int_{\gamma_R} g(w) dw + \int_{-\gamma_r} g(w) dw = \int_{\gamma_R} g(w) dw - \int_{\gamma_r} g(w) dw$$

This results from using  $-\gamma_r$  on the small circle so the small circle has the opposite orientation indicated in the picture. Now let

$$g(w) = \frac{f(w)}{w-z}$$

<sup>2</sup>Giacinto Morera 1856-1909. This theorem or one like it dates from around 1886

This has derivative outside the small disk and inside some open set containing the large disk. Also

$$\int_{\gamma_R} \frac{f(w)}{w-z} dw = \int_{\gamma_r} \frac{f(w)}{w-z} dw$$

Now, since  $\gamma_r$  is oriented positively,

$$\left| \frac{1}{2\pi i} \int_{\gamma_r} \frac{f(w)}{w-z} dw - f(z) \right| = \left| \frac{1}{2\pi i} \int_{\gamma_r} \frac{f(w) - f(z)}{w-z} dw \right| \quad (35.6)$$

Since  $f'(z)$  exists,

$$\begin{aligned} \frac{f(w) - f(z)}{w-z} &= \frac{f(z) + f'(z)(w-z) + o(w-z) - f(z)}{w-z} \\ &= f'(z) + \frac{o(w-z)}{w-z} \end{aligned}$$

Now  $f'(z)$  is a constant and so it has a primitive, namely  $w \rightarrow f'(z)w$ . Thus  $\int_{\gamma_r} f'(z) dw = 0$ . It follows that if  $r$  is sufficiently small, then

$$\left| \frac{1}{2\pi i} \int_{\gamma_r} \frac{f(w) - f(z)}{w-z} dw \right| \leq \frac{1}{2\pi} 2\pi r \epsilon \frac{1}{r} = \epsilon$$

Thus, as  $r \rightarrow 0$ , the right term in 35.6 converges to 0. It follows that

$$\frac{1}{2\pi i} \int_{\gamma_R} \frac{f(w)}{w-z} dw = \lim_{r \rightarrow 0} \frac{1}{2\pi i} \int_{\gamma_r} \frac{f(w) - f(z)}{w-z} dw + f(z) = f(z) \blacksquare$$

This is the Cauchy integral formula for a disk. This remarkable formula is sufficient to show that if a function has a derivative, then it has infinitely many and in fact, the function can be represented as a power series. When this is shown, it will be easy to give the general Cauchy integral formula for an arbitrary piecewise smooth simple closed curve. Let  $z_0$  be the center of the large circle.

In the situation of Theorem 35.6.2,

$$f(z) = \frac{1}{2\pi i} \int_{\gamma_R} \frac{f(w)}{w-z_0 - (z-z_0)} dw = \frac{1}{2\pi i} \int_{\gamma_R} \frac{1}{w-z_0} \frac{f(w)}{1 - \frac{z-z_0}{w-z_0}} dw$$

Now  $\left| \frac{z-z_0}{w-z_0} \right| = \frac{|z-z_0|}{R} < 1$  for all  $w \in \gamma_R^*$ . Therefore, the above equals

$$\frac{1}{2\pi i} \int_{\gamma_R} \sum_{k=0}^{\infty} \frac{f(w)(z-z_0)^k}{(w-z_0)^{k+1}} dw = \frac{1}{2\pi i} \int_{\gamma_R} \left( \sum_{k=0}^{\infty} \frac{(z-z_0)^k}{(w-z_0)^{k+1}} \right) f(w) dw$$

If the partial sums of the above series converge uniformly on  $\gamma_R^*$  then by Lemma 35.3.3,

$$\begin{aligned} & \frac{1}{2\pi i} \int_{\gamma_R} \left( \sum_{k=0}^{\infty} \frac{(z-z_0)^k}{(w-z_0)^{k+1}} \right) f(w) dw \\ &= \lim_{p \rightarrow \infty} \frac{1}{2\pi i} \int_{\gamma_R} \left( \sum_{k=0}^p \frac{(z-z_0)^k}{(w-z_0)^{k+1}} \right) f(w) dw \end{aligned}$$

$$= \lim_{p \rightarrow \infty} \frac{1}{2\pi i} \sum_{k=0}^p \int_{\gamma_R} \frac{(z-z_0)^k}{(w-z_0)^{k+1}} f(w) dw \quad (35.7)$$

Which by definition is

$$\sum_{k=0}^{\infty} \left( \frac{1}{2\pi i} \int_{\gamma_R} \frac{1}{(w-z_0)^{k+1}} f(w) dw \right) (z-z_0)^k$$

It is assumed that  $f$  is continuous on  $U_i \cup \gamma_R^*$ . Thus there is an upper bound for  $|f(w)|$ , called  $M$  thanks to the extreme value theorem. Then for  $w \in \gamma_R^*$ ,

$$\left| \frac{f(w)(z-z_0)^k}{(w-z_0)^{k+1}} \right| \leq \frac{1}{|z-z_0|} M \left( \frac{|z-z_0|}{R} \right)^{k+1} < \frac{M}{R} \left( \frac{|z-z_0|}{R} \right)^k$$

and  $|z-z_0|/R < 1$  so the right side is summable. Therefore, by Theorem 13.8.3, convergence is indeed uniform on  $\gamma_R^*$  and so

$$f(z) = \sum_{k=0}^{\infty} \left( \frac{1}{2\pi i} \int_{\gamma_R} \frac{1}{(w-z_0)^{k+1}} f(w) dw \right) (z-z_0)^k \equiv \sum_{k=0}^{\infty} a_k (z-z_0)^k$$

This proves part of the next theorem which says, among other things, that when  $f$  has one derivative on the interior of a disk, then it must have all derivatives.

**Theorem 35.6.3** *Suppose  $z_0 \in U$ , an open set in  $\mathbb{C}$  and  $f : U \rightarrow X$  has a derivative for each  $z \in U$ . Then if  $B(z_0, R) \subseteq U$ , then for each  $z \in B(z_0, R)$ ,*

$$f(z) = \sum_{n=0}^{\infty} a_n (z-z_0)^n. \quad (35.8)$$

where

$$a_n \equiv \frac{1}{2\pi i} \int_{\gamma_R} \frac{1}{(w-z_0)^{n+1}} f(w) dw$$

and  $\gamma_R$  is a positively oriented parametrization for the circle bounding  $B(z_0, R)$ . Then

$$f^{(k)}(z_0) = k! a_k, \quad (35.9)$$

$$\limsup_{n \rightarrow \infty} |a_n|^{1/n} |z-z_0| < 1, \quad (35.10)$$

$$f^{(k)}(z) = \sum_{n=k}^{\infty} n(n-1) \cdots (n-k+1) a_n (z-z_0)^{n-k}, \quad (35.11)$$

**Proof:** 35.8 follows from the above argument. Now consider 35.10. The above argument based on the Cauchy integral formula for a disk shows that if  $R > |\hat{z}-z_0| > |z-z_0|$ , then

$$f(\hat{z}) = \sum_{n=0}^{\infty} a_n (\hat{z}-z_0)^n$$

and so, by the root test, Theorem 13.7.1,

$$1 \geq \limsup_{n \rightarrow \infty} |a_n|^{1/n} |\hat{z}-z_0| > \limsup_{n \rightarrow \infty} |a_n|^{1/n} |z-z_0|$$



Consider 35.11 which involves identifying the  $a_n$  in terms of the derivatives of  $f$ . This is obvious if  $k = 0$ . Suppose it is true for  $k$ . Then for small  $h \in \mathbb{C}$ ,

$$\begin{aligned}
 & \frac{1}{h} \left( f^{(k)}(z+h) - f^{(k)}(z) \right) \\
 &= \frac{1}{h} \sum_{n=k}^{\infty} n(n-1) \cdots (n-k+1) a_n \left( (z+h-z_0)^{n-k} - (z-z_0)^{n-k} \right) \\
 &= \frac{1}{h} \sum_{n=k+1}^{\infty} n(n-1) \cdots (n-k+1) a_n \left( \sum_{j=0}^{n-k} \binom{n-k}{j} h^j (z-z_0)^{(n-k)-j} - (z-z_0)^{n-k} \right) \\
 &= \sum_{n=k+1}^{\infty} n(n-1) \cdots (n-k+1) a_n \left( \sum_{j=1}^{n-k} \binom{n-k}{j} h^{j-1} (z-z_0)^{(n-k)-j} \right) \\
 &= \sum_{n=k+1}^{\infty} n(n-1) \cdots (n-k+1) (n-k) a_n (z-z_0)^{(n-k)-1} \\
 &+ h \sum_{n=k+1}^{\infty} n(n-1) \cdots (n-k+1) a_n \left( \sum_{j=2}^{n-k} \binom{n-k}{j} h^{j-2} (z-z_0)^{(n-k)-j} \right) \quad (35.12)
 \end{aligned}$$

By what was shown earlier,

$$\begin{aligned}
 & \limsup_{n \rightarrow \infty} \left( n(n-1) \cdots (n-k+1) |a_n| |z-z_0|^{n-k} \right)^{1/n} \\
 &= \limsup_{n \rightarrow \infty} |a_n|^{1/n} |z-z_0| < 1 \quad (35.13)
 \end{aligned}$$

Consider the part of 35.12 which multiplies  $h$ . Does the infinite series converge? Yes it does. In fact it converges absolutely.

$$\begin{aligned}
 & \sum_{n=k+1}^{\infty} \left| n(n-1) \cdots (n-k+1) a_n \left( \sum_{j=2}^{n-k} \binom{n-k}{j} h^{j-2} (z-z_0)^{(n-k)-j} \right) \right| \\
 &\leq \sum_{n=k+1}^{\infty} n(n-1) \cdots (n-k+1) |a_n| |z-z_0|^{(n-2)-k} \sum_{j=2}^{n-k} \binom{n-k}{j} \frac{|h|^{j-2}}{|z-z_0|^{j-2}} \\
 &\leq \sum_{n=k+1}^{\infty} n(n-1) \cdots (n-k+1) |a_n| |z-z_0|^{(n-2)-k} \left( 1 + \frac{|h|}{|z-z_0|} \right)^{n-k}
 \end{aligned}$$

For all  $h$  small enough, this series converges, the infinite sum being decreasing in  $|h|$ . Indeed,

$$\begin{aligned}
 & \limsup_{n \rightarrow \infty} \left( n(n-1) \cdots (n-k+1) |a_n| |z-z_0|^{(n-2)-k} \left( 1 + \frac{|h|}{|z-z_0|} \right)^{n-k} \right)^{1/n} \\
 &= \limsup_{n \rightarrow \infty} |a_n|^{1/n} |z-z_0| \left( 1 + \frac{|h|}{|z-z_0|} \right) < 1
 \end{aligned}$$

if  $|h|$  is small enough. Thus we can take a limit as  $h \rightarrow 0$  in 35.12 and conclude that

$$f^{(k+1)}(z) = \sum_{n=k+1}^{\infty} n(n-1) \cdots (n-k+1) (n-k) a_n (z-z_0)^{n-(k+1)} \blacksquare$$

**Corollary 35.6.4** Suppose  $f$  is continuous on  $\partial B(z_0, r)$  and suppose that for all  $z \in B(z_0, r)$ ,

$$f(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(w)}{w-z} dw,$$

where  $\gamma$  is the positively oriented boundary of the circular disk, conveniently given as  $\gamma(t) \equiv z_0 + re^{it}, t \in [0, 2\pi]$ . Then  $f$  is analytic on  $B(z_0, r)$  and in fact has infinitely many derivatives on  $B(z_0, r)$ .

**Proof:** This is just a repeat of the above arguments. You show that  $f(z)$  is given by a power series for  $|z - z_0| < r$  and from this, the result follows. ■

The following is very different than what is expected in real analysis. It says that uniform convergence tends to take with it differentiability.

**Lemma 35.6.5** Let  $\gamma(t) = z_0 + re^{it}$ , for  $t \in [0, 2\pi]$ , suppose  $f_n \rightarrow f$  uniformly on  $\overline{B(z_0, r)}$ , and suppose

$$f_n(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f_n(w)}{w-z} dw \quad (35.14)$$

for  $z \in B(z_0, r)$ . Then

$$f(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(w)}{w-z} dw, \quad (35.15)$$

implying that  $f$  is analytic on  $B(z_0, r)$ .

**Proof:** From 35.14 and the uniform convergence of  $f_n$  to  $f$  on  $\gamma([0, 2\pi])$ , the integrals in 35.14 converge to

$$\frac{1}{2\pi i} \int_{\gamma} \frac{f(w)}{w-z} dw.$$

Therefore, the formula 35.15 follows. ■

Because of the above result, from now on, the term analytic will be used interchangeably with “has a derivative”. This has shown that if the function has one derivative on an open set, then it has all of them. Now here is another version of Morera’s theorem.

**Corollary 35.6.6** Let  $\Omega$  be an open set and suppose that whenever

$$\gamma(z_1, z_2, z_3, z_1)$$

is a closed curve bounding a triangle  $T$ , which is contained in  $\Omega$ , and  $f$  is a continuous function defined on  $\Omega$ , it follows that

$$\int_{\gamma(z_1, z_2, z_3, z_1)} f(z) dz = 0,$$

then  $f$  is analytic on  $\Omega$ .

**Proof:** As in the proof of Morera’s theorem, let  $\overline{B(z_0, r)} \subseteq \Omega$  and use the given condition to construct a primitive,  $F$  for  $f$  on  $B(z_0, r)$ . Then  $F$  is analytic and so by Theorem 35.6.3, it follows that  $F$  and hence  $f$  have infinitely many derivatives, implying that  $f$  is analytic on  $B(z_0, r)$ . Since  $z_0$  is arbitrary, this shows  $f$  is analytic on  $\Omega$ . ■

The following observation is useful to keep in mind.

**Observation 35.6.7** Suppose  $\sum_{n=0}^{\infty} a_n h^n$  converges for  $|h| < r$ . Then

$$\lim_{h \rightarrow 0} \frac{1}{h^k} \sum_{n=k+1}^{\infty} a_n h^n = 0$$

To see this, note the expression is  $h \sum_{n=k+1}^{\infty} a_n h^{n-(k+1)}$ . Now the sum of the absolute values is  $\sum_{n=k+1}^{\infty} |a_n| |h|^{n-(k+1)}$  and it converges because there exists  $\hat{h}$ , such that  $r > |\hat{h}| > |h|$  and by the root test, Theorem 13.7.1,  $\limsup_{n \rightarrow \infty} |a_n|^{1/n} |\hat{h}| \leq 1$  so

$$\limsup_{n \rightarrow \infty} |a_n|^{1/n} |h| < 1$$

Now applying this to the sum in question,

$$\limsup_{n \rightarrow \infty} |a_n|^{1/n} |h|^{\frac{n-(k+1)}{n}} = \limsup_{n \rightarrow \infty} |a_n|^{1/n} |h| < 1$$

Also the sum decreases in  $|h|$  and so

$$\lim_{h \rightarrow 0} \left| h \sum_{n=k+1}^{\infty} a_n h^{n-(k+1)} \right| \leq \lim_{h \rightarrow 0} |h| \sum_{n=k+1}^{\infty} |a_n| |h|^{n-(k+1)} = 0$$

The tail of the series just described is sometimes referred to as “higher order terms”.

The following is a remarkable result about the zeros of an analytic function on a connected open set. It turns out that if the set of zeros have a limit point, then the function ends up being zero. It is an illustration of how analytic functions are a lot like polynomials which have finitely many zeros unless they are identically zero.

**Definition 35.6.8** Suppose  $f$  is an analytic function defined near a point,  $\alpha$  where  $f(\alpha) = 0$ . Thus  $\alpha$  is a zero of the function  $f$ . The zero is of order  $m$  if  $f(z) = (z - \alpha)^m g(z)$  where  $g$  is an analytic function which is not equal to zero at  $\alpha$ .

**Theorem 35.6.9** Let  $\Omega$  be a connected open set (region) and let  $f : \Omega \rightarrow \mathbb{C}$  be analytic. Then the following are equivalent.

1.  $f(z) = 0$  for all  $z \in \Omega$
2. There exists  $z_0 \in \Omega$  such that  $f^{(n)}(z_0) = 0$  for all  $n$ .
3. There exists  $z_0 \in \Omega$  which is a limit point of the set,

$$Z \equiv \{z \in \Omega : f(z) = 0\}.$$

**Proof:** It is clear the first condition implies the second two. Suppose the third holds. Then for  $z$  near  $z_0$

$$f(z) = \sum_{n=k}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z - z_0)^n$$

where  $k \geq 1$  since  $z_0$  is a zero of  $f$ . Suppose  $k < \infty$ . Then,

$$f(z) = (z - z_0)^k g(z)$$

where  $g(z_0) \neq 0$ . Letting  $z_n \rightarrow z_0$  where  $z_n \in Z, z_n \neq z_0$ , it follows

$$0 = (z_n - z_0)^k g(z_n)$$

which implies  $g(z_n) = 0$ . Then by continuity of  $g$ , we see that  $g(z_0) = 0$  also, contrary to the choice of  $k$ . Therefore,  $k$  cannot be less than  $\infty$  and so  $z_0$  is a point satisfying the second condition, all derivatives at  $z_0$  are zero.

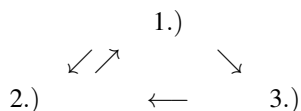
Now suppose the second condition and let

$$S \equiv \left\{ z \in \Omega : f^{(n)}(z) = 0 \text{ for all } n \right\}.$$

It is clear that  $S$  is a closed set which by assumption is nonempty. However, this set is also open. To see this, let  $z \in S$ . Then for all  $w$  close enough to  $z$ ,

$$f(w) = \sum_{k=0}^{\infty} \frac{f^{(k)}(z)}{k!} (w-z)^k = 0.$$

Thus  $f$  is identically equal to zero near  $z \in S$ . Therefore, all points near  $z$  are contained in  $S$  also, showing that  $S$  is an open set. Now  $\Omega = S \cup (\Omega \setminus S)$ , the union of two disjoint open sets,  $S$  being nonempty. It follows the other open set,  $\Omega \setminus S$ , must be empty because  $\Omega$  is connected. Therefore, the first condition is verified. This proves the theorem. (See the following diagram.)



Note how radically different this is from the theory of functions of a real variable. Consider, for example the function

$$f(x) \equiv \begin{cases} x^2 \sin\left(\frac{1}{x}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

which has a derivative for all  $x \in \mathbb{R}$  and for which 0 is a limit point of the set  $Z$ , even though  $f$  is not identically equal to zero.

Here is a very important application called Euler's formula. Recall that

$$e^z \equiv e^x (\cos(y) + i \sin(y)) \quad (35.16)$$

Is it also true that  $e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}$ ?

**Theorem 35.6.10 (Euler's Formula)** Let  $z = x + iy$ . Then

$$e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}.$$

**Proof:** The Cauchy Riemann equations show that  $e^z$  given by 35.16 is analytic. So is  $\exp(z) \equiv \sum_{k=0}^{\infty} \frac{z^k}{k!}$ . In fact the power series converges for all  $z \in \mathbb{C}$ . Furthermore the two functions,  $e^z$  and  $\exp(z)$  agree on the real line which is a set which contains a limit point. Therefore, they agree for all values of  $z \in \mathbb{C}$ . ■

This formula shows the famous two identities,

$$e^{i\pi} = -1 \text{ and } e^{2\pi i} = 1.$$

This properties of zeros of an analytic function can be used to verify with no effort that identities which hold for  $z$  real continue to hold for  $z$  complex and this can be done with no effort.

## 35.7 Liouville's Theorem

Now the following is the general Cauchy integral formula.

**Theorem 35.7.1** *Let  $U$  along with its boundary  $\Gamma$  satisfy satisfy Green's theorem and let  $f$  be analytic on an open set  $V$  containing  $U \cup \Gamma$  and let  $\gamma$  be an orientation of  $\Gamma$  such that Green's theorem holds. Thus,*

$$n(\gamma, z) \equiv \frac{1}{2\pi i} \int_{\gamma} \frac{1}{w-z} dw = 1$$

Then if  $z \in U$ ,

$$f(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(w)}{w-z} dw$$

**Proof:** Consider the function

$$g(w) \equiv \begin{cases} \frac{f(w)-f(z)}{w-z} & \text{if } w \neq z \\ f'(z) & \text{if } w = z \end{cases} \quad (35.17)$$

It remains to consider whether  $g'(z)$  exists for  $z \in V$ . Then from the Theorem 35.6.3, we can write  $f(z+h)$  as a power series in  $h$  whenever  $h$  is suitably small.

$$\begin{aligned} & \frac{\frac{f(z+h)-f(z)}{h} - f'(z)}{h} = \\ & \frac{1}{h} \left( \frac{1}{h} \left( f'(z)h + \frac{1}{2!}f''(z)h^2 + \frac{1}{3!}f'''(z)h^3 + \dots \right) - f'(z) \right) \\ & = \frac{1}{h} \left( \left( f'(z) + \frac{1}{2!}f''(z)h + \frac{1}{3!}f'''(z)h^2 + \dots \right) - f'(z) \right) \\ & = \frac{1}{2!}f''(z) + \frac{1}{3!}f'''(z)h + \text{higher order terms} \end{aligned}$$

Thus the limit of the difference quotient exists and is  $\frac{1}{2!}f''(z)$ . It follows that

$$\begin{aligned} 0 &= \frac{1}{2\pi i} \int_{\gamma} g(w) dw = \frac{1}{2\pi i} \int_{\gamma} \frac{f(w)}{w-z} dw - \frac{1}{2\pi i} \int_{\gamma} \frac{f(z)}{w-z} dw \\ &= \frac{1}{2\pi i} \int_{\gamma} \frac{f(w)}{w-z} dw - f(z) \blacksquare \end{aligned}$$

The following is a spectacular application. It is Liouville's theorem.

**Theorem 35.7.2** Suppose  $f$  is analytic on  $\mathbb{C}$  and that  $|f(z)|$  is bounded for  $z \in \mathbb{C}$ . Then  $f$  is constant.

**Proof:** It was shown above that if  $\gamma_r$  is a counter clockwise oriented parametrization for the circle of radius  $r$  centered at  $z$ , then

$$f'(z) = \frac{1}{2\pi i} \int_{\gamma_r} \frac{f(w)}{(w-z)^2} dw \text{ if } |z| < r$$

and so

$$|f'(z)| \leq \frac{1}{2\pi} C 2\pi r \frac{1}{r^2}$$

where  $|f(z)| < C$  for all  $z$  and this is true for any  $r$  so let  $r \rightarrow \infty$  and you can conclude that  $f'(z) = 0$  for all  $z \in \mathbb{C}$ . However, this shows that  $f^{(k)}(z) = 0$  for all  $z$  and for each  $k \geq 1$ . Thus the power series for  $f(z)$ , which exists by Theorem 35.6.3, is

$$f(z) = f(0) + \sum_{k=1}^{\infty} \frac{f^{(k)}(0)}{k!} z^k = f(0). \blacksquare$$

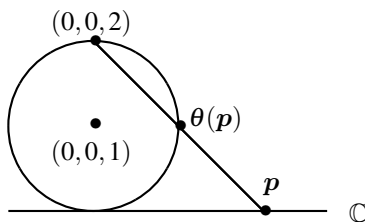
This leads right away to the shortest proof of the fundamental theorem of algebra.

**Theorem 35.7.3** Let  $p(z)$  be a non constant polynomial with complex coefficients. Then  $p(z) = 0$  for some  $z \in \mathbb{C}$ . That is,  $p(z)$  has a root in  $\mathbb{C}$ .

**Proof:** Suppose not. Then  $1/p(z)$  is analytic on  $\mathbb{C}$ . Also, the leading order term dominates the others and so  $1/p(z)$  must be bounded. Indeed,  $\lim_{|z| \rightarrow \infty} (1/|p(z)|) = 0$  and the continuous function  $z \rightarrow 1/|p(z)|$  achieves a maximum on any bounded ball centered at 0 by the extreme value theorem. By Liouville's theorem, this quotient must be constant. However, by assumption, this does not take place. Hence there is a root of  $p(z)$ .  $\blacksquare$

## 35.8 Riemann Sphere

I do not wish to emphasize the Riemann sphere in this book but some mention of it is appropriate. Consider the unit sphere,  $S^2$  given by  $(z-1)^2 + y^2 + x^2 = 1$ . Define a map from the complex plane to the surface of this sphere as follows. Extend a line from the point,  $p$  in the complex plane to the point  $(0,0,2)$  on the top of this sphere and let  $\theta(p)$  denote the point of this sphere which the line intersects. Define  $\theta(\infty) \equiv (0,0,2)$ .



Then  $\theta^{-1}$  is sometimes called stereographic projection. The mapping  $\theta$  is clearly continuous because it takes converging sequences, to converging sequences. Furthermore, it is clear that  $\theta^{-1}$  is also continuous. In terms of the extended complex plane  $\hat{\mathbb{C}}$ , consisting of

$\mathbb{C}$  along with a point called  $\infty$ , a sequence,  $z_n$  converges to  $\infty$  if and only if  $\theta z_n$  converges to  $(0,0,2)$  if and only if  $|z_n|$  converges to  $\infty$  in the usual manner from calculus and a sequence,  $z_n$  converges to  $z \in \mathbb{C}$  if and only if  $\theta(z_n) \rightarrow \theta(z)$ . This is interesting because of this last part. It gives a meaning for a sequence of complex numbers to converge to something called  $\infty$ . To do this properly, we should define a metric on  $\hat{\mathbb{C}}$  and word everything in terms of this metric. However, it amounts to the same thing as saying what it means for sequences to converge. Then, with this definition of what it means for a sequence of complex numbers to converge to  $\infty$ , the usual definition of connected sets and separated sets is identical with what was given earlier.

**Definition 35.8.1** Let  $S \subseteq \hat{\mathbb{C}}$  the extended complex plane in which this extra point  $\infty$  has been included as just described. Then  $S$  is separated if there exist  $A, B$  not both empty such that  $S = A \cup B$ ,  $A \cap B = \emptyset$  and no point of  $A$  is a limit of any sequence of points of  $B$  while no point of  $B$  is the limit of any sequence of points of  $A$ . If  $S$  is not separated, then it is called connected.

**Example 35.8.2** Consider the open set  $S \equiv \{z \in \mathbb{C} \text{ such that } \text{Im}(z) > 0\}$ . Then  $S \cup \{\infty\} \equiv \hat{S}$  is connected in  $\hat{\mathbb{C}}$ .

It is obvious that  $S$  is connected in  $\mathbb{C}$  because it is arcwise connected. Suppose  $\hat{S} = A \cup B$  where these two new sets separate  $\hat{S}$  in  $\hat{\mathbb{C}}$ . Then one of them, say  $B$  must contain  $\infty$ . Therefore,  $A$  is bounded since otherwise there would be a sequence of points of  $A$  converging to  $\infty$  which is assumed not to happen. Then  $S = A \cup (B \setminus \{\infty\})$  and  $A, B \setminus \{\infty\}$  would separate  $S$  unless one is empty. If  $B \setminus \{\infty\} = \emptyset$ , then  $S$  would be bounded which is not the case. Hence  $A = \emptyset$ . Thus  $\hat{S}$  is connected.

**Definition 35.8.3** Let  $S \subseteq \mathbb{C}$ . It is said to be simply connected if the set is connected and  $\mathbb{C} \setminus S \cup \{\infty\}$  is connected in  $\hat{\mathbb{C}}$ . Written more compactly,  $S$  is simply connected means  $S$  is connected and also  $\hat{\mathbb{C}} \setminus S$  is connected in  $\hat{\mathbb{C}}$ .

When looking at a set  $S$  in  $\mathbb{C}$ , how do you determine whether it is simply connected? You consider  $\theta(S^c)$  in  $S^2$  and ask whether it is connected with the convention that if  $S^c$  is unbounded, you must include  $(0,0,2)$  in the image of  $\theta$ .

**Example 35.8.4** Consider the set  $S \equiv \{z \in \mathbb{C} \text{ such that } |z| > 1\}$ . This is a connected set, but it is not simply connected because  $\hat{\mathbb{C}} \setminus S$  is not connected. On  $S^2$  it consists of a piece near the bottom of the sphere and the point  $(0,0,2)$  at the top.

**Example 35.8.5** Consider  $S \equiv \{z \in \mathbb{C} \text{ such that } |z| \leq 1\}$ . This connected set is simply connected because  $\hat{\mathbb{C}} \setminus S$  corresponds to a connected set on  $S^2$ .

## 35.9 Exercises

In the following exercises, the term “simple closed curve” will be used repeatedly. Assume that such curves  $\Gamma$  have an inside  $U_i$  and an outside and that Green’s theorem applies for  $U_i$  with its boundary  $\Gamma$  if the boundary is oriented appropriately. This can be proved, but is not in this book. It is one of these things which is mainly of mathematical interest. In the examples of interest, it is typically not an issue.

1. Suppose you have  $U \subseteq \mathbb{C}$  an open set and  $f : U \rightarrow \mathbb{C}$  is analytic but has only real values. Find all possible  $f$  with these properties.
2. Suppose  $f$  is an entire function (analytic on  $\mathbb{C}$ ) and suppose  $\operatorname{Re} f$  is never 0. Show that  $f$  must be constant. **Hint:** Consider  $U = \{(x, y) : \operatorname{Re} f(x, y) > 0\}$ ,  $V = \{(x, y) : \operatorname{Re} f(x, y) < 0\}$ . These are open and disjoint so one must be empty. If  $V$  is empty, consider  $1/e^{f(z)}$ . Use Liouville's theorem.
3. Suppose  $f : \mathbb{C} \rightarrow \mathbb{C}$  is analytic. Suppose also there is an estimate

$$|f(z)| \leq M(1 + |z|^\alpha), \alpha > 0$$

Show that  $f$  must be a polynomial. **Hint:** Consider the formula for the derivative in which  $\gamma_r$  is positively oriented and a circle of radius  $r$  for  $r$  very large centered at 0,

$$f^{(n)}(z) = \frac{n!}{2\pi i} \int_{\gamma_r} \frac{f(w)}{(w-z)^{n+1}}$$

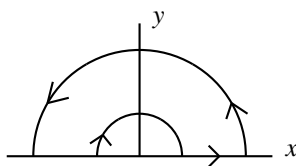
and pick large  $n$ . Then let  $r \rightarrow \infty$ .

4. Define for  $z \in \mathbb{C}$   $\sin z \equiv \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k+1}}{(2k+1)!}$ . That is, you just replace  $x$  with  $z$ . Give a similar definition for  $\cos z$ , and  $e^z$ . Show that the series converges for  $\sin z$  and that a corresponding series converges for  $\cos z$ . Then show that

$$\sin z = \frac{e^{iz} - e^{-iz}}{2i}, \quad \cos z = \frac{e^{iz} + e^{-iz}}{2}$$

Show that it is not longer true that the functions  $\sin z, \cos z$  must be bounded in absolute value by 1. **Hint:** This is a very easy problem if you use the theorem about the zeros of an analytic function, Theorem 35.6.9.

5. Verify the identities  $\cos(z-w) = \cos z \cos w + \sin z \sin w$  and similar identities. **Hint:** This is a very easy problem if you use the theorem about the zeros of an analytic function, Theorem 35.6.9.
6. Consider the following contour in which the large semicircle has radius  $R$  and the small one has radius  $r \equiv 1/R$ .



The function  $z \rightarrow \frac{e^{iz}}{z}$  is analytic on the curve and on its inside. Therefore, the contour integral with respect to the given orientation is 0. Use this contour and the Cauchy integral theorem to verify that  $\int_0^\infty \frac{\sin z}{z} dz = \pi/2$  where this improper integral is defined as

$$\lim_{R \rightarrow \infty} \int_{-1/R}^R \frac{\sin z}{z} dz$$

The function is actually not absolutely integrable and so the precise description of its meaning just given is important. To do this, show that the integral over the large



circle of  $\int_{C_R} \frac{e^{-z}}{z} dz \rightarrow 0$  as  $R \rightarrow \infty$  and verify that you get something else like  $-\pi$  for the integral over the small integral as  $r \rightarrow 0$ .

7. A set  $U$  is star shaped if there exists a single point  $z_0 \in U$  such that every segment from  $z_0$  to  $z$  is contained in  $U$ . Now suppose that  $U$  is star shaped and  $f : U \rightarrow \mathbb{C}$  is analytic. Show that  $f$  has a primitive on  $U$ .
8. Let  $U$  be what remains of  $\mathbb{C}$  after  $(-\infty, 0]$  is deleted. Explain why  $U$  is star shaped. Letting  $\gamma(1, z)$  be the straight line segment from 1 to  $z$ , let  $f(z) = \int_{\gamma(1, z)} \frac{1}{w} dw$ . Explain why  $f'(z) = \frac{1}{z}$ ,  $f(1) = 0$ . Now explain why  $f$  is analytic and why  $e^{f(z)} = z$  for all  $z \in U$ . Also formulate an assertion which says  $f(e^z) = z$  for suitable  $z$ . This  $f$  is the principal logarithm, denoted  $\log(z)$ .
9. Explain why one could delete any ray starting at 0 and obtain a function  $f(z)$  which is a primitive of  $1/z$ .
10. For  $z \in \mathbb{C} \setminus (-\infty, 0]$ , let  $\arg(z) \equiv \theta \in (-\pi, \pi)$  such that  $z = |z|e^{i\theta}$ . Show that

$$\log(z) = \ln|z| + i\arg(z).$$

11. Suppose  $f(z) = u(x, y) + iv(x, y)$  is analytic. Show that both  $u, v$  satisfy Laplace's equation,  $u_{xx} + u_{yy} = 0$ .
12. Suppose you have two complex numbers  $z = a + ib$  and  $w = x + iy$ . Show that the dot product of the two vectors  $(a, b) \cdot (x, y)$  is  $\operatorname{Re}((a + ib)(x - iy)) = \operatorname{Re}(z\bar{w})$ .
13.  $\uparrow$  Suppose you have two curves  $t \rightarrow z(t)$  and  $s \rightarrow w(s)$  which intersect at some point  $z_0$  corresponding to  $t = t_0$  and  $s = s_0$ . Show that the cosine of the angle  $\theta$  between these two curves at this point is

$$\cos(\theta) = \frac{\operatorname{Re}\left(z'(t_0)\overline{w'(s_0)}\right)}{|z'(t_0)||w'(s_0)|}$$

Now suppose  $z \rightarrow f(z)$  is analytic. Thus there are two curves  $t \rightarrow f(z(t))$  and  $s \rightarrow f(w(s))$  which intersect when  $t = t_0$  and  $s = s_0$ . Show that the angle between these two new curves at their point of intersection is also  $\theta$ . This shows that analytic mappings preserve the angles between curves.

14. Suppose  $z = x + iy$  and  $f(z) = u(x, y) + iv(x, y)$  where  $f$  is analytic. Explain why level curves of  $u$  and  $v$  intersect in right angles.
15. Let  $\Gamma$  be a simple closed piecewise  $C^1$  curve in  $\mathbb{C}$ . Let  $\gamma$  be a parametrization of  $\Gamma$  which has positive orientation. Thus  $n(\gamma, z) = 1$  for all  $z$  inside  $\Gamma$ . Also suppose  $f$  is an analytic function on a connected open set containing  $\Gamma$  and its inside  $U_i$ . Suppose  $f$  is not identically zero and has no zeros on  $\Gamma$ . Explain why  $f$  has finitely many zeros on the inside of  $\Gamma$ . A zero  $a$  has multiplicity  $m$  if  $f(z) = (z - a)^m g(z)$  where  $g(z) \neq 0$  on  $U_i$ . Let the zeros of  $f$  in  $U_i$  be  $\{a_1, \dots, a_m\}$  where there might be repeated numbers in this list, zeros of multiplicity higher than 1. Show that

$$m = \frac{1}{2\pi i} \int_{\gamma} \frac{f'(z)}{f(z)} dz \quad (35.18)$$

Thus you can count the zeros of an analytic function inside a simple closed curve by doing an integral! **Hint:** First of all,  $m$  is finite since if not, Theorem 35.6.9 implies that  $f(z) = 0$  for all  $z$  since there would be a limit point or else a zero of infinite order. Now argue that  $f(z) = \prod_{k=1}^m (z - a_k) g(z)$  where  $g(z)$  is analytic and nonzero on  $U_i$ . Use the product rule to simplify  $\frac{f'(z)}{f(z)}$ . Then use the fact that  $n(\gamma, z) = 1$ .

16. Suppose now you have a piecewise  $C^1$  simple closed curve  $\Gamma$  and on  $\Gamma^*$ ,  $|f(z)| > |g(z)|$  where  $f, g$  are analytic on an open set containing  $\Gamma^*$ . Suppose also that  $f$  has no zeros on  $\Gamma^*$ . In particular,  $f$  is not identically 0. Let  $\lambda \in [0, 1]$ .

(a) Verify that for  $\lambda \in [0, 1]$ ,  $f + \lambda g$  has no zeros on  $\Gamma^*$ .

(b) Verify that on  $\Gamma^*$ ,  $\left| \frac{f'(z) + \lambda g'(z)}{f(z) + \lambda g(z)} - \frac{f'(z) + \mu g'(z)}{f(z) + \mu g(z)} \right| \leq C |\mu - \lambda|$ .

(c) Use Theorem 35.3.3 to show that for  $\gamma$  a positively oriented parametrization of  $\Gamma$ ,

$$\lambda \rightarrow \frac{1}{2\pi i} \int_{\gamma} \frac{f'(z) + \lambda g'(z)}{f(z) + \lambda g(z)} dz$$

is continuous.

(d) Now explain why this shows that the number of zeros of  $f + \lambda g$  on the inside of  $\Gamma$  is the same as the number of zeros of  $f$  on the inside of  $\Gamma$ . This is a version of Rouché's theorem.

17. Give an extremely easy proof of the fundamental theorem of algebra as follows. Let  $\gamma_R$  be a parametrization of the circle centered at 0 having radius  $R$  which has positive orientation so  $n(\gamma, z) = 1$ . Let  $p(z)$  be a polynomial  $a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$ . Now explain why you can choose  $R$  so large that  $|a_n z^n| > |a_{n-1} z^{n-1} + \cdots + a_1 z + a_0|$  for all  $|z| \geq R$ . Using Problem 16 above explain why all zeros of  $p(z)$  are inside  $\gamma_R^*$  and why there are exactly  $n$  of them counted according to multiplicity.

18. The polynomial  $z^5 + z^4 - z^3 - 3z^2 - 5z + 1 = p(z)$  has no rational roots. You can check this by applying the rational root theorem from algebra. However, it has five complex roots. Also

$$|z^4 - z^3 - 3z^2 - 5z + 1| \leq |z|^4 + |z|^3 + 3|z|^2 + 5|z| + 1$$

By graphing, observe that  $x^5 - (x^4 + x^3 + 3x^2 + 5x + 1) > 0$  for all  $x \geq 2.4$ . Explain why the roots of  $p(z)$  are inside the circle  $|z| = 2.4$ .

19. This problem will feature the situation where the radius of the simple closed curve is sufficiently small. The zero counting integral can be used to prove an open mapping theorem for analytic functions. Suppose you have  $f(z) = f(z_0) + \phi(z)^m$  for  $z \in V$  an open set containing  $z_0$  and  $\phi(z_0) = 0$ ,  $|\phi'(z_0)| = 2r \neq 0$ , and  $m \in \mathbb{N}$ . Let  $C(a, \rho)$  denote the positively oriented circle centered at  $a$  which has radius  $\rho$ .

(a) Explain why there exists  $\delta > 0$  such that if  $|z - z_0| = \delta$ , then  $B(z_0, \delta) \subseteq V$  and

$$\left| \frac{\phi(z)}{z - z_0} \right| \geq r, \quad |\phi(z)| \geq r|z - z_0| = r\delta$$

Therefore, if  $|w| < r\delta$ , then if  $|z - z_0| = \delta$ ,  $|\phi(z) - w| \neq 0$ .

- (b) Use continuity of  $w \rightarrow \frac{1}{2\pi i} \int_{C(z_0, \delta)} \frac{\phi'(z)}{\phi(z)-w} dz$  for  $|w| < \delta/2$  and Problem 15 to conclude that there exists  $\varepsilon < r\delta$  such that if  $|w| < \varepsilon$  there is one zero for  $\phi(z) - w$  in  $B(z_0, \delta)$ . In other words,  $\phi(B(z_0, \delta)) \supseteq B(0, \varepsilon)$ . Then also  $\phi^m(B(z_0, \delta)) \supseteq B(0, \varepsilon^m)$ . **Hint:** If you have  $w \in B(0, \varepsilon^m)$ , then there are  $m$   $m^{\text{th}}$  roots of  $w$  equally spaced around  $B(0, |w|^{1/m})$ . Thus these roots are on a circle of radius less than  $\varepsilon$ . Pick one. Call it  $\hat{w}$ . Then there exists  $z \in B(z_0, \delta)$  such that  $\phi(z) = \hat{w}$ . Then  $\phi^m(z) = w$ . Fill in details.
- (c) Explain why  $f(B(z_0, \delta)) \supseteq f(z_0) + B(0, \varepsilon^m)$  and why for  $w \in f(z_0) + B(0, \varepsilon^m)$  there are  $m$  different points in  $B(z_0, \delta)$ ,  $z_1, \dots, z_m$  such that  $f(z_j) = w$ .
20. †Let  $\Omega$  be an open connected set. Let  $f: \Omega \rightarrow \mathbb{C}$  be analytic. Suppose  $f(\Omega)$  is not a single point. Then pick  $z_0 \in \Omega$ . Explain why  $f(z) = f(z_0) + (z - z_0)^m g(z)$  for all  $z \in V$  an open ball contained in  $\Omega$  which contains  $z_0$  and  $g(z) \neq 0$  in  $V$ ,  $g(z)$  analytic. If this were not so, then  $z_0$  would be a zero of infinite order and by the theorem on zeros, Theorem 35.6.9,  $f(z) = f(z_0)$  for all  $z \in \Omega$  which is assumed not to happen. Thus, every  $z_0$  in  $\Omega$  has this property that near  $z_0$ ,  $f(z) = f(z_0) + (z - z_0)^m g(z)$  for nonzero  $g(z)$ . Now explain why  $f(z) = f(z_0) + \phi^m(z)$  where  $\phi(z_0) = 0$  but  $\phi'(z_0) \neq 0$  and  $\phi(z)$  is some analytic function. Thus from Problem 19 above, there is  $\delta$  such that  $f(\Omega) \supseteq f(z_0) + B(0, \varepsilon^m)$ . Hence  $f(\Omega)$  is open since each  $f(z_0)$  is an interior point of  $f(\Omega)$ . You only need to show that there is  $G(z)$  such that  $G(z)^m = g(z)$  and then  $\phi(z) \equiv (z - z_0)G(z)$  will work fine. When you have done this, Problem 19 will yield a proof of the open mapping theorem which says that if  $f$  is analytic on  $\Omega$  a connected open set, then  $f(\Omega)$  is either an open set or a single point. So here are some steps for doing this.
- (a) Consider  $z \rightarrow \frac{g'(z)}{g(z)}$ . It is analytic on the open ball  $V$  and so it has a primitive on  $V$ . In fact, you could take  $h(z) \equiv \int_{\gamma(z_0, z)} \frac{g'(w)}{g(w)} dw$ .
- (b) Let the primitive be  $h(z)$ . Then consider  $(g(z)e^{-h(z)})'$ . Show this equals 0. Then explain why this requires it to be constant. Explain why there is  $a + ib$  such that  $g(z) = e^{h(z)+a+ib}$ . Then use the primitive  $h(z) + a + ib$  instead of the original one. Call it  $h(z)$ . Then
- $$g(z) = e^{h(z)}$$
- You can then complete the argument by letting  $g(z)^{1/m} \equiv e^{h(z)/m}$  and
- $$G(z) \equiv (z - z_0)g(z)^{1/m}$$
- (c) Show that this theorem is certainly not true when considering functions of a real variable by considering  $f(x) = x^2$ .
21. If you have an open set  $U$  in  $\mathbb{C}$  show that for all  $z \in U$ ,  $|z| < \sup\{|w| : w \in U\}$ . In other words,  $z \rightarrow |z|$  never achieves its maximum on any open set  $U \in \mathbb{C}$ .
22. Let  $f$  be analytic on  $U$  and let  $B(z, r) \subseteq U$ . Let  $\gamma_r$  be the positively oriented boundary of  $B(z, r)$ . Explain, using the Cauchy integral formula why

$$|f(z)| \leq \max\{|f(w)| : w \in \gamma_r^*\} \equiv m_r$$

Show that if equality is achieved, then  $|f(w)|$  must be constantly equal to  $m_r$  on  $\gamma_r^*$ .

23. The maximum modulus theorem says that if  $\Omega$  is a bounded connected open set and  $f : \Omega \rightarrow \mathbb{C}$  is analytic and  $f : \overline{\Omega} \rightarrow \mathbb{C}$  is continuous, then if  $|f|$  achieves its maximum at any point of  $\Omega$  then  $f$  is equal to a constant on  $\overline{\Omega}$ . Thus  $|f|$  achieves its maximum on the boundary of  $\Omega$  in every case. **Hint:** Suppose the maximum is achieved at a point of  $\Omega$ ,  $z_0$ . Then let  $B(z_0, r) \subseteq \Omega$ . Show that if  $f$  is constant on  $B(z_0, r)$ , then it equals this constant on all of  $\Omega$  using Theorem 35.6.9. However, if it is not constant, then from the open mapping theorem of Problem 20,  $f(B(z_0, r))$  is an open set. Then use Problem 21 above to obtain a contradiction. Alternatively, use Problem 22 to verify that the set where  $|f|$  achieves its maximum is both open and closed.

24. Let  $f : \mathbb{C} \rightarrow \mathbb{C}$  be analytic with  $f'(z) \neq 0$  for all  $z$ . Say  $f(x + iy) = u(x, y) + iv(x, y)$ . Thus the mapping  $(x, y) \rightarrow \begin{pmatrix} u(x, y) \\ v(x, y) \end{pmatrix}$  is a  $C^1$  mapping of  $\mathbb{R}^2$  to  $\mathbb{R}^2$ . Show that at any point

$$\begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} \neq 0$$

Therefore, by the inverse function theorem, Theorem 26.0.3, this mapping is locally one to one. However, the function does not need to be globally one to one. Give an easy example using the complex exponential which shows this to be the case.

25. Let  $\Gamma$  be a simple closed piecewise  $C^1$  curve and let  $\{f_n\}$  be a sequence of functions which are analytic near  $U_i \cup \Gamma^*$ . Then if  $\gamma$  is a parametrization of  $\Gamma$  with  $n(\gamma, z) = 1$  for  $z \in U_i$ , then

$$f_n(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f_n(w)}{w - z} dw$$

This is by the Cauchy integral formula presented above. Suppose  $f_n$  converges uniformly on  $\Gamma^*$  to a continuous function  $f$ . Show that then, for  $z \in U_i$ , and  $f(z)$  defined as

$$f(z) \equiv \frac{1}{2\pi i} \int_{\gamma} \frac{f(w)}{w - z} dw$$

It follows that  $f_n(z) \rightarrow f(z)$  for each  $z \in U_i$  and also  $f$  is analytic on  $U_i$ . **Hint:** You might use Theorem 35.3.3. This is very different than what happens with functions of a real variable in which uniform convergence of polynomials  $p_n$  to  $f$  does not necessarily confer differentiability on  $f$ . For example, to approximate  $f$ , a continuous function having no derivatives or even a very easy function like  $f(x) = |x - (1/2)|$  for  $x \in [0, 1]$ .

26. The Schwarz lemma is as follows: Suppose  $F : B(0, 1) \rightarrow B(0, 1)$ ,  $F$  is analytic, and  $F(0) = 0$ . Then for all  $z \in B(0, 1)$ ,

$$|F(z)| \leq |z|, \quad (35.19)$$

and

$$|F'(0)| \leq 1. \quad (35.20)$$

If equality holds in 35.20 then there exists  $\lambda \in \mathbb{C}$  with  $|\lambda| = 1$  and

$$F(z) = \lambda z. \quad (35.21)$$

Prove the Schwarz lemma. **Hint:** Since  $F$  has a power series of the form  $\sum_{k=1}^{\infty} a_k z^k$ , it follows that  $F(z)/z$  equals an analytic function  $g(z)$  for all  $z \in B(0, 1)$ . By the maximum modulus theorem, Problem 23 above, applied to  $g(z)$ , if  $|z| < r < 1$ ,

$$\left| \frac{F(z)}{z} \right| \leq \max_{t \in [0, 2\pi]} \frac{|F(re^{it})|}{r} \leq \frac{1}{r}.$$

Explain why this implies

$$|g(z)| = \left| \frac{F(z)}{z} \right| \leq 1$$

Now explain why  $\lim_{z \rightarrow 0} \frac{F(z)}{z} = F'(0) = g(0)$  and so  $|F'(0)| \leq 1$ . It only remains to verify that if  $|F'(0)| = 1$ , then  $F(z)$  is just a rotation as described. If  $|F'(0)| = 1$ , then the analytic function  $g(z)$  has the property that it achieves its maximum at an interior point. Apply Problem 23 to conclude that  $g(z)$  must be a constant. Explain why this requires  $\left| \frac{F(z)}{z} \right| = 1$  for all  $z$ . Use this to conclude the proof.

27. Sketch an example of two differentiable functions defined on  $[0, 1]$  such that their product is 0 but neither function is 0. Explain why this never happens for the set of analytic functions defined on an open connected set. In other words, if you have  $fg = 0$  where  $f, g$  are analytic on  $D$  an open connected set, then either  $f = 0$  or  $g = 0$ . For those who like to classify algebraically, this says that the set of analytic functions defined on an open connected set is an integral domain. It is clear that this set of functions is a ring with the usual operations. The extra ingredient is this observation that there are no nonzero zero divisors. **Hint:** To show this, consider  $D \setminus f^{-1}(0)$  an open set. If  $f^{-1}(0) = D$ , then you are done. Otherwise, you have  $g$  is 0 on an open set. Now use Theorem 35.6.9.
28. For  $D \equiv \{z \in \mathbb{C} : |z| < 1\}$ , consider the function  $\sin\left(\frac{1}{1-z}\right)$ . Show that this function has infinitely many zeros in  $D$ . Thus there is a limit point to the set of zeros, but its limit point is not in  $D$ . It is good to keep this example in mind when considering Theorem 35.6.9.



## Chapter 36

# Isolated Singularities and Analytic Functions

### 36.1 Open Mapping Theorem for Complex Valued Functions

The open mapping theorem is for an analytic function with values in  $\mathbb{C}$ . It is even more surprising result than the theorem about the zeros of an analytic function. The following proof of this important theorem uses an interesting local representation of the analytic function.

**Theorem 36.1.1** (*Open mapping theorem*) Let  $\Omega$  be a region in  $\mathbb{C}$  and suppose  $f : \Omega \rightarrow \mathbb{C}$  is analytic. Then  $f(\Omega)$  is either a point or a region. In the case where  $f(\Omega)$  is a region, it follows that for each  $z_0 \in \Omega$ , there exists an open set  $V$  containing  $z_0$  and  $m \in \mathbb{N}$  such that for all  $z \in V$ ,

$$f(z) = f(z_0) + \phi(z)^m \quad (36.1)$$

where  $\phi : V \rightarrow B(0, \delta)$  is one to one, analytic and onto,  $\phi(z_0) = 0$ ,  $\phi'(z) \neq 0$  on  $V$  and  $\phi^{-1}$  analytic on  $B(0, \delta)$ . If  $f$  is one to one then  $m = 1$  for each  $z_0$  and  $f^{-1} : f(\Omega) \rightarrow \Omega$  is analytic.

**Proof:** Suppose  $f(\Omega)$  is not a point. Then if  $z_0 \in \Omega$  it follows there exists  $r > 0$  such that  $f(z) \neq f(z_0)$  for all  $z \in B(z_0, r) \setminus \{z_0\}$ . Otherwise,  $z_0$  would be a limit point of the set,

$$\{z \in \Omega : f(z) - f(z_0) = 0\}$$

which would imply from Theorem 35.6.9 that  $f(z) = f(z_0)$  for all  $z \in \Omega$ . Therefore, making  $r$  smaller if necessary and using the power series of  $f$ ,

$$f(z) = f(z_0) + (z - z_0)^m g(z) \stackrel{?}{=} f(z_0) + \left( (z - z_0) g(z)^{1/m} \right)^m$$

for all  $z \in B(z_0, r)$ , where  $g(z) \neq 0$  on  $B(z_0, r)$ . As implied in the above formula, one wonders if you can take the  $m^{\text{th}}$  root of  $g(z)$ .

$\frac{g'}{g}$  is an analytic function on  $B(z_0, r)$  and so by Morera's theorem, Theorem 35.6.1, it has a primitive on  $B(z_0, r)$  called  $h$ . Therefore by the product rule and the chain rule,

$$\begin{aligned} (ge^{-h})' &= g'(e^{-h}) + g(-e^{-h})h' \\ &= g'(e^{-h}) + g(-e^{-h})\frac{g'}{g} = 0 \end{aligned}$$

and so there exists a constant,  $C = e^{a+ib}$  such that on  $B(z_0, r)$ ,

$$ge^{-h} = e^{a+ib}.$$

Therefore,

$$g(z) = e^{h(z)+a+ib}$$

and so, modifying  $h$  by adding in the constant,  $a+ib$  it is still a primitive of  $g'/g$  and now  $g(z) = e^{h(z)}$  where  $h'(z) = \frac{g'(z)}{g(z)}$  on  $B(z_0, r)$ . Letting

$$\phi(z) = (z - z_0)e^{\frac{h(z)}{m}}$$

implies formula 36.1 is valid on  $B(z_0, r)$ . Now  $\phi(z_0) = 0$  but

$$\phi'(z_0) = e^{\frac{h(z_0)}{m}} \neq 0.$$

Shrinking  $r$  if necessary you can assume  $\phi'(z) \neq 0$  on  $B(z_0, r)$ . Is there an open set  $V$  contained in  $B(z_0, r)$  such that  $\phi$  maps  $V$  onto  $B(0, \delta)$  for some  $\delta > 0$ ?

Let  $\phi(z) = u(x, y) + iv(x, y)$  where  $z = x + iy$ . Consider the mapping

$$\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \begin{pmatrix} u(x, y) \\ v(x, y) \end{pmatrix}$$

where  $u, v$  are  $C^1$  because  $\phi$  is given to be analytic. The Jacobian of this map at  $(x, y) \in B(z_0, r)$  is

$$\begin{aligned} \begin{vmatrix} u_x(x, y) & u_y(x, y) \\ v_x(x, y) & v_y(x, y) \end{vmatrix} &= \begin{vmatrix} u_x(x, y) & -v_x(x, y) \\ v_x(x, y) & u_x(x, y) \end{vmatrix} \\ &= u_x(x, y)^2 + v_x(x, y)^2 = |\phi'(z)|^2 \neq 0. \end{aligned}$$

This follows from a use of the Cauchy Riemann equations. Also

$$\begin{pmatrix} u(x_0, y_0) \\ v(x_0, y_0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Therefore, by the inverse function theorem there exists an open set  $V$ , containing  $z_0$  and  $\delta > 0$  such that  $(u, v)^T$  maps  $V$  one to one onto  $B(0, \delta)$ . Thus  $\phi$  is one to one onto  $B(0, \delta)$  as claimed. Applying the same argument to other points  $z$  of  $V$  and using the fact that  $\phi'(z) \neq 0$  at these points, it follows  $\phi$  maps open sets to open sets. In other words,  $\phi^{-1}$  is continuous.



It also follows that  $\phi^m$  maps  $V$  onto  $B(0, \delta^m)$ . Indeed,

$$|\phi(z)|^m = |\phi(z)^m|.$$

Therefore, the formula 36.1 implies that  $f$  maps the open set  $V$ , containing  $z_0$  to an open set. This shows  $f(\Omega)$  is an open set because  $z_0$  was arbitrary. It is connected because  $f$  is continuous and  $\Omega$  is connected. Thus  $f(\Omega)$  is a region (open and connected). It remains to verify that  $\phi^{-1}$  is analytic on  $B(0, \delta)$ . Since  $\phi^{-1}$  is continuous,

$$\lim_{\phi(z_1) \rightarrow \phi(z)} \frac{\phi^{-1}(\phi(z_1)) - \phi^{-1}(\phi(z))}{\phi(z_1) - \phi(z)} = \lim_{z_1 \rightarrow z} \frac{z_1 - z}{\phi(z_1) - \phi(z)} = \frac{1}{\phi'(z)}.$$

Therefore,  $\phi^{-1}$  is analytic as claimed.

It only remains to verify the assertion about the case where  $f$  is one to one. If  $m > 1$ , then  $e^{\frac{2\pi i}{m}} \neq 1$  and so for  $z_1 \in V$ ,

$$e^{\frac{2\pi i}{m}} \phi(z_1) \neq \phi(z_1). \quad (36.2)$$

But  $e^{\frac{2\pi i}{m}} \phi(z_1) \in B(0, \delta)$  and so there exists  $z_2 \neq z_1$  (since  $\phi$  is one to one) such that  $\phi(z_2) = e^{\frac{2\pi i}{m}} \phi(z_1)$ . But then

$$\phi(z_2)^m = \left(e^{\frac{2\pi i}{m}} \phi(z_1)\right)^m = e^{2\pi i} \phi(z_1)^m = \phi(z_1)^m$$

implying  $f(z_2) = f(z_1)$  contradicting the assumption that  $f$  is one to one. Thus  $m = 1$  and  $f'(z) = \phi'(z) \neq 0$  on  $V$ . Since  $f$  maps open sets to open sets, it follows that  $f^{-1}$  is continuous and so

$$\begin{aligned} (f^{-1})'(f(z)) &= \lim_{f(z_1) \rightarrow f(z)} \frac{f^{-1}(f(z_1)) - f^{-1}(f(z))}{f(z_1) - f(z)} \\ &= \lim_{z_1 \rightarrow z} \frac{z_1 - z}{f(z_1) - f(z)} = \frac{1}{f'(z)}. \blacksquare \end{aligned}$$

You can dispense with the appeal to the inverse function theorem by using Problem 19 on Page 722.

One does not have to look very far to find that this sort of thing does not hold for functions mapping  $\mathbb{R}$  to  $\mathbb{R}$ . Take for example, the function  $f(x) = x^2$ . Then  $f(\mathbb{R})$  is neither a point nor a region. In fact  $f(\mathbb{R})$  fails to be open.

**Corollary 36.1.2** Suppose in the situation of Theorem 36.1.1  $m > 1$  for the local representation of  $f$  given in this theorem. Then there exists  $\delta > 0$  such that if  $w \in B(f(z_0), \delta) = f(V)$  for  $V$  an open set containing  $z_0$ , then  $f^{-1}(w)$  consists of  $m$  distinct points in  $V$ . ( $f$  is  $m$  to one on  $V$ )

**Proof:** Let  $w \in B(f(z_0), \delta)$ . Then  $w = f(\hat{z})$  where  $\hat{z} \in V$ . Thus  $f(\hat{z}) = f(z_0) + \phi(\hat{z})^m$ . Consider the  $m$  distinct numbers,  $\left\{e^{\frac{2k\pi i}{m}} \phi(\hat{z})\right\}_{k=1}^m$ . Then each of these numbers is in  $B(0, \delta)$  and so since  $\phi$  maps  $V$  one to one onto  $B(0, \delta)$ , there are  $m$  distinct numbers in  $V$ ,  $\{z_k\}_{k=1}^m$  such that  $\phi(z_k) = e^{\frac{2k\pi i}{m}} \phi(\hat{z})$ . Then

$$\begin{aligned} f(z_k) &= f(z_0) + \phi(z_k)^m = f(z_0) + \left(e^{\frac{2k\pi i}{m}} \phi(\hat{z})\right)^m \\ &= f(z_0) + e^{2k\pi i} \phi(\hat{z})^m = f(z_0) + \phi(\hat{z})^m = f(\hat{z}) = w \blacksquare \end{aligned}$$

**Example 36.1.3** Consider the open connected set  $D \equiv \mathbb{R} + i(a - \pi, a + \pi)$ . Then  $z \rightarrow e^z$  is one to one and analytic on  $D$ . It maps  $D$  onto  $\mathbb{C} \setminus l$  where  $l$  is the ray starting from 0 whose angle is  $a$ . Therefore, it has an analytic inverse defined on  $\mathbb{C} \setminus l$ . This is a branch of the logarithm. It is of the form

$$\log(z) = \ln|z| + i \arg_a(z)$$

where  $\arg_a(z)$  is the angle in  $(a - \pi, a + \pi)$  with the property that

$$e^{\ln|z| + i \arg_a(z)} = z$$

We usually let  $a = 0$  and then the inverse is what is usually called the logarithm and is denoted by  $\log$ . As in Problem 10 this is  $\ln(|z|) + i \arg(z)$  where  $\arg(z)$  is the angle between  $-\pi$  and  $\pi$  corresponding to  $z \in \mathbb{C} \setminus (-\infty, 0]$ .

With the open mapping theorem, the maximum modulus theorem is fairly easy.

**Theorem 36.1.4** Let  $\Omega$  be an open connected, bounded set in  $\mathbb{C}$  and let  $f : \Omega \rightarrow \mathbb{C}$  be analytic. Let  $\partial\Omega \equiv \bar{\Omega} \setminus \Omega$ . Then

$$\max\{|f(z)| : z \in \bar{\Omega}\} = \max\{|f(z)| : z \in \partial\Omega\}$$

and if the maximum of  $|f(z)|$  is achieved at a point of  $\Omega$ , then  $f$  is a constant.

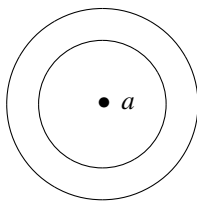
**Proof:** Suppose  $f(\Omega)$  is not a single point. That is,  $f$  is not constant. Then by the open mapping theorem,  $f(\Omega)$  is an open connected subset of  $\mathbb{C}$  and so  $z \rightarrow |f(z)|$  has no maximum. Therefore, the maximum of  $|f(z)|$  for  $z \in \bar{\Omega}$  is on  $\partial\Omega$ . If  $f(\Omega)$  is a single point, then the equation still holds. ■

## 36.2 Functions Analytic on an Annulus

First consider the definition of an annulus.

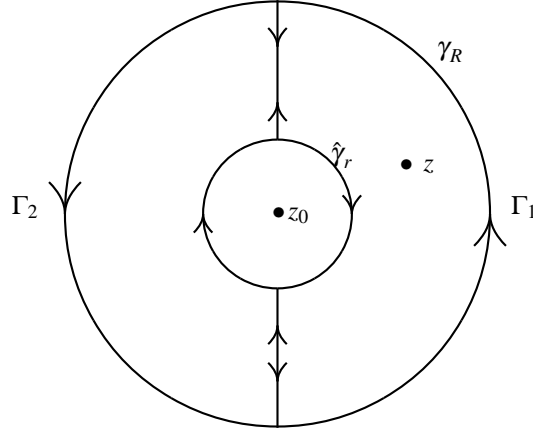
**Definition 36.2.1** Define  $\text{ann}(a, r, R) \equiv \{z : r < |z - a| < R\}$ .

Thus  $\text{ann}(a, 0, R)$  would denote the punctured ball,  $B(a, R) \setminus \{a\}$  and when  $r > 0$ , the annulus looks like the following.



The annulus consists of the points between the two circles.

In the following picture, let there be two parametrizations  $\gamma_R$  for the large circle and  $\gamma_r$  for the small one with orientation as shown. There are also two line segments oriented as shown which miss  $z \in \text{ann}(z_0, r, R)$  and constitute the intersection of the two simple closed curves  $\Gamma_1, \Gamma_2$ . These two simple closed curves are oriented as shown. Thus each of  $\Gamma_i$  is positively oriented. Let  $f$  be analytic near  $\overline{\text{ann}(z_0, r, R)}$ .



It follows from Theorem 35.7.1, that for  $z$  in the annulus,

$$\frac{1}{2\pi i} \int_{\gamma_R} \frac{f(w)}{w-z} dw + \frac{1}{2\pi i} \int_{\hat{\gamma}_r} \frac{f(w)}{w-z} dw = f(z)$$

This is because the contributions to the line integrals along those straight lines is 0 since they cancel off because of opposite orientations. Let  $\gamma_r$  be the opposite orientation from  $\hat{\gamma}_r$ . Then this reduces to

$$\int_{\gamma_R} \frac{f(w)}{w-z} dw - \int_{\gamma_r} \frac{f(w)}{w-z} dw = 2\pi i f(z)$$

Thus

$$\begin{aligned} f(z) &= \frac{1}{2\pi i} \left[ \int_{\gamma_R} \frac{f(w)}{w-z_0-(z-z_0)} dw + \int_{\gamma_r} \frac{f(w)}{(z-z_0)-(w-z_0)} dw \right] \\ &= \frac{1}{2\pi i} \left[ \int_{\gamma_R} \frac{1}{w-z_0} \frac{f(w)}{1-\frac{z-z_0}{w-z_0}} dw + \int_{\gamma_r} \frac{1}{z-z_0} \frac{f(w)}{1-\frac{w-z_0}{z-z_0}} dw \right] \end{aligned}$$

Now note that for  $z$  in the annulus between the two circles and  $w \in \gamma_R^*$ ,  $\left| \frac{z-z_0}{w-z_0} \right| < 1$ , and for  $w \in \gamma_r^*$ ,  $\left| \frac{w-z_0}{z-z_0} \right| < 1$ . In fact, in each case, there is  $b < 1$  such that

$$w \in \gamma_R^*, \left| \frac{z-z_0}{w-z_0} \right| < b < 1, \quad w \in \gamma_r^*, \left| \frac{w-z_0}{z-z_0} \right| < b < 1 \quad (36.3)$$

Thus you can use the formula for the sum of an infinite geometric series and conclude

$$f(z) = \frac{1}{2\pi i} \left[ \int_{\gamma_R} f(w) \frac{1}{w-z_0} \sum_{n=0}^{\infty} \left( \frac{z-z_0}{w-z_0} \right)^n dw + \int_{\gamma_r} f(w) \frac{1}{(z-z_0)} \sum_{n=0}^{\infty} \left( \frac{w-z_0}{z-z_0} \right)^n dw \right]$$

Then from the uniform estimates of 36.3, one can conclude uniform convergence of the partial sums for  $w \in \gamma_R^*$  or  $\gamma_r^*$ , and so by the Weierstrass M test, Theorem 13.8.3, one can

interchange the summation with the integral and write

$$\begin{aligned} f(z) &= \sum_{n=0}^{\infty} \left( \frac{1}{2\pi i} \int_{\gamma_R} f(w) \frac{1}{(w-z_0)^{n+1}} dw \right) (z-z_0)^n \\ &\quad + \sum_{n=0}^{\infty} \left( \frac{1}{2\pi i} \int_{\gamma_r} f(w) (w-z_0)^n dw \right) \frac{1}{(z-z_0)^{n+1}}, \end{aligned}$$

both series converging absolutely. Thus there are  $a_n, b_n \in X$  such that

$$f(z) = \sum_{n=0}^{\infty} a_n (z-z_0)^n + \sum_{n=1}^{\infty} b_n (z-z_0)^{-n}$$

This proves most of the following theorem.

**Theorem 36.2.2** *Let  $z \in \text{ann}(z_0, r, R)$  and let  $f : \text{ann}(z_0, r, R) \rightarrow X$  be analytic near  $\overline{\text{ann}(z_0, r, R)}$ . Then for any  $z \in \text{ann}(z_0, r, R)$ ,*

$$f(z) = \sum_{n=0}^{\infty} a_n (z-z_0)^n + \sum_{n=1}^{\infty} b_n (z-z_0)^{-n} \quad (36.4)$$

where

$$\begin{aligned} a_n &= \frac{1}{2\pi i} \int_{\gamma_R} f(w) \frac{1}{(w-z_0)^{n+1}} dw \\ b_n &= \frac{1}{2\pi i} \int_{\gamma_r} f(w) (w-z_0)^{n-1} dw \end{aligned}$$

and both of these series in 36.4 converge absolutely. If  $r < \hat{r} < \hat{R} < R$ , then convergence of both series is absolute and uniform for  $z \in \text{ann}(z_0, \hat{r}, \hat{R})$ .

**Proof:** Consider the sum with the negative exponents. The other is similar. Let  $|f(w)| \leq M$  on the closure of the annulus.

$$\sum_{n=1}^{\infty} b_n (z-z_0)^{-n}, \quad b_n = \left( \frac{1}{2\pi i} \int_{\gamma_r} f(w) (w-z_0)^n dw \right)$$

Therefore,  $\|b_n\| \leq 2\pi r M r^n$  and  $|z-z_0| \geq \hat{r} > r$ . Thus

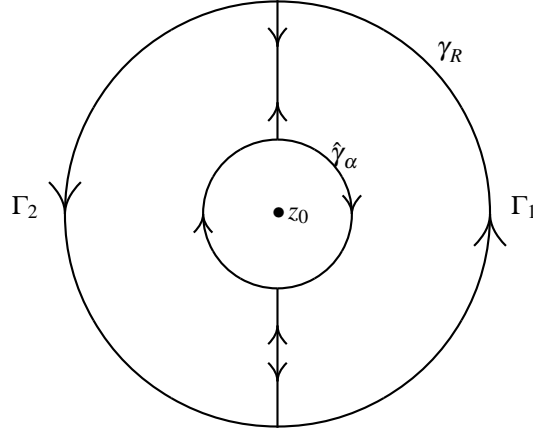
$$\sum_{n=p}^q \|b_n\| |z-z_0|^{-n} \leq \sum_{n=p}^q 2\pi \hat{r} M \frac{r^n}{\hat{r}^n} < \varepsilon$$

if  $p$  is large enough. Therefore, the partial sums are a uniformly Cauchy sequence so the sum converges absolutely and uniformly on the set  $\{z : \hat{r} \leq |z-z_0| \leq \hat{R}\}$ . ■

Note that for arbitrary  $\alpha$  with  $r \leq \alpha \leq R$ ,

$$\frac{1}{2\pi i} \int_{\gamma_R} f(w) \frac{1}{(w-z_0)^{n+1}} dw = \frac{1}{2\pi i} \int_{\gamma_\alpha} f(w) \frac{1}{(w-z_0)^{n+1}} dw \quad (36.5)$$

This is a simple application of the Cauchy integral theorem applied to the union of two simple closed curves of the sort used to prove Theorem 36.2.2. You consider the annulus  $\text{ann}(z_0, \alpha, R)$  and the following diagram.



The integrand is analytic on the inside of the two simple closed curves  $\Gamma_1$  and  $\Gamma_2$ . Letting  $\gamma_1$  and  $\gamma_2$  be oriented parametrizations for these and using the argument that the integrals over the straight lines cancel, this yields

$$\frac{1}{2\pi i} \int_{\gamma_R} f(w) \frac{1}{(w-z_0)^{n+1}} dw + \frac{1}{2\pi i} \int_{\gamma_\alpha} f(w) \frac{1}{(w-z_0)^{n+1}} dw = 0$$

and so, letting  $\gamma_\alpha \equiv -\hat{\gamma}_\alpha$ , yields formula 36.5.

Similar considerations apply to  $b_n$ .

**Corollary 36.2.3** *The  $a_n$  and  $b_n$  are uniquely determined.*

**Proof:** Let  $\alpha \in (r, R)$  and let  $\gamma_\alpha$  be a parametrization of the circle centered at  $z_0$  of radius  $\alpha$  which is counterclockwise. We have

$$f(w) = \sum_{n=0}^{\infty} a_n (w-z_0)^n + \sum_{n=1}^{\infty} b_n (w-z_0)^{-n}$$

for  $w$  in the annulus. Then for  $k \geq 1$ ,

$$f(w) (w-z_0)^{k-1} = \sum_{n=0}^{\infty} a_n (w-z_0)^{n+k-1} + \sum_{n=1}^{\infty} b_n (w-z_0)^{-n+k-1}$$

By uniform convergence,

$$\begin{aligned} \int_{\gamma_\alpha} f(w) (w-z_0)^{k-1} dw &= \sum_{n=0}^{\infty} a_n \int_{\gamma_\alpha} (w-z_0)^{n+k-1} dw \\ &\quad + \sum_{n=1}^{\infty} b_n \int_{\gamma_\alpha} (w-z_0)^{-n+k-1} dw \end{aligned}$$

Now in the sums, all integrals are 0 except the one when  $n = k$  in the second sum. Therefore,

$$\int_{\gamma_\alpha} f(w) (w-z_0)^{k-1} dw = b_k \int_{\gamma_\alpha} (w-z_0)^{-1} dw = 2\pi i b_k$$

This shows that for any  $\alpha, r < \alpha < R$ ,

$$b_k = \frac{1}{2\pi i} \int_{\gamma_\alpha} f(w) (w - z_0)^{k-1} dw$$

Similar reasoning gives

$$a_n = \frac{1}{2\pi i} \int_{\gamma_\alpha} f(w) \frac{1}{(w - z_0)^{n+1}} dw$$

and as explained above, nothing changes when  $\alpha$  is changed. ■

**Definition 36.2.4** For  $f$  near the closure of an annulus as just described, it follows that on the annulus,  $f$  can be written as the sum of a power series and a series involving  $(z - z_0)$  raised to negative powers. This is called the *Laurent series*. The series involving negative powers of  $(z - z_0)$  is called the *principal part of the Laurent series*.

Note that if  $f$  is analytic near  $z_0$ , but possibly not at  $z_0$  then the  $r$  in  $\gamma_r$  can be taken as small as desired.

### 36.3 Isolated Singularities

This is about the situation where the Laurent series of  $f$  has nonzero principal part. When this occurs, we say that  $z_0$  is a singularity. The singularities are isolated if each is the center of a ball such that  $f$  is analytic except for the center of the ball.

**Definition 36.3.1** Let  $B'(a, r) \equiv \{z \in \mathbb{C} \text{ such that } 0 < |z - a| < r\}$ . Thus this is the usual ball without the center. A function is said to have an *isolated singularity* at the point  $a \in \mathbb{C}$  if  $f$  is analytic on  $B'(a, r)$  for some  $r > 0$ .

It turns out isolated singularities can be neatly classified into three types, removable singularities, poles, and essential singularities. The next theorem deals with the case of a removable singularity.

**Definition 36.3.2** An isolated singularity of  $f$  is said to be *removable* if there exists an analytic function  $g$  analytic at  $a$  and near  $a$  such that  $f = g$  at all points near  $a$ .

**Theorem 36.3.3** Let  $f : B'(a, r) \rightarrow X$  be analytic. Thus  $f$  has an isolated singularity at  $a$ . Then  $a$  is a removable singularity if and only if

$$\lim_{z \rightarrow a} f(z)(z - a) = 0.$$

Thus the above limit occurs if and only if there exists a unique analytic function,  $g : B(a, r) \rightarrow X$  such that  $g = f$  on  $B'(a, r)$ . In other words, you can re define  $f$  at  $a$  so that the resulting function is analytic.

**Proof:**  $\Rightarrow$  Let  $h(z) \equiv (z - a)^2 f(z)$ ,  $h(a) \equiv 0$ . Then  $h$  is analytic on  $B(a, r)$  because it is easy to see that  $h'(a) = 0$ . It follows  $h$  is given by a power series,

$$h(z) = \sum_{k=2}^{\infty} a_k (z - a)^k$$

where  $a_0 = a_1 = 0$  because of the observation above that  $h'(a) = h(a) = 0$ . It follows that for  $|z - a| > 0$

$$f(z) = \sum_{k=2}^{\infty} a_k (z-a)^{k-2} \equiv g(z).$$

⇐The converse is obvious. ■

What of the other case where the singularity is not removable? This situation is dealt with by the amazing Casorati Weierstrass theorem.

**Theorem 36.3.4** (Casorati Weierstrass) *Let  $a$  be an isolated singularity and suppose for some  $r > 0$ ,  $f(B'(a, r))$  is not dense in  $\mathbb{C}$ . Then either  $a$  is a removable singularity or there exist finitely many  $b_1, \dots, b_M$  for some finite number,  $M$  such that for  $z$  near  $a$ ,*

$$f(z) = g(z) + \sum_{k=1}^M \frac{b_k}{(z-a)^k} \quad (36.6)$$

where  $g(z)$  is analytic near  $a$ .

Such an  $a$  satisfying 36.6 is called a pole.

**Proof:** Suppose  $B(z_0, \delta)$  has no points of  $f(B'(a, r))$ . Such a ball must exist if  $f(B'(a, r))$  is not dense. Then for  $z \in B'(a, r)$ ,  $|f(z) - z_0| \geq \delta > 0$ . It follows from Theorem 36.3.3 that  $\frac{1}{f(z) - z_0}$  has a removable singularity at  $a$ . Hence, there exists  $h$  an analytic function such that for  $z$  near  $a$ ,

$$h(z) = \frac{1}{f(z) - z_0}. \quad (36.7)$$

There are two cases. First suppose  $h(a) = 0$ . Then  $\sum_{k=1}^{\infty} a_k (z-a)^k = \frac{1}{f(z) - z_0}$  for  $z$  near  $a$ . If all the  $a_k = 0$ , this would be a contradiction because then the left side would equal zero for  $z$  near  $a$  but the right side could not equal zero. Therefore, there is a first  $m$  such that  $a_m \neq 0$ . Hence there exists an analytic function,  $k(z)$  which is not equal to zero in some ball,  $B(a, \varepsilon)$  such that

$$k(z)(z-a)^m = \frac{1}{f(z) - z_0}.$$

Hence, taking both sides to the  $-1$  power,

$$f(z) - z_0 = \frac{1}{(z-a)^m} \sum_{k=0}^{\infty} b_k (z-a)^k$$

and so 36.6 holds.

The other case is that  $h(a) \neq 0$ . In this case, raise both sides of 36.7 to the  $-1$  power and obtain

$$f(z) - z_0 = h(z)^{-1},$$

a function analytic near  $a$ . Therefore, the singularity is removable. ■

This theorem is the basis for the following definition which describes isolated singularities.

## 36.4 Meromorphic Functions

In short, meromorphic functions have only isolated singularities and the singularities are either poles or removable. Thus this collection of functions includes the analytic functions. Analytic functions are all like polynomials. Meromorphic functions are all like rational functions. This observation can be made much more precise but this is roughly the idea. In fact, functions meromorphic on the Riemann sphere are rational functions but this is not developed in this book. There is so much available in complex analysis that I don't wish to try and include it all.

**Definition 36.4.1** *Let  $a$  be an isolated singularity of  $f$ . When 36.6 holds for  $z$  near  $a$ , then  $a$  is called a pole. The order of the pole in 36.6 is  $M$ . Essential singularities are those which have infinitely many nonzero terms in the principal part of the Laurent series. When a function  $f$  is analytic except for isolated singularities and the isolated singularities are all poles, and there are finitely many of these poles in every compact set, the function is called meromorphic.*

Actually, if you insist only that the singularities are isolated and poles, then you can prove that there are finitely many in any compact set so part of the above definition is actually redundant as shown in the following lemma.

**Lemma 36.4.2** *If  $f$  has a pole at  $a$ , then  $\lim_{z \rightarrow a} |f(z)| = \infty$ . Also if  $f \in \mathcal{M}(\Omega)$  for  $\Omega$  an open set, then the poles cannot have a limit point in  $\Omega$  and there are finitely many poles in every  $B(0, R)$ . For  $f \in \mathcal{M}(\Omega)$ ,  $\alpha$  is a pole if and only if  $\lim_{z \rightarrow \alpha} |f(z)| = \infty$ . Also  $\alpha$  is a zero if and only if  $\lim_{z \rightarrow \alpha} |f(z)| = 0$ .*

**Proof:** We know by definition that

$$f(z) = g(z) + \sum_{k=1}^n \frac{b_k}{(z-a)^k}$$

where  $b_n \neq 0$ . Thus by the triangle inequality,

$$\begin{aligned} |f(z)| &\geq \frac{|b_n|}{|z-a|^n} - \left( |g(z)| + \sum_{k=1}^{n-1} \frac{|b_k|}{|z-a|^k} \right) \\ &= \frac{|b_n|}{|z-a|^n} \left( 1 - \left( |g(z)| |z-a|^n / |b_n| + \frac{1}{|b_n|} \sum_{k=1}^{n-1} |b_k| |z-a|^{n-k} \right) \right) \\ &\geq \frac{|b_n|}{|z-a|^n} \frac{1}{2} \end{aligned}$$

for  $|z-a|$  small enough. Thus the claim is verified.

Consider the second claim. Suppose  $\alpha_m$  is a pole and  $\lim_{m \rightarrow \infty} \alpha_m = \alpha \in \Omega$  and  $f$  is analytic at  $\alpha$ . Then from the first part, there exists  $\beta_m$  such that  $|\alpha_m - \beta_m| < 1/m$  but  $|f(\beta_m)| > m$ . Then  $\beta_m \rightarrow \alpha$  but  $\lim_{m \rightarrow \infty} |f(\beta_m)|$  does not exist. In particular,  $|f(\beta_m)|$  fails to converge to  $f(\alpha)$  showing that  $f$  cannot be analytic at  $\alpha$ . Thus it must be the case that  $\alpha$  is a pole. But now this is not allowed either because at a pole the function is analytic on a deleted ball centered at the pole and it is assumed here that  $\lim_{m \rightarrow \infty} \alpha_m = \alpha$ . Thus there are finitely many poles in every compact subset of  $\Omega$  including  $B(0, R)$ .



Finally, consider the last claim. It is obvious that  $\alpha$  is a zero if and only if  $\lim_{z \rightarrow \alpha} f(z) = 0$ . It was shown above that at poles  $\lim_{z \rightarrow \alpha} |f(z)| = \infty$ . Then suppose the limit condition holds. Why is  $\alpha$  a pole? This happens because of the Casorati Weierstrass theorem, Theorem 36.3.4. Every singularity is isolated for a meromorphic function by definition. Thus there is a Laurent expansion for  $f$  near  $\alpha$ . If the principal part is an infinite series, then by this theorem, the values of  $f$  near  $\alpha$  are dense in  $\mathbb{C}$  and so  $\lim_{z \rightarrow \alpha} |f(z)|$  does not even exist. Therefore, this principal part must be a finite sum and so  $\alpha$  is a pole. ■

What follows is the definition of something called a residue. This pertains to a singularity which has a pole at an isolated singularity.

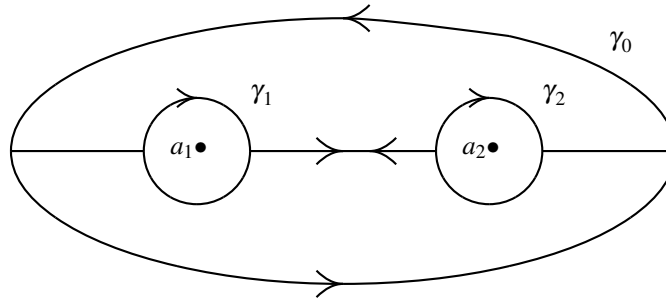
**Definition 36.4.3** *The residue of  $f$  at an isolated singularity  $\alpha$  which is a pole, written  $\text{res}(f, \alpha)$  is the coefficient of  $(z - \alpha)^{-1}$  where*

$$f(z) = g(z) + \sum_{k=1}^m \frac{b_k}{(z - \alpha)^k}.$$

Thus  $\text{res}(f, \alpha) = b_1$  in the above.

## 36.5 The Residue Theorem

We have in mind finitely many poles enclosed by a simple closed, piecewise  $C^1$  curve as in the following picture which shows the case of two poles.



You have a simple closed curve, positively oriented. Say  $\gamma$  is a parametrization for this curve. Then inside there are finitely many singularities  $\{a_k\}_{k=1}^n$ . Enclose each with a circle oriented in the clockwise direction, parameterized by  $\hat{\gamma}_k$  and connect them with straight lines as shown. Then you have two simple closed curves which intersect in these finitely many straight line segments. Orient them oppositely so that line integrals over the straight line segments cancel and each of the two simple closed curves is oriented positively. Then if  $f$  is analytic except at the points shown, the Cauchy integral theorem implies

$$\int_{\gamma} f(z) dz + \sum_{k=1}^n \int_{\hat{\gamma}_k} f(z) dz = 0$$

Letting  $\gamma_k \equiv -\hat{\gamma}_k$ ,

$$\int_{\gamma} f(z) dz = \sum_{k=1}^n \int_{\gamma_k} f(z) dz \quad (36.8)$$

Now on the inside of  $\gamma_k$ ,

$$f(z) = g_k(z) + \sum_{n=1}^{M_k} \frac{b_n}{(z-a_k)^n} \quad (36.9)$$

Thus  $\int_{\gamma_1} f(z) dz = b_1$  because all the other terms have primitives. Indeed, if  $n \neq -1$ ,  $(z-a)^n$  has  $\frac{(z-a)^{n+1}}{n+1}$  as a primitive. However,

$$\int_{\gamma_k} \frac{b_1}{z-a_k} dz = 2\pi i b_1$$

**Definition 36.5.1** Suppose  $f(z) = g(z) + \sum_{n=1}^M \frac{b_n}{(z-a)^n}$  for  $z$  near  $a$ . Then  $\text{res}(f, a) \equiv b_1$ .

Using this notation, by analogy to the above,  $\int_{\gamma_k} f(z) dz = 2\pi i \text{res}(f, a_k)$ . Then from 36.8,

$$\int_{\gamma} f(z) dz = 2\pi i \sum_{k=1}^n \text{res}(f, a_k)$$

In words, the contour integral is  $2\pi i$  times the sum of the residues.

So is there a way to find the residues? The answer is yes.

**PROCEDURE 36.5.2** Say you want to find  $\text{res}(f, a) = b_1$  in

$$f(z) = g(z) + \sum_{n=1}^M \frac{b_n}{(z-a)^n}, \quad g \text{ analytic}$$

This is the case where you have a pole of order  $M$  at  $a$ . You would multiply by  $(z-a)^M$ . This would give

$$f(z)(z-a)^M = g(z)(z-a)^M + \sum_{n=1}^M b_n (z-a)^{M-n}$$

Then you would take  $M-1$  derivatives and then take the limit as  $z \rightarrow a$ . This would give  $(M-1)!b_1$ .

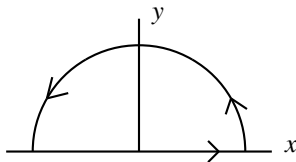
You can see from the formula that this will work and so there is no question that the limit exists. Because of this, you could use L'Hospital's rule to formally find this limit. This rule pertains only to real functions of a real variable so it is somewhat unjustified to use it. However, since you know the limit exists in this case, you can pick a one dimensional direction and apply L'Hospital to the real and imaginary parts to identify the limit which is typically what needs to be done. It is a nice illustration of the difference between real analysis which is characterized by pathology and complex analysis which is much more agreeable. Difficult mathematical questions about whether something exists are often less the issue in complex analysis.

## 36.6 Evaluation of Improper Integrals

You can use the above method of residues to evaluate obnoxious integrals of the form

$$\int_{-\infty}^{\infty} \frac{p(x)}{q(x)} dx \equiv \lim_{R \rightarrow \infty} \int_{-R}^R \frac{p(x)}{q(x)} dx$$

provided the degree of  $p(x)$  is two less than the degree of  $q(x)$ . This can be done by using the contour  $\gamma_R$  which goes from  $(-R, 0)$  to  $(R, 0)$  along the real line and then on the semicircle of radius  $R$  from  $(R, 0)$  to  $(-R, 0)$ .



Letting  $C_R$  be the circular part of this contour, for large  $R$ ,

$$\left| \int_{C_R} \frac{p(z)}{q(z)} dz \right| \leq \pi R \frac{CR^k}{R^{k+2}}$$

which converges to 0 as  $R \rightarrow \infty$ . Therefore, it is only a matter of taking large enough  $R$  to enclose all the roots of  $q(z)$  which are in the upper half plane, finding the residues at these points and then computing the contour integral. Then you would let  $R \rightarrow \infty$  and the part of the contour on the semicircle will disappear leaving the Cauchy principal value integral which is desired. There are other situations which will work just as well. You simply need to have the case where the integral over the curved part of the contour converges to 0 as  $R \rightarrow \infty$ .

Here is an easy example.

**Example 36.6.1** Find  $\int_{-\infty}^{\infty} \frac{1}{x^2+1} dx$

You know from calculus that the answer is  $\pi$ . Lets use the method of residues to find this. The function  $\frac{1}{z^2+1}$  has poles at  $i$  and  $-i$ . We don't need to consider  $-i$ . It seems clear that the pole at  $i$  is of order 1 and so all we have to do is take

$$\lim_{z \rightarrow i} \frac{z-i}{1+z^2} = \frac{1}{(x-i)(x+i)} (x-i) = \frac{1}{2i}$$

Then the integral equals  $2\pi i \left(\frac{1}{2i}\right) = \pi$ .

That one is easy. Now here is a genuinely obnoxious integral.

**Example 36.6.2** Find  $\int_{-\infty}^{\infty} \frac{1}{1+x^4} dx$

It will have poles at the roots of  $1+x^4$ . These are

$$\left(\frac{1}{2} - \frac{1}{2}i\right)\sqrt{2}, -\left(\frac{1}{2} + \frac{1}{2}i\right)\sqrt{2}, -\left(\frac{1}{2} - \frac{1}{2}i\right)\sqrt{2}, \left(\frac{1}{2} + \frac{1}{2}i\right)\sqrt{2}$$

Using the above contour, we only need consider

$$-\left(\frac{1}{2} - \frac{1}{2}i\right)\sqrt{2}, \left(\frac{1}{2} + \frac{1}{2}i\right)\sqrt{2}$$

Since they are all distinct, the poles at these two will be of order 1. To find the residues at these points, you would need to have

$$\lim_{z \rightarrow -\left(\frac{1}{2} + \frac{1}{2}i\right)\sqrt{2}} \frac{\left(z - \left(-\left(\frac{1}{2} - \frac{1}{2}i\right)\sqrt{2}\right)\right)}{1+z^4}$$

factoring  $1+x^4$  and computing the limit, you could get the answer. Applying L'Hospital's rule to identify the limit you know is there,

$$\lim_{z \rightarrow -(\frac{1}{2} + \frac{1}{2}i)\sqrt{2}} \frac{1}{4z^3} = \left(\frac{1}{8} - \frac{1}{8}i\right)\sqrt{2}$$

Similarly, the residue at  $(\frac{1}{2} + \frac{1}{2}i)\sqrt{2}$  is

$$-\left(\frac{1}{8} + \frac{1}{8}i\right)\sqrt{2}$$

Then the contour integral is

$$2\pi i \left( \left(\frac{1}{8} - \frac{1}{8}i\right)\sqrt{2} \right) + 2\pi i \left( -\left(\frac{1}{8} + \frac{1}{8}i\right)\sqrt{2} \right) = \frac{1}{2}\sqrt{2}\pi$$

You might observe that this is a lot easier than doing the usual partial fractions and trig substitutions etc. Now here is another tedious example.

**Example 36.6.3** Find  $\int_{-\infty}^{\infty} \frac{x+2}{(x^2+1)(x^2+4)^2} dx$

The poles of interest are located at  $i, 2i$ . The pole at  $2i$  is of order 2 and the one at  $i$  is of order 1. In this case, the partial fractions expansion is

$$\frac{\frac{1}{9}x + \frac{2}{9}}{x^2 + 1} - \frac{\frac{1}{3}x + \frac{2}{3}}{(x^2 + 4)^2} - \frac{\frac{1}{9}x + \frac{2}{9}}{x^2 + 4}$$

The pole at  $i$  would be

$$\lim_{z \rightarrow i} \frac{\left(\frac{1}{9}z + \frac{2}{9}\right)(z-i)}{(z+i)(z-i)} = \frac{\left(\frac{1}{9}i + \frac{2}{9}\right)}{(i+i)} = \frac{1}{18} - \frac{1}{9}i$$

Now consider the pole at  $2i$  by consideration of the next two terms in the partial fractions expansion. You must multiply it by  $(x-2i)^2$ , take the derivative and then take a limit as  $x \rightarrow 2i$ . Multiplying and taking the derivative yields

$$D_x \left( \frac{\frac{1}{3}x + \frac{2}{3}}{(x+2i)^2} \right) = -\frac{1}{3(x+2i)^3} (x+4-2i)$$

Then you have to take a limit as  $x \rightarrow 2i$  which is

$$-\frac{1}{48}i$$

Finally, consider the last term which has a pole of order 1.

$$\lim_{x \rightarrow 2i} \frac{\left(\frac{1}{9}x + \frac{2}{9}\right)(x-2i)}{(x-2i)(x+2i)} = \frac{1}{18} - \frac{1}{18}i$$

Then adding in the minus sign, we have the following for the integral.

$$2\pi i \left( \frac{1}{18} - \frac{1}{9}i \right) + 2\pi i \left( -\left( \frac{1}{18} - \frac{1}{18}i \right) \right) - 2\pi i \left( -\frac{1}{48}i \right) = \frac{5}{72}\pi$$

Sometimes you don't blow up the curves and take limits. Sometimes the problem of interest reduces directly to a complex integral over a closed curve. Here is an example of this.

**Example 36.6.4** *The integral is*

$$\int_0^\pi \frac{\cos \theta}{2 + \cos \theta} d\theta$$

This integrand is even and so it equals

$$\frac{1}{2} \int_{-\pi}^\pi \frac{\cos \theta}{2 + \cos \theta} d\theta.$$

For  $z$  on the unit circle,  $z = e^{i\theta}$ ,  $\bar{z} = \frac{1}{z}$  and therefore,  $\cos \theta = \frac{1}{2} \left( z + \frac{1}{z} \right)$ . Thus  $dz = ie^{i\theta} d\theta$  and so  $d\theta = \frac{dz}{iz}$ . Note that this is done in order to get a complex integral which reduces to the one of interest. It follows that a complex integral which reduces to the integral of interest is

$$\frac{1}{2i} \int_\gamma \frac{\frac{1}{2} \left( z + \frac{1}{z} \right)}{2 + \frac{1}{2} \left( z + \frac{1}{z} \right)} \frac{dz}{z} = \frac{1}{2i} \int_\gamma \frac{z^2 + 1}{z(4z + z^2 + 1)} dz$$

where  $\gamma$  is the unit circle oriented counter clockwise. Now the integrand has poles of order 1 at those points where  $z(4z + z^2 + 1) = 0$ . These points are

$$0, -2 + \sqrt{3}, -2 - \sqrt{3}.$$

Only the first two are inside the unit circle. It is also clear the function has simple poles at these points. Therefore,

$$\text{res}(f, 0) = \lim_{z \rightarrow 0} z \left( \frac{z^2 + 1}{z(4z + z^2 + 1)} \right) = 1.$$

$$\text{res}(f, -2 + \sqrt{3}) =$$

$$\lim_{z \rightarrow -2 + \sqrt{3}} \left( z - (-2 + \sqrt{3}) \right) \frac{z^2 + 1}{z(4z + z^2 + 1)} = -\frac{2}{3}\sqrt{3}.$$

It follows

$$\begin{aligned} \int_0^\pi \frac{\cos \theta}{2 + \cos \theta} d\theta &= \frac{1}{2i} \int_\gamma \frac{z^2 + 1}{z(4z + z^2 + 1)} dz \\ &= \frac{1}{2i} 2\pi i \left( 1 - \frac{2}{3}\sqrt{3} \right) \\ &= \pi \left( 1 - \frac{2}{3}\sqrt{3} \right). \end{aligned}$$

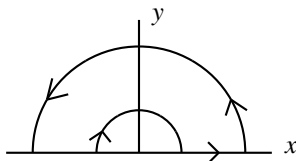
Other rational functions of the trig functions will work out by this method also.

Sometimes we have to be clever about which version of an analytic function that reduces to a real function we should use. The following is such an example.

**Example 36.6.5** *The integral here is*

$$\int_0^\infty \frac{\ln x}{1 + x^4} dx.$$

It is natural to try and use the contour in the following picture in which the small circle has radius  $r$  and the large one has radius  $R$ .



However, this will create problems with the log since the usual version of the log is not defined on the negative real axis. This difficulty may be eliminated by simply using another branch of the logarithm as in Example 36.1.3. Leave out the ray from 0 along the negative y axis and use this example to define  $L(z)$  on this set. Thus  $L(z) = \ln|z| + i\arg_1(z)$  where  $\arg_1(z)$  will be the angle  $\theta$ , between  $-\frac{\pi}{2}$  and  $\frac{3\pi}{2}$  such that  $z = |z|e^{i\theta}$ . Then the function used is  $f(z) \equiv \frac{L(z)}{1+z^4}$ . Now the only singularities contained in this contour are

$$\frac{1}{2}\sqrt{2} + \frac{1}{2}i\sqrt{2}, -\frac{1}{2}\sqrt{2} + \frac{1}{2}i\sqrt{2}$$

and the integrand  $f$  has simple poles at these points. Thus  $\text{res}\left(f, \frac{1}{2}\sqrt{2} + \frac{1}{2}i\sqrt{2}\right) =$

$$\begin{aligned} & \lim_{z \rightarrow \frac{1}{2}\sqrt{2} + \frac{1}{2}i\sqrt{2}} \frac{\left(z - \left(\frac{1}{2}\sqrt{2} + \frac{1}{2}i\sqrt{2}\right)\right) (\ln|z| + i\arg_1(z))}{1+z^4} \\ &= \lim_{z \rightarrow \frac{1}{2}\sqrt{2} + \frac{1}{2}i\sqrt{2}} \frac{(\ln|z| + i\arg_1(z)) + \left(z - \left(\frac{1}{2}\sqrt{2} + \frac{1}{2}i\sqrt{2}\right)\right) (1/z)}{4z^3} \\ &= \frac{\ln\left(\sqrt{\frac{1}{2} + \frac{1}{2}}\right) + i\frac{\pi}{4}}{4\left(\frac{1}{2}\sqrt{2} + \frac{1}{2}i\sqrt{2}\right)^3} = \left(\frac{1}{32} - \frac{1}{32}i\right)\sqrt{2}\pi \end{aligned}$$

Similarly

$$\begin{aligned} & \text{res}\left(f, -\frac{1}{2}\sqrt{2} + \frac{1}{2}i\sqrt{2}\right) = \\ & \frac{3}{32}\sqrt{2}\pi + \frac{3}{32}i\sqrt{2}\pi. \end{aligned}$$

Of course it is necessary to consider the integral along the small semicircle of radius  $r$ . This reduces to

$$\int_{\pi}^0 \frac{\ln|r| + it}{1 + (re^{it})^4} (rie^{it}) dt$$

which clearly converges to zero as  $r \rightarrow 0$  because  $r \ln r \rightarrow 0$ . Therefore, taking the limit as  $r \rightarrow 0$ ,

$$\begin{aligned} & \int_{\text{large semicircle}} \frac{L(z)}{1+z^4} dz + \lim_{r \rightarrow 0+} \int_{-R}^{-r} \frac{\ln(-t) + i\pi}{1+t^4} dt + \\ & \lim_{r \rightarrow 0+} \int_r^R \frac{\ln t}{1+t^4} dt = 2\pi i \left( \frac{3}{32}\sqrt{2}\pi + \frac{3}{32}i\sqrt{2}\pi + \frac{1}{32}\sqrt{2}\pi - \frac{1}{32}i\sqrt{2}\pi \right). \end{aligned}$$

Observing that  $\int_{\text{large semicircle}} \frac{L(z)}{1+z^4} dz \rightarrow 0$  as  $R \rightarrow \infty$ ,

$$e(R) + 2 \lim_{r \rightarrow 0+} \int_r^R \frac{\ln t}{1+t^4} dt + i\pi \int_{-\infty}^0 \frac{1}{1+t^4} dt = \left(-\frac{1}{8} + \frac{1}{4}i\right) \pi^2 \sqrt{2}$$

where  $e(R) \rightarrow 0$  as  $R \rightarrow \infty$ . From an earlier example this becomes

$$e(R) + 2 \lim_{r \rightarrow 0+} \int_r^R \frac{\ln t}{1+t^4} dt + i\pi \left(\frac{\sqrt{2}}{4} \pi\right) = \left(-\frac{1}{8} + \frac{1}{4}i\right) \pi^2 \sqrt{2}.$$

Now letting  $r \rightarrow 0+$  and  $R \rightarrow \infty$ ,

$$\begin{aligned} 2 \int_0^\infty \frac{\ln t}{1+t^4} dt &= \left(-\frac{1}{8} + \frac{1}{4}i\right) \pi^2 \sqrt{2} - i\pi \left(\frac{\sqrt{2}}{4} \pi\right) \\ &= -\frac{1}{8} \sqrt{2} \pi^2, \end{aligned}$$

and so

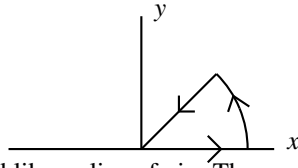
$$\int_0^\infty \frac{\ln t}{1+t^4} dt = -\frac{1}{16} \sqrt{2} \pi^2,$$

which is probably not the first thing you would think of. You might try to imagine how this could be obtained using elementary techniques.

**Example 36.6.6** The Fresnel integrals are

$$\int_0^\infty \cos(x^2) dx, \int_0^\infty \sin(x^2) dx.$$

To evaluate these integrals we will consider  $f(z) = e^{iz^2}$  on the curve which goes from the origin to the point  $r$  on the  $x$  axis and from this point to the point  $r \left(\frac{1+i}{\sqrt{2}}\right)$  along a circle of radius  $r$ , and from there back to the origin as illustrated in the following picture.



Thus the curve is shaped like a slice of pie. The angle is  $45^\circ$ . Denote by  $\gamma_r$  the curved part. Since  $f$  is analytic,

$$\begin{aligned} 0 &= \int_{\gamma_r} e^{iz^2} dz + \int_0^r e^{ix^2} dx - \int_0^r e^{i\left(t\left(\frac{1+i}{\sqrt{2}}\right)\right)^2} \left(\frac{1+i}{\sqrt{2}}\right) dt \\ &= \int_{\gamma_r} e^{iz^2} dz + \int_0^r e^{ix^2} dx - \int_0^r e^{-t^2} \left(\frac{1+i}{\sqrt{2}}\right) dt \\ &= \int_{\gamma_r} e^{iz^2} dz + \int_0^r e^{ix^2} dx - \frac{\sqrt{\pi}}{2} \left(\frac{1+i}{\sqrt{2}}\right) + e(r) \end{aligned}$$

where  $e(r) \rightarrow 0$  as  $r \rightarrow \infty$ . This used  $\int_0^\infty e^{-t^2} dt = \frac{\sqrt{\pi}}{2}$ . Now examine the first of these integrals.

$$\begin{aligned} \left| \int_{\gamma_r} e^{iz^2} dz \right| &= \left| \int_0^{\frac{\pi}{4}} e^{i(re^{it})^2} rie^{it} dt \right| \\ &\leq r \int_0^{\frac{\pi}{4}} e^{-r^2 \sin 2t} dt \\ &= \frac{r}{2} \int_0^1 \frac{e^{-r^2 u}}{\sqrt{1-u^2}} du \\ &= \frac{r}{2} \int_0^{r^{-(3/2)}} \frac{1}{\sqrt{1-u^2}} du + \frac{r}{2} \left( \int_0^1 \frac{1}{\sqrt{1-u^2}} du \right) e^{-(r^{1/2})} \end{aligned}$$

which converges to zero as  $r \rightarrow \infty$ . Therefore, taking the limit as  $r \rightarrow \infty$ ,

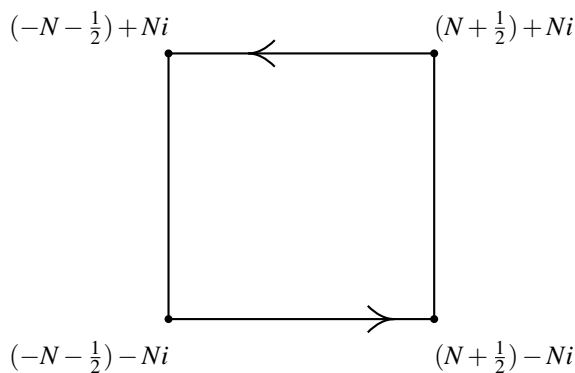
$$\frac{\sqrt{\pi}}{2} \left( \frac{1+i}{\sqrt{2}} \right) = \int_0^\infty e^{ix^2} dx$$

and so the Fresnel integrals are given by

$$\int_0^\infty \sin x^2 dx = \frac{\sqrt{\pi}}{2\sqrt{2}} = \int_0^\infty \cos x^2 dx.$$

The following example is one of the most interesting. By an auspicious choice of the contour it is possible to obtain a very interesting formula for  $\cot \pi z$  known as the Mittag Leffler expansion of  $\cot \pi z$ .

**Example 36.6.7** Let  $\gamma_N$  be the contour which goes from  $-N - \frac{1}{2} - Ni$  horizontally to  $N + \frac{1}{2} - Ni$  and from there, vertically to  $N + \frac{1}{2} + Ni$  and then horizontally to  $-N - \frac{1}{2} + Ni$  and finally vertically to  $-N - \frac{1}{2} - Ni$ . Thus the contour is a large rectangle and the direction of integration is in the counter clockwise direction.



Consider the following integral.

$$I_N \equiv \int_{\gamma_N} \frac{\pi \cos \pi z}{(\alpha^2 - z^2) \sin \pi z} dz$$

where  $\alpha$  is not an integer. This will be used to verify the formula of Mittag Leffler,

$$\frac{1}{\alpha^2} + \sum_{n=1}^{\infty} \frac{2}{\alpha^2 - n^2} = \frac{\pi \cot \pi \alpha}{\alpha}. \quad (36.10)$$



It is left as an exercise to verify that  $\cot \pi z$  is bounded on this contour and that therefore,  $I_N \rightarrow 0$  as  $N \rightarrow \infty$ . Now compute the residues of the integrand at  $\pm\alpha$  and at  $n$  where  $|n| < N + \frac{1}{2}$  for  $n$  an integer. These are the only singularities of the integrand in this contour and therefore,  $I_N$  can be obtained by using these. First consider the residue at  $\pm\alpha$ . These are obviously poles of order 1 and so to get the one at  $\alpha$ , you take

$$\lim_{z \rightarrow \alpha} \frac{(z - \alpha) \pi \cos \pi z}{(\alpha^2 - z^2) \sin \pi z} = \lim_{z \rightarrow \alpha} \frac{-\pi \cos \pi z}{(\alpha + z) \sin \pi z} = \frac{-\pi \cos \pi \alpha}{2\alpha \sin \pi \alpha}$$

You get the same thing at  $-\alpha$ . Next consider the residue at  $n$ . If you consider the power series, you will see that this should also be a pole of order 1. Thus it is

$$\begin{aligned} \lim_{z \rightarrow n} \frac{(z - n) \pi \cos \pi z}{(\alpha^2 - z^2) \sin \pi z} &= \lim_{z \rightarrow n} \frac{\pi \cos \pi z - (z - n) \pi^2 \sin(\pi z)}{-2z \sin \pi z + (\alpha^2 - z^2) \pi \cos(\pi z)} \\ &= \frac{\pi (-1)^n}{(\alpha^2 - n^2) \pi (-1)^n} = \frac{1}{\alpha^2 - n^2} \end{aligned}$$

Therefore,

$$0 = \lim_{N \rightarrow \infty} I_N = \lim_{N \rightarrow \infty} 2\pi i \left[ \sum_{n=-N}^N \frac{1}{\alpha^2 - n^2} - \frac{\pi \cot \pi \alpha}{\alpha} \right]$$

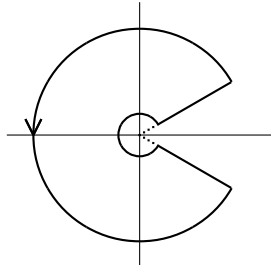
which establishes the following formula of Mittag Leffler.

$$\lim_{N \rightarrow \infty} \sum_{n=-N}^N \frac{1}{\alpha^2 - n^2} = \frac{\pi \cot \pi \alpha}{\alpha}.$$

Writing this in a slightly nicer form, we obtain 36.10.

The next example illustrates the technique of a branch cut.

**Example 36.6.8** For  $p \in (0, 1)$ , find  $\int_0^\infty \frac{x^{p-1}}{1+x} dx$ . This example illustrates the use of something called a branch cut. The idea is you need to pick a single determination of  $z^{p-1}$  which converges to  $x^{p-1}$  for  $x$  real and  $z$  getting close to  $x$ . It will make use of the following contour. In this contour, the radius of the large circle is  $R$  and the radius of the small one is  $r$ . The angle between the straight lines and the  $x$  axis is  $\varepsilon$ . Denote this contour by  $\gamma_{R,r,\varepsilon}$ .



Choose a branch of the logarithm of the form  $\log(z) = \ln|z| + iA(z)$  where  $A(z)$  is the angle of  $z$  in  $(0, 2\pi)$ . Thus

$$z^{p-1} = e^{(p-1)(\ln|z| + iA(z))}$$

The straight lines, the one on top.  $re^{i\varepsilon} + t(Re^{i\varepsilon}) = z, t \in [0, 1]$ .

Contour integral:

$$\int_0^1 \frac{|re^{i\varepsilon} + t(Re^{i\varepsilon})|^{p-1} e^{(p-1)i\varepsilon}}{1 + re^{i\varepsilon} + t(Re^{i\varepsilon})} Re^{i\varepsilon} dt$$

The one on the bottom:  $re^{i(2\pi-\varepsilon)} + t(Re^{i(2\pi-\varepsilon)}) = z, t \in [0, 1]$

Contour integral:

$$-\int_0^1 \frac{|re^{i(2\pi-\varepsilon)} + t(Re^{i(2\pi-\varepsilon)})|^{p-1} e^{(p-1)i(2\pi-\varepsilon)}}{1 + re^{i(2\pi-\varepsilon)} + t(Re^{i(2\pi-\varepsilon)})} Re^{i(2\pi-\varepsilon)} dt$$

The integral over the small circle:  $z = re^{it}, t \in [\varepsilon, 2\pi - \varepsilon]$

Contour integral:

$$-\int_{\varepsilon}^{2\pi-\varepsilon} \frac{r^{p-1} e^{(p-1)it}}{1 + re^{it}} rie^{it} dt$$

The integral over the large circle:  $z = Re^{it}, t \in [\varepsilon, 2\pi - \varepsilon]$

Contour integral:

$$\int_{\varepsilon}^{2\pi-\varepsilon} \frac{R^{p-1} e^{(p-1)it}}{1 + Re^{it}} Rie^{it} dt$$

$$2\pi i e^{i\pi(p-1)} = \int_{\gamma_{R,\varepsilon}} \frac{z^{p-1}}{1+z} dz$$

The residue at  $-1$  of the function is  $e^{i\pi(p-1)}$  and so the contour integral on the right equals the sum of those other integrals above. Now let  $\varepsilon \rightarrow 0$ . This yields

$$2\pi i e^{i\pi(p-1)} = \int_{\gamma_{R,r}} \frac{z^{p-1}}{1+z} dz$$

where the integral on the right equals the sum

$$\begin{aligned} & \int_0^1 \frac{(r+tR)^{p-1}}{1+r+tR} R dt + \left( - \int_0^1 \frac{(r+tR)^{p-1} e^{(p-1)i(2\pi)}}{1+r+tR} R dt \right) \\ & + \int_0^{2\pi} \frac{r^{p-1} e^{(p-1)it}}{1+re^{it}} rie^{it} dt + \int_0^{2\pi} \frac{R^{p-1} e^{(p-1)it}}{1+Re^{it}} Rie^{it} dt \end{aligned}$$

The last two integrals converge to 0 as  $r \rightarrow 0$  and  $R \rightarrow \infty$ . This follows easily from the form of the integrands. You can change the variable in the first two to write them as

$$\int_r^R \frac{x^{p-1}}{1+x} dx, -e^{(p-1)i(2\pi)} \int_r^R \frac{x^{p-1}}{1+x} dx$$

Thus

$$2\pi i e^{i\pi(p-1)} = \int_r^R \frac{x^{p-1}}{1+x} dx \left( 1 - e^{(p-1)i(2\pi)} \right) + E_1(r) + E_2(R)$$

where  $E(r), E(R)$  converges to 0 as  $r \rightarrow 0$  and  $R \rightarrow \infty$ . There is no hope of taking a limit as  $R \rightarrow \infty$  while keeping  $r > 0$  fixed, but if we let both variables converge at the same time, then we could get something. Let  $r \rightarrow 0+$  and let  $R = 1/r$ .

$$2\pi i e^{i\pi(p-1)} = \int_r^{1/r} \frac{x^{p-1}}{1+x} dx \left(1 - e^{(p-1)i(2\pi)}\right) + E(r, 1/r)$$

and now, as  $r \rightarrow 0$ , which equals the sum of the two integrals over the straight lines added to the integrals over the small circles which converge to 0 as  $r \rightarrow 0$  and  $R \rightarrow \infty$ . Top straight line converges as  $r \rightarrow 0$  to

$$\int_0^1 \frac{(Rt)^{p-1}}{1+tR} R dt$$

Bottom integral converges as  $r \rightarrow 0$  to

$$- \int_0^1 \frac{|tR|^{p-1} e^{(p-1)i(2\pi)}}{1+tR} R dt$$

Of course change variables letting  $x = tR$  and the two integrals which must be summed are

$$\int_0^R \frac{x^{p-1}}{1+x} dx, -e^{(p-1)i(2\pi)} \int_0^R \frac{x^{p-1}}{1+x} dx$$

Thus you get the following as  $R \rightarrow \infty$ .

$$\int_0^\infty \frac{x^{p-1}}{1+x} dx \left(1 - e^{2\pi(p-1)i}\right) = 2\pi i e^{i\pi(p-1)}$$

Then what you get is

$$\begin{aligned} \int_0^\infty \frac{x^{p-1}}{1+x} dx &= \frac{2\pi i e^{i\pi(p-1)}}{1 - e^{2\pi(p-1)i}} \\ &= \frac{-2\pi i e^{i\pi p}}{1 - e^{2\pi p i}} = \frac{-2\pi i}{(1 - e^{2\pi p i}) e^{-i\pi p}} = \frac{-2\pi i}{e^{-i\pi p} - e^{i\pi p}} \\ &= \frac{-2\pi i}{(\cos \pi p - i \sin(\pi p)) - (\cos \pi p + i \sin(\pi p))} = \frac{\pi}{\sin(p\pi)} \end{aligned}$$

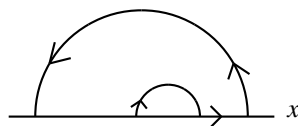
I think this is quite an amazing result.

Actually, people typically are a little more informal in the consideration of such integrals. They regard the bottom side of the line  $x \geq 0$  as being associated with  $\theta = 2\pi$  and the top side being associated with  $\theta = 0$  and leave out the fuss with taking limits as  $\varepsilon \rightarrow 0$  and so forth.

## 36.7 Exercises

1. Find the following improper integral.  $\int_{-\infty}^{\infty} \frac{\cos x}{1+x^4} dx$  **Hint:** Use upper semicircle contour and consider instead  $\int_{-\infty}^{\infty} \frac{e^{ix}}{1+x^4} dx$ . This is because the integral over the semicircle will converge to 0 as  $R \rightarrow \infty$  if you have  $e^{iz}$  but this won't happen if you use  $\cos z$  because  $\cos z$  will be unbounded. Just write down and check and you will see why this happens. Thus you should use  $\frac{e^{iz}}{1+z^4}$  and take real part. I think the standard calculus techniques will not work for this horrible integral.

2. Find  $\int_{-\infty}^{\infty} \frac{\cos(x)}{(1+x^2)^2} dx$ . **Hint:** Do the same as above replacing  $\cos x$  with  $e^{ix}$ .
3. Consider the following contour.

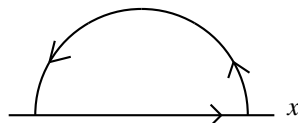


The small semicircle has radius  $r$  and is centered at  $(1,0)$ . The large semicircle has radius  $R$  and is centered at  $(0,0)$ . Use the method of residues to compute

$$\lim_{r \rightarrow 0} \left( \lim_{R \rightarrow \infty} \int_r^R \frac{x}{1-x^3} dx + \int_{-R}^{-r} \frac{x}{1-x^3} dx \right)$$

This is called the Cauchy principal value for  $\int_{-\infty}^{\infty} \frac{x}{1-x^3} dx$ . The integral makes no sense in terms of a real honest integral. The function has a pole on the  $x$  axis. Another instance of this was in Problem 6 on Page 720 where  $\int_0^{\infty} \sin(x)/x dx$  was determined similarly. However, you can define such a Cauchy principal value. Rather than belabor this issue, I will illustrate with this example. These principal value integrals occur because of cancelation. They depend on a particular way of taking a limit. They are not mathematically respectable but are certainly interesting. They are in that general area of finding something by taking a certain kind of symmetric limit. Such problems include the Lebesgue fundamental theorem of calculus with the symmetric derivative.

4. Find  $\int_0^{2\pi} \frac{\cos(\theta)}{1+\sin^2(\theta)} d\theta$ .
5. Find  $\int_0^{2\pi} \frac{d\theta}{2-\sin \theta}$ .
6. Find  $\int_{-\pi/2}^{\pi/2} \frac{d\theta}{2-\sin \theta}$ .
7. Suppose you have a function  $f(z)$  which is the quotient of two polynomials in which the degree of the top is two less than the degree of the bottom and you consider the contour.



Then define

$$\int_{\gamma_R} f(z) e^{isz} dz$$

in which  $s$  is real and positive. Explain why the integral makes sense and why the part of it on the semicircle converges to 0 as  $R \rightarrow \infty$ . Use this to find

$$\int_{-\infty}^{\infty} \frac{e^{isx}}{k^2 + x^2} dx, \quad k > 0.$$

8. Show using methods from real analysis that for  $b \geq 0$ ,

$$\int_0^\infty e^{-x^2} \cos(2bx) dx = \frac{\sqrt{\pi}}{2} e^{-b^2}$$

**Hint:** Let  $F(b) \equiv \int_0^\infty e^{-x^2} \cos(2bx) dx - \frac{\sqrt{\pi}}{2} e^{-b^2}$ . Then from Problem 13 on Page 383,  $F(0) = 0$ . Using the mean value theorem on difference quotients, explain why

$$\begin{aligned} F'(b) &= \int_0^\infty -2xe^{-x^2} \sin(2bx) dx + 2b \frac{\sqrt{\pi}}{2} e^{-b^2} \\ F'(b) &= 2b \left( \int_0^\infty e^{-x^2} \cos(2bx) dx + \frac{\sqrt{\pi}}{2} e^{-b^2} \right) \\ &= 2b \left( F(b) + \frac{\sqrt{\pi}}{2} e^{-b^2} + \frac{\sqrt{\pi}}{2} e^{-b^2} \right) \\ &= 2bF(b) + \sqrt{\pi} 2be^{-b^2} \end{aligned}$$

Now use the integrating factor method for solving linear differential equations from beginning differential equations to solve the ordinary differential equation.

$$\frac{d}{db} \left( e^{-b^2} F(b) \right) = \sqrt{\pi} 2be^{-2b^2}$$

Then

$$\begin{aligned} e^{-b^2} F(b) - 0 &= -\frac{1}{2} e^{-2b^2} \sqrt{\pi} + \frac{1}{2} \sqrt{\pi} \\ F(b) &= -\frac{1}{2} e^{-b^2} + \frac{1}{2} \sqrt{\pi} e^{-b^2} = 0 \end{aligned}$$

You fill in the details. This is meant to be a review of real variable techniques.

9. For  $b > 0$ , use the contour which goes from  $-a$  to  $a$  to  $a + ib$  to  $-a + ib$  to  $-a$ . Then let  $a \rightarrow \infty$  and show that the integral of  $e^{-z^2}$  over the vertical parts of this contour converge to 0. **Hint:** You know from an earlier problem what happens on the bottom part of the contour. Also for  $z = x + ib$ ,  $e^{-z^2} = e^{-(x^2 - b^2 + 2ixb)} = e^{b^2} e^{-x^2} (\cos(2xb) + i \sin(2xb))$ .

10. Consider the circle of radius 1 oriented counter clockwise. Evaluate

$$\int_\gamma z^{-6} \cos(z) dz$$

11. Consider the circle of radius 1 oriented counter clockwise. Evaluate

$$\int_\gamma z^{-7} \cos(z) dz$$

12. Find  $\int_0^\infty \frac{2+x^2}{1+x^4} dx$ .

13. Find  $\int_0^\infty \frac{x^{1/3}}{1+x^2} dx$

14. Suppose  $f$  is an entire function and that it has no zeros. Show there must exist an entire function  $g$  such that  $f(z) = e^{g(z)}$ . **Hint:** Letting  $\gamma(0, z)$  be the line segment which goes from 0 to  $z$ , let  $\hat{g}(z) \equiv \int_{\gamma(0, z)} \frac{f'(w)}{f(w)} dw$ . Then show that  $\hat{g}'(z) = \frac{f'(z)}{f(z)}$ . Then  $\left(e^{-\hat{g}(z)} f(z)\right)' = e^{-\hat{g}(z)} \frac{-f'(z)}{f(z)} f(z) + e^{-\hat{g}(z)} f'(z) = 0$ . Now when you have an entire function whose derivative is 0, it must be a constant. Modify  $\hat{g}(z)$  to make  $f(z) = e^{g(z)}$ .
15. Let  $f$  be an entire function with zeros  $\{\alpha_1, \dots, \alpha_n\}$  listed according to multiplicity. Thus you might have repeats in this list. Show that there is an analytic function  $g(z)$  such that for all  $z \in \mathbb{C}$ ,

$$f(z) = \prod_{k=1}^n (z - \alpha_k) e^{g(z)}$$

**Hint:** You know  $f(z) = \prod_{k=1}^n (z - \alpha_k) h(z)$  where  $h(z)$  has no zeros. To see this, note that near  $\alpha_1$ ,  $f(z) = a_1(z - \alpha_1) + a_2(z - \alpha_1)^2 + \dots$  and so  $f(z) = (z - \alpha_1) f_1(z)$  where  $f_1(z) \neq 0$  at  $\alpha_1$ . Now do the same for  $f_1$  and continue till  $f_n = h$ . Now use the above problem.

## Chapter 37

# Some Fundamental Functions and Transforms

### 37.1 Gamma Function

This chapter is on some fundamental ideas related to Fourier and Laplace transforms and the Gamma function. The symbol  $\int_a^\infty f(t) dt$  will always mean

$$\lim_{R \rightarrow \infty} \int_a^R f(t) dt$$

provided  $f$  is piecewise continuous on  $[a, \infty)$  whenever the limit exists. It is the standard improper integral from calculus.  $\int_{-\infty}^a f(t) dt$  is defined similarly. However, if  $f$  is unbounded at 0, the symbol will mean

$$\lim_{R \rightarrow \infty} \int_{a+1}^R f(t) dt + \lim_{\delta \rightarrow 0+} \int_{a+\delta}^1 f(t) dt$$

or more simply, when  $f(t) \geq 0$ ,

$$\lim_{\delta \rightarrow 0} \int_{a+\delta}^{a+\delta^{-1}} f(t) dt$$

First is a very important function defined in terms of an integral. Also recall that the value of the Riemann integral does not depend on the value of the function at single points. All this is more satisfactory if you do it in the context of the Lebesgue integral. Here it is assumed that all functions are piecewise continuous having finitely many jumps in every finite interval so there will be no difficulty in writing the Riemann integral.

**Definition 37.1.1** *The gamma function is defined by*

$$\Gamma(\alpha) \equiv \int_0^\infty e^{-t} t^{\alpha-1} dt$$

*whenever  $\alpha > 0$ .*

**Lemma 37.1.2** *The integral is finite for each  $\alpha > 0$ .*

**Proof:** By the monotone convergence theorem, for  $n \in \mathbb{N}$

$$\begin{aligned}\Gamma(\alpha) &= \lim_{n \rightarrow \infty} \int_{1/n}^n e^{-t} t^{\alpha-1} dt \leq \limsup_{n \rightarrow \infty} \left( \int_{1/n}^1 t^{\alpha-1} dt + \int_1^n C e^{-t/2} dt \right) \\ &\leq \frac{1}{\alpha} + \lim_{n \rightarrow \infty} \left( -2C e^{-\frac{1}{2}n} + 2C e^{-\frac{1}{2}} \right) < \infty\end{aligned}$$

The explanation for the constant is as follows. For  $t \geq 1$  and  $m$  a positive integer larger than  $\alpha$ ,

$$\frac{t^{\alpha-1}}{e^{t/2}} < \frac{t^{m-1}}{e^{t/2}}$$

which converges to 0 as  $t \rightarrow \infty$  which is easily shown by an appeal to L'Hospital's rule. Hence

$$t^{\alpha-1} e^{-t} \leq C e^{t/2} e^{-t} = C e^{-t/2}. \blacksquare$$

**Proposition 37.1.3** For  $n$  a positive integer,  $n! = \Gamma(n+1)$ . In general,  $\Gamma(1) = 1, \Gamma(\alpha+1) = \alpha\Gamma(\alpha)$

**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.  $\blacksquare$

## 37.2 Laplace Transform

Everything holds for a much more general set of assumptions if you have a more modern version of the integral. This is why I am using notation which corresponds to this more general situation. All of the functions considered here are assumed piecewise continuous with finitely many jumps in every finite interval. Then such a function  $f$  is said to be in  $L^1([0, \infty))$  if

$$\int_0^{\infty} |f(t)| dt < \infty$$

Similar usages of this symbol are defined synonymously. Sometimes I will just write  $L^1$  to indicate that the absolute value of the function is integrable. Here is the definition of a Laplace transform.



**Definition 37.2.1** A function  $\phi$  has exponential growth on  $[0, \infty)$  if there are positive constants  $\lambda, C$  such that  $|\phi(t)| \leq Ce^{\lambda t}$  for all  $t$ . Then for  $s > \lambda$ , one defines the Laplace transform  $\mathcal{L}\phi(s) \equiv \int_0^\infty \phi(t) e^{-st} dt$ .

**Theorem 37.2.2** If  $s$  is a complex number and  $\operatorname{Re} s > \lambda$  where  $|\phi(t)| \leq Ce^{\lambda t}$ , and

$$f(s) \equiv \int_0^\infty e^{-st} \phi(t) dt$$

then for  $\operatorname{Re} s > \lambda$ ,

$$\lim_{h \rightarrow 0} \frac{f(s+h) - f(s)}{h} \equiv f'(s) = \int_0^\infty (-t) e^{-st} \phi(t) dt$$

Thus  $s \rightarrow f(s)$  is analytic on  $\operatorname{Re} s > \lambda$ .

**Proof:** Let  $\operatorname{Re} s > \lambda$ .  $s$  will be complex as will  $h$ .

$$\int_0^\infty \frac{e^{-(s+h)t} - e^{-st}}{h} \phi(t) dt + \int_0^\infty t e^{-st} \phi(t) dt = \int_0^\infty e^{-st} \left( \frac{e^{-ht} - 1}{h} + t \right) \phi(t) dt$$

Then

$$\begin{aligned} \frac{e^{-ht} - 1}{h} + t &= \frac{1}{h} \left( \sum_{k=0}^\infty (-1)^k h^k t^k - 1 \right) + t \\ &= \left( \sum_{k=1}^\infty (-1)^k h^{k-1} t^k \right) + t \\ &= h \sum_{k=2}^\infty (-1)^k h^{k-2} t^k \end{aligned}$$

Thus

$$\left| \left( \frac{e^{-ht} - 1}{h} + t \right) \right| \leq |h| t^2 e^{|h|t}$$

and so

$$\left| \int_0^\infty \frac{e^{-(s+h)t} - e^{-st}}{h} \phi(t) dt + \int_0^\infty t e^{-st} \phi(t) dt \right| \leq \int_0^\infty |h| t^2 e^{|h|t} e^{-\operatorname{Re}(s)t} e^{\lambda t} dt$$

which clearly converges to 0 since for all  $|h|$  sufficiently small,

$$e^{|h|t} e^{-\operatorname{Re}(s)t} e^{\lambda t} \leq e^{-(\operatorname{Re}(s) - (\lambda + \varepsilon))t}$$

where  $\varepsilon$  is small enough that  $\operatorname{Re}(s) > \lambda + \varepsilon$ . Thus the integral is finite for all  $|h|$  small enough and it is multiplied by  $|h|$ . ■

This shows that  $f$  is analytic on  $\operatorname{Re}(s) > \lambda$ . Hence it has all derivatives. In fact, you can do a similar computation to the above and verify that

$$f^{(k)}(s) = \int_0^\infty (-t)^k e^{-st} \phi(t) dt$$

### 37.3 Fourier Transform

**Definition 37.3.1** The Fourier transform is defined as follows for  $f \in L^1(\mathbb{R})$  meaning that

$$\lim_{R \rightarrow \infty} \int_{-R}^R |f(t)| dt < \infty$$

$f$  is piecewise continuous on  $[-R, R]$ . Then the Fourier transform is given by

$$Ff(t) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-itx} f(x) dx$$

The inverse Fourier transform is defined the same way except you delete the minus sign in the complex exponential.

$$F^{-1}f(t) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} f(x) dx$$

Does it deserve to be called the “inverse” Fourier transform? This question will be explored somewhat below.

There is a very important improper integral involving  $\sin(x)/x$ . You can show with a little estimating that  $x \rightarrow \sin(x)/x$  is not in  $L^1(0, \infty)$ . Nevertheless, a lot can be said about improper integrals involving this function.

**Theorem 37.3.2** The following hold

1.  $\int_0^{\infty} \frac{\sin u}{u} du = \frac{\pi}{2}$
2.  $\lim_{r \rightarrow \infty} \int_{\delta}^{\infty} \frac{\sin(ru)}{u} du = 0$  whenever  $\delta > 0$ .
3. If  $f \in L^1(\mathbb{R})$ , then  $\lim_{r \rightarrow \infty} \int_{\mathbb{R}} \sin(ru) f(u) du = 0$ . This is called the Riemann Lebesgue lemma.

**Proof:** The first claim follows from Problem 6 on Page 720 above.

Now consider  $\int_{\delta}^{\infty} \frac{\sin(ru)}{u} du$ . It equals  $\int_0^{\infty} \frac{\sin(ru)}{u} du - \int_0^{\delta} \frac{\sin(ru)}{u} du$  which can be seen from the definition of what the improper integral means. Also, you can change the variable. Let  $ru = t$  so  $rdu = dt$  and the above reduces to

$$\int_0^{\infty} \frac{\sin(t)}{t} \frac{1}{r} dt - \int_0^{r\delta} \frac{\sin(t)}{t} dt = \int_{\delta}^{\infty} \frac{\sin(ru)}{u} du$$

Thus

$$\frac{\pi}{2} - \int_0^{r\delta} \frac{\sin(t)}{t} dt = \int_{\delta}^{\infty} \frac{\sin(ru)}{u} du$$

and so  $\lim_{r \rightarrow \infty} \int_{\delta}^{\infty} \frac{\sin(ru)}{u} du = 0$  from the first part.

Now consider the third claim, the Riemann Lebesgue lemma. For  $I$  an interval let

$$\mathcal{X}_I(t) \equiv \begin{cases} 1 & \text{if } t \in I \\ 0 & \text{if } t \notin I \end{cases}$$

Then for  $f \in L^1$ , let  $f_R(t) \equiv \mathcal{X}_{[-R,R]}(t) f(t)$ . Then for  $R$  large,

$$\int_{-\infty}^{\infty} |f(t) - f_R(t)| dt = \int_R^{\infty} |f(t)| dt + \int_{-\infty}^{-R} |f(t)| dt < \varepsilon$$

Now  $f_R$  is Riemann integrable and so there is a step function  $s(t) = \sum_{i=1}^n a_i \mathcal{X}_{I_i}(t)$  such that  $|s(t)| \leq |f_R(t)|$  and

$$\int_{-R}^R |f_R(t) - s(t)| dt = \int_{-R}^R |f_R(t) - s(t)| dt < \varepsilon$$

This follows from the definition of the Riemann integral as a limit of integrals of step functions, details are left for you. Therefore,

$$\int_{-\infty}^{\infty} |s(t) - f(t)| dt < 2\varepsilon$$

Now

$$\begin{aligned} \left| \int_{-\infty}^{\infty} f(t) \sin(rt) dt \right| &\leq \int_{-\infty}^{\infty} |(f(t) - s(t)) \sin(rt)| dt + \left| \int_{-\infty}^{\infty} s(t) \sin(rt) dt \right| \\ &\leq 2\varepsilon + \left| \int_{-\infty}^{\infty} s(t) \sin(rt) dt \right| \end{aligned} \quad (37.1)$$

It remains to verify that  $\lim_{r \rightarrow \infty} \int_{-\infty}^{\infty} s(t) \sin(rt) dt = 0$ . Since  $s(t)$  is a sum of scalars times  $\mathcal{X}_I$  for  $I$  an interval, it suffices to verify that  $\lim_{r \rightarrow \infty} \int_{-\infty}^{\infty} \mathcal{X}_{[a,b]}(t) \sin(rt) dt = 0$ . However, this integral is just

$$\int_a^b \sin(rt) dt = \frac{-1}{r} \cos(rb) + \frac{1}{r} \cos(ra)$$

which clearly converges to 0 as  $r \rightarrow \infty$ . Therefore, for  $r$  large enough, 37.1 implies

$$\left| \int_{-\infty}^{\infty} f(t) \sin(rt) dt \right| < 3\varepsilon$$

Since  $\varepsilon$  is arbitrary, this shows that 3. holds. ■

**Definition 37.3.3** *The following notation will be used assuming the limits exist.*

$$\lim_{r \rightarrow 0+} g(x+r) \equiv g(x+), \quad \lim_{r \rightarrow 0+} g(x-r) \equiv g(x-)$$

**Theorem 37.3.4** *Suppose that  $g \in L^1(\mathbb{R})$  and that at some  $x$ ,  $g$  is locally Holder continuous from the right and from the left. This means there exist constants  $K, \delta > 0$  and  $r \in (0, 1]$  such that for  $|x-y| < \delta$ ,*

$$|g(x+) - g(y)| < K|x-y|^r \quad (37.2)$$

for  $y > x$  and

$$|g(x-) - g(y)| < K|x-y|^r \quad (37.3)$$

for  $y < x$ . Then

$$\begin{aligned} \lim_{r \rightarrow \infty} \frac{2}{\pi} \int_0^{\infty} \frac{\sin(ur)}{u} \left( \frac{g(x-u) + g(x+u)}{2} \right) du \\ = \frac{g(x+) + g(x-)}{2}. \end{aligned}$$

**Proof:** As in the proof of Theorem 37.3.2, changing variables shows that  $\frac{2}{\pi} \int_0^\infty \frac{\sin(ru)}{u} du = 1$ . Therefore,

$$\begin{aligned} & \frac{2}{\pi} \int_0^\infty \frac{\sin(ur)}{u} \left( \frac{g(x-u) + g(x+u)}{2} \right) du - \frac{g(x+) + g(x-)}{2} \\ &= \frac{2}{\pi} \int_0^\infty \frac{\sin(ur)}{u} \left( \frac{g(x-u) - g(x-) + g(x+u) - g(x+)}{2} \right) du \\ &= \frac{2}{\pi} \int_0^\delta \frac{\sin(ur)}{u} \left( \frac{g(x-u) - g(x-)}{2u} + \frac{g(x+u) - g(x+)}{2u} \right) du \\ & \quad + \frac{2}{\pi} \int_\delta^\infty \frac{\sin(ur)}{u} \left( \frac{g(x-u) - g(x-)}{2} + \frac{g(x+u) - g(x+)}{2} \right) du \quad (37.4) \end{aligned}$$

**Second Integral:** It equals

$$\begin{aligned} & \frac{2}{\pi} \int_\delta^\infty \frac{\sin(ur)}{u} \left( \frac{g(x-u) + g(x+u)}{2} - \frac{g(x-) + g(x+)}{2} \right) du \\ &= \frac{2}{\pi} \int_\delta^\infty \frac{\sin(ur)}{u} \left( \frac{g(x-u) + g(x+u)}{2} \right) du \\ & \quad - \frac{2}{\pi} \int_\delta^\infty \frac{\sin(ur)}{u} \left( \frac{g(x-) + g(x+)}{2} \right) du \quad (37.5) \end{aligned}$$

From part 2 of Theorem 37.3.2,

$$\lim_{r \rightarrow \infty} \frac{2}{\pi} \int_\delta^\infty \frac{\sin(ur)}{u} \frac{g(x-) + g(x+)}{2} du = 0$$

Thus consider the first integral in 37.4.

$$\begin{aligned} & \frac{2}{\pi} \int_\delta^\infty \frac{\sin(ur)}{u} \left( \frac{g(x-u) + g(x+u)}{2} \right) du \\ &= \frac{1}{\pi} \int_\delta^\infty \frac{\sin(ur)}{u} g(x-u) du + \frac{1}{\pi} \int_\delta^\infty \frac{\sin(ur)}{u} g(x+u) du \\ &= \frac{1}{\pi} \left( \int_{-\infty}^{-\delta} \frac{\sin(ur)}{u} g(x+u) du + \int_\delta^\infty \frac{\sin(ur)}{u} g(x+u) du \right) \end{aligned}$$

Now

$$\int_{-\infty}^{-\delta} \frac{\sin(ur)}{u} g(x+u) du = \int_{-\infty}^{-\delta} \sin(ur) \frac{g(x+u)}{u} du$$

and  $\left| \frac{g(x+u)}{u} \right| \leq \frac{1}{\delta} |g(x+u)|$  for  $u < -\delta$ . Thus  $u \rightarrow \frac{g(x+u)}{u}$  is in  $L^1((-\infty, -\delta))$ . Indeed,

$$\int_{-\infty}^{-\delta} \left| \frac{g(x+u)}{u} \right| du \leq \frac{1}{\delta} \int_{\mathbb{R}} |g(x+u)| du = \frac{1}{\delta} \int_{\mathbb{R}} |g(y)| dy < \infty$$

It follows from the Riemann Lebesgue lemma

$$\lim_{r \rightarrow \infty} \int_{-\infty}^{-\delta} \sin(ur) \frac{g(x+u)}{u} du = \lim_{r \rightarrow \infty} \int_\delta^\infty \sin(ur) \frac{g(x+u)}{u} du = 0$$

**First Integral in 37.4:** This converges to 0 as  $r \rightarrow \infty$  because of the Riemann Lebesgue lemma. Indeed, for  $0 \leq u \leq \delta$ ,

$$\left| \frac{g(x-u) - g(x-)}{2u} \right| \leq K \frac{1}{u^{1-r}}$$

which is integrable on  $[0, \delta]$ . The other quotient also is integrable by similar reasoning. ■

The next theorem justifies the terminology above which defines  $F^{-1}$  and calls it the inverse Fourier transform. Roughly it says that the inverse Fourier transform of the Fourier transform equals the mid point of the jump. Thus if the original function is continuous, it restores the original value of this function. Surely this is what you would want by calling something the inverse Fourier transform.

Now for certain special kinds of functions, the Fourier transform is indeed in  $L^1$  and one can show that it maps this special kind of function to another function of the same sort. This can be used as the basis for a general theory of Fourier transforms. However, the following does indeed give adequate justification for the terminology that  $F^{-1}$  is called the inverse Fourier transform.

**Theorem 37.3.5** *Let  $g \in L^1(\mathbb{R})$  and suppose  $g$  is locally Holder continuous from the right and from the left at  $x$  as in 37.2 and 37.3. Then*

$$\lim_{R \rightarrow \infty} \frac{1}{2\pi} \int_{-R}^R e^{ixt} \int_{-\infty}^{\infty} e^{-ity} g(y) dy dt = \frac{g(x+) + g(x-)}{2}.$$

**Proof:** Note that

$$\begin{aligned} \int_{-R}^R e^{ixt} \int_{-\infty}^{\infty} e^{-ity} g(y) dy dt &= \int_{-\infty}^{\infty} e^{-ity} g(y) dy \int_{-R}^R e^{ixt} dt \\ &= \int_{-\infty}^{\infty} e^{-ity} g(y) \int_{-R}^R e^{ixt} dy dt \end{aligned}$$

One merely takes a constant outside the integral and then moves a constant inside an integral. Consider the following manipulations.

$$\begin{aligned} &\frac{1}{2\pi} \int_{-R}^R e^{ixt} \int_{-\infty}^{\infty} e^{-ity} g(y) dy dt = \\ &\frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-R}^R e^{ixt} e^{-ity} g(y) dt dy = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-R}^R e^{i(x-y)t} g(y) dt dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} g(y) \left( \int_0^R e^{i(x-y)t} dt + \int_0^R e^{-i(x-y)t} dt \right) dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} g(y) \left( \int_0^R 2 \cos((x-y)t) dt \right) dy \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} g(y) \frac{\sin R(x-y)}{x-y} dy = \frac{1}{\pi} \int_{-\infty}^{\infty} g(x-y) \frac{\sin Ry}{y} dy \\ &= \frac{1}{\pi} \int_0^{\infty} (g(x-y) + g(x+y)) \frac{\sin Ry}{y} dy \\ &= \frac{2}{\pi} \int_0^{\infty} \left( \frac{g(x-y) + g(x+y)}{2} \right) \frac{\sin Ry}{y} dy \end{aligned}$$

From Theorem 37.3.4,

$$\begin{aligned}
 & \lim_{R \rightarrow \infty} \frac{1}{2\pi} \int_{-R}^R e^{ixt} \int_{-\infty}^{\infty} e^{-iy} g(y) dy dt \\
 &= \lim_{R \rightarrow \infty} \frac{2}{\pi} \int_0^{\infty} \left( \frac{g(x-y) + g(x+y)}{2} \right) \frac{\sin Ry}{y} dy \\
 &= \frac{g(x+) + g(x-)}{2}. \blacksquare
 \end{aligned}$$

## 37.4 The Inversion of Laplace Transforms

How does the Fourier transform relate to the Laplace transform? This is considered next. Recall that from Theorem 4.1.5 if  $g$  has exponential growth  $|g(t)| \leq Ce^{\lambda t}$ , then if  $\operatorname{Re}(s) > \lambda$ , one can define  $\mathcal{L}g(s)$  as

$$\mathcal{L}g(s) \equiv \int_0^{\infty} e^{-su} g(u) du$$

and also  $s \rightarrow \mathcal{L}g(s)$  is differentiable on  $\operatorname{Re}(s) > \lambda$  in the sense that if  $h \in \mathbb{C}$  and  $G(s) \equiv \mathcal{L}g(s)$ , then

$$\lim_{h \rightarrow 0} \frac{G(s+h) - G(s)}{h} = G'(s) = - \int_0^{\infty} u e^{-su} g(u) du$$

Thus  $G$  is analytic and has all derivatives. Then the next theorem shows how to invert the Laplace transform. It is another one of those results which says that you get the mid point of the jump when you do a certain process. It is like what happens in Fourier series where the Fourier series converges to the midpoint of the jump under suitable conditions and like what was just shown for the inverse Laplace transform. For a fairly elementary discussion of this kind of thing related to Fourier series, see the single variable advanced calculus book on my web page.

The next theorem gives a more specific version of what is contained in Theorem 4.2.3 presented later. However, this theorem does assume a Holder continuity condition which is not needed for Theorem 4.2.3. I think that it is usually the case that the needed Holder condition will be available.

**Theorem 37.4.1** *Let  $g$  be a piecewise continuous function defined on  $(0, \infty)$  which has exponential growth*

$$|g(t)| \leq Ce^{\lambda t} \text{ for some real } \lambda$$

*and is Holder continuous from the right and left as in 37.2 and 37.3. For  $\operatorname{Re}(s) > \lambda$*

$$\mathcal{L}g(s) \equiv \int_0^{\infty} e^{-su} g(u) du$$

*Then for any  $\gamma > \lambda$ ,*

$$\lim_{R \rightarrow \infty} \frac{1}{2\pi} \int_{-R}^R e^{(\gamma+iy)t} \mathcal{L}g(\gamma+iy) dy = \frac{g(t+) + g(t-)}{2} \quad (37.6)$$

**Proof:** This follows from plugging in the formula for the Laplace transform of  $g$  and then using the above. Thus

$$\frac{1}{2\pi} \int_{-R}^R e^{(\gamma+iy)t} \mathcal{L}g(\gamma+iy) dy =$$

$$\begin{aligned}
& \frac{1}{2\pi} \int_{-R}^R e^{(\gamma+iy)t} \int_{-\infty}^{\infty} e^{-(\gamma+iy)u} g(u) du dy \\
&= \frac{1}{2\pi} \int_{-R}^R e^{\eta t} e^{iyt} \int_{-\infty}^{\infty} e^{-(\gamma+iy)u} g(u) du dy \\
&= e^{\eta t} \frac{1}{2\pi} \int_{-R}^R e^{iyt} \int_{-\infty}^{\infty} e^{-iyu} e^{-\gamma u} g(u) du dy
\end{aligned}$$

Now apply Theorem 37.3.5 to conclude that

$$\begin{aligned}
& \lim_{R \rightarrow \infty} \frac{1}{2\pi} \int_{-R}^R e^{(\gamma+iy)t} \mathcal{L}g(\gamma+iy) dy \\
&= e^{\eta t} \lim_{R \rightarrow \infty} \frac{1}{2\pi} \int_{-R}^R e^{iyt} \int_{-\infty}^{\infty} e^{-iyu} e^{-\gamma u} g(u) du dy \\
&= e^{\eta t} \frac{g(t+)e^{-\eta+} + g(t-)e^{-\eta-}}{2} = \frac{g(t+) + g(t-)}{2}. \blacksquare
\end{aligned}$$

In particular, this shows that if  $\mathcal{L}g(s) = \mathcal{L}h(s)$  for all  $s$  large enough, both  $g, h$  having exponential growth, then  $f, g$  must be equal except for jumps and in fact, at any point where they are both Holder continuous from right and left, the mid point of their jumps is the same.

This answers the question raised earlier about whether the Laplace transform method even makes sense to use because it shows that if two functions have the same Laplace transform, then they are the same function except at jumps where the midpoint of the jumps coincide.

Next is a systematic way to invert the Laplace transform. It will be no harder than what is usually done in standard differential equations courses but differs from this material in being completely general.

## 37.5 The Bromwich Integral

First pick  $\gamma > \lambda$  and write the integral on the left in 37.6 as a contour integral. Thus  $z = \gamma + iy$  and  $dz = i dy$  and this is just the contour integral

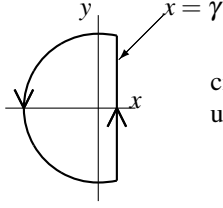
$$\frac{1}{2\pi i} \int_{\gamma-iR}^{\gamma+iR} e^{ut} \mathcal{L}g(u) du$$

where the contour is the straight line from  $\gamma - iR$  to  $\gamma + iR$ . Indeed, if you parametrize this contour as  $z = \gamma + iy$  and use the procedures for evaluation of contour integrals, you get the integral in 37.6. Then taking the limit as  $R \rightarrow \infty$  it is customary to write this limit as

$$\frac{1}{2\pi i} \int_{\gamma-i\infty}^{\gamma+i\infty} e^{ut} \mathcal{L}g(u) du$$

This is called the Bromwich integral and as shown earlier it recovers the mid point of the jump of  $g$  at  $t$  for every point  $t$  where  $g$  is Holder continuous from the right and from the left. Remember  $t \geq 0$ . Now  $u \rightarrow e^{ut} \mathcal{L}g(u)$  is analytic for  $\operatorname{Re}(u) > \eta$  and in particular for  $\operatorname{Re}(u) \geq \gamma$  therefore, all of the poles of  $u \rightarrow \mathcal{L}g(u)$  are contained in the set  $\operatorname{Re}(u) < \gamma$ . Indeed, in practice,  $u \rightarrow \mathcal{L}g(u)$  ends up being represented by a formula which is clearly a meromorphic function, one which is analytic except for isolated poles.

So how do you compute this Bromwich integral? This is where the method of residues is very useful. Consider the following contour.



Let  $\eta_R$  be the above contour oriented as shown. The radius of the circular part is  $R$ . Let  $C_R$  be the curved part. Then one can show that under suitable assumptions

$$\lim_{R \rightarrow \infty} \frac{1}{2\pi i} \int_{C_R} e^{ut} F(u) du = 0 \quad (37.7)$$

The needed condition is that for all  $|z|$  large enough,

$$|F(z)| \leq \frac{C}{|z|^\alpha}, \text{ some } \alpha > 0. \quad (37.8)$$

Note that this assumption implies there are finitely many poles for  $F(z)$  because if  $w$  is a pole, you have  $\lim_{z \rightarrow w} |F(z)| = \infty$ . Thus all poles are in some disk of suitable radius. Also recall that poles have no limit point. Thus there are only finitely many in a suitably large disk and this accounts for all of them.

**Lemma 37.5.1** *Let the contour be as shown and assume 37.8. Then the above limit in 37.7 exists.*

**Proof:** Assume  $c \geq 0$  as shown and let  $\theta$  be the angle between the positive  $x$  axis and a point on  $C_R$ . Let  $0 < \beta < \alpha$ . Then the contour integral over  $C_R$  will be broken up into three pieces, two pieces around the  $y$  axis

$$\theta \in \left[ \frac{\pi}{2} - \arcsin\left(\frac{c}{R}\right), \frac{\pi}{2} + \arcsin\left(\frac{c}{R^{1-\beta}}\right) \right], \\ \left[ \frac{3\pi}{2} - \arcsin\left(\frac{c}{R^{1-\beta}}\right), \frac{3\pi}{2} + \arcsin\left(\frac{c}{R}\right) \right],$$

and the third having

$$\theta \in \left( \frac{\pi}{2} + \arcsin\left(\frac{c}{R^{1-\beta}}\right), \frac{3\pi}{2} - \arcsin\left(\frac{c}{R^{1-\beta}}\right) \right)$$

Then,

$$\int_{C_R} e^{tz} F(z) dz = \int_{\frac{\pi}{2} + \arcsin(\frac{c}{R^{1-\beta}})}^{\frac{3\pi}{2} - \arcsin(\frac{c}{R^{1-\beta}})} e^{(R \cos \theta + iR \sin \theta)t} F(Re^{i\theta}) Rie^{i\theta} d\theta + \quad (37.9) \\ + \int_{\frac{\pi}{2} - \arcsin(\frac{c}{R})}^{\frac{\pi}{2} + \arcsin(\frac{c}{R^{1-\beta}})} e^{(R \cos \theta + iR \sin \theta)t} F(Re^{i\theta}) Rie^{i\theta} d\theta \\ + \int_{\frac{3\pi}{2} - \arcsin(\frac{c}{R^{1-\beta}})}^{\frac{3\pi}{2} + \arcsin(\frac{c}{R})} e^{(R \cos \theta + iR \sin \theta)t} F(Re^{i\theta}) Rie^{i\theta} d\theta$$

Consider the last two integrals first. For large  $|z|$ , with  $z \in C_R^*$ , the sum of the absolute values of these is no more than

$$\left| \int_{\frac{\pi}{2} - \arcsin(\frac{c}{R})}^{\frac{\pi}{2} + \arcsin(\frac{c}{R^{1-\beta}})} e^{R(\cos \theta)t} \frac{C}{R^\alpha} R d\theta \right| + \left| \int_{\frac{3\pi}{2} - \arcsin(\frac{c}{R^{1-\beta}})}^{\frac{3\pi}{2} + \arcsin(\frac{c}{R})} e^{R(\cos \theta)t} \frac{C}{R^\alpha} R d\theta \right| \\ \leq C e^{R(\cos(\frac{\pi}{2} - \arcsin(\frac{c}{R})))t} \left( \arcsin\left(\frac{c}{R^{1-\beta}}\right) + \arcsin\left(\frac{c}{R}\right) \right) R^{1-\alpha} \\ + C e^{R(\cos(\frac{3\pi}{2} + \arcsin(\frac{c}{R})))t} \left( \arcsin\left(\frac{c}{R^{1-\beta}}\right) + \arcsin\left(\frac{c}{R}\right) \right) R^{1-\alpha}$$



Now from trig. identities,  $\cos\left(\frac{\pi}{2} - \arcsin(\theta)\right) = \theta$ ,  $\cos\left(\frac{3\pi}{2} + \arcsin(\theta)\right) = \theta$ , and so the above reduces to

$$2Ce^{ct} \left( \arcsin\left(\frac{c}{R^{1-\beta}}\right) + \arcsin\left(\frac{c}{R}\right) \right) R^{1-\alpha}$$

which converges to 0 as  $R \rightarrow \infty$ . Recall  $0 < \beta < \alpha$ . It remains to consider the integral in 37.9. For large  $|z|$ , the absolute value of this integral is no more than

$$\int_{\frac{\pi}{2} + \arcsin\left(\frac{c}{R}\right)}^{\frac{3\pi}{2} - \arcsin\left(\frac{c}{R}\right)} e^{R(\cos \theta)t} \frac{C}{R^\alpha} R d\theta \leq C\pi e^{Rt \cos\left(\frac{\pi}{2} + \arcsin\left(\frac{c}{R^{1-\beta}}\right)\right)} R^{1-\alpha} = C\pi R^{1-\alpha} e^{-cR^\beta}$$

which converges to 0 as  $R \rightarrow \infty$ . ■

**Corollary 37.5.2** *Let the contour be as shown and assume 37.8 for meromorphic  $F(s)$ . Then the above limit in 37.7 exists. Also  $f(t)$ , given by the Bromwich integral, is continuous on  $(0, \infty)$  and its Laplace transform is  $F(s)$ .*

**Proof:** It only remains to verify continuity. Let  $R$  be so large that the above contour  $\eta_R^*$  encloses all poles of  $F$ . Then for such large  $R$ , the contour integrals are not changing because all the poles are enclosed. Thus

$$\begin{aligned} f(\hat{t}) &= \lim_{R \rightarrow \infty} \frac{1}{2\pi i} \int_{\eta_R} e^{\hat{u}t} F(u) du = \frac{1}{2\pi i} \int_{\eta_R} e^{\hat{u}t} F(u) du \\ |f(\hat{t}) - f(t)| &\leq \left| f(\hat{t}) - \frac{1}{2\pi i} \int_{\eta_R} e^{\hat{u}t} F(u) du \right| \\ &\quad + \left| \frac{1}{2\pi i} \int_{\eta_R} e^{\hat{u}t} F(u) du - \frac{1}{2\pi i} \int_{\eta_R} e^{ut} F(u) du \right| \\ &\quad + \left| \frac{1}{2\pi i} \int_{\eta_R} e^{ut} F(u) du - f(t) \right| \\ &= \left| \frac{1}{2\pi i} \int_{\eta_R} e^{\hat{u}t} F(u) du - \frac{1}{2\pi i} \int_{\eta_R} e^{ut} F(u) du \right| \end{aligned}$$

Since  $\eta_R$  is fixed, it follows that if  $|\hat{t} - t|$  is small enough, then  $|f(\hat{t}) - f(t)|$  is also small. ■

It follows from Lemma 37.5.1 that

$$\begin{aligned} f(t) &\equiv \lim_{R \rightarrow \infty} \frac{1}{2\pi i} \int_{\gamma-iR}^{\gamma+iR} e^{ut} F(u) du \\ &= \lim_{R \rightarrow \infty} \left( \frac{1}{2\pi i} \int_{\gamma-iR}^{\gamma+iR} e^{ut} F(u) du + \frac{1}{2\pi i} \int_{C_R} e^{ut} F(u) du \right) \\ &= \frac{1}{2\pi i} 2\pi i (\text{sum of residues of the poles of } e^{zt} F(z)) \\ &= \text{sum of residues.} \end{aligned}$$

The following procedure shows how the Bromwich integral can be computed to obtain an actual formula for a function. However, the integral itself will make sense and could be numerically computed to solve for the inverse Laplace transform.

**PROCEDURE 37.5.3** Suppose  $F(s)$  is a Laplace transform and is meromorphic on  $\mathbb{C}$  and satisfies 37.8. (This situation is quite typical) Then to compute the function of  $t$ ,  $f(t)$  whose Laplace transform gives  $F(s)$ , do the following. Find the sum of the residues of  $e^{st}F(z)$  for  $\operatorname{Re} z < \gamma$  where all the poles of  $F(z)$  have real part less than  $\gamma$ . This yields the midpoint of the jump of  $f(t)$  at each  $t$  where  $f$  is Holder continuous from the left and right. (Note there are no jumps by Corollary 37.5.2 so if  $f$  is Holder continuous at every point, then  $f(t)$  is recovered.)

**Example 37.5.4** Suppose  $F(s) = \frac{s}{(s^2+1)^2}$ . Find  $f(t)$  such that  $F(s)$  is the Laplace transform of  $f(t)$ .

There are two residues of this function, one at  $i$  and one at  $-i$ . At both of these points the poles are of order two and so we find the residue at  $i$  by

$$\operatorname{res}(f, i) = \lim_{s \rightarrow i} \frac{d}{ds} \left( \frac{e^{ts} s (s-i)^2}{(s^2+1)^2} \right) = \frac{-ite^{it}}{4}$$

and the residue at  $-i$  is

$$\operatorname{res}(f, -i) = \lim_{s \rightarrow -i} \frac{d}{ds} \left( \frac{e^{ts} s (s+i)^2}{(s^2+1)^2} \right) = \frac{ite^{-it}}{4}$$

From the above procedure, the function  $f(t)$  is the sum of these.

$$\begin{aligned} \frac{ite^{-it}}{4} + \frac{-ite^{it}}{4} &= \frac{1}{4}it(e^{-it} - e^{it}) \\ &= \frac{1}{4}it(\cos(t) - i\sin t - (\cos t + i\sin t)) \\ &= \frac{1}{2}t \sin t \end{aligned}$$

You should verify that this actually works giving  $\mathcal{L}(f) = \frac{s}{(s^2+1)^2}$ .

**Example 37.5.5** Find  $f(t)$  if  $F(s)$ , the Laplace transform is  $e^{-s}/s$ .

You need to compute the residues of  $\frac{e^{st}e^{-s}}{s}$ . The function equals

$$\frac{1}{s} \sum_{k=0}^{\infty} \frac{(-1)^k (t-1)^k s^k}{k!}.$$

Thus the residue is 1. However, this fails to be the function whose Laplace transform is  $F(s)$ . What is wrong? The problem with this is the failure of the estimate on  $F(s)$  to hold for large  $s$ . Indeed, if  $s = -n$ , you would have  $e^n/n$  but it would need to be less than  $C/n^\alpha$  which is not possible. The estimate requires  $F(s) \rightarrow 0$  as  $|s| \rightarrow \infty$  and this does not happen here. You can verify directly that the function which works is  $u_1(t)$  which is 0 for  $t < 1$  and 1 for  $t \geq 1$ . Thus if the estimate does not hold, the procedure does not necessarily hold either.

If  $\operatorname{Re} p < \gamma$  for all  $p$  a pole of  $F(s)$  and if  $F(s)$  is meromorphic and satisfies the growth condition 37.8, and if  $f(t)$  is defined by that Bromwich integral, is it true that  $F(s)$  is the Laplace transform of  $f(t)$  for large  $s$ ? Thus

$$f(t) \equiv \lim_{R \rightarrow \infty} \frac{1}{2\pi} \int_{-R}^R e^{(\gamma+iy)t} F(\gamma+iy) dy = \frac{1}{2\pi i} \int_{\gamma-i\infty}^{\gamma+i\infty} e^{zt} F(z) dz$$

The limit must exist because, as discussed above,

$$\lim_{R \rightarrow \infty} \frac{1}{2\pi i} \left( \int_{\gamma-iR}^{\gamma+iR} e^{zt} F(z) dz + \int_{C_R} e^{zt} F(z) dz \right)$$

is eventually constant because the contour will have enclosed all poles of  $F(z)$ , but as  $R$  continues to increase, the integral over the curved part  $C_R$  converges to 0. Let  $\operatorname{Re} s$  be larger than  $\gamma$ . One needs to consider

$$\begin{aligned} \mathcal{L}(f)(s) &= \int_0^\infty e^{-st} \frac{1}{2\pi i} \int_{\gamma-i\infty}^{\gamma+i\infty} e^{zt} F(z) dz dt \\ &= \frac{1}{2\pi i} \int_0^\infty e^{-st} \lim_{R \rightarrow \infty} \int_{\eta_R} e^{zt} F(z) dz dt \end{aligned}$$

This equals

$$\lim_{r \rightarrow \infty} \frac{1}{2\pi i} \int_0^r e^{-st} \lim_{R \rightarrow \infty} \int_{\eta_R} e^{zt} F(z) dz dt$$

Eventually, for all  $R$  large enough, the contour includes all of the finitely many poles of  $F(z)$ . There are only finitely many poles because of the estimate on  $F(z)$ . Thus we can pick  $R$  large enough that the limit on the inside equals the contour integral. Thus

$$\begin{aligned} \mathcal{L}(f)(s) &= \frac{1}{2\pi i} \int_0^\infty e^{-st} \int_{\eta_R} e^{zt} F(z) dz dt \\ &= \lim_{r \rightarrow \infty} \frac{1}{2\pi i} \int_0^r e^{-st} \int_{\eta_R} e^{zt} F(z) dz dt \end{aligned}$$

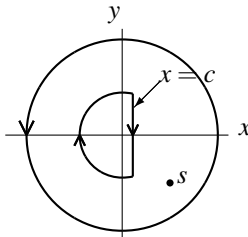
Interchanging the two contour integrals, Theorem 35.3.8,

$$\begin{aligned} &= \lim_{r \rightarrow \infty} \frac{1}{2\pi i} \int_{\eta_R} \int_0^r e^{-(s-z)t} F(z) dz dt \\ &= \lim_{r \rightarrow \infty} \frac{1}{2\pi i} \int_{\eta_R} \left( \frac{1}{s-z} - \frac{e^{-(s-z)r}}{s-z} \right) F(z) dz \\ &= \frac{1}{2\pi i} \int_{\eta_R} \frac{F(z)}{s-z} dz \end{aligned}$$

Now this contour integral is not zero because  $F(z)$  is not analytic on the inside of  $\eta_R^*$ . Let the orientation of  $\eta_R$  be switched and call the new contour  $\hat{\eta}_R$ . Then

$$\mathcal{L}(f)(s) = \frac{1}{2\pi i} \int_{\hat{\eta}_R} \frac{F(z)}{z-s} dz$$

Is this equal to  $F(s)$ ? Consider a large circular contour of radius  $M$  where  $M > |s|$  and orient it counter clockwise about  $s$  as shown in the following picture. Denote this oriented curve as  $\eta_M$ .



Then from the estimate assumed on  $F$ ,

$$\left| \int_{\eta_M} \frac{F(z)}{z-s} dz \right| \leq \frac{C}{M^\alpha} \frac{1}{M-|s|} 2\pi M$$

Now as  $M \rightarrow \infty$ , this converges to 0. Therefore, from the usual Cauchy integral formula,

$$F(s) = \frac{1}{2\pi i} \left( \int_{\eta_R} \frac{F(z)}{z-s} dz + \int_{\eta_M} \frac{F(z)}{z-s} dz \right)$$

Now take a limit of both sides as  $M \rightarrow \infty$  and you obtain

$$F(s) = \frac{1}{2\pi i} \int_{\eta_R} \frac{F(z)}{z-s} dz = \frac{1}{2\pi i} \int_{\eta_R} \frac{F(z)}{s-z} dz$$

Thus this shows the following interesting proposition. This proposition shows conditions under which a meromorphic function is the Laplace transform of a function which happens to be given by the Bromwich integral and they are the conditions used earlier.

**Proposition 37.5.6** *If  $\operatorname{Re} p < \gamma$  for all  $p$  a pole of  $F(s)$  and if  $F(s)$  is meromorphic and satisfies the growth condition 37.8, and if  $f(t)$  is defined by the Bromwich integral, then  $F(s)$  is the Laplace transform of  $f(t)$  for large  $s$ .*

## 37.6 Exercises

1. Let  $F(s) = \frac{2}{(s-1)^2+4}$  so it is the Laplace transform of some  $f(t)$ . Use the method of residues to determine  $f(t)$ .
2. This problem is about finding the fundamental matrix for a system of ordinary differential equations

$$\Phi'(t) = A\Phi(t), \quad \Phi(0) = I$$

having constant coefficients. Here  $A$  is an  $n \times n$  matrix and  $I$  is the identity matrix. A matrix,  $\Phi(t)$  satisfying the above is called a fundamental matrix for  $A$ . In the following,  $s$  will be large, larger than the magnitude of all poles of  $(sI - A)^{-1}$ .

- (a) Show that  $\mathcal{L} \left( \int_0^{(\cdot)} f(u) du \right) (s) = \frac{1}{s} F(s)$  where  $F(s) \equiv \mathcal{L}(f)(s)$
- (b) Show that  $\mathcal{L}(I) = \frac{1}{s} I$  where  $I$  is the identity matrix.

- (c) Show that there exists an  $n \times n$  matrix  $\Phi(t)$  such that  $\mathcal{L}(\Phi)(s) = (sI - A)^{-1}$ .

**Hint:** From linear algebra

$$\left((sI - A)^{-1}\right)_{ij} = \frac{\text{cof}(sI - A)_{ji}}{\det(sI - A)}$$

Show that the  $ij^{\text{th}}$  entry of  $(sI - A)^{-1}$  satisfies the conditions of Proposition 37.5.6 and so there exists  $\Phi(t)$  such that  $\mathcal{L}(\Phi)(s) = (sI - A)^{-1}$ . By Corollary 37.5.2, this  $t \rightarrow \Phi(t)$  is continuous.

- (d) Thus  $(sI - A)\mathcal{L}(\Phi)(s) = I$ . Then explain why  $(I - \frac{1}{s}A)\mathcal{L}(\Phi)(s) = \frac{1}{s}I = \mathcal{L}(I)$  and

$$\begin{aligned}\mathcal{L}(\Phi)(s) - \frac{1}{s}\mathcal{L}(A\Phi)(s) &= \mathcal{L}(I) \\ \mathcal{L}(\Phi) - \mathcal{L}\left(\int_0^{(\cdot)} A\Phi(u) du\right) &= \mathcal{L}(I)\end{aligned}$$

so

$$\Phi(t) - \int_0^t A\Phi(u) du = I$$

and so  $\Phi$  is a fundamental matrix.

- (e) Next explain why  $\Phi$  must be unique by showing that if  $\Phi(t)$  is a fundamental matrix, then its Laplace transform must be  $(sI - A)^{-1}$  and use the theorem which says that if the two continuous functions have the same Laplace transform, then they are the same function.
3. In the situation of the above problem, show that there is one and only one solution to the initial value problem

$$x'(t) = Ax(t) + f(t), x(0) = x_0, t \geq 0$$

and it is given by

$$x(t) = \Phi(t)x_0 + \int_0^t \Phi(t-u)f(u) du$$

**Hint:** Verify that  $\mathcal{L}\left(\int_0^{(\cdot)} \Phi(t-u)f(u) du\right)(s) = \mathcal{L}(\Phi)(s)\mathcal{L}(f)(s)$ . Thus if  $x$  is given by the variation of constants formula just listed, then

$$\begin{aligned}\mathcal{L}(x)(s) &= (sI - A)^{-1}x_0 + (sI - A)^{-1}\mathcal{L}(f)(s) \\ (sI - A)\mathcal{L}(x)(s) &= x_0 + \mathcal{L}(f)\end{aligned}$$

Now divide by  $s$  and verify  $x(t) = x_0 + \int_0^t Ax(u) du + \int_0^t f(u) du$ . You could also simply differentiate the variation of constants formula using chain rule and verify it works.



**Part IV**

**Probability and Statistics**





# Chapter 38

## Probability

### 38.1 Improper Integrals

If  $f$  is Riemann integrable on  $[0, R]$  for each  $R$ , then

$$\int_0^\infty f(x) dx \equiv \lim_{R \rightarrow \infty} \int_0^R f(x) dx$$

if this limit exists. Otherwise the improper integral is not defined. If  $f$  is only Riemann integrable on  $[\delta, R]$  for each  $\delta < R$ , then

$$\int_0^\infty f(x) dx \equiv \lim_{(\delta, R) \rightarrow (0, \infty)} \int_\delta^R f(x) dx$$

provided this limit exists. This expression means: There exists  $I \equiv \int_0^\infty f(x) dx$  such that for each  $\varepsilon > 0$  there is  $R_0$  and  $\delta_0$  such that if  $\delta < \delta_0$  and  $R > R_0$ , then

$$\left| \int_\delta^R f(x) dx - I \right| < \varepsilon$$

Otherwise we don't give a definition of the improper integral. Integrals of the form  $\int_{-\infty}^0 f(x) dx$  are defined similarly. As to  $\int_{-\infty}^\infty f(x) dx$ , it equals

$$\int_0^\infty f(x) dx + \int_{-\infty}^0 f(x) dx$$

provided these last two exist. As an application of polar coordinates, here is an important theorem.

**Theorem 38.1.1**  $\int_0^\infty e^{-x^2} dx = \frac{1}{2}\sqrt{\pi}$  and  $\int_{-\infty}^\infty e^{-x^2} dx = \sqrt{\pi}$ .

**Proof:** Let  $I_R \equiv \int_0^R e^{-x^2} dx$ . Then  $I_R I_R = \int_0^R \int_0^R e^{-x^2} e^{-y^2} dx dy$ . Also

$$I \equiv \lim_{R \rightarrow \infty} I_R$$

also exists. This is left as an exercise. Let  $D_R$  be the quarter circle centered at  $(0,0)$  with radius  $R$ . Then using polar coordinates to write  $\int_{D_R} e^{-(x^2+y^2)} dx$ ,

$$I_R^2 = \int_0^R \int_0^{\pi/2} e^{-r^2} r d\theta dr + \int_0^R \int_{\sqrt{R^2-x^2}}^R e^{-(x^2+y^2)} dy dx$$

That second integral satisfies

$$\begin{aligned} 0 &\leq \int_0^R \int_{\sqrt{R^2-x^2}}^R e^{-(x^2+y^2)} dy dx \leq \int_0^R \int_{\sqrt{R^2-x^2}}^R e^{-(x^2+R^2-x^2)} dy dx \\ &\leq \int_0^R \int_0^R e^{-R^2} dy dx = R^2 e^{-R^2} \end{aligned}$$

which converges to 0 as  $R \rightarrow \infty$ . Therefore,

$$I^2 = \lim_{R \rightarrow \infty} \int_0^R \int_0^{\pi/2} e^{-r^2} r d\theta dr = \lim_{R \rightarrow \infty} \frac{\pi}{2} \frac{e^{-r^2}}{2} \Big|_0^R = \frac{\pi}{4}$$

and so  $I = \frac{\sqrt{\pi}}{2}$ . Then the other integral is obviously equal to  $\sqrt{\pi}$ . ■

An alternative way to establish this integral is as follows.

$$\begin{aligned} F(x) &\equiv \left( \int_0^x e^{-t^2} dt \right)^2, F'(x) = 2 \int_0^x e^{-t^2} dt e^{-x^2} \\ &= 2x \int_0^1 e^{-x^2 t^2} dt e^{-x^2} = \int_0^1 2x e^{-x^2(t^2+1)} dt \end{aligned}$$

$$\begin{aligned} F(x) &= \int_0^x \int_0^1 2y e^{-y^2(t^2+1)} dt dy = \int_0^1 \int_0^x 2y e^{-y^2(t^2+1)} dy dt \\ &= \int_0^1 \left( -\frac{e^{-y^2(t^2+1)}}{t^2+1} \Big|_0^x \right) \\ &= \int_0^1 \left( \frac{1}{1+t^2} - \frac{e^{-x^2(t^2+1)}}{t^2+1} \right) dt = \frac{\pi}{4} - e(x) \end{aligned}$$

where  $|e(x)| < e^{-x^2}$ . It follows on taking a limit that  $\left( \int_0^\infty e^{-t^2} dt \right)^2 = \frac{\pi}{4}$ .

**Corollary 38.1.2**  $\Gamma(1/2) = \sqrt{\pi}$

**Proof:** By definition it is  $\int_0^\infty e^{-t} t^{-1/2} dt$ . Let  $t = u^2$  so  $dt = 2u du$ . Then, changing the variables,

$$\Gamma(1/2) = \int_0^\infty e^{-u^2} u^{-1} 2u du = 2 \int_0^\infty e^{-u^2} du = 2 \frac{1}{2} \sqrt{\pi} = \sqrt{\pi} \quad \blacksquare$$

## 38.2 Combinations

The fundamental problem is to find the number of ways of selecting a subset of  $k \leq n$  elements from a set having  $n$  elements. For example, consider the set  $S = \{1, 2, 3\}$ . How many subsets having two elements are there? In this case, you can simply list them. Here they are

$$\{1, 2\}, \{1, 3\}, \{2, 3\}$$

This seems easy enough, but what if you had a set of 52 things like a deck of cards and you wanted the number of ways of picking a set of 5 things from it. Then it would be a little harder. Here is some standard notation.

**Definition 38.2.1** Let  $0 \leq k \leq n$ . Then  $\binom{n}{k}$  denotes the number of subsets of a set having  $n$  elements which have  $k$  elements.

Here are some obvious assertions.

$$\binom{n}{0} = \binom{n}{n} = 1, \binom{n}{1} = n \quad (38.1)$$

The first says there is one subset which has no elements in it. Of course it is the empty set. The next says there is one subset of a set having  $n$  things which has  $n$  things in it. Of course, this would be the whole set itself. The last says there are  $n$  subsets which have a single element of the set in them. Now to get a formula for  $\binom{n}{k}$ , here is a lemma.

**Lemma 38.2.2** Let  $n$  be a positive integer and let  $1 \leq k \leq n$ . Then

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$$

**Proof:** Letting  $1 \leq k \leq n$ , suppose your set of  $n+1$  things is

$$\{a_1, \dots, a_n, a_{n+1}\}$$

Here  $a_i$  denotes the  $i^{\text{th}}$  element of the set and this is just a list of the elements of the set. Then there are two ways to select a set of  $k$  things from this set depending on whether  $a_{n+1}$  is in the set of  $k$  things. If it is, there are exactly  $\binom{n}{k-1}$  ways to obtain such a set of  $k$  things because it must be the number of ways of selecting the remaining  $k-1$  elements from the first  $n$  elements in the set. The other case is where all  $k$  elements are selected from the first  $n$  elements of the set. By definition, there are  $\binom{n}{k}$  ways to do this. Thus

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$$

■

**Definition 38.2.3** Let  $0! \equiv 1$  and for  $n \in \mathbb{N}$ ,  $n! \equiv n(n-1)(n-2) \cdots 1$ . This is called the factorial symbol. We say  $n!$  as  $n$  factorial.

With this definition, it is easy to give a simple description of  $\binom{n}{k}$ .

**Theorem 38.2.4** Let  $0 \leq k \leq n$ . Then

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

**Proof:** You see easily this is true if  $n = 1$ . In this case, the only possibilities for  $k$  are 0, 1 the the formula gives the right answer in either of these cases. Assume the formula holds for  $n$ . Then by Lemma 38.2.2 and the induction hypothesis, if  $1 \leq k \leq n$

$$\begin{aligned} \binom{n+1}{k} &= \binom{n}{k-1} + \binom{n}{k} \\ &= \frac{n!}{(k-1)!(n-k+1)!} + \frac{n!}{k!(n-k)!} \\ &= \frac{kn!}{k!(n-k+1)!} + \frac{n!(n-k+1)}{k!(n-k)!(n-k+1)} \\ &= \frac{kn!}{k!(n-k+1)!} + \frac{(n-k+1)n!}{k!(n-k+1)!} = \frac{(n+1)n!}{k!(n+1-k)!} = \frac{(n+1)!}{k!(n+1-k)!} \end{aligned}$$

and so this proves the formula in the case that  $1 \leq k \leq n$ . If  $k = 0$  or  $n+1$ , the definition of the factorial symbol and the obvious observations of 38.1 shows the formula holds in these cases also. ■

Notice that

$$\binom{n}{k} = \binom{n}{n-k}.$$

### 38.3 The Binomial Theorem

The Binomial theorem is one of the most useful and fundamental theorems in algebra. It is easy to prove from the above using induction. Here it is.

**Theorem 38.3.1** Let  $n \in \mathbb{N}$ . Then

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

**Proof:** In case  $n = 1$ , both sides reduce to  $a+b$  so it works in this case. Suppose now it works for  $n$ . Then by induction,

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

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

By Lemma 38.2.2 this reduces to

$$\begin{aligned}
&a^{n+1} + \sum_{k=1}^n \binom{n+1}{k} a^{n+1-k} b^k + b^{n+1} \\
&= \sum_{k=0}^{n+1} \binom{n+1}{k} a^{n+1-k} b^k
\end{aligned}$$

which shows that when the formula holds for  $n$  it also holds for  $n+1$ . ■

Another way to verify this important formula is as follows. For  $n$  a positive integer  $(a+b)^n$  must be of the form  $(a+b)(a+b)\cdots(a+b)$  and it must consist of a sum of terms of the form  $a^k b^{n-k}$ . How many are there for a given  $k$ ? This involves the number of ways to pick  $k$  factors in the product which contribute  $a$  and the remaining factors contributing  $b$ . Thus the coefficient of this term is  $\binom{n}{k}$ . As to the case where  $k=0$ , this means all factors contribute  $b$  and so there is only one way to obtain this term  $a^0 b^n$  and this is  $\binom{n}{0}$ . Thus the above product of terms reduces to

$$\sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$$

**Example 38.3.2** Find the coefficient which multiplies  $x^3 y^6$  in  $(x+y)^9$ .

By the binomial theorem, this is

$$\binom{9}{3} = \frac{9!}{3!6!} = 84$$

Thus  $(x+y)^9$  is the sum of terms  $c_k x^k y^{9-k}$  and the  $c_k$  which corresponds to  $k=3$  is 84.

**Example 38.3.3** Find the constant coefficient of  $(2x+3x^{-3})^8$ .

You have that this is the sum of constants times  $x^{8-k} (x^{-3})^k$  and so you need to have  $8-k-3k=0$  so  $k=2$ . It follows that this term is of the form

$$\binom{8}{2} (2x)^6 (3x^{-3})^2 = \frac{8!}{2!6!} 2^6 3^2 = 16128$$

### 38.4 Exercises

1. Use the binomial theorem to expand or simplify the following.

(a)  $(x+y)^5$

(b)  $(x-y)^5$

(c)  $(x-y)^4$

(d)  $(x+h)^3 - x^3$

(e)  $(x+h)^4 - x^4$

(f)  $h^{-1} \left( (x+h)^5 - x^5 \right)$

(g)  $h^{-1} \left( (x+h)^6 - x^6 \right)$

2. Show that for a positive integer and  $x > 0$ ,  $(1+x)^n \geq 1+nx$ .

3. Show that  $\sum_{k=0}^n \binom{n}{k} = 2^n$ .

4. Approximate  $100(1.005)^{12}$ . This would be the amount in the bank after one year if interest is 6% compounded monthly.

5. Show that for  $k \geq 1$ ,

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

### 38.5 Counting and Basic Probability

You do an experiment  $n$  times and there are two possible outcomes to this experiment each time it occurs, a “success” having probability  $p$ , a positive number less than 1 and a “failure” having probability  $(1-p)$ . For example, you could have  $p$  be the probability of getting a 4 when you roll a pair of fair dice. What would this probability be? Here is a table of possible outcomes for the pair of dice.

|     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|
| 1,1 | 1,2 | 1,3 | 1,4 | 1,5 | 1,6 |
| 2,1 | 2,2 | 2,3 | 2,4 | 2,5 | 2,6 |
| 3,1 | 3,2 | 3,3 | 3,4 | 3,5 | 3,6 |
| 4,1 | 4,2 | 4,3 | 4,4 | 4,5 | 4,6 |
| 5,1 | 5,2 | 5,3 | 5,4 | 5,5 | 5,6 |
| 6,1 | 6,2 | 6,3 | 6,4 | 6,5 | 6,6 |

The first number represents the one on the first die and the second represents the number on the second die. (die is singular for dice) How many ways are there to get a 4? From the table, there are exactly 3 ways,  $(3,1)$ ,  $(2,2)$ ,  $(1,3)$ . How many possible outcomes are

there? There are 36. Thus if every outcome is as likely as any other, the probability of rolling a 4 is  $3/36$  or  $1/12$ .

Now in a succession of rolls of the dice, the probability of a particular outcome on roll  $k$  is not affected by what happened on earlier rolls of the dice. Each time the dice are rolled, the probability of rolling a four is  $1/12$  and the probability rolling a non four is  $11/12$ .

What is the probability of rolling a 5 twice in a row? In this case there would be  $36^2$  possible outcomes and only  $4^2$  of them are favorable to rolling two fives in succession. (Four possibilities for the first roll of the dice and for each of these, four for the second.) Thus the probability of this occurring is

$$\frac{4^2}{36^2} = \frac{1}{81}$$

What about the probability of a five on the first roll and a non five on the second? This probability is

$$\frac{4}{36} \cdot \frac{32}{36} = \frac{8}{81}.$$

You can determine this the same way by counting the ways favorable to the desired outcome and dividing this by the number of possible outcomes.

$$\frac{4 \cdot 32}{36^2} = \frac{8}{81}$$

Similarly, the probability of rolling a non five followed by a five would be

$$\frac{32}{36} \cdot \frac{4}{36} = \frac{8}{81}$$

More generally, the probability of getting  $k$  fives and  $n - k$  non fives in a particular order would be

$$\left(\frac{4}{36}\right)^k \left(\frac{32}{36}\right)^{n-k}.$$

More generally, you have a situation where the probability of  $k$  success with probability  $p$  and  $(n - k)$  failures happening with probability  $q \equiv (1 - p)$  in any particular order is  $p^k q^{n-k}$ . What is the probability of having  $k$  successes in  $n$  trials? This is known as the binomial distribution. How many ways can  $k$  success happen in  $n$  trials? It can happen

exactly the number of ways there are of selecting  $k$  of the  $n$  trials. There are  $\binom{n}{k}$  ways

for this to happen. Therefore, since each of these has the same probability,  $p^k q^{n-k}$ , the probability of  $k$  successes in  $n$  trials is

$$\binom{n}{k} p^k q^{n-k}$$

This motivates the following definition of the binomial distribution and the idea of a random variable.

**Definition 38.5.1** Define a “random variable”  $X$  to be the number of successes, each having probability  $p$  in  $n$  trials. Thus  $X$  has values  $0, 1, \dots, n$ . If the probability that  $X$  has value  $k$ , written

$$P(X = k)$$

is given by

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

then  $X$  is said to have a binomial distribution.

Note that the sum

$$\sum_{k=0}^n P(X = k)$$

needs to equal 1 because the random variable must achieve one of the numbers  $0, 1, 2, \dots, n$ . This occurs by the binomial theorem,

$$\sum_{k=0}^n P(X = k) = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = (p + q)^n = 1^n = 1.$$

There is a general principle of counting which should be mentioned. Suppose you have  $m$  “positions” and  $n$  different things. How many ways are there to fill the  $m$  positions with the  $n$  things? There are  $n$  choices for the first, and having filled this position, there are  $n - 1$  left to place in the second. Thus the number of ways to fill the first two positions is  $n(n - 1)$ . Then, having filled these two, there are now  $n - 2$  things left to place in the third position and so there are  $n(n - 1)(n - 2)$  ways to fill the first three of these positions. Continue this way till you run out of positions to fill. How many ways of filling them do you obtain? You see that there are  $n(n - 1)(n - 2) \cdots (n - m + 1)$  ways to do it. This is called permutations of  $n$  things taken  $m$  at a time. See the exercise below.

**Example 38.5.2** *In a class of 12 students who are arranged in three rows of four students, what is the probability that the particular four students, Eliphaz, Elihu, Zophar, and Bildad will occupy the front four seats?*

There are  $4!$  ways for them to occupy these four seats in some order. There are  $12 \cdot 11 \cdot 10 \cdot 9$  ways to fill these seats in some order. Therefore, the probability is

$$\frac{4!}{12(11)(10)(9)} = \frac{1}{495}$$

Of course, you don’t care about order in this problem so you could also do this in terms of combinations of  $n$  things taken  $m$  at a time.

$$\frac{1}{\binom{12!}{4!8!}} = \frac{1}{495}$$

There is exactly one way to select these four students for the first four seats and then there are  $12! / (4!8!)$  ways to fill these seats.

**Example 38.5.3** *In the above example involving 12 students, it is absolutely necessary for disciplinary reasons that Eliphaz must not sit next to Elihu. If the students file in and sit down randomly, what is the probability that Eliphaz ends up on the front right seat when viewed by the teacher and is not sitting next to Elihu?*



There is one way to fill the front right seat with Eliphaz. Then there are 10 favorable ways to fill the seat on the left side of Eliphaz with someone other than Eliphaz. There are now 10 students left who can fill the remaining seats in any order because you have used two. Thus there are  $1 \times 10 \times 10 \times 9$  ways to have a favorable outcome. There are  $12 \times 11 \times 10 \times 9$  ways for them to select seats at random. Therefore, the probability is

$$\frac{1 \times 10 \times 10 \times 9}{12 \times 11 \times 10 \times 9} = \frac{5}{66}$$

## 38.6 Exercises

1. Let  $k \leq n$  where  $k$  and  $n$  are natural numbers.  $P(n, k)$ , permutations of  $n$  things taken  $k$  at a time, is defined to be the number of different ways to form an ordered list of  $k$  of the numbers  $\{1, 2, \dots, n\}$ . Show

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

2. Now consider the word “mississippi”. By rearranging the letters, how many distinctly different words can you obtain? Note that for each list of these letters the four different s are indistinguishable. There are therefore,  $4!$  ways which are not really different.
3. Using Problem 1, show the number of ways of selecting a set of  $k$  things from a set of  $n$  things is  $\frac{n!}{(n-k)!k!}$ .
4. Prove by induction that  $n < 2^n$  for all natural numbers  $n \geq 1$ .
5. Prove by the binomial theorem and Problem 3 that the number of subsets of a given finite set containing  $n$  elements is  $2^n$ .
6. Show that for  $p \in (0, 1)$ ,  $\sum_{k=0}^n \binom{n}{k} k p^k (1-p)^{n-k} = np$ .
7. 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,

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

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.

8. Let  $n$  be a natural number and let  $k_1 + k_2 + \cdots + k_r = n$  where  $k_i$  is a non negative integer. The symbol

$$\binom{n}{k_1 k_2 \cdots k_r}$$

denotes the number of ways of selecting  $r$  subsets of  $\{1, \dots, n\}$  which contain  $k_1, k_2, \dots, k_r$  elements in them. Find a formula for this number.

9. Is it ever the case that  $(a+b)^n = a^n + b^n$  for  $a$  and  $b$  positive real numbers?
10. Is it ever the case that  $\sqrt{a^2 + b^2} = a + b$  for  $a$  and  $b$  positive real numbers?
11. Is it ever the case that  $\frac{1}{x+y} = \frac{1}{x} + \frac{1}{y}$  for  $x$  and  $y$  positive real numbers?
12. Derive a formula for the multinomial expansion,  $(\sum_{k=1}^p a_k)^n$  which is analogous to the binomial expansion. **Hint:** See Problem 8.
13. Let  $X$  be a binomial random variable. Thus  $P(X=k) = \binom{n}{k} p^k q^{n-k}$  where  $p$  is the probability of success and  $q = 1 - p$  is the probability of failure. The expected value of  $X$  denoted as  $E(X)$ , is defined as

$$\sum_{k=0}^n k P(X=k).$$

Show the expected value of  $X$  equals  $np$ .

14. The variance of the random variable in the above problem is defined as

$$\sigma^2 \equiv \sum_{k=0}^n (k - E(X))^2 P(X=k)$$

Find  $\sigma^2$ . You should get  $npq$ .

15. Find the probability of drawing from a shuffled deck of playing cards four hearts. **Hint:** Use principles of counting to find the number of ways of drawing four hearts. There are 13 of these. Now how many ways can you pull out four of them? Then note there are 52 cards in all. How many ways can you pull out four cards from these.
16. Find the probability of obtaining 2 clubs and three spades from a shuffled deck of cards.
17. A pond has  $N$  fish and 120 of these are marked fish. What is the probability in terms of  $N$  of catching 10 fish, two of which are marked and 8 of which are unmarked?
18. Show that in general, for  $k \leq m < N$

$$\sum_{j=0}^k \frac{\binom{m}{j} \binom{N-m}{k-j}}{\binom{N}{k}} = 1$$

If  $X$  is a random variable having values in  $\{0, 1, \dots, k\}$  such that the probability that  $X = j$  is given by the  $j^{\text{th}}$  term of the above sum, then  $X$  is said to have a hypergeometric distribution. Much much more can be said about this topic. **Hint:** If you pick  $k$  things from  $N$  things  $m$  of which are marked and  $N - m$  unmarked, there are various ways to do it determined by the value of  $j$ , the number of marked things out of your sample of  $k$  things.

19. Suppose a pair of dice has one blue and the other one red. What is the probability that when they are rolled the blue die delivers a strictly larger number than the red die? Now what is the probability that either this happened **or** a 6 is rolled? What is the probability that the blue is greater than the red **and** a 6 is rolled?
20. Recall the following table illustrating the possible outcomes of rolling a pair of dice.

|      |      |      |      |      |      |
|------|------|------|------|------|------|
| 1, 1 | 1, 2 | 1, 3 | 1, 4 | 1, 5 | 1, 6 |
| 2, 1 | 2, 2 | 2, 3 | 2, 4 | 2, 5 | 2, 6 |
| 3, 1 | 3, 2 | 3, 3 | 3, 4 | 3, 5 | 3, 6 |
| 4, 1 | 4, 2 | 4, 3 | 4, 4 | 4, 5 | 4, 6 |
| 5, 1 | 5, 2 | 5, 3 | 5, 4 | 5, 5 | 5, 6 |
| 6, 1 | 6, 2 | 6, 3 | 6, 4 | 6, 5 | 6, 6 |

Find the probability that you roll a 7 before you roll either a 3 or an 11. **Hint:** It can happen in infinitely many distinct ways. You don't roll either a 3 or an 11 for  $k$  rolls and then on the  $k^{\text{th}}$  roll you get a 7. Here  $k = 0, 1, 2, 3, \dots$  so you need to take a limit of the partial sums associated with the different values of  $k$  and then take a limit. So what is the probability of getting a 7 on try  $k + 1$  and not getting either a 3 or an 11 before this? Argue it is  $\left(\frac{13}{18}\right)^k \frac{1}{6}$ .

21. Explain why in general,  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  where  $A, B$  are two events such as in the above problem having the blue be larger than the red die or rolling a 6.
22. Let  $X$  be the random variable which gives the number of heads when you flip a coin 6 times. Which value of  $X$  has the highest probability? What is the expected value of  $X$ ? What is the variance of  $X$ . For these last parts, see Problem 13 and 14 above.
23. You have a class of 12 students who will be seated in four rows consisting of three students in each row. Jeroboam, Nadab, Baasha, and Elah must sit in the front for disciplinary reasons. Also, you absolutely must not have Baasha sitting next to Nadab because Baasha is a thug who will attack Nadab. If Baasha is to sit on the front left seat as viewed by the teacher, what is the probability that an acceptable outcome will occur if the students take their seats completely at random?

## 38.7 General Considerations Probability

As mentioned earlier,  $X$  is a random variable if a probability is associated with  $X$  being found in some set of possible values. The following examples have been discussed either in the chapter or in the exercises. These are examples of discrete random variables because the

random variable takes values in some subset of the integers. The following two examples consider situations where  $X$  can only take finitely many values.

**Example 38.7.1** *Let an experiment be performed  $n$  times. Each time the experiment is performed, the probability of a “success” is  $p$  and the probability of a “failure” is  $q$ ,  $p + q = 1$ . Then let  $X$  be the number of successes in the  $n$  experiments. The probability that  $X = k$ ,  $P(X = k)$  is*

$$\binom{n}{k} p^k q^{n-k}$$

*A distribution of this sort is called a binomial distribution.*

**Example 38.7.2** *Let  $k \leq m < N$ . If  $X$  is a random variable such  $P(X = j)$ ,  $j \leq k$ , is given by*

$$P(X = j) \equiv \frac{\binom{m}{j} \binom{N-m}{k-j}}{\binom{N}{k}}$$

*this is called a hypergeometric distribution. This is when you have  $m$  marked fish and you take a sample of  $k$  fish. Then  $X$  is the number of marked fish you get in your sample of  $k$  fish. The probability it equals  $j$  is given by the above. Thus as explained in Problem 18 on Page 778,*

$$\sum_{j=0}^k \frac{\binom{m}{j} \binom{N-m}{k-j}}{\binom{N}{k}} = 1$$

There are  $\binom{m}{j} \binom{N-m}{k-j}$  ways to get exactly  $j$  marked fish from a sample of  $k$  fish. You have  $\binom{m}{j}$  ways to get  $j$  marked fish from the set of  $m$  marked fish and for each of these, there are exactly  $\binom{N-m}{k-j}$  ways to fill the set of  $k$  fish with non marked fish. Thus

$$\sum_{j=0}^k \binom{m}{j} \binom{N-m}{k-j} = \binom{N}{k}$$

where the last is the total number of ways of selecting  $k$  fish from the  $N$  fish. Thus the above claim is verified.

Now sometimes a random variable can take values from the set of all nonnegative integers. Suppose you have a binomial distribution in which the probability of a success is extremely small and the number of trials is very large. Say  $pn = \lambda$  where  $n$  is large. Then the probability of success in the  $n$  trials is

$$P(X = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Thus, as  $n$  gets increasingly large and  $p$  correspondingly small,

$$\begin{aligned} P(X = k) &\approx \lim_{n \rightarrow \infty} \frac{n!}{k!(n-k)!} \lambda^k \frac{1}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n \rightarrow \infty} \frac{n(n-1) \cdots (n-k+1)}{k! n^k} \lambda^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

Note that

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

**Example 38.7.3** A random variable has Poisson distribution if for  $k$  a nonnegative integer,

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

The sort of thing this models is the probability of being kicked by a mule  $k$  times in some time interval of moderate length or the probability that  $k$  customers arrive at the check out of a store in some 1 minute interval.

These random variables just discussed take values in a set of integers but often the random variable takes values in the real numbers or  $\mathbb{R}^n$ . When this is the case, you must use an integral to determine the probability that the random variable is in some set. These are called continuous random variables when you use a Riemann integral to determine the probability that a random variable is in some set.

**Example 38.7.4** Let

$$f(x) \equiv \begin{cases} x/2 & \text{if } x \in [0, 2] \\ 0 & \text{if } x \notin [0, 2] \end{cases}$$

Thus  $\int_{-\infty}^{\infty} f(x) dx = 1$ . Then  $f(x)$  is a distribution function for the random variable  $X$  if  $P(X \in [a, b]) = \int_a^b f(x) dx$ . More generally, for all “suitable” sets  $F$ ,

$$P(X \in F) = \int_F f(x) dx \equiv \int \mathcal{X}_F(x) f(x) dx$$

where

$$\mathcal{X}_F(x) \equiv \begin{cases} 1 & \text{if } x \in F \\ 0 & \text{if } x \notin F \end{cases}$$

You really need the notions of measure spaces and Lebesgue integrals to do this right. Now here is some terminology.

**Definition 38.7.5** Two random variables  $X, Y$  are said to have the same distribution if for all intervals  $I$ ,

$$P(X \in I) = P(Y \in I)$$

**Example 38.7.6** Let  $\alpha > 0$  and let  $f(x) \equiv \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}$ . Then the random variable  $X$  having values in  $[0, \infty)$  has this as its distribution function if

$$\int_a^b f(x) dx = P(X \in [a, b]).$$

Note that  $\int_0^{\infty} f(x) dx = 1$  from the definition of the gamma function.

A modification of this density function gives the very important  $\mathcal{X}^2(r)$ , chi squared, distribution in the next example.

**Example 38.7.7** Let  $r$  be a positive integer. The  $\mathcal{X}^2(r)$  distribution with  $r$  degrees of freedom for a random variable having values in  $[0, \infty)$  is given by

$$f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{(r/2)-1} e^{-x/2}$$

Thus there are infinitely many of these, one for each  $r$ . I have no idea why they refer to  $r$  as “degrees of freedom”.

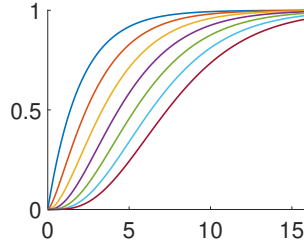
It is necessary to verify that the above is a probability density. Thus it is necessary to show that its integral is 1. This involves changing the variable. Let  $\frac{x}{2} = t$  so  $dx = 2dt$  then

$$\begin{aligned} \int_0^\infty \frac{1}{\Gamma(r/2)2^{r/2}} x^{(r/2)-1} e^{-x/2} dx &= \int_0^\infty \frac{1}{\Gamma(r/2)2^{r/2}} (2t)^{(r/2)-1} e^{-t} 2dt \\ &= \int_0^\infty \frac{1}{\Gamma(r/2)2^{r/2}} 2^{r/2} t^{(r/2)-1} e^{-t} dt = 1 \end{aligned}$$

from the definition of  $\Gamma(r/2)$ . The following picture gives the graphs of

$$F_r(x) \equiv \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} t^{(r/2)-1} e^{-t/2} dt$$

for  $r = 2, 3, \dots, 8$ . Thus  $F_r(x)$  equals  $P(X \leq x)$  where  $X$  is a  $\mathcal{X}^2(r)$  random variable.



As the number of degrees of freedom  $r$  increases, the graph becomes increasingly flat near 0. This is good. Having many “degrees of freedom” is a fine thing because this chi squared distribution can be used to estimate the variance of a normal distribution and having the graph flat near 0 ends up meaning that you can be confident in a smaller upper bound for the variance. This will be discussed more later. It turns out that having  $r$  large is associated with having a large sample size. In other words, you are considering many identically distributed random variables.

**Example 38.7.8** Let  $\mathbf{X}$  have values in  $\mathbb{R}^p$ . Then the density function of  $\mathbf{X}$  will be  $f(\mathbf{x})$  where

$$P(\mathbf{X} \in A) = \int_A f(\mathbf{x}) dV \equiv \int \mathcal{X}_A(\mathbf{x}) f(\mathbf{x}) dV$$

$$\mathcal{X}_A(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A \\ 0 & \text{if } \mathbf{x} \notin A \end{cases}$$

where  $dV$  refers to the  $p$  dimensional volume. Thus the expression on the right is an integral of a function of  $p$  variables.  $\mathbf{x} = (x_1, \dots, x_p)$ . We usually write  $dV$  as  $dx_1 dx_2 \cdots dx_p$ . Of course the case  $p = 3$  was discussed earlier and the higher dimensional case is exactly similar. When it is desired to emphasize that  $X$  has values in  $\mathbb{R}^p$  it will be referred to as a random vector and may be written in bold face.

The most important distribution is the normal distribution. It has two parameters and is given as follows.

**Example 38.7.9** Let  $\mu, \sigma > 0$  then a random variable  $X$  having values in  $\mathbb{R}$  is normally distributed if

$$P(X \in (a, b)) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

The density function is then

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

It is necessary to verify that this really is a density function. To do this, let

$$y = \frac{1}{2} \frac{x-\mu}{\sigma}$$

Then, changing the variables in

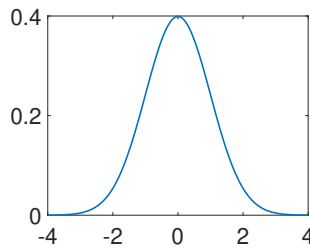
$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

yields

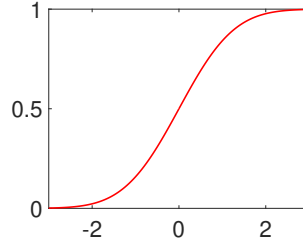
$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2} \sqrt{2} du = \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-u^2} du$$

and from Theorem 38.1.1, this equals 1.

Often one reduces to the case that  $\sigma = 1$  and  $\mu = 0$ . Thus the density is  $\frac{1}{\sqrt{2\pi}} e^{-(x^2/2)}$ . I have heard people refer to random variables with this distribution as “standard normal deviates”. Its graph is as follows.



You observe that if a random variable has this distribution defined by this probability density function, it is very likely to assume a value between  $-2$  and  $2$ . The graph of  $F(x) \equiv P(X \leq x)$  for  $X$  a normally distributed random variable with  $\mu = 0$  and  $\sigma = 1$  follows.



You can see that the probability that  $X \leq 3$  is very close to 1.

**Example 38.7.10** The multivariate normal is as follows. The random variable  $\mathbf{X}$  has values in  $\mathbb{R}^p$  and its density function is of the form

$$\frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^* \Sigma^{-1}(\mathbf{x}-\mathbf{m})}$$

Here  $\Sigma$  is the covariance matrix. It is a symmetric matrix with positive eigenvalues and  $\mathbf{m} \in \mathbb{R}^p$  is the mean. Thus

$$P(\mathbf{X} \in A) = \int_A \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^* \Sigma^{-1}(\mathbf{x}-\mathbf{m})} d\mathbf{x}$$

You integrate over the set  $A$  the density function. Just as in the case of three dimensions, this is easier said than done. However, if  $A$  has a simple form  $\prod_{k=1}^p (-\infty, a_k]$ , then  $P(\mathbf{X} \in A) =$

$$\int_{-\infty}^{a_1} \int_{-\infty}^{a_2} \cdots \int_{-\infty}^{a_p} \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^* \Sigma^{-1}(\mathbf{x}-\mathbf{m})} dx_p \cdots dx_1$$

Assuming there are no mathematical difficulties, the following is the definition of what is meant by expectation.

**Definition 38.7.11** Let  $X$  be a discrete random variable such that  $P(X = j) = f(j)$ . Then if  $g$  is some function defined on the values of  $X$ ,  $E(g(X)) \equiv \sum_j g(j) f(j)$  assuming the sum makes sense. It is called the expected value of  $g(X)$  or simply the expectation of  $g(X)$ . In case  $X$  is a continuously distributed random variable with density  $f(x)$ , the expectation of  $g(X)$  is  $E(g(X)) \equiv \int g(x) f(x) dx$ , assuming the integral makes sense.

The two cases considered above are the discrete and continuously distributed cases for random variables. However, this does not include all cases. To do this right, one needs the notion of the Lebesgue integral and measure spaces and one defines exactly what a random variable is, a measurable function defined on a measure space, instead of referring to it vaguely in terms of the probability “it” has certain values or lies in some set called an “event”. What does always happen is that, assuming everything makes sense,

$$E(aX + bY) = aE(X) + bE(Y)$$

for two random variables  $X, Y$  and scalars  $a, b$ . Also, for any random variable  $X$  it may or may not have a valid expectation, denoted as  $E(X)$  but in every case, it makes sense to speak of  $P(X \in E)$  where  $E$  is some interval or more generally something called a Borel set.



**Observation 38.7.12** *In every case, if  $a, b$  are numbers, then if everything makes sense, and  $X, \hat{X}$  are two random variables having the same probability distribution, meaning that  $P(X \in F) = P(\hat{X} \in F)$  for all  $F$  an interval, then*

$$E(aX + b\hat{X}) = aE(X) + bE(\hat{X})$$

Suppose you had many random variables  $X_i$  each having the same distribution and the collection of random variables independent, explained below. If you averaged  $X_i$  for all  $i$ , what you would get is probably close to  $E(X)$ . This is why taking the expectation is of interest. I will give a brief explanation why this is so.

Where do independent random variables come from? In practice, you have independent observations from an underlying probability space, meaning that it makes sense to ask for the probability that a random variable is in suitable subsets of  $\mathbb{R}$  or  $\mathbb{R}^n$ . These observations are independent in the sense that the outcome of an observation does not depend on the outcome of the others. Then the numerical values are called independent random variables. A more precise description is given below.

First, here is an important formula. I will be considering only the case of a continuous distribution in explaining this inequality, but it all works in general. The inequality is called the Chebychev inequality.

**Proposition 38.7.13** *Let  $X$  be a random variable. Then for  $\varepsilon > 0$ ,*

$$P(|g(X)| \geq \varepsilon) \leq \frac{1}{\varepsilon} E(|g(X)|)$$

**Proof:** By definition of what is meant by a distribution function, if  $E \equiv |g|^{-1}([\varepsilon, \infty)) \equiv \{x : |g(x)| \geq \varepsilon\}$ , then on this set,  $|g(x)|/\varepsilon \geq 1$  and off this set,  $|g(x)|f(x) \geq 0$ . Thus

$$P(|g(X)| \geq \varepsilon) = \int_E f(x) dx \leq \frac{1}{\varepsilon} \int_{\mathbb{R}} |g(x)| f(x) dx = \frac{1}{\varepsilon} E(|g(X)|) \blacksquare$$

Now suppose you have  $X_i$  a random variable having distribution function  $f(x)$  and suppose  $\mu = E(X)$ ,  $\sigma^2 = E((X - \mu)^2)$  both exist. Suppose  $X_i, i = 1, \dots$  all these random variables are independent as in the next definition.

**Definition 38.7.14** *Let there be random variables  $X_1, \dots$ , having well defined mean  $\mu \equiv E(X_k)$  and variance  $\sigma^2 = E((X - \mu)^2)$ . Then to say these are independent implies that  $E((X_i - \mu)(X_j - \mu)) = E(X_i - \mu)E(X_j - \mu) = 0$  whenever  $i \neq j$ . The more complete meaning of independence is as follows: For each  $m$ ,*

$$P(X_i \in E_i \text{ for each } i \leq m) = \prod_{i=1}^m P(X_i \in E_i).$$

Here the  $E_i$  can be considered intervals. The idea is that what happens in terms of probability involving each  $X_j$  for  $j \neq i$  does not affect probability involving  $X_i$ . Also, if you have  $X_1, \dots$ , independent, then if  $g$  is some continuous function, then  $g(X_1), \dots$  will also be independent. This is because  $g(X_i) \in E_i$  if and only if  $X_i \in g^{-1}(E_i)$ . Then there is a significant observation.

**Proposition 38.7.15** Suppose  $X_i$  for  $i = 1, \dots, m$  are independent random variables and  $E(X_i) = \mu$  while  $E((X_i - \mu)^2) = \sigma^2$ . Then if  $Z \equiv \frac{1}{m} \sum_{i=1}^m X_i$  is their average, then  $E(Z) = \mu$  and  $E((Z - \mu)^2) = \sigma^2/m$ .

**Proof:**  $E(Z) = E(\frac{1}{m} \sum_{i=1}^m X_i) = \frac{1}{m} \sum_{i=1}^m E(X_i) = \frac{1}{m} \sum_{i=1}^m \mu = \mu$ . Also, using the independence of these random variables,

$$\begin{aligned} E((Z - \mu)^2) &= E\left(\left(\frac{1}{m} \sum_{k=1}^m X_k - \mu\right)^2\right) \\ &= E\left(\left(\sum_{k=1}^m \left(\frac{X_k - \mu}{m}\right)\right)^2\right) \\ &= \frac{1}{m^2} E\left(\sum_{k,l} (X_k - \mu)(X_l - \mu)\right) \\ &= \frac{1}{m^2} \sum_{k,l} E((X_k - \mu)(X_l - \mu)) \\ &= \frac{1}{m^2} \sum_{k=1}^m E((X_k - \mu)^2) = \frac{\sigma^2}{m} \blacksquare \end{aligned}$$

Then it follows from this proposition and Proposition 38.7.13 the following important result which says that the average of observations of independent random variables which have the same mean and same variance converges in probability to 0 as more and more independent observations are taken. Actually much more can be said, but this is enough here.

**Proposition 38.7.16** Let  $X_i, i = 1, \dots, m$  be independent random variables with common mean  $\mu$  and common variance  $\sigma^2$  then for  $Z_m \equiv \frac{1}{m} \sum_{k=1}^m X_k$ , the average of the first  $m$ ,

$$\lim_{m \rightarrow \infty} P((Z_m - \mu)^2 \geq \epsilon) = 0$$

**Proof:** This follows from the above propositions which imply

$$P((Z_m - \mu)^2 \geq \epsilon) \leq \frac{1}{\epsilon} E((Z_m - \mu)^2) = \frac{1}{\epsilon m} \sigma^2 \blacksquare$$

In words, this says that if you average independent observations (That is, the  $i^{\text{th}}$  observation does not depend on the others. For example, you throw the marked fish back into the lake and let them swim around before taking another observation.) then as you take more and more of them, the probability that this average differs by very much from the true mean becomes very small. This is a version of the law of large numbers. In words, the average is probably close to the true mean if you average many independent observations.

**Example 38.7.17** Let  $X$  have the hypergeometric distribution.

$$P(X = j) = \frac{\binom{m}{j} \binom{N-m}{k-j}}{\binom{N}{k}}, k \geq 1$$

where  $j \leq k \leq m < N$ . Find a formula for  $E(X)$ .

This follows from a computation. The details are left to you.

$$\begin{aligned}
 E(X) &\equiv \sum_{j=0}^k \frac{\binom{m}{j} \binom{N-m}{k-j} j}{\binom{N}{k}} = \sum_{j=1}^k \frac{\binom{m}{j} \binom{N-m}{k-j} j}{\binom{N}{k}} \\
 &= \sum_{j=1}^k \frac{\frac{m}{j} \binom{m-1}{j-1} \binom{N-m}{k-j} j}{\binom{N}{k}} = m \sum_{j=1}^k \frac{\binom{m-1}{j-1} \binom{N-m}{k-j}}{\binom{N}{k}} \\
 &= m \sum_{j=1}^k \frac{\binom{m-1}{j-1} \binom{N-1-(m-1)}{(k-1)-(j-1)}}{\frac{N}{k} \binom{N-1}{k-1}} \\
 &= \frac{mk}{N} \sum_{j=1}^k \frac{\binom{m-1}{j-1} \binom{N-1-(m-1)}{(k-1)-(j-1)}}{\binom{N-1}{k-1}} \\
 &= \frac{mk}{N} \sum_{j=0}^{k-1} \frac{\binom{m-1}{j} \binom{N-1-(m-1)}{(k-1)-j}}{\binom{N-1}{k-1}} = \frac{mk}{N}
 \end{aligned}$$

**Example 38.7.18** *There are 100 fish in a pond and there are 30 fish which are marked. You scoop up 20 fish with a large net and then throw them back after counting the number of marked fish. If this is done repeatedly, then on average, about how many marked fish do you expect to get?*

From the formula, it would be  $\frac{(30)(20)}{100} = 6$ . Note that if you did this a lot, you could estimate how many fish are in the pond. You know there are 30 marked fish and so you know  $m, k$  therefore,  $N$  should be such that your observed average is close to  $\frac{(30)(20)}{N}$ .

The expectations of most interest are  $E(X^k)$  where  $k$  is a positive integer. These are called the moments.

## 38.8 Moment Generating Functions

A very convenient gimmick for computing moments is the moment generating function. Assuming there are no fussy mathematical issues, the moment generating function is

$$M(t) \equiv E(e^{tX})$$

Then, unless there are pathologies, you could write

$$M'(t) = E(Xe^{tX}), M''(t) = E(X^2e^{tX}), \text{ etc.}$$

Then you would simply let  $t = 0$  and find various moments. The  $k^{\text{th}}$  moment is  $E(X^k)$ . In all cases, you are using the fact that  $E$  is linear, either a sum or some sort of integral and you interchange the derivative with the sum or integral. Of course the legitimacy of this operation is in question, but in most cases of interest, there is no problem.

This will suffice for what is considered in this introduction, but the moment generating function has some deficiencies. In particular, it might not exist.

A much better approach is the characteristic function

$$\phi_X(t) \equiv E(e^{itX})$$

because it always exists. It can be shown, although it won't be attempted here, that the distribution of the random variable is completely determined by the characteristic function. However, you can see why this is so in case there is a continuous density function. Say

$$\int_{\mathbb{R}} e^{itx} f(x) dx = \int_{\mathbb{R}} e^{itx} g(x) dx$$

Then

$$\int_{\mathbb{R}} e^{itx} (f(x) - g(x)) dx = 0$$

and by the Fourier inversion theorem adapted slightly, Theorem 37.3.5,

$$f(y) - g(y) = \lim_{R \rightarrow \infty} \frac{1}{2\pi} \int_{-R}^R e^{-iyt} \int_{\mathbb{R}} e^{itx} (f(x) - g(x)) dx dt = 0$$

Thus it is not unreasonable to believe this assertion that if the two characteristic functions coincide, then the densities are the same.

However, it is less trouble to use the moment generating functions because it does not require fussing with complex numbers, and it can be shown that if two random variables have the same moment generating function, then they have the same density, although it has not been done in this book. Everything could be done just as well with the more general characteristic functions.

The following definition includes the case where  $X$  is a random vector and gives the above discussion as a special case.

**Definition 38.8.1** Let  $\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_p \end{pmatrix}$  be a random vector. The moment generating function is defined as  $E(e^{t \cdot \mathbf{X}})$ .

As mentioned above, if two random variables have the same moment generating function, then they will have the same distribution.

**Example 38.8.2** Find the moment generating function for a binomial random variable  $X$  and use to find some moments.

$$\sum_{k=0}^n \binom{n}{k} e^{tk} p^k q^{n-k} = (q + pe^t)^n = M(t)$$

Then  $M'(t) = npe^t(q + pe^t)^{n-1}$ . Then let  $t = 0$  and you get  $M'(0) = np$  which is  $E(X)$ . Also

$$M''(t) = npe^t(q + pe^t)^{n-2}(q + npe^t)$$

so  $E(X^2) = np(q + np)$ .

Recall the following definition.

**Definition 38.8.3** The variance of  $X$  is defined as  $E((X - E(X))^2)$ . The mean is defined as  $E(X)$ .

This is a measure of how spread out the distribution is.

**Example 38.8.4** Find the variance of  $X$  if  $X$  is a binomial random variable.

Note that in general,

$$\begin{aligned} E((X - E(X))^2) &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

Thus the variance of  $X$  for  $X$  a binomial random variable is  $np(q + np) - (np)^2 = npq$ .

**Example 38.8.5** Let  $X$  be normally distributed with parameters  $\mu, \sigma^2$ . Find the mean and variance. In fact, show that the mean is  $\mu$  and the variance is  $\sigma^2$ . Determine the moment generating function.

This will be done by using a moment generating function as above. For  $X$  normally distributed,

$$\begin{aligned} M(t) &\equiv E(e^{tX}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{tx} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}((x - (\mu + t\sigma^2))^2 - (t^2\sigma^4 + 2t\sigma^2\mu))\right) dx \end{aligned}$$

after simplification and completing the square. Thus this equals

$$\exp\left(\frac{1}{2}t^2\sigma^2 + \mu t\right) \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}((x - (\mu + t\sigma^2))^2)\right) dx$$

Change the variable in that integral on the right. Let

$$\frac{1}{\sqrt{2}\sigma}(x - (\mu + t\sigma^2)) = u$$

so  $\frac{1}{\sqrt{2}\sigma}dx = du$ . Then the expression becomes

$$\begin{aligned} &\exp\left(\frac{1}{2}t^2\sigma^2 + \mu t\right) \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp(-u^2) dx \sqrt{2}\sigma \\ &= \exp\left(\frac{1}{2}t^2\sigma^2 + \mu t\right) = M(t) \end{aligned} \tag{38.2}$$

In this case there are no mathematical pathologies and so

$$M'(t) = e^{\frac{1}{2}t^2\sigma^2 + \mu t} (t\sigma^2 + \mu)$$

so letting  $t = 0$  yields  $E(X) = \mu$ . Then also

$$M''(t) = e^{\frac{1}{2}t^2\sigma^2 + \mu t} (t^2\sigma^4 + 2t\sigma^2\mu + \sigma^2 + \mu^2)$$

and so  $E(X^2) = M''(0) = \sigma^2 + \mu^2$ . Then the variance is  $E(X^2) - E(X)^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$  showing the identification of these parameters.

**Example 38.8.6** Let  $X$  be a  $\mathcal{X}^2(r)$  distribution. Find the moment generating function valid for  $t$  in some interval containing 0.

By definition, this is

$$\int_0^\infty \frac{1}{\Gamma(r/2)2^{r/2}} x^{(r/2)-1} e^{-x/2} e^{tx} dx = \int_0^\infty \frac{1}{\Gamma(r/2)2^{r/2}} x^{(r/2)-1} e^{-x(\frac{1}{2}-t)} dx$$

so change the variable letting  $u = x(\frac{1}{2} - t)$  so  $du = (\frac{1}{2} - t) dx$ . Let  $|t| < \frac{1}{2}$ . Then the integral is

$$\begin{aligned} & \int_0^\infty \frac{1}{\Gamma(r/2)2^{r/2}} \left( \frac{u}{(1/2-t)} \right)^{(r/2)-1} e^{-u} \frac{1}{(1/2-t)} du \\ &= \frac{1}{2^{r/2}(1/2-t)^{r/2}} \frac{1}{\Gamma(r/2)} \int_0^\infty u^{r/2-1} e^{-u} du = \frac{1}{(1-2t)^{r/2}} \end{aligned}$$

Now with this, you can find all the moments desired.

**Proposition 38.8.7** Suppose  $X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ . Then  $\frac{X-\mu}{\sigma}$  is normally distributed with mean 0 and variance 1.

**Proof:** This is real easy to do with the moment generating technique.

$$E\left(\exp\left(\frac{X-\mu}{\sigma}\right)\right) = E\left(\exp\left(\frac{t}{\sigma}X - \frac{t\mu}{\sigma}\right)\right) = E\left(\exp\left(\frac{t}{\sigma}X\right)\exp\left(-\frac{t\mu}{\sigma}\right)\right)$$

Now  $\frac{t}{\sigma}X$  and  $-\frac{t\mu}{\sigma}$  are independent. (Check the definition.) Therefore, the above reduces to

$$E\left(\exp\left(\frac{t}{\sigma}X\right)\right)E\left(\exp\left(-\frac{t\mu}{\sigma}\right)\right) = e^{\frac{t}{\sigma}\mu}e^{\frac{1}{2}\sigma^2\left(\frac{t}{\sigma}\right)^2}e^{-\frac{t\mu}{\sigma}} = e^{\frac{1}{2}t^2}$$

which is the moment generating function of a random variable which is normally distributed with mean 0 and variance 1. ■

You might call  $\frac{X-\mu}{\sigma}$  a standard normal deviate.

**Corollary 38.8.8** Let  $X$  be normally distributed with mean  $\mu$  and variance  $\sigma^2$ . Then  $\left(\frac{X-\mu}{\sigma}\right)^2$  is distributed as  $\mathcal{X}^2(1)$ .

**Proof:** From Proposition 38.8.7,  $\left(\frac{X-\mu}{\sigma}\right)$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ . If  $f(t)$  is the density of  $\left(\frac{X-\mu}{\sigma}\right)^2$ , then

$$\begin{aligned} F(x) &\equiv \int_0^x f(t) dt \equiv P\left(\left(\frac{X-\mu}{\sigma}\right)^2 < x\right) = P\left(-\sqrt{x} < \frac{X-\mu}{\sigma} < \sqrt{x}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-\frac{1}{2}t^2} dt = \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^{\sqrt{x}} e^{-\frac{1}{2}t^2} dt \end{aligned}$$

change variables. Let  $\frac{t^2}{2} = u$  so  $t dt = du, dt = \frac{du}{\sqrt{2u}}$ . Then the above is

$$\frac{\sqrt{2}}{\sqrt{2}\sqrt{\pi}} \int_0^{x/2} u^{-1/2} e^{-u} du$$

Then, taking the derivative will yield the density. This is

$$\begin{aligned} \frac{1}{\sqrt{\pi}} \frac{1}{2} \left(\frac{x}{2}\right)^{-1/2} e^{-x/2} &= \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{2}\sqrt{x}} e^{-x/2} \\ &= \frac{1}{\Gamma(1/2)2^{1/2}} x^{1/2-1} e^{-x/2} \end{aligned}$$

because of Corollary 38.1.2, which is the density for  $\mathcal{X}^2(1)$  as claimed. ■

## 38.9 Independence and Conditional Probability

In the above, the concept of Probability that a random variable is in some set has been considered. More generally, you have a set and a collection of subsets of this set and a function which assigns a number between 0 and 1 to sets in this collection. This is the probability function. The following has to do with conditional probability.

**Definition 38.9.1** Let  $\mathcal{C}$  be a collection of sets contained in some universal set  $U$ . These could be intervals on the real line for example, and  $U$  could be  $\mathbb{R}$ . Let  $P : \mathcal{C} \rightarrow [0, 1]$ . Thus for  $A$  a set,  $P(A) \in [0, 1]$ . It satisfies the following conditions.

1. If  $A_i$  are disjoint sets in  $\mathcal{C}$ , then  $P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$ . More generally, if you have infinitely many such disjoint sets,  $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ .
2. If  $A \in \mathcal{C}$ , then  $P(A) + P(U \setminus A) = 1$ .

Then one defines the conditional probability as follows. If  $P(B) \neq 0$ ,

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$

The sets in  $\mathcal{C}$  are called **events**.

To do this right, you should be using  $\sigma$  algebras and measures on an abstract probability space. However, these things are not discussed in this book.

The words used when you write  $P(A|B)$  are: probability of  $A$  given  $B$ . In other words, if you are considering random variables, you know that  $X \in B$  where  $B$  is some possibly smaller set than  $U$ . For example, you might know that a normally distributed random variable is in  $[1, 5]$  and given this knowledge, the appropriate probability function would be defined as

$$P(X \in A|X \in [1, 5]) = \frac{P(X \in A \text{ and } X \in [1, 5])}{P(X \in [1, 5])}$$

This really restricts the set  $U$  to  $B$  and  $A \rightarrow P(A|B)$  is a probability function defined on  $B$ .

As indicated earlier, events  $A_1, \dots, A_n$  are said to be independent if

$$P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$$

Note that if  $n = 2$ , and  $A_1, A_2$  are independent, this says that

$$P(A_1|A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)} = \frac{P(A_1)P(A_2)}{P(A_2)} = P(A_1)$$

**Example 38.9.2** You roll a die  $n$  times. Let  $X_k$  be the value on the die on the  $k^{\text{th}}$  roll of the die. Let  $A_k$  be the event that the value of  $X_k$  is in  $S_k$  where  $S_k$  is some set of numbers from 1 to 6. Then the value of  $X_j$  for  $j \neq k$  has absolutely no bearing on whether  $X_k$  is in  $S_k$ .

$$P(X_1 \in S_1 \text{ and } X_2 \in S_2 \cdots \text{ and } X_n \in S_n) = \prod_{k=1}^n P(X_k \in S_k).$$

Indeed, this follows from noting that if  $|S_k|$  is the number of outcomes in  $S_k$  then the expression on the left is

$$\frac{\prod_{k=1}^n |S_k|}{6^n} = \prod_{k=1}^n \frac{|S_k|}{6} = \prod_{k=1}^n P(X_k \in S_k) \quad (38.3)$$

This is a typical way of getting independent events. Just do experiments in which the outcome of any experiment is totally unaffected by the outcome of all the others. This also illustrates what it means for random variables to be independent. Recall the following definition.

**Definition 38.9.3** Let  $X_i, i = 1, 2, \dots, n$  be random variables. They are independent means that for  $I_k$  a suitable set,

$$P(X_1 \in I_1, \dots, X_n \in I_n) = \prod_{k=1}^n P(X_k \in I_k)$$

It follows that if the  $X_i$  are independent and if  $g_i$  is a continuous function, then the  $g_i(X_i)$  are also independent. All this does is change the sets  $I_k$  in the above definition, but to do this right, you need more mathematical machinery.

When you have a random vector  $\mathbf{X} = (X_1, \dots, X_p)$  with density function  $f$  what does it mean for the components of this random vector to be independent? It means that there



are nonnegative functions  $x_i \rightarrow f_i(x_i)$  such that  $f(\mathbf{x}) = \prod_{i=1}^p f_i(x_i)$ . Note how this gives the conclusion of the above theorem.

$$\begin{aligned}
 P(X_1 \in I_1, \dots, X_p \in I_p) &= \int_{\prod_{i=1}^p I_i} f(\mathbf{x}) d\mathbf{x} \\
 &= \int_{I_1} \cdots \int_{I_p} f_1(x_1) \cdots f_p(x_p) dx_p \cdots dx_1 \\
 &= \int_{I_1} f_1(x_1) dx_1 \cdots \int_{I_p} f_p(x_p) dx_p \\
 &= \prod_{i=1}^p P(X_i \in I_i)
 \end{aligned}$$

In fact, this is a specialization of what always happens in every situation. Note that the same argument shows that if these components are an independent set, then if you consider

$$(g_1(X_1), \dots, g_p(X_p))$$

these would also be independent random variables. In this case,

$$\begin{aligned}
 &P(g_1(X_1) \in I_1, \dots, g_p(X_p) \in I_p) \\
 &= P(X_1 \in g_1^{-1}(I_1), \dots, X_p \in g_p^{-1}(I_p)) \\
 &= \int_{\prod_{i=1}^p g_i^{-1}(I_i)} f(\mathbf{x}) d\mathbf{x} \\
 &= \int_{g_1^{-1}(I_1)} \cdots \int_{g_p^{-1}(I_p)} f_1(x_1) \cdots f_p(x_p) dx_p \cdots dx_1 \\
 &= \int_{g_1^{-1}(I_1)} f_1(x_1) dx_1 \cdots \int_{g_p^{-1}(I_p)} f_p(x_p) dx_p \\
 &= \prod_{i=1}^p P(X_i \in g_i^{-1}(I_i)) = \prod_{i=1}^p P(g_i(X_i) \in I_i)
 \end{aligned}$$

This proves the following.

**Proposition 38.9.4** *Let  $\mathbf{X} = (X_1, \dots, X_p)$  and suppose there is a density function  $f(\mathbf{x})$ . Then the components are independent random variables if  $f$  has the following form.*

$$f(\mathbf{x}) = \prod_{i=1}^p f_i(x_i)$$

*If these are independent random variables, then so are  $\{g_i(X_i)\}_{i=1}^p$  whenever  $g_i$  are continuous functions.*

This is actually an equivalence so there is no loss of generality in taking it as the definition of independence. However, it gets much more involved because some random variables are neither discrete nor continuous. Nevertheless, this kind of thing will hold if appropriately generalized. To do this in full generality requires a much better mathematical theory than any contemplated in this book. It will end up involving differentiation theory of something called a Radon measure. Having noted this, all of the techniques discussed here were developed with nothing more than the Riemann integral in the early 1900's.

Also recall the definition of mean and variance given above.

**Definition 38.9.5** Let  $X$  be a random variable. Its mean is defined as  $\mu \equiv E(X)$ . The variance is defined as  $E((X - \mu)^2)$ . The mean is a weighted average. It is what you would expect to see if you took many random samples from this distribution and averaged them. (In fact there is a theorem which says this.) The variance is a description of how spread out the probability density is. If the variance is small, then the random variable will be close to  $\mu$  with high probability and if it is large, then it is not as certain the random variable is close to  $\mu$ .

Now with this definition of mean and variance, why is the normal distribution so important? It is because of the central limit theorem. Suppose  $E(X_k^2) < \infty$  where  $X_k$  is a random variable.

**Theorem 38.9.6** Let  $\{X_k\}_{k=1}^\infty$  be random variables satisfying  $E(X_k^2) < \infty$ , which are independent and identically distributed with mean  $\mu = E(X_k)$  and positive variance  $0 < \sigma^2 \equiv E((X_k - \mu)^2)$ . Let

$$Z_n \equiv \sum_{j=1}^n \frac{X_j - \mu}{\sigma \sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \quad (38.4)$$

where  $\bar{X}$  is the average of the  $X_k$   $\frac{1}{n} \sum_{k=1}^n X_k$ . Then for  $Z$  a normally distributed random variable having mean 0 and variance 1,

$$\lim_{n \rightarrow \infty} P(Z_n \in A) = P(Z \in A)$$

where  $A$  is a suitable set.

Of course this begs the question: What are  $\mu, \sigma$ ? Much that is done in statistics has to do with determination of these or other parameters. They both give interesting information if they can be estimated.

How does independence relate to moment generating functions?

**Proposition 38.9.7** Let  $\mathbf{X}_k$  be a random vector with values in  $\mathbb{R}^{m_k}$ . Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{pmatrix}$$

and let the moment generating function for  $\mathbf{X}$  exist

$$M(\mathbf{t}) \equiv M(t_1, \dots, t_p) \equiv E(e^{\mathbf{t} \cdot \mathbf{X}}) = E\left(\exp\left(\sum_{k=1}^p t_k \cdot \mathbf{X}_k\right)\right)$$

Then the  $\mathbf{X}_k$  are independent if and only if

$$M(\mathbf{t}) = \prod_{k=1}^p M(\mathbf{0}, \dots, \mathbf{0}, t_k, \mathbf{0}, \dots, \mathbf{0}) \quad (38.5)$$

**Proof:** First suppose the  $\mathbf{X}_k$  are independent. Then the density function for  $\mathbf{X}$  is of the form

$$f(\mathbf{x}) = f_1(\mathbf{x}_1) f_2(\mathbf{x}_2) \cdots f_p(\mathbf{x}_p)$$

Therefore,

$$\begin{aligned}
 M(t) &= \int_{\mathbb{R}^{m_p}} \int_{\mathbb{R}^{m_{p-1}}} \cdots \int_{\mathbb{R}^{m_1}} f_1(x_1) f_2(x_2) \\
 &\quad \cdots f_p(x_p) \exp\left(\sum_{k=1}^p t_k \cdot X_k\right) dx_1 \cdots dx_p \\
 &= \int_{\mathbb{R}^{m_p}} \int_{\mathbb{R}^{m_{p-1}}} \cdots \int_{\mathbb{R}^{m_1}} f_1(x_1) f_2(x_2) \cdots f_p(x_p) \prod_{k=1}^p \exp(t_k \cdot x_k) \\
 &= \prod_{k=1}^p \int_{\mathbb{R}^{m_k}} f_k(x_k) \exp(t_k \cdot x_k) dx_k = \prod_{k=1}^p M(0, \dots, 0, t_k, 0, \dots, 0)
 \end{aligned}$$

Conversely, suppose the other condition. Then

$$\begin{aligned}
 M(t) &= \prod_{k=1}^p M(0, \dots, 0, t_k, 0, \dots, 0) = \\
 &\quad \prod_{k=1}^p \int_{\mathbb{R}^{m_p}} \int_{\mathbb{R}^{m_{p-1}}} \cdots \int_{\mathbb{R}^{m_1}} f(x) \exp(t_k \cdot x_k)
 \end{aligned}$$

by Fubini's theorem,

$$\begin{aligned}
 &\prod_{k=1}^p \int_{\mathbb{R}^{m_k}} \exp(t_k \cdot x_k) \cdots \int_{\mathbb{R}^{m_j}} \cdots \int_{\mathbb{R}^{m_1}} f(x) dx_1 \cdots dx_j \cdots dx_k \\
 &\quad \equiv \prod_{k=1}^p \int_{\mathbb{R}^{m_k}} \exp(t_k \cdot x_k) f_k(x_k) dx_k
 \end{aligned} \tag{38.6}$$

where  $f_k(x_k)$  is called the marginal distribution and is obtained as

$$f_k(x_k) \equiv \int_{\mathbb{R}^{m_p}} \cdots \widehat{\int_{\mathbb{R}^{m_k}}} \cdots \int_{\mathbb{R}^{m_1}} f(x) dx_1 \cdots \widehat{dx_k} \cdots dx_{m_p}$$

where the hat indicates the thing is being omitted. Thus,

$$\int_{\mathbb{R}^{m_p}} \cdots \int_{\mathbb{R}^{m_1}} \prod_{k=1}^p f_k(x_k) dx_1 \cdots dx_p = 1$$

and with respect to the density  $\prod_{k=1}^p f_k(x_k)$ ,  $E(t \cdot X)$  yields 38.6. But, as noted above, if two densities deliver the same moment generating function, then they are the same. Hence the  $X_k$  are independent because the density is the product of functions  $f_k$  of the  $x_k$ . ■

**Proposition 38.9.8** Suppose  $\{X_k\}_{k=1}^r$  are independent and each  $n(0, 1)$ , normal with 0 mean and variance 1. Then  $\sum_{k=1}^r X_k^2$  is  $\mathcal{X}^2(r)$ .

**Proof:** It follows from Corollary 38.8.8 that  $X_k^2$  is  $\mathcal{X}^2(1)$ . Then using independence,

$$\begin{aligned}
 E\left(\exp\left(t \sum_{k=1}^r X_k^2\right)\right) &= E\left(\prod_{k=1}^r \exp(t X_k^2)\right) = \prod_{k=1}^r E(t X_k^2) \\
 &= \prod_{k=1}^r \frac{1}{(1-2t)^{1/2}} = \frac{1}{(1-2t)^{r/2}}
 \end{aligned}$$

which is the moment generating function of  $\mathcal{X}^2(r)$ . ■

**Corollary 38.9.9** Suppose  $\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}^T$  where  $\mathbf{X}$  has a moment generating function of the form

$$M(\mathbf{t}) = e^{\frac{1}{2}\mathbf{t}^T A \mathbf{t}}$$

where  $A$  is real and symmetric having rank  $r \leq n$  and eigenvalues 0 or 1. Then  $\mathbf{X}^T A \mathbf{X}$  is  $\mathcal{X}^2(r)$ . (When  $r < n$ , this is a moment generating function of a random variable which is said to be a singular multivariate normal. )

**Proof:** By Theorem 11.4.7 there is orthogonal  $U$  such that  $U^T A U = D$  where  $D$  is of the form  $\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$  where  $I$  is an  $r \times r$  identity matrix. Then let  $\mathbf{Y} = U^T \mathbf{X}$ . What is the distribution of  $\mathbf{Y}$ ?

$$\begin{aligned} E(\exp(\mathbf{t} \cdot \mathbf{Y})) &\equiv E(\exp(\mathbf{t} \cdot U^T \mathbf{X})) = E(\exp(U \mathbf{t} \cdot \mathbf{X})) \\ &= \exp\left(\frac{1}{2}(U \mathbf{t})^T A (U \mathbf{t})\right) = \exp\left(-\frac{1}{2}\mathbf{t}^T U^T A U \mathbf{t}\right) \\ &= \exp\left(\frac{1}{2}\mathbf{t}^T D \mathbf{t}\right) = \prod_{k=1}^r \exp\left(\frac{1}{2}t_k^2\right) \end{aligned}$$

Now  $\exp(tY_k) = 1$  and so  $Y_k = 0$  if  $k > r$  and otherwise,  $Y_k$  is  $n(0, 1)$ , normal with mean 0 and variance 1. Thus Theorem 38.9.7 implies that these random variables are independent and each  $n(0, 1)$ . Hence by Proposition 38.9.8,

$$\mathbf{X}^T A \mathbf{X} = \mathbf{Y}^T U^T A U \mathbf{Y} = \mathbf{Y}^T D \mathbf{Y} = \sum_{k=1}^r Y_k^2 \text{ which is } \mathcal{X}^2(r). \blacksquare$$

## Chapter 39

# Statistical Tests

In this chapter, are various tests for determining parameters and answering other questions with a certain probability associated with the answers. This is all based on the notion of random variables of various forms, called statistics, for which there is a known distribution. The pattern is to compute the statistic which is based on a random sample and then to use its known distribution to make statistical inferences. This is always what you do when you know that the samples are coming from a probability distribution involving unknown parameters.

For example, it is reasonable to believe that the weight of adult men in Arkansas is normally distributed. However, you don't know the mean  $\mu$  and the variance  $\sigma^2$  and these are what you want because you want to know the probability that some man weighs between 140 and 180 pounds. You pick randomly 40 males and record their weights. These weights are the values of independent random variables. Then, you estimate  $\mu$  and  $\sigma^2$  from this sample, and things like an interval where you have a probability of .95 that the weight of a person will lie in this interval. A hypothesis you might want to test for would be that the average weight of men in Arkansas is the same as the average weight of men in Alabama. If you are interested in something other than weight, you would adjust accordingly. You could be interested in errors produced by a machine when it makes bolts for example. How sure are you that some measurement is acceptable? If you were an insurance company, you would want to know with some confidence an interval containing the lifespan of a person or an interval and probability associated with it which gives the number of accidents that people age 17-30 will have. One could go on and on.

In addition to this, each application of these methods would need to be examined carefully to be sure that the assumptions on the underlying distribution are not unreasonable.

As suggested, there are two main forms these inferences take. One involves something called a confidence interval and the other involves rejecting or accepting a given hypothesis, called a null hypothesis. I admit to being prejudiced toward confidence intervals because they deliver a straight forward affirmation that with a certain probability something happens and involve less jargon. However, it is sometimes appropriate to test for a hypothesis which is either true or false and you may be able to identify a probability that the so called null hypothesis, that which is being tested, is false and this is also very useful. These notions will be developed on specific examples. I think this will make the ideas easier to understand than to focus first on generalities laden with jargon. The statistics of interest in the following will be those which have  $\chi^2(r)$ ,  $T$ , or  $F$  distributions. The first of these has

already been considered. The last two are described below.

An interesting observation about all of this is that there is a gap between the theory and the applications like those mentioned above. To really understand the mathematical theory, you need much more advanced mathematics than what is encountered in this book. This happens as soon as you start asking fundamental questions about what a random variable is independent of some application or why certain limits exist and in what sense they exist. Some of the most fundamental questions come from the Kolmogorov extension theorem which has to do with measures defined on infinite products.

### 39.1 The Distribution of $nS^2/\sigma^2$

In all of this,  $X_k$  is a random variable and we assume  $X_1, X_2, \dots$  are independent. For example, you might have a large population of people and the weight of a person is normally distributed. Then  $X_i$  would be the  $i^{\text{th}}$  observation of a randomly selected person's weight.

**Definition 39.1.1** The symbol  $S^2$  denotes the sample variance of  $X_k, k = 1, \dots, n$ , which is of the form  $\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2$  where  $\bar{X}$  is the sample average of the random variables  $X_1, \dots, X_n$ ,  $\bar{X} \equiv \frac{1}{n} \sum_{i=1}^n X_i$ .

When the sample is taken from a normal distribution having mean  $\mu$  and variance  $\sigma^2$ , it turns out that the random variable  $nS^2/\sigma^2$  has a chi-squared distribution. When this is shown, it becomes possible to estimate the variance along with a probability that the variance is really in some interval called a confidence interval. One can also use this in terms of a hypothesis test. For example, you might reject the hypothesis that the variance is very large. This fact that  $nS^2/\sigma^2$  is  $\chi^2(n-1)$  which is shown below is very significant because the statistic  $nS^2/\sigma^2$  does not involve  $\mu$ . The following proposition makes this possible. It is a statement about independence of the sample mean  $\bar{X}$  and the random vector of deviations from the sample mean.

**Proposition 39.1.2** Let  $X_k, k = 1, 2, \dots, n$  be independent random variables all having a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Let  $\bar{X} \equiv \frac{1}{n} \sum_{k=1}^n X_k$ , called the sample mean. Then  $\bar{X}$  and the random vector  $\begin{pmatrix} X_1 - \bar{X} & \dots & X_n - \bar{X} \end{pmatrix}$  are independent.

**Proof:** This is done most easily with the moment generating technique.

$$E \left( e^{t\bar{X} + \sum_{k=1}^n t_k (X_k - \bar{X})} \right) = E \left( e^{(t - \sum_{k=1}^n t_k) \bar{X} + \sum_{k=1}^n t_k X_k} \right) \quad (39.1)$$

It is necessary to verify that this equals  $E \left( e^{t\bar{X}} \right) E \left( e^{\sum_{k=1}^n t_k (X_k - \bar{X})} \right)$ . However, 39.1 equals

$$\begin{aligned} &= E \left( e^{\left( \frac{1}{n} t - \sum_{k=1}^n \frac{1}{n} t_k \right) \sum_{j=1}^n X_j + \sum_{k=1}^n t_k X_k} \right) \\ &= E \left( e^{\sum_{j=1}^n \left( \frac{1}{n} t - \sum_{k=1}^n \frac{1}{n} t_k \right) X_j + \sum_{j=1}^n t_j X_j} \right) \\ &= E \left( e^{\sum_{j=1}^n \left( \frac{1}{n} t - \sum_{k=1}^n \frac{1}{n} t_k + t_j \right) X_j} \right) \\ &= E \left( \prod_{j=1}^n \exp \left( \left( \frac{1}{n} t - \sum_{k=1}^n \frac{1}{n} t_k + t_j \right) X_j \right) \right) \end{aligned}$$

In that last term you have the product of continuous functions of independent random variables and so, by 38.2 which gives the moment generating function for a normally distributed random variable, it equals

$$\begin{aligned} & \prod_{j=1}^n E \left( e^{\left(\frac{1}{n}t - \sum_{k=1}^n \frac{1}{n}t_k + t_j\right)X_j} \right) \\ &= \prod_{j=1}^n \exp \left( \left( \left( \frac{1}{n}t - \sum_{k=1}^n \frac{1}{n}t_k + t_j \right) \mu + \frac{1}{2} \sigma^2 \left( \frac{1}{n}t - \sum_{k=1}^n \frac{1}{n}t_k + t_j \right)^2 \right) \right) \\ &= \exp \left( \sum_j \left( \left( \frac{1}{n}t - \sum_{k=1}^n \frac{1}{n}t_k + t_j \right) \mu + \sum_j \frac{1}{2} \sigma^2 \left( \frac{1}{n}t - \sum_{k=1}^n \frac{1}{n}t_k + t_j \right)^2 \right) \right) \end{aligned}$$

Simple algebra shows that  $\sum_j \left( \frac{1}{n}t - \sum_{k=1}^n \frac{1}{n}t_k + t_j \right) = t$ . Thus the above is

$$= \exp \left( t\mu + \sum_j \frac{1}{2} \sigma^2 \left( \frac{1}{n}t - \sum_{k=1}^n \frac{1}{n}t_k + t_j \right)^2 \right)$$

Now  $-\sum_{k=1}^n \frac{1}{n}t_k + t_j = \sum_{k=1}^n \frac{1}{n}(t_j - t_k)$  so the above reduces to

$$= \exp \left( t\mu + \frac{1}{2} \sigma^2 \sum_j \left( \frac{1}{n}t + \sum_k \frac{1}{n}(t_j - t_k) \right)^2 \right)$$

Consider the mixed term in that last summand above.

$$\sum_j 2 \frac{t}{n} \sum_k \frac{1}{n}(t_j - t_k) = 2 \frac{t}{n} \sum_j \sum_k \frac{1}{n}(t_j - t_k) = 0$$

Hence the above reduces to

$$\begin{aligned} & \exp \left( t\mu + \frac{1}{2} \sigma^2 \left( \sum_j \left( \frac{1}{n}t \right)^2 + \sum_j \left( \sum_k \frac{1}{n}(t_j - t_k) \right)^2 \right) \right) \\ &= \exp \left( t\mu + \frac{1}{2} \sigma^2 \frac{t^2}{n} \right) \exp \left( \sum_j \sum_k \frac{1}{2} \sigma^2 \frac{1}{n} (t_j - t_k)^2 \right) \end{aligned} \quad (39.2)$$

So you see, the moment generating function splits up the first factor depending only on  $t$  and the second depending only on the  $t_k$ .

$$\begin{aligned} E(e^{t\bar{X}}) &= E \left( \exp \left( \frac{1}{n} \sum_k t X_k \right) \right) = E \left( \prod_k \exp \left( \frac{t}{n} X_k \right) \right) \\ &= \prod_k E \left( \exp \left( \frac{t}{n} X_k \right) \right) = \prod_k e^{\frac{t}{n} \mu + \frac{1}{2} \sigma^2 \frac{t^2}{n^2}} = e^{t\mu + \frac{1}{2} \sigma^2 \frac{t^2}{n}} \end{aligned}$$

Thus the first term in 39.2 is the moment generating function of  $\bar{X}$ . Some computations show that the second term is the moment generating function of the vector

$$\begin{pmatrix} X_1 - \bar{X} & \cdots & X_n - \bar{X} \end{pmatrix}$$

Indeed,

$$\begin{aligned} E \left( \exp \sum_{k=1}^n t_k (X_k - \bar{X}) \right) &= E \left( \exp \left( \sum_k t_k X_k - \sum_k t_k \frac{1}{n} \sum_j X_j \right) \right) \\ &= E \left( \exp \left( \sum_k t_k X_k - \sum_j t_j \frac{1}{n} \sum_k X_k \right) \right) = E \left( \exp \left( \sum_j \sum_k \frac{t_k - t_j}{n} X_k \right) \right) \\ &= \prod_j E \left( \exp \left( \sum_k \frac{t_k - t_j}{n} X_k \right) \right) = \prod_j \prod_k E \left( \exp \left( \frac{t_k - t_j}{n} X_k \right) \right) \\ &= \prod_j \prod_k \left( \exp \left( \left( \frac{t_k - t_j}{n} \mu \right) + \frac{1}{2} \left( \frac{t_k - t_j}{n} \right)^2 \sigma^2 \right) \right) \\ &= \exp \left( \sum_j \sum_k \frac{1}{2} \left( \frac{t_k - t_j}{n} \right)^2 \sigma^2 \right) \end{aligned}$$

Therefore, by Proposition 38.9.7,  $\bar{X}$  and this random vector are linearly independent. ■

The above proposition leads to something interesting, the distribution of  $nS^2/\sigma^2$ . Let the  $X_k$  be independent and normally distributed with mean  $\mu$  and variance  $\sigma^2$ . Then

$$S^2 \equiv \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2$$

The distribution of  $nS^2/\sigma^2 = \sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2}$  will be considered. If we know this, then since  $S^2$  is experimentally determined, it will follow that we could estimate  $\sigma^2$ . First note that

$$X_k - \bar{X} = X_k - \mu + \mu - \bar{X}$$

and so

$$\begin{aligned} (X_k - \bar{X})^2 &= (X_k - \mu)^2 - 2(X_k - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2 \\ 2 \sum_{k=1}^n (X_k - \mu)(\bar{X} - \mu) &= 2n(\bar{X} - \mu)(\bar{X} - \mu) \end{aligned}$$

Therefore,

$$\sum_{k=1}^n (X_k - \bar{X})^2 + n(\bar{X} - \mu)^2 = \sum_{k=1}^n (X_k - \mu)^2$$

Then

$$\sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2} = \sum_{k=1}^n \frac{(X_k - \mu)^2}{\sigma^2}$$



From what was shown above,  $\frac{n(\bar{X}-\mu)^2}{\sigma^2}$  and the vector  $\begin{pmatrix} X_1 - \bar{X} & \cdots & X_n - \bar{X} \end{pmatrix}$  are independent. From this, it follows that

$$\frac{n(\bar{X}-\mu)^2}{\sigma^2}, \sum_{k=1}^n \frac{t(X_k - \bar{X})^2}{\sigma^2}$$

are independent.

Using this and the known distribution of  $\frac{(X_k - \mu)^2}{\sigma^2}$ ,

$$\begin{aligned} E\left(t \sum_{k=1}^n \frac{(X_k - \mu)^2}{\sigma^2}\right) &= E\left(\exp\left(\sum_{k=1}^n \frac{t(X_k - \bar{X})^2}{\sigma^2} + t \frac{n(\bar{X} - \mu)^2}{\sigma^2}\right)\right) \\ &= E\left(\exp\left(t \sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2}\right) \exp\left(t \frac{n(\bar{X} - \mu)^2}{\sigma^2}\right)\right) \end{aligned}$$

By independence, this is

$$= E\left(\exp\left(t \sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2}\right)\right) E\left(\exp\left(t \frac{n(\bar{X} - \mu)^2}{\sigma^2}\right)\right) \quad (39.3)$$

Of course the thing we want is  $E\left(\exp\left(t \sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2}\right)\right)$ , but the expression on the left  $E\left(t \sum_{k=1}^n \frac{(X_k - \mu)^2}{\sigma^2}\right)$ , and the factor on the right are known or easy to find. Consider the factor on the right.

$$\begin{aligned} t \frac{n(\bar{X} - \mu)^2}{\sigma^2} &= t \frac{n\left(\frac{1}{n} \sum_{k=1}^n (X_k - \mu)\right)^2}{\sigma^2} \\ &= t \left(\sum_{k=1}^n \left(\frac{X_k - \mu}{\sqrt{n}\sigma}\right)\right)^2 \end{aligned}$$

What is the distribution of  $\sum_{k=1}^n \left(\frac{X_k - \mu}{\sqrt{n}\sigma}\right)$ ? By independence, its moment generating function is

$$\begin{aligned} E\left(\exp\left(t \sum_{k=1}^n \left(\frac{X_k - \mu}{\sqrt{n}\sigma}\right)\right)\right) &= E\left(\prod_{k=1}^n \exp\left(t \left(\frac{X_k - \mu}{\sqrt{n}\sigma}\right)\right)\right) \\ &= \prod_{k=1}^n E\left(\exp\left(t (\sqrt{n})^{-1} \left(\frac{X_k - \mu}{\sigma}\right)\right)\right) = \prod_{k=1}^n e^{\frac{1}{2} \left(\frac{t}{\sqrt{n}}\right)^2} = e^{\frac{1}{2} t^2} \end{aligned}$$

so it is a normal distribution having variance 1 and mean 0. It follows from Corollary 38.8.8 that the square of this random variable is  $\mathcal{X}^2(1)$ . Since we know the moment generating function for chi squared distributions, it follows that we know all the terms in 39.3 except for the one we want. It just a matter of filling in the expressions. Recall the moment generating function for  $\mathcal{X}^2(r)$  is

$$\frac{1}{(1-2t)^{r/2}}$$

Thus from 39.3,

$$\begin{aligned} E \left( \exp \left( t \sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2} \right) \right) \frac{1}{(1-2t)^{1/2}} &= \prod_{k=1}^n E \left( \exp \left( t \frac{(X_k - \mu)^2}{\sigma^2} \right) \right) \\ &= E \left( t \sum_{k=1}^n \frac{(X_k - \mu)^2}{\sigma^2} \right) \end{aligned}$$

This last term is the moment generating function of the sum of squares of standard normal deviates and so its moment generating function is known by Proposition 38.9.8 equals  $\frac{1}{(1-2t)^{n/2}}$ . Thus, dividing both sides by  $\frac{1}{(1-2t)^{1/2}}$  we get

$$E \left( \exp \left( t \sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2} \right) \right) = \frac{1}{(1-2t)^{(n-1)/2}}$$

which shows that  $nS^2/\sigma^2$  is distributed as  $\mathcal{X}^2(n-1)$ . This proves the following major theorem.

**Theorem 39.1.3** Suppose  $\{X_1, \dots, X_n\}$  are independent and they are normally distributed with variance  $\sigma^2$ . Let  $S^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2$ , called the sample variance, where  $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ . Then the random variable  $nS^2/\sigma^2$  is distributed as  $\mathcal{X}^2(n-1)$ .

This is really interesting. Note that we don't know the mean and yet this allows an estimation of the variance based on observations of the  $X_i$ .

### 39.1.1 Confidence Intervals for Variance

Here is the concept of a confidence interval for the variance of a normally distributed random variable.

**Example 39.1.4** Here are 11 numbers:

$$1, 2, 3, -3, 1.5, 0, -1, -2, -.7, -1.5, -2.2$$

These are independent samples taken from a normal distribution. Find a confidence interval for the variance of this normal distribution.

First, what in the world is a "confidence interval".

**Definition 39.1.5** An interval  $[a, b]$  is a .95 confidence interval for a parameter  $v$  means that the probability that  $v$  lies in  $[a, b]$  is .95.

Of course, if the probability that the parameter lies in  $[a, b]$  is .9, then it would be a .9 confidence interval and so forth.

Now consider the above example of 11 numbers. The sample mean or average of these numbers is  $-.26364$ . Then  $11S^2$  for these numbers is 37.065. This just follows from a computation. Then from Theorem 39.1.3,  $11S^2/\sigma^2$  is a  $\mathcal{X}^2(10)$  random variable. We find an interval such that the probability that such a  $\mathcal{X}^2(10)$  random variable is in this interval. This is easy to do from tables. However, you can also use the distribution

$F(x) \equiv P(X \leq x)$ . You can obtain your own table of this using MATLAB or you can use the graph of this function using MATLAB. Here is an easy way to do it. I am sure there are more elegant ways to obtain this graph but I am picking one which seems to minimize the fussiness. MATLAB knows how to do numerical integration. The following tells it to integrate up to  $n * .05$  and place a dot there at the point  $(n * .05, y)$  where  $y$  is the integral up to  $n * .05$ .

```
>> hold on
r=10;
for n=1:1:1200
f=@(t)[1/(gamma(r/2)*2^(r/2))*t.^((r/2)-1).*exp(-t/2)];
y=integral(f,0,n*.05);
plot(n*.05,y,'.','Linewidth',2,'color','black')
end
```

This will produce a nice graph of  $F(x) \equiv P(X \leq x)$ , called the probability distribution function, and so you identify an interval for which the area under the curve is no more than .95. Click on the icon on the tool bar for the graph which says: "data cursor". This will give you a little cross which you can move around and click on points of the graph and it will tell you coordinates, an  $x$  coordinate and a  $y$  coordinate which is the probability that  $X \leq x$ . This allows you to avoid hunting for things in a table. In fact, MATLAB can essentially produce the tables for you. In ancient times, we had to use tables and we even used tables of trig. functions and logarithms. There was a whole set of specialized techniques which are now obsolete which we suffered with long ago. Now of course, there is software which can do all of it for you so it is important to understand what the software is doing.

If you have scientific notebook, it is even easier. In this case, all you have to do is type in math mode

$$\int_0^x \frac{1}{\Gamma(10/2) 2^{10/2}} t^4 e^{-t/2} dt$$

and ask it to graph this function of  $x$ . It will do so. Then you click on the icon which lets you identify coordinates just like you can do in MATLAB. The quality of the graph is not as good as what you get in MATLAB, but it does the job quite well with less hassle. I like looking at pictures better than rummaging through tables. However, if you like to look at tables, try this. Matlab will make a table for you.

```
>> hold on
T=[]; r=10;
for n=1:1:100
f=@(t)[(gamma(r/2)*2^(r/2))^(-1)*t.^((r/2)-1).*exp(-t/2)];
x=n*.5;
y=integral(f,0,x);
T=[T; x y];
end
T
```

I found this on line which is where I usually go for questions about MATLAB. It will produce a table having two columns, one for  $x$  and the next for  $y$  which will be the probability up to  $x$ .

Here are two points on the graph: (3.45, .0312) and (21.85, .9841). Thus the probability that  $X \leq 21.85$  is .9841 and the probability that  $X \leq 3.45$  is .0312. It follows that the probability that  $X \in [3.45, 21.85]$  is  $.9841 - .0312 = 0.9529$ . Therefore, the probability

that  $X = 11S^2/\sigma^2$  is in  $[3.45, 21.85]$  is better than .95. Thus

$$3.45 \leq \frac{(37.065)}{\sigma^2} \leq 21.85$$

so the probability is better than .95 that

$$\frac{1}{3.45} \geq \frac{\sigma^2}{(37.065)} \geq \frac{1}{21.85}$$

In other words, the probability is better than .95 that

$$10.743 \geq \sigma^2 \geq 1.6963$$

What if we only wanted to know with probability .7 where  $\sigma^2$  is? Then one could get a much smaller interval. Two points on the graph are (5.1, .1156) and (14.5, .8486). Then the same process yields with probability better than .7

$$\frac{1}{5.1} \geq \frac{\sigma^2}{(37.065)} \geq \frac{1}{14.5}$$

$$7.2676 \geq \sigma^2 \geq 2.5562$$

This is a much shorter interval but we can't be as sure that the variance is in this interval. If you say more about something, it is hardly surprising that you can't be quite as sure about your assertion.

**PROCEDURE 39.1.6** *To find a .95 confidence interval for the variance using a random sample*

$$X_1, X_2, \dots, X_n$$

*from a normal distribution of mean  $\mu$  and variance  $\sigma^2$  do the following.*

1. *Using a table or graph, determine an interval  $[a, b]$ ,  $a > 0$  such that the probability that a  $\chi^2(n-1)$  random variable is in  $[a, b]$  is at least .95.*
2. *Compute the sample mean  $\bar{X} \equiv \frac{1}{n} \sum_{k=1}^n X_k$  and  $nS^2 \equiv \sum_{k=1}^n (X_k - \bar{X})^2$ .*
3. *The .95 confidence interval is determined by solving the following inequality for  $\sigma^2$ .*

$$a \leq \frac{nS^2}{\sigma^2} \leq b$$

4. *Thus the .95 confidence interval is*

$$\frac{nS^2}{a} \geq \sigma^2 \geq \frac{nS^2}{b}$$

*The same procedure is followed if you want some other probability than .95.*

Incidentally, if you knew the mean  $\mu$  you could replace the sample mean with this and use a chi-squared distribution with one more degree of freedom which of course will result in a better confidence interval. However, I don't think you could have a good reason for thinking you know the mean, so such observations are mainly theoretical at this point.

## 39.2 The T and F Distributions

These are really interesting. They both involve independent random variables which are distributed as normal or  $\mathcal{X}^2$  distributions. These involve combinations of these other random variables and the idea is to find the density of these combinations. It is a nice application of the change of variables theorem.

### 39.2.1 The T Distribution

Here there are two **independent** random variables,  $W$  which is normally distributed with mean 0 and variance 1 and  $V$  which is  $\mathcal{X}^2(r)$ . Thus, as explained above,

$$P((V, W) \in A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} \frac{1}{\Gamma(r/2) 2^{r/2}} v^{(r/2)-1} e^{-v/2} dw dv$$

thus  $(V, W) \in (0, \infty) \times (-\infty, \infty)$ . The idea is to find the probability density of the statistic

$$T = \frac{W}{\sqrt{V/r}}$$

It is a random variable which has a known distribution. This involves changing the variable. Let

$$t = \frac{w}{\sqrt{v/r}}, u = v, \begin{pmatrix} u \\ t \end{pmatrix} = \mathbf{r} \begin{pmatrix} v \\ w \end{pmatrix}$$

This maps  $(0, \infty) \times (-\infty, \infty)$  one to one onto  $(0, \infty) \times (-\infty, \infty)$  as can be seen with a short computation. Let the density function of  $(t, u)$  be  $f(t, u)$ .

$$P((t, u) \in U) = P((v, w) \in \mathbf{r}^{-1}(U))$$

By the change of variables formula for multiple integrals if  $U$  is some open set in  $\mathbb{R}^2$ ,

$$\begin{aligned} \int_U f(t, u) du dt &= \int_{\mathbf{r}^{-1}(U)} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} \frac{1}{\Gamma(r/2) 2^{r/2}} v^{(r/2)-1} e^{-v/2} dw dv \\ &= \int_{\mathbf{r}^{-1}(U)} f\left(\frac{w}{\sqrt{v/r}}, v\right) J(v, w) dw dv \end{aligned}$$

where

$$J(v, w) = \left| \det \begin{pmatrix} 1 & 0 \\ -\frac{1}{2r} \frac{w}{(\frac{1}{r}v)^{\frac{3}{2}}} & \frac{1}{\sqrt{\frac{1}{r}v}} \end{pmatrix} \right| = \frac{1}{\sqrt{\frac{1}{r}v}}$$

Thus

$$f\left(\frac{w}{\sqrt{v/r}}, v\right) \frac{1}{\sqrt{\frac{1}{r}v}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} \frac{1}{\Gamma(r/2) 2^{r/2}} v^{(r/2)-1} e^{-v/2}$$

Then

$$f\left(\frac{w}{\sqrt{v/r}}, v\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} \frac{1}{\Gamma(r/2) 2^{r/2}} \sqrt{\frac{1}{r}v} v^{(r/2)-1} e^{-v/2}$$

Now it is necessary to invert the transformations and solve for  $v, w$  in terms of  $t, u$ .

$$t = \frac{w}{\sqrt{v/r}}, u = v$$

So  $w = t\sqrt{\frac{u}{r}}, v = u$ . Thus

$$\begin{aligned} f(t, u) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2 u}{2r}} \frac{1}{\Gamma(r/2) 2^{r/2}} \sqrt{\frac{1}{r}} u u^{(r/2)-1} e^{-u/2} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2r} t^2 u} e^{-\frac{1}{2} u} \frac{1}{\Gamma(r/2) 2^{r/2}} u^{\frac{1}{2}r - \frac{1}{2}} \sqrt{\frac{1}{r}} \end{aligned}$$

Now this is the density for a random vector  $(T, U)$  and it is desired to find the density for  $T$ . This means  $U$  can be anywhere in  $(0, \infty)$  and so to get this density we do the following integral.

$$\frac{1}{\sqrt{2\pi}} \sqrt{\frac{1}{r}} \frac{1}{\Gamma(r/2) 2^{r/2}} \int_0^\infty e^{-\frac{1}{2r} t^2 u} e^{-\frac{1}{2} u} u^{\frac{1}{2}r - \frac{1}{2}} du$$

Consider the integral. It is

$$\int_0^\infty e^{-u\left(\frac{t^2}{2r} + \frac{1}{2}\right)} u^{\frac{1}{2}(r-1)} du$$

Change variables letting  $x = u\left(\frac{t^2}{2r} + \frac{1}{2}\right), dx = \left(\frac{t^2}{2r} + \frac{1}{2}\right) du$ . Then it equals

$$\begin{aligned} &\int_0^\infty e^{-x} \left(\frac{x}{\frac{t^2}{2r} + \frac{1}{2}}\right)^{\frac{1}{2}(r-1)} \frac{1}{\frac{t^2}{2r} + \frac{1}{2}} dx \\ &= \left(\frac{1}{\frac{t^2}{2r} + \frac{1}{2}}\right)^{\frac{1}{2}r + \frac{1}{2}} \int_0^\infty e^{-x} x^{\frac{1}{2}(r-1)} dx \end{aligned}$$

Let  $\alpha - 1 = \frac{1}{2}(r-1)$ . Then the above equals

$$\left(\frac{1}{\frac{t^2}{2r} + \frac{1}{2}}\right)^{\frac{1}{2}r + \frac{1}{2}} \Gamma(\alpha) = \left(\frac{1}{\frac{t^2}{2r} + \frac{1}{2}}\right)^{\frac{1}{2}r + \frac{1}{2}} \Gamma\left(\frac{1}{2}r + \frac{1}{2}\right)$$

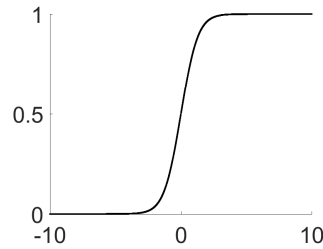
Therefore, the density function for  $T$  is

$$\begin{aligned} &= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{1}{r}} \frac{1}{\Gamma(r/2) 2^{r/2}} \left(\frac{2}{(t^2/r + 1)}\right)^{\frac{1}{2}r + \frac{1}{2}} \Gamma\left(\frac{1}{2}r + \frac{1}{2}\right) \\ &= \frac{1}{\sqrt{\pi}} \sqrt{\frac{1}{r}} \frac{\Gamma\left(\frac{1}{2}r + \frac{1}{2}\right)}{\Gamma(r/2)} \left(\frac{1}{(t^2/r + 1)}\right)^{\frac{1}{2}r + \frac{1}{2}} \end{aligned}$$

Then

$$f(t) \equiv \frac{1}{\sqrt{\pi}} \sqrt{\frac{1}{r}} \frac{\Gamma\left(\frac{1}{2}r + \frac{1}{2}\right)}{\Gamma(r/2)} \left(\frac{1}{(t^2/r + 1)}\right)^{\frac{1}{2}r + \frac{1}{2}}$$

is the density for the  $T$  distribution. Here  $t \in \mathbb{R}$ . Here is a graph of  $F(x) = P(X \leq x)$  for  $X$  distributed as a  $T$  distribution in which  $r = 10$ .



To get this graph I was tricky. I wanted to integrate from  $-\infty$  to some positive point. I used the fact that the density function is even. I didn't want to consider  $\int_{-\infty}^x f(t) dt$  so instead considered  $.5 + \int_0^x f(t) dt$  for  $x \geq 0$  and then this gave the right thing for positive  $x$ . A similar adjustment took care of the graph for  $x < 0$ . In the syntax, you can pick  $r$ . I have shown it for  $r = 10$ .

```
hold on
for n=1:1:1000
r=10;
a=((r*pi)^(-1/2))*(gamma(.5*(r+1))/gamma(r/2));
f=@(t)[a*(((t.^2)/r)+1).^(-.5*(r+1))];
y=integral(f,0,n*.01);
hold on
plot(n*.01,y+.5,'.','Linewidth',2,'color','black')
plot(-n*.01,-y+.5,'.','Linewidth',2,'color','black')
end
```

If you wanted a table of  $x \rightarrow P(X \leq x)$ , you can do the following.

```
hold on
T=[]; r=10;
for n=1:1:1500
a=((r*pi)^(-1/2))*(gamma(.5*(r+1))/gamma(.5*r));
f=@(t)[a*(((t.^2)/r)+1).^(-.5*(r+1))];
x=-10+(n*.1);
y=integral(f,-10,x);
T=[T; x y];
end
T
```

This will produce a table for the  $T$  distribution with  $r = 10$ . You can follow the same pattern to get a table for other values of  $r$ . Just change the statement  $r = 10$  to  $r = 5$  for example. I started the integral at  $-10$  because if  $x < -10$ ,  $P(X \leq x)$  is considered 0 due to round off error so there is no point in trying to take  $\int_{-\infty}^x f(t) dt$  when  $\int_{-10}^x f(t) dt$  is going to give the same thing as far as can be ascertained. You might want to change where you start the integral depending on  $r$ . The table should start at 0 and end at 1 or something close to it.

### 39.2.2 Confidence Intervals for the Mean

Say you have  $r + 1$  independent samples from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Earlier a method was given which would allow one to give a confidence interval for the variance. Now a method will be given for finding a confidence interval for the mean.

This involves the  $T$  distribution. We will assume  $X_k, k \leq r+1$  is an independent sample from a normal distribution having mean  $\mu$  and variance  $\sigma^2$ .

**Lemma 39.2.1** *Let  $X_k$  be  $r+1$  independent random variables normally distributed with mean  $\mu$  and variance  $\sigma^2$ . Then*

$$\frac{1}{\sqrt{r+1}} \sum_{k=1}^{r+1} \frac{X_k - \mu}{\sigma}$$

*is normally distributed with mean 0 and variance 1.*

**Proof:**

$$E \left( \exp \left( t \frac{1}{\sqrt{r+1}} \sum_{k=1}^{r+1} \frac{X_k - \mu}{\sigma} \right) \right) = E \left( \prod_{k=1}^{r+1} \frac{t}{\sqrt{r+1}} \left( \frac{X_k - \mu}{\sigma} \right) \right)$$

Now recall that  $\frac{X_k - \mu}{\sigma}$  is normally distributed with mean 0 and variance 1. Therefore, by independence, this equals

$$\prod_{k=1}^{r+1} \exp \left( -\frac{1}{2} \frac{t^2}{r+1} \right) = \exp \left( -\frac{1}{2} t^2 \right)$$

which is the moment generating function for a normally distributed random variable with mean 0 and variance 1. ■

Recall that  $(r+1)S^2/\sigma^2$  is  $\mathcal{X}^2(r)$ . By the above discussion of the  $T$  distribution,

$$\frac{\frac{1}{\sqrt{r+1}} \sum_{k=1}^{r+1} \frac{X_k - \mu}{\sigma}}{\sqrt{\frac{\sum_{k=1}^{r+1} (X_k - \bar{X})^2}{r\sigma^2}}}$$

is a  $T$  random variable with  $r$  degrees of freedom discussed above. However, this expression simplifies quite a bit. It becomes

$$\frac{\frac{\sqrt{r}}{\sqrt{r+1}} \sum_{k=1}^{r+1} (X_k - \mu)}{\sqrt{\sum_{k=1}^{r+1} (X_k - \bar{X})^2}} = \frac{\frac{\sqrt{r}}{\sqrt{r+1}} (r+1)(\bar{X} - \mu)}{\sqrt{\sum_{k=1}^{r+1} (X_k - \bar{X})^2}} = \frac{\sqrt{r}\sqrt{r+1}(\bar{X} - \mu)}{\sqrt{\sum_{k=1}^{r+1} (X_k - \bar{X})^2}}$$

Notice how the  $\sigma$  disappeared leaving only  $\mu$ .

**Example 39.2.2** *Here are 11 numbers from an independent random sample of a normal distribution having variance  $\sigma^2$  and mean  $\mu$ . Find a .95 confidence interval for the mean  $\mu$ . The numbers are*

3, 4, 5, 6, 2, 3.5, 5, 4, 6, 2, 4.2

After some computations, we find  $\bar{X} = 4.0636$  and  $11S^2 = 19.245$ . Then the statistic above is of the form

$$\frac{\sqrt{10}\sqrt{11}(4.0636 - \mu)}{\sqrt{19.245}}$$

Using the data cursor in the graph of the function  $F(x) = P(X \leq x)$  for  $X$  a  $T$  random variable with  $r = 10$ , we can find an interval corresponding to probability at least .95. A



point on this graph is (2.48, .9837). Another point is (−2.52, .01519) and so the probability that  $X$  is in this interval is  $.9837 - .01519 = 0.96851$ . Thus with probability at least .95

$$-2.52 \leq \frac{\sqrt{10}\sqrt{11}(4.0636 - \mu)}{\sqrt{19.245}} \leq 2.48$$

Of course you could arrange the interval to be symmetric about 0 because the distribution is symmetric. I just used the data cursor to identify a couple of points. Thus with probability at least .95,

$$1.0541 \geq (\mu - 4.0636) \geq -1.0373$$

$$5.1177 \geq \mu \geq 3.0263$$

Note that theoretically we could have used

$$\frac{\frac{X_1 - \mu}{\sigma}}{\sqrt{\frac{\sum_{k=1}^{r+1} (X_k - \bar{X})^2}{r\sigma^2}}}$$

but this would not give us such a good result because instead of dividing by  $\sqrt{r}\sqrt{r+1}$  we would end up dividing by only  $\sqrt{r}$ . The interval would be much longer. Of course this is not surprising. If you use more information, you should get better results. You might try this to see what happens.

**PROCEDURE 39.2.3** *To find a .95 confidence interval for the mean of a normal distribution having variance  $\sigma^2$  and mean  $\mu$  based on a random sample  $X_1, \dots, X_n$ , do the following.*

1. *Using a table or graph, determine an interval  $[a, b]$  such that for  $X$  distributed as a  $T$  random variable with  $r = n - 1$ , such that  $P(X \in [a, b]) \geq .95$ .*
2. *Find the sample mean  $\bar{X} \equiv \frac{1}{n} \sum_{k=1}^n X_k$ .*
3. *The .95 confidence interval for  $\mu$  is determined by solving the following inequality for  $\mu$ .*

$$a \leq \frac{\sqrt{n-1}\sqrt{n}(\bar{X} - \mu)}{\sqrt{\sum_{k=1}^n (X_k - \bar{X})^2}} \leq b$$

4. *Thus the .95 confidence interval is*

$$\bar{X} - \frac{a\sqrt{\sum_{k=1}^n (X_k - \bar{X})^2}}{\sqrt{n-1}\sqrt{n}} \geq \mu \geq \bar{X} - \frac{b\sqrt{\sum_{k=1}^n (X_k - \bar{X})^2}}{\sqrt{n-1}\sqrt{n}}$$

*If you want some other probability than .95, just find  $[a, b]$  associated with this other probability for  $X \in [a, b]$  where  $X$  is  $\mathcal{X}^2(n-1)$  and do the same thing.*

### 39.2.3 Testing For Two Different Means

Suppose you have two different normal distributions and you take samples from each. You might ask whether the two means are different. This involves the general notion of hypothesis testing. First of all, there is no way you will ever know that two means are exactly equal based on random samples from the two, but you might be able to conclude that you are very sure that the two are not equal. The following is some general terminology.

**Definition 39.2.4** *The hypothesis to be tested is called the null hypothesis, often denoted as  $H_0$ . For example, you might have equality of two means be the null hypothesis. Rejection of the hypothesis depends on whether some statistic, depending on the validity of  $H_0$  is in a region for which we agree to reject the hypothesis. Usually this is done based on the probability of the statistic being in this region. The set of values for which we don't reject the hypothesis is called the acceptance region.*

**Lemma 39.2.5** *Let  $X, Y$  be independent random variables taken from two different normal distributions, respectively  $n(\mu_1, \sigma_1^2)$  and  $n(\mu_2, \sigma_2^2)$ . Then  $X - Y$  is  $n(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ .*

**Proof:** Let  $M(t)$  be the moment generating function of  $X - Y$ .

$$\begin{aligned} M(t) &\equiv E(\exp(t(X - Y))) = E(\exp(tX)\exp(-tY)) \\ &= E(\exp(tX))E(\exp(-tY)) \\ &= e^{t\mu_1}e^{\frac{1}{2}t^2\sigma_1^2}e^{-t\mu_2}e^{\frac{1}{2}t^2\sigma_2^2} = e^{t(\mu_1 - \mu_2)}e^{\frac{1}{2}t^2(\sigma_1^2 + \sigma_2^2)} \end{aligned}$$

which is the moment generating function of a random variable with distribution

$$n(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

■

Now suppose that  $\{X_k\}_{k=1}^{r+1}$  and  $\{Y_k\}_{k=1}^{r+1}$  are two random samples taken from  $n(\mu_1, \sigma_1^2)$  and  $n(\mu_2, \sigma_2^2)$  respectively. Thus from Lemma 39.2.5,

$$X_k - Y_k$$

is  $n(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ . Letting  $Z_k \equiv X_k - Y_k$ , then it follows as in Lemma 39.2.1

$$\frac{1}{\sqrt{r+1}} \sum_{k=1}^{r+1} \frac{Z_k - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}} \text{ is } n(0, 1)$$

Then, just as before,

$$\frac{\frac{1}{\sqrt{r+1}} \sum_{k=1}^{r+1} \frac{Z_k - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}}{\sqrt{\frac{\sum_{k=1}^{r+1} (Z_k - \bar{Z})^2}{r(\sigma_1^2 + \sigma_2^2)}}}$$

is a  $T$  random variable with parameter equal to  $r, T(r)$ . As before, we can simplify this to obtain that the following is a  $T$  random variable.

$$\frac{\sqrt{r}\sqrt{r+1}(\bar{Z} - (\mu_1 - \mu_2))}{\sqrt{\sum_{k=1}^{r+1} (Z_k - \bar{Z})^2}} \quad (39.4)$$

Let  $H_0$  be the hypothesis that  $\mu_1 = \mu_2$ . Then with this assumption, the above  $T(r)$  is

$$T \equiv \frac{\sqrt{r}\sqrt{r+1}\bar{Z}}{\sqrt{\sum_{k=1}^{r+1} (Z_k - \bar{Z})^2}} \quad (39.5)$$

**Example 39.2.6** You have two random samples from normal distributions. The first  $\{X_k\}$  is  $(2, 3, -2, -5, 7, 9)$ . The second  $\{Y_k\}$  is  $(-2, -4, 1, 3, 4, 5)$  these taken in the order indicated. Then the corresponding list of normal random variables  $Z_k \equiv X_k - Y_k$  is

$$(4, 7, -3, -8, 3, 4)$$

Lets agree to reject  $H_0$  that the two means are equal if  $T$  in 39.5 is either too large or too small, meaning that  $T$  is in a region which is associated with small probability, thus the imperative to reject the Hypothesis. Let  $a$  be such that  $P(T > a) < .05$ . Then reject  $H_0$  if  $T$  is either larger than  $a$  or smaller than  $-a$ .

First find  $\bar{Z}$ . It equals 1.1667. Next find  $\sqrt{\sum_{k=1}^{r+1} (Z_k - \bar{Z})^2}$ . It equals 12.443. Thus  $T =$

$$\frac{\sqrt{5}\sqrt{6}(1.1667)}{12.443} = .51356$$

Now we need to go to a table or use MATLAB or something to find out information about  $T(5)$ . From a table,  $P(T > 2.015) = .05$  and so we do not reject  $H_0$ . In other words, we “accept” the hypothesis that the two means are equal.

**PROCEDURE 39.2.7** To test the hypothesis  $H_0$  that two means from two different normal distributions are equal, do the following:

1. Take random samples  $(X_1, Y_1), (X_2, Y_2), \dots, (X_{r+1}, Y_{r+1})$  where  $X_k$  is from  $n(\mu_1, \sigma_1^2)$  and  $Y_k$  is from  $n(\mu_2, \sigma_2^2)$ .
2. Letting  $T(r)$  be a  $T$  random variable with parameter  $r$ , determine  $a$  such that

$$P(|T(r)| > a)$$

is smaller than .05. (You could pick any other number in  $(0, 1)$  here depending on how sure you want to be that a rejection of  $H_0$  is warranted.)

3. Compute

$$\frac{\sqrt{r}\sqrt{r+1}\bar{Z}}{\sqrt{\sum_{k=1}^{r+1} (Z_k - \bar{Z})^2}}$$

where  $Z_k = X_k - Y_k$  and  $\bar{Z}$  is the average of the  $Z_k$ .

4. Reject  $H_0$  if  $|T(r)| > a$ . Otherwise “Accept  $H_0$ ”.

Now what would be the outcome if we looked for a confidence interval.

**Example 39.2.8** Find a .54 confidence interval for  $|\mu_1 - \mu_2|$  in the above example.

Here we use 39.4. From MATLAB, for  $T$  the  $T$  statistic with  $r = 5$  being used here,

$$P(|T| < .8) = .54$$

Thus with probability .54 you get

$$-.8 < \frac{\sqrt{r}\sqrt{r+1}(\bar{Z} - (\mu_1 - \mu_2))}{\sqrt{\sum_{k=1}^{r+1} (Z_k - \bar{Z})^2}} < .8$$

$$.8 > \frac{\sqrt{r}\sqrt{r+1}((\mu_1 - \mu_2) - \bar{Z})}{\sqrt{\sum_{k=1}^{r+1} (Z_k - \bar{Z})^2}} > -.8$$

Then

$$.8 > \frac{\sqrt{30}((\mu_1 - \mu_2) - 1.1667)}{12.443} > -.8$$

$$2.9841 > (\mu_1 - \mu_2) > -.65072$$

with probability .54. By replacing .54 with a smaller number, this could be changed and we could conclude with a reasonable probability that  $\mu_1 - \mu_2 > 0$ . Thus not rejecting  $H_0$  isn't really the same as saying that  $H_0$  is true.

### 39.2.4 The $F$ Distribution

In this case, you have two independent random variables  $U, V$  which are respectively  $\mathcal{X}^2(r_1)$  and  $\mathcal{X}^2(r_2)$ . Thus the density for the random variable  $(U, V)$  is

$$\frac{1}{\Gamma(r_1/2) 2^{r_1/2}} u^{(r_1/2)-1} e^{-u/2} \frac{1}{\Gamma(r_2/2) 2^{r_2/2}} v^{(r_2/2)-1} e^{-v/2}, u, v > 0$$

Here we consider the density function for

$$F = \frac{U/r_1}{V/r_2}$$

Change the variables as done above.

$$f = \frac{ur_2}{vr_1}, k = v$$

inverting the transformations gives

$$u = \frac{fvr_1}{r_2}, v = k, \begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{r} \begin{pmatrix} f \\ k \end{pmatrix}$$

$$J(u, v) = \left| \det \begin{pmatrix} \frac{1}{vr_1} r_2 & -\frac{u}{v^2 r_1} r_2 \\ 0 & 1 \end{pmatrix} \right| = \frac{1}{vr_1} r_2$$

Then by the change of variables formula, and letting  $g$  denote the density for  $(F, K)$ ,

$$\int_{\mathbf{r}(U)} g(f, k) df dk = \int_U g\left(\frac{ur_2}{vr_1}, v\right) |J(u, v)| du dv$$

$$= \int_U g\left(\frac{ur_2}{vr_1}, v\right) \frac{1}{vr_1} r_2 du dv$$

However, the left side is

$$\int_U \frac{1}{\Gamma(r_1/2) 2^{r_1/2}} u^{(r_1/2)-1} e^{-u/2} \frac{1}{\Gamma(r_2/2) 2^{r_2/2}} v^{(r_2/2)-1} e^{-v/2} du dv$$

because the probability that  $(f, k)$  is in  $\mathbf{r}(U)$  is the same as the probability that  $(u, v)$  is in  $U$ . Thus

$$g\left(\frac{ur_2}{vr_1}, v\right) = \frac{vr_1}{r_2} \frac{1}{\Gamma(r_1/2) 2^{r_1/2}} u^{(r_1/2)-1} e^{-u/2} \frac{1}{\Gamma(r_2/2) 2^{r_2/2}} v^{(r_2/2)-1} e^{-v/2}$$

Now write in terms of  $f, k$

$$g(f, k) = \frac{kr_1}{r_2} \frac{1}{\Gamma(r_1/2) 2^{r_1/2}} \left(\frac{fkr_1}{r_2}\right)^{(r_1/2)-1} e^{-\left(\frac{fkr_1}{2r_2}\right)} \frac{1}{\Gamma(r_2/2) 2^{r_2/2}} k^{(r_2/2)-1} e^{-k/2}$$

Of course  $k \in (0, \infty)$  and so if we want the density of  $F$ , all that is needed is to integrate the above from 0 to  $\infty$  with respect to  $k$ . Then this integral is

$$\frac{1}{\Gamma(r_1/2) 2^{r_1/2}} \frac{1}{\Gamma(r_2/2) 2^{r_2/2}} f^{(r_1/2)-1} \left(\frac{r_1}{r_2}\right)^{r_1/2} \int_0^\infty k^{\left(\frac{r_1+r_2}{2}-1\right)} e^{-\left(\frac{fr_1}{2r_2} + \frac{1}{2}\right)k} dk$$

Change the variable in the integral. Let  $u = \left(\frac{fr_1}{2r_2} + \frac{1}{2}\right)k$  so

$$dk = \frac{du}{\left(\frac{fr_1}{2r_2} + \frac{1}{2}\right)}$$

then the integral is

$$\begin{aligned} & \int_0^\infty \left(\frac{u}{\left(\frac{fr_1}{2r_2} + \frac{1}{2}\right)}\right)^{\left(\frac{r_1+r_2}{2}-1\right)} e^{-u} \frac{du}{\left(\frac{fr_1}{2r_2} + \frac{1}{2}\right)} \\ &= \frac{1}{\left(\frac{fr_1}{2r_2} + \frac{1}{2}\right)^{\frac{r_1+r_2}{2}}} \int_0^\infty u^{\left(\frac{r_1+r_2}{2}-1\right)} e^{-u} du \\ &= \frac{1}{\left(\frac{fr_1}{2r_2} + \frac{1}{2}\right)^{\frac{r_1+r_2}{2}}} \Gamma\left(\frac{r_1+r_2}{2}\right) \end{aligned}$$

thus we end up with the following for the density for  $F$ .

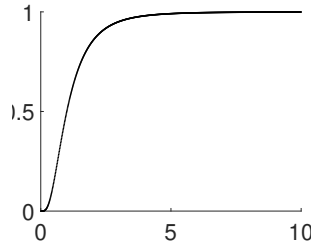
$$\frac{1}{\Gamma(r_1/2) 2^{r_1/2}} \frac{1}{\Gamma(r_2/2) 2^{r_2/2}} f^{(r_1/2)-1} \left(\frac{r_1}{r_2}\right)^{r_1/2} \frac{\Gamma\left(\frac{r_1+r_2}{2}\right)}{\left(\frac{fr_1}{2r_2} + \frac{1}{2}\right)^{\frac{r_1+r_2}{2}}}$$

$$= \frac{\Gamma\left(\frac{r_1+r_2}{2}\right) \left(\frac{r_1}{r_2}\right)^{r_1/2}}{\Gamma(r_1/2)\Gamma(r_2/2)} \frac{f^{(r_1/2)-1}}{\left(\frac{fr_1}{r_2} + 1\right)^{\frac{r_1+r_2}{2}}}, f > 0$$

Note that if  $r_1 = r_2 = r$ , this is much less ugly. It then reduces to

$$\frac{\Gamma(r)}{\Gamma(r/2)^2} \frac{f^{(r/2)-1}}{(f+1)^r}$$

You can probably see that this  $F$  distribution could be used to test the ratio of variances coming from two normal densities and obtain a confidence interval for this ratio. If this interval did not contain 1, then you could conclude that with a certain probability the two variances are different. Here is a graph of  $F(x) \equiv P(X \leq x)$  where  $X$  is an  $F$  random variable with  $r = r_1 = r_2 = 10$ .



### 39.2.5 Confidence Intervals for the Ratio of Two Variances

Suppose you have two random samples of length  $r$  taken from two normal distributions having variances  $\sigma_1^2$  and  $\sigma_2^2$ . The symbols which represent such normal distributions are  $n(\mu_i, \sigma_i^2)$ . Let these be  $\{X_1, \dots, X_r\}$  from  $n(\mu_1, \sigma_1^2)$  and  $\{Y_1, \dots, Y_r\}$  from  $n(\mu_2, \sigma_2^2)$ . Let

$$rS_1^2 \equiv \sum_{k=1}^r (X_k - \bar{X})^2, \quad rS_2^2 \equiv \sum_{k=1}^r (Y_k - \bar{Y})^2$$

Then the  $rS_i^2/\sigma_i^2$  is  $\chi^2(r-1)$ . It follows from the above that

$$\frac{rS_1^2/\sigma_1^2}{rS_2^2/\sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2} \frac{rS_1^2}{rS_2^2}$$

is distributed as an  $F$  random variable with the parameter equal to  $r-1$ .

**Example 39.2.9** Find a .9 confidence interval for the ratio  $\frac{\sigma_2^2}{\sigma_1^2}$  where you have two random samples, of length 11 taken respectively from two normal distributions.  $n(\mu_1, \sigma_1^2)$  and  $n(\mu_2, \sigma_2^2)$ . These samples are  $\{-3, 2, -1, 0, 1, -2, .5, .4, -.2, -.5, .3\}$  and  $\{-4, -7, 7, 10, 15, 5, -8, 11, 12, -12, -5\}$ . You can see that the second sample is much more spread out than the first. Thus, they should have different variances. Does the confidence interval predict this?

Some computations show that  $11S_1^2 = 19.22$  and  $11S_2^2 = 906.64$ . Now, from the graph of  $F(x) \equiv P(X \leq x)$  where  $X$  has  $F$  distribution with  $r_1 = r_2 = r = 10$  given above, using

the data cursor, two ordered pairs on this curve are  $(.33, .047)$  and  $(3.01, .9516)$ . Thus the associated probability for  $X$  in  $(.33, 3.01)$  is  $.9516 - .047 = 0.9046$ . Thus

$$.33 \leq \frac{\sigma_2^2}{\sigma_1^2} \frac{19.22}{906.64} \leq 3.01$$

with probability larger than .9. Thus, with probability larger than .9,

$$15.567 \leq \frac{\sigma_2^2}{\sigma_1^2} \leq 141.99$$

It is obvious from this that the ratio is much larger than 1 so, just as you might have guessed, the two variances are very different. In fact, we could have asserted this with much higher probability.

From the data cursor, we find the ordered pair  $(.16, .0038)$ . thus the probability that

$$.16 \leq \frac{\sigma_2^2}{\sigma_1^2} \frac{19.22}{906.64}$$

is no more than  $1 - .0038 = 0.9962$ . Therefore, the probability

$$.16 \left( \frac{19.22}{906.64} \right)^{-1} = 7.5475 \leq \frac{\sigma_2^2}{\sigma_1^2}$$

is at least .9962. There can be no doubt that the second variance is much larger than the first. Of course we would have thought that, but in general, the samples might not exhibit such extreme differences in how spread out they are.

If you believe that the two means are the same, then you can add one degree of freedom to the  $\mathcal{X}^2$  distributions and replace  $rS^2$  with  $\sum_{k=1}^r (X_k - \mu)^2$ . This means you can get better confidence intervals, but why should you believe this? I think that in general, you wouldn't know this, so I have emphasized the case where the means are not known and the sample mean is used instead. There is seemingly no end to complicated tests on statistics which can be used to draw conclusions about the parameters of underlying distributions. This book is not the place to explore each and every such technical procedure. To do this, you should see specialized texts on statistics.

### PROCEDURE 39.2.10 *Suppose you have two normal distributions*

$$n(\mu, \sigma^2), n(\hat{\mu}, \hat{\sigma}^2)$$

and two random samples  $\{X_1, \dots, X_n\}, \{Y_1, \dots, Y_n\}$  respectively from these two distributions. To find a .95 confidence interval for the ratio  $\frac{\hat{\sigma}^2}{\sigma^2}$ , do the following.

1. For  $X$  a random variable distributed as an  $F$  distribution with  $r = r_1 = r_2 = n - 1$ , determine an interval  $[a, b]$  such that  $P(X \in [a, b]) \geq .95$ .
2. Determine the sample means  $\bar{X} \equiv \sum_{k=1}^n X_k$ ,  $\bar{Y} \equiv \sum_{k=1}^n Y_k$ . Then find

$$nS^2 \equiv \sum_{k=1}^n (X_k - \bar{X})^2, \quad n\hat{S}^2 \equiv \sum_{k=1}^n (Y_k - \bar{Y})^2$$

3. The confidence interval is determined by solving for the ratio  $\frac{\hat{\sigma}^2}{\sigma^2}$  in the inequality

$$a \leq \frac{\hat{\sigma}^2}{\sigma^2} \frac{nS^2}{n\hat{S}^2} \leq b$$

4. Thus the confidence interval is

$$a \frac{n\hat{S}^2}{nS^2} \leq \frac{\hat{\sigma}^2}{\sigma^2} \leq b \frac{n\hat{S}^2}{nS^2}$$

or in other words,

$$a \frac{\sum_{k=1}^n (Y_k - \bar{Y})^2}{\sum_{k=1}^n (X_k - \bar{X})^2} \leq \frac{\hat{\sigma}^2}{\sigma^2} \leq b \frac{\sum_{k=1}^n (Y_k - \bar{Y})^2}{\sum_{k=1}^n (X_k - \bar{X})^2}$$

You do the same thing if you want a different probability. Just identify a different interval corresponding to the different probability and do the above.

### 39.3 Maximum Likelihood Estimates

These estimates give a simple way to estimate various parameters. Unlike the above material on confidence intervals and hypothesis testing, you don't get from these procedures a confidence interval associated with a probability that the parameter is in this interval or a direction to reject a hypothesis. You just get the best estimate for the parameter in terms of maximizing likelihood. I think it is best to illustrate the technique using specific examples.

**Example 39.3.1** You know a random variable is a binomial random variable. Thus

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

where  $q = 1 - p$ . Find the maximum likelihood estimate for  $p$ .

What you do is to write down the “likelihood” obtained by taking a sample  $X_1, \dots, X_m$ . Then the likelihood is

$$L(p) \equiv \prod_{k=1}^m p^{X_k} (1-p)^{1-X_k}$$

To find an estimate, you seek to pick  $p$  in order to maximize this likelihood. Obviously it would be better to maximize  $\ln(L(p))$  which equals

$$\ln(L(p)) = \sum_{k=1}^m [X_k \ln(p) + (1 - X_k) \ln(1 - p)]$$

Then from beginning calculus, we take a derivative with respect to  $p$  and set equal to 0 and solve for  $p$ . This is the maximum likelihood estimate for  $p$ .

$$\sum_{k=1}^m \frac{X_k}{p} + (1 - X_k) \frac{-1}{1-p} = 0$$



Thus

$$\begin{aligned}\frac{1}{p} \left( \sum_{k=1}^m X_k \right) &= \frac{1}{1-p} \sum_{k=1}^m 1 - X_k = \frac{m}{1-p} - \left( \sum_{k=1}^m X_k \right) \frac{1}{1-p} \\ \frac{1}{p(1-p)} \left( \sum_{k=1}^m X_k \right) &= \frac{m}{1-p}\end{aligned}$$

and so

$$\frac{1}{p} \left( \sum_{k=1}^m X_k \right) = m, \quad p = \frac{1}{m} \sum_{k=1}^m X_k$$

Surely this makes sense. Recall that  $p$  was the probability of a success in a Bernoulli trial and the random variable  $X$  is the sum of these successes in  $n$  trials. To emphasize that this is an estimate, people will write

$$\hat{p} = \frac{1}{m} \sum_{k=1}^m X_k.$$

The above trick in which you maximize  $\ln(L)$  is typically used. It is generally a good idea because the likelihood involves taking a product and when you take the  $\ln$  of a product, you end up with a sum which is a lot easier to work with than the original product.

**Example 39.3.2** Find a maximum likelihood estimate for  $\mu$  and  $\sigma$  based on a random sample  $X_1, \dots, X_n$  taken from a normal distribution.

In this and other cases of random variables having a density function,  $f(x)$ , you choose the parameters to maximize the likelihood  $\prod_{k=1}^n f(X_k)$ . Thus, in this case, you maximize

$$L(\mu, \sigma) \equiv \prod_{k=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} (X_k - \mu)^2\right)$$

You can delete the  $1/\sqrt{2\pi}$ . Then maximize the  $\ln$  of this. Thus you want to maximize

$$\sum_{k=1}^n -\ln(\sigma) + \left(-\frac{1}{2\sigma^2} (X_k - \mu)^2\right)$$

First take the partial with respect to  $\mu$ . Cancelling out the  $\sigma^2$ ,

$$\sum_{k=1}^n (X_k - \mu) = 0 = \sum_{k=1}^n X_k - n\mu$$

and so

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n X_k \equiv \bar{X}$$

which is the sample mean. Next take partial derivative with respect to  $\sigma$

$$\sum_{k=1}^n -\frac{1}{\sigma} + \frac{1}{\sigma^3} (X_k - \mu)^2 = 0$$

Thus, from the first part where  $\hat{\mu}$  was found,

$$\sigma^2 n = \sum_{k=1}^n (X_k - \bar{X})^2$$

and so

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2 \equiv S^2$$

This is the maximum likelihood estimate for the variance.

It should be noticed that there is a problem with this. The estimate is biased. This means that  $E(\hat{\sigma}^2) = E(S^2) \neq \sigma^2$ . To see this, recall that it was shown above that  $nS^2/\sigma^2$  is a  $\mathcal{X}^2(n-1)$  random variable. We know the moment generating function of such a random variable. It is

$$M(t) = \frac{1}{(1-2t)^{(n-1)/2}}$$

and so we can find the expectation of  $nS^2/\sigma^2$ .

$$M'(t) = \frac{2}{(1-2t)^{\frac{1}{2}n+\frac{1}{2}}} \left( \frac{1}{2}n - \frac{1}{2} \right)$$

Now letting  $t = 0$ , you get  $E(nS^2/\sigma^2) = n-1$ . Thus  $E\left(\frac{S^2}{\sigma^2}\right) = \frac{n-1}{n} \neq 1$ . For this reason, people often use  $\frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2$  as an estimate for the variance.

## 39.4 Quadratic Forms

When you have a symmetric real matrix  $A = A^T$ , a quadratic form is an expression of the form  $\mathbf{x}^T A \mathbf{x}$ . What is considered here are two symmetric, real matrices  $A, B$  which are  $n \times n$  and independent random variables  $X_1, \dots, X_n$  which have identical normal distribution  $n(0, \sigma^2)$ . For example, it is a random sample from such a normal distribution. The question of interest is whether  $\mathbf{X}^T A \mathbf{X}$  and  $\mathbf{X}^T B \mathbf{X}$  are independent random variables.

Why might this be of interest? Recall from Corollary 38.8.8, the distribution of  $\frac{(X_k - \mu)^2}{\sigma^2}$  is  $\mathcal{X}^2(1)$  and since these are independent, the distribution of  $\sum_{k=1}^n \frac{(X_k - \mu)^2}{\sigma^2}$  is  $\mathcal{X}^2(n)$ . This is a quadratic form in the independent variables  $\frac{X_k - \mu}{\sigma}$  in which the symmetric matrix is just  $I$ . It was important earlier to consider  $\bar{X}$  and the random vector

$$\begin{pmatrix} (X_1 - \bar{X}) & \cdots & (X_n - \bar{X}) \end{pmatrix}$$

and it was shown, using the special form of the normal distribution that this random variable and random vector are independent. This is what made it possible to determine the moment generating function and distribution of  $nS^2/\sigma^2$  which made possible a whole collection of statistical tests and motivated the  $T$  and  $F$  distributions. However, what was really needed were independence of the quadratic forms  $\sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2}$  and  $\bar{X}^2$  being independent. Thus we have already been using quadratic forms evaluated at random samples of the normal distribution.

The idea now is to just extend this to more general situations in which the symmetric matrix is perhaps not  $I$ . To do this, I will first consider the moment generating function for  $\mathbf{X}^T A \mathbf{X}$  where  $A$  is symmetric and  $\mathbf{X}$  is a random vector whose components are distributed as  $n(0, \sigma^2)$ . To save space let  $dx_1 \cdots dx_n = d\vec{x}$

$$M(t) \equiv \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \int_{\mathbb{R}^n} e^{t\mathbf{x}^T A \mathbf{x}} e^{-\frac{1}{2} \frac{\mathbf{x} \cdot \mathbf{x}}{\sigma^2}} dx_1 \cdots dx_n$$

$$= \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^n \int_{\mathbb{R}^n} e^{\sigma^2 t \frac{\mathbf{x}^T A \mathbf{x}}{\sigma^2}} e^{-\frac{1}{2} \frac{\mathbf{x} \cdot \mathbf{x}}{\sigma^2}} d\vec{x} = \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^n \int_{\mathbb{R}^n} e^{-\frac{1}{2\sigma^2} \mathbf{x}^T (I - 2\sigma^2 t A) \mathbf{x}} d\vec{x}$$

Now let  $\mathbf{y} \equiv U^T \mathbf{x}$  where  $U$  is an orthogonal matrix such that  $U^T (I - 2\sigma^2 t A) U = D$ , a diagonal matrix having all positive diagonal entries. We can get such a thing whenever  $|t|$  is small enough because then the expression  $I - 2\sigma^2 t A$  will have all positive eigenvalues. Now  $\det(U) = \pm 1$  because it is orthogonal. Changing variables to  $\mathbf{y}$  and using the change of variables formula,

$$\begin{aligned} M(t) &= \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^n \int_{\mathbb{R}^n} e^{-\frac{1}{2\sigma^2} \mathbf{y}^T U^T (I - 2\sigma^2 t A) U \mathbf{y}} d\vec{y} \\ &= \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^n \int_{\mathbb{R}^n} e^{-\frac{1}{2\sigma^2} \mathbf{y}^T D(t) \mathbf{y}} d\vec{y} \end{aligned}$$

where  $D(t)$  is the diagonal matrix which has the positive eigenvalues  $\lambda_k^2(t)$  down the diagonal. Then the above expression splits into factors of the form

$$\frac{1}{\sqrt{2\pi\sigma}} \int_{\mathbb{R}} e^{-\frac{1}{2\sigma^2} y_k^2 \lambda_k^2} dy_k$$

So let  $u = \lambda_k y_k$  and this becomes

$$\frac{1}{\sqrt{2\pi\sigma}} \frac{1}{\lambda_k} \int_{\mathbb{R}} e^{-\frac{1}{2\sigma^2} u^2} du = \frac{1}{\lambda_k}.$$

Hence,

$$\begin{aligned} M(t) &= \prod_{k=1}^n \frac{1}{\lambda_k} = \left( \prod_{k=1}^n \frac{1}{\lambda_k^2} \right)^{1/2} \\ &= \left( \frac{1}{\det(D(t))} \right)^{1/2} = \frac{1}{\det(I - 2\sigma^2 t A)^{1/2}}. \end{aligned}$$

Remember the determinant is the product of the eigenvalues of the matrix and  $I - 2\sigma^2 t A$  and  $D$  are similar so they have the same eigenvalues, namely the diagonal entries of  $D$ . See Corollary 11.4.5 to Schur's theorem. This is stated as the following lemma.

**Lemma 39.4.1** *Let  $A$  be symmetric and let  $X_1, \dots, X_n$  be independent and  $n(0, \sigma^2)$ . Then the moment generating function of  $\mathbf{X}^T A \mathbf{X}$  is  $M(t) = \det(I - 2\sigma^2 t A)^{-1/2}$  for all  $|t|$  sufficiently small.*

Now suppose you have two symmetric matrices  $A, B$  and independent  $n(0, \sigma^2)$  random variables  $X_1, \dots, X_n$ . Then there are two random variables  $\mathbf{X}^T A \mathbf{X}, \mathbf{X}^T B \mathbf{X}$  and we want to determine when these two are independent. Then using similar reasoning to the above, it follows that for  $|s|, |t|$  both small enough,

$$M(t, s) \equiv E(\exp(t \mathbf{X}^T A \mathbf{X} + s \mathbf{X}^T B \mathbf{X})) = \frac{1}{\det(I - 2\sigma^2 t A - 2\sigma^2 s B)^{1/2}}$$

If  $AB = 0$ , then

$$\begin{aligned} (I - 2\sigma^2 t A)(I - 2\sigma^2 s B) &= I - 2\sigma^2 t A - 2\sigma^2 s B + 4\sigma^4 t s AB \\ &= I - 2\sigma^2 t A - 2\sigma^2 s B \end{aligned}$$

and so

$$\begin{aligned} M(t, s) &= \frac{1}{\det(I - 2\sigma^2 tA - 2\sigma^2 sB)^{1/2}} = \frac{1}{\det(I - 2\sigma^2 tA)^{1/2}} \frac{1}{\det(I - 2\sigma^2 sB)^{1/2}} \\ &= M(t, 0)M(0, s) \end{aligned}$$

which shows that the two quadratic forms  $X^T A X, X^T B X$  are independent.

In fact this is true the other direction. Suppose the two quadratic forms are independent. Thus  $M(t, 0)M(0, s) = M(t, s)$ . Then this requires

$$\frac{1}{\det(I - 2\sigma^2 tA - 2\sigma^2 sB)^{1/2}} = \frac{1}{\det(I - 2\sigma^2 tA)^{1/2}} \frac{1}{\det(I - 2\sigma^2 sB)^{1/2}}$$

and so

$$\begin{aligned} \det(I - 2\sigma^2 tA - 2\sigma^2 sB) &= \det(I - 2\sigma^2 tA) \det(I - 2\sigma^2 sB) \\ &= \det(I - 2\sigma^2 tA - 2\sigma^2 sB + 4\sigma^4 tsAB) \end{aligned}$$

This is to hold for all  $|t|, |s|$  small enough. However, if  $AB \neq 0$ , the polynomial on the right will be of degree  $2n$  while the one on the left will be of degree  $n$ . Therefore, these cannot be equal. The details follow.

From the definition of the determinant, the left side is

$$\sum_{r_1 \cdots r_n} \text{sgn}(r_1 \cdots r_n) (\delta_{1r_1} - 2\sigma^2 tA_{1r_1} - 2\sigma^2 sB_{1r_1}) \cdots (\delta_{nr_1} - 2\sigma^2 tA_{nr_1} - 2\sigma^2 sB_{nr_1})$$

This expression is of the form  $\sum_{p+q \leq n} a_k t^p s^q$ . However, similar reasoning gives

$$\det(I - 2\sigma^2 tA - 2\sigma^2 sB + 4\sigma^4 tsAB)$$

is of the form  $\sum_{p+q \leq 2n} b_k t^p s^q$  and if  $AB \neq 0$ , there will be nonzero terms  $bt^p s^q$  where  $b \neq 0$  and  $p+q = 2n$ . These two polynomials cannot be equal if this happens. Say  $\hat{p} + \hat{q} = 2n$  and the second has a term  $bt^{\hat{p}} s^{\hat{q}}$ . Then one of  $\hat{p}, \hat{q}$  is larger than  $n$ . Say  $\hat{q} > n$ . Differentiate  $\sum_{p+q \leq n} a_k t^p s^q$  with respect to  $s$   $\hat{q} + 1$  times and the result causes  $\det(I - 2\sigma^2 tA - 2\sigma^2 sB)$  to vanish but does not cause  $\det(I - 2\sigma^2 tA - 2\sigma^2 sB + 4\sigma^4 tsAB)$  to vanish. Hence  $AB = 0$ . This proves the following very interesting result.

**Proposition 39.4.2** *Let  $X_1, \dots, X_n$  be a random independent sample from  $n(0, \sigma^2)$  and for*

$$X \equiv \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}^T$$

*and  $A, B$  two symmetric real matrices, then  $X^T A X$  and  $X^T B X$  are independent if and only if  $AB = 0$ .*

Note that for  $A, B$  symmetric,  $AB = 0$  if and only if  $B^T A^T = BA = 0$ .

If you have more than two of these, say  $A_k, k \leq m$  the result would end up being similar although you would need to have  $A_j A_k = 0$  whenever  $j \neq k$ .

As an interesting observation, from linear algebra, this condition that the products give 0 implies that the matrices  $\{A_k\}$  are a commuting family of diagonalizable matrices and so they are simultaneously diagonalizable, meaning that there exists a single invertible matrix  $S$  such that  $S^{-1} A_k S = D_k$  where  $D_k$  is a diagonal matrix. However, more is assumed here in saying that the product is 0. In particular, you can't have a repeated nonzero matrix in the list of matrices.

**Corollary 39.4.3** Let  $X_1, \dots, X_n$  be a random independent sample from  $n(0, \sigma^2)$  and for

$$\mathbf{X} \equiv \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}^T$$

and let  $\{A_k\}_{k=1}^m$  be symmetric real matrices. Then the random variables  $\{\mathbf{X}^T A_k \mathbf{X}\}_{k=1}^m$  are independent if and only if  $A_k A_j = 0$  whenever  $k \neq j$ .

Recall that for  $X_1, \dots, X_n$  independent random variables which are  $n(\mu, \sigma^2)$ , their sum  $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$  is  $\mathcal{X}^2(n)$ . As noted, this is a quadratic form in the independent random variables  $\left\{\frac{X_i - \mu}{\sigma}\right\}_{i=1}^n$ . You just let the symmetric matrix  $A$  be the identity.

What about  $\frac{\mathbf{X}^T A \mathbf{X}}{\sigma^2}$ ? When will this be distributed as a  $\mathcal{X}^2(r)$  random variable? For simplicity, assume the random variables are  $n(0, \sigma^2)$ . It was shown above that the moment generating function for  $\mathbf{X}^T A \mathbf{X}$  is

$$\left(\det(I - 2\sigma^2 t A)^{-1/2}\right)$$

Therefore, the moment generating function of  $\frac{\mathbf{X}^T A \mathbf{X}}{\sigma^2}$  is

$$M(t) \equiv E\left(e^{t \frac{\mathbf{X}^T A \mathbf{X}}{\sigma^2}}\right) = \left(\det\left(I - 2\sigma^2 \frac{t}{\sigma^2} A\right)^{-1/2}\right) = \left(\det(I - 2tA)^{-1/2}\right)$$

Of course, it was shown some time ago that the moment generating function for  $\mathcal{X}^2(r)$  is  $\frac{1}{(1-2t)^{r/2}}$ . Let  $U$  be an orthogonal matrix such that  $U^T A U = D$ , a diagonal matrix. Thus

$$M(t) = \det(I - 2tD)^{-1/2}$$

Now if there is anything other than 1 or 0 on the diagonal of  $D$  then  $M(t)$  cannot possibly be of the form  $\frac{1}{(1-2t)^{r/2}}$ . Lets consider why this is. Suppose the diagonal entries of  $D$  are  $d_1, \dots, d_n$ . Then

$$M(t) = \left(\prod_{i=1}^n (1 - 2td_i)\right)^{-1/2}$$

If you have a factor  $(1 - 2td_i)$  for some  $d_i \notin \{0, 1\}$ , then it simply does not have the right form to be the moment generating function for  $\mathcal{X}^2(r)$ . On the other hand, if each  $d_i$  is either 0 or 1, then  $M(t)$  will have the right form and the  $r$  will be the number of eigenvalues equal to 1, the rank of  $A$ .

Is there a simple way to describe this condition that  $\mathbf{X}^T A \mathbf{X}$  is  $\mathcal{X}^2(r)$ ? Yes there is. The eigenvalues of the symmetric matrix  $A$  are either 1 or 0.

**Lemma 39.4.4** Let  $A$  be a real symmetric matrix. Then  $A^2 = A$  if and only if the eigenvalues of  $A$  are either 0 or 1.

**Proof:** Suppose the eigenvalues are 0 or 1. Since  $A$  is symmetric, there is an orthonormal basis of eigenvectors.  $\{v_k\}_{k=1}^n$ . See Theorem 11.4.7 from the early material on linear algebra. Then say  $A v_k = \lambda v_k$  either  $\lambda$  is 0 or 1 so either  $A v_k = v_k$  or  $A v_k = \mathbf{0}$ . In the first case,  $A^2 v_k = A v_k$  so  $(A^2 - A) v_k = \mathbf{0}$ . In the second case,  $A^2 v_k = A \mathbf{0} = \mathbf{0}$  and so

$(A^2 - A)v_k = \mathbf{0} - \mathbf{0} = \mathbf{0}$ . Thus  $A^2 - A = 0$  because this matrix sends every vector in a basis to  $\mathbf{0}$ .

Conversely, suppose  $A^2 = A$ . Why are all eigenvalues 1 or 0? Say  $Av = \lambda v$  and say  $\lambda \neq 0$ . Then for each  $v$  an eigenvector,  $A^2v = Av = \lambda Av$  and so  $A(1 - \lambda)v = \mathbf{0}$ . If  $\lambda \neq 1$ , then  $Av = \mathbf{0}$  which is assumed not to be so. Hence  $\lambda = 1$ . Thus all eigenvalues are either 0 or 1. ■

This proves the following interesting theorem.

**Theorem 39.4.5** *Let  $X_1, \dots, X_n$  be independent and  $n(0, \sigma^2)$ . Let  $A$  be a real symmetric matrix. Then for  $\mathbf{X} = \begin{pmatrix} X_1 & \dots & X_n \end{pmatrix}^T$ ,  $\frac{\mathbf{X}^T A \mathbf{X}}{\sigma^2}$  is  $\mathcal{X}^2(r)$  for some  $r \leq n$  if and only if  $A^2 = A$  if and only if the eigenvalues of  $A$  are 0 or 1. In fact,  $r$  is the rank of  $A$ .*

At this point, it might be good to recall that the distribution of

$$nS^2/\sigma^2 \equiv \sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2}$$

is  $\mathcal{X}^2(n-1)$  where

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$$

In showing this, first there was some algebra.

$$\overbrace{\sum_{k=1}^n \frac{(X_k - \mu)^2}{\sigma^2}}^{\mathcal{X}^2(n)} = \sum_{k=1}^n \frac{((X_k - \bar{X}) + (\bar{X} - \mu))^2}{\sigma^2}$$

After some simple manipulations,

$$\begin{aligned} &= \sum_{k=1}^n \frac{(X_k - \bar{X})^2}{\sigma^2} + \sum_{k=1}^n \frac{(\bar{X} - \mu)^2}{\sigma^2} = \frac{nS^2}{\sigma^2} + \sum_{k=1}^n \frac{\left(\frac{1}{n} \sum_{j=1}^n (X_j - \mu)\right)^2}{\sigma^2} \\ &= \frac{nS^2}{\sigma^2} + \sum_{k=1}^n \left(\frac{1}{n} \sum_{j=1}^n \frac{(X_k - \mu)}{\sigma}\right)^2 = \left(\overbrace{\sum_{j=1}^n \frac{(X_k - \mu)}{\sqrt{n}\sigma}}^{\mathcal{X}^2(1)}\right)^2 + \frac{nS^2}{\sigma^2} \end{aligned}$$

Then it was proved that the two random variables at the end are independent. This was done by using the special form of the normal distribution. Then from this, we obtained on looking at the moment generating functions,

$$\begin{aligned} \left(\frac{1}{1-2t}\right)^{n/2} &= E\left(\exp\left(t \frac{nS^2}{\sigma^2}\right)\right) E\left(\exp\left(t \left(\sum_{j=1}^n \frac{(X_k - \mu)}{\sqrt{n}\sigma}\right)^2\right)\right) \\ &= E\left(\exp\left(t \frac{nS^2}{\sigma^2}\right)\right) \frac{1}{(1-2t)^{1/2}} \end{aligned}$$

and so the moment generating function for  $nS^2/\sigma^2$  is

$$M(t) = \left( \frac{1}{1-2t} \right)^{(n-1)/2}$$

so  $\frac{nS^2}{\sigma^2}$  is  $\mathcal{X}^2(n-1)$ . Notice how important independence was in doing this last step.

Also notice that all the terms in the above were quadratic forms. This hopefully motivates the following very interesting result and shows why it is interesting. It is going to look a lot like what was done earlier with a difference. The independence of  $\mathbf{x}^T B \mathbf{x}, \mathbf{x}^T C \mathbf{x}$  will be obtained directly from an assumption that  $\mathbf{X}^T B \mathbf{X}$  is  $\mathcal{X}^2(r_1)$ .

**Theorem 39.4.6** *Let  $A, B, C$  be real symmetric  $n \times n$  matrices. Let*

$$\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}^T$$

*where  $\{X_1, \dots, X_n\}$  is a independent random sample from  $n(0, \sigma^2)$ . Let*

$$\mathbf{x}^T A \mathbf{x} = \mathbf{x}^T B \mathbf{x} + \mathbf{x}^T C \mathbf{x} \quad (39.6)$$

*and suppose  $\mathbf{X}^T A \mathbf{X}$  is  $\mathcal{X}^2(r)$ ,  $\mathbf{X}^T B \mathbf{X}$  is  $\mathcal{X}^2(r_1)$  for  $r_1 < r$ . Then the two random variables on the right are independent and  $\mathbf{X}^T C \mathbf{X}$  is  $\mathcal{X}^2(r-r_1)$ .*

**Proof:** Since 39.6 is a statement about quadratic forms for arbitrary  $\mathbf{x}$ , it follows that  $A = B + C$ . Now there is an orthogonal matrix  $U$  such that  $U^T A U = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$  where  $I$  is  $r \times r$  for  $r$  the rank of  $A$ . This follows from Theorem 11.4.7 presented much earlier in the material on linear algebra and the fact that  $\mathbf{X}^T A \mathbf{X}$  is  $\mathcal{X}^2(r)$  which implies, from Theorem 39.4.5 the eigenvalues of  $A$  are 1 or 0. Therefore,

$$\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} P & P_{12} \\ P_{21} & P_{22} \end{pmatrix} + \begin{pmatrix} Q & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \quad (39.7)$$

where  $P, Q$  are  $r \times r$  matrices and  $U^T B U = \begin{pmatrix} P & P_{12} \\ P_{21} & P_{22} \end{pmatrix}, U^T C U = \begin{pmatrix} Q & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$ .

Now multiply on both sides of 39.7 by  $\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$ . This yields

$$\begin{aligned} \begin{pmatrix} P & P_{12} \\ P_{21} & P_{22} \end{pmatrix} + \begin{pmatrix} Q & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} &= \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} &= \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Thus  $P_{12}, P_{21}, P_{22}, Q_{12}, Q_{21}, Q_{22}$  are all 0 and

$$U^T A U = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} = U^T (B + C) U = \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} \quad (39.8)$$

Note that the symmetry of  $B, C$  implies  $P, Q$  are symmetric also. It is given that  $\mathbf{X}^T B \mathbf{X}$  is  $\mathcal{X}^2(r^1)$  which happens if and only if  $B^2 = B$  thanks to Theorem 39.4.5. In other words,  $B$  has eigenvalues either 0 or 1. It follows that  $P^2 = P$ .

Now multiply on the left in 39.8 by  $\begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix}$ .

$$\begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} P^2 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} PQ & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} PQ & 0 \\ 0 & 0 \end{pmatrix}$$

Thus, comparing the ends,  $PQ = QP = 0$ . It follows that  $BC = 0$ . By Corollary 39.4.3,  $\mathbf{X}^T B \mathbf{X}, \mathbf{X}^T C \mathbf{X}$  are independent. It follows from this independence

$$\begin{aligned} E(\exp(t\mathbf{X}^T A \mathbf{X})) &= E(\exp(t\mathbf{X}^T B \mathbf{X} + t\mathbf{X}^T C \mathbf{X})) \\ &= E(\exp(t\mathbf{X}^T B \mathbf{X}) \exp(t\mathbf{X}^T C \mathbf{X})) \\ &= E(\exp(t\mathbf{X}^T B \mathbf{X})) E(\exp(t\mathbf{X}^T C \mathbf{X})) \end{aligned}$$

and so, by assumption,

$$\frac{1}{(1-2t)^{r/2}} = \frac{1}{(1-2t)^{r_1/2}} E(\exp(t\mathbf{X}^T C \mathbf{X}))$$

showing that

$$E(\exp(t\mathbf{X}^T C \mathbf{X})) = \frac{1}{(1-2t)^{(r-r_1)/2}}$$

which implies  $\mathbf{X}^T C \mathbf{X}$  is  $\mathcal{X}^2(r-r_1)$ . ■

Something should be pointed out here. It is that  $\mathbf{x}^T C \mathbf{x} \geq 0$  and this follows from linear algebra considerations. From the above argument,

$$\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix}, P^2 = P$$

Thus  $P$  has eigenvalues 0 or 1. Then also,  $0 \leq ((I-P)^2 \mathbf{x}, \mathbf{x}) = ((I-2P+P^2) \mathbf{x}, \mathbf{x}) = ((I-P) \mathbf{x}, \mathbf{x})$  and so  $I-P=Q$  has all nonnegative eigenvalues. Hence

$$(\mathbf{x}^T C \mathbf{x}) = \mathbf{x}^T U \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix} U^T \mathbf{x} \geq 0$$

You can extend this to more than two quadratic forms on the right.

**Corollary 39.4.7** Let  $A, \{A_k\}_{k=1}^m$  be real symmetric  $n \times n$  matrices. Let

$$\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}^T$$

where  $\{X_1, \dots, X_n\}$  is a random sample from  $n(0, \sigma^2)$ . Let

$$\mathbf{x}^T A \mathbf{x} = \sum_{k=1}^m \mathbf{x}^T A_k \mathbf{x} \quad (39.9)$$

and suppose  $\mathbf{X}^T A \mathbf{X}$  is  $\mathcal{X}^2(r)$ ,  $\mathbf{X}^T A_k \mathbf{X}$  is  $\mathcal{X}^2(r_k)$  for  $\sum_{k=1}^{m-1} r_k < r$ . Then the random variables  $\{\mathbf{X}^T A_k \mathbf{X}\}_{k=1}^{m-1}$  on the right are independent and  $\mathbf{X}^T A_m \mathbf{X}$  is  $\mathcal{X}^2(r - \sum_{k=1}^{m-1} r_k)$ .



**Proof:** Suppose the corollary is true for  $m-1$ , where  $m-1 \geq 2$ .

$$A = A_1 + (A_2 + \cdots + A_m) \equiv A_1 + B.$$

Then doing the same argument as above, you find that  $A_1 B = 0$  and  $B^2 = B$  since  $\mathbf{X}^T B \mathbf{X}$  is  $\mathcal{X}^2(r - r_1)$ . Now

$$B = A_2 + \cdots + A_m$$

and there are only  $m-1$  in the sum on the right. By induction,  $A_m$  is

$$\mathcal{X}^2\left(r - r_1 - \left(\sum_{k=2}^{m-1} r_k\right)\right)$$

and  $\mathbf{X}^T A_k \mathbf{X}$  are independent for  $k$  between 2 and  $m$  because  $A_k A_j = 0$  for such  $k, j$ . Recall Proposition 39.4.2. It only remains to verify that  $A_1 A_k$  for  $2 \leq k \leq m$ . You could do the same argument in the form  $A = A_2 + (A_1 + A_3 + \cdots + A_m)$  and conclude that  $A_1 A_k = 0$  for  $3 \leq k \leq m-1$ . Then all that is left is  $A_1 A_2$ . Just do the argument again for  $A = A_3 + (A_1 + A_2 + A_4 + \cdots + A_m)$  and conclude in particular that  $A_1 A_2 = 0$ . Thus all mixed products are 0 and so the quadratic forms are independent. ■

**Summary 39.4.8** *The following are the main ideas in this section. In all of this,*

$$\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}^T$$

where the  $X_i$  are independent and  $n(0, \sigma^2)$  and the matrices are symmetric

1. For  $\{A_k\}$  symmetric matrices,  $\{\mathbf{X}^T A_k \mathbf{X}\}$  are independent if and only if  $A_k A_j = 0$  for each  $k \neq j$ .
2.  $\frac{\mathbf{X}^T A \mathbf{X}}{\sigma^2}$  is  $\mathcal{X}^2(r)$  if and only if  $A^2 = A$  and the rank of  $A$  is  $r$ .
3. If  $\mathbf{x}^T A \mathbf{x} = \sum_{k=1}^m \mathbf{x}^T A_k \mathbf{x}$  and  $\mathbf{X}^T A \mathbf{X}$  is  $\mathcal{X}^2(r)$ ,  $\mathbf{X}^T A_k \mathbf{X}$  is  $\mathcal{X}^2(r_k)$  where  $k \leq m-1$  and  $\sum_{k=1}^{m-1} r_k < r$ , then  $\mathbf{X}^T A_m \mathbf{X}$  is  $\mathcal{X}^2(r - \sum_{k=1}^{m-1} r_k)$  and the random variables  $\{\mathbf{X}^T A_k \mathbf{X}\}$  are independent.

## 39.5 Linear Regression

This will be an interesting and important application of the above theory of quadratic forms. The idea is you have finitely many times  $t_i, t_1 \leq t_2 \leq \cdots \leq t_n$  and there are independent random variables  $X_1, \dots, X_n$ , corresponding to these  $t_i$ . More generally, the  $t_i$  are simply real numbers, but often the interpretation is time. Assume the following condition.

**Condition 39.5.1** *The random variables  $X_i$  associated with time  $t_k$  are*

$$n(\alpha + \beta(t_k - \bar{t}), \sigma^2)$$

where  $\bar{t} \equiv \frac{1}{n} \sum_{k=1}^n t_k$ . There may be many  $X_i$  associated with a single  $t_k$  but it is assumed that they are independent and normally distributed with a mean which depends on the  $t_k$  but the variance is constant. Note that the  $t_k$  might be repeated in the list. Typically they are repeated because one is taking a sample larger than one for each  $t_k$ .

Should the above condition be assumed? I am not sure, from the point of view of rigorous math, but this kind of thing is often assumed in experimental work and leads to useful conclusions and is not unreasonable since it is just an assumption that the random variables for each  $t_i$  are normally distributed.

For example suppose someone is developing vaccines for antiplasmosis, a disease in animals which causes anemia. You can measure anemia easily by keeping track of the packed cell volume. He has 15 animals which are infected by the disease and every week, he takes a small sample of blood from each and measures this packed cell volume using a centrifuge and capillary tubes. These measurements yield the  $X_k$ . If he had another group of animals say 20 which have been given a vaccine, how would he tell if the vaccine was effective? He would look for differences in the two different values of  $\beta$ . If he has a confidence interval for each  $\beta$ , the one for the vaccinated cows and the one for the unvaccinated ones, he could possibly conclude that his vaccine was working. An ordinary least squares approach would approximate the data with a straight line for each group of animals and would give a slope  $\hat{\beta}$  based solely on geometric conditions. This may suggest that the vaccine is working, but to be sure the pictures mean something, he needs a confidence interval or something similar involving a probability for the two parameters  $\beta$ , not just the estimate  $\hat{\beta}$ . It would of course also be very interesting to estimate the variance. The machinery for doing these estimates will be considered in this section. It is a very nice application of the results of the last section in which the distribution of quadratic forms was considered.

By independence, the probability density of the vector  $\mathbf{X} = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}^T$  is

$$\left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \prod_{k=1}^n e^{-\frac{1}{2} \frac{(x_k - (\alpha + \beta(t_k - \bar{t})))^2}{\sigma^2}}$$

Here  $t_k$  is the time which goes with  $X_k$ . Thus the  $t_k$  may be repeated because for a given  $t_i$ , there are at least one  $X_k$ , maybe more. The reason for writing the mean as  $\alpha + \beta(t_k - \bar{t})$  rather than more simply  $a + tb$  is that certain formulas come out looking much simpler if it is written this way and the maximum likelihood estimates for  $\alpha, \beta$  turn out to be  $\mathcal{X}^2(1)$ .

First consider the maximum likelihood estimates for  $\alpha, \beta, \sigma^2$ . Forget about the  $\sqrt{2\pi}$  and work with  $\ln$  of the expression.

$$n \ln(\sigma) + \sum_{k=1}^n \frac{1}{2} \frac{(X_k - (\alpha + \beta(t_k - \bar{t})))^2}{\sigma^2} = L(\sigma, \alpha, \beta) \quad (39.10)$$

Now take partial with respect to  $\alpha$  and set equal to 0

$$\sum_{k=1}^n (X_k - (\alpha + \beta(t_k - \bar{t}))) = 0$$

Next take partial with respect to  $\beta$  and set equal to 0. Denote by  $\hat{\alpha}, \hat{\beta}$  the solutions. These are the maximum likelihood estimates.

$$\sum_{k=1}^n (X_k - (\alpha + \beta(t_k - \bar{t}))) (t_k - \bar{t}) = 0$$

Thus

$$\sum_k X_k - n\alpha = 0,$$

and so

$$\hat{\alpha} = \frac{1}{n} \sum_k X_k \equiv \bar{X}$$

Then also

$$\sum_{k=1}^n X_k (t_k - \bar{t}) - \beta \sum_k (t_k - \bar{t})^2 = 0$$

and so

$$\hat{\beta} = \frac{\sum_{k=1}^n X_k (t_k - \bar{t})}{\sum_k (t_k - \bar{t})^2} = \frac{\sum_{k=1}^n (X_k - \bar{X}) (t_k - \bar{t})}{\sum_k (t_k - \bar{t})^2}$$

because  $\sum_{k=1}^n \bar{X} (t_k - \bar{t}) = 0$ . It remains to find the maximum likelihood estimate for  $\sigma^2$ . Using 39.10,

$$\begin{aligned} \frac{n}{\sigma} - \sum_{k=1}^n \frac{\left( X_k - \left( \hat{\alpha} + \hat{\beta} (t_k - \bar{t}) \right) \right)^2}{\sigma^3} &= 0 \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{k=1}^n \left( X_k - \left( \hat{\alpha} + \hat{\beta} (t_k - \bar{t}) \right) \right)^2 \end{aligned}$$

Now consider

$$\sum_{k=1}^n \frac{(X_k - (\alpha + \beta (t_k - \bar{t})))^2}{\sigma^2} \quad (39.11)$$

I will add in  $\hat{\alpha} + \hat{\beta} (t_k - \bar{t})$  and subtract it and then write this as a sum of quadratic forms. First of all, note that it is the sum of the squares of independent random variables in  $n(0, 1)$  and so it is  $\mathcal{X}^2(n)$ . It equals

$$\sum_{k=1}^n \frac{\left( X_k - \left( \hat{\alpha} + \hat{\beta} (t_k - \bar{t}) \right) + \left( \left( \hat{\alpha} + \hat{\beta} (t_k - \bar{t}) \right) - (\alpha + \beta (t_k - \bar{t})) \right) \right)^2}{\sigma^2} \quad (39.12)$$

This will be expanded. I need to consider the mixed term in which I will use the above descriptions of  $\hat{\alpha}$  and  $\hat{\beta}$ .

$$\begin{aligned} &\sum_k \left( X_k - \left( \hat{\alpha} + \hat{\beta} (t_k - \bar{t}) \right) \right) \left( \left( \hat{\alpha} + \hat{\beta} (t_k - \bar{t}) \right) - (\alpha + \beta (t_k - \bar{t})) \right) \\ &= \sum_k \left[ (X_k - \bar{X}) - \left( \hat{\beta} (t_k - \bar{t}) \right) \right] \left[ \left( \bar{X} + \hat{\beta} (t_k - \bar{t}) \right) - (\alpha + \beta (t_k - \bar{t})) \right] \end{aligned}$$

First note that

$$\sum_k (X_k - \bar{X}) \bar{X} = \sum_k (X_k - \bar{X}) \alpha = \sum_k \hat{\beta} (t_k - \bar{t}) \bar{X} = \sum_k \hat{\beta} (t_k - \bar{t}) \alpha = 0$$

Thus the mixed term is

$$\begin{aligned} &\sum_k (X_k - \bar{X}) \hat{\beta} (t_k - \bar{t}) - \beta \sum_k (X_k - \bar{X}) (t_k - \bar{t}) \\ &\quad - \hat{\beta}^2 \sum_k (t_k - \bar{t})^2 + \hat{\beta} \beta \sum_k (t_k - \bar{t})^2 \\ &= (\hat{\beta} - \beta) \sum_k (X_k - \bar{X}) (t_k - \bar{t}) + \hat{\beta} (\beta - \hat{\beta}) \sum_k (t_k - \bar{t})^2 \\ &= (\hat{\beta} - \beta) \hat{\beta} \sum_j (t_j - \bar{t})^2 + \hat{\beta} (\beta - \hat{\beta}) \sum_k (t_k - \bar{t})^2 = 0 \end{aligned}$$

It follows from the vanishing of the mixed term that 39.11 equals

$$\begin{aligned}
 & \sum_{k=1}^n \frac{(X_k - (\alpha + \beta(t_k - \bar{t})))^2}{\sigma^2} \\
 = & \sum_{k=1}^n \frac{\left(X_k - (\hat{\alpha} + \hat{\beta}(t_k - \bar{t}))\right)^2 + \left(\hat{\alpha} + \hat{\beta}(t_k - \bar{t})\right) - (\alpha + \beta(t_k - \bar{t}))^2}{\sigma^2} \\
 = & \frac{1}{\sigma^2} \sum_k \left(\hat{\alpha} + \hat{\beta}(t_k - \bar{t})\right) - (\alpha + \beta(t_k - \bar{t}))^2 + \frac{n\hat{\sigma}^2}{\sigma^2} \\
 = & \frac{1}{\sigma^2} \sum_k \left((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)(t_k - \bar{t})\right)^2 + \frac{n\hat{\sigma}^2}{\sigma^2}
 \end{aligned}$$

Consider the mixed term in the first sum.

$$\sum_k (\hat{\alpha} - \alpha) (\hat{\beta} - \beta) (t_k - \bar{t}) = 0$$

Therefore, from 39.11,

$$\begin{aligned}
 & \sum_{k=1}^n \frac{(X_k - (\alpha + \beta(t_k - \bar{t})))^2}{\sigma^2} = \\
 & \frac{n}{\sigma^2} (\hat{\alpha} - \alpha)^2 + \frac{(\hat{\beta} - \beta)^2}{\sigma^2} \sum_k (t_k - \bar{t})^2 + \frac{n\hat{\sigma}^2}{\sigma^2}
 \end{aligned} \tag{39.13}$$

At this point, we need to consider what we have.

**Lemma 39.5.2** Suppose  $X_i$  is  $n(\mu_i, \sigma_i^2)$  and the  $X_i$  are independent for  $i \leq n$ . Then  $\sum_i a_i X_i$  is  $n(\sum_i a_i \mu_i, \sum_i a_i^2 \sigma_i^2)$ .

**Proof:** Consider the moment generating function.

$$\begin{aligned}
 M(t) & \equiv E \left( \exp \left( t \sum_i a_i X_i \right) \right) = E \left( \prod_{i=1}^n \exp(t a_i X_i) \right) \\
 & = \prod_{i=1}^n E(\exp(t a_i X_i)) = \prod_{i=1}^n e^{\frac{1}{2} t^2 a_i^2 \sigma_i^2} e^{t a_i \mu_i} \\
 & = e^{\frac{1}{2} t^2 (\sum_i a_i^2 \sigma_i^2)} e^{t \sum_i a_i \mu_i} \blacksquare
 \end{aligned}$$

Thus  $\hat{\alpha} = \bar{X}$  is  $n\left(\frac{1}{n} \sum_{i=1}^n (\alpha + \beta(t_i - \bar{t})), \sum_i \frac{1}{n^2} \sigma^2\right) = n\left(\alpha, \frac{\sigma^2}{n}\right)$ . Also  $\frac{\hat{\alpha} - \alpha}{(\sigma/\sqrt{n})} = \frac{\bar{X} - \alpha}{(\sigma/\sqrt{n})}$  is  $n(0, 1)$  and so the square root of the first term on the right in 39.13 is  $n(0, 1)$  so that first term is  $\mathcal{X}^2(1)$ . Similarly, consider the second term or rather its square root. This is

$$\frac{\hat{\beta} - \beta}{\left(\sigma / \left(\sum_k (t_k - \bar{t})^2\right)^{1/2}\right)} \tag{39.14}$$

Consider the moment generating function for  $\hat{\beta}$ .

$$\begin{aligned}
 M(t) &= E\left(\exp t\hat{\beta}\right) = E\left(\exp\left(t\frac{\sum_{k=1}^n X_k(t_k - \bar{t})}{\sum_i (t_i - \bar{t})^2}\right)\right) \\
 &= E\left(\prod_{k=1}^n \exp \frac{t(t_k - \bar{t})}{\sum_i (t_i - \bar{t})^2} X_k\right) \\
 &= \prod_{k=1}^n \exp\left(\frac{\frac{1}{2} \frac{t^2 (t_k - \bar{t})^2}{(\sum_i (t_i - \bar{t})^2)^2} \sigma^2}{+\frac{t(t_k - \bar{t})}{\sum_i (t_i - \bar{t})^2} (\alpha + \beta(t_k - \bar{t}))}\right) \\
 &= \exp\left(\frac{1}{2} t^2 \frac{\sigma^2}{\sum_i (t_i - \bar{t})^2} + \beta t\right)
 \end{aligned}$$

So  $\hat{\beta}$  is  $n\left(\beta, \frac{\sigma^2}{\sum_i (t_i - \bar{t})^2}\right)$ . It follows that the random variable in 39.14 is  $n(0, 1)$  and so the second term on the right in 39.13 is  $\mathcal{X}^2(1)$ . The last term in 39.13 is

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{k=1}^n \left(X_k - (\hat{\alpha} + \hat{\beta}(t_k - \bar{t}))\right)^2$$

Thus we have

$$\begin{aligned}
 \overbrace{\sum_{k=1}^n \frac{(X_k - (\alpha + \beta(t_k - \bar{t})))^2}{\sigma^2}}^{\mathcal{X}^2(n)} &= \overbrace{\frac{n}{\sigma^2} (\hat{\alpha} - \alpha)^2}^{\mathcal{X}^2(1)} + \overbrace{\frac{(\hat{\beta} - \beta)^2}{\sigma^2} \sum_k (t_k - \bar{t})^2}^{\mathcal{X}^2(1)} \\
 &\quad + \frac{1}{\sigma^2} \sum_{k=1}^n \left(X_k - (\hat{\alpha} + \hat{\beta}(t_k - \bar{t}))\right)^2 \quad (39.15)
 \end{aligned}$$

In fact, the terms on the right are quadratic forms in the variables  $X_k - (\alpha + \beta(t_k - \bar{t}))$  although it does not look like it. Consider the first term.

$$\begin{aligned}
 \hat{\alpha} - \alpha &= \frac{1}{n} \sum_k X_k - \alpha = \frac{1}{n} \sum_k (X_k - \alpha) \\
 &= \frac{1}{n} \sum_k (X_k - (\alpha + \beta(t_k - \bar{t})))
 \end{aligned}$$

This is squared and that is why this term is a quadratic form in the variables

$$X_k - (\alpha + \beta(t_k - \bar{t}))$$

Note that the terms added in sum to 0. A similar trick will apply to the other terms. Consider the second term.

$$\hat{\beta} - \beta = \frac{\sum_{k=1}^n X_k(t_k - \bar{t})}{\sum_k (t_k - \bar{t})^2} - \beta = \frac{\sum_{k=1}^n X_k(t_k - \bar{t}) - \beta \sum_k (t_k - \bar{t})^2}{\sum_k (t_k - \bar{t})^2}$$

$$= \frac{\sum_{k=1}^n (X_k - \beta(t_k - \bar{t}))(t_k - \bar{t})}{\sum_k (t_k - \bar{t})^2} = \frac{\sum_{k=1}^n (X_k - (\alpha + \beta(t_k - \bar{t}))(t_k - \bar{t}))}{\sum_k (t_k - \bar{t})^2}$$

Note that the terms added in which include  $\alpha$  sum to 0. Thus this second term is a constant times the square of the above which is a quadratic form in the variables  $X_k - (\alpha + \beta(t_k - \bar{t}))$ .

Finally, consider the last term. Since all the other terms are quadratic forms in the variables  $X_k - (\alpha + \beta(t_k - \bar{t}))$ , this one must also be so because it is equal to a linear combination of these terms. Alternatively, you could verify this in a similar manner. However, I will stop here and not wade in sorrow to massage the complicated expression into the right form. This is what is needed for the following proposition.

**Proposition 39.5.3** *In 39.15 the various terms are chi-squared as indicated in the formula and the last term is  $\mathcal{X}^2(n-2)$ . Also, the three terms on the right are independent random variables.*

**Proof:** This follows from Corollary 39.4.7 or Summary 39.4.8. ■

Something else should be noted which involves the maximum likelihood estimates.

$$\begin{aligned} \hat{\alpha} + \hat{\beta}(t_i - \bar{t}) &= \frac{1}{n} \sum_k X_k - \frac{1}{n} \hat{\beta} \sum_k t_k + \hat{\beta} t_i \\ &\equiv a + b t_i \end{aligned}$$

Thus

$$\begin{aligned} b &= \hat{\beta} = \frac{\sum_{k=1}^n (X_k)(t_k - \bar{t})}{\sum_k (t_k - \bar{t})^2} = \frac{\sum_k X_k t_k - \bar{t} \sum_k X_k}{\sum_k t_k^2 - 2\bar{t} \sum_k t_k + n\bar{t}^2} \\ &= \frac{\sum_k X_k t_k - \bar{t} \sum_k X_k}{\sum_k t_k^2 - n\bar{t}^2} \end{aligned}$$

Recall how the least squares line is  $a + bt$  where  $b =$

$$\begin{aligned} &= \frac{-(\sum_{k=1}^n t_k)(\sum_{k=1}^n X_k) + (\sum_{k=1}^n t_k X_k)n}{(\sum_{k=1}^n t_k^2)n - (\sum_{k=1}^n t_k)^2} \\ &= \frac{(\sum_{k=1}^n t_k X_k)n - n\bar{t}(\sum_{k=1}^n X_k)}{(\sum_{k=1}^n t_k^2)n - n^2\bar{t}^2} \\ &= \frac{(\sum_{k=1}^n t_k X_k) - \bar{t}(\sum_{k=1}^n X_k)}{(\sum_{k=1}^n t_k^2) - n\bar{t}^2} \end{aligned}$$

This is the same as  $\hat{\beta}$ . Similarly, more computations will show that  $a \equiv \frac{1}{n} \sum_k X_k - \frac{1}{n} \hat{\beta} \sum_k t_k$  will end up being the least squares estimate. Thus, if you have a confidence interval for  $\hat{\alpha}$  and  $\hat{\beta}$ , this delivers a confidence interval for  $a, b$ .

Now suppose you want a confidence interval for  $\sigma^2$ . It was shown above that

$$\frac{1}{\sigma^2} \sum_{k=1}^n \left( X_k - \left( \hat{\alpha} + \hat{\beta}(t_k - \bar{t}) \right) \right)^2$$

is  $\mathcal{X}^2(n-2)$  and so you can use this statistic and a table or graph of the appropriate chi squared distribution to obtain a confidence interval for  $\sigma^2$ . Since  $\hat{\beta} = \frac{\sum_{j=1}^n X_j(t_j - \bar{t})}{\sum_j (t_j - \bar{t})^2}$ , the

above expression is

$$\begin{aligned} & \frac{1}{\sigma^2} \sum_{k=1}^n \left( X_k - \left( \hat{\alpha} + \hat{\beta} (t_k - \bar{t}) \right) \right)^2 \\ &= \frac{1}{\sigma^2} \sum_{k=1}^n \left( X_k - \left( \bar{X} + \left( \frac{\sum_{j=1}^n X_j (t_j - \bar{t})}{\sum_{j=1}^n (t_j - \bar{t})^2} \right) (t_k - \bar{t}) \right) \right)^2 \end{aligned}$$

Thus, to find a confidence interval for the variance, do the following.

**PROCEDURE 39.5.4** *In the above situation, to find a .95 confidence interval for the variance which is unknown do this:*

1. *If there are  $n$  observations,  $n$  fairly large, certainly larger than 2, find an interval  $[a, b]$  such that if  $V$  is a  $\mathcal{X}^2(n-2)$  random variable  $P(V \in [a, b]) \geq .95$ .*
2. *Find  $\bar{X}$  the sample mean and  $\bar{t}$  the average of the  $t$  values. Then fill in to find*

$$S \equiv \sum_{k=1}^n \left( X_k - \left( \bar{X} + \left( \frac{\sum_{j=1}^n X_j (t_j - \bar{t})}{\sum_{j=1}^n (t_j - \bar{t})^2} \right) (t_k - \bar{t}) \right) \right)^2$$

3. *Then the .95 confidence interval for  $\sigma^2$  is determined by*

$$a \leq \frac{S}{\sigma^2} \leq b$$

*In other words, with probability .95, the variance  $\sigma^2$  satisfies*

$$\frac{S}{b} \leq \sigma^2 \leq \frac{S}{a}$$

I think one is even more interested in  $\beta$ , the slope of the line for the mean. To find a confidence interval for  $\beta$ , recall the  $T$  test. The  $T$  distribution was the distribution of  $\frac{W}{\sqrt{V/r}}$  where  $V$  was  $\mathcal{X}^2(r)$  and  $W$  was  $n(0, 1)$ . Do we have such random variables above?

Recall it was shown above that  $\hat{\beta}$  is  $n\left(\beta, \frac{\sigma^2}{\sum_i (t_i - \bar{t})^2}\right)$ . Therefore,

$$\frac{\hat{\beta} - \beta}{\left( \sigma / \sqrt{\sum_i (t_i - \bar{t})^2} \right)} \text{ is } n(0, 1)$$

Therefore,

$$\frac{\frac{\hat{\beta} - \beta}{\left( \sigma / \sqrt{\sum_i (t_i - \bar{t})^2} \right)}}{\sqrt{\frac{1}{\sigma^2} \sum_{k=1}^n \left( X_k - \left( \hat{\alpha} + \hat{\beta} (t_k - \bar{t}) \right) \right)^2 / (n-2)}}$$

is a  $T$  random variable with  $r = n - 2$ . Simplifying the above gives

$$\begin{aligned} & \frac{\sqrt{n-2} \left( \frac{\sum_{j=1}^n X_j (t_j - \bar{t})}{\sum_j (t_j - \bar{t})^2} - \beta \right) \sqrt{\sum_i (t_i - \bar{t})^2}}{\sqrt{\sum_{k=1}^n \left( X_k - \left( \bar{X} + \frac{\sum_{j=1}^n X_j (t_j - \bar{t})}{\sum_j (t_j - \bar{t})^2} (t_k - \bar{t}) \right) \right)^2}} \\ &= \frac{\sqrt{n-2} \left( \sum_{j=1}^n X_j (t_j - \bar{t}) - \beta \sum_{j=1}^n (t_j - \bar{t})^2 \right)}{\sqrt{\sum_j (t_j - \bar{t})^2} \sqrt{\sum_{k=1}^n \left( X_k - \left( \bar{X} + \frac{\sum_{j=1}^n X_j (t_j - \bar{t})}{\sum_j (t_j - \bar{t})^2} (t_k - \bar{t}) \right) \right)^2}} \end{aligned}$$

**PROCEDURE 39.5.5** To find a .95 confidence interval for  $\beta$ , do the following.

1. If there are  $n$  observations,  $n$  fairly large, certainly larger than 2, find an interval for the  $T$  distribution  $[a, b]$  such that if  $X$  is a random variable with this distribution,  $P(X \in [a, b]) \geq .95$ .
2. Compute  $\bar{t}$  the average  $t$  value and  $\bar{X}$  the sample mean. Then compute

$$S \equiv \sqrt{\sum_j (t_j - \bar{t})^2} \sqrt{\sum_{k=1}^n \left( X_k - \left( \bar{X} + \frac{\sum_{j=1}^n X_j (t_j - \bar{t})}{\sum_j (t_j - \bar{t})^2} (t_k - \bar{t}) \right) \right)^2}$$

and  $P \equiv \sqrt{n-2} \sum_{j=1}^n X_j (t_j - \bar{t})$ . Then the confidence interval is obtained by solving the following inequality for  $\beta$ .

$$a \leq \frac{P - \sqrt{n-2} \beta \sum_{j=1}^n (t_j - \bar{t})^2}{S} \leq b$$

Then the confidence interval for  $\beta$  is

$$\frac{P - aS}{\sqrt{n-2} \sum_{j=1}^n (t_j - \bar{t})^2} \geq \beta \geq \frac{P - Sb}{\sqrt{n-2} \sum_{j=1}^n (t_j - \bar{t})^2}$$

You can also find a confidence interval for  $\alpha$ . Here you would use a  $T$  distribution involving  $\frac{\hat{\alpha} - \alpha}{(\hat{\sigma}/\sqrt{n})}$  which was shown above to be  $n(0, 1)$  along with the distribution of the  $\mathcal{X}^2(n-2)$  random variable  $\frac{1}{\hat{\sigma}^2} \sum_{k=1}^n \left( X_k - \left( \hat{\alpha} + \hat{\beta} (t_k - \bar{t}) \right) \right)^2$  in the description of the  $T$  statistic. This is left to the interested reader. It is just like the above.

## 39.6 Goodness of Fit

In all of the above, it was about identifying parameters, usually for the normal distribution which has two parameters  $\mu$  and  $\sigma^2$ . What if you wonder whether a given random sample is consistent with the sample coming from some probability distribution? What then? You have now abandoned the assumption that the sample comes from say  $n(\mu, \sigma^2)$ . This question is of course much more speculative. However, there are methods for considering



it. These methods come from Pearson around 1900. Again, they involve massaging things to use a known distribution, this time a chi-squared distribution.

To do this right, you should be using characteristic functions, but everything of interest in this book will have a moment generating function and it is just less fussy to do everything in terms of moment generating functions. However, the complex variable material in this book is sufficient for you to do in terms of characteristic functions, except even then, there are more advanced and theoretical theorems needed which involve much harder techniques. These are related to convergence of the characteristic functions leading to convergence of the distributions. Because of these considerations, I will give a discussion to make the main result plausible based on moment generating functions. To see a full discussion of the theory about to be presented, see [9]. The following is the situation of interest.

1.  $F(x) \equiv P(X \leq x)$  so  $F$  is the distribution function of  $X$  which has a density  $f(x)$ .
2. There are  $r$  disjoint intervals  $S_1, \dots, S_r$  whose union is  $\mathbb{R}$  and  $p_i \equiv P(X \in S_i)$ .
3. For  $n$  large, there is a random sample  $X_1, \dots, X_n$  and  $V_i$  will denote the number of these samples which end up in  $S_i$ . Thus the expected number for  $V_i$  would be  $np_i$ .
4. The determination whether it is reasonable to consider the  $X_k$  as coming from the probability distribution  $F$  is dependent on consideration of

$$Q(n) \equiv \sum_{k=1}^r \frac{(V_k - np_k)^2}{np_k}$$

If this is small, then there isn't much difference between the observed value  $V_i$  and the expected value  $np_i$  and it would be reasonable to think that the sample is from the given probability distribution. On the other hand, if it is large, then it would not be reasonable to consider the sample as coming from the given distribution.

Of course, the problem is in quantifying these issues and this involves the next major proposition. First is a lemma about counting.

**Lemma 39.6.1** *The number of ways of selecting subsets having  $v_1, \dots, v_r$  elements where  $\sum_k v_k = n$ , from a set having  $n$  elements is*

$$\frac{n!}{v_1!v_2! \cdots v_r!}$$

Also

$$(a_1 + a_2 + \cdots + a_r)^n = \sum_{v_1 + \cdots + v_r = n} \frac{n!}{v_1!v_2! \cdots v_r!} a_1^{v_1} a_2^{v_2} \cdots a_r^{v_r}$$

**Proof:** There is nothing to prove if  $r = 1$ . In case  $r = 2$ , it was shown earlier. Recall that  $\frac{n!}{v_1!(n-v_1)!} = \frac{n!}{v_1!v_2!}$  is the number of ways to select a set having  $v_1$  elements and a set having  $v_2$  elements from a set having  $n$  elements. In general, the number of ways to obtain subsets of size  $v_1, v_2, \dots, v_r$  is the number of ways to select subsets of size  $v_1, v_2, \dots, v_{r-1}$  from a set of size  $n - v_r$  times the number of ways to select the set of size  $n - v_r$  which, by induction is

$$\frac{(n - v_r)!}{v_1!v_2! \cdots v_{r-1}!} \frac{n!}{(n - v_r)!v_r!} = \frac{n!}{v_1!v_2! \cdots v_r!}.$$

As to the last assertion,  $(a_1 + a_2 + \cdots + a_r)^n =$

$$(a_1 + a_2 + \cdots + a_r)(a_1 + a_2 + \cdots + a_r) \cdots (a_1 + a_2 + \cdots + a_r)$$

where there are  $n$  products. Thus this product equals a sum of terms of the form  $a_1^{v_1} a_2^{v_2} \cdots a_r^{v_r}$  where  $\sum_k v_k = n$ . How many are there for a given choice of exponents  $v_1, v_2, \dots, v_r$ ? it is the number of ways of picking  $v_1$  factors from the above product to go with  $a_1^{v_1}$ ,  $v_2$  factors to go with  $a_2^{v_2}$  and so forth. Thus the total number associated with a particular term of this form is  $\frac{n!}{v_1! v_2! \cdots v_r!}$  and this proves the second part. ■

**Proposition 39.6.2** *Let  $p_k, V_k$  be as described above where  $F(x)$  is a given distribution function. Then if  $n$  is large,  $Q(n)$  is distributed approximately as  $\mathcal{X}^2(r-1)$  where there are  $r$  disjoint intervals covering  $\mathbb{R}$ .*

**Proof:** Assume the  $X_k$  are samples from  $F(x)$ . Then the  $V_k$  have a multinomial distribution. That is

$$P(V_1 = v_1, V_2 = v_2, \dots, V_r = v_r) = \frac{n!}{v_1! v_2! \cdots v_r!} p_1^{v_1} \cdots p_r^{v_r}, \sum_k v_k = n.$$

Indeed, the probability that any  $X_k$  is in  $S_i$  is  $p_i$  and so the probability that there are  $v_i$  of them in  $S_i$  is as claimed above. Then consider the moment generating function

$$\begin{aligned} M(t_1, \dots, t_r) &\equiv E\left(\exp \sum_{k=1}^r t_k V_k\right) \\ &= \sum_{v_1 + \dots + v_r = n} \frac{n!}{v_1! v_2! \cdots v_r!} p_1^{v_1} \cdots p_r^{v_r} e^{t_1 v_1} \cdots e^{t_r v_r} \end{aligned}$$

By Lemma 39.6.1 this equals

$$(p_1 e^{t_1} + \cdots + p_r e^{t_r})^n \quad (39.16)$$

Now consider the moment generating function of the vector

$$\left( \frac{V_1 - np_1}{\sqrt{np_1}} \quad \cdots \quad \frac{V_k - np_k}{\sqrt{np_k}} \right).$$

Then

$$E\left(\sum_k t_k \frac{V_k - np_k}{\sqrt{np_k}}\right) = e^{-\sum_k t_k \frac{np_k}{\sqrt{np_k}}} E\left(\sum_k \frac{t_k}{\sqrt{np_k}} V_k\right)$$

From 39.16, this equals

$$M_n(t_1, \dots, t_r) = e^{-\sum_k t_k \frac{np_k}{\sqrt{np_k}}} \left( p_1 e^{\frac{t_1}{\sqrt{np_1}}} + \cdots + p_r e^{\frac{t_r}{\sqrt{np_r}}} \right)^n$$

Taking ln of this, and simplifying a little,

$$\ln(M_n) = -\sum_k t_k \sqrt{np_k} + n \ln \left( p_1 e^{\frac{t_1}{\sqrt{np_1}}} + \cdots + p_r e^{\frac{t_r}{\sqrt{np_r}}} \right)$$

Now replace each  $e^{\frac{t_k}{\sqrt{np_k}}}$  with the first few terms of its power series. Then the inside of ( ) above becomes

$$p_1 \left( 1 + \frac{t_1}{\sqrt{np_1}} + \frac{\left( \frac{t_1}{\sqrt{np_1}} \right)^2}{2!} \right) + \cdots + p_r \left( 1 + \frac{t_r}{\sqrt{np_r}} + \frac{\left( \frac{t_r}{\sqrt{np_r}} \right)^2}{2!} \right) + O\left( \frac{1}{n^{3/2}} \right)$$

That last term indicates that what is left over is just a lot of stuff times powers of  $\frac{1}{n^{3/2}}$ . Since the sum of the  $p_i$  is one, this yields  $\ln(M_n) =$

$$-\sum_k t_k \sqrt{p_k n} + n \ln \left( 1 + \frac{p_1 t_1}{\sqrt{np_1}} + \frac{p_1 \left( \frac{t_1}{\sqrt{np_1}} \right)^2}{2!} + \cdots + \frac{p_r t_r}{\sqrt{np_r}} + \frac{p_r \left( \frac{t_r}{\sqrt{np_r}} \right)^2}{2!} + O\left( \frac{1}{n^{3/2}} \right) \right)$$

Now  $\ln(1+x) = 0 + x - \frac{1}{2}x^2 + O(x^3)$ . Of course the  $x$  here is the material in the above which comes after the 1. The  $O(x^3)$  terms are all  $O\left(\frac{1}{n^{3/2}}\right)$  and there are a few terms in the  $x^2$  which are not, which are included in  $\left(\sum_k \frac{p_k t_k}{\sqrt{np_k}}\right)^2$ . I will retain these terms in the following. Thus  $\ln M_n =$

$$-\sum_k t_k \sqrt{p_k n} + n \left( \frac{p_1 t_1}{\sqrt{np_1}} + \frac{p_1 \left( \frac{t_1}{\sqrt{np_1}} \right)^2}{2!} + \cdots + \frac{p_r t_r}{\sqrt{np_r}} + \frac{p_r \left( \frac{t_r}{\sqrt{np_r}} \right)^2}{2!} - \frac{1}{2} \left( \sum_k \frac{p_k t_k}{\sqrt{np_k}} \right)^2 + O\left( \frac{1}{n^{3/2}} \right) \right)$$

Of course this simplifies. When you multiply by the  $n$  you get  $\ln(M_n) =$

$$\begin{aligned} & -\sum_k t_k \sqrt{p_k n} + \left( \sqrt{n} \sqrt{p_1} t_1 + \frac{(t_1)^2}{2} + \cdots + \sqrt{n} \sqrt{p_r} t_r + \frac{(t_r)^2}{2} - \frac{1}{2} \left( n \sum_k \frac{p_k t_k}{\sqrt{np_k}} \right)^2 + O\left( \frac{1}{n^{1/2}} \right) \right) \\ & = \left( \frac{(t_1)^2}{2} + \cdots + \frac{(t_r)^2}{2} - \frac{1}{2} \left( \sum_k \sqrt{p_k} t_k \right)^2 + O\left( \frac{1}{n^{1/2}} \right) \right) \end{aligned}$$

Therefore,

$$M_n = \exp \left( \frac{1}{2} \left[ \sum_k t_k^2 - \left( \sum_k \sqrt{p_k} t_k \right)^2 \right] \right) e^{O(1/\sqrt{n})}$$

For large  $n$  this is very close to

$$M_n = \exp \left( \frac{1}{2} \left[ \sum_k t_k^2 - \left( \sum_k \sqrt{p_k} t_k \right)^2 \right] \right)$$

which is of the form  $\exp\left(\frac{1}{2} \mathbf{t}^T A \mathbf{t}\right)$  where  $\mathbf{t} \in \mathbb{R}^r$ . What is  $A$ ?

$$\sum_k t_k^2 - \left( \sum_k \sqrt{p_k} t_k \right)^2 = \sum_k t_k^2 - \sum_{i,j} \sqrt{p_i p_j} t_i t_j$$

and so

$$A = \begin{pmatrix} 1-p_1 & -\sqrt{p_1 p_2} & \cdots & -\sqrt{p_1 p_r} \\ -\sqrt{p_1 p_2} & 1-p_2 & \cdots & -\sqrt{p_2 p_r} \\ \vdots & & \ddots & \vdots \\ -\sqrt{p_r p_1} & -\sqrt{p_r p_2} & \cdots & 1-p_r \end{pmatrix}$$

$$A = I - \begin{pmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \vdots \\ \sqrt{p_r} \end{pmatrix} \begin{pmatrix} \sqrt{p_1} & \sqrt{p_2} & \cdots & \sqrt{p_r} \end{pmatrix}$$

This is of the form  $I - \mathbf{a}\mathbf{a}^T$  where  $\mathbf{a}^T \mathbf{a} = |\mathbf{a}|^2 = 1$ . Now

$$(I - \mathbf{a}\mathbf{a}^T)(I - \mathbf{a}\mathbf{a}^T) = I - 2\mathbf{a}\mathbf{a}^T + \mathbf{a}\mathbf{a}^T \mathbf{a}\mathbf{a}^T = I - 2\mathbf{a}\mathbf{a}^T + \mathbf{a}\mathbf{a}^T = I - \mathbf{a}\mathbf{a}^T$$

Thus  $A^2 = A$  and so by Lemma 39.4.4 the matrix  $A$  has only 0 and 1 as eigenvalues. Now note that

$$(I - \mathbf{a}\mathbf{a}^T)\mathbf{a} = \mathbf{0}$$

and so there is a 0 eigenvalue. In fact multiples of this single eigenvector are the only ones which deliver 0 as an eigenvalue. I show this now. If  $(I - \mathbf{a}\mathbf{a}^T)\mathbf{b} = \mathbf{0}$ , then  $\mathbf{b} = \mathbf{a}\mathbf{a}^T \mathbf{b}$  and so  $|\mathbf{b}| = |\mathbf{a}| |\mathbf{a}^T \mathbf{b}| \leq |\mathbf{a}|^2 |\mathbf{b}| = |\mathbf{b}|$  and so you must have  $(\mathbf{a} \cdot \mathbf{b}) = \mathbf{a}^T \mathbf{b} = |\mathbf{a}| |\mathbf{b}|$  which means that the only eigenvectors  $\mathbf{b}$  which have 0 as an eigenvalue are multiples of  $\mathbf{a}$ . Recall that this was the condition for equality in the Cauchy Schwarz inequality. Therefore, the rank of  $A$  is  $r - 1$ . This is because  $A$  is symmetric so there is a basis of eigenvectors. By Corollary 38.9.9, and the fact that if the moment generating functions converge, then so do the random variables having the given moment generating function, it follows that for large  $n$ , the distribution of  $\sum_{k=1}^r \frac{(V_k - np_k)^2}{np_k}$  is  $\mathcal{X}^2(r - 1)$  as claimed. ■

**Example 39.6.3** A die is half of a pair of dice. It is cubic and has numbers from 1 to 6 on the sides. They are suppose to come up with equal probability. Now you have a die and it is rolled 60 times. The following table summarizes the outcomes.

|   |   |   |    |    |   |
|---|---|---|----|----|---|
| 1 | 2 | 3 | 4  | 5  | 6 |
| 6 | 5 | 9 | 10 | 23 | 7 |

Is it reasonable to conclude that the die is fair, doing what it is supposed to do by giving the same probability to each possible outcome? Well, obviously not, but lets see how to quantify this conclusion.

Let  $S_1 = (-\infty, 1], S_2 = (1, 2], S_3 = (2, 3], S_4 = (3, 4], S_5 = (4, 5], S_6 = (5, \infty)$ . Then the  $p_i$  are each  $1/6$ . We compute the thing which will have a  $\mathcal{X}^2(5)$  distribution.

$$\sum_{k=1}^r \frac{(V_k - np_k)^2}{np_k} = \left( \frac{(6-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(23-10)^2}{10} + \frac{(7-10)^2}{10} \right) = 22$$

Now from a table or using MATLAB and data cursor you find that the probability that a  $\mathcal{X}^2(5)$  random variable is less than 14.35 is .98. However, here we have that the random variable is 22 so I can say with probability .98 that the die is unfair.

Actually, there is a fly in the ointment. You know the random variable is not exactly  $\mathcal{X}^2(5)$ . After all, you only used a sample of  $60 = n$ . The idea is to let  $n \rightarrow \infty$ . Cramér says that if  $n$  is large enough that the expected numbers  $np_i \geq 10$  for each  $S_i$ , then the approximation will be good enough for ordinary applications. Of course you end up being more sure if you take  $n$  larger. More information is typically better.

This process illustrates another example of Hypothesis testing. In this case, the “null hypothesis” is that the die is fair and the density function is  $1/6$  for each outcome. The above process indicates that we should reject the “null hypothesis” with probability .98. This is a general notion in statistics called hypothesis testing. There is quite a bit of jargon associated with this, but the main idea is illustrated by the above example. Here is another example where it is not so clear.

**Example 39.6.4** *A die is half of a pair of dice. It is cubic and has numbers from 1 to 6 on the sides. They are suppose to come up with equal probability. Now you have a die and it is rolled 60 times. The following table summarizes the outcomes.*

|   |    |   |    |    |    |
|---|----|---|----|----|----|
| 1 | 2  | 3 | 4  | 5  | 6  |
| 9 | 11 | 9 | 11 | 10 | 10 |

*Is it reasonable to conclude that the die is fair, doing what it is supposed to do by giving the same probability to each possible outcome?*

We can do this the same way. Lets agree to reject the null hypothesis that the die is fair if the statistic used above which measures discrepancy is in a region  $x > a$  where  $P(X \leq a) = .6$ . The  $\mathcal{X}^2(5)$  variable is

$$\frac{1}{10} + \frac{1}{10} + \frac{1}{10} = .3$$

It is clearly not in a region associated with smaller than probability .4. In fact, from the graph or a table,  $\mathcal{X}^2 > .3$ , occurs with probability almost 1, certainly larger than .98. Therefore, it is totally unsurprising that this random variable would be larger than .3. Therefore, we don't reject the hypothesis. It is reasonable to think that the die is fair.

**PROCEDURE 39.6.5** *A random sample  $\{X_k\}_{k=1}^n$  is taken where  $n$  is large. To test whether the sample is taken from a distribution function  $F(x) \equiv P(X \leq x)$ , do the following. Partition  $\mathbb{R}$  into  $r$  disjoint intervals  $S_1, \dots, S_r$  such that  $P(X \in S_i) = p_i > 0$  and assume  $n$  is large enough that  $np_i \geq 10$  for each  $i$ . Letting  $V_k$  be the number of times some  $X_i$  is in  $S_k$  form*

$$D \equiv \sum_{k=1}^r \frac{(V_k - np_k)^2}{np_k}$$

*Since  $n$  is large, this is distributed as  $\mathcal{X}^2(r-1)$ . The null hypothesis  $H_0$  is that  $F(x)$  is the distribution for the sample. We reject  $H_0$  with probability .95 if  $D \geq \alpha$  where  $P(D \leq \alpha) \geq .95$ .*

The following example illustrates a situation which is more typical in which there are parameters. The question is whether grades are normally distributed. Of course there are two parameters  $\mu$  and  $\sigma^2$  and what you are asking is whether some choice of  $\mu$  and  $\sigma^2$  results in a normal distribution from which a random sample of test scores can be considered drawn. The way to deal with this is to regard  $D$  in the above as  $\mathcal{X}^2((r-1) - s)$  where  $s$  is

the number of unknown parameters and to replace each parameter with its maximum likelihood estimate. The proof of this is very technical and you can see it discussed in Cramér, [9].

**Example 39.6.6** *A certain university assigns grades in calculus according to an assumption that the grades on the final will be normally distributed. This is a great idea because you don't have to ask whether students have learned a well defined set of outcomes and it removes the onus of having to tell the students that they didn't learn anything, thus improving course evaluations, which will please university administrators who regard the job of the university as pleasing the customers. It also removes the responsibility of ensuring that the exam is reasonable. The magic curve will take care of any problems. Here are outcomes from a final exam. I am making these up of course.*

|                                     |                 |            |            |            |            |        |
|-------------------------------------|-----------------|------------|------------|------------|------------|--------|
| The $S_k$                           | $(-\infty, 50]$ | $(50, 60]$ | $(60, 70]$ | $(70, 80]$ | $(80, 90]$ | $> 90$ |
| number in $S_k$                     | 120             | 150        | 60         | 20         | 10         | 40     |
| average grade in $S_k$              | 40              | 55         | 65         | 75         | 85         | 95     |
| average grade <sup>2</sup> in $S_k$ | 1369            | 3136       | 4489       | 5929       | 7396       | 9409   |

First we need to compute the maximum likelihood estimates. There are 400 exams.

$$\bar{X} = \frac{1}{400} \left( 40 \times 120 + 55 \times 150 + 65 \times 60 + 75 \times 20 + 85 \times 10 + 95 \times 40 \right) = 57.75$$

Then the estimate for variance is

$$\frac{1}{400} \sum_{k=1}^{400} (X_i - 56.5)^2 = \frac{1}{400} \sum_{k=1}^{400} X_i^2 - \bar{X}^2$$

To find the first term,  $\frac{1}{400} \sum_{k=1}^{400} X_i^2 =$

$$\frac{1}{400} \left( 1369 \times 120 + 3136 \times 150 + 4489 \times 60 + 5929 \times 20 + 7396 \times 10 + 9409 \times 40 \right) = 3682.3$$

Thus the sample variance is

$$3682.3 - (57.75)^2 = 347.24$$

Now we can compute the  $p_i$ .

$$p_1 = \frac{1}{\sqrt{2\pi}\sqrt{347.24}} \int_{-\infty}^{50} e^{-\frac{1}{2} \frac{(x-56.5)^2}{347.24}} dx = 0.36361$$

$$p_2 = \frac{1}{\sqrt{2\pi}\sqrt{347.24}} \int_{50}^{60} e^{-\frac{1}{2} \frac{(x-56.5)^2}{347.24}} dx = 0.21088$$

$$p_3 = \frac{1}{\sqrt{2\pi}\sqrt{347.24}} \int_{60}^{70} e^{-\frac{1}{2} \frac{(x-56.5)^2}{347.24}} dx = 0.19112$$

$$p_4 = \frac{1}{\sqrt{2\pi}\sqrt{347.24}} \int_{70}^{80} e^{-\frac{1}{2} \frac{(x-56.5)^2}{347.24}} dx = 0.13075$$

$$p_5 = \frac{1}{\sqrt{2\pi}\sqrt{347.24}} \int_{80}^{90} e^{-\frac{1}{2} \frac{(x-56.5)^2}{347.24}} dx = 6.7526 \times 10^{-2}$$

$$p_6 = \frac{1}{\sqrt{2\pi}\sqrt{347.24}} \int_{90}^{\infty} e^{-\frac{1}{2} \frac{(x-56.5)^2}{347.24}} dx = 3.6108 \times 10^{-2}$$

Assemble  $D$ . After some computations one finds

$$D = 134.96$$

Now since there are two parameters,  $D$  is  $\chi^2(3)$ ,  $(6-1)-2$ . If you use the data cursor on a graph of the distribution function, you find the pair  $(15.3, .9984)$ . Thus if the null hypothesis is true that this sample is normally distributed, it would involve a probability of less than .01. Therefore, the null hypothesis can be rejected with probability larger than .99. This isn't quite true of course. Clearly if you used more disjoint intervals, you should be more sure that the approximation is good, so there is a little fuzziness in this goodness of fit test. However, it does help to quantify the appearance that the distribution is not normal. In the above example, this seems fairly clear just from looking at the scores, but this allows you to give numbers to justify its lack of normality.

Incidentally, if you are not careful, you can get such a distribution of scores on a final exam. All you need are faculty who wax creative rather than focussing on published outcomes. One wonders whether it is reasonable to base assigning letter grades on an assumption that the final exam scores are normally distributed, if the hypothesis that this is so can be rejected with high probability according to the above procedure. However, I think that it is often the case that people who follow these automatic procedures do not do goodness of fit tests like that just described.

**PROCEDURE 39.6.7** *If the distribution depends on  $s$  parameters, modify Procedure 39.6.5 as follows. First replace each parameter with its maximum likelihood estimate then do exactly the same thing to define  $D$  only this time, it is  $\chi^2((r-1)-s)$ .*

## 39.7 Contingency Tables

Another application of Proposition 39.6.2 and its generalization to when the distribution depends on parameters is to the notion of contingency tables. These can come in any size including more than two dimensions, but I will give a simple example to illustrate and leave to the reader whatever generalization is appropriate. Here is such a table.

|       | $A_1$    | $A_2$    | $A_3$    |
|-------|----------|----------|----------|
| $B_1$ | $p_{11}$ | $p_{12}$ | $p_{13}$ |
| $B_2$ | $p_{21}$ | $p_{22}$ | $p_{23}$ |

For example, you might be looking at the people in a city and  $B_1$  is the event that the person is female and  $B_2$  the event that the person is male while  $A_1$  might be that the person is a democrat,  $A_2$  the person is a republican and  $A_3$  the person is neither one. A given person will be in exactly one of the  $A_i \cap B_j$ .

The numbers  $p_{ij}$  are probabilities and  $\sum_j \sum_i p_{ij} = 1$ . Thus there is a random variable  $Z$  and  $p_{ij}$  is  $P(Z \in A_i \cap B_j) = P(A_i \cap B_j)$ . Denote as  $p_{.j}$  the marginal probability  $\sum_i p_{ij}$  and

$p_{i\cdot}$  the marginal probability  $\sum_j p_{ij}$ . Thus  $p_{\cdot j} = P(Z \in A_j) = P(A_j)$  and  $p_{i\cdot} = P(Z \in B_i) = P(B_i)$ .

The problem of interest is whether the events  $A_i$  and  $B_j$  are independent. Is it the case that

$$P(A_i \cap B_j) = P(A_i)P(B_j)?$$

In other words, is  $p_{ij} = p_{\cdot j}p_{i\cdot}$ ? If you knew each  $p_{ij}$  this would be no problem but you have no idea about  $p_{ij}$ .

The null hypothesis will be that  $p_{ij} = p_{\cdot j}p_{i\cdot}$ . Then the  $p_{\cdot j}, p_{i\cdot}$  are to be considered as parameters. The first item is to find the maximum likelihood estimates for these based on a random sample of size  $n$   $X_{ij}, i = 1, 2$  and  $j = 1, 2, 3$ . Here the sample has been indexed according to which "cell"  $B_i \cap A_j$  contains the sample point. Then the likelihood based on the null hypothesis is to maximize

$$\prod_{i=1}^2 \prod_{j=1}^3 p_{\cdot j}^{X_{ij}} p_{i\cdot}^{X_{ij}}$$

subject to the constraint that  $\sum_i p_{i\cdot} = 1$  as usual, it works best to maximize the  $\ln$  of the above. Thus maximize

$$\sum_{i=1}^2 \sum_{j=1}^3 (X_{ij} \ln(p_{\cdot j}) + X_{ij} \ln(p_{i\cdot})), \quad \sum_i p_{i\cdot} = 1$$

First consider the  $p_{i\cdot}$ . This amounts to maximizing

$$\sum_{i=1}^2 S_i \ln(p_{i\cdot}), \quad \sum_i p_{i\cdot} = 1$$

where  $S_i \equiv \sum_j X_{ij}$ . Using the method of Lagrange multipliers, we need

$$\left( \frac{S_1}{p_{1\cdot}} \quad \frac{S_2}{p_{2\cdot}} \right) = \lambda \left( 1 \quad 1 \right)$$

thus  $\lambda p_{1\cdot} = S_1, \lambda p_{2\cdot} = S_2$ . Thus  $\frac{S_2}{\lambda} + \frac{S_1}{\lambda} = 1$  and so  $\frac{\sum_i \sum_j X_{ij}}{\lambda} = 1$  and so  $\lambda = n$ . Then

$$\hat{p}_{i\cdot} = \frac{S_i}{n} = \frac{\sum_j X_{ij}}{n}$$

where this is the maximum likelihood estimate for  $p_{i\cdot}$ . Similar reasoning shows that

$$\hat{p}_{\cdot j} = \frac{\sum_i X_{ij}}{n}.$$

Now form

$$D \equiv \sum_{i,j} \frac{(X_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

By what was explained above, this is  $\chi^2((2 \times 3 - 1) - 3)$ . The reason there is a 3 there rather than a 5 is that there are only 3 unknown parameters due to the fact that  $\sum_i p_{i\cdot} = 1, \sum_j p_{\cdot j} = 1$ . In general, if the table is  $r \times s$ , the above expression would be

$$\chi^2((rs - 1) - (r + s - 2))$$

This justifies the following proposition.



**Proposition 39.7.1** *Let there be an  $r \times s$  contingency table such that the random variable is in exactly one of  $B_i \cap A_j$  for  $i = 1, \dots, r, j = 1, \dots, s$ . If  $P(B_i \cap A_j) = P(B_i)P(A_j)$  for all  $i, j$ , then if a sample is taken of size  $n$  and  $X_{ij}$  is the observed number in  $B_i \cap A_j$ , then when  $n$  is large,*

$$D \equiv \sum_{i,j} \frac{(X_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

*is distributed as  $\mathcal{X}^2((rs-1) - (r+s-2)) = \mathcal{X}^2(rs-s-r+1)$ .*

Assuming the null hypothesis that the events  $B_i$  and  $A_j$  are independent, one can now test this hypothesis by using a graph or table for  $\mathcal{X}^2(rs-s-r+1)$ .

**Example 39.7.2** *You have a  $3 \times 2$  contingency table, three rows and two columns. Also the number in a random sample is 900. The numbers of observations found in the various positions are illustrated in the following.*

|     |     |
|-----|-----|
| 120 | 300 |
| 180 | 80  |
| 100 | 120 |

*Determine whether the underlying contingency table has the property that the events could be independent. If the probability is no more than .01 that the events are independent, reject the null hypothesis. Otherwise conclude that the events might be independent.*

In the above,  $n = 900$ . Now let's find the  $\hat{p}$ .

$$\hat{p}_{1\cdot} = \frac{420}{900}, \hat{p}_{2\cdot} = \frac{260}{900}, \hat{p}_{3\cdot} = \frac{220}{900}$$

$$\hat{p}_{\cdot 1} = \frac{400}{900}, \hat{p}_{\cdot 2} = \frac{500}{900}$$

Now assemble  $D$ .

$$D = \frac{(120 - 900(\frac{420}{900})(\frac{400}{900}))^2}{900(\frac{420}{900})(\frac{400}{900})} + \frac{(300 - 900(\frac{420}{900})(\frac{500}{900}))^2}{900(\frac{420}{900})(\frac{500}{900})}$$

$$+ \frac{(180 - 900(\frac{260}{900})(\frac{400}{900}))^2}{900(\frac{260}{900})(\frac{400}{900})} + \frac{(80 - 900(\frac{260}{900})(\frac{500}{900}))^2}{900(\frac{260}{900})(\frac{500}{900})}$$

$$+ \frac{(100 - 900(\frac{220}{900})(\frac{400}{900}))^2}{900(\frac{220}{900})(\frac{400}{900})} + \frac{(120 - 900(\frac{220}{900})(\frac{500}{900}))^2}{900(\frac{220}{900})(\frac{500}{900})}$$

Now compute this.

$$D = 107.64$$

This is way too big to accept the null hypothesis. The events are not independent. The statistic is distributed as  $\mathcal{X}^2(2)$  and a table gives probability 1 that the variable is less than 10. Yet  $D$  is larger than 100.

**Example 39.7.3** You have a  $2 \times 2$  contingency table, three rows and two columns. Also the number in a random sample is 1500. The numbers of observations found in the various positions are illustrated in the following.

|     |     |
|-----|-----|
| 297 | 196 |
| 600 | 407 |

Determine whether the underlying contingency table has the property that the events could be independent. If the probability is no more than .5 that the events are independent, reject the null hypothesis. Otherwise conclude that the events might be independent.

First, what are the  $\hat{p}$ ?

$$\hat{p}_{1\cdot} = \frac{493}{1500}, \hat{p}_{2\cdot} = \frac{1007}{1500}, \hat{p}_{\cdot 1} = \frac{897}{1500}, \hat{p}_{\cdot 2} = \frac{201}{500}$$

Now find  $D$ .

$$\begin{aligned} D = & \frac{(297 - 1500(\frac{493}{1500})(\frac{897}{1500}))^2}{1500(\frac{493}{1500})(\frac{897}{1500})} + \frac{(196 - 1500(\frac{493}{1500})(\frac{201}{500}))^2}{1500(\frac{493}{1500})(\frac{201}{500})} \\ & + \frac{(600 - 1500(\frac{1007}{1500})(\frac{897}{1500}))^2}{1500(\frac{493}{1500})(\frac{201}{500})} + \frac{(407 - 1500(\frac{1007}{1500})(\frac{201}{500}))^2}{1500(\frac{1007}{1500})(\frac{201}{500})} \\ & D = 7.6237 \times 10^{-2} \end{aligned}$$

This is distributed as  $\mathcal{X}^2((4-1)-(2)) = \mathcal{X}^2(1)$ . From a table or graph, (.1, .248) is on the graph of the distribution function. Therefore, since  $D$  is far smaller than .1, we don't reject the hypothesis that the sets are independent. Not being independent is indicated by  $D$  being larger than some number  $a$  where the probability that a  $\mathcal{X}^2(1)$  random variable is larger than  $a$  is very small, but this  $D$  is very small, so it is highly probable under the null hypothesis that  $\mathcal{X}^2(1) > .1$ . **This does not mean that the sets are independent. It only means we don't reject the possibility that they are. This is termed acceptance of the hypothesis but you might fail to have independence even though you accept the hypothesis.**

## Appendix A

# The Theory Of The Riemannnn Integral\*



### A.1 An Important Warning

If you read and understand this appendix on the Riemann integral you will become abnormal if you are not already that way. You will laugh at atrocious puns. You will be unpopular with well adjusted confident people, especially religious people who love to accept on faith inconsistent decrees of authority figures. Furthermore, your confidence will be completely shattered. Virtually nothing will be obvious to you ever again. Consider whether it would be better to accept the superficial presentation given earlier than to attempt to acquire deep understanding of the integral, risking your self esteem and confidence, before proceeding further. This is only here for those who need explanations and are not content to accept on faith. This chapter is one of the worst things I have seen and I don't know how to improve it without losing the rigor. I think it is a good illustration why, if you want to do integration, you should approach it as a Lebesgue integral. This is found in my book Calculus of functions of real and complex variables or Calculus of One and Many Variables. It will be much more abstract, but much less filled with mind numbing technicalities.

### A.2 Basic Definition

The definition of the Riemannnn integral of a function of  $n$  variables uses the following definition.

**Definition A.2.1** For  $i = 1, \dots, n$ , let  $\{\alpha_k^i\}_{k=-\infty}^{\infty}$  be points on  $\mathbb{R}$  which satisfy

$$\lim_{k \rightarrow \infty} \alpha_k^i = \infty, \lim_{k \rightarrow -\infty} \alpha_k^i = -\infty, \alpha_k^i < \alpha_{k+1}^i. \quad (1.1)$$

For such sequences, define a grid on  $\mathbb{R}^n$  denoted by  $\mathcal{G}$  or  $\mathcal{F}$  as the collection of boxes of the form

$$Q = \prod_{i=1}^n [\alpha_{j_i}^i, \alpha_{j_i+1}^i]. \quad (1.2)$$

If  $\mathcal{G}$  is a grid,  $\mathcal{F}$  is called a refinement of  $\mathcal{G}$  if every box of  $\mathcal{G}$  is the union of boxes of  $\mathcal{F}$ .

**Lemma A.2.2** *If  $\mathcal{G}$  and  $\mathcal{F}$  are two grids, they have a common refinement, denoted here by  $\mathcal{G} \vee \mathcal{F}$ .*

**Proof:** Let  $\{\alpha_k^i\}_{k=-\infty}^{\infty}$  be the sequences used to construct  $\mathcal{G}$  and let  $\{\beta_k^i\}_{k=-\infty}^{\infty}$  be the sequence used to construct  $\mathcal{F}$ . Now let  $\{\gamma_k^i\}_{k=-\infty}^{\infty}$  denote the union of  $\{\alpha_k^i\}_{k=-\infty}^{\infty}$  and  $\{\beta_k^i\}_{k=-\infty}^{\infty}$ . It is necessary to show that for each  $i$  these points can be arranged in order. To do so, let  $\gamma_0^i \equiv \alpha_0^i$ . Now if

$$\gamma_{-j}^i, \dots, \gamma_0^i, \dots, \gamma_j^i$$

have been chosen such that they are in order and all distinct, let  $\gamma_{j+1}^i$  be the first element of

$$\{\alpha_k^i\}_{k=-\infty}^{\infty} \cup \{\beta_k^i\}_{k=-\infty}^{\infty} \quad (1.3)$$

which is larger than  $\gamma_j^i$  and let  $\gamma_{-(j+1)}^i$  be the last element of (1.3) which is strictly smaller than  $\gamma_{-j}^i$ . The assumption (1.1) insures such a first and last element exists. Now let the grid  $\mathcal{G} \vee \mathcal{F}$  consist of boxes of the form

$$Q \equiv \prod_{i=1}^n [\gamma_{j_i}^i, \gamma_{j_i+1}^i]. \quad \blacksquare$$

The Riemannn integral is only defined for functions  $f$  which are bounded and are equal to zero off some bounded set  $D$ . In what follows  $f$  will always be such a function.

**Definition A.2.3** *Let  $f$  be a bounded function which equals zero off a bounded set  $D$ , and let  $\mathcal{G}$  be a grid. For  $Q \in \mathcal{G}$ , define*

$$M_Q(f) \equiv \sup \{f(x) : x \in Q\}, \quad m_Q(f) \equiv \inf \{f(x) : x \in Q\}. \quad (1.4)$$

Also define for  $Q$  a box, the volume of  $Q$ , denoted by  $v(Q)$  by

$$v(Q) \equiv \prod_{i=1}^n (b_i - a_i), \quad Q \equiv \prod_{i=1}^n [a_i, b_i].$$

Now define upper sums,  $\mathcal{U}_{\mathcal{G}}(f)$  and lower sums,  $\mathcal{L}_{\mathcal{G}}(f)$  with respect to the indicated grid, by the formulas

$$\mathcal{U}_{\mathcal{G}}(f) \equiv \sum_{Q \in \mathcal{G}} M_Q(f) v(Q), \quad \mathcal{L}_{\mathcal{G}}(f) \equiv \sum_{Q \in \mathcal{G}} m_Q(f) v(Q).$$

A function of  $n$  variables is Riemannn integrable when there is a unique number between all the upper and lower sums. This number is the value of the integral.

Note that in this definition,  $M_Q(f) = m_Q(f) = 0$  for all but finitely many  $Q \in \mathcal{G}$  so there are no convergence questions to be considered here.

**Lemma A.2.4** *If  $\mathcal{F}$  is a refinement of  $\mathcal{G}$  then*

$$\mathcal{U}_{\mathcal{G}}(f) \geq \mathcal{U}_{\mathcal{F}}(f), \mathcal{L}_{\mathcal{G}}(f) \leq \mathcal{L}_{\mathcal{F}}(f).$$

*Also if  $\mathcal{F}$  and  $\mathcal{G}$  are two grids,*

$$\mathcal{L}_{\mathcal{G}}(f) \leq \mathcal{U}_{\mathcal{F}}(f).$$

**Proof:** For  $P \in \mathcal{G}$  let  $\hat{P}$  denote the set

$$\{Q \in \mathcal{F} : Q \subseteq P\}.$$

Then  $P = \cup \hat{P}$  and

$$\begin{aligned} \mathcal{L}_{\mathcal{F}}(f) &\equiv \sum_{Q \in \mathcal{F}} m_Q(f) v(Q) = \sum_{P \in \mathcal{G}} \sum_{Q \in \hat{P}} m_Q(f) v(Q) \\ &\geq \sum_{P \in \mathcal{G}} m_P(f) \sum_{Q \in \hat{P}} v(Q) = \sum_{P \in \mathcal{G}} m_P(f) v(P) \equiv \mathcal{L}_{\mathcal{G}}(f). \end{aligned}$$

Similarly, the other inequality for the upper sums is valid.

To verify the last assertion of the lemma, use Lemma A.2.2 to write

$$\mathcal{L}_{\mathcal{G}}(f) \leq \mathcal{L}_{\mathcal{G} \vee \mathcal{F}}(f) \leq \mathcal{U}_{\mathcal{G} \vee \mathcal{F}}(f) \leq \mathcal{U}_{\mathcal{F}}(f). \quad \blacksquare$$

This lemma makes it possible to define the Riemann integral.

**Definition A.2.5** *Define an upper and a lower integral as follows.*

$$\bar{I}(f) \equiv \inf \{ \mathcal{U}_{\mathcal{G}}(f) : \mathcal{G} \text{ is a grid} \},$$

$$\underline{I}(f) \equiv \sup \{ \mathcal{L}_{\mathcal{G}}(f) : \mathcal{G} \text{ is a grid} \}.$$

**Lemma A.2.6**  $\bar{I}(f) \geq \underline{I}(f)$ .

**Proof:** From Lemma A.2.4 it follows for any two grids  $\mathcal{G}$  and  $\mathcal{F}$ ,

$$\mathcal{L}_{\mathcal{G}}(f) \leq \mathcal{U}_{\mathcal{F}}(f).$$

Therefore, taking the supremum for all grids on the left in this inequality,

$$\underline{I}(f) \leq \mathcal{U}_{\mathcal{F}}(f)$$

for all grids  $\mathcal{F}$ . Taking the infimum in this inequality, yields the conclusion of the lemma.  $\blacksquare$

**Definition A.2.7** *A bounded function  $f$  which equals zero off a bounded set  $D$ , is said to be Riemann integrable, written as  $f \in \mathcal{R}(\mathbb{R}^n)$  exactly when  $\underline{I}(f) = \bar{I}(f)$ . In this case define*

$$\int f dV \equiv \int f dx = \bar{I}(f) = \underline{I}(f).$$

As in the case of integration of functions of one variable, one obtains the Riemann criterion which is stated as the following theorem.

**Theorem A.2.8** (Riemannn criterion)  $f \in \mathcal{R}(\mathbb{R}^n)$  if and only if for all  $\varepsilon > 0$  there exists a grid  $\mathcal{G}$  such that

$$\mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) < \varepsilon.$$

**Proof:** If  $f \in \mathcal{R}(\mathbb{R}^n)$ , then  $\bar{I}(f) = \underline{I}(f)$  and so there exist grids  $\mathcal{G}$  and  $\mathcal{F}$  such that

$$\mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{F}}(f) \leq \bar{I}(f) + \frac{\varepsilon}{2} - \left( \underline{I}(f) - \frac{\varepsilon}{2} \right) = \varepsilon.$$

Then letting  $\mathcal{H} = \mathcal{G} \vee \mathcal{F}$ , Lemma A.2.4 implies

$$\mathcal{U}_{\mathcal{H}}(f) - \mathcal{L}_{\mathcal{H}}(f) \leq \mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{F}}(f) < \varepsilon.$$

Conversely, if for all  $\varepsilon > 0$  there exists  $\mathcal{G}$  such that

$$\mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) < \varepsilon,$$

then

$$\bar{I}(f) - \underline{I}(f) \leq \mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) < \varepsilon.$$

Since  $\varepsilon > 0$  is arbitrary, this proves the theorem. ■

### A.3 Basic Properties

It is important to know that certain combinations of Riemannn integrable functions are Riemannn integrable. The following theorem will include all the important cases.

**Theorem A.3.1** Let  $f, g \in \mathcal{R}(\mathbb{R}^n)$  and let  $\phi : K \rightarrow \mathbb{R}$  be continuous where  $K$  is a compact set in  $\mathbb{R}^2$  containing  $f(\mathbb{R}^n) \times g(\mathbb{R}^n)$ . Also suppose that  $\phi(0, 0) = 0$ . Then defining

$$h(\mathbf{x}) \equiv \phi(f(\mathbf{x}), g(\mathbf{x})),$$

it follows that  $h$  is also in  $\mathcal{R}(\mathbb{R}^n)$ .

**Proof:** Let  $\varepsilon > 0$  and let  $\delta_1 > 0$  be such that if  $(y_i, z_i), i = 1, 2$  are points in  $K$ , such that  $|z_1 - z_2| \leq \delta_1$  and  $|y_1 - y_2| \leq \delta_1$ , then

$$|\phi(y_1, z_1) - \phi(y_2, z_2)| < \varepsilon.$$

Let  $0 < \delta < \min(\delta_1, \varepsilon, 1)$ . Let  $\mathcal{G}$  be a grid with the property that for  $Q \in \mathcal{G}$ , the diameter of  $Q$  is less than  $\delta$  and also for  $k = f, g$ ,

$$\mathcal{U}_{\mathcal{G}}(k) - \mathcal{L}_{\mathcal{G}}(k) < \delta^2. \tag{1.5}$$

Then defining for  $k = f, g$ ,

$$\mathcal{P}_k \equiv \{Q \in \mathcal{G} : M_Q(k) - m_Q(k) > \delta\},$$

it follows

$$\delta^2 > \sum_{Q \in \mathcal{G}} (M_Q(k) - m_Q(k)) v(Q) \geq$$

$$\sum_{\mathcal{P}_k} (M_Q(k) - m_Q(k)) v(Q) \geq \delta \sum_{\mathcal{P}_k} v(Q)$$

and so for  $k = f, g$ ,

$$\varepsilon > \delta > \sum_{\mathcal{P}_k} v(Q). \quad (1.6)$$

Suppose for  $k = f, g$ ,

$$M_Q(k) - m_Q(k) \leq \delta.$$

Then if  $x_1, x_2 \in Q$ ,

$$|f(x_1) - f(x_2)| < \delta, \text{ and } |g(x_1) - g(x_2)| < \delta.$$

Therefore,

$$|h(x_1) - h(x_2)| \equiv |\phi(f(x_1), g(x_1)) - \phi(f(x_2), g(x_2))| < \varepsilon$$

and it follows that

$$|M_Q(h) - m_Q(h)| \leq \varepsilon.$$

Now let

$$\mathcal{S} \equiv \{Q \in \mathcal{G} : 0 < M_Q(k) - m_Q(k) \leq \delta, k = f, g\}.$$

Thus the union of the boxes in  $\mathcal{S}$  is contained in some large box,  $R$ , which depends only on  $f$  and  $g$  and also, from the assumption that  $\phi(0, 0) = 0$ ,  $M_Q(h) - m_Q(h) = 0$ , unless  $Q \subseteq R$ . Then

$$\begin{aligned} \mathcal{U}_g(h) - \mathcal{L}_g(h) &\leq \sum_{Q \in \mathcal{P}_f} (M_Q(h) - m_Q(h)) v(Q) + \\ &\quad \sum_{Q \in \mathcal{P}_g} (M_Q(h) - m_Q(h)) v(Q) + \sum_{Q \in \mathcal{S}} \delta v(Q). \end{aligned}$$

Now since  $K$  is compact, it follows  $\phi(K)$  is bounded and so there exists a constant  $C$ , depending only on  $h$  and  $\phi$  such that  $M_Q(h) - m_Q(h) < C$ . Therefore, the above inequality implies

$$\mathcal{U}_g(h) - \mathcal{L}_g(h) \leq C \sum_{Q \in \mathcal{P}_f} v(Q) + C \sum_{Q \in \mathcal{P}_g} v(Q) + \sum_{Q \in \mathcal{S}} \delta v(Q),$$

which by (1.6) implies

$$\mathcal{U}_g(h) - \mathcal{L}_g(h) \leq 2C\varepsilon + \delta v(R) \leq 2C\varepsilon + \varepsilon v(R).$$

Since  $\varepsilon$  is arbitrary, the Riemann criterion is satisfied and so  $h \in \mathcal{R}(\mathbb{R}^n)$ . ■

**Corollary A.3.2** *Let  $f, g \in \mathcal{R}(\mathbb{R}^n)$  and let  $a, b \in \mathbb{R}$ . Then  $af + bg$ ,  $fg$ , and  $|f|$  are all in  $\mathcal{R}(\mathbb{R}^n)$ . Also,*

$$\int_{\mathbb{R}^n} (af + bg) dx = a \int_{\mathbb{R}^n} f dx + b \int_{\mathbb{R}^n} g dx, \quad (1.7)$$

and

$$\int_{\mathbb{R}^n} |f| dx \geq \left| \int_{\mathbb{R}^n} f dx \right|. \quad (1.8)$$

**Proof:** Each of the combinations of functions described above is Riemann integrable by Theorem A.3.1. For example, to see  $af + bg \in \mathcal{R}(\mathbb{R}^n)$  consider  $\phi(y, z) \equiv ay + bz$ . This is clearly a continuous function of  $(y, z)$  such that  $\phi(0, 0) = 0$ . To obtain  $|f| \in \mathcal{R}(\mathbb{R}^n)$ , let  $\phi(y, z) \equiv |y|$ . It remains to verify the formulas. To do so, let  $\mathcal{G}$  be a grid with the property that for  $k = f, g, |f|$  and  $af + bg$ ,

$$\mathcal{U}_{\mathcal{G}}(k) - \mathcal{L}_{\mathcal{G}}(k) < \varepsilon. \quad (1.9)$$

Consider (1.7). For each  $Q \in \mathcal{G}$  pick a point in  $Q$ ,  $\mathbf{x}_Q$ . Then

$$\sum_{Q \in \mathcal{G}} k(\mathbf{x}_Q) v(Q) \in [\mathcal{L}_{\mathcal{G}}(k), \mathcal{U}_{\mathcal{G}}(k)]$$

and so

$$\left| \int k dx - \sum_{Q \in \mathcal{G}} k(\mathbf{x}_Q) v(Q) \right| < \varepsilon.$$

Consequently, since

$$\begin{aligned} & \sum_{Q \in \mathcal{G}} (af + bg)(\mathbf{x}_Q) v(Q) \\ &= a \sum_{Q \in \mathcal{G}} f(\mathbf{x}_Q) v(Q) + b \sum_{Q \in \mathcal{G}} g(\mathbf{x}_Q) v(Q), \end{aligned}$$

it follows

$$\begin{aligned} & \left| \int (af + bg) dx - a \int f dx - b \int g dx \right| \leq \\ & \left| \int (af + bg) dx - \sum_{Q \in \mathcal{G}} (af + bg)(\mathbf{x}_Q) v(Q) \right| + \\ & \left| a \sum_{Q \in \mathcal{G}} f(\mathbf{x}_Q) v(Q) - a \int f dx \right| + \left| b \sum_{Q \in \mathcal{G}} g(\mathbf{x}_Q) v(Q) - b \int g dx \right| \\ & \leq \varepsilon + |a| \varepsilon + |b| \varepsilon. \end{aligned}$$

Since  $\varepsilon$  is arbitrary, this establishes (1.7) and shows the integral is linear.

It remains to establish the inequality (1.8). By (1.9), and the triangle inequality for sums,

$$\begin{aligned} & \int |f| dx + \varepsilon \geq \sum_{Q \in \mathcal{G}} |f(\mathbf{x}_Q)| v(Q) \\ & \geq \left| \sum_{Q \in \mathcal{G}} f(\mathbf{x}_Q) v(Q) \right| \geq \left| \int f dx \right| - \varepsilon. \end{aligned}$$

Then since  $\varepsilon$  is arbitrary, this establishes the desired inequality. ■



## A.4 Which Functions Are Integrable?

Which functions are in  $\mathcal{R}(\mathbb{R}^n)$ ? As in the case of integrals of functions of one variable, this is an important question. It turns out the Riemann integrable functions are characterized by being continuous except on a very small set. This has to do with Jordan content.

**Definition A.4.1** A bounded set  $E$ , has Jordan content 0 or content 0 if for every  $\varepsilon > 0$  there exists a grid  $\mathcal{G}$  such that

$$\sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon.$$

This symbol says to sum the volumes of all boxes from  $\mathcal{G}$  which have nonempty intersection with  $E$ .

Next it is necessary to define the oscillation of a function.

**Definition A.4.2** Let  $f$  be a function defined on  $\mathbb{R}^n$  and let

$$\omega_{f,r}(x) \equiv \sup \{|f(z) - f(y)| : z, y \in B(x, r)\}.$$

This is called the oscillation of  $f$  on  $B(x, r)$ . Note that this function of  $r$  is decreasing in  $r$ . Define the oscillation of  $f$  as

$$\omega_f(x) \equiv \lim_{r \rightarrow 0+} \omega_{f,r}(x).$$

Note that as  $r$  decreases, the function  $\omega_{f,r}(x)$  decreases. It is also bounded below by 0, so the limit must exist and equals  $\inf \{\omega_{f,r}(x) : r > 0\}$ . (Why?) Then the following simple lemma whose proof follows directly from the definition of continuity gives the reason for this definition.

**Lemma A.4.3** A function  $f$  is continuous at  $x$  if and only if  $\omega_f(x) = 0$ .

This concept of oscillation gives a way to define how discontinuous a function is at a point. The discussion will depend on the following fundamental lemma which gives the existence of something called the Lebesgue number.

**Definition A.4.4** Let  $\mathcal{C}$  be a set whose elements are sets of  $\mathbb{R}^n$  and let  $K \subseteq \mathbb{R}^n$ . The set  $\mathcal{C}$  is called a cover of  $K$  if every point of  $K$  is contained in some set of  $\mathcal{C}$ . If the elements of  $\mathcal{C}$  are open sets, it is called an open cover.

**Lemma A.4.5** Let  $K$  be sequentially compact and let  $\mathcal{C}$  be an open cover of  $K$ . Then there exists  $r > 0$  such that whenever  $x \in K$ ,  $B(x, r)$  is contained in some set of  $\mathcal{C}$ .

**Proof:** Suppose this is not so. Then letting  $r_n = 1/n$ , there exists  $x_n \in K$  such that  $B(x_n, r_n)$  is not contained in any set of  $\mathcal{C}$ . Since  $K$  is sequentially compact, there is a subsequence,  $x_{n_k}$  which converges to a point  $x \in K$ . But there exists  $\delta > 0$  such that  $B(x, \delta) \subseteq U$  for some  $U \in \mathcal{C}$ . Let  $k$  be so large that  $1/k < \delta/2$  and  $|x_{n_k} - x| < \delta/2$  also. Then if  $z \in B(x_{n_k}, r_{n_k})$ , it follows

$$|z - x| \leq |z - x_{n_k}| + |x_{n_k} - x| < \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

and so  $B(x_{n_k}, r_{n_k}) \subseteq U$  contrary to supposition. Therefore, the desired number exists after all. ■

**Theorem A.4.6** Let  $f$  be a bounded function which equals zero off a bounded set and let  $W$  denote the set of points where  $f$  fails to be continuous. Then  $f \in \mathcal{R}(\mathbb{R}^n)$  if  $W$  has content zero. That is, for all  $\varepsilon > 0$  there exists a grid  $\mathcal{G}$  such that

$$\sum_{Q \in \mathcal{G}_W} v(Q) < \varepsilon \quad (1.10)$$

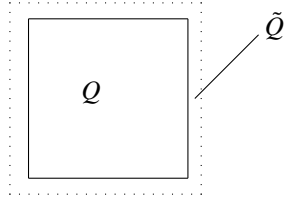
where

$$\mathcal{G}_W \equiv \{Q \in \mathcal{G} : Q \cap W \neq \emptyset\}.$$

**Proof:** Let  $W$  have content zero. Also let  $|f(x)| < C/2$  for all  $x \in \mathbb{R}^n$ , let  $\varepsilon > 0$  be given, and let  $\mathcal{G}$  be a grid which satisfies (1.10). Since  $f$  equals zero off some bounded set, there exists  $R$  such that  $f$  equals zero off of  $B(\mathbf{0}, \frac{R}{2})$ . Thus  $W \subseteq B(\mathbf{0}, \frac{R}{2})$ . Also note that if  $\mathcal{G}$  is a grid for which (1.10) holds, then this inequality continues to hold if  $\mathcal{G}$  is replaced with a refined grid. Therefore, you may assume the diameter of every box in  $\mathcal{G}$  which intersects  $B(\mathbf{0}, R)$  is less than  $\frac{R}{3}$  and so all boxes of  $\mathcal{G}$  which intersect the set where  $f$  is nonzero are contained in  $B(\mathbf{0}, R)$ . Since  $W$  is bounded,  $\mathcal{G}_W$  contains only finitely many boxes. Letting

$$Q \equiv \prod_{i=1}^n [a_i, b_i]$$

be one of these boxes, enlarge the box slightly as indicated in the following picture.



The enlarged box is an open set of the form,

$$\tilde{Q} \equiv \prod_{i=1}^n (a_i - \eta_i, b_i + \eta_i)$$

where  $\eta_i$  is chosen small enough that if

$$\prod_{i=1}^n (b_i + \eta_i - (a_i - \eta_i)) \equiv v(\tilde{Q}),$$

and  $\widetilde{\mathcal{G}}_W$  denotes those  $\tilde{Q}$  for  $Q \in \mathcal{G}$  which have nonempty intersection with  $W$ , then

$$\sum_{\tilde{Q} \in \widetilde{\mathcal{G}}_W} v(\tilde{Q}) < \varepsilon \quad (1.11)$$

where  $\tilde{Q}$  is the box,

$$\prod_{i=1}^n ((a_i - 2\eta_i), b_i + 2\eta_i).$$

For each  $x \in \mathbb{R}^n$ , let  $r_x < \min(\eta_1/2, \dots, \eta_n/2)$  be such that

$$\omega_{f, r_x}(x) < \varepsilon + \omega_f(x). \quad (1.12)$$

Now let  $\mathfrak{C}$  denote all intersections of the form  $\tilde{Q} \cap B(x, r_x)$  such that  $x \in \overline{B(0, R)}$  so that  $\mathfrak{C}$  is an open cover of the compact set  $\overline{B(0, R)}$ . Let  $\delta$  be a Lebesgue number for this open cover of  $\overline{B(0, R)}$  and let  $\mathcal{F}$  be a refinement of  $\mathcal{G}$  such that every box in  $\mathcal{F}$  has diameter less than  $\delta$ . Now let  $\mathcal{F}_1$  consist of those boxes of  $\mathcal{F}$  which have nonempty intersection with  $B(0, R/2)$ . Thus all boxes of  $\mathcal{F}_1$  are contained in  $B(0, R)$  and each one is contained in some set of  $\mathfrak{C}$ . Let  $\mathfrak{C}_W$  be those open sets of  $\mathfrak{C}$ ,  $\tilde{Q} \cap B(x, r_x)$ , for which  $x \in W$ . Thus each of these sets is contained in some  $\tilde{Q}$  where  $Q \in \mathcal{G}_W$ . Let  $\mathcal{F}_W$  be those sets of  $\mathcal{F}_1$  which are subsets of some set of  $\mathfrak{C}_W$ . Thus

$$\sum_{Q \in \mathcal{F}_W} v(Q) < \varepsilon. \quad (1.13)$$

because each  $Q$  in  $\mathcal{F}_W$  is contained in a set  $\tilde{Q}$  described above and the sum of the volumes of these is less than  $\varepsilon$  by (1.11). Then

$$\begin{aligned} \mathcal{U}_{\mathcal{F}}(f) - \mathcal{L}_{\mathcal{F}}(f) &= \sum_{Q \in \mathcal{F}_W} (M_Q(f) - m_Q(f)) v(Q) \\ &+ \sum_{Q \in \mathcal{F}_1 \setminus \mathcal{F}_W} (M_Q(f) - m_Q(f)) v(Q). \end{aligned}$$

If  $Q \in \mathcal{F}_1 \setminus \mathcal{F}_W$ , then  $Q$  must be a subset of some set of  $\mathfrak{C} \setminus \mathfrak{C}_W$  since it is not in any set of  $\mathfrak{C}_W$ . Say  $Q \subseteq \tilde{Q} \cap B(x, r_x)$  where  $x \notin W$ . Therefore, from (1.12) and the observation that  $x \notin W$ , it follows  $\omega_f(x) = 0$  and so

$$M_Q(f) - m_Q(f) \leq \varepsilon.$$

Therefore, from (1.13) and the estimate on  $f$ ,

$$\begin{aligned} \mathcal{U}_{\mathcal{F}}(f) - \mathcal{L}_{\mathcal{F}}(f) &\leq \sum_{Q \in \mathcal{F}_W} C v(Q) + \sum_{Q \in \mathcal{F}_1 \setminus \mathcal{F}_W} \varepsilon v(Q) \\ &\leq C\varepsilon + \varepsilon(2R)^n, \end{aligned}$$

the estimate of the second sum coming from the fact that

$$B(0, R) \subseteq \prod_{i=1}^n [-R, R].$$

Since  $\varepsilon$  is arbitrary, this proves the theorem.<sup>1</sup> ■

**Definition A.4.7** A bounded set  $E$  is a Jordan set in  $\mathbb{R}^n$ , also called a contented set in  $\mathbb{R}^n$  if  $\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$ . The symbol  $\mathcal{X}_E$  means

$$\mathcal{X}_E(x) = \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{if } x \notin E \end{cases}$$

<sup>1</sup>In fact one cannot do any better. It can be shown that if a function is Riemann integrable, then it must be the case that for all  $\varepsilon > 0$ , (1.10) is satisfied for some grid  $\mathcal{G}$ . This along with what was just shown is known as Lebesgue's theorem after Lebesgue who discovered it in the early years of the twentieth century. Actually, he also invented a far superior integral which made the Riemann integral which is the topic of this appendix obsolete.

It is called the indicator function because it indicates whether  $x$  is in  $E$  according to whether it equals 1. For a function  $f \in \mathcal{R}(\mathbb{R}^n)$  and  $E$  a contented set,  $f\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$  by Corollary A.3.2. Then

$$\int_E f dV \equiv \int f \mathcal{X}_E dV.$$

So what are examples of contented sets?

**Theorem A.4.8** Suppose  $E$  is a bounded contented set in  $\mathbb{R}^n$  and  $f, g : E \rightarrow \mathbb{R}$  are two functions satisfying  $f(x) \geq g(x)$  for all  $x \in E$  and  $f\mathcal{X}_E$  and  $g\mathcal{X}_E$  are both in  $\mathcal{R}(\mathbb{R}^n)$ . Now define

$$P \equiv \{(x, x_{n+1}) : x \in E \text{ and } g(x) \leq x_{n+1} \leq f(x)\}.$$

Then  $P$  is a contented set in  $\mathbb{R}^{n+1}$ .

**Proof:** Let  $\mathcal{G}$  be a grid such that for  $k = f\mathcal{X}_E, g\mathcal{X}_E$ ,

$$\mathcal{U}_{\mathcal{G}}(k) - \mathcal{L}_{\mathcal{G}}(k) < \varepsilon/4. \quad (1.14)$$

Also let  $K \geq \sum_{j=1}^m v_n(Q_j)$  where the  $Q_j$  are the boxes which intersect  $E$ . Let  $\{a_i\}_{i=-\infty}^{\infty}$  be a sequence on  $\mathbb{R}$ ,  $a_i < a_{i+1}$  for all  $i$ , which includes

$$\begin{aligned} &M_{Q_j}(f\mathcal{X}_E) + \frac{\varepsilon}{4mK}, M_{Q_j}(f\mathcal{X}_E), M_{Q_j}(g\mathcal{X}_E), \\ &m_{Q_j}(f\mathcal{X}_E), m_{Q_j}(g\mathcal{X}_E), m_{Q_j}(g\mathcal{X}_E) - \frac{\varepsilon}{4mK} \end{aligned}$$

for all  $j = 1, \dots, m$ . Now define a grid on  $\mathbb{R}^{n+1}$  as follows.

$$\mathcal{G}' \equiv \{Q \times [a_i, a_{i+1}] : Q \in \mathcal{G}, i \in \mathbb{Z}\}$$

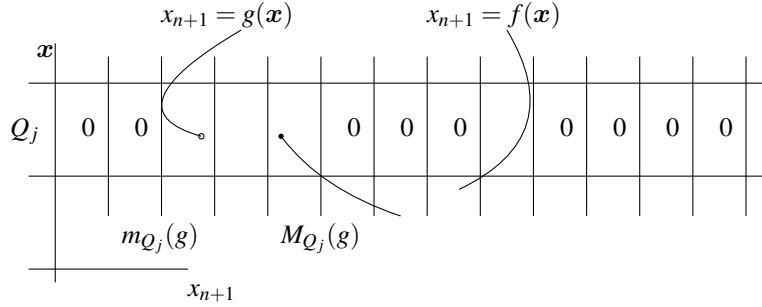
In words, this grid consists of all possible boxes of the form  $Q \times [a_i, a_{i+1}]$  where  $Q \in \mathcal{G}$  and  $a_i$  is a term of the sequence just described. It is necessary to verify that for  $P \in \mathcal{G}'$ ,  $\mathcal{X}_P \in \mathcal{R}(\mathbb{R}^{n+1})$ . This is done by showing that  $\mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P) - \mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P) < \varepsilon$  and then noting that  $\varepsilon > 0$  was arbitrary. For  $\mathcal{G}'$  just described, denote by  $Q'$  a box in  $\mathcal{G}'$ . Thus  $Q' = Q \times [a_i, a_{i+1}]$  for some  $i$ .

$$\begin{aligned} \mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P) - \mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P) &\equiv \sum_{Q' \in \mathcal{G}'} (M_{Q'}(\mathcal{X}_P) - m_{Q'}(\mathcal{X}_P)) v_{n+1}(Q') \\ &= \sum_{i=-\infty}^{\infty} \sum_{j=1}^m (M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P)) v_n(Q_j) (a_{i+1} - a_i) \end{aligned}$$

and all sums are bounded because the functions  $f$  and  $g$  are given to be bounded. Therefore, there are no limit considerations needed here. Thus

$$\begin{aligned} &\mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P) - \mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P) = \\ &\sum_{j=1}^m v_n(Q_j) \sum_{i=-\infty}^{\infty} (M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P)) (a_{i+1} - a_i). \end{aligned}$$

Consider the inside sum with the aid of the following picture.



In this picture, the little rectangles represent the boxes  $Q_j \times [a_i, a_{i+1}]$  for fixed  $j$ . The part of  $P$  having  $x$  contained in  $Q_j$  is between the two surfaces,  $x_{n+1} = g(x)$  and  $x_{n+1} = f(x)$  and there is a zero placed in those boxes for which

$$M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) = 0.$$

You see,  $\mathcal{X}_P$  has either the value of 1 or the value of 0 depending on whether  $(x, y)$  is contained in  $P$ . For the boxes shown with 0 in them, either all of the box is contained in  $P$  or none of the box is contained in  $P$ . Either way,

$$M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) = 0$$

on these boxes. However, on the boxes intersected by the surfaces, the value of

$$M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P)$$

is 1 because there are points in this box which are not in  $P$  as well as points which are in  $P$ . Because of the construction of  $\mathcal{G}'$  which included all values of

$$M_{Q_j}(f\mathcal{X}_E) + \frac{\varepsilon}{4mK}, M_{Q_j}(f\mathcal{X}_E), \\ M_{Q_j}(g\mathcal{X}_E), m_{Q_j}(f\mathcal{X}_E), m_{Q_j}(g\mathcal{X}_E)$$

for all  $j = 1, \dots, m$ ,

$$\sum_{i=-\infty}^{\infty} \left( M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) \right) (a_{i+1} - a_i) \leq \\ \sum_{\{i: m_{Q_j}(g\mathcal{X}_E) \leq a_i < M_{Q_j}(g\mathcal{X}_E)\}} 1(a_{i+1} - a_i) + \sum_{\{i: m_{Q_j}(f\mathcal{X}_E) \leq a_i < M_{Q_j}(f\mathcal{X}_E)\}} 1(a_{i+1} - a_i) \quad (1.15)$$

The first of the sums in (1.15) contains all possible terms for which

$$M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P)$$

might be 1 due to the graph of the bottom surface  $g\mathcal{X}_E$  while the second sum contains all possible terms for which the expression might be 1 due to the graph of the top surface  $f\mathcal{X}_E$ .

$$\leq \left( M_{Q_j}(g\mathcal{X}_E) + \frac{\varepsilon}{4mK} - m_{Q_j}(g\mathcal{X}_E) \right) + \left( M_{Q_j}(f\mathcal{X}_E) + \frac{\varepsilon}{4mK} - m_{Q_j}(f\mathcal{X}_E) \right)$$

$$= (M_{Q_j}(g \mathcal{X}_E) - m_{Q_j}(g \mathcal{X}_E)) + (M_{Q_j}(f \mathcal{X}_E) - m_{Q_j}(f \mathcal{X}_E)) + \frac{\varepsilon}{2m} \left( \sum_{j=1}^m v(Q_j) \right)^{-1}.$$

Therefore, by (1.14),

$$\begin{aligned} & \mathcal{U}_{g'}(\mathcal{X}_P) - \mathcal{L}_{g'}(\mathcal{X}_P) \leq \\ & \sum_{j=1}^m v_n(Q_j) [(M_{Q_j}(g \mathcal{X}_E) - m_{Q_j}(g \mathcal{X}_E)) + (M_{Q_j}(f \mathcal{X}_E) - m_{Q_j}(f \mathcal{X}_E))] \\ & + \sum_{j=1}^m v(Q_j) \frac{\varepsilon}{2m} \left( \sum_{j=1}^m v(Q_j) \right)^{-1} \\ & = \mathcal{U}_g(f) - \mathcal{L}_g(f) + \mathcal{U}_g(g) - \mathcal{L}_g(g) + \frac{\varepsilon}{2} \\ & < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Since  $\varepsilon > 0$  is arbitrary, this proves the theorem. ■

**Corollary A.4.9** Suppose  $f$  and  $g$  are continuous functions defined on  $E$ , a contented set in  $\mathbb{R}^n$  and that  $g(\mathbf{x}) \leq f(\mathbf{x})$  for all  $\mathbf{x} \in E$ . Then

$$P \equiv \{(\mathbf{x}, x_{n+1}) : \mathbf{x} \in E \text{ and } g(\mathbf{x}) \leq x_{n+1} \leq f(\mathbf{x})\}$$

is a contented set in  $\mathbb{R}^n$ .

**Proof:** Since  $E$  is contented, meaning  $\mathcal{X}_E$  is integrable, it follows from Theorem A.4.6 the set of discontinuities of  $\mathcal{X}_E$  has Jordan content 0. But the set of discontinuities of  $\mathcal{X}_E$  is  $\partial E$  defined as those points  $\mathbf{x}$  such that  $B(\mathbf{x}, r)$  contains points of  $E$  and points of  $E^C$  for every  $r > 0$ . Extend  $f$  and  $g$  to equal 0 off  $E$ . Then the set of discontinuities of these extended functions still denoted as  $f, g$  is  $\partial E$  which has Jordan content 0. This reduces to the situation of Theorem A.4.8. ■

As an example of how this can be applied, it is obvious a closed interval is a contented set in  $\mathbb{R}$ . Therefore, if  $f, g$  are two continuous functions with  $f(x) \geq g(x)$  for  $x \in [a, b]$ , it follows from the above theorem or its corollary that the set

$$P_1 \equiv \{(x, y) : g(x) \leq y \leq f(x)\}$$

is a contented set in  $\mathbb{R}^2$ . Now using the theorem and corollary again, suppose  $f_1(x, y) \geq g_1(x, y)$  for  $(x, y) \in P_1$  and  $f, g$  are continuous. Then the set

$$P_2 \equiv \{(x, y, z) : g_1(x, y) \leq z \leq f_1(x, y)\}$$

is a contented set in  $\mathbb{R}^3$ . Clearly you can continue this way obtaining examples of contented sets. ■

Note that as a special case, it follows that every box is a contented set. Therefore, if  $B_i$  is a box, functions of the form

$$\sum_{i=1}^m a_i \mathcal{X}_{B_i}$$

are integrable. These functions are called step functions.

The following theorem is analogous to the fact that in one dimension, when you integrate over a point, the answer is 0.

**Theorem A.4.10** *If a bounded set  $E$ , has Jordan content 0, then  $E$  is a Jordan (contented) set and if  $f$  is any bounded function defined on  $E$ , then  $f\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$  and*

$$\int_E f dV = 0.$$

**Proof:** Let  $m$  be a lower bound for  $f$  and let  $M$  be an upper bound. Let  $\mathcal{G}$  be a grid with

$$\sum_{Q \cap E \neq \emptyset} v(Q) < \frac{\varepsilon}{1 + (M - m)}.$$

Then

$$\mathcal{U}_{\mathcal{G}}(f\mathcal{X}_E) \leq \sum_{Q \cap E \neq \emptyset} Mv(Q) \leq \frac{\varepsilon M}{1 + (M - m)}$$

and

$$\mathcal{L}_{\mathcal{G}}(f\mathcal{X}_E) \geq \sum_{Q \cap E \neq \emptyset} mv(Q) \geq \frac{\varepsilon m}{1 + (M - m)}$$

and so

$$\begin{aligned} \mathcal{U}_{\mathcal{G}}(f\mathcal{X}_E) - \mathcal{L}_{\mathcal{G}}(f\mathcal{X}_E) &\leq \sum_{Q \cap E \neq \emptyset} Mv(Q) - \sum_{Q \cap E \neq \emptyset} mv(Q) \\ &= (M - m) \sum_{Q \cap E \neq \emptyset} v(Q) < \frac{\varepsilon(M - m)}{1 + (M - m)} < \varepsilon. \end{aligned}$$

This shows  $f\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$ . Now also,

$$m\varepsilon \leq \int f\mathcal{X}_E dV \leq M\varepsilon$$

and since  $\varepsilon$  is arbitrary, this shows

$$\int_E f dV \equiv \int f\mathcal{X}_E dV = 0$$

Why is  $E$  contented? Let  $\mathcal{G}$  be a grid for which

$$\sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon$$

Then for this grid,

$$\mathcal{U}_{\mathcal{G}}(\mathcal{X}_E) - \mathcal{L}_{\mathcal{G}}(\mathcal{X}_E) \leq \sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon$$

and this proves the theorem. ■

**Corollary A.4.11** *If  $f\mathcal{X}_{E_i} \in \mathcal{R}(\mathbb{R}^n)$  for  $i = 1, 2, \dots, r$  and for all  $i \neq j$ ,  $E_i \cap E_j$  is either the empty set or a set of Jordan content 0, then letting  $F \equiv \cup_{i=1}^r E_i$ , it follows  $f\mathcal{X}_F \in \mathcal{R}(\mathbb{R}^n)$  and*

$$\int f\mathcal{X}_F dV \equiv \int_F f dV = \sum_{i=1}^r \int_{E_i} f dV.$$

**Proof:** This is true if  $r = 1$ . Suppose it is true for  $r$ . It will be shown that it is true for  $r + 1$ . Let  $F_r = \cup_{i=1}^r E_i$  and let  $F_{r+1}$  be defined similarly. By the induction hypothesis,  $f|_{F_r} \in \mathcal{R}(\mathbb{R}^n)$ . Also, since  $F_r$  is a finite union of the  $E_i$ , it follows that  $F_r \cap E_{r+1}$  is either empty or a set of Jordan content 0.

$$-f|_{F_r \cap E_{r+1}} + f|_{F_r} + f|_{E_{r+1}} = f|_{F_{r+1}}$$

and by Theorem A.4.10 each function on the left is in  $\mathcal{R}(\mathbb{R}^n)$  and the first one on the left has integral equal to zero. Therefore,

$$\int f|_{F_{r+1}} dV = \int f|_{F_r} dV + \int f|_{E_{r+1}} dV$$

which by induction equals

$$\sum_{i=1}^r \int_{E_i} f dV + \int_{E_{r+1}} f dV = \sum_{i=1}^{r+1} \int_{E_i} f dV$$

and this proves the corollary. ■

In particular, for

$$Q = \prod_{i=1}^n [a_i, b_i], \quad Q' = \prod_{i=1}^n (a_i, b_i]$$

both are contented sets and

$$\int \mathcal{X}_Q dV = \int_{Q'} \mathcal{X}_{Q'} dV = v(Q). \quad (1.16)$$

This is because

$$Q \setminus Q' = \cup_{i=1}^n a_i \times \prod_{j \neq i} (a_j, b_j]$$

a finite union of sets of content 0. It is obvious  $\int \mathcal{X}_Q dV = v(Q)$  because you can use a grid which has  $Q$  as one of the boxes and then the upper and lower sums are the same and equal to  $v(Q)$ . Therefore, the claim about the equality of the two integrals in (1.16) follows right away from Corollary A.4.11. That  $\mathcal{X}_{Q'}$  is integrable follows from

$$\mathcal{X}_{Q'} = \mathcal{X}_Q - \mathcal{X}_{Q \setminus Q'}$$

and each of the two functions on the right is integrable thanks to Theorem A.4.10.

In fact, here is an interesting version of the Riemann criterion which depends on these half open boxes.

**Lemma A.4.12** *Suppose  $f$  is a bounded function which equals zero off some bounded set. Then  $f \in \mathcal{R}(\mathbb{R}^n)$  if and only if for all  $\varepsilon > 0$  there exists a grid  $\mathcal{G}$  such that*

$$\sum_{Q \in \mathcal{G}} (M_Q(f) - m_Q(f)) v(Q) < \varepsilon. \quad (1.17)$$

**Proof:** Since  $Q' \subseteq Q$ ,

$$M_{Q'}(f) - m_{Q'}(f) \leq M_Q(f) - m_Q(f)$$



and therefore, the only if part of the equivalence is obvious.

Conversely, let  $\mathcal{G}$  be a grid such that (1.17) holds with  $\varepsilon$  replaced with  $\frac{\varepsilon}{2}$ . It is necessary to show that there is a grid such that (1.17) holds with no primes on the  $Q$ . Let  $\mathcal{F}$  be a refinement of  $\mathcal{G}$  obtained by adding the points  $\alpha_k^i + \eta_k$  where  $\eta_k \leq \eta$  and is also chosen so small that for each  $i = 1, \dots, n$ ,

$$\alpha_k^i + \eta_k < \alpha_{k+1}^i.$$

You only need to have  $\eta_k > 0$  for the finitely many boxes of  $\mathcal{G}$  which intersect the bounded set where  $f$  is not zero. Then for

$$Q \equiv \prod_{i=1}^n [\alpha_{k_i}^i, \alpha_{k_i+1}^i] \in \mathcal{G},$$

Let

$$\widehat{Q} \equiv \prod_{i=1}^n [\alpha_{k_i}^i + \eta_{k_i}, \alpha_{k_i+1}^i]$$

and denote by  $\widehat{\mathcal{G}}$  the collection of these smaller boxes. For each set  $Q$  in  $\mathcal{G}$  there is the smaller set  $\widehat{Q}$  along with  $n$  boxes,  $B_k, k = 1, \dots, n$ , one of whose sides is of length  $\eta_k$  and the remainder of whose sides are shorter than the diameter of  $Q$  such that the set  $Q$  is the union of  $\widehat{Q}$  and these sets  $B_k$ . Now suppose  $f$  equals zero off the ball  $B(\mathbf{0}, \frac{R}{2})$ . Then without loss of generality, you may assume the diameter of every box in  $\mathcal{G}$  which has nonempty intersection with  $B(\mathbf{0}, R)$  is smaller than  $\frac{R}{3}$ . (If this is not so, simply refine  $\mathcal{G}$  to make it so, such a refinement leaving (1.17) valid because refinements do not increase the difference between upper and lower sums in this context either.) Suppose there are  $P$  sets of  $\mathcal{G}$  contained in  $B(\mathbf{0}, R)$  (So these are the only sets of  $\mathcal{G}$  which could have nonempty intersection with the set where  $f$  is nonzero.) and suppose that for all  $\mathbf{x}$ ,  $|f(\mathbf{x})| < C/2$ . Then

$$\begin{aligned} \sum_{Q \in \mathcal{F}} (M_Q(f) - m_Q(f)) v(Q) &\leq \sum_{\widehat{Q} \in \widehat{\mathcal{G}}} (M_{\widehat{Q}}(f) - m_{\widehat{Q}}(f)) v(Q) \\ &+ \sum_{Q \in \mathcal{F} \setminus \widehat{\mathcal{G}}} (M_Q(f) - m_Q(f)) v(Q) \end{aligned}$$

The first term on the right of the inequality in the above is no larger than  $\varepsilon/2$  because  $M_{\widehat{Q}}(f) - m_{\widehat{Q}}(f) \leq M_{Q'}(f) - m_{Q'}(f)$  for each  $Q$ . Therefore, the above is dominated by

$$\leq \varepsilon/2 + CPnR^{n-1}\eta < \varepsilon$$

whenever  $\eta$  is small enough. Since  $\varepsilon$  is arbitrary,  $f \in \mathcal{R}(\mathbb{R}^n)$  as claimed. ■

## A.5 Iterated Integrals

To evaluate an  $n$  dimensional Riemannn integral, one uses iterated integrals. Formally, an iterated integral is defined as follows. For  $f$  a function defined on  $\mathbb{R}^{n+m}$ ,

$$\mathbf{y} \rightarrow f(\mathbf{x}, \mathbf{y})$$

is a function of  $\mathbf{y}$  for each  $\mathbf{x} \in \mathbb{R}^n$ . Therefore, it might be possible to integrate this function of  $\mathbf{y}$  and write

$$\int_{\mathbb{R}^m} f(\mathbf{x}, \mathbf{y}) dV_{\mathbf{y}}.$$

Now the result is clearly a function of  $\mathbf{x}$  and so, it might be possible to integrate this and write

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f(\mathbf{x}, \mathbf{y}) dV_{\mathbf{y}} dV_{\mathbf{x}}.$$

This symbol is called an iterated integral, because it involves the iteration of two lower dimensional integrations. Under what conditions are the two iterated integrals equal to the integral

$$\int_{\mathbb{R}^{n+m}} f(\mathbf{z}) dV?$$

**Definition A.5.1** Let  $\mathcal{G}$  be a grid on  $\mathbb{R}^{n+m}$  defined by the  $n+m$  sequences,

$$\{\alpha_k^i\}_{k=-\infty}^{\infty} \quad i = 1, \dots, n+m.$$

Let  $\mathcal{G}_n$  be the grid on  $\mathbb{R}^n$  obtained by considering only the first  $n$  of these sequences and let  $\mathcal{G}_m$  be the grid on  $\mathbb{R}^m$  obtained by considering only the last  $m$  of the sequences. Thus a typical box in  $\mathcal{G}_m$  would be

$$\prod_{i=n+1}^{n+m} [\alpha_{k_i}^i, \alpha_{k_i+1}^i], \quad k_i \geq n+1$$

and a box in  $\mathcal{G}_n$  would be of the form

$$\prod_{i=1}^n [\alpha_{k_i}^i, \alpha_{k_i+1}^i], \quad k_i \leq n.$$

**Lemma A.5.2** Let  $\mathcal{G}$ ,  $\mathcal{G}_n$ , and  $\mathcal{G}_m$  be the grids defined above. Then

$$\mathcal{G} = \{R \times P : R \in \mathcal{G}_n \text{ and } P \in \mathcal{G}_m\}.$$

**Proof:** If  $Q \in \mathcal{G}$ , then  $Q$  is clearly of this form. On the other hand, if  $R \times P$  is one of the sets described above, then from the above description of  $R$  and  $P$ , it follows  $R \times P$  is one of the sets of  $\mathcal{G}$ . ■

Now let  $\mathcal{G}$  be a grid on  $\mathbb{R}^{n+m}$  and suppose

$$\phi(\mathbf{z}) = \sum_{Q \in \mathcal{G}} \phi_Q \mathcal{X}_Q(\mathbf{z}) \quad (1.18)$$

where  $\phi_Q$  equals zero for all but finitely many  $Q$ . Thus  $\phi$  is a step function. Recall that for

$$Q = \prod_{i=1}^{n+m} [a_i, b_i], \quad Q' \equiv \prod_{i=1}^{n+m} (a_i, b_i]$$

The function

$$\phi = \sum_{Q \in \mathcal{G}} \phi_Q \mathcal{X}_Q$$

is integrable because it is a finite sum of integrable functions, each function in the sum being integrable because the set of discontinuities has Jordan content 0. (why?) Letting  $(\mathbf{x}, \mathbf{y}) = \mathbf{z}$ ,

$$\begin{aligned}\phi(\mathbf{z}) &= \phi(\mathbf{x}, \mathbf{y}) = \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{R' \times P'}(\mathbf{x}, \mathbf{y}) \\ &= \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{R'}(\mathbf{x}) \mathcal{X}_{P'}(\mathbf{y}).\end{aligned}\quad (1.19)$$

For a function of two variables  $h$ , denote by  $h(\cdot, \mathbf{y})$  the function  $\mathbf{x} \rightarrow h(\mathbf{x}, \mathbf{y})$  and  $h(\mathbf{x}, \cdot)$  the function  $\mathbf{y} \rightarrow h(\mathbf{x}, \mathbf{y})$ . The following lemma is a preliminary version of Fubini's theorem.

**Lemma A.5.3** *Let  $\phi$  be a step function as described in (1.18). Then*

$$\phi(\mathbf{x}, \cdot) \in \mathcal{R}(\mathbb{R}^m), \quad (1.20)$$

$$\int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y}) dV_{\mathbf{y}} \in \mathcal{R}(\mathbb{R}^n), \quad (1.21)$$

and

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \phi(\mathbf{x}, \mathbf{y}) dV_{\mathbf{y}} dV_{\mathbf{x}} = \int_{\mathbb{R}^{n+m}} \phi(\mathbf{z}) dV. \quad (1.22)$$

**Proof:** To verify (1.20), note that  $\phi(\mathbf{x}, \cdot)$  is the step function

$$\phi(\mathbf{x}, \mathbf{y}) = \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{P'}(\mathbf{y}).$$

Where  $\mathbf{x} \in R'$  and this is a finite sum of integrable functions because each has set of discontinuities with Jordan content 0. From the description in (1.19),

$$\begin{aligned}\int_{\mathbb{R}^m} \phi(\mathbf{x}, \mathbf{y}) dV_{\mathbf{y}} &= \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{R'}(\mathbf{x}) v(P) \\ &= \sum_{R \in \mathcal{G}_n} \left( \sum_{P \in \mathcal{G}_m} \phi_{R \times P} v(P) \right) \mathcal{X}_{R'}(\mathbf{x}),\end{aligned}\quad (1.23)$$

another step function. Therefore,

$$\begin{aligned}\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \phi(\mathbf{x}, \mathbf{y}) dV_{\mathbf{y}} dV_{\mathbf{x}} &= \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} v(P) v(R) \\ &= \sum_{Q \in \mathcal{G}} \phi_Q v(Q) = \int_{\mathbb{R}^{n+m}} \phi(\mathbf{z}) dV.\end{aligned}\quad \blacksquare$$

From (1.23),

$$\begin{aligned}M_{R'_1} \left( \int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y}) dV_{\mathbf{y}} \right) &\equiv \sup \left\{ \sum_{R \in \mathcal{G}_n} \left( \sum_{P \in \mathcal{G}_m} \phi_{R \times P} v(P) \right) \mathcal{X}_{R'}(\mathbf{x}) : \mathbf{x} \in R'_1 \right\} \\ &= \sum_{P \in \mathcal{G}_m} \phi_{R_1 \times P} v(P)\end{aligned}\quad (1.24)$$

because  $\int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y}) dV_{\mathbf{y}}$  has the constant value given in (1.24) for  $\mathbf{x} \in R'_1$ . Similarly,

$$\begin{aligned} m_{R'_1} \left( \int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y}) dV_{\mathbf{y}} \right) &\equiv \inf \left\{ \sum_{R \in \mathcal{G}_n} \left( \sum_{P \in \mathcal{G}_m} \phi_{R \times P} v(P) \right) \mathcal{X}_{R'}(\mathbf{x}) : \mathbf{x} \in R'_1 \right\} \\ &= \sum_{P \in \mathcal{G}_m} \phi_{R_1 \times P} v(P). \end{aligned} \quad (1.25)$$

**Theorem A.5.4 (Fubini)** Let  $f \in \mathcal{R}(\mathbb{R}^{n+m})$  and suppose also that  $f(\mathbf{x}, \cdot) \in \mathcal{R}(\mathbb{R}^m)$  for each  $\mathbf{x}$ . Then

$$\int_{\mathbb{R}^m} f(\cdot, \mathbf{y}) dV_{\mathbf{y}} \in \mathcal{R}(\mathbb{R}^n) \quad (1.26)$$

and

$$\int_{\mathbb{R}^{n+m}} f(\mathbf{z}) dV = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f(\mathbf{x}, \mathbf{y}) dV_{\mathbf{y}} dV_{\mathbf{x}}. \quad (1.27)$$

**Proof:** Let  $\mathcal{G}$  be a grid such that  $\mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) < \varepsilon$  and let  $\mathcal{G}_n$  and  $\mathcal{G}_m$  be as defined above. Let

$$\phi(\mathbf{z}) \equiv \sum_{Q \in \mathcal{G}} M_{Q'}(f) \mathcal{X}_{Q'}(\mathbf{z}), \quad \psi(\mathbf{z}) \equiv \sum_{Q \in \mathcal{G}} m_{Q'}(f) \mathcal{X}_{Q'}(\mathbf{z}).$$

Observe that  $M_{Q'}(f) \leq M_Q(f)$  and  $m_{Q'}(f) \geq m_Q(f)$ . Then

$$\mathcal{U}_{\mathcal{G}}(f) \geq \int \phi dV, \quad \mathcal{L}_{\mathcal{G}}(f) \leq \int \psi dV.$$

Also  $f(\mathbf{z}) \in (\psi(\mathbf{z}), \phi(\mathbf{z}))$  for all  $\mathbf{z}$ . Thus from (1.24),

$$M_{R'} \left( \int_{\mathbb{R}^m} f(\cdot, \mathbf{y}) dV_{\mathbf{y}} \right) \leq M_{R'} \left( \int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y}) dV_{\mathbf{y}} \right) = \sum_{P \in \mathcal{G}_m} M_{R' \times P'}(f) v(P)$$

and from (1.25),

$$m_{R'} \left( \int_{\mathbb{R}^m} f(\cdot, \mathbf{y}) dV_{\mathbf{y}} \right) \geq m_{R'} \left( \int_{\mathbb{R}^m} \psi(\cdot, \mathbf{y}) dV_{\mathbf{y}} \right) = \sum_{P \in \mathcal{G}_m} m_{R' \times P'}(f) v(P).$$

Therefore,

$$\begin{aligned} \sum_{R \in \mathcal{G}_n} \left[ M_{R'} \left( \int_{\mathbb{R}^m} f(\cdot, \mathbf{y}) dV_{\mathbf{y}} \right) - m_{R'} \left( \int_{\mathbb{R}^m} f(\cdot, \mathbf{y}) dV_{\mathbf{y}} \right) \right] v(R) &\leq \\ \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} [M_{R' \times P'}(f) - m_{R' \times P'}(f)] v(P) v(R) &\leq \mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) < \varepsilon. \end{aligned}$$

This shows, from Lemma A.4.12 and the Riemannn criterion, that  $\int_{\mathbb{R}^m} f(\cdot, \mathbf{y}) dV_{\mathbf{y}} \in \mathcal{R}(\mathbb{R}^n)$ . It remains to verify (1.27). First note

$$\int_{\mathbb{R}^{n+m}} f(\mathbf{z}) dV \in [\mathcal{L}_{\mathcal{G}}(f), \mathcal{U}_{\mathcal{G}}(f)].$$

Next,

$$\mathcal{L}_{\mathcal{G}}(f) \leq \int_{\mathbb{R}^{n+m}} \psi dV = \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \psi dV_{\mathbf{y}} dV_{\mathbf{x}} \leq \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f(\mathbf{x}, \mathbf{y}) dV_{\mathbf{y}} dV_{\mathbf{x}}$$

$$\leq \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \phi(\mathbf{x}, \mathbf{y}) dV_y dV_x = \int_{\mathbb{R}^{n+m}} \phi dV \leq \mathcal{U}_g(f).$$

Therefore,

$$\left| \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f(\mathbf{x}, \mathbf{y}) dV_y dV_x - \int_{\mathbb{R}^{n+m}} f(z) dV \right| \leq \varepsilon$$

and since  $\varepsilon > 0$  is arbitrary, this proves Fubini's theorem<sup>2</sup>. ■

**Corollary A.5.5** Suppose  $E$  is a bounded contented set in  $\mathbb{R}^n$  and let  $\phi, \psi$  be continuous functions defined on  $E$  such that  $\phi(\mathbf{x}) \geq \psi(\mathbf{x})$ . Also suppose  $f$  is a continuous bounded function defined on the set

$$P \equiv \{(\mathbf{x}, y) : \psi(\mathbf{x}) \leq y \leq \phi(\mathbf{x})\},$$

It follows  $f \mathcal{X}_P \in \mathcal{R}(\mathbb{R}^{n+1})$  and

$$\int_P f dV = \int_E \int_{\psi(\mathbf{x})}^{\phi(\mathbf{x})} f(\mathbf{x}, y) dy dV_x.$$

**Proof:** Since  $f$  is continuous, there is no problem in writing  $f(\mathbf{x}, \cdot) \mathcal{X}_{[\psi(\mathbf{x}), \phi(\mathbf{x})]}(\cdot) \in \mathcal{R}(\mathbb{R}^1)$ . Also,  $f \mathcal{X}_P \in \mathcal{R}(\mathbb{R}^{n+1})$  because  $P$  is contented thanks to Corollary A.4.9. Therefore, by Fubini's theorem

$$\begin{aligned} \int_P f dV &= \int_{\mathbb{R}^n} \int_{\mathbb{R}} f \mathcal{X}_P dy dV_x \\ &= \int_E \int_{\psi(\mathbf{x})}^{\phi(\mathbf{x})} f(\mathbf{x}, y) dy dV_x \end{aligned}$$

proving the corollary. ■

Other versions of this corollary are immediate and should be obvious whenever encountered.

## A.6 The Change Of Variables Formula

First recall Theorem 26.3.2 on Page 492 which is listed here for convenience.

**Theorem A.6.1** Let  $\mathbf{h} : U \rightarrow \mathbb{R}^n$  be a  $C^1$  function with  $\mathbf{h}(\mathbf{0}) = \mathbf{0}, D\mathbf{h}(\mathbf{0})^{-1}$  exists. Then there exists an open set  $V \subseteq U$  containing  $\mathbf{0}$  flips,  $\mathbf{F}_1, \dots, \mathbf{F}_{n-1}$ , and primitive functions  $\mathbf{G}_n, \mathbf{G}_{n-1}, \dots, \mathbf{G}_1$  such that for  $\mathbf{x} \in V$ ,

$$\mathbf{h}(\mathbf{x}) = \mathbf{F}_1 \circ \dots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \mathbf{G}_{n-1} \circ \dots \circ \mathbf{G}_1(\mathbf{x}).$$

Also recall Theorem 14.6.5 on Page 272.

**Theorem A.6.2** Let  $\phi : [a, b] \rightarrow [c, d]$  be one to one and suppose  $\phi'$  exists and is continuous on  $[a, b]$ . Then if  $f$  is a continuous function defined on  $[a, b]$ ,

$$\int_c^d f(s) ds = \int_a^b f(\phi(t)) |\phi'(t)| dt$$

<sup>2</sup>Actually, Fubini's theorem usually refers to a much more profound result in the theory of Lebesgue integration.

The following is a simple corollary to this theorem.

**Corollary A.6.3** *Let  $\phi : [a, b] \rightarrow [c, d]$  be one to one and suppose  $\phi'$  exists and is continuous on  $[a, b]$ . Then if  $f$  is a continuous function defined on  $[a, b]$ ,*

$$\int_{\mathbb{R}} \mathcal{X}_{[a,b]}(\phi^{-1}(x)) f(x) dx = \int_{\mathbb{R}} \mathcal{X}_{[a,b]}(t) f(\phi(t)) |\phi'(t)| dt$$

**Lemma A.6.4** *Let  $\mathbf{h} : V \rightarrow \mathbb{R}^n$  be a  $C^1$  function and suppose  $H$  is a compact subset of  $V$ . Then there exists a constant  $C$  independent of  $\mathbf{x} \in H$  such that*

$$|D\mathbf{h}(\mathbf{x})\mathbf{v}| \leq C|\mathbf{v}|.$$

**Proof:** Consider the compact set  $H \times \partial B(\mathbf{0}, 1) \subseteq \mathbb{R}^{2n}$ . Let  $f : H \times \partial B(\mathbf{0}, 1) \rightarrow \mathbb{R}$  be given by  $f(\mathbf{x}, \mathbf{v}) = |D\mathbf{h}(\mathbf{x})\mathbf{v}|$ . Then let  $C$  denote the maximum value of  $f$ . It follows that for  $\mathbf{v} \in \mathbb{R}^n$ ,

$$\left| D\mathbf{h}(\mathbf{x}) \frac{\mathbf{v}}{|\mathbf{v}|} \right| \leq C$$

and so the desired formula follows when you multiply both sides by  $|\mathbf{v}|$ . ■

**Definition A.6.5** *Let  $A$  be an open set. Write  $C^k(A; \mathbb{R}^n)$  to denote a  $C^k$  function whose domain is  $A$  and whose range is in  $\mathbb{R}^n$ . Let  $U$  be an open set in  $\mathbb{R}^n$ . Then  $\mathbf{h} \in C^k(\overline{U}; \mathbb{R}^n)$  if there exists an open set  $V \supseteq \overline{U}$  and a function  $\mathbf{g} \in C^1(V; \mathbb{R}^n)$  such that  $\mathbf{g} = \mathbf{h}$  on  $\overline{U}$ .  $f \in C^k(\overline{U})$  means the same thing except that  $f$  has values in  $\mathbb{R}$ . Also recall that  $\mathbf{x} \in \partial U$  means that every open set which contains  $\mathbf{x}$  contains points of  $U$  and points of  $U^C$*

**Theorem A.6.6** *Let  $U$  be a bounded open set such that  $\partial U$  has zero content and let  $\mathbf{h} \in C(\overline{U}; \mathbb{R}^n)$  be one to one and  $D\mathbf{h}(\mathbf{x})^{-1}$  exists for all  $\mathbf{x} \in U$ . Then  $\mathbf{h}(\partial U) = \partial(\mathbf{h}(U))$  and  $\partial(\mathbf{h}(U))$  has zero content.*

**Proof:** Let  $\mathbf{x} \in \partial U$  and let  $\mathbf{g} = \mathbf{h}$  where  $\mathbf{g}$  is a  $C^1$  function defined on an open set containing  $\overline{U}$ . By the inverse function theorem,  $\mathbf{g}$  is locally one to one and an open mapping near  $\mathbf{x}$ . Thus  $\mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{x})$  and is in an open set containing points of  $\mathbf{g}(U)$  and points of  $\mathbf{g}(U^C)$ . These points of  $\mathbf{g}(U^C)$  cannot equal any points of  $\mathbf{h}(U)$  because  $\mathbf{g}$  is one to one locally. Thus  $\mathbf{h}(\mathbf{x}) \in \partial(\mathbf{h}(U))$  and so  $\mathbf{h}(\partial U) \subseteq \partial(\mathbf{h}(U))$ . Now suppose  $\mathbf{y} \in \partial(\mathbf{h}(U))$ . By the inverse function theorem  $\mathbf{y}$  cannot be in the open set  $\mathbf{h}(U)$ . Since  $\mathbf{y} \in \partial(\mathbf{h}(U))$ , every ball centered at  $\mathbf{y}$  contains points of  $\mathbf{h}(U)$  and so  $\mathbf{y} \in \overline{\mathbf{h}(U)} \setminus \mathbf{h}(U)$ . Thus there exists a sequence,  $\{\mathbf{x}_n\} \subseteq U$  such that  $\mathbf{h}(\mathbf{x}_n) \rightarrow \mathbf{y}$ . But then, by the continuity of  $\mathbf{h}^{-1}$  which comes from the inverse function theorem,  $\mathbf{x}_n \rightarrow \mathbf{h}^{-1}(\mathbf{y})$  and so  $\mathbf{h}^{-1}(\mathbf{y}) \notin U$  but is in  $\overline{U}$ . Thus  $\mathbf{h}^{-1}(\mathbf{y}) \in \partial U$ . (Why?) Therefore,  $\mathbf{y} \in \mathbf{h}(\partial U)$ , and this proves the two sets are equal. It remains to verify the claim about content.

First let  $H$  denote a compact set whose interior contains  $\overline{U}$  which is also in the interior of the domain of  $\mathbf{g}$ . Now since  $\partial U$  has content zero, it follows that for  $\varepsilon > 0$  given, there exists a grid  $\mathcal{G}$  such that if  $\mathcal{G}'$  are those boxes of  $\mathcal{G}$  which have nonempty intersection with  $\partial U$ , then

$$\sum_{Q \in \mathcal{G}'} v(Q) < \varepsilon$$

and by refining the grid if necessary, no box of  $\mathcal{G}$  has nonempty intersection with both  $\overline{U}$  and  $H^C$ . Refining this grid still more, you can also assume that for all boxes in  $\mathcal{G}'$ ,

$$\frac{l_i}{l_j} < 2$$

where  $l_i$  is the length of the  $i^{\text{th}}$  side. (Thus the boxes are not too far from being cubes.)

Let  $C$  be the constant of Lemma A.6.4 applied to  $g$  on  $H$ .

Now consider one of these boxes,  $Q \in \mathcal{G}'$ . If  $\mathbf{x}, \mathbf{y} \in Q$ , it follows from the chain rule that

$$g(\mathbf{y}) - g(\mathbf{x}) = \int_0^1 Dg(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) dt$$

By Lemma A.6.4 applied to  $H$

$$\begin{aligned} |g(\mathbf{y}) - g(\mathbf{x})| &\leq \int_0^1 |Dg(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})| dt \\ &\leq C \int_0^1 |\mathbf{x} - \mathbf{y}| dt \leq C \operatorname{diam}(Q) \\ &= C \left( \sum_{i=1}^n l_i^2 \right)^{1/2} \leq C\sqrt{n}L \end{aligned}$$

where  $L$  is the length of the longest side of  $Q$ . Thus  $\operatorname{diam}(g(Q)) \leq C\sqrt{n}L$  and so  $g(Q)$  is contained in a cube having sides equal to  $C\sqrt{n}L$  and volume equal to

$$C^n n^{n/2} L^n \leq C^n n^{n/2} 2^n l_1 l_2 \cdots l_n = C^n n^{n/2} 2^n v(Q).$$

Denoting by  $P_Q$  this cube, it follows

$$h(\partial U) \subseteq \bigcup_{Q \in \mathcal{G}'} v(P_Q)$$

and

$$\sum_{Q \in \mathcal{G}'} v(P_Q) \leq C^n n^{n/2} 2^n \sum_{Q \in \mathcal{G}'} v(Q) < \varepsilon C^n n^{n/2} 2^n.$$

Since  $\varepsilon > 0$  is arbitrary, this shows  $h(\partial U)$  has content zero as claimed. ■

**Theorem A.6.7** Suppose  $f \in C(\overline{U})$  where  $U$  is a bounded open set with  $\partial U$  having content 0. Then  $f\mathcal{X}_U \in \mathcal{R}(\mathbb{R}^n)$ .

**Proof:** Let  $H$  be a compact set whose interior contains  $\overline{U}$  which is also contained in the domain of  $g$  where  $g$  is a continuous functions whose restriction to  $U$  equals  $f$ . Consider  $g\mathcal{X}_U$ , a function whose set of discontinuities has content 0. Then  $g\mathcal{X}_U = f\mathcal{X}_U \in \mathcal{R}(\mathbb{R}^n)$  as claimed. This is by the big theorem which tells which functions are Riemannn integrable. ■

The symbol  $U - \mathbf{p}$  is defined as  $\{\mathbf{x} - \mathbf{p} : \mathbf{x} \in U\}$ . It merely slides  $U$  by the vector  $\mathbf{p}$ . The following lemma is obvious from the definition of the integral.

**Lemma A.6.8** Let  $U$  be a bounded open set and let  $f\mathcal{X}_U \in \mathcal{R}(\mathbb{R}^n)$ . Then

$$\int f(\mathbf{x} + \mathbf{p}) \mathcal{X}_{U - \mathbf{p}}(\mathbf{x}) d\mathbf{x} = \int f(\mathbf{x}) \mathcal{X}_U(\mathbf{x}) d\mathbf{x}$$

A few more lemmas are needed.

**Lemma A.6.9** Let  $S$  be a nonempty subset of  $\mathbb{R}^n$ . Define

$$f(\mathbf{x}) \equiv \operatorname{dist}(\mathbf{x}, S) \equiv \inf\{|\mathbf{x} - \mathbf{y}| : \mathbf{y} \in S\}.$$

Then  $f$  is continuous.

**Proof:** Consider  $|f(x) - f(x_1)|$  and suppose without loss of generality that  $f(x_1) \geq f(x)$ . Then choose  $y \in S$  such that  $f(x) + \varepsilon > |x - y|$ . Then

$$\begin{aligned} |f(x_1) - f(x)| &= f(x_1) - f(x) \leq f(x_1) - |x - y| + \varepsilon \\ &\leq |x_1 - y| - |x - y| + \varepsilon \\ &\leq |x - x_1| + |x - y| - |x - y| + \varepsilon \\ &= |x - x_1| + \varepsilon. \end{aligned}$$

Since  $\varepsilon$  is arbitrary, it follows that  $|f(x_1) - f(x)| \leq |x - x_1|$  and this proves the lemma. ■

**Theorem A.6.10** (Urysohn's lemma for  $\mathbb{R}^n$ ) Let  $H$  be a closed subset of an open set  $U$ . Then there exists a continuous function  $g : \mathbb{R}^n \rightarrow [0, 1]$  such that  $g(x) = 1$  for all  $x \in H$  and  $g(x) = 0$  for all  $x \notin U$ .

**Proof:** If  $x \notin C$ , a closed set, then  $\text{dist}(x, C) > 0$  because there exists  $\delta > 0$  such that  $B(x, \delta) \cap C = \emptyset$ . This is because, since  $C$  is closed, its complement is open. Therefore,  $\text{dist}(x, H) + \text{dist}(x, U^C) > 0$  for all  $x \in \mathbb{R}^n$ . Now define a continuous function  $g$  as

$$g(x) \equiv \frac{\text{dist}(x, U^C)}{\text{dist}(x, H) + \text{dist}(x, U^C)}.$$

It is easy to see this verifies the conclusions of the theorem and this proves the theorem. ■

**Definition A.6.11** Define  $\text{spt}(f)$  (support of  $f$ ) to be the closure of the set  $\{x : f(x) \neq 0\}$ . If  $V$  is an open set,  $C_c(V)$  will be the set of continuous functions  $f$ , defined on  $\mathbb{R}^n$  having  $\text{spt}(f) \subseteq V$ .

**Definition A.6.12** If  $K$  is a compact subset of an open set  $V$ , then  $K \prec \phi \prec V$  if

$$\phi \in C_c(V), \phi(K) = \{1\}, \phi(\mathbb{R}^n) \subseteq [0, 1].$$

Also for  $\phi \in C_c(\mathbb{R}^n)$ ,  $K \prec \phi$  if

$$\phi(\mathbb{R}^n) \subseteq [0, 1] \text{ and } \phi(K) = 1.$$

and  $\phi \prec V$  if

$$\phi(\mathbb{R}^n) \subseteq [0, 1] \text{ and } \text{spt}(\phi) \subseteq V.$$

**Theorem A.6.13** (Partition of unity) Let  $K$  be a compact subset of  $\mathbb{R}^n$  and suppose

$$K \subseteq V = \bigcup_{i=1}^n V_i, \text{ } V_i \text{ open and bounded.}$$

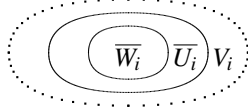
Then there exist  $\psi_i \prec V_i$  with

$$\sum_{i=1}^n \psi_i(x) = 1$$

for all  $x \in K$ .



**Proof:** Let  $K_1 = K \setminus \bigcup_{i=2}^n V_i$ . Thus  $K_1$  is compact because it is the intersection of a closed set with a compact set and  $K_1 \subseteq V_1$ . Let  $K_1 \subseteq W_1 \subseteq \bar{W}_1 \subseteq V_1$  with  $\bar{W}_1$  compact. To obtain  $W_1$ , use Theorem A.6.10 to get  $f$  such that  $K_1 \prec f \prec V_1$  and let  $W_1 = \{x : f(x) \neq 0\}$ . Thus  $W_1, V_2, \dots, V_n$  covers  $K$  and  $\bar{W}_1 \subseteq V_1$ . Let  $K_2 = K \setminus (\bigcup_{i=3}^n V_i \cup W_1)$ . Then  $K_2$  is compact and  $K_2 \subseteq V_2$ . Let  $K_2 \subseteq W_2 \subseteq \bar{W}_2 \subseteq V_2$  with  $\bar{W}_2$  compact. Continue this way finally obtaining  $W_1, \dots, W_n, K \subseteq W_1 \cup \dots \cup W_n$ , and  $\bar{W}_i \subseteq V_i; \bar{W}_i$  compact. Now let  $\bar{W}_i \subseteq U_i \subseteq \bar{U}_i \subseteq V_i, \bar{U}_i$  compact.



By Theorem A.6.10, there exist functions  $\phi_i, \gamma$  such that  $\bar{U}_i \prec \phi_i \prec V_i, \bigcup_{i=1}^n \bar{W}_i \prec \gamma \prec \bigcup_{i=1}^n U_i$ . Define

$$\psi_i(x) = \begin{cases} \gamma(x)\phi_i(x)/\sum_{j=1}^n \phi_j(x) & \text{if } \sum_{j=1}^n \phi_j(x) \neq 0, \\ 0 & \text{if } \sum_{j=1}^n \phi_j(x) = 0. \end{cases}$$

If  $x$  is such that  $\sum_{j=1}^n \phi_j(x) = 0$ , then  $x \notin \bigcup_{i=1}^n \bar{U}_i$ . Consequently  $\gamma(y) = 0$  for all  $y$  near  $x$  and so  $\psi_i(y) = 0$  for all  $y$  near  $x$ . Hence  $\psi_i$  is continuous at such  $x$ . If  $\sum_{j=1}^n \phi_j(x) \neq 0$ , this situation persists near  $x$  and so  $\psi_i$  is continuous at such points. Therefore  $\psi_i$  is continuous. If  $x \in K$ , then  $\gamma(x) = 1$  and so  $\sum_{j=1}^n \psi_j(x) = 1$ . Clearly  $0 \leq \psi_i(x) \leq 1$  and  $\text{spt}(\psi_j) \subseteq V_j$ . ■

The next lemma contains the main ideas. See [35] and [31] for similar proofs.

**Lemma A.6.14** *Let  $U$  be a bounded open set with  $\partial U$  having content 0. Also let  $h \in C^1(\bar{U}; \mathbb{R}^n)$  be one to one on  $U$  with  $Dh(x)^{-1}$  exists for all  $x \in U$ . Let  $f \in C(\bar{U})$  be nonnegative. Then*

$$\int \mathcal{H}_{h(U)}(z) f(z) dV_n = \int \mathcal{H}_U(x) f(h(x)) |\det Dh(x)| dV_n$$

**Proof:** Let  $\varepsilon > 0$  be given. Then by Theorem A.6.7,

$$x \rightarrow \mathcal{H}_U(x) f(h(x)) |\det Dh(x)|$$

is Riemann integrable. Therefore, there exists a grid  $\mathcal{G}$  such that, letting

$$g(x) = \mathcal{H}_U(x) f(h(x)) |\det Dh(x)|,$$

$$\mathcal{L}_{\mathcal{G}}(g) + \varepsilon > \mathcal{U}_{\mathcal{G}}(g).$$

Let  $K$  denote the union of the boxes  $Q$  of  $\mathcal{G}$  which intersect  $\bar{U}$ . Thus  $K$  is a compact subset of  $V$  where  $V$  is a bounded open set containing  $\bar{U}$ , and it is only the terms from these boxes which contribute anything nonzero to the lower sum. By Theorem 26.3.2 on Page 492 which is stated above and the inverse function theorem, it follows that for  $p \in K$ , there exists an open set contained in  $U$  which contains  $p$ , denoted as  $O_p$  such that for  $x \in O_p - p$ ,

$$h(x+p) - h(p) = F_1 \circ \dots \circ F_{n-1} \circ G_n \circ \dots \circ G_1(x)$$

where the  $G_i$  are primitive functions, and the  $F_j$  are flips. Also  $h(O_j)$  is an open set.

Finitely many of these open sets  $\{O_j\}_{j=1}^q$  cover  $K$ . Let the distinguished point for  $O_j$  be denoted by  $p_j$ . Now refine  $\mathcal{G}$  if necessary, such that the diameter of every cell of the new  $\mathcal{G}$  which intersects  $\bar{U}$  is smaller than a Lebesgue number for this open cover. Denote by  $\mathcal{G}'$  those boxes of the new  $\mathcal{G}$  which intersect  $\bar{U}$ . Thus the union of these boxes of  $\mathcal{G}'$  equals the set  $K$  and every box of  $\mathcal{G}'$  is contained in one of these  $O_j$ . By Theorem A.6.13, there exists a partition of unity  $\{\psi_j\}$  on  $h(K)$  such that  $\psi_j \prec h(O_j)$ . Then

$$\begin{aligned} \mathcal{L}_{\mathcal{G}}(g) &\leq \sum_{Q \in \mathcal{G}'} \int \mathcal{X}_Q(x) f(h(x)) |\det Dh(x)| dx \\ &= \sum_{Q \in \mathcal{G}'} \sum_{j=1}^q \int \mathcal{X}_Q(x) (\psi_j f)(h(x)) |\det Dh(x)| dx. \end{aligned} \quad (1.28)$$

Consider the term  $\int \mathcal{X}_Q(x) (\psi_j f)(h(x)) |\det Dh(x)| dx$ . By Lemma A.6.8 and Fubini's theorem this equals

$$\begin{aligned} &\int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \mathcal{X}_{Q-p_j}(x) (\psi_j f)(h(p_i) + F_1 \circ \dots \circ F_{n-1} \circ G_n \circ \dots \circ G_1(x)) \cdot \\ &|DF(G_n \circ \dots \circ G_1(x))| |DG_n(G_{n-1} \circ \dots \circ G_1(x))| \cdot \\ &|DG_{n-1}(G_{n-2} \circ \dots \circ G_1(x))| \cdot \end{aligned} \quad (1.29)$$

$$\dots |DG_2(G_1(x))| |DG_1(x)| dx_1 dV_{n-1}. \quad (1.30)$$

The vertical lines in the above signify the absolute value of the determinant of the matrix on the inside. Here  $dV_{n-1}$  is with respect to the variables  $x_2, \dots, x_n$ . Also  $F$  denotes  $F_1 \circ \dots \circ F_{n-1}$ . Now

$$G_1(x) = (\alpha(x), x_2, \dots, x_n)^T$$

and is one to one. Therefore, fixing  $x_2, \dots, x_n$ ,  $x_1 \rightarrow \alpha(x)$  is one to one. Also

$$|DG_1(x)| = |\alpha_{x_1}(x)|$$

Fixing  $x_2, \dots, x_n$ , change the variable,

$$y_1 = \alpha(x_1, x_2, \dots, x_n), \quad dy_1 = \alpha_{x_1}(x_1, x_2, \dots, x_n) dx_1$$

Thus

$$x = (x_1, x_2, \dots, x_n)^T = G_1^{-1}(y_1, x_2, \dots, x_n) \equiv G_1^{-1}(x')$$

Then in (1.30) you can use Corollary A.6.3 to write (1.30) as

$$\begin{aligned} &\int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \mathcal{X}_{Q-p_j}(G_1^{-1}(x')) (\psi_j f) \\ &(h(p_i) + F_1 \circ \dots \circ F_{n-1} \circ G_n \circ \dots \circ G_1(G_1^{-1}(x'))) \\ &\cdot |DF(G_n \circ \dots \circ G_1(G_1^{-1}(x')))| |DG_n(G_{n-1} \circ \dots \circ G_1(G_1^{-1}(x')))| \cdot \\ &|DG_{n-1}(G_{n-2} \circ \dots \circ G_1(G_1^{-1}(x')))| \dots |DG_2(G_1(G_1^{-1}(x')))| dy_1 dV_{n-1} \end{aligned}$$

which reduces to

$$\int_{\mathbb{R}^n} \mathcal{X}_{Q-p_j}(G_1^{-1}(x')) (\psi_j f)(h(p_i) + F_1 \circ \dots \circ F_{n-1} \circ G_n \circ \dots \circ G_2(x'))$$

$$\cdot |DF(G_n \circ \cdots \circ G_2(x'))| |DG_n(G_{n-1} \circ \cdots \circ G_2(x'))| \cdot \\ |DG_{n-1}(G_{n-2} \circ \cdots \circ G_2(x'))| \cdots |DG_2(x')| dV_n.$$

Now use Fubini's theorem again to make the inside integral taken with respect to  $x_2$ . Note that the term  $|DG_1(x)|$  disappeared. Exactly the same process yields

$$\int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \mathcal{X}_{Q-p_j}(G_1^{-1} \circ G_2^{-1}(x'')) (\psi_j f) \\ (h(p_i) + F_1 \circ \cdots \circ F_{n-1} \circ G_n \circ \cdots \circ G_3(x'')) \\ \cdot |DF(G_n \circ \cdots \circ G_3(x''))| |DG_n(G_{n-1} \circ \cdots \circ G_3(x''))| \cdot \\ |DG_{n-1}(G_{n-2} \circ \cdots \circ G_3(x''))| \cdots dy_2 dV_{n-1}.$$

Now  $F$  is just a composition of flips, so  $|DF(G_n \circ \cdots \circ G_3(x''))| = 1$ , and so this term can be replaced with 1. Continuing this process, eventually yields an expression of the form

$$\int_{\mathbb{R}^n} \mathcal{X}_{Q-p_j}(G_1^{-1} \circ \cdots \circ G_{n-2}^{-1} \circ G_{n-1}^{-1} \circ G_n^{-1} \circ F^{-1}(y)) (\psi_j f) (h(p_i) + y) dV_n. \quad (1.31)$$

Denoting by  $G^{-1}$  the expression,  $G_1^{-1} \circ \cdots \circ G_{n-2}^{-1} \circ G_{n-1}^{-1} \circ G_n^{-1}$ ,

$$\mathcal{X}_{Q-p_j}(G^{-1} \circ \cdots \circ G_{n-2}^{-1} \circ G_{n-1}^{-1} \circ G_n^{-1} \circ F^{-1}(y)) = 1$$

exactly when  $G^{-1} \circ F^{-1}(y) \in Q - p_j$ . Now recall that

$$h(p_j + x) - h(p_j) = F \circ G(x)$$

and so the above holds exactly when

$$y = h(p_j + G^{-1} \circ F^{-1}(y)) - h(p_j) \in h(p_j + Q - p_j) - h(p_j) \\ = h(Q) - h(p_j).$$

Thus (1.31) reduces to

$$\int_{\mathbb{R}^n} \mathcal{X}_{h(Q)-h(p_j)}(y) (\psi_j f) (h(p_i) + y) dV_n \\ = \int_{\mathbb{R}^n} \mathcal{X}_{h(Q)}(z) (\psi_j f) (z) dV_n.$$

It follows from (1.28),

$$\begin{aligned} \mathcal{U}_g(g) - \varepsilon &\leq \mathcal{L}_g(g) \leq \int \mathcal{X}_U(x) f(h(x)) |\det Dh(x)| dV_n \\ &\leq \sum_{Q \in \mathcal{G}'} \int \mathcal{X}_Q(x) f(h(x)) |\det Dh(x)| dx \\ &= \sum_{Q \in \mathcal{G}'} \sum_{j=1}^q \int \mathcal{X}_Q(x) (\psi_j f) (h(x)) |\det Dh(x)| dx \\ &= \sum_{Q \in \mathcal{G}'} \sum_{j=1}^q \int_{\mathbb{R}^n} \mathcal{X}_{h(Q)}(z) (\psi_j f) (z) dV_n \\ &= \sum_{Q \in \mathcal{G}'} \int_{\mathbb{R}^n} \mathcal{X}_{h(Q)}(z) f(z) dV_n = \int \mathcal{X}_{h(U)}(z) f(z) dV_n \end{aligned}$$

which implies the inequality,

$$\int \mathcal{X}_U(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| dV_n \leq \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) dV_n$$

But now you can use the same information just derived to obtain equality.

$$\mathbf{x} = \mathbf{h}^{-1}(\mathbf{z})$$

and so from what was just done,

$$\begin{aligned} & \int \mathcal{X}_U(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| dV_n \\ &= \int \mathcal{X}_{\mathbf{h}^{-1}(\mathbf{h}(U))}(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| dV_n \\ &\geq \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) |\det D\mathbf{h}(\mathbf{h}^{-1}(\mathbf{z}))| |\det D\mathbf{h}^{-1}(\mathbf{z})| dV_n \\ &= \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) dV_n \end{aligned}$$

from the chain rule. In fact,

$$I = D\mathbf{h}(\mathbf{h}^{-1}(\mathbf{z})) D\mathbf{h}^{-1}(\mathbf{z}),$$

so

$$1 = |\det D\mathbf{h}(\mathbf{h}^{-1}(\mathbf{z}))| |\det D\mathbf{h}^{-1}(\mathbf{z})|. \quad \blacksquare$$

The change of variables theorem follows.

**Theorem A.6.15** *Let  $U$  be a bounded open set with  $\partial U$  having content 0. Also let  $\mathbf{h} \in C^1(\overline{U}; \mathbb{R}^n)$  be one to one on  $U$  and  $D\mathbf{h}(\mathbf{x})^{-1}$  exists for all  $\mathbf{x} \in U$ . Let  $f \in C(\overline{U})$ . Then*

$$\int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} = \int \mathcal{X}_U(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| d\mathbf{x}$$

**Proof:** You note that the formula holds for  $f^+ \equiv \frac{|f|+f}{2}$  and  $f^- \equiv \frac{|f|-f}{2}$ . Now  $f = f^+ - f^-$  and so

$$\begin{aligned} & \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \\ &= \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f^+(\mathbf{z}) d\mathbf{z} - \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f^-(\mathbf{z}) d\mathbf{z} \\ &= \int \mathcal{X}_U(\mathbf{x}) f^+(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| d\mathbf{x} - \int \mathcal{X}_U(\mathbf{x}) f^-(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| d\mathbf{x} \\ &= \int \mathcal{X}_U(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| d\mathbf{x}. \quad \blacksquare \end{aligned}$$

## A.7 Some Observations

Some of the above material is very technical. This is because it gives complete answers to the fundamental questions on existence of the integral and related theoretical considerations. However, most of the difficulties are artifacts. They should not even be considered! It was realized early in the twentieth century that these difficulties occur because, from the point of view of mathematics, this is not the right way to define an integral! Better results are obtained much more easily using the Lebesgue integral. Many of the technicalities related to Jordan content disappear almost magically when the right integral is used. However, the Lebesgue integral is more abstract than the Riemannnn integral and it is not traditional to consider it in a beginning calculus course. If you are interested in the fundamental properties of the integral and the theory behind it, you should abandon the Riemannnn integral which is an antiquated relic and begin to study the integral of the last century. An introduction to it is in [31]. Another very good source is [16]. This advanced calculus text does everything in terms of the Lebesgue integral and never bothers to struggle with the inferior Riemannnn integral. A more general treatment is found in [26], [27], [32], and [28]. There is also a still more general integral called the generalized Riemannnn integral. A recent book on this subject is [5]. It is far easier to define than the Lebesgue integral but the convergence theorems are much harder to prove. An introduction is also in [27].



## Appendix B

# A Rigid Body Rotating About a Point

Imagine a rigid body which is rotating about a point fixed in space. For example, you could consider a bicycle wheel rotating about its axis which is held still. More generally, we let the point about which the body rotates move also. In this case, the point is usually the center of mass of the body. However, in this section, this point will be regarded as fixed. Let  $B(t)$  denote the set of points in three dimensional space which the body occupies at time  $t$ . We will refer to the points in three dimensional space occupied by the body at time  $t = 0$  as the material points of the body.

Recall Theorem 24.3.2 about the existence of the angular velocity vector. The idea is that you have a material point  $\mathbf{x}_0$  in the body and some right handed orthonormal system of basis vectors  $\{\mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t)\}$  which moves with the body such if  $\mathbf{x}(t, \mathbf{x})$  is the vector from  $\mathbf{x}_0$  to the point where  $\mathbf{x}$  is at time  $t$ , then  $\mathbf{x}(t, \mathbf{x}) = a\mathbf{e}_1(t) + b\mathbf{e}_2(t) + c\mathbf{e}_3(t)$  where  $a, b, c$  are constants. Note that here it is assumed that  $\mathbf{x}_0$  does not change. Thus it is not moving through space. Then this theorem is summarized in the following lemma.

**Lemma B.0.1** *For a body which undergoes rigid body motion about a fixed point in three dimensional space, if  $\mathbf{x}(t, \mathbf{x})$  denotes the position vector of the point  $\mathbf{x}$  at time  $t$ , from some fixed point in the body, then there exists a time dependent vector  $\boldsymbol{\omega}(t)$  such that the velocity of this point at time  $t$ ,  $\mathbf{x}_t(t, \mathbf{x})$  is given by*

$$\mathbf{x}_t(t, \mathbf{x}) = \boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x}).$$

*In particular, letting  $\mathbf{x} = \mathbf{e}_i$ , we see that  $\mathbf{e}_i'(t) = \boldsymbol{\omega}(t) \times \mathbf{e}_i(t)$ .*

**Definition B.0.2** *The vector,  $\boldsymbol{\omega}(t)$  whose existence is given by the above lemma is called the angular velocity vector.*

We are now ready to write the total angular momentum of the rigid body. In doing so, we assume the density equals  $\rho(\mathbf{x})$ . Thus at time  $t$  the total angular momentum,  $\boldsymbol{\Omega}$ , would be given by the three dimensional integral,

$$\begin{aligned} \boldsymbol{\Omega} &= \int_{B(0)} \mathbf{x}(t, \mathbf{x}) \times \rho(\mathbf{x}) \mathbf{x}_t(t, \mathbf{x}) d\mathbf{x} \\ &= \int_{B(0)} \rho(\mathbf{x}) \mathbf{x}(t, \mathbf{x}) \times (\boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x})) d\mathbf{x}. \end{aligned} \quad (2.1)$$

In terms of the material basis,  $\{e_1(t), e_2(t), e_3(t)\}$  which is fixed with the body,

$$(\boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x})) = \begin{vmatrix} e_1(t) & e_2(t) & e_3(t) \\ \omega_1(t) & \omega_2(t) & \omega_3(t) \\ x_1 & x_2 & x_3 \end{vmatrix}$$

where the  $\omega_i$  are the components of  $\boldsymbol{\omega}$  taken with respect to  $\{e_1(t), e_2(t), e_3(t)\}$  and as we observed earlier,  $\{x_1, x_2, x_3\}$  are the coordinates of the vector  $\mathbf{x}(t, \mathbf{x})$  taken with respect to the  $\{e_1(t), e_2(t), e_3(t)\}$ . To simplify the integrand in 2.1 that long cross product is simplified.

**Lemma B.0.3** *Let  $\mathbf{a}, \mathbf{b}, \mathbf{c}$  be three dimensional vectors. Then*

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c}) \mathbf{b} - (\mathbf{a} \cdot \mathbf{b}) \mathbf{c}.$$

**Proof:** Let an orthonormal right handed coordinate system  $\{e_1, e_2, e_3\}$  be given. Then

$$\begin{aligned} \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) &= \varepsilon_{ijk} a_j (\mathbf{b} \times \mathbf{c})_k e_i \\ &= \varepsilon_{ijk} \varepsilon_{kpq} a_j b_p c_q e_i \\ &= \varepsilon_{kij} \varepsilon_{kpq} a_j b_p c_q e_i \\ &= (\delta_{ip} \delta_{jq} - \delta_{jp} \delta_{iq}) a_j b_p c_q e_i \\ &= (a_j b_i c_j - a_j b_j c_i) e_i \\ &= (\mathbf{a} \cdot \mathbf{c}) \mathbf{b} - (\mathbf{a} \cdot \mathbf{b}) \mathbf{c}. \blacksquare \end{aligned}$$

Now simplify the integrand using this lemma.

$$\begin{aligned} &\mathbf{x}(t, \mathbf{x}) \times (\boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x})) \\ &= (\mathbf{x}(t, \mathbf{x}) \cdot \mathbf{x}(t, \mathbf{x})) \boldsymbol{\omega}(t) - (\mathbf{x}(t, \mathbf{x}) \cdot \boldsymbol{\omega}(t)) \mathbf{x}(t, \mathbf{x}). \end{aligned}$$

Writing  $\mathbf{x}(t, \mathbf{x})$  and  $\boldsymbol{\omega}(t)$  in terms of the material coordinates,

$$\begin{aligned} \boldsymbol{\omega}(t) &= \omega_1 e_1(t) + \omega_2 e_2(t) + \omega_3 e_3(t), \\ \mathbf{x}(t, \mathbf{x}) &= x_1 e_1(t) + x_2 e_2(t) + x_3 e_3(t), \end{aligned}$$

and so

$$\begin{aligned} &\mathbf{x}(t, \mathbf{x}) \times (\boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x})) = \\ &\sum_i |\mathbf{x}|^2 \omega_i e_i(t) - \left( \sum_i \sum_j x_j \omega_j x_i e_i(t) \right). \end{aligned} \quad (2.2)$$

Thus, listing the components of  $\mathbf{x}(t, \mathbf{x}) \times (\boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x}))$  with respect to the material basis yields the following in which  $\mathbf{x}(t, \mathbf{x}) \times (\boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x}))$  is written as a column vector.

$$\begin{pmatrix} (x_1^2 + x_2^2 + x_3^2) \omega_1 - (x_1^2 \omega_1 + x_2 x_1 \omega_2 + x_3 x_1 \omega_3) \\ (x_1^2 + x_2^2 + x_3^2) \omega_2 - (x_2 x_1 \omega_1 + x_2^2 \omega_2 + x_3 x_2 \omega_3) \\ (x_1^2 + x_2^2 + x_3^2) \omega_3 - (x_3 x_1 \omega_1 + x_3 x_2 \omega_2 + x_3^2 \omega_3) \end{pmatrix}$$



Written as a matrix, this is

$$\begin{pmatrix} x_2^2 + x_3^2 & -x_1x_2 & -x_1x_3 \\ -x_1x_2 & x_1^2 + x_3^2 & -x_2x_3 \\ -x_1x_3 & -x_2x_3 & x_1^2 + x_2^2 \end{pmatrix} \begin{pmatrix} \omega_1(t) \\ \omega_2(t) \\ \omega_3(t) \end{pmatrix}. \quad (2.3)$$

Therefore, the components of angular momentum taken with respect to the material basis are

$$\begin{pmatrix} \Omega_1(t) \\ \Omega_2(t) \\ \Omega_3(t) \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix} \begin{pmatrix} \omega_1(t) \\ \omega_2(t) \\ \omega_3(t) \end{pmatrix}. \quad (2.4)$$

Where

$$\begin{aligned} I_{kk} &= \int_{B(0)} \left( \sum_{j \neq k} x_j^2 \right) \rho(x_1, x_2, x_3) dx \\ I_{ij} &= - \int_{B(0)} x_i x_j \rho(x_1, x_2, x_3) dx, \quad i \neq j. \end{aligned}$$

Thus the matrix in 2.4 is symmetric. Because of the choice of coordinates, this matrix is also time independent. It is called the moment of inertia tensor and the off diagonal terms are called the products of inertia. Now recall that

$$\Omega = \int_{B(0)} \mathbf{x}(t, \mathbf{x}) \times \rho(\mathbf{x}) \mathbf{x}_t(t, \mathbf{x}) dx.$$

Taking the time derivative on both sides, (We do not worry about mathematical details related to differentiating under the integral sign here.)

$$\begin{aligned} \Omega' &= \int_{B(0)} \mathbf{x}_t(t, \mathbf{x}) \times \rho(\mathbf{x}) \mathbf{x}_t(t, \mathbf{x}) dx \\ &\quad + \int_{B(0)} \mathbf{x}(t, \mathbf{x}) \times \frac{d}{dt} (\rho(\mathbf{x}) \mathbf{x}_t(t, \mathbf{x})) dx \\ &= \int_{B(0)} \mathbf{x}(t, \mathbf{x}) \times \frac{d}{dt} (\rho(\mathbf{x}) \mathbf{x}_t(t, \mathbf{x})) dx. \end{aligned}$$

Now from Newton's second law, the force on the chunk of mass,  $\rho(\mathbf{x}) dx$  at time  $t$ , denoted here by  $\mathbf{F}(\mathbf{x}(t, \mathbf{x})) dx$  is just  $\frac{d}{dt} (\rho(\mathbf{x}) \mathbf{x}_t(t, \mathbf{x})) dx$ . Therefore,

$$\Gamma(t) \equiv \Omega'(t) = \int_{B(0)} \mathbf{x}(t, \mathbf{x}) \times \mathbf{F}(\mathbf{x}(t, \mathbf{x})) dx$$

which is the total torque acting on the body at time  $t$ . Note it has units of distance times units of force. Now differentiate the angular momentum to find the torque, this in terms of the moment of inertia tensor of 2.4  $\Omega = I\omega$ . There is a slight complication due to the fact that we have the angular momentum expressed in terms of a basis which is time dependent. Therefore, when we take the derivative of this vector we must include this fact. From 2.4 we see

$$\Omega(t) = \sum_i \sum_j I_{ij} \omega_j(t) \mathbf{e}_i(t).$$

Now recall by Lemma B.0.1,  $\mathbf{e}_i'(t) = \boldsymbol{\omega}(t) \times \mathbf{e}_i(t)$ . Therefore,

$$\begin{aligned} \Gamma(t) &= \Omega'(t) = \\ &= \sum_i \sum_j I_{ij} \omega_j'(t) \mathbf{e}_i(t) + \sum_i \sum_j I_{ij} \omega_j(t) (\boldsymbol{\omega}(t) \times \mathbf{e}_i(t)). \end{aligned} \quad (2.5)$$

This is called Euler's equation for the torque. There are three equations hidden in the above formula, one for each  $\mathbf{e}_i$  for  $i = 1, 2$ , and  $3$ . If you want, you can write them down but there is a simpler way to proceed. Recall the matrix,  $(I_{ij})$  is symmetric and real. Therefore, it can be diagonalized by a unitary real matrix. See Theorem 11.4.7. If we let the columns of this unitary matrix be the  $\mathbf{e}_i$ , it follows the moment of inertia tensor is a diagonal matrix,  $\text{diag}(I_1, I_2, I_3)$  and 2.5 becomes

$$\Gamma(t) = \sum_i I_i \omega_i'(t) \mathbf{e}_i(t) + \sum_i I_i \omega_i(t) (\boldsymbol{\omega}(t) \times \mathbf{e}_i(t))$$

Writing the right side out,  $\Gamma(t) =$

$$\begin{aligned} &I_1 \omega_1' \mathbf{e}_1 + I_2 \omega_2' \mathbf{e}_2 + I_3 \omega_3' \mathbf{e}_3 + I_1 \omega_1 \left( \overbrace{\omega_3 \mathbf{e}_2 - \omega_2 \mathbf{e}_3}^{\boldsymbol{\omega} \times \mathbf{e}_1} \right) + \\ &I_2 \omega_2 \left( \overbrace{\omega_1 \mathbf{e}_3 - \omega_3 \mathbf{e}_1}^{\boldsymbol{\omega} \times \mathbf{e}_2} \right) + I_3 \omega_3 \left( \overbrace{\omega_2 \mathbf{e}_1 - \omega_1 \mathbf{e}_2}^{\boldsymbol{\omega} \times \mathbf{e}_3} \right) \end{aligned}$$

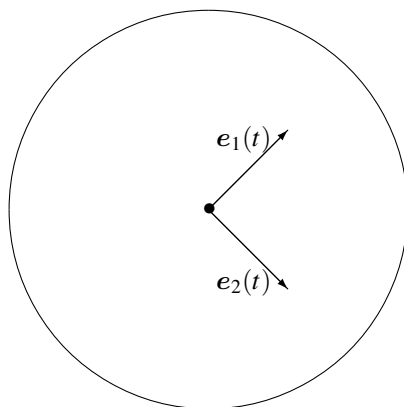
and now, collecting terms,  $\Gamma(t) = \Gamma_1(t) \mathbf{e}_1(t) + \Gamma_2(t) \mathbf{e}_2(t) + \Gamma_3(t) \mathbf{e}_3(t)$  where

$$\begin{aligned} \Gamma_1(t) &= I_1 \omega_1' + \omega_3 \omega_2 (I_3 - I_2) \\ \Gamma_2(t) &= I_2 \omega_2' + \omega_1 \omega_3 (I_1 - I_3) \\ \Gamma_3(t) &= I_3 \omega_3' + \omega_1 \omega_2 (I_2 - I_1). \end{aligned} \quad (2.6)$$

These are called Euler's equations for the torque. Although I invoked the theorem that Hermitian or symmetric matrices can be diagonalized by a unitary transformation in order to get axes with respect to which the moment of inertia tensor is diagonal, it is usually much easier than this. Often there are symmetry considerations which make it obvious how to choose these axes and when this is done 2.6 allows us to compute the torque which results from a given angular velocity.

**Example B.0.4** Consider a disc having negligible thickness and radius  $R$  with constant density  $\rho$  taken with respect to area which spins around its center. How should we choose the material bases to get a nice diagonal moment of inertia tensor?

Consider the following picture in which the vectors  $\mathbf{e}_1(t)$  and  $\mathbf{e}_2(t)$  are shown fixed with the disc which is assumed to be rotating.



We let  $e_3(t) = e_1(t) \times e_2(t)$  so that we have a right handed orthonormal system of basis vectors. We calculate the moment of inertia tensor first.

$$I_{11} \equiv \rho \int_{B(0)} x_2^2 dx = \rho \int_0^{2\pi} \int_0^R (r \sin \theta)^2 r dr d\theta = \frac{1}{4} R^4 \pi \rho$$

By symmetry, we see that  $I_{22} = \frac{1}{4} R^4 \pi \rho$  also. Now

$$I_{33} \equiv \rho \int_{B(0)} (x_2^2 + x_1^2) dx = \rho \int_0^{2\pi} \int_0^R r^3 dr d\theta = \frac{1}{2} \rho \pi R^4.$$

Now by symmetry considerations,  $I_{12} = 0$  as are all the other off diagonal terms. Those that have a 3 in the subscript are zero because we are assuming for the sake of simplicity that the disc has negligible thickness. However, if we didn't assume this we would still get zero for these terms by the symmetry of the shape with respect to the other variable. Therefore, the moment of inertia tensor is

$$\begin{pmatrix} \frac{1}{4} \rho \pi R^4 & 0 & 0 \\ 0 & \frac{1}{4} \rho \pi R^4 & 0 \\ 0 & 0 & \frac{1}{2} \rho \pi R^4 \end{pmatrix}.$$

It follows that for  $\omega = \omega_1(t) e_1(t) + \omega_2(t) e_2(t) + \omega_3(t) e_3(t)$  we can find the Torque by Euler's equations.

$$\begin{aligned} \Gamma_1(t) &= \frac{1}{4} \rho \pi R^4 \omega'_1 + \omega_3 \omega_2 \left( \frac{1}{4} \rho \pi R^4 \right) \\ \Gamma_2(t) &= \frac{1}{4} \rho \pi R^4 \omega'_2 + \omega_1 \omega_3 \left( -\frac{1}{4} \rho \pi R^4 \right) \\ \Gamma_3(t) &= I_3 \omega'_3. \end{aligned} \tag{2.7}$$

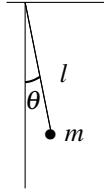
The physical interpretation of  $\omega$  given above is that the term  $\omega_3(t) e_3(t)$  represents the angular velocity about the axis determined by  $e_3(t)$ . Thus it is a measure of how fast

and in what direction the disc is spinning about this axis. If the disc were spinning very fast we would have  $\omega_3(t)$  very large. The other terms of angular velocity,  $\omega_1(t)e_1(t) + \omega_2(t)e_2(t)$ , yield a vector which is in the plane determined by  $e_1(t)$  and  $e_2(t)$  and so it is a measure of the angular velocity about this axis. If we assumed  $\omega'_i(t) = 0$  for each  $i = 1, 2, 3$ , and  $\omega_2$  and  $\omega_1$  are moderate, note that we would still have substantial components of torque,  $\Gamma_2(t)$  and  $\Gamma_1(t)$ . Much more could be said about this problem and more examples could be given but this much will suffice here.

## Appendix C

# Lagrangian Mechanics

Let  $\mathbf{y} = \mathbf{y}(x, t)$  where  $t$  signifies time and  $x \in U \subseteq \mathbb{R}^m$  for  $U$  an open set, while  $\mathbf{y} \in \mathbb{R}^n$  and suppose  $x$  is a function of  $t$ . Physically, this corresponds to an object moving over a surface in  $\mathbb{R}^n$ , its position being  $\mathbf{y}(x, t)$ . If we know about  $x(t)$  then we also know  $\mathbf{y}$ . More generally, we might have  $M$  masses, the position of mass  $\alpha$  being  $\mathbf{y}_\alpha$ . For example, consider the pendulum in which there is only one mass.



in which  $n = 2$ ,  $l$  is fixed and  $y^1 = l \sin \theta$ ,  $y^2 = l - l \cos \theta$ . Thus, in this simple example,  $m = 1$  and  $x = \theta$ . If  $l$  were changing in a known way with respect to  $t$ , then this would be of the form  $\mathbf{y} = \mathbf{y}(x, t)$ . We seek differential equations for  $x$ .

The kinetic energy is defined as

$$T \equiv \frac{1}{2} \sum_{\alpha} m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \dot{\mathbf{y}}_{\alpha} \quad (*)$$

where the dot on the top signifies differentiation with respect to  $t$ . Thus, from the chain rule,  $T$  is a function of  $\dot{x}$ . The following lemma is an important observation.

**Lemma C.0.1** *The following formula holds.*

$$\frac{\partial T}{\partial \dot{x}^k} = \sum_{\alpha} m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k}.$$

**Proof:** From the chain rule,

$$\dot{\mathbf{y}}_{\alpha} = \sum_k \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k} \dot{x}^k + \frac{\partial \mathbf{y}_{\alpha}}{\partial t} \quad (**)$$

and so

$$\frac{\partial \dot{\mathbf{y}}_{\alpha}}{\partial \dot{x}^k} = \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k}.$$

Therefore,

$$\frac{\partial T}{\partial \dot{x}^k} = \sum_{\alpha} m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \frac{\partial \dot{\mathbf{y}}_{\alpha}}{\partial \dot{x}^k} = \sum_{\alpha} m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k} \blacksquare$$

It follows from the above and the product and chain rule that

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}^k} \right) = \sum_{\alpha} m_{\alpha} \ddot{\mathbf{y}}_{\alpha} \cdot \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k} +$$

$$\sum_{\alpha} m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \sum_r \frac{\partial^2 \mathbf{y}_{\alpha}}{\partial x^r \partial x^k} \dot{x}^r + \sum_{\alpha} m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \frac{\partial^2 \mathbf{y}_{\alpha}}{\partial t \partial x^k}. \quad (***)$$

Also from the product rule,

$$\frac{\partial T}{\partial x^k} = \sum_{\alpha} m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \left( \frac{\partial \dot{\mathbf{y}}_{\alpha}}{\partial x^k} \right)$$

But from \*\*,

$$\frac{\partial \dot{\mathbf{y}}_{\alpha}}{\partial x^k} = \sum_r \frac{\partial^2 \mathbf{y}_{\alpha}}{\partial x^k \partial x^r} \dot{x}^r + \frac{\partial^2 \mathbf{y}_{\alpha}}{\partial x^k \partial t}$$

Thus

$$\begin{aligned} \frac{\partial T}{\partial x^k} &= \sum_{\alpha} m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \left( \sum_r \frac{\partial^2 \mathbf{y}_{\alpha}}{\partial x^k \partial x^r} \dot{x}^r + \frac{\partial^2 \mathbf{y}_{\alpha}}{\partial x^k \partial t} \right) \\ &= \sum_{\alpha} \sum_r \left( m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \frac{\partial^2 \mathbf{y}_{\alpha}}{\partial x^k \partial x^r} \dot{x}^r \right) + m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \frac{\partial^2 \mathbf{y}_{\alpha}}{\partial x^k \partial t} \end{aligned}$$

From this and \*\*\*,

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}^k} \right) - \frac{\partial T}{\partial x^k} = \sum_{\alpha} m_{\alpha} \ddot{\mathbf{y}}_{\alpha} \cdot \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k}$$

Now  $\ddot{\mathbf{y}}_{\alpha}$  denotes the acceleration of the  $\alpha^{th}$  mass and so, by Newton's second law, if  $\mathbf{F}$  is the force acting on the object,

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}^k} \right) - \frac{\partial T}{\partial x^k} = \sum_{\alpha} \mathbf{F}_{\alpha} \cdot \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k} \quad (3.1)$$

This is a particularly agreeable formula in case  $\mathbf{F}_{\alpha} = \nabla \Phi_{\alpha}(\mathbf{y}) + \mathbf{g}_{\alpha}$  where  $\mathbf{g}_{\alpha}$  is a force of constraint which causes motion to remain in the surface  $\mathbf{x} \rightarrow \mathbf{y}_{\alpha}(\mathbf{x})$ . Thus  $\mathbf{g}_{\alpha} \cdot \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k} = 0$ . In this special case, you have

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}^k} \right) - \frac{\partial T}{\partial x^k} = \sum_{\alpha} \nabla \Phi_{\alpha}(\mathbf{y}) \cdot \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k} = \sum_{\alpha} \frac{\partial}{\partial x^k} (\Phi_{\alpha}(\mathbf{y}))$$

Let  $\Phi$  denote the total potential energy so  $\Phi = \sum_{\alpha} \Phi_{\alpha}$ . Now  $\Phi_{\alpha}(\mathbf{y})$  does not depend on  $\dot{\mathbf{x}}$ , only on  $\mathbf{x}$ . Hence  $\frac{\partial \Phi_{\alpha}(\mathbf{y})}{\partial \dot{x}_k} = 0$ . It follows that in this special case,

$$\frac{d}{dt} \left( \frac{\partial (T - \Phi)}{\partial \dot{x}_k} \right) - \frac{\partial (T - \Phi)}{\partial x^k} = 0, \quad (3.2)$$

this for each  $k$ . This formula is due to Lagrange.<sup>1</sup>

<sup>1</sup>Joseph Louis Lagrange (1736-1813) was born in Italy but lived much of his life in France which is where he died. He made major contributions to analysis, number theory, and mechanics. His most famous work is likely *Mécanique analytique*. He invented the method of variation of parameters used earlier. With Euler, he invented the calculus of variations and also the method of Lagrange multipliers in order to include constraints. Lagrange was also involved in the development of the metric system.

**Theorem C.0.2** Let  $\mathbf{y}_\alpha(\mathbf{x}, t)$  denote the position of an object of mass  $m_\alpha$  where  $\mathbf{x}$  is a function of  $t$ . Let the kinetic energy be defined by

$$T \equiv \frac{1}{2} \sum_{\alpha} m_{\alpha} \dot{\mathbf{y}}_{\alpha} \cdot \dot{\mathbf{y}}_{\alpha}.$$

Let the mass  $m_\alpha$  be acted on by a force  $\mathbf{F}_\alpha$ . Then Newton's second law implies

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}^k} \right) - \frac{\partial T}{\partial x^k} = \sum_{\alpha} \mathbf{F}_{\alpha} \cdot \frac{\partial \mathbf{y}_{\alpha}}{\partial x^k} \quad (3.3)$$

In case  $\mathbf{F}_\alpha = \nabla \Phi_\alpha + \mathbf{g}_\alpha$  where  $\mathbf{g}_\alpha$  is a force of constraint so the total force comes from forces of constraint and the gradient of a potential function, then

$$\frac{d}{dt} \left( \frac{\partial (T - \Phi)}{\partial \dot{x}^k} \right) - \frac{\partial (T - \Phi)}{\partial x^k} = 0$$

Also, the above 3.3 implies Newton's second law.

**Proof:** The above derivation shows that Newton's law implies the above two formulas. On the other hand, if 3.3 holds, then in the case of one mass, the first part of the derivation which depended only on the chain rule and product rule shows

$$\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}^k} \right) - \frac{\partial T}{\partial x^k} = m \ddot{\mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial x^k}$$

Thus if 3.3 and there is no force of constraint, then  $\mathbf{F} = m\ddot{\mathbf{y}}$  which is Newton's second law. ■

**Example C.0.3** In the case of the simple pendulum,  $\mathbf{x} = \theta$  as shown in the picture and

$$\begin{pmatrix} y^1 \\ y^2 \end{pmatrix} = \begin{pmatrix} l \sin \theta \\ l - l \cos \theta \end{pmatrix}$$

the force acting on weight being  $mg(-\mathbf{j}) = \nabla(-mgy^2)$ . Find the equation of motion of this pendulum.

$$T = \frac{1}{2} m \begin{pmatrix} l \cos(\theta) \theta' \\ l \sin(\theta) \theta' \end{pmatrix} \cdot \begin{pmatrix} l \cos(\theta) \theta' \\ l \sin(\theta) \theta' \end{pmatrix} = \frac{1}{2} m l^2 (\theta')^2$$

Then  $\Phi = -mg(l - l \cos \theta)$ .  $T - \Phi = \frac{1}{2} m l^2 (\theta')^2 + mg(l - l \cos \theta)$ . Thus the equation of motion of this pendulum is

$$\frac{d}{dt} (m l^2 \theta') - m g l (-\sin(\theta)) = 0$$

so

$$\theta'' + \frac{g}{l} \sin \theta = 0$$

This is an equation which doesn't have a simple analytic solution in terms of standard calculus type functions.

**Example C.0.4** In the above simple pendulum, suppose there is a friction force  $-k(\mathbf{y})\dot{\mathbf{y}}$  acting to impede the motion. What are equations of motion in this case?

The following is from the chain rule.

$$\dot{\mathbf{y}} = \begin{pmatrix} l \cos \theta \\ l \sin \theta \end{pmatrix} \theta'$$

Denote  $k(l \sin \theta, l - l \cos \theta)$  as  $k(\theta)$  to save notation. Then it follows from 3.3 and the previous example that

$$\begin{aligned} & \frac{d}{dt} \left( \frac{\partial}{\partial \theta'} \left( \frac{1}{2} m l^2 (\theta')^2 + m g (l - l \cos \theta) \right) \right) - \\ & \frac{\partial}{\partial \theta} \left( \frac{1}{2} m l^2 (\theta')^2 + m g (l - l \cos \theta) \right) = -k(\theta) \theta' \begin{pmatrix} l \cos \theta \\ l \sin \theta \end{pmatrix} \cdot \begin{pmatrix} l \cos \theta \\ l \sin \theta \end{pmatrix} \end{aligned}$$

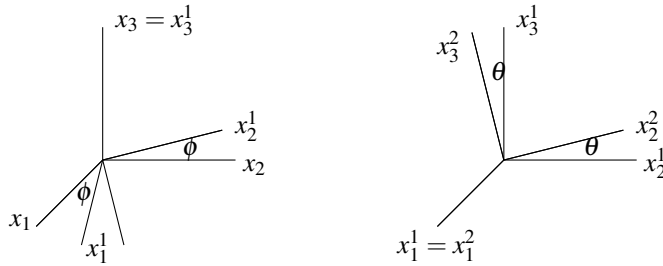
and so

$$\begin{aligned} & \frac{d}{dt} (m l^2 \theta') + m g l \sin(\theta) = -k(\theta) \theta' l^2 \\ & \theta'' + \frac{k(\theta)}{m} \theta' + \frac{g}{l} \sin(\theta) = 0 \end{aligned}$$

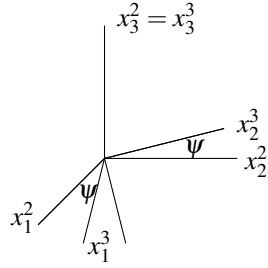
This is another equation for which we don't have a good way to obtain a simple analytic solution.

## C.1 The Spinning Top and the Euler Angles

This material is discussed in [8]. It is due to Lagrange. He was doing this kind of thing in the second half of the 1700's. However, he didn't use pictures to illustrate what he was doing. Here we consider the techniques for considering the motion of rigid bodies rotating about a fixed point in space using Lagrangian mechanics. Earlier, it was done in a way which made it convenient to find the torque given the angular velocity. Here we consider a rigid body as a very large number of point masses which satisfy equations of constraint which cause them to remain at a constant distance from each other. It follows we can consider the motion using only three parameters. The ones we use are called the Euler angles. To describe the Euler angles consider the following picture in which  $x_1, x_2$  and  $x_3$  are the usual coordinate axes fixed in space and the axes labeled with a superscript denote other coordinate axes. Here is the picture.







We obtain  $\phi$  by rotating about the fixed  $x_3$  axis. Next we rotate about the  $x_1^1$  axis which results from the first rotation. This gives  $\theta$ . Finally, we rotate about the  $x_3^2$  axis by  $\psi$ . This can realize any rotation about the origin in this manner. In practice one knows  $\theta'$ ,  $\phi'$  and  $\psi'$  and you want to find a formula for the kinetic energy in terms of these quantities because this will allow you to write a Lagrangian and obtain the equations of motion. A little thought will show that a choice of these angles determines another right handed orthogonal coordinate system,  $x_1^3$ ,  $x_2^3$ , and  $x_3^3$  and that every such system is determined by a suitable choice of the Euler angles. In the context of Lagrangian mechanics above, define  $G_\alpha(\phi, \theta, \psi)$  to be the point in space whose coordinates in  $x_1^3$ ,  $x_2^3$ , and  $x_3^3$  are the same as the coordinates of this point in  $x_1$ ,  $x_2$ , and  $x_3$  and since the body is rigid, the constraints require that  $G(\phi, \theta, \psi) \equiv (G_1(\phi, \theta, \psi), \dots, G_N(\phi, \theta, \psi))$ . Now recall Lemma B.0.1 listed here for convenience.

**Lemma C.1.1** *For a body which undergoes rigid body motion about a fixed point in three dimensional space, if we let  $\mathbf{x}(t, \mathbf{x})$  denote the position vector of the point,  $\mathbf{x}$  at time  $t$ , then there exists a time dependent vector  $\boldsymbol{\omega}(t)$  such that the velocity of this point at time  $t$ ,  $\mathbf{x}_t(t, \mathbf{x})$  is given by*

$$\mathbf{x}_t(t, \mathbf{x}) = \boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x}).$$

*In particular, letting  $\mathbf{x} = \mathbf{e}_i$ , we see that  $\mathbf{e}_i'(t) = \boldsymbol{\omega}(t) \times \mathbf{e}_i(t)$ .*

It follows from this lemma that the total kinetic energy of the rigid body is

$$\frac{1}{2} \int_{B(0)} \rho(\mathbf{x}) |\mathbf{x}_t(t, \mathbf{x})|^2 dx = \frac{1}{2} \int_{B(0)} \rho(\mathbf{x}) |\boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x})|^2 dx.$$

As discussed above, when the Euler angles change, this results in new coordinate axes that come from rotating the original axes. If we let these new axes be fixed with the moving body and call the new axes,  $x_1(t)$ ,  $x_2(t)$ , and  $x_3(t)$  with  $\mathbf{e}_i(t)$  a unit vector in the positive  $x_i(t)$  direction, it follows the coordinates of  $\mathbf{x}(t, \mathbf{x})$  with respect to these new axes are the same as the coordinates of  $\mathbf{x}$  with respect to  $x_1(0)$ ,  $x_2(0)$ , and  $x_3(0)$ , the axes at time  $t = 0$ . We can compute  $|\boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x})|^2$  as follows.

$$\begin{aligned} |\boldsymbol{\omega}(t) \times \mathbf{x}(t, \mathbf{x})|^2 &= \varepsilon_{ijk} \omega^j(t) x^k \varepsilon_{ipq} \omega^p(t) x^q \\ &= (\delta_{jp} \delta_{kq} - \delta_{jq} \delta_{kp}) \omega^j(t) x^k \omega^p(t) x^q \\ &= \omega^j(t) x^k \omega_j(t) x_k - \omega^j(t) x^k \omega_k(t) x_j \\ &= |\mathbf{x}|^2 |\boldsymbol{\omega}|^2 - (\mathbf{x} \cdot \boldsymbol{\omega})^2. \end{aligned}$$

Here  $\omega_i$  are the components of  $\omega$  taken with respect to the  $e_i(t) = e^i(t)$ . Thus, as in Section B,  $|\omega(t) \times x(t, x)|^2 =$

$$\omega^T(t) \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{12} & I_{22} & I_{23} \\ I_{13} & I_{23} & I_{33} \end{pmatrix} \omega(t)$$

Where

$$\begin{aligned} I_{kk} &= \int_{B(0)} \left( \sum_{j \neq k} x_j^2 \right) \rho(x_1, x_2, x_3) dx \\ I_{ij} &= - \int_{B(0)} x_i x_j \rho(x_1, x_2, x_3) dx, \quad i \neq j. \end{aligned}$$

As in this section, choose  $x_1(0)$ ,  $x_2(0)$  and  $x_3(0)$  such that  $I_{ij} = 0$  whenever  $i \neq j$ . Therefore, the kinetic energy in terms of the components of  $\omega$  taken with respect to the axes,  $x_i(t)$  is seen to be

$$T = \frac{1}{2} \sum_{k=1}^3 I_k \omega^k(t)^2.$$

Note that  $I_k$  is independent of  $t$  and the  $\omega^k$  are the components of  $\omega$  taken with respect to the axes,  $x_i(t)$ . While this is a nice formula, we want to relate it to the Euler angles because the Euler angles have more geometric significance. Therefore, what we need to find is  $\omega^k(t)$  in terms of the time derivatives of the Euler angles. Referring to the above picture of the Euler angles, we see that  $\phi'$  contributes a term, to the angular velocity vector which is of the form  $(0, 0, \phi')$  where these are the components taken with respect to  $x_1^1, x_2^1$  and  $x_3^1$ . Writing this vector in terms of the axes,  $x_1^2, x_2^2$  and  $x_3^2$ , we get  $(0, \phi' \sin(\theta), \cos(\theta) \phi')$ . Now to this we add the angular velocity vector contributed by  $\theta'$  which with respect to the axes,  $x_1^2, x_2^2$  and  $x_3^2$  is  $(\theta', 0, 0)$ . Therefore, in terms of  $x_1^2, x_2^2$  and  $x_3^2$ , we have the total angular velocity vector resulting from  $\theta$  and  $\phi$  is  $(\theta', \phi' \sin(\theta), \cos(\theta) \phi')$ . Now we write this vector in terms of the final coordinate system,  $x_1^3, x_2^3$  and  $x_3^3 = x_1(t), x_2(t)$  and  $x_3(t)$ . This yields  $(\cos(\psi) \theta' + \sin(\psi) \sin(\theta) \phi', \cos(\psi) \sin(\theta) \phi' - \sin(\psi) \theta', \cos(\theta) \phi')$ . To this we must add the contribution to the angular velocity from  $\psi'$  which in terms of this last system of coordinate axes is just  $(0, 0, \psi')$ . Therefore, in terms of  $x_1(t), x_2(t)$  and  $x_3(t)$  we have the angular velocity is

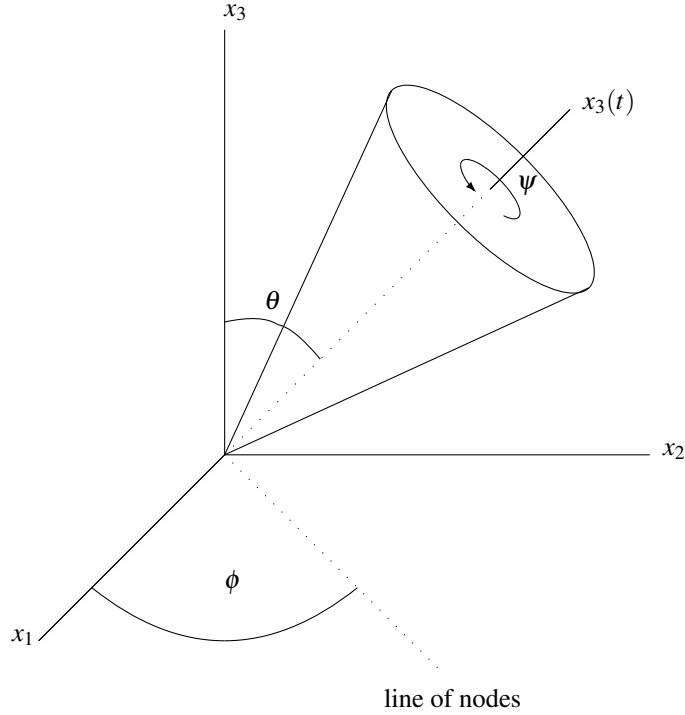
$$\omega = (\cos(\psi) \theta' + \sin(\psi) \sin(\theta) \phi', \cos(\psi) \sin(\theta) \phi' - \sin(\psi) \theta', \cos(\theta) \phi' + \psi').$$

Therefore, the kinetic energy is

$$\begin{aligned} T = \frac{1}{2} & \left( I_1 (\cos(\psi) \theta' + \sin(\psi) \sin(\theta) \phi')^2 + \right. \\ & \left. I_2 (\cos(\psi) \sin(\theta) \phi' - \sin(\psi) \theta')^2 + I_3 (\cos(\theta) \phi' + \psi')^2 \right). \end{aligned} \quad (3.4)$$

Now we will consider a spinning top or gyroscope. Consider the following picture. There are two planes through the origin, one perpendicular to the  $x_3$  axis, and one perpendicular to the  $x_3(t)$  axis. They intersect in the line of nodes shown in the picture. Also, in the above discussion of the Euler angles, we see the  $x_1^1$  axis is in the plane perpendicular to  $x_3$  and also is in the plane perpendicular to  $x_3^3 = x_3(t)$  here. Therefore,  $\phi$  is as shown in the

picture and the other angles are as shown there as well. We see therefore, that  $\phi'$  gives the angular speed of the line of nodes as the axis,  $x_3(t)$  moves around the  $x_3$  axis. Thus  $\phi'$  is a measure of the speed the top or gyroscope moves around the fixed  $x_3$  axis.



We will assume our top has the property that  $I_1 = I_2$ . This would happen, for example if the density is a constant and if the cross sections perpendicular to the  $x_3(t)$  axis are circles. Then the potential energy of the top would be of the form  $Mgl \cos \theta$  where  $M$  is the total mass,  $g$  is the acceleration of gravity, and  $l$  is the distance along the  $x_3(t)$  axis to the center of mass. Then the Lagrangian is of the form

$$L = \frac{1}{2}I_1 [\sin^2(\theta) (\phi')^2 + (\theta')^2] + \frac{1}{2}I_3 [(\cos^2 \theta) (\phi')^2 + 2(\cos \theta) \phi' \psi' + (\psi')^2] - Mgl \cos \theta$$

and therefore, the equations of motion are

$$(I_1 \sin^2(\theta) \phi')' + (I_3 \cos^2(\theta) \phi' + I_3 \cos(\theta) \psi')' = 0 \quad (3.5)$$

$$(I_3 \cos(\theta) \phi' + I_3 \psi')' = 0 \quad (3.6)$$

$$I_1 \theta'' + (\phi')^2 \cos(\theta) \sin(\theta) (I_3 - I_1) + I_3 \sin(\theta) \phi' \psi' - Mgl \sin(\theta) = 0 \quad (3.7)$$

The conservation of energy yields

$$\frac{1}{2}I_1 [\sin^2(\theta) (\phi')^2 + (\theta')^2] + \frac{I_3}{2} [(\cos \theta) (\phi') + \psi']^2 +$$

$$+Mgl \cos \theta = C. \quad (3.8)$$

We can use the conservation of energy along with the equations of motion to gain understanding of the spinning top. From 3.6 we see there is a constant,  $P$  such that  $I_3 \cos(\theta) \phi' + I_3 \psi' = P$  and so

$$\psi' = \frac{P - I_3 \cos(\theta) \phi'}{I_3} \quad (3.9)$$

and from 3.5 there is a constant,  $Q$  such that  $I_1 \sin^2(\theta) \phi' + I_3 \cos^2(\theta) \phi' + I_3 \cos(\theta) \psi' = Q$ . This along with 3.9 implies  $I_1 \sin^2(\theta) \phi' + P \cos(\theta) = Q$  and so we also have

$$\phi' = \frac{Q - P \cos(\theta)}{I_1 \sin^2(\theta)}. \quad (3.10)$$

Therefore, from the conservation of energy,

$$\frac{1}{2} I_1 \left[ \sin^2(\theta) (\phi')^2 + (\theta')^2 \right] + \frac{I_3}{2} P^2 + Mgl \cos \theta = C$$

and using 3.10 to find  $\phi'$ , and adjusting the constant,

$$I_1 (\theta')^2 + I_1 \frac{(Q - P \cos(\theta))^2}{I_1 \sin^2(\theta)} + I_3 P^2 + 2Mgl \cos \theta = C$$

The expression,  $f(\theta) = I_1 \frac{(Q - P \cos(\theta))^2}{I_1 \sin^2(\theta)} + I_3 P^2 + 2Mgl \cos \theta$  is concave up and has some asymptotes. If  $C$  happens to equal the minimum value of  $f$  then we must have  $\theta' = 0$  and so the top will circle around the  $x_3$  axis with  $\theta$  a constant. Thus we would observe the angle between the axis of the top and the  $x_3$  axis would be constant. If  $C$  is not the minimum value of  $f$  then we will have  $\theta$  changing between two values. This is called nutation. Also, from 3.10 we see that  $\phi'$  is probably not zero. Thus the line of nodes moves around the  $x_3$  axis. Even  $\psi'$  may change due to 3.9. If  $\psi'$  were known to be constant, then you could use 3.9 to conclude  $\phi' = \frac{C}{\cos \theta}$ .

# Bibliography

- [1] **Apostol, T. M.**, *Calculus second edition*, Wiley, 1967.
- [2] **Apostol T.** *Calculus Volume II Second edition*, Wiley 1969.
- [3] **Apostol, T. M.**, *Mathematical Analysis*, Addison Wesley Publishing Co., 1974.
- [4] **Baker, Roger**, *Linear Algebra*, Rinton Press 2001.
- [5] **Bartle R.G.**, *A Modern Theory of Integration*, Grad. Studies in Math., Amer. Math. Society, Providence, RI, 2000.
- [6] **Boas, M.** *Mathematical Methods in Physical Science*, John Wiley and Sons, 1966.
- [7] **Boyce, W. and DiPrima, R.** *Elementary Differential Equations and Boundary Value Problems*, John Wiley and Sons, 2005.
- [8] **Bradbury T.C.** *Theoretical Mechanics*, John Wiley and Sons, 1967.
- [9] **Cramér H.**, *Mathematical Methods of Statistics*, Princeton University Press, 1957.
- [10] **Chahal J. S.** , *Historical Perspective of Mathematics* 2000 B.C. - 2000 A.D.
- [11] **Davis, H. and Snider, D.** *Introduction to Vector Analysis*, William C. Brown 1997.
- [12] **D'Angelo, J. and West D.** *Mathematical Thinking Problem Solving and Proofs*, Prentice Hall 1997.
- [13] **Edwards C.H.** *Advanced Calculus of several Variables*, Dover 1994.
- [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] **Gray, A. and Mathews, G. B.** *A Treatise on Bessel functions and Their Applications to Physics*, Macmillan, N.Y. 1952.
- [18] **Greenberg, M.** *Advanced Engineering Mathematics*, Second edition, Prentice Hall, 1998
- [19] **Gurtin M.** *An introduction to continuum mechanics*, Academic press 1981.

- [20] **Hardy G.**, *A Course Of Pure Mathematics, Tenth edition*, Cambridge University Press 1992.
- [21] **Hog R. and Craig A.**, *Introduction to Mathematical Statistics*, third edition, Macmillan Publishing co. 1970.
- [22] **Horn R. and Johnson C.** *matrix Analysis*, Cambridge University Press, 1985.
- [23] **Ince, E.L.** *Ordinary Differential Equations*, Dover 1956.
- [24] **Karlin S. and Taylor H.** *A First Course in Stochastic Processes*, Academic Press, 1975.
- [25] **Kuttler K.L.**, *Elementary Differential Equations*, CRC Press 2018. 573 pages.
- [26] **Kuttler K. L.**, *Basic Analysis*, Rinton
- [27] **Kuttler K.L.**, *Modern Analysis* CRC Press 1998.
- [28] **Lang S.** *Real and Functional analysis* third edition Springer Verlag 1993. Press, 2001.
- [29] **Leighton, W.** *An Introduction to the Theory of Differential Equations*, McGraw Hill, 1952.
- [30] **Nobel B. and Daniel J.** *Applied Linear Algebra*, Prentice Hall, 1977.
- [31] **Rudin, W.**, *Principles of mathematical analysis*, McGraw Hill third edition 1976
- [32] **Rudin W.**, *Real and Complex Analysis*, third edition, McGraw-Hill, 1987.
- [33] **Salas S. and Hille E.**, *Calculus One and Several Variables*, Wiley 1990.
- [34] **Sears and Zemansky**, *University Physics, Third edition*, Addison Wesley 1963.
- [35] **Spivak M.**, *Calculus On Manifolds*, Benjamin 1965.
- [36] **Tierney John**, *Calculus and Analytic Geometry*, fourth edition, Allyn and Bacon, Boston, 1969.
- [37] **Widder, D.** *Advanced Calculus*, second edition, Prentice Hall 1961.
- [38] **Yosida, K.** *Lectures on Differential and Integral Equations*, Dover 1991.

# Index

- $C^1$ , 306
- $C^k$ , 306
- $\Delta$ , 400
- $\cap$ , 15
- $\cup$ , 15
- $\mathbb{R}^n$ , 53
- $\nabla^2$ , 400
  
- Abel's formula, 522
- absolute value
  - complex number, 20
- acceptance region, 810
- adjoint, 185
- adjugate, 181, 506, 534
- affine linear equations, 562
- affine linear procedure, 564
- agony, pain and suffering, 353
- analytic, 615, 616, 698
- analytic functions
  - integral domain, 725
  - uniform convergence, 724
- analytic mappings
  - preservation of angles, 721
- angle between planes, 97
- angle between vectors, 78
- angular velocity, 93
- angular velocity vector, 454, 871
- annulus, 730
- anti-derivative, 42
- arc length, 262
- arcwise connected, 250
  - connected, 250
- area of a parallelogram, 86
- arithmetic mean, 344
- asymptotically stable, 551
- augmented matrix, 107
- autonomous, 553
  
- back substitution, 107
  
- balance of momentum, 412
- basic variables, 115
- basis, 163
- basis, 465
- basis of eigenvectors
  - diagonalizable, 197
- beams
  - buckling, 661
- Bernoulli equation, 545
- Bessel equation
  - $\nu$  an integer, 628
  - eigenvalues, 663
  - parameter not integer, 627
- Bessel equations, 626
- Bessel function
  - integral identity, 632
- Bessel functions
  - addition formula, 631
  - generating function, 631
  - zeros, 663
- Bezier curves, 261
- biased estimate, 818
- binomial distribution, 776, 780
- binomial theorem, 28, 772
- binormal, 280
- boundary problem, 637
- boundary value problem
  - Sturm-Liouville, 661
- bounded, 232
- box product, 89
- Bromwich integral, 759
- buckling beams
  - eigenvalues, 661
  
- Cartesian coordinates, 54
- Cartesian product, 231
- Casorati Weierstrass theorem, 735
- catenary, 549
- Cauchy integral theorem, 709

- Cauchy principal value, 748
- Cauchy Riemann equations, 699
- Cauchy Schwarz inequality, 62, 76, 83
- Cauchy sequence, 239
- Cauchy stress, 415
- Cauchy-Schwarz inequality, 644
- Cavendish, 449, 572
- Cayley Hamilton theorem, 513
- Ceasaro means
  - convergence, 652
  - uniform convergence, 653
- center of mass, 92
- central force, 283
- central force field, 449
- central limit theorem, 794
- centripetal force, 448
- chain rule, 319
- change of variables formula, 381
- characteristic function, 788
- characteristic polynomial, 183, 509, 513
- chemical reactions
  - balancing, 117
- chi-squared
  - moment generating function, 790
  - sample variance and variance, 802
- chi-squared distribution, 782
- circular helix, 281
- classical adjoint, 181, 506
- closed set, 229
- closure of a set, 247
  - limit points, 247
- coefficient of thermal conductivity, 324
- cofactor, 175, 176, 499, 500, 532
- cofactor matrix, 176, 500
- column space, 159
- combinations, 771
- compact, 252
- complement, 229
- completeness, 31
- complex conjugate, 19
- complex numbers, 18
- complex numbers
  - arithmetic, 18
  - roots, 22
  - triangle inequality, 20
- component, 69, 84
- component of a force, 80
- components of a matrix, 130
- compound interest, 542
- computer algebra systems
  - MATLAB, 567
- conditional probability, 791
- confidence interval
  - mean, 808
  - ratio of variances, 814
- confidence intervals, 802
- conjugate
  - of a product, 29
- connected, 247
  - extended complex plane, 719
  - open balls, 250
- connected component, 248
- connected components, 248
  - equivalence class, 249
  - equivalence relation, 249
  - open sets, 249
- connected set
  - continuous function, 251
  - continuous image, 248
- connected sets
  - intersection, 248
  - intervals, 249
  - real line, 249
  - union, 248
- conservation of mass, 412
- conservative, 269, 442
- consistent, 117
- constitutive laws, 418
- contented set, 851
- continuity
  - limit of a sequence, 241
- continuous function, 221
- continuous functions
  - properties, 226
- contour integral, 700
- converge, 239
- convergence
  - Fourier series, 647
  - infinite series, 246
  - midpoint of jump, 647
  - pointwise, 243
  - uniform, 243
- convolution integral, 595
- Coordinates, 53
- Coriolis acceleration
  - earth, 459



- Coriolis force, 448
- counting zeros, 722
- counting zeros of analytic function, 722
- countour integrals
  - definition, 700
- Cramer's rule, 512, 535
- critical point, 330
- cross product, 86
  - area of parallelogram, 86
  - coordinate description, 87
  - geometric description, 86
- cross product
  - general curvilinear coordinates, 483
- curl, 399
- curl
  - general curvilinear coordinates, 483
- curvature, 275, 280
  - independence of parametrization, 276
- D'Alembert, 296
- De Moivre's theorem, 21
- defective, 523
- deformation gradient, 413
- degrees of freedom, 782
- density and mass, 353
- dependent, 162
- derivative, 300
  - has all derivatives, 712
- derivative of a function, 254
- determinant, 527
  - alternating property, 530
  - cofactor, 175, 498
  - cofactor expansion, 532
  - expanding along row or column, 176, 499
  - expansion along row (column), 532
  - matrix inverse formula, 181, 506, 534
  - minor, 174, 497
  - product, 179, 503, 531
  - row operations, 178, 502
  - transpose, 529
- diameter, 238
- difference quotient, 254
- differentiable, 297
- differential, 302
- differential equation
  - different notation, 541
  - linear uniqueness, 540
- differential equations
  - homogeneous, 554
  - separable, 546
- differential equations, 473
- differential equations of motion
  - derivation, 879
- differentiation rules, 257
- dimension, 163
- directed line segment, 58
- direction vector, 58
- directional derivative, 288
- directrix, 85
- Dirichlet kernel, 648
- distance formula, 60
- distribution function, 803, 833
- divergence, 399
- divergence, 478
  - general curvilinear coordinates, 479
- divergence theorem, 404
- donut, 392
- dot product, 75
  - geometric description, 78
- dual basis, 466
- dual basis, 470
- echelon form, 109
- eigenfunctions
  - Bessel functions, 663
- eigenspace, 209
- eigenvalue, 187, 343
- eigenvalues, 513
- eigenvector, 187
- Einstein summation convention, 95
- elementary matrix, 166
- elementary operations, 105
- empty set, 16
- entries of a matrix, 130
- equal area rule, 450
- equality of mixed partial derivatives, 294
- equations of motion
  - spherical coordinates, 880
- equilibrium point, 551
  - stable, 551
- Euler equation, 611
- Euler's equations, 874
- Eulerian coordinates, 413
- even extension, 658
- events, 791

- exact equations, 556
  - procedure, 557
- expansion
  - Legendre polynomials, 665
- exponential growth, 581
- extended complex plane, 719
  - separated sets, 719
- F distribution, 814
- factorial, 772
- Fejer kernel, 652
- Fick's law, 324, 427
- field axioms, 19
- finding fundamental matrix
  - Laplace transforms, 592
- first order systems
  - Laplace transform, 591
- flip, 492
- focus, 66
- force, 67
- force field, 265, 449
- Foucault pendulum, 460
- Fourier law of heat conduction, 324
- Fourier series, 206, 643
  - integrating term by term, 654
  - pointwise convergence, 647
  - term by term differentiation, 658
  - term by term integration, 656
- Fourier transform, 754
  - inverse, 754
- Fredholm alternative, 199
- free variables, 115
- Frenet Serret formulas, 281
- Frobenius norm, 216
- fundamental matrix, 586
- fundamental theorem line integrals, 443
- fundamental theorem of algebra, 23, 26
- fundamental theorem of algebra
  - plausibility argument, 26
  - rigorous proof, 27
- fundamental theorem of calculus, 38, 41
- gamma function, 770
  - existence and convergence, 45, 751
  - factorials, 752
  - properties, 46, 752
- Gauss Elimination, 117
- Gauss elimination, 107, 108
- Gauss Jordan method for inverses, 147
- Gauss's theorem, 404
- general solution, 599
- geometric mean, 344
- geometric multiplicity, 522
- gradient, 290
- gradient
  - contravariant components, 478
  - covariant components, 478
- Gram Schmidt process, 204
- Green's theorem, 431, 432, 442, 704
- grids, 843
- hanging chain, 548
- harmonic, 295
- heat equation, 295
  - derivation, 670
  - solutions, 672
- Heine Borel theorem, 252
- Hermitian, 194
- Hessian matrix, 332, 346
- Holder continuous, 227
- homogeneous coordinates, 158
- homogeneous equation
  - procedure, 555
- homogeneous equations, 554
- hypergeometric, 780
- hypergeometric distribution, 779
- hypergeometric equation, 635
- imaginary part, 698
- implicit function theorem, 485, 489
- inconsistent, 114, 117
- increment of volume
  - increment of area, 381
- independence
  - moment generating functions, 794
  - quadratic forms, 820
- independent, 162
- independent random variables
  - density, 793
  - functions of, 793
- index
  - lowering, 467
  - raising, 467
- indicial equation, 612
- initial condition, 538
- initial value problem, 538

- inner product, 75
- integral
  - uniform convergence, 44
- integral curves, 547
- integral domain, 725
- integrating factor, 538, 544, 557
  - Euler, 559
  - procedure, 561
- intercepts, 99
- interior point, 229
- intermediate value theorem, 249
- intersection, 15
- intervals
  - notation, 16
- inverse
  - left inverse, 534
  - right inverse, 534
- inverse Fourier transform, 754
- inverse function theorem, 488, 490
- inverses and determinants, 507, 533
- invertible, 144
- isolated singularity, 734
- iterated integral, 349
  
- Jacobian determinant, 381
- Jordan content, 849
- Jordan set, 851
- joule, 81
  
- Kepler's first law, 450
- Kepler's laws, 450
- Kepler's third law, 453
- kilogram, 92
- kinetic energy, 877
- kinetic energy, 472
- Kirchoff's law, 127, 128
- Kroneker delta, 95
- Kroneker symbol, 144
  
- Lagrange
  - mechanics, 879
- Lagrange multipliers, 341, 490, 491
- Lagrange remainder, 345, 346
- Lagrangian coordinates, 413
- Lagrangian formalism, 473
- lake pollution, 542
- Laplace equation
  - circular disk, 686
  - rectangles, 683
- Laplace expansion, 500, 532
- Laplace transform, 47, 581, 752
  - linear, 581
  - obvious properties, 48
  - step function, 582
  - table, 582
- Laplacian, 294
  - cylindrical coordinates, 669
  - polar coordinates, 322
  - spherical coordinates, 669
- Laplacian
  - general curvilinear coordinates, 480
- Laurent series, 734
- law of large numbers, 786
- leading entry, 108
- least squares regression, 295
- Lebesgue number, 252, 849
- Lebesgue's theorem, 851
- Legendre equation
  - eigenvalues, 665
  - existence of bounded solutions, 665
- Legendre polynomials
  - expansion, 665
- length of smooth curve, 263
- Leontief model, 158
- level curves
  - intersecting at right angles, 721
- limit of a function, 223
- limit point, 233, 285
- limits and continuity, 226
- line integral, 266
- line of nodes, 882
- linear and nonlinear equations
  - differences, 567
- linear combination, 132, 159, 530
- linear differential equations
  - zeros of solutions, 636
- linear equation
  - procedure, 544
- linear equations
  - solution, 564
- linear regression
  - slope confidence interval, 832
  - variance confidence interval, 831
- linear system
  - general solution, 594
- linear transformations, 135
- lines

- parametric equation, 58
- Liouville
  - transformation, 663
- Liouville's theorem, 717
- Lipschitz, 227, 228
- lizards
  - surface area, 390
- local extremum, 329
- local maximum, 329
- local minimum, 329
- locally one to one, 724
- logistic equation, 550
- Lotka Volterra equations, 221
- lower sum, 844
- main diagonal, 177, 501
- mass balance, 412
- material coordinates, 413
- math induction, 17
- mathematical induction, 17
- matrix, 88, 129
  - identity, 144
  - inverse, 144
  - left inverse, 534
  - lower triangular, 177, 501
  - right inverse, 534
  - rotation about given vector, 142
  - self adjoint, 212
  - upper triangular, 177, 501
- matrix inverse
  - finding it, 145, 147
- maximum likelihood estimates, 816
- maximum modulus theorem, 724
- mean, 789
- mean square
  - poinwise, 667
- mean square norm, 643
- meromorphic, 736
- metric tensor, 466, 468
- metric tensor, 477
- minimal polynomial, 171
- minimum polynomial, 187
- minor, 174, 176, 499, 500, 532
- Mittag Leffler, 744
- mixed partial derivatives, 293
- moment generating function, 787, 788
- moment of a force, 91
- moment of inertia tensor, 873
- moments, 788
- motion, 413
- multi-index, 222
- multinomial expansion, 778
- Navier, 428
- nested interval lemma, 32
- Neuman series, 158
- Newton, 70
- Newton's second law, 473
- nilpotent, 518
- nonhomogeneous problems
  - solutions, 680
- nonlinear equation
  - no uniqueness, 566
- nonlinear equations
  - local solutions, 565
- norm
  - mean square, 643
  - uniform, 643
- normal vector to plane, 97
- null hypothesis, 810, 837
- nutation, 884
- obnoxious integrals, 738
- odd extension, 658
- one to one
  - rank, 170
- open cover, 252
- open mapping theorem, 723, 727
- open set, 229
- order of a pole, 736
- order of a zero, 715
- orientable, 440
- orientation, 265
  - positive, 706
- oriented curve, 265
- origin, 53
- orthogonal matrix, 191, 518
- orthogonality conditions, 641
- orthonormal, 191
- osculating plane, 276, 280
- parallelepiped
  - definition, 89
  - volume, 89
- parameter, 58
- parametric equation, 58
- parametrization, 262

- partial derivative, 290
- partial differential equations
  - nonhomogeneous, 678
- partition, 33
- partition of unity, 864
- periodic, 641
- permutation, 527
- permutation symbol, 95
- permutations and combinations, 777
- permutations of  $n$  things taken  $m$  at a time, 776
- perp, 198
- perpendicular, 79
- piecewise continuous, 654
- Piola Kirchhoff stress, 417
- pivot, 114
- pivot column, 109
- pivot columns, 110
- pivot position, 109
- pivot positions, 110
- planes, 97
- Poincare Bendixon theorem, 608
- point at infinity, 635
- pointwise
  - mean square, 667
- pointwise convergence, 42
- Poisson distribution, 781
- Poisson's equation
  - representation, 689
- polar form complex number, 21
- pole, 735
- polynomial, 25
  - addition, 25
  - degree, 25
  - division, 25
  - equality, 25
  - multiplication, 25
- polynomial
  - leading term, 25
  - monic, 25
- polynomials
  - factoring, 23
- polynomials in  $n$  variables, 222
- position vector, 56, 67
- power series
  - Cauchy product, 615
  - of inverse, 615
- primitive, 492, 706
- principal invariants, 510
- principal logarithm, 721
- principal normal, 275, 280
- principal part
  - Laurent series, 734
- probability
  - conditional, 791
  - properties, 791
- product of inertia, 873
- product rule
  - cross product, 258
  - dot product, 258
- projection, 167
- projection of a vector, 80
- projection onto a subspace, 201
- pumpkin
  - flying, 543
- quadratic form
  - distribution, 822
  - moment generating function, 819
- quadratic forms
  - independence, 821
  - summary of properties, 825
- quadratic formula, 23
- radius of curvature, 275, 280
- random variables
  - density, 793
  - independent, 792
- rank, 164
  - existence of solutions, 198
- rank and singular values, 208
- real part, 698
- refinement of grids, 843
- regular singular point, 615
  - Euler equation, 616
  - exponents of singularity, 620
  - solution procedure, 625
- regular singular points
  - Euler equations, 616
  - finding them, 616
  - indicial equation, 620
- removable singularity, 734
- residue, 737
- resultant, 69
- Reynolds transport formula, 420
- Riccati equation, 571, 662

- Riemann criterion, 845
- Riemann integrable, 351, 356
- Riemann integral, 351, 356, 845
- Riemann sphere, 718
- Riemann sums, 350
- Riemann-Lebesgue lemma, 646
- right handed system, 85
- right inverse, 147
- ring, 725
- root test, 244
- rot, 399
- Rouche's theorem, 722
- row operations, 108, 178, 502
- row reduced echelon form, 109
- Runge-Kutta method
  - syntax, 602
- saddle point, 332
- sample average, 798
- sample mean, 798
- sample variance, 798
- scalar field, 399
- scalar multiplication, 55
- scalar potential, 443
- scalar product, 75
- scalars, 55, 129
- Schur's theorem, 192, 193
- Schwarz lemma, 724
- second derivative test, 348
- semi-stable
  - equilibrium point, 573
- separable differential equations, 546
- separable equations
  - procedure, 554
- separated sets, 247
- separation of variables, 672
- sequence, 239
- sequential compactness, 240
- sequentially compact, 240
- set notation, 15
- sgn, 525
  - uniqueness, 527
- sign of a permutation, 527
- simply connected, 719
- singular point, 330
- singular value decomposition, 208
- singular values, 207
- skew symmetric, 138, 153
- smooth curve, 262
- solving heat equation, 672
- spacial coordinates, 413
- span, 159, 530
- speed, 70
- spherical coordinates, 471
- stability from graphs, 553
- stable, 551
- stable equilibrium point, 551
- standard normal deviate, 790
- standard position, 68
- star shaped, 721
- statistic, 797, 805
- statistics, 797
- stereographic projection, 718
- Stoke's theorem, 437
- Stokes, 428
- Sturm
  - separation theorem, 662
- Sturm-Liouville problem, 661
- subsequence, 239
- subspace, 160
- support of a function, 864
- symmetric, 138, 153
- symmetric form of a line, 59
- T distribution, 806
- Taylor polynomial
  - sine, 346
- Taylor's formula, 345
- term by term differentiation
  - Fourier series, 658
- term by term integration
  - Fourier series, 656
- torque vector, 91
- torsion, 280, 281
- torus, 392
- trace, 216
  - sum of eigenvalues, 216
- traces, 100
- transformation rules, 474
- transpose
  - dot product, 198
- triangle inequality, 63, 77
  - complex numbers, 20
- trig. identities, 641
- two point boundary value problem, 637
- uniform continuity, 243

- uniform convergence, 42
  - infinite sums, 246
- uniform norm, 243, 643
- uniformly Cauchy sequence of functions, 44
- uniformly close, 643
- uniformly continuous, 228, 251
- union, 15
- unit tangent vector, 275, 280
- unitary matrix, 191
- upper sum, 844
- Urysohn's lemma, 252
  
- variance, 789
  - meaning, 794
- vector
  - scalar multiplication, 55
  - vector addition, 55
- vector
  - contravariant components, 467
  - covariant components, 467
- vector field, 265, 399
- vector fields, 218
- vector potential, 401
- vector space axioms, 131
- vector valued function
  - continuity, 222
  - derivative, 254
  - integral, 254
  - limit theorems, 223
- vector valued functions, 217
- vectors, 53, 67
- velocity, 70
- volume element, 381
- volume increment, 381
- volume of parallelepiped, 379
- volume of unit ball in  $n$  dimensions, 424
  
- wave equation, 295
  - derivation, 675
- well ordered, 17
- well ordering, 16
- winding number
  - orientation, 706
  - simple closed curve, 706
- work, 266
- Wronskian, 521
- Wronskian alternative, 599
  
- zero matrix, 130
- zeros of analytic function
  - counting them, 722