

Basic Analysis II: A Modern Calculus in Many Variables



The cephalopods were eager to learn. Now both squid
and octopi are coming to the lectures.

James K. Peterson
Department of Mathematical Sciences
Clemson University
email: petersj@clemson.edu
© James K. Peterson First Edition
Gneural Gnome Press

Version 07.10.2018 : Compiled July 10, 2018

Dedication I dedicate this work to all of my students who have learned these ideas of analysis through my courses. I have learned as much from them as I hope they have from me. I am a firm believer that all my students are capable of excellence and that the only path to excellence is through discipline and study. I am always proud of my students for doing so well on this journey. I hope these notes in turn make you proud of my efforts.

Abstract This book introduces you to more ideas from multidimensional calculus. In addition to classical approaches, we will also discuss things like one forms, a little algebraic topology and a reasonable coverage of the multidimensional versions of the one variable Fundamental Theorem of Calculus ideas. We also continue your training in the *abstract* way of looking at the world. We feel that is a most important skill to have when your life's work will involve quantitative modeling to gain insight into the real world.

Acknowledgements Since we regained the ability to run in 2009, we are less grumpy than usual. Running on the local trails in the forest helps with the stress levels and counteracts our most important philosophical view: "*There is always room for a donut!*" Of course, now that Jim is a bit older, all those donuts, cookies and so forth have had to be cut back a bit. It turns out that as we age, our physiology slows to a crawl and even a cookie crumb expands into several pounds along the old waistline. However, we still look wistfully at all the sweets even though we have fewer samples!

Our daily coffee, half decaf now, is still good though and holding a cup in our hands, we wish to thank all of you who have been students for helping us by listening to what we say in the lectures, finding our typographical errors and other mistakes. We are always hopeful that our efforts help students to some extent and also impart some of our obvious enthusiasm for the subject. Of course, the reality is that we have been damaging students for years by forcing them to learn these abstract things. This is why we tell people at parties we are a roofer or electrician. If we are identified as a mathematician, it could go badly given the terrible things we inflict on the students in the classes. Who knows whom they have told about our tortuous methods. Hence, anonymity is best.

Still, by writing these notes, we have gone public. Sigh. This could be bad. So before we are taken out with a very public hit, we, of course, want to thank our family for all their help in many small and big ways. It is amazing to me that we have been teaching this material since our children were in grade school!

We also want to acknowledge the great debt we have to our wife, Pauli, for her patience in dealing with those vacant stares and the long hours spent in typing and thinking. You are the love of my life.

The cover for this book is an original painting by me which was done in July 2017. It shows the moment when cephalopods first reached out to ask to be taught advanced mathematics. We were awed by their trust in picking us.

History Based On:
Handwritten Notes For Modern Analysis
MATH 9740 2007
MATH 4500 2015

Table Of Contents

I	Introduction	1
1	Beginning Remarks	3
1.1	Table of Contents	4
II	Linear Mappings	7
2	Preliminaries	9
2.1	The Basics	9
2.2	Some Topology in \mathbb{R}^2	18
2.2.1	Homework	20
2.3	Bolzano - Weierstrass in \mathbb{R}^2	20
2.3.1	Homework	22
3	Vector Spaces	23
3.1	Introduction	23
3.2	Vector Spaces over a Field	24
3.2.1	The Span of a Set of Vectors	26
3.3	Inner Products	29
3.3.1	Homework	31
3.4	Examples	31
3.4.1	Two Dimensional Vectors in the Plane	31
3.4.2	The Connection between Two Orthonormal Bases	34
3.4.3	The Invariance of The Inner Product	36
3.4.4	Two Dimensional Vectors As Functions	37
3.4.5	Three Dimensional Vectors in Space	44
3.4.6	The Solution Space of Higher Dimensional ODE Systems	48
3.5	Best Approximation in a Vector Space With Inner Product . . .	52
3.5.1	Homework	54
4	Linear Transformations	55
4.1	Organizing Point Cloud Data	55
4.1.1	Homework	56
4.2	Linear Transformations	56
4.2.1	Homework	56
4.3	Sequence Spaces Revisited	57
4.3.1	Homework	63
4.4	Linear Transformations Between Normed Linear Spaces . . .	64
4.4.1	Basic Properties	65

4.4.2	Mappings between Finite Dimensional Vector Spaces	66
4.5	Magnitudes of Linear Transformations	71
5	Symmetric Matrices	75
5.1	The General Two by Two Symmetric Matrix	75
5.1.1	Examples	76
5.1.2	A Canonical Form	77
5.1.3	Two Dimensional Rotations	80
5.1.4	Homework	81
5.2	Rotating Surfaces	81
5.2.1	Homework	84
5.3	An Complex ODE System Example	84
5.3.1	The General Real and Complex Solution	84
5.3.2	Rewriting The Real Solution	86
5.3.3	Signed Definite Matrices	88
5.3.4	Summarizing	89
6	Matrix Sequences and Ordinary Differential Equations	91
6.1	Linear Systems of ODE	91
6.1.1	The Characteristic Equation	92
6.1.2	Finding The General Solution	94
6.1.3	Homework	97
6.2	Symmetric Systems of ODEs	97
6.2.1	Writing The Solution Another Way	99
6.2.2	Homework	101
7	Continuity and Topology	103
7.1	Topology in n Dimensions	103
7.1.1	Homework	106
7.2	Cauchy Sequences	106
7.3	Compactness	107
7.4	Functions of Many Variables	112
7.4.1	Limits and Continuity for Functions of Many Variables	112
7.4.2	Homework	116
8	Abstract Symmetric Matrices	119
8.1	Input - Output Ratios for Matrices	119
8.1.1	Homework	120
8.2	The Norm of a Symmetric Matrix	120
8.2.1	Constructing Eigenvalues	122
8.3	What Does This Mean?	125
8.3.1	A Worked Out Example	127
8.3.2	Homework	128
8.4	Signed Definite Matrices Again	128
8.4.1	Homework	128

TABLE OF CONTENTS

v

9 Rotations and Orbital Mechanics	129
9.1 Introduction	129
9.1.1 Homework	131
9.2 Orbital Planes	132
9.2.1 Orbital Constants	132
9.2.2 The Orbital Motion is a Conic	134
9.2.3 The Constant B Vector	134
9.2.4 The Orbital Conic	136
9.3 Three Dimensional Rotations	139
9.3.1 Homework	141
9.3.2 Drawing Rotations	141
9.3.3 Rotated Ellipses	146
9.4 Drawing the Orbital Plane	148
9.4.1 The Perifocal Coordinate System	149
9.4.2 Orbital Elements	150
9.5 Drawing Orbital Plane In MatLab	152
9.5.1 Homework	159
10 Determinants and Matrix Manipulations	161
10.1 Determinants	161
10.1.1 Consequences One	162
10.1.2 Homework	164
10.1.3 Consequences Two	164
10.1.4 Homework	170
10.2 Matrix Manipulation	170
10.2.1 Elementary Row Operations and Determinants	170
10.2.2 MatLab Implementations	173
10.2.3 Matrix Inverse Calculations	177
10.3 Back To Definite Matrices	181
10.3.1 Matrix Minors	182
III Calculus of Many Variables	185
11 Differentiability	187
11.1 Partial Derivatives	187
11.1.1 Homework	188
11.2 Tangent Planes	189
11.3 Derivatives for Scalar Functions of n Variables	192
11.3.1 The Chain Rule for Scalar Functions of n Variables	194
11.4 Partials and Differentiability	198
11.4.1 Partials Can Exist but not be Continuous	198
11.4.2 Higher Order Partials	202
11.4.3 When Do Mixed Partials Match?	204
11.5 Derivatives for Vector Functions of n Variables	208
11.5.1 The Chain Rule For Vector Valued Functions	211
11.6 Tangent Plane Error	213
11.6.1 The Mean Value Theorem	213
11.6.2 Hessian Approximations	214
11.7 A Specific Coordinate Transformation	219
11.7.1 Homework	221

12 Multivariable Extremal Theory	223
12.1 Differentiability and Extremals	223
12.2 Second Order Extremal Conditions	224
12.2.1 Positive and Negative Definite Hessians	225
12.2.2 Saddles	226
12.2.3 Expressing Conditions in Terms of Partials	227
13 The Inverse and Implicit Function Theorems	233
13.1 Mappings	233
13.2 Invertibility Results	238
13.2.1 Homework	244
13.3 Implicit Function Results	244
13.3.1 Homework	250
13.4 Constrained Optimization	250
13.4.1 What Does the Lagrange Multiplier Mean?	254
13.4.2 Homework	255
14 Linear Approximation Applications	257
14.1 Linear Approximations to Nonlinear ODE	257
14.1.1 An Insulin Model	257
14.1.2 An Autoimmune Model	274
14.2 Finite Difference Approximations in PDE	275
14.2.1 First Order Approximations	276
14.2.2 Second Order Approximations	278
14.2.3 Homework	278
14.3 FD Approximations	279
14.3.1 Error Analysis	281
14.3.2 Homework	283
14.4 FD Diffusion Code	284
14.4.1 Homework	286
IV Integration	289
15 Integration in Multiple Dimensions	291
15.1 The Darboux Integral	291
15.1.1 Homework	302
15.2 The Riemann Integral in n Dimensions	302
15.2.1 Homework	305
15.3 Volume Zero and Measure Zero	305
15.3.1 Measure Zero	305
15.3.2 Volume Zero	309
15.4 When is a Function Riemann Integrable?	310
15.4.1 Homework	315
15.5 Integration and Sets of Measure Zero	315
16 Change of Variables and Fubini's Theorem	319
16.1 Linear Maps	319
16.1.1 Homework	323
16.2 The Change of Variable Theorem	324
16.2.1 Homework	328
16.3 Fubini Type Results	329

<i>TABLE OF CONTENTS</i>	vii
16.3.1 Fubini on a Rectangle	329
16.3.2 Homework	337
17 Line Integrals	339
17.1 Paths	339
17.1.1 Homework	343
17.2 Conservative Force Fields	343
17.2.1 Homework	346
17.3 Potential Functions	347
17.3.1 Homework	349
17.4 Finding Potential Functions	349
17.4.1 Homework	349
17.5 Green's Theorem	349
17.5.1 Homework	354
17.6 Green's Theorem for Images of the unit square	354
17.6.1 Homework	358
17.7 Motivational Notation	358
17.7.1 Homework	359
V Differential Forms	361
18 Differential Forms	363
18.1 One Forms	363
18.1.1 A Simple Example	369
18.2 Exact and Closed 1 - Forms	371
18.3 Two and Three Forms	376
18.4 Exterior Derivatives	378
18.5 Integration of Forms	378
18.6 Green's Theorem Revisited	378
VI Applications	379
19 The Exponential Matrix	381
19.1 The Exponential Matrix	381
19.1.1 Homework	383
19.2 The Jordan Canonical Form	383
19.2.1 Homework	387
19.3 Exponential Matrix Calculations	387
19.3.1 Jordan Block Matrices	390
19.3.2 General Matrices	392
19.3.3 Homework	393
19.4 Applications to Linear ODE	393
19.4.1 The Homogeneous Solution	395
19.4.2 Homework	397
19.5 The Non homogeneous Solution	397
19.5.1 Homework	399
19.6 A Diagonalizable Test Problem	399
19.6.1 Homework	401
19.7 A Non diagonalizable Test Problem	401
19.7.1 Homework	402

19.8 A Non homogeneous Problem	402
19.8.1 Louiville’s Formula	403
19.8.2 Finding the Particular Solution	404
20 Nonlinear Parametric Optimization Theory	407
20.1 The More Precise and Careful Way	408
20.2 Unconstrained Parametric Optimization	411
20.3 Constrained Parametric Optimization	414
20.3.1 Hessian Error Estimates	415
20.3.2 A First Look at Constraint Satisfaction	418
20.3.3 Constraint Satisfaction and the Implicit Function Theorem	419
20.4 Lagrange Multipliers	425
VII Summing It All Up	427
21 Summing It All Up	429
VIII References	431
IX Detailed Indices	435

Part I

Introduction

Chapter 1

Beginning Remarks

The first primer essentially discusses the abstract concepts underlying the study of calculus on the real line. A few higher dimensional concepts are touched on such as the development of rudimentary topology in \mathbb{R}^2 and \mathbb{R}^3 , compactness and the tests for extrema for functions of two variables, but that is not a proper study of calculus concepts in two or more variables. A full discussion of the \mathbb{R}^n based calculus is quite complex and even this second primer can not cover all the important things. The chosen focus here is on differentiation in \mathbb{R}^n and important concepts about mappings from \mathbb{R}^n to \mathbb{R}^m such as the inverse and implicit function theorem and change of variable formulae for multidimensional integration. These topics alone require much discussion and setup. These topics intersect nicely with many other important applied and theoretical areas which are no longer covered in mathematical science curricula. The knowledge here allows a quantitatively inclined biologists and other scientists to more properly develop multivariable nonlinear ODE models for themselves and physicists to learn the proper background to study differential geometry and manifolds among many other applications.

However, this course is just not taught at all anymore. It is material that students at all levels must figure out on their own. Most of the textbooks here are extremely terse and hard to follow as they assume a lot of abstract sophistication from their readers. This text is designed to be a self - study guide to this material. It is also designed to be taught from, but at my institution, it would be very difficult to find the requisite 10 students to register so that the course could be taught. Students who are coming in as Master’s Students generally do not have the ability to allocate a semester course like this to learn this material. Instead, even if they have a solid introduction from the primer on analysis, they typically jump to Primer Three which is an introduction to very abstract concepts such as metric spaces, normed linear spaces and inner products spaces along with many other needed deeper ideas. Such a transition is always problematic and the student is always trying to catch up on the wholes in their background.

Also, many multivariable concepts and the associated theory are used in probability, operations research and optimization to name a few. In those courses, \mathbb{R}^n based ideas must be introduced and used despite the students not having a solid foundational course in such things. Hence, good students are reading about this themselves. This primer is intended to give them a reasonable book to read and study from.

We also provide an nice to line integrals and Green’s Theorem in the plane and a re branding of this material in terms of differential forms. New models of signals applied to a complex system use these sorts of ideas and it is a nice way to plant such a pointer to future things. We also have many examples of how we use these ideas in optimization and approximation algorithms.

Some key features of this approach to explaining this material are

- A very careful discussion of **some** of the many important concepts as we believe it is more important to teach **well** an important subset of this material rather than teaching a large quantity

poorly. Hence, the coverage of the textbook is chosen to provide detailed coverage of an appropriate roadmap through this block of material.

- The discussion and associated online lecture material is also designed for self-study as a greater mathematical maturity and training is desperately needed in the biological sciences and physical sciences at this critical junior - senior college level.
- Many pointers to ideas and concepts that are extensions of what is covered in the text so that students know this is just the first step of the journey.
- The emphasis is on learning how to think as no one can take all the courses they need in a tidy linear order. Hence, the ability to read and assess on their own is an essential skill to foster.
- An emphasis on learning how to understand the consequences of assumptions using a variety of tools to provide the proofs of propositions.
- An emphasis on learning how to learn how to use abstraction as a key skill in solving problems.

In what follows, I will outline a set of five primers on analysis. Each course has a web site for it which contains presentations and other material so feel free to download and use any material on those sites. These courses are

- First Half of Primer One for Analysis: MATH 4530 here. The web site is at [The Primer On Analysis I: Part One Home Page](#) .
- Second Half of Primer One for Analysis: MATH 4540 here. The web site is at [The Primer On Analysis I: Part Two Home Page](#) .
- Second Primer on Analysis: this is material we do not teach: essentially calculus in \mathbb{R}^n ideas. The web site is at [The Primer On Analysis II Home Page](#) .
- Third Primer on Analysis: MATH 8210 here. The web site is at [The Primer On Analysis III: Linear Analysis Home Page](#) .
- Fourth Primer on Analysis: MATH 8220 here. The web site is at [The Primer On Analysis IV: Measure Theory Home Page](#) .
- Fifth Primer on Analysis: MATH 9270 here. The web site is at [The Primer On Analysis V: Functional Analysis and Topology Home Page](#) .

All of these books will be published in 2019 as the set (Peterson (8) 2019), (Peterson (10) 2019), (Peterson (9) 2019), (Peterson (7) 2019) and (Peterson (6) 2019).

1.1 Table of Contents

Part One is concerned with some preliminary material: **Preliminaries**, Chapter 2.

Part Two is devoted to **Linear Mappings** and begins with a careful explanation of the ideas of vector spaces in **Vector Spaces**, Chapter 3. We use the ideas of vector spaces so much it is important to see this structure.

In **Linear Transformations**, Chapter 4, we look in detail at how we manage point cloud data and what linear transformations between finite dimensional vector spaces mean. Then, since the special class of symmetric matrices is so important, we start a detailed discussion of their properties in **Symmetric Matrices**, Chap 5 for the two dimensional case. This includes their use in rotating surfaces

1.1. TABLE OF CONTENTS

5

and ODE models. We then look at how these ideas are applied in linear systems of ODEs in **Matrix Sequences and Ordinary Differential Equations**.

In **Continuity and Topology**, Chapter 7, we discuss continuity and compactness for functions of many variables. We then look at the n dimensional symmetric matrix and derive its eigenvalue and eigenvector structure carefully. This is almost the same proof we would use for the case of a self - adjoint linear operator in functional analysis (see (Peterson (9) 2019) and (Peterson (6) 2019)) but the tools we have now are set in \mathbb{R}^n so we can not do that more general case here.

In **Rotations and Orbital Mechanics**, Chapter 9, we go over a complicated example of how we use two dimensional spaces in a real setting which involves determining the motion of a satellite around the earth. We finish part One, with a full treatment of the determinant function and include a fair bit of code to help you see nontrivial examples.

In **Calculus of Many Variables**, Part Three, we discuss the idea of derivative for functions of many variables in **Differentiability**, Chapter 11. This sets the stage for a treatment of higher dimensional extremal theory, **Multivariable Extremal Theory** in Chapter 12. We then cover the important topics of the inverse and implicit function theorems in **The Inverse and Implicit Function Theorem** in Chapter 13 and finish with how we apply these ideas to some applications in **Linear Approximation Applications**, Chapter 14.

In **Integration**, Part Four, we go over the theory of integration in \mathbb{R}^n which is understandably more difficult than the treatment of Riemann Integration on an interval in \mathbb{R} . This is done in **Integration in Multiple Dimensions**, Chapter 15. We must include a careful discussion of measure zero extended to \mathbb{R}^n and the idea of a multidimensional volume for a subset which is not a rectangle. We then go over **Change of Variables and Fubini's Theorem** in Chapter 16. We finish with a long discussion of line integrals and their connection of Green's Theorem in **Line Integrals**, Chapter 17.

The notion of a line integral is then redone as a differential form in Part Five, **Differential Forms** as Chapter 18.

Applications, Part Six is then about ODE applications again and optimization. In The Exponential Matrix, Chapter 19 we define e^A for a matrix and show how it is useful as a matrix of solutions to a linear system of ODEs. We introduce and use the idea of the Jordan Canonical Form here. We finish the text with a proof of a standard result in Constrained Optimization in **Nonlinear Parametric Optimization Theory**, Chapter 20.

There is much more we could have done, but these topics are a nice introduction into the further use of abstraction in modeling.

Jim Peterson
Department of Mathematical Sciences
Clemson University
July 10, 2018

Part II

Linear Mappings

Chapter 2

Preliminaries

We begin with some preliminaries. We do discuss these things in (Peterson (8) 2019) but a recap is always good and it is nice to keep stuff reasonably self contained.

2.1 The Basics

Let S be a set of real numbers. We say S is bounded if there is a number $M > 0$ so that $x \leq M$ for all $x \in S$. For example, if $S = \{y : y = \tanh(x), x \in \mathbb{R}\}$, then many numbers can play the role of a bound; e.g. $M = 4, 3, 2$ and 1 , among others. Let’s be more precise.

Definition 2.1.1 Bounded Sets

*We say the set S of real numbers is bounded above if there is a number M so that $x \leq M$ for all $x \in S$. Any such number M is called an **upper bound** of S . We usually abbreviate this by the letters **u.b.** or just **ub**.*

*We say this same set S is bounded below if there is a number m so that $x \geq m$ for all $x \in S$. Any such number m is called a **lower bound** of S ; we abbreviate this a **l.b.** or simply **lb**.*

The next idea is more abstract; the idea of a **least upper bound** and **greatest lower bound**. We have to be careful to define this carefully.

Definition 2.1.2 Least Upper Bound and Greatest Lower Bound

*The least upper bound or **lub** or **supremum** or **sup** of a nonempty set S of real numbers is a number U satisfying*

1. *U itself is an upper bound of S*
2. *if M is any other upper bound of S , $U \leq M$.*

*Similarly, the greatest lower bound or **glb** or **infimum** or **inf** of the nonempty set S of real numbers is a number L satisfying*

1. *L itself is a lower bound of S*
2. *if m is any other lower bound of S , $m \leq L$.*

We often use the notation $\inf(S)$ and $\sup(S)$ to denote these numbers.

Note, $\inf(S)$ and $\sup(S)$ need not be in the set S itself!

We use the following conventions:

1. if S has no lower bound, we set $\inf(S) = -\infty$. If S has no upper bound, we set $\sup(S) = \infty$.
2. We can extend these definitions to an empty set which is sometimes nice to do. If $S = \emptyset$, since all positive numbers are not lower bounds, we can argue we should set $\inf(\emptyset) = \infty$. Similarly, we set $\sup(\emptyset) = -\infty$.

When the $\inf(S)$ and $\sup(S)$ belong to S , this situation is special. It is worthy of a definition of a new set of terms.

Definition 2.1.3 The Minimum and Maximum of a Set

*Let S be a set of real numbers. We say Q is a **maximum** of S if $x \leq Q$ for all $x \in S$ and $Q \in S$. Then, we say q is a **minimum** of S when $x \geq q$ for all $x \in S$ and $q \in S$. We also say q is a **minimal element** of S and Q is a **maximal element** of S .*

Note if Q is a maximum of S , by definition, Q is an upper bound of S . Hence, $\sup(S) \leq Q$. On the other hand, since $Q \in S$, we must also have $Q \leq \sup(S)$. We conclude $Q = \sup(S)$. So, when the supremum of a set is achieved by an element of the set, that element is the supremum. We can argue in a similar way to show $m = \inf(S)$.

Now, there are things we can not prove about the set of real numbers \mathbb{R} . Instead we must assume a certain property called the **Completeness Axiom**.

Axiom 1 The Completeness Axiom

Let S be a set of real numbers which is nonempty and bounded above. The $\sup(S)$ exists and is finite. If S is bounded below, then $\inf(S)$ exists and is finite.

So bounded sets of real numbers always have a finite infimum and a finite supremum. Of course, they need not possess a minimum or maximum! We can prove some useful things now.

Theorem 2.1.1 S has a maximal element if and only if sup(S) is in S

*Let S be a nonempty set of real numbers which is bounded above. Then S has a **maximal element** if and only $\sup(S) \in S$.*

Proof 2.1.1

(\Leftarrow):

Assume $\sup(S) \in S$. By definition, $\sup(S)$ is an upper bound of S and so $x \leq \sup(S)$ for all $x \in S$. This shows $\sup(S)$ is the maximum.

(\Rightarrow):

Let Q denote the maximum of S . Then, by definition, $x \leq Q$ for all $x \in S$ which shows Q is an upper bound of S . Then, from the definition of the supremum of S , we must have $\sup(S) \leq Q$. However, $\sup(S)$ is also an upper bound of S and so we also have $Q \leq \sup(S)$. Combining these inequalities, we see $Q = \sup(S)$. ■

We can prove a similar theorem about the minimum.

Theorem 2.1.2 S has a minimal element if and only if inf(S) is in S

*Let S be a nonempty set of real numbers which is bounded below. Then S has a **minimal element** if and only $\inf(S) \in S$.*

Proof 2.1.2

The argument is similar to the one we did for the maximum, so we'll leave this one to you. ■

Now the next two lemmas are very important in modern analysis.

Lemma 2.1.3 Supremum Tolerance Lemma

Let S be a nonempty set that is bounded above. Then we know $\sup(S)$ exists and is finite by the completeness axiom. Moreover, for all $\epsilon > 0$, there is an element $y \in S$ so that $\sup(S) - \epsilon < y \leq \sup(S)$.

Proof 2.1.3

We argue by contradiction. Suppose for a given $\epsilon > 0$, we can not find such a y . Then we must have $y \leq \sup(S) - \epsilon$ for all $y \in S$. This tells us $\sup(S) - \epsilon$ is an upper bound of S and so by definition, $\sup(S) \leq \sup(S) - \epsilon$ which is not possible. Hence, our assumption is wrong and we can find such a y . ■

Lemma 2.1.4 Infimum Tolerance Lemma

Let S be a nonempty set that is bounded below. Then we know $\inf(S)$ exists and is finite by the completeness axiom. Moreover, for all $\epsilon > 0$, there is an element $y \in S$ so that $\inf(S) \leq y \leq \inf(S) + \epsilon$.

Proof 2.1.4

We argue by contradiction. Suppose for a given $\epsilon > 0$, we can not find such a y . Then we must have $y \geq \inf(S) + \epsilon$ for all $y \in S$. This tells us $\inf(S) + \epsilon$ is a lower bound of S and so by definition, $\inf(S) \geq \inf(S) + \epsilon$ which is not possible. Hence, our assumption is wrong and we can find such a y . ■

We often abbreviate these lemma as the **STL** and **IFT**, respectively.

Homework

Exercise 2.1.1**Exercise 2.1.2****Exercise 2.1.3****Exercise 2.1.4****Exercise 2.1.5**

There is also the notion of **compactness** which we have discussed thoroughly in (Peterson (8) 2019). However, let's go over it again as this is an important concept. We begin with **sequential compactness**. Recall

Definition 2.1.4 Sequential Compactness

*Let S be a set of real numbers. We say S is **sequentially compact** if every sequence (x_n) in S has at least one subsequence which converges to an element of S .*

This definition has many consequences. The first is a way to characterize the sequentially compact subsets of the real line.

Theorem 2.1.5 A Set S is Sequentially Compact if and only if it is Closed and Bounded

Let S be a set of real numbers. Then S is sequentially compact if and only if it is a bounded and closed set.

Proof 2.1.5

($\Rightarrow:$) If we assume S is sequentially compact, we must show S is closed and bounded.

- a: To show S is closed, we show S contains its limit points. Let x be a limit point of S . Then there is a sequence (x_n) in S so that $x_n \rightarrow x$. Since (x_n) is in S , the sequential compactness of S says there is a subsequence (x_n^1) of (x_n) and a $y \in S$ so that $x_n^1 \rightarrow y$. But the subsequential limits of a converging sequence must converge to the same limit. Hence, $y = x$. This tells us $x \in S$ and so the limit point x is in S . Since the choice of limit point x was arbitrary, this implies S is a closed set.
- b: To show S is bounded, we assume it is not. Then there must be a sequence (x_n) in S with $|x_n| > n$ for all positive integers n . Now apply the sequential compactness of S again. This sequence must contain a convergent subsequence (x_n^1) which converges to an $x \in S$. But this implies the elements of the sequence (x_n^1) must be bounded as there is a theorem which tells us convergent sequences must be bounded. But this subsequence is from an unbounded sequence which is monotonic in absolute value. This is a contradiction. So our assumption that S is unbounded is wrong and we conclude S must be bounded.

($\Leftarrow:$) We assume S is closed and bounded. We want to show S is sequentially compact. let (x_n) be any sequence in S . Since S is a bounded set, by the Bolzano - Weierstrass Theorem, there is at least one subsequence (x_n^1) in x_n and a real number x with $x_n^1 \rightarrow x$. This says x is a limit point of S and since S is closed, we must have $x \in S$. Hence the sequence (x_n) in S has a subsequence which converges to a point in S . Since our choice of sequence in S is arbitrary, this shows S is sequentially compact. ■

Homework

Exercise 2.1.6

Exercise 2.1.7

Exercise 2.1.8

Exercise 2.1.9

Exercise 2.1.10

Now add continuity to the mix. Here is a typical result.

Theorem 2.1.6 The range of a continuous function on a sequentially compact domain is also sequentially compact

Let S be a nonempty sequentially compact set of real numbers. Assume $f : S \rightarrow \mathbb{R}$ is continuous on S . Then the range $f(S)$ is sequentially compact.

Proof 2.1.6

We know $f(S) = \{f(x) | x \in S\}$. Let (y_n) be sequence in $f(S)$. Then there is an associated sequence (x_n) in S so that $f(x_n) = y_n$. But S is sequentially compact, so there is a subsequence x_n^1 which converges to a point $x \in S$. Since f is continuous at x , $f(x_n^1) \rightarrow f(x)$. Let y denote the number $f(x)$. We have shown there is a subsequence $(y_n^1 = f(x_n^1))$ with $y_n^1 \rightarrow y$ with $y \in f(S)$. Hence, we have shown $f(S)$ is sequentially compact. ■

This leads to our first **extremal theorem**.

Theorem 2.1.7 Continuous functions on a sequentially compact domains have a minimum and maximum value

Let S be a sequentially compact set of real numbers and let $f : S \rightarrow \mathbb{R}$ be continuous on S . Then there are two sequences x_n^m and x_n^M which converge to the points x_m and x_M respectively and $f(x) \geq f(x_m)$ and $f(x) \leq f(x_M)$ for all $x \in S$. Thus, $f(x_m)$ is the minimum value of f on S which is achieved at x_m . Moreover, $f(x_M)$ is the maximum value of f on S which is achieved at x_M .

We call the sequences (x_n^m) and (x_n^M) minimizing and maximizing sequences for f on S .

Proof 2.1.7

By Theorem 2.1.6, $f(S)$ is sequentially compact and nonempty. Hence, $f(S)$ is a closed and bounded set of real numbers by Theorem 2.1.5. We thus know $M = \sup(f(S))$ and $m = \inf(f(S))$ both exist and are finite. By the IFT and SFT, satisfying

$$\begin{aligned} \exists x_1^M \in S \ni M - 1 &< f(x_1^M) \leq M \\ \exists x_2^M \in S \ni M - 1/2 &< f(x_2^M) \leq M \\ \exists x_3^M \in S \ni M - 1/3 &< f(x_3^M) \leq M \\ &\vdots \\ \exists x_n^M \in S \ni M - 1/n &< f(x_n^M) \leq M \\ &\vdots \end{aligned}$$

and

$$\begin{aligned} \exists x_1^m \in S \ni m &\leq f(x_1^m) < m + 1 \\ \exists x_2^m \in S \ni m &\leq f(x_2^m) < m + 1/2 \\ \exists x_3^m \in S \ni m &\leq f(x_3^m) < m + 1/3 \\ &\vdots \\ \exists x_n^m \in S \ni m &\leq f(x_n^m) < m + 1/n \\ &\vdots \end{aligned}$$

We see $f(x_n^M) \rightarrow M$ and $f(x_n^m) \rightarrow m$. Hence $(f(x_n^M))$ and $(f(x_n^m))$ are convergent sequences in the closed set $f(S)$. So their limit values must be in $f(S)$. Hence, we know $m \in f(S)$ and $M \in f(S)$. Now apply Theorem 2.1.1 and Theorem 2.1.2 to see there are elements x_m and x_M in S so that $f(x_m) = m \leq f(x)$ and $f(x_M) = M \geq f(x)$ for all $x \in S$. ■

Homework

Exercise 2.1.11

Exercise 2.1.12

Exercise 2.1.13

Exercise 2.1.14

Exercise 2.1.15

Now add differentiation.

Theorem 2.1.8 Derivative of f at an interior point local extremum is zero

Let $f : [a, b] \rightarrow \mathbb{R}$ where $[a, b]$ is a finite interval. Assume f has a local extrema at an interior point $p \in [a, b]$. Then if f is differentiable at p , $f'(p) = 0$.

Proof 2.1.8

Since p is an interior point, there is an $r_1 > 0$ so that $-x - p| < r_1 \implies x \in [a, b]$. Also, since $f(p)$ is a local extreme value, there is another $r_2 > 0$ so that $f(p) \leq f(x)$ for $x \in B_{r_2}(p)$ with $B_{r_2}(p) \subset [a, b]$ if $f(p)$ is a local maximum or $f(p) \geq f(x)$ for $x \in B_{r_2}(p)$ with $B_{r_2}(p) \subset [a, b]$ if $f(p)$ is a local minimum. For convenience, we will assume $f(p)$ is a local maximum and let you argue the local minimum case for yourself.

Combining these facts, we see

$$|x - p| < r < \min(r_1, r_2) \implies f(x) \geq f(p)$$

Thus, for $p < x < p + r$, $f(x) - f(p) \leq 0$ and $x - p > 0$ telling us the ratio $\frac{f(x) - f(p)}{x - p} \leq 0$. Hence, $f'(p^+) = \lim_{x \rightarrow p^+} \frac{f(x) - f(p)}{x - p} \leq 0$. Since f is differentiable at p , we know $f'(p^+) = f'(p)$. Thus, we have $f'(p) \leq 0$.

Next, for $p - r < x < p$, we have $f(x) - f(p) \leq 0$ and $x - p < 0$. Thus, the ratio $\frac{f(x) - f(p)}{x - p} \geq 0$. Hence, $f'(p^-) = \lim_{x \rightarrow p^-} \frac{f(x) - f(p)}{x - p} \geq 0$. Since f is differentiable at p , we know $f'(p^-) = f'(p)$. Thus, we have $f'(p) \geq 0$.

Combining, we see $f'(p) = 0$. You can do a similar argument for the case of a local minimum. ■

Homework

Exercise 2.1.16**Exercise 2.1.17****Exercise 2.1.18****Exercise 2.1.19****Exercise 2.1.20**

We will often be in situations where the classical limits fail to exist. However, even if this is the case, **limit inferiors** and **limit superiors** do exist! Let (a_n) be a sequence of real numbers. Let (a_n^1) be any subsequence of (a_n) which converges. If (a_n) is a bounded sequence, the Bolzano - Weierstrass Theorem guarantees there is at least one such subsequence. Consider the set

$$S = \{\alpha \in \mathbb{R} | \exists (a_n^1) \subset (a_n) \ni a_n^1 \rightarrow \alpha\}$$

The set S is called the **set of subsequential limits** of the sequence (a_n) . We define the limit inferior and limit superior of a sequence using S .

Definition 2.1.5 Limit Inferior and Limit Superior of a Sequence

*The **limit inferior** of the sequence (a_n) is denoted by $\liminf(a_n) = \underline{\lim}(a_n)$ and defined by $\liminf(a_n) = \inf(S)$. In a similar fashion, the **limit superior** of the sequence (a_n) is denoted by $\limsup(a_n) = \overline{\lim}(a_n)$ and defined by $\limsup(a_n) = \sup(S)$.*

2.1. THE BASICS

15

Recall, it is possible for a sequence to converge to $\pm\infty$. For example, if $(a_n) = (n)$ for $n \geq 1$, we know the sequence is not bounded above and $\lim_n(a_n) = \lim_n(n) = \infty$. Similarly, if $(a_n) = (-n)$ for $n \geq 1$, we know the sequence is not bounded below and $\lim_n(a_n) = \lim_n(-n) = -\infty$. There are many other examples! Thus, it is possible to have sequences whose limits are not finite real numbers. The set of **extended real numbers** is denoted $\bar{\mathbb{R}}$ and consists of all finite real numbers plus two additional symbols $+\infty$ and $-\infty$. Note, we use these symbols to connect with our intuition about *converging to* $\pm\infty$ but we could just as well used the symbols α and β . But for us, we have defined

$$\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$$

We use the following conventions: for addition we have

$$\begin{aligned} x + (+\infty) &= +\infty, \forall x \in \mathbb{R} \\ x - (+\infty) &= -\infty, \forall x \in \mathbb{R} \\ (+\infty) + x &= +\infty, \forall x \in \mathbb{R} \\ (-\infty) + x &= -\infty, \forall x \in \mathbb{R} \\ (-\infty) + (-\infty) &= -\infty, \\ (+\infty) + (+\infty) &= +\infty \end{aligned}$$

For multiplication,

$$\begin{aligned} (+\infty) \cdot x &= x \cdot (+\infty) = +\infty, \forall x > 0 \\ (+\infty) \cdot 0 &= 0 \cdot (+\infty) = 0, \\ (+\infty) \cdot x &= x \cdot (+\infty) = -\infty, \forall x < 0 \\ (-\infty) \cdot x &= x \cdot (-\infty) = -\infty, \forall x > 0 \\ (-\infty) \cdot 0 &= 0 \cdot (-\infty) = 0, \\ (-\infty) \cdot x &= x \cdot (-\infty) = +\infty, \forall x < 0 \\ (+\infty) \cdot (+\infty) &= +\infty \\ (-\infty) \cdot (-\infty) &= +\infty \\ (+\infty) \cdot (-\infty) &= -\infty \\ (-\infty) \cdot (+\infty) &= -\infty \end{aligned}$$

For division

$$\begin{aligned} \frac{x}{+\infty} &= 0, \forall x \in \mathbb{R} \\ \frac{x}{-\infty} &= 0, \forall x \in \mathbb{R} \\ \frac{+\infty}{x} &= +\infty, \forall x > 0 \\ \frac{+\infty}{x} &= -\infty, \forall x < 0 \\ \frac{-\infty}{x} &= -\infty, \forall x > 0 \\ \frac{-\infty}{x} &= \infty, \forall x < 0 \end{aligned}$$

There are several operations that are not defined. These are $0/0$, $(+\infty + (-\infty))$, $(+\infty)/(-\infty)$, $(-\infty)/(+\infty)$, $(+\infty)/(+\infty)$ and $(-\infty)/(-\infty)$. However, with the above conventions, arithmetic

in $\bar{\mathfrak{R}}$ is well-defined. With these definitions, we can extend limit inferiors and limit superiors to the extended reals. Let (a_n) be any sequence in $\bar{\mathfrak{R}}$. The set of subsequential limits now is

$$S = \{\alpha \in \bar{\mathfrak{R}} | \exists (a_n^1) \subset (a_n) \ni a_n^1 \rightarrow \alpha\}$$

where we now allow the possibility that $\pm\infty \in S$. We need to think carefully about limiting behavior here.

- If (a_n^1) is a subsequence which is in the reals, then $a_n^1 \rightarrow +\infty$ implies for sufficiently large n^1 , $1/a_n^1 \rightarrow 0$. This means for any $\epsilon > 0$, there is an N so that $n^1 > N$ implies $|1/a_n^1| < \epsilon$ or $|a_n^1| > 1/\epsilon$. As $\epsilon \rightarrow 0^+$, $1/\epsilon \rightarrow +\infty$. So this is the same as saying for all $R > 0$, there is an N so that $|a_n^1| > R$ when $n^1 > N$.
- If the subsequence $(a_n^1) \rightarrow -\infty$, this is handled like we did above.
- If the subsequence has an infinite number of $+\infty$ values, then our tests for convergence which require us to look at the difference $a_n^1 - \alpha$ where α is the limit fail when we look at the term $\infty - \infty$ which is not a defined operation. In this case, we will simply posit that $\alpha = \infty$ or $\alpha = -\infty$ depending on the structure of the subsequence.

Here are some examples:

1. $(a_n) = (\tanh(n))$. Then $\underline{\lim}(a_n) = -1$ and $\overline{\lim}(a_n) = +1$.
2. $(a_n) = (n^2)$. Then $\underline{\lim}(a_n) = +\infty$ and $\overline{\lim}(a_n) = +\infty$.
3. $(a_n) = (\cos(n\pi/3))$ for $n \geq 0$. Then the range of this sequence is a set of repeating blocks

$$(a_n) = \{\textbf{Block}, \textbf{Block}, \dots\}$$

where

$$\textbf{Block} = \{1(n=0), 1/2(n=1), -1/2(n=2), -1(n=3), -1/2(n=4), 1/2(n=5)\}$$

We see for integers k , we have subsequences that converges to each element in this block:

$$\begin{aligned} \cos((6k+0)\pi/3) &= 1, \implies a_{6k+0} \rightarrow 1 \\ \cos((6k+1)\pi/3) &= 1/2, \implies a_{6k+1} \rightarrow 1/2 \\ \cos((6k+2)\pi/3) &= -1/2, \implies a_{6k+2} \rightarrow -1/2 \\ \cos((6k+3)\pi/3) &= -1, \implies a_{6k+3} \rightarrow -1 \\ \cos((6k+4)\pi/3) &= -1/2, \implies a_{6k+4} \rightarrow -1/2 \\ \cos((6k+5)\pi/3) &= 1/2, \implies a_{6k+5} \rightarrow 1/2 \end{aligned}$$

A little thought then allows us to conclude $S = \{-1, -1/2, 1/2, 1\}$ and so $\underline{\lim}(a_n) = -1$ and $\overline{\lim}(a_n) = +1$.

We can extend these ideas to functional limits. Consider the function f defined on a finite interval $[a, b]$. Let $p \in [a, b]$. we say $\alpha \in \bar{\mathfrak{R}}$ is a **cluster point** of f as we approach p , if there is a sequence (x_n) in $[a, b]$ with $x_n \rightarrow p$ and $f(x_n) \rightarrow \alpha$. We must now think of convergence in the sense of the extended reals, of course. We define the set of cluster points $S(p)$ then as

$$S(p) = \{\exists (x_n) \subset [a, b] \ni x_n \rightarrow p \text{ and } f(x_n) \rightarrow \alpha\}$$

and we define the limit inferior as x approaches p of f to be $\underline{\lim}_{x \rightarrow p} f(x) = \inf(S(p))$. Similarly, the limit superior as x approaches p of f is $\overline{\lim}_{x \rightarrow p} f(x) = \sup(S(p))$. In this version of these limits, we do not have an unbounded domain and so we cannot have a sequence $x_n \rightarrow \pm\infty$. The most general version has the domain of f simply being the set Ω and

$$S(p) = \{\exists(x_n) \subset \Omega \ni x_n \rightarrow p \text{ and } f(x_n) \rightarrow \alpha\}$$

In this case, we can have the extended real value limits $\pm\infty$ but the definitions of the limits above are still the same. We also use the notation $\underline{\lim}_{x \rightarrow p} f(x) = \liminf_{x \rightarrow p} f(x)$ and $\overline{\lim}_{x \rightarrow p} f(x) = \limsup_{x \rightarrow p} f(x)$.

Example 2.1.1 Find the cluster points of $f(x) = \cos(1/x)$ for $x \neq 0$.

Solution Here f is defined in \mathfrak{R} not $\bar{\mathfrak{R}}$. Since $-1 \leq \cos(y) \leq 1$, given any $\alpha \in [-1, 1]$, the sequence (s_n) defined by $x_n = 1/(2n\pi + \cos^{-1}(\alpha))$ satisfies $x_n \rightarrow 0$ and $f(x_n) = \alpha$ for all n . Hence, $S(0) = [-1, 1]$ and $\underline{\lim}_{x \rightarrow 0} f(x) = \inf(S(0)) = -1$ and $\overline{\lim}_{x \rightarrow 0} f(x) = \sup(S(0)) = 1$. Because f is continuous at all $p \neq 0$, $S(p) = \{\cos(1/p)\}$ there and $\underline{\lim}_{x \rightarrow p} f(x) = \inf(S(p)) = \cos(1/p)$ and $\overline{\lim}_{x \rightarrow p} f(x) = \sup(S(p)) = \cos(1/p)$. Hence, $\lim_{x \rightarrow p} f(x) = f(p)$ at those points.

We can prove an alternate definition of the limit inferior and limit superior of a sequence.

Theorem 2.1.9 Alternate Definition of the limit inferior and limit superior of a sequence

Let (a_n) be a sequence of real numbers. Then

$$\underline{\lim}(a_n) = \sup_k \inf_{n \geq k} a_n \text{ and } \overline{\lim}(a_n) = \inf_k \sup_{n \geq k} a_n$$

To see this clearer, let

$$z_n = \inf\{a_k, a_{k+1}, \dots\} = \inf_{n \geq k} a_n \text{ and } w_n = \sup\{a_k, a_{k+1}, \dots\} = \sup_{n \geq k} a_n$$

Then,

$$z_1 \leq z_2 \leq \dots \leq z_n \leq \dots \leq w_n \leq w_{n-1} \leq \dots \leq w_1$$

The (z_n) sequence is monotone increasing and the (w_n) sequence is monotone decreasing. Hence, $\lim_n z_n \uparrow \underline{\lim}(a_n)$ and $\lim_n w_n \downarrow \overline{\lim}(a_n)$.

Proof 2.1.9

Careful proofs of these ideas are in (Peterson (8) 2019). You should go back and review them. ■

Homework

Exercise 2.1.21

Exercise 2.1.22

Exercise 2.1.23

Exercise 2.1.24

Exercise 2.1.25

2.2 Some Topology in \mathbb{R}^2

If you look at Figure 2.1, you can see the big difference between simple set topology in \mathbb{R} and in \mathbb{R}^2 .

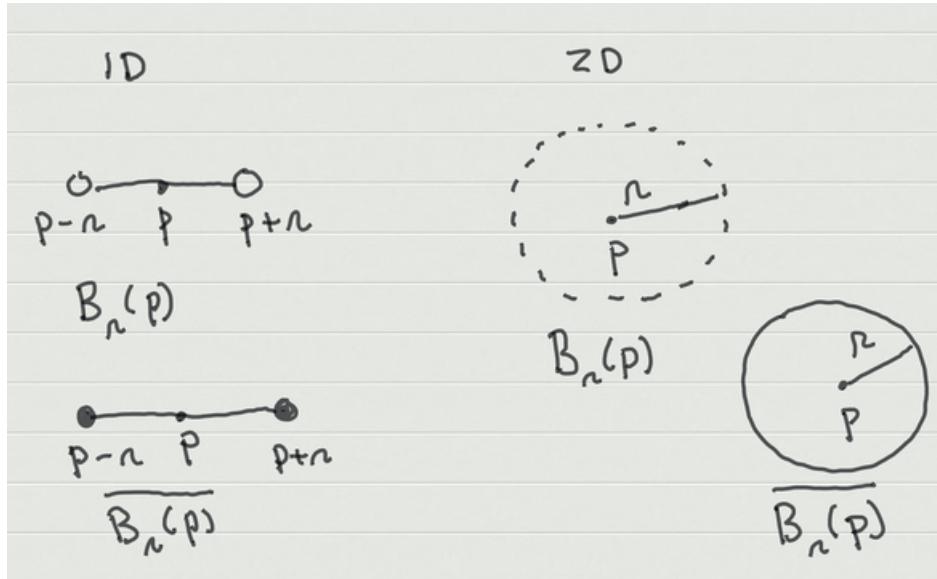


Figure 2.1: Open and Closed Balls in One and Two D

The open and closed balls in \mathbb{R} are simply open intervals on the x -axis. You draw the interval centered at p and then it is an open ball if the endpoints are not included and a closed ball otherwise. We define this as $B_r(p) = \{x \in \mathbb{R} : p - r \leq x \leq p + r\}$. In \mathbb{R}^2 , the open intervals become 2D circles centered at a point in \mathbb{R}^2 . Thus, $B_r(p) = \{x \in \mathbb{R}^2 : \|x - p\| \leq r\}$ where x is the vector $x = [x_1 // x_2]$ and the center is the vector $p = [p_1 // p_2]$. Get familiar with this by drawing lots of pictures. From our discussions of topology in \mathbb{R} in (Peterson (8) 2019), the important concepts to move into \mathbb{R}^2 are

1. If S is a subset of \mathbb{R}^2 , S^C is the **complement** of S and consists of all x not in S .
2. p is an **interior point** of the set S in \mathbb{R}^2 if there is a positive r so that $B_r(p) \subset S$. We define S to be an **open set** if all of its points are **interior points**. Note a curve in \mathbb{R}^2 has no interior! Draw the graph of $y = x^2$ for $x \in [-1, 1]$ and this arc has no interior points.
3. p is a **boundary point** of S if all balls $B_r(p)$ intersect both S and S^C in at least one point. In Figure 2.2, we see boundary point examples in both \mathbb{R} and \mathbb{R}^2 .
4. A set S is said to be **closed** if S^C is open.

A sequence in \mathbb{R}^2 is a set of two dimensional vectors

$$\left(\begin{bmatrix} x_n \\ y_n \end{bmatrix} \right)$$

We say this sequence converges to the vector $\begin{bmatrix} x \\ y \end{bmatrix}$ if

$$\forall \epsilon > 0, \exists N \ni \|X_n - X\| < \epsilon, \forall n > N$$

2.2. SOME TOPOLOGY IN \mathbb{R}^2

19

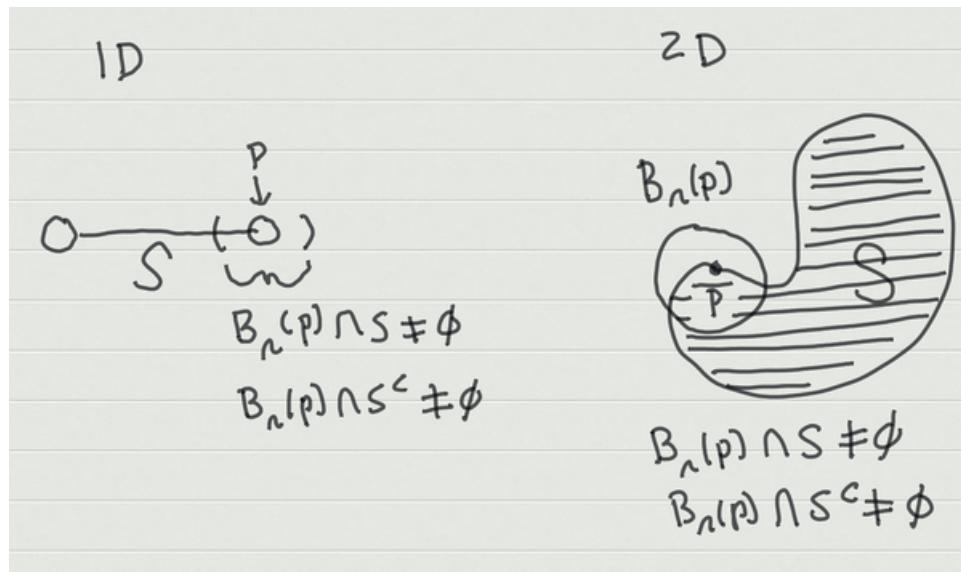


Figure 2.2: Boundary Points in One and Two D

where we let $\mathbf{X}_n = \begin{pmatrix} x_n \\ y_n \end{pmatrix}$ and $\mathbf{X} = \begin{pmatrix} x \\ y \end{pmatrix}$. Further, we use the norm

$$\|\mathbf{X}_n - \mathbf{X}\| = \sqrt{(x_n - x)^2 + (y_n - y)^2}$$

as our way to measuring the distance between vectors in \mathbb{R}^2 . Of course, other norms are possible but we will simply use the standard Euclidean norm for now. Now that we know what a sequence of vectors is and what we mean by the sequence converging, we can define some other special points.

1. A point p is an **accumulation point** of S if every ball $B_r(p)$ contains a point of S different from p . This means there is a sequence (\mathbf{x}_n) from $B_{1/n}(p) \setminus \{p\}$ which converges to p .
2. A point p is a **cluster point** of S if there is a sequence (\mathbf{x}_n) from S with each $\mathbf{x}_n \neq p$ so that $\mathbf{x}_n \rightarrow p$. Note cluster points are accumulation points and accumulation points are cluster points.
3. A point p is a **limit point** of S if there is a sequence (\mathbf{x}_n) from S that converges to p . Note an **isolated point** of a set is a point which is a positive distance from any other point in the set and an isolated point is a limit point but not an accumulation point.
4. S' is the set S and its limit points.

Note a set S need not be open or closed and \mathbb{R}^2 is both open and closed. We can prove useful things essentially in the same way as we did in \mathbb{R} .

Theorem 2.2.1 Basic Topology Results in \mathbb{R}^2

Given S in \mathbb{R}^2 :

1. S is closed if and only if $S = S'$; i.e. S is closed if and only if it contains its limit points.
2. S is closed if and only if S contains its boundary points.
3. Finite intersections of open sets are open.
4. Finite unions of open sets are open.
5. Countable unions of open sets are open. Countable intersections of open sets can be closed.
6. Finite intersections of closed sets are closed.
7. Finite unions of closed sets are closed.
8. Countable intersections of closed sets are closed. Countable unions of closed sets need not be open.
9. DeMorgan's Laws hold: i.e. $(\bigcup_n S_n)^C = \bigcap_n S_n^C$ and $(\bigcap_n S_n)^C = \bigcup_n S_n^C$

Proof 2.2.1

You can mimic the one dimensional proofs fairly easily. ■

2.2.1 Homework

Exercise 2.2.1

Exercise 2.2.2

Exercise 2.2.3

Exercise 2.2.4

Exercise 2.2.5

2.3 Bolzano - Weierstrass in \mathbb{R}^2

We have discussed the Bolzano - Weierstrass theorem very carefully in \mathbb{R} and we mentioned casually how we would go about handling the proof in \mathbb{R}^n for $n \geq 2$ in (Peterson (8) 2019), but now we would like to show the details for the 2D proof. The proof is then very similar in 3D and so on and we just refer to this one below and say, “well, it is similar...”!

Theorem 2.3.1 Bolzano - Weierstrass Theorem in 2D

Every bounded sequence in \mathbb{R}^2 has at least one convergent subsequence.

Proof 2.3.1

Let's assume the range of this sequence is infinite. If it were finite, there would be subsequences of it that converge to each of the values in the finite range. To make it easier to read this argument, we will use boldface font to indicate vectors in \mathbb{R}^2 and it is assumed they will have two

components labeled with subscripts 1 and 2. So here the sequence is (\mathbf{a}_n) with $\mathbf{a}_n = \begin{bmatrix} a_{n1} \\ a_{n2} \end{bmatrix}$. We assume the sequences start at $n = 1$ for convenience and by assumption, there is a positive number B so that $\|\mathbf{a}_n\| \leq B/2$ for all $n \geq 1$. Hence, all the components of this sequence live in the square $[-B/2, B/2] \times [-B/2, B/2] \subset \mathbb{R}^2$. Let's use a special notation for a **box** of this sort. Let $J_0 = I_0^1 \times I_0^2$ where the interval $I_0^1 = [\alpha_{01}, \beta_{01}]$ and $I_0^2 = [\alpha_{02}, \beta_{02}]$ also. Here, $\alpha_{0i} = -B/2$ and $\beta_{0i} = B/2$. Note the area of J_0 , denoted by ℓ_0 is B^2 .

Let \mathcal{S} be the range of the sequence which has infinitely many points and for convenience, we will let the phrase infinitely many points be abbreviated to **IMPs**.

Step 1:

Bisect each axis interval of J_0 into two pieces giving four subregions of J_0 all of which have area $B^2/4$. Now at least one of the subregions contains IMPs of \mathcal{S} as otherwise each subregion has only finitely many points and that contradicts our assumption that \mathcal{S} has IMPs. Now all may contain IMPs so select one such subregion containing IMPs and call it J_1 . Then $J_1 = I_1^1 \times I_1^2$ where the interval $I_1^1 = [\alpha_{11}, \beta_{11}]$ and $I_1^2 = [\alpha_{12}, \beta_{12}]$ also. Note the area of J_1 , denoted by ℓ_1 is $B^2/4$. We see $J_1 \subset J_0$ and

$$-B/2 = \alpha_{0i} \leq \alpha_{1i} \leq \beta_{1i} \leq \beta_{0i} = B/2$$

Since J_1 contains IMPs, we can select a sequence vector \mathbf{a}_{n_1} from J_1 .

Step 2:

Now subdivide J_1 into four subregions just as before. At least one of these subregions contain IMPs of \mathcal{S} . Choose one such subregion and call it J_2 . Then $J_2 = I_2^1 \times I_2^2$ where the interval $I_2^1 = [\alpha_{21}, \beta_{21}]$ and $I_2^2 = [\alpha_{22}, \beta_{22}]$ also. Note the area of J_2 , denoted by ℓ_2 is $B^2/(16)$.

We see $J_2 \subset J_1$ and

$$-B/2 = \alpha_{0i} \leq \alpha_{1i} \leq \alpha_{2i} \leq \beta_{2i} \leq \beta_{1i} \leq \beta_{0i} = B/2$$

Since J_2 contains IMPs, we can select a sequence vector \mathbf{a}_{n_2} from J_2 . It is easy to see this value can be chosen different from \mathbf{a}_{n_1} , our previous choice.

You should be able to see that we can continue this argument using induction.

Proposition:

$\forall p \geq 1, \exists$ an interval $J_p = I_p^1 \times I_p^2$ with $I_p^i = [\alpha_{pi}, \beta_{pi}]$ with the area of J_p , $\ell_p = B^2/(2^{2p})$ satisfying $J_p \subseteq J_{p-1}$, J_p contains IMPs of \mathcal{S} and

$$\alpha_{0i} \leq \dots \leq \alpha_{p-1,i} \leq \alpha_{pi} \leq \beta_{pi} \leq \beta_{p-1,i} \leq \dots \leq \beta_{0i}$$

Finally, there is a sequence vector \mathbf{a}_{n_p} in J_p , different from $\mathbf{a}_{n_1}, \dots, \mathbf{a}_{n_{p-1}}$.

Proof We have already established the proposition is true for the basis step J_1 and indeed also for the next step J_2 .

Inductive: We assume the interval J_q exists with all the desired properties. Since by assumption, J_q contains IMPs, bisect J_q into four subregions as we have done before. At least one of these subregions contains IMPs of \mathcal{S} . Choose one of the subregions and call it J_{q+1} and label $J_{q+1} = I_{q+1}^1 \times I_{q+1}^2$ where $I_{q+1}^i = [\alpha_{q+1,i}, \beta_{q+1,i}]$ for $i = 1, 2$. We see immediately $\ell_{q+1} = B^2/2^{2(q+1)}$ with $\alpha_q^i \leq \alpha_{q+1}^i \leq \beta_{q+1}^i \leq \beta_q^i$. This shows the nested inequality we want is satisfied. Finally, since J_{q+1} contains IMPs, we can choose $\mathbf{a}_{n_{q+1}}$ distinct from the other \mathbf{a}_{n_i} 's. So the inductive step is satisfied and by the POMI, the proposition is true for all n . \square

From our proposition, we have proven the existence of sequences, (α_{pi}) , (β_{pi}) and (ℓ_p) which have various properties. The sequence ℓ_p satisfies $\ell_p = (1/4)\ell_{p-1}$ for all $p \geq 1$. Since $\ell_0 = B^2$, this means $\ell_1 = B^2/4$, $\ell_2 = B^2/16$, $\ell_3 = B^2/(2^2)^3$ leading to $\ell_p = B^2/(2^2)^p$ for $p \geq 1$. Further, we have the inequality chain

$$\begin{aligned} -B/2 &= \alpha_{0i} \leq \alpha_{1i} \leq \alpha_{2i} \leq \dots \leq \alpha_{pi} \\ &\leq \dots \leq \\ \beta_{pi} &\leq \dots \leq \beta_{2i} \leq \dots \leq \beta_{0i} = B/2 \end{aligned}$$

The rest of this argument is almost identical to the one we did for the case of a bounded sequence in \mathbb{R} . Note (α_{pi}) is bounded above by $B/2$ and $(\beta_{pi})_{p \geq 0}$ is bounded below by $-B/2$. Hence, by the completeness axiom, $\inf(\beta_{pi})$ exists and equals the finite number β^i ; also $\sup(\alpha_{pi})$ exists and is the finite number α^i .

So if we fix p , it should be clear the number β_{pi} is an upper bound for all the α_{pi} values (look at our inequality chain again and think about this). Thus β_{pi} is an upper bound for (α_{pi}) and so by definition of a supremum, $\alpha^i \leq \beta_{pi}$ for all p . Of course, we also know since α^i is a supremum, that $\alpha_{pi} \leq \alpha^i$. Thus, $\alpha_{pi} \leq \alpha^i \leq \beta_{pi}$ for all p . A similar argument shows if we fix p , the number α_p^i is an lower bound for all the β_{pi} values and so by definition of an infimum, $\alpha_{pi} \leq \beta^i \leq \beta_{pi}$ for all the α_{pi} values. This tells us α^i and β^i are in $[\alpha_{pi}, \beta_{pi}]$ for all p . Next we show $\alpha^i = \beta^i$.

Let $\epsilon > 0$ be arbitrary. Since α^i and β^i are in an interval whose length is $\ell_p = (1/2^{2p})B^2$, we have $|\alpha^i - \beta^i| \leq (1/2^{2p})B^2$. Pick P so that $1/(2^{2P}B^2) < \epsilon$. Then $|\alpha^i - \beta^i| < \epsilon$. But $\epsilon > 0$ is arbitrary.

Hence, $\alpha^i - \beta^i = 0$ implying $\alpha^i = \beta^i$. Finally, define the vector $\theta = \begin{bmatrix} \alpha^0 \\ \alpha^1 \end{bmatrix}$.

We now must show $a_{n_k} \rightarrow \theta$. This shows we have found a subsequence which converges to θ . We know $\alpha_{pi} \leq a_{n_p}^i \leq \beta_{pi}$ and $\alpha_{pi} \leq \alpha^i \leq \beta_{pi}$ for all p . Pick $\epsilon > 0$ arbitrarily. Given any p , we have

$$\begin{aligned} |a_{n_p,1}^i - \alpha^i| &= |a_{n_p,i}^i - \alpha_{pi} + \alpha_{pi} - \alpha^i|, \quad \text{add and subtract trick} \\ &\leq |a_{n_p,i}^i - \alpha_{pi}| + |\alpha_{pi} - \alpha^i| \quad \text{triangle inequality} \\ &\leq |\beta_{pi} - \alpha_{pi}| + |\alpha_{pi} - \beta_{pi}| \quad \text{definition of length} \\ &= 2|\beta_{pi} - \alpha_{pi}| = 2(1/2^{2p})B^2. \end{aligned}$$

Choose P so that $(1/2^{2P})B^2 < \epsilon/2$. Then, $p > P$ implies $|a_{n_p,1}^i - \alpha^i| < 2\epsilon/2 = \epsilon$. Thus, $a_{n_k,i} \rightarrow \alpha^i$. This shows the subsequence converges to θ . ■

Note this argument is messy but quite similar to the one dimensional case. It should be easy for you to see how to extend this to \mathbb{R}^3 and even \mathbb{R}^n . It is more a problem of correct labeling than intellectual difficulty!

2.3.1 Homework

Exercise 2.3.1

Exercise 2.3.2

Exercise 2.3.3

Exercise 2.3.4

Exercise 2.3.5

Chapter 3

Vector Spaces

We are now going to explore vector spaces in some generality.

3.1 Introduction

Let's go back and think about vectors in \mathbb{R}^2 . As you know, we think of these as arrows with a tail fixed at the origin of the two dimensional coordinate system we call the $x - y$ plane. They also have a length or magnitude and this arrow makes an angle with the positive x axis. Suppose we look at two such vectors, \mathbf{E} and \mathbf{F} . Each vector has an x and a y component so that we can write

$$\mathbf{E} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} c \\ d \end{bmatrix}$$

The cosine of the angle between them is proportional to the inner product $\langle \mathbf{E}, \mathbf{F} \rangle = ac + bd$. If this angle is 0 or π , the two vectors lie along the same line. In any case, the angle associated with \mathbf{E} is $\tan^{-1}(\frac{b}{a})$ and for \mathbf{F} , $\tan^{-1}(\frac{d}{c})$. Hence, if the two vectors lie on the same line, \mathbf{E} must be a multiple of \mathbf{F} . This means there is a number β so that

$$\mathbf{E} = \beta \mathbf{F}.$$

We can rewrite this as

$$\begin{bmatrix} a \\ b \end{bmatrix} = \beta \begin{bmatrix} c \\ d \end{bmatrix}$$

Now let the number 1 in front of \mathbf{E} be called $-\alpha$. Then the fact that \mathbf{E} and \mathbf{F} lie on the same line implies there are 2 constants α and β , both not zero, so that

$$\alpha \mathbf{E} + \beta \mathbf{F} = \mathbf{0}.$$

where $\mathbf{0}$ is the zero vector. Note we could argue this way for vectors in \mathbb{R}^3 and even in \mathbb{R}^n . Of course, our ability to think of these things in terms of lying on the same line and so forth needs to be extended to situations we can no longer draw, but the idea is essentially the same. Instead of thinking of our two vectors as lying on the same line or not, we can *rethink* what is happening here and try to identify what is happening in a more abstract way. If our two vectors lie on the same line, they are not *independent* things in the sense one is a multiple of the other. As we saw above, this implies there was a linear equation connecting the two vectors which had to add up to 0. Hence, we might say the vectors were *not linearly independent* or simply, they are *linearly dependent*. Phrased this way, we

are on to a way of stating this idea which can be used in many more situations. We state this as a definition. We will be general here: we are looking at objects of some *type* which we can add and scalar multiply. For the moment, think of these as vectors in say \mathbb{R}^2 but soon enough we will make this more abstract.

Definition 3.1.1 Two Linearly Independent Objects

*Let E and F be two mathematical objects for which addition and scalar multiplication is defined. We say E and F are **linearly dependent** if we can find non zero constants α and β so that*

$$\alpha E + \beta F = 0.$$

*Otherwise, we say they are **linearly independent**.*

We can then easily extend this idea to any finite collection of such objects as follows.

Definition 3.1.2 Finitely many Linearly Independent Objects

*Let $\{E_i : 1 \leq i \leq N\}$ be N mathematical objects for which addition and scalar multiplication is defined. We say E and F are **linearly dependent** if we can find non zero constants α_1 to α_N , not all 0, so that*

$$\alpha_1 E_1 + \dots + \alpha_N E_N = 0.$$

Note we have changed the way we define the constants a bit. When there are more than two objects involved, we can't say, in general, that all of the constants must be non zero.

3.2 Vector Spaces over a Field

If we have a set of objects u with a way to add them to create new objects in the set and a way to *scale* them to make new objects, this is formally called a **Vector Space** with the set denoted by V . For our purposes, we scale such objects with either real or complex numbers. If the scalars are real numbers, we say v is a vector space over the reals; otherwise, it is a vector space over the complex field. There are other fields, of course, but we will focus on these two.

Definition 3.2.1 Abstract Vector Space

Let \mathbf{V} be a set of objects \mathbf{u} with an additive operation \oplus and a scaling method \odot . Formally, this means

VS1: Given any \mathbf{u} and \mathbf{v} , the operation of adding them together is written $\mathbf{u} \oplus \mathbf{v}$ and results in the creation of a new object in the vector space. This operation is commutative which means the order of the operation is not important; so $\mathbf{u} \oplus \mathbf{v}$ and $\mathbf{v} \oplus \mathbf{u}$ give the same result. Also, this operation is associative as we can group any two objects together first, perform this addition \oplus and then do the others and the order of the grouping does not matter.

VS2: Given any \mathbf{u} and any number c (either real or complex, depending on the type of vector space we have), the operation $c \odot \mathbf{u}$ creates a new object. We call such numbers scalars.

VS3: The scaling and additive operations are nicely compatible in the sense that order and grouping is not important. These are called the distributive laws for scaling and addition. They are

$$\begin{aligned} c \odot (\mathbf{u} \oplus \mathbf{v}) &= (c \odot \mathbf{u}) \oplus (c \odot \mathbf{v}) \\ (c + d) \odot \mathbf{u} &= (c \odot \mathbf{u}) \oplus (d \odot \mathbf{u}). \end{aligned}$$

VS4: There is a special object called \mathbf{o} which functions as a zero so we always have $\mathbf{o} \oplus \mathbf{u} = \mathbf{u} \oplus \mathbf{o} = \mathbf{u}$.

VS5: There are additive inverses which means to each \mathbf{u} there is a unique object \mathbf{u}^\dagger so that $\mathbf{u} \oplus \mathbf{u}^\dagger = \mathbf{o}$.

Comment 3.2.1 These laws imply

$$(0 + 0) \odot \mathbf{u} = (0 \odot \mathbf{u}) \oplus (0 \odot \mathbf{u})$$

which tells us $0 \odot \mathbf{u} = 0$. A little further thought then tells us that since

$$\begin{aligned} \mathbf{0} &= (1 - 1) \odot \mathbf{u} \\ &= (1 \odot \mathbf{u}) \oplus (-1 \odot \mathbf{u}) \end{aligned}$$

we have the additive inverse $\mathbf{u}^\dagger = -1 \odot \mathbf{u}$.

Comment 3.2.2 We usually say this much simpler. The set of objects \mathbf{V} is a vector space over its scalar field if there are two operations which we denote by $\mathbf{u} + \mathbf{v}$ and $c\mathbf{u}$ which generate new objects in the vector space for any \mathbf{u} , \mathbf{v} and scalar c . We then just add that these operations satisfy the usual commutative, associative and distributive laws and there are unique additive inverses.

Comment 3.2.3 The objects are often called vectors and sometimes we denote them by \mathbf{u} although this notation is often too cumbersome.

Comment 3.2.4 To give examples of vector spaces, it is usually enough to specify how the additive and scaling operations are done.

- Vectors in \mathbb{R}^2 , \mathbb{R}^3 and so forth are added and scaled by components.
- Matrices of the same size are added and scaled by components.
- A set of functions of similar characteristics uses as its additive operator, pointwise addition. The new function $(f \oplus g)$ is defined pointwise by $(f \oplus g)(t) = f(t) + g(t)$. Similarly, the new

function $c \odot f$ is defined by $c \odot f$ is the function whose value at t is $(cf)(t) = cf(t)$. Classic examples are

1. $C([a, b])$ is the set of all functions whose domain is $[a, b]$ that are continuous on the domain.
2. $C^1([a, b])$ is the set of all functions whose domain is $[a, b]$ that are continuously differentiable on the domain.
3. $RI([a, b])$ is the set of all functions whose domain is $[a, b]$ that are Riemann integrable on the domain.

There are many more, of course.

Vector spaces have two other important ideas associated with them. We have already talked about linearly independent objects in \mathbb{R}^2 or \mathbb{R}^3 . In general, for any vector space, we would use the same definitions with just a few changes:

Definition 3.2.2 Two Linearly Independent Objects in A Vector Space

Let \mathbf{E} and \mathbf{F} be two objects in a vector space \mathbf{V} . We say \mathbf{E} and \mathbf{F} are **linearly dependent** if we can find non zero scalars α and β so that

$$\alpha \odot \mathbf{E} + \beta \odot \mathbf{F} = \mathbf{0}.$$

Otherwise, we say they are **linearly independent**.

We can then easily extend this idea to any finite collection of such objects as follows.

Definition 3.2.3 Finitely many Linearly Independent Objects

Let $\{\mathbf{E}_i : 1 \leq i \leq N\}$ be N objects in a vector \mathbf{V} . We say \mathbf{E} and \mathbf{F} are **linearly dependent** if we can find non zero constants α_1 to α_N , not all 0, so that

$$\alpha_1 \odot \mathbf{E}_1 + \dots + \alpha_N \odot \mathbf{E}_N = \mathbf{0}.$$

Again, note we have changed the way we define the constants. When there are more than two objects involved, we can't say, in general, that all of the constants must be non zero.

Homework:

Exercise 3.2.1

Exercise 3.2.2

Exercise 3.2.3

Exercise 3.2.4

Exercise 3.2.5

3.2.1 The Span of a Set of Vectors

The next concept is that of the **span** of a collection of vectors.

Definition 3.2.4 The Span Of A Set Of Vectors

Given a finite set of vectors in a vector space \mathbf{V} , $\mathcal{W} = \{\mathbf{u}_1, \dots, \mathbf{u}_N\}$ for some positive integer N , the span of \mathcal{W} is the collection of all new vectors of the form $\sum_{i=1}^N c_i \mathbf{u}_i$ for any choices of scalars c_1, \dots, c_N . It is easy to see \mathcal{W} is a vector space itself and since it is a subset of \mathcal{V} , we call it a vector subspace. The span of the set \mathcal{W} is denoted by $Sp\mathcal{W}$. If the set of vectors \mathcal{W} is not finite, the definition is similar but we say the span of \mathcal{W} is the set of all vectors which can be written as $\sum_{i=1}^N c_i \mathbf{u}_i$ for some finite set of vectors $\mathbf{u}_1, \dots, \mathbf{u}_N$ from \mathcal{W} .

Let's talk about subspaces of a vector space in more detail. If we have a vector space \mathbf{V} and we are given a finite collection of objects $\{V_1, \dots, V_n\}$ in the vector space, we have already noted the span of this collection satisfies the definition of a vector space itself. Since it is a subset of the original vector space, we distinguish it from the vector space \mathbf{V} it is inside and call it a **subspace**. There are many examples and it is useful to look at some.

Example 3.2.1 If we look at the vector space $C([a, b])$ of continuous functions on the interval $[a, b]$, the span of the function $\mathbf{1}$, i.e. the constant function $x(t) = 1$ on $[a, b]$ is the set of all real numbers. This is a subspace of $C([a, b])$. Note $x(t) = t^0$.

The span of $\{t^0, t^1, t^2, \dots, t^n\}$ is all functions of the form $p(t) = \sum_{i=0}^n a_i t^i$ for any value of the real numbers a_i . This is the set of all polynomials of degree n over the real numbers. It is straightforward to show these polynomials are linearly independent in the vector space. The linear independence equation is

$$c_0 \odot \mathbf{1} \oplus c_1 \odot t^1 \oplus \dots \oplus c_n \odot t^n = \mathbf{0}$$

which means

$$c_0 + c_1 t^1 + \dots + c_n t^n = 0, \quad a \leq t \leq b$$

Now take n derivatives of this equation which will tell us $c_n = 0$ and then back substitute up through the n derivative equations to see the rest of the constants are zero also. Thus, this collection of functions is linearly dependent. Note every n^{th} degree polynomial is therefore uniquely expressed as a member of the span of $\{t^0, t^1, t^2, \dots, t^n\}$. Let P_n denote the span of $\{t^0, t^1, t^2, \dots, t^n\}$. Then $\{t^0, t^1, t^2, \dots, t^n\}$ is a linearly independent set that spans P_n . For future reference, note it has $n + 1$ elements.

Example 3.2.2 The set of all solutions to the ordinary differential equation $x'' + 4x = 0$ is the span of the two linearly independent functions $\{\cos(2t), \sin(2t)\}$. Note this linearly independent spanning set consists of two functions.

Comment 3.2.5 Let \mathbf{V} and \mathbf{W} be two linearly independent vectors in a vector space. Is it possible to find another vector \mathbf{Q} independent from \mathbf{V} and \mathbf{W} in the span of \mathbf{V} and \mathbf{W} ? Since \mathbf{Q} is in the span of these two vectors, we know we can find constants a and b so that $\mathbf{Q} = a\mathbf{V} + \mathbf{W}$. But this says the set $\{\mathbf{V}, \mathbf{W}, \mathbf{Q}\}$ is a linearly dependent set. So it is not possible to find a third independent vector in the span.

Comment 3.2.6 The above comment can be extended to the case of n linearly independent vectors and their associated span. We see the maximum number of linearly independent vectors in the span is n .

The comments above lead to the notation of a *basis* for a vector space.

Definition 3.2.5 A Basis For A Finite Dimensional Vector Space

If the set of vectors $\{V_1, \dots, V_n\}$ is a linearly independent spanning set for the vector space \mathcal{V} , we say this set is a **basis** for \mathcal{V} and that the **dimension** of \mathcal{V} is n .

Comment 3.2.7 From our comments, we know the maximum number of linear independent objects in a vector space of dimension n is n .

We can this to vectors spaces that do not have a finite basis. First, we extend the idea of linear independence to sets that are not necessarily finite.

Definition 3.2.6 Linear Independence For Non Finite Sets

Given a set of vectors in a vector space \mathcal{V} , \mathcal{W} , we say \mathcal{W} is a linearly independent subset if every finite set of vectors from \mathcal{W} is linearly independent in the usual manner.

We have already defined the span of a set of vectors that is not finite. So as a final matter, we want to define what we mean by a **basis** for a vector space which does not have a finite basis.

Definition 3.2.7 A Basis For An Infinite Dimensional Vector Space

Given a set of vectors in an infinite dimensional vector space \mathcal{V} , \mathcal{W} , we say \mathcal{W} is a basis for \mathcal{V} if the span of \mathcal{W} is all of \mathcal{V} and if the vectors in \mathcal{W} are linearly independent. Hence, a basis is a linearly independent spanning set for \mathcal{V} . Recall, this means any vector in \mathcal{V} can be expressed as a finite linear combination of elements from \mathcal{W} and any finite collection of objects from \mathcal{W} is linearly independent in the usual sense for a finite dimensional vector space.

Comment 3.2.8 In a vector space like \mathbb{R}^n , the maximum size of a set of linearly independent vectors is n , the dimension of the vector space.

Comment 3.2.9 Let's look at the vector space $C([0, 1])$, the set of all continuous functions on $[0, 1]$. Let \mathcal{W} be the set of all powers of t , $\{1, t, t^2, t^3, \dots\}$. We can use the derivative technique to show this set is linearly independent even though it is infinite in size. Take any finite subset from \mathcal{W} . Label the resulting powers as $\{n_1, n_2, \dots, n_p\}$. Write down the linear dependence equation

$$c_1 t^{n_1} + c_2 t^{n_2} + \dots + c_p t^{n_p} = 0.$$

Take n_p derivatives to find $c_p = 0$ and then backtrack to find the other constants are zero also. Hence $C[0, 1]$ is an infinite dimensional vector space. It is also clear that \mathcal{W} does not span $C[0, 1]$ as if this was true, every continuous function on $[0, 1]$ would be a polynomial of some finite degree. This is not true as $\sin(t)$, e^{-2t} and many others are not finite degree polynomials.

Homework

Exercise 3.2.6

Exercise 3.2.7

Exercise 3.2.8

Exercise 3.2.9

Exercise 3.2.10

3.3 Inner Products

Now there is an important idea that we use a lot in applied work. If we have an object \mathbf{u} in a Vector Space V , we often want to find to *approximate* \mathbf{u} using an element from a given subspace \mathcal{W} of the vector space. To do this, we need to add another property to the vector space. This is the notion of an **inner product**. A vector space with an inner product is called an **Inner Product Space**. We already know what an inner product is in a simple vector space like \mathbb{R}^2 . Many vector spaces can have an inner product structure added easily. For example, in $C([a, b])$, since each object is continuous, each object is Riemann integrable. Hence, given two functions f and g from $C[a, b]$, the real number given by $\int_a^b f(s)g(s)ds$ is well - defined. It satisfies all the usual properties that the inner product for finite dimensional vectors in \mathbb{R}^n does also. These properties are so common we will codify them into a definition for what an inner product for a vector space \mathcal{V} should behave like.

Definition 3.3.1 Real Inner Product

Let \mathcal{V} be a vector space with the reals as the scalar field. Then a mapping ω which assigns a pair of objects to a real number is called an inner product on \mathcal{V} if

IP1: $\omega(\mathbf{u}, \mathbf{v}) = \omega(\mathbf{v}, \mathbf{u})$; that is, the order is not important for any two objects.

IP2: $\omega(c \odot \mathbf{u}, \mathbf{v}) = c\omega(\mathbf{u}, \mathbf{v})$; that is, scalars in the first slot can be pulled out.

IP3: $\omega(\mathbf{u} \oplus \mathbf{w}, \mathbf{v}) = \omega(\mathbf{u}, \mathbf{v}) + \omega(\mathbf{w}, \mathbf{v})$, \$Y. for any three objects.

IP4: $\omega(\mathbf{u}, \mathbf{u}) \geq 0$ and $\omega(\mathbf{u}, \mathbf{u}) = 0$ if and only if $\mathbf{u} = 0$.

*These properties imply that $\omega(\mathbf{u}, c \odot \mathbf{v}) = c\omega(\mathbf{u}, \mathbf{v})$ as well. A vector space \mathcal{V} with an inner product is called an **Inner Product Space**.*

Comment 3.3.1 The inner product is usually denoted with the symbol \langle , \rangle instead of $\omega(,)$. We will use this notation from now on.

Comment 3.3.2 When we have an inner product, we can measure the size or magnitude of an object, as follows. We define the analogue of the euclidean norm of an object \mathbf{u} using the usual $\| \|$ symbol as

$$\| \mathbf{u} \| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}.$$

This is called the norm induced by the inner product of the object. In $C([a, b])$, with the inner product $\langle f, g \rangle = \int_a^b f(s)g(s)ds$, the norm of a function f is thus $\| f \| = \sqrt{\int_a^b f^2(s)ds}$. This has been discussed in the earlier volumes so you can go back and review there to get up to speed. This is called the L_2 norm of f and is denoted $\| \cdot \|_2$.

It is possible to prove the Cauchy-Schwartz inequality in this more general setting also.

Theorem 3.3.1 General Cauchy-Schwartz Inequality

If \mathcal{V} is an inner product space with inner product \langle , \rangle and induced norm $\| \|$, then

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \| \mathbf{u} \| \| \mathbf{v} \|$$

with equality occurring if and only if \mathbf{u} and \mathbf{v} are linearly dependent.

Proof 3.3.1

If $\mathbf{u} = t\mathbf{v}$, then

$$\begin{aligned} |\langle \mathbf{u}, \mathbf{v} \rangle| &= |\langle t\mathbf{v}, \mathbf{v} \rangle| = |t\langle \mathbf{v}, \mathbf{v} \rangle| = |t\|\mathbf{v}\|^2| \\ &= (t\|\mathbf{v}\|)(\|\mathbf{v}\|) = \|\mathbf{u}\|\|\mathbf{v}\| \end{aligned}$$

Hence, the result holds if \mathbf{u} and \mathbf{v} are linearly dependent. In general, if \mathbf{u} and \mathbf{v} are linearly independent, look at the subspace spanned by \mathbf{v} . Any element in this subspace can be written as $a\mathbf{v}\|\mathbf{v}\|$. Consider the function

$$f(a) = \langle \mathbf{u} - a\mathbf{v}\|\mathbf{v}\|, \mathbf{u} - a\mathbf{v}\|\mathbf{v}\| \rangle + \|\mathbf{u}\| - 2a\langle \mathbf{u}, \mathbf{v}\|\mathbf{v}\| \rangle + a^2$$

Then

$$f'(a) = -2\langle \mathbf{u}, \mathbf{v}\|\mathbf{v}\| \rangle + 2a$$

which implies the critical point is $a = -\langle \mathbf{u}, \mathbf{v}\|\mathbf{v}\| \rangle$. Since $f''(a) = 2 > 0$, the critical point is a global minimum. Thus, the vector whose minimum norm distance to the subspace generated by \mathbf{v} is $\mathbf{P}_{\mathbf{u}\mathbf{v}} = \mathbf{u} - \langle \mathbf{u}, \mathbf{v}\|\mathbf{v}\| \rangle \mathbf{v}\|\mathbf{v}\|$ which has length $|\langle \mathbf{u}, \mathbf{v}\|\mathbf{v}\| \rangle|$. The vector $\mathbf{P}_{\mathbf{u}\mathbf{v}}$ is called the projection of \mathbf{u} onto \mathbf{v} . We can then use GSO to find the orthonormal basis for the subspace spanned by \mathbf{u} and \mathbf{v} .

$$\begin{aligned} \mathbf{Q}_1 &= \mathbf{v}/\|\mathbf{v}\| \\ \mathbf{W} &= \mathbf{u} - \langle \mathbf{u}, \mathbf{Q}_1 \rangle \mathbf{Q}_1, \quad \mathbf{Q}_2 = \mathbf{W}/\|\mathbf{W}\| \end{aligned}$$

Thus we have the decomposition $\mathbf{u} = \langle \mathbf{u}, \mathbf{Q}_1 \rangle \mathbf{Q}_1 + \langle \mathbf{u}, \mathbf{Q}_2 \rangle \mathbf{Q}_2$. Hence, if we assumed \mathbf{u} and \mathbf{v} were linearly independent with

$$|\langle \mathbf{u}, \mathbf{v} \rangle| = \|\mathbf{u}\|\|\mathbf{v}\| \implies \|\mathbf{P}_{\mathbf{u}\mathbf{v}}\| = \left| \left\langle \mathbf{u}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\rangle \right| = \|\mathbf{u}\|$$

this forces $\langle \mathbf{u}, \frac{\mathbf{v}}{\|\mathbf{v}\|} \rangle = 0$ telling us $\mathbf{u} = \|\mathbf{u}\|\mathbf{Q}_1$. But, of course, this say \mathbf{u} is a multiple of \mathbf{v} which is not possible as they are assumed independent. Hence, we conclude if $|\langle \mathbf{u}, \mathbf{v} \rangle| = \|\mathbf{u}\|\|\mathbf{v}\|$, \mathbf{u} and \mathbf{v} must be dependent.

However, if \mathbf{u} and \mathbf{v} are linearly independent, we have the decomposition $\mathbf{u} = \langle \mathbf{u}, \mathbf{Q}_1 \rangle \mathbf{Q}_1 + \langle \mathbf{u}, \mathbf{Q}_2 \rangle \mathbf{Q}_2$. which implies $\|\mathbf{u}\|^2 = |\langle \mathbf{u}, \mathbf{Q}_1 \rangle|^2 + |\langle \mathbf{u}, \mathbf{Q}_2 \rangle|^2$. This immediately tells us

$$|\langle \mathbf{u}, \mathbf{v}/\|\mathbf{v}\| \rangle|^2 \leq \|\mathbf{u}\|^2 \implies |\langle \mathbf{u}, \mathbf{v}/\|\mathbf{v}\| \rangle| \leq \|\mathbf{u}\| \implies |\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\|\|\mathbf{v}\|$$

which is the result we wanted to show. ■

Comment 3.3.3 We can use the Cauchy-Schwartz inequality to define a notion of angle between objects exactly like we would do in \mathbb{R}^2 . We define the angle θ between \mathbf{u} and \mathbf{v} via its cosine as usual.

$$\cos(\theta) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|\|\mathbf{v}\|}.$$

Hence, objects can be perpendicular or orthogonal even if we can not interpret them as vectors in \mathbb{R}^2 . We see two objects are orthogonal if their inner product is 0.

Comment 3.3.4 If \mathcal{W} is a finite dimensional subspace, a basis for \mathcal{W} is said to be an orthonormal basis if each object in the basis has L_2 norm 1 and all of the objects are mutually orthogonal. This means $\langle \mathbf{u}_i, \mathbf{u}_j \rangle$ is 1 if $i = j$ and 0 otherwise. We typically let the Kronecker delta symbol

δ_{ij} be defined by $\delta_{ij} = 1$ if $i = j$ and 0 otherwise so that we can say this more succinctly as $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{ij}$.

Let's go back to \mathbb{R}^2 . You have gone over these ideas quite a bit in your last few courses, so we will just refresh you mind here. Grab three pencils for the \mathbb{R}^2 examples coming up and four pencils for the \mathbb{R}^2 examples. We will try to look at these examples in a new light so you can build on earlier levels of understanding.

3.3.1 Homework

Exercise 3.3.1

Exercise 3.3.2

Exercise 3.3.3

Exercise 3.3.4

Exercise 3.3.5

3.4 Examples

Let's work through a number of important examples ranging from finite dimensional to infinite dimensional cases.

3.4.1 Two Dimensional Vectors in the Plane

Take two pencils and hold them in your hand. They don't have to be the same size. These are two vectors in the plane. It doesn't matter how you hold them relative to the room you are in either. These two pencils or vectors are linearly independent if they point in different directions. The only way they are linearly dependent is if they lay on top of one another or point directly opposite to one another. In the first case, the angle between the vectors is 0 radians and in the second the angle is π . Now these two pencils can be thought of as sitting in a sheet of paper which is our approximation of the plane determined by the two vectors represented by the pencils. Note as long the vectors are linearly independent we can represent any other vector (now use that third pencil!) \mathbf{W} in terms of the first two which we now call \mathbf{E} and \mathbf{F} . We want to find the scalars a and b so that $a\mathbf{E} + b\mathbf{F} = \mathbf{W}$ where we have stopped representing scalar multiplication by \odot . Taking the usual representation of these vectors, we have

$$\begin{aligned}\mathbf{E} &= E_{11} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + E_{12} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ \mathbf{F} &= F_{11} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + F_{12} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ \mathbf{W} &= W_{11} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + W_{12} \begin{bmatrix} 0 \\ 1 \end{bmatrix}\end{aligned}$$

We assume, of course, our vectors are non zero. Recall we usually just write this as

$$\mathbf{E} = \begin{bmatrix} E_{11} \\ E_{12} \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} F_{11} \\ F_{12} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} W_{11} \\ W_{12} \end{bmatrix}$$

and it is very important to notice that we have **hidden** our way of representing the components of our vectors inside these column vectors. Using matrix and vector notation, we can write our problem of

finding a and b as

$$\begin{bmatrix} E_{11} & F_{11} \\ E_{12} & F_{12} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} W_{11} \\ W_{12} \end{bmatrix}$$

Call this 2×2 matrix \mathbf{T} . Now we all know how to take the determinant of a 2×2 matrix. Note if $\det(\mathbf{T}) = E_{11}F_{12} - E_{12}F_{11}$. If this determinant was zero, then that would tell us $\frac{E_{11}}{E_{12}} = \frac{F_{11}}{F_{12}}$ in the case all the components are non zero. But this tells us \mathbf{E} and \mathbf{F} point in the same direction. Hence, they can't be linearly independent. The argument going the other way is similar. So the $\det(\mathbf{T}) = 0$ if and only if these two vectors are linearly dependent. Since we have assumed they are the opposite: i.e. linearly independent, we must have $\det(\mathbf{T}) \neq 0$. But then Cramer's rule tells us how to find a and b . Note each vector must have at least one non zero component and it is easy to alter our argument above to handle the zero's we might see. We will let you do that yourself! Using Cramer's rule, we have

$$a = \frac{\begin{bmatrix} W_{11} & F_{11} \\ W_{12} & F_{12} \end{bmatrix}}{\det(\mathbf{T})}, \quad b = \frac{\begin{bmatrix} E_{11} & W_{11} \\ E_{12} & W_{12} \end{bmatrix}}{\det(\mathbf{T})}$$

Of course, if the vectors \mathbf{E} and \mathbf{F} were orthonormal, things are much simpler. First, note

$$a\mathbf{E} + b\mathbf{F} = \mathbf{W} \implies \langle \mathbf{E}, a\mathbf{E} + b\mathbf{F} \rangle = \langle \mathbf{E}, \mathbf{W} \rangle = a \langle \mathbf{E}, \mathbf{E} \rangle + b \langle \mathbf{E}, \mathbf{F} \rangle = \langle \mathbf{E}, \mathbf{W} \rangle$$

But $\langle \mathbf{E}, \mathbf{F} \rangle = 0$ and $\langle \mathbf{E}, \mathbf{E} \rangle = 1$, so we have $a = \langle \mathbf{E}, \mathbf{W} \rangle$. In a similar way, we find

$$a\mathbf{E} + b\mathbf{F} = \mathbf{W} \implies \langle \mathbf{F}, a\mathbf{E} + b\mathbf{F} \rangle = \langle \mathbf{F}, \mathbf{W} \rangle = a \langle \mathbf{F}, \mathbf{E} \rangle + b \langle \mathbf{F}, \mathbf{F} \rangle = \langle \mathbf{E}, \mathbf{W} \rangle$$

which tells us $b = \langle \mathbf{F}, \mathbf{W} \rangle$. So we have found the unique solution is

$$\mathbf{W} = \begin{bmatrix} \langle \mathbf{W}, \mathbf{E} \rangle \\ \langle \mathbf{W}, \mathbf{F} \rangle \end{bmatrix} = \begin{bmatrix} E_{11}W_{11} + E_{12}W_{12} \\ F_{11}W_{11} + F_{12}W_{12} \end{bmatrix}$$

Now using standard notation, we can also write this as

$$\begin{bmatrix} a \\ b \end{bmatrix} = [\mathbf{E}, \mathbf{F}] \mathbf{W}$$

It is easy to see, using the orthonormality of \mathbf{E} and \mathbf{F} that we also have

$$\begin{bmatrix} \mathbf{E}^T \\ \mathbf{F}^T \end{bmatrix} [\mathbf{E}, \mathbf{F}] = \mathbf{I}$$

where \mathbf{I} is the 2×2 identity matrix. Since we are in two dimensions, we note there is a nice relationship between the components of \mathbf{E} and \mathbf{F} because they are orthogonal. First, the slope of \mathbf{F} is the negative reciprocal of the slope of \mathbf{E} and hence we must have

$$\mathbf{E} = \begin{bmatrix} E_{11} \\ E_{12} \end{bmatrix} = \mathbf{F} = \begin{bmatrix} -E_{12} \\ E_{11} \end{bmatrix}$$

with $E_{11}^2 + E_{12}^2 = 1$. We can thus identify $E_{11} = \cos(\theta)$ and $E_{12} = \sin(\theta)$ for some angle θ . This shows us we can represent \mathbf{E} and \mathbf{F} as

$$\mathbf{E} = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} = \mathbf{F} = \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix}$$

3.4. EXAMPLES

33

Thus, we see

$$\begin{aligned} [\mathbf{E}, \mathbf{F}] \begin{bmatrix} \mathbf{E}^T \\ \mathbf{F}^T \end{bmatrix} &= \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \\ &= \begin{bmatrix} \cos^2(\theta) + \sin(\theta)^2 & \cos(\theta)\sin(\theta) - \cos(\theta)\sin(\theta) \\ \cos(\theta)\sin(\theta) - \cos(\theta)\sin(\theta) & \cos^2(\theta) + \sin(\theta)^2 \end{bmatrix} = \mathbf{I} \end{aligned}$$

You probably recognize this representation of \mathbf{T} as the usual 2×2 matrix which denotes the rotation of a vector by the angle θ . We note the determinant here is always +1. From what we have just done, it is now clear that the matrix $[\mathbf{E}^T, \mathbf{F}^T]$ is the inverse of the matrix \mathbf{T} . Indeed, it is clear the inverse is \mathbf{T}^T ; i.e. the inverse of \mathbf{T} is its own transpose. Hence, any two dimensional vector in \mathbb{R}^2 is a unique linear combination of \mathbf{E} and \mathbf{F} .

What we have shown here is that any vector \mathbf{W} can be written as $a\mathbf{E} + b\mathbf{F}$ for a unique a and b . This kind of representation is what we have called a **linear combination of the vectors \mathbf{E} and \mathbf{F}** . Also note we can do this whether \mathbf{E} and \mathbf{F} form an orthonormal set. Clearly, the set of all possible linear combinations of \mathbf{E} and \mathbf{F} gives all of \mathbb{R}^2 ; i.e. **the span of \mathbf{E} and \mathbf{F} is \mathbb{R}^2** . Further, since \mathbf{E} and \mathbf{F} are linearly independent, we say **\mathbf{E} and \mathbf{F} form a linearly independent spanning set or basis for \mathbb{R}^2** .

This method of attack will not be very useful for three dimensional vectors, of course as we will lose our intuition of dependent vectors being a simple scalar multiple of one another. Another way of showing $[\mathbf{E} \ \mathbf{F}]^T$ is the inverse of \mathbf{T} is as follows. It does not use rotations at all. Let

$$\mathbf{A} = [\mathbf{E}, \mathbf{F}] \begin{bmatrix} \mathbf{E}^T \\ \mathbf{F}^T \end{bmatrix}$$

We already know that any vector \mathbf{W} can be expressed as the unique linear combination $a\mathbf{E} + b\mathbf{F}$ where

$$\begin{bmatrix} a \\ b \end{bmatrix} = [\mathbf{E}, \mathbf{F}] \mathbf{W}$$

Thus

$$\begin{aligned} \mathbf{A} \mathbf{W} &= [\mathbf{E}, \mathbf{F}] \begin{bmatrix} \mathbf{E}^T \\ \mathbf{F}^T \end{bmatrix} (a\mathbf{E} + b\mathbf{F}) \\ &= a [\mathbf{E}, \mathbf{F}] \begin{bmatrix} \mathbf{E}^T \\ \mathbf{F}^T \end{bmatrix} \mathbf{E} + b [\mathbf{E}, \mathbf{F}] \begin{bmatrix} \mathbf{E}^T \\ \mathbf{F}^T \end{bmatrix} \mathbf{F} \\ &= a [\mathbf{E}, \mathbf{F}] \begin{bmatrix} \mathbf{E}^T \mathbf{E} \\ \mathbf{F}^T \mathbf{E} \end{bmatrix} + b [\mathbf{E}, \mathbf{F}] \begin{bmatrix} \mathbf{E}^T \mathbf{F} \\ \mathbf{F}^T \mathbf{F} \end{bmatrix} \\ &= a [\mathbf{E}, \mathbf{F}] \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b [\mathbf{E}, \mathbf{F}] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = a\mathbf{E} + b\mathbf{F} = \mathbf{W} \end{aligned}$$

Since $\mathbf{A}\mathbf{W} = \mathbf{W}$ for all \mathbf{W} , this tells us $\mathbf{A} = \mathbf{I}$. Combining we have

$$\begin{bmatrix} \mathbf{E}^T \\ \mathbf{F}^T \end{bmatrix} [\mathbf{E}, \mathbf{F}] = [\mathbf{E}, \mathbf{F}] \begin{bmatrix} \mathbf{E}^T \\ \mathbf{F}^T \end{bmatrix} = \mathbf{I}$$

And so the inverse of \mathbf{T} is \mathbf{T}^T but this time we proved the result without resorting to rotation angles and slopes of vectors.

Of course, if \mathbf{E} and \mathbf{F} were not orthonormal, we could apply Graham - Schmidt Orthogonalization to create the orthonormal basis $\mathbf{U}_1, \mathbf{U}_2$ as follows:

$$\begin{aligned} \mathbf{U}_1 &= \frac{\mathbf{E}}{\|\mathbf{E}\|} \\ \mathbf{V} &= \mathbf{F} - \langle \mathbf{F}, \mathbf{U}_1 \rangle \mathbf{U}_1 \\ \mathbf{U}_2 &= \frac{\mathbf{V}}{\|\mathbf{V}\|} \end{aligned}$$

and $\{\mathbf{U}_1, \mathbf{U}_2\}$ would be an orthonormal basis for \Re^2 .

3.4.1.1 Homework

Exercise 3.4.1

Exercise 3.4.2

Exercise 3.4.3

Exercise 3.4.4

Exercise 3.4.5

3.4.2 The Connection between Two Orthonormal Bases

Given the orthonormal basis $\{\mathbf{E}_1, \mathbf{E}_2\}$, we know from our remarks above that the matrix we form by using these vectors as columns, $\mathbf{T}_{\mathbf{E}} = [\mathbf{E}_1, \mathbf{E}_2]$ has an inverse which is its own transpose; i.e. $\mathbf{T}_{\mathbf{E}}^{-1} = (\mathbf{T}_{\mathbf{E}})^T$. For convenience, we now denote this orthonormal basis by \mathbf{E} . Hence, the orthonormal basis $\{\mathbf{F}_1, \mathbf{F}_2\}$ is denoted by \mathbf{F} . We now introduce notation we will use later when we think carefully about transformations from \Re^n to \Re^m . Given a vector \mathbf{A} it has a representation relative to any choice of basis. We need to recognize that this representation will depend on the choice of basis, so given that $\mathbf{A} = a\mathbf{E}_1 + b\mathbf{E}_2$ for unique scalars a and b , the vector $\begin{bmatrix} a \\ b \end{bmatrix}$ is a convenient way to handle this representation. We use the subscript \mathbf{E} to remind us these components depend on our choice of basis \mathbf{E} . When we write a two dimensional column vector $\begin{bmatrix} a \\ b \end{bmatrix}$ have you ever thought about what the numbers a and b mean? You need to always remember that the idea of a two dimensional vector space is quite abstract. We can't really see these objects until we make a choice of basis. Thus we can't calculate inner products until we choose a basis also. The simplest basis is the usual one

$$\mathbf{i} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{j} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

which is called the **standard basis** which we will denote by \mathbf{I} . So all of our discussions above were using the standard basis \mathbf{I} . Now this is very important. With respect to the orthonormal basis \mathbf{E} , the representation of \mathbf{E}_1 is $\begin{bmatrix} E_{11} \\ E_{12} \end{bmatrix}_{\mathbf{I}}$ but with respect to the basis \mathbf{E} , its components would be $\begin{bmatrix} 1 \\ 0 \end{bmatrix}_{\mathbf{E}}$. This is what we are doing when we set up the cartesian coordinate system. We choose what we call the x and y axis oriented so they are 90° apart. The vectors \mathbf{i} and \mathbf{j} are then chosen along the positive x and y axis. However, rotating this choice of \mathbf{i} and \mathbf{j} by θ° gives us a new x' and y' axis still 90° apart. So when we want to be very explicit, we write

$$[\mathbf{A}]_{\mathbf{E}} = \begin{bmatrix} a \\ b \end{bmatrix}_{\mathbf{E}}$$

3.4. EXAMPLES

35

to indicate this representation with respect to the basis \mathbf{E} . Now \mathbf{A} also has a representation due to the basis \mathbf{F} .

$$[\mathbf{A}]_{\mathbf{F}} = \begin{bmatrix} c \\ d \end{bmatrix}_{\mathbf{F}}$$

From our first discussion, now explicitly realizing we were using representations relative to the standard basis, we should write

$$[\mathbf{A}_I]_{\mathbf{F}_I} = \begin{bmatrix} c \\ d \end{bmatrix}_{\mathbf{F}_I}$$

This is very cumbersome. The subscript I here reminds us that the components we see in the vectors are relative to the standard basis and that the components of \mathbf{E} and \mathbf{F} are also given with respect to the standard basis. If we are expressing everything in terms of the \mathbf{E} basis remember \mathbf{E}_1 and \mathbf{E}_2 themselves have the usual standard basis components; i.e. we are finding components relative to a new x' and y' axis. So with all that said, given some component representation for the components of \mathbf{E} , we can find the components of the vectors in \mathbf{F} with respect to the basis \mathbf{E} as follows. Remember, the basis \mathbf{F} corresponds to a new rotation to an x'' and y'' coordinate axis system. Using these ideas, we see we can also express \mathbf{F} in terms of \mathbf{E} . We have

$$\begin{aligned} \mathbf{F}_1 &= \langle \mathbf{F}_1, \mathbf{E}_1 \rangle \mathbf{E}_1 + \langle \mathbf{F}_1, \mathbf{E}_2 \rangle \mathbf{E}_2 = T_{11}\mathbf{E}_1 + T_{12}\mathbf{E}_2 \\ \mathbf{F}_2 &= \langle \mathbf{F}_2, \mathbf{E}_1 \rangle \mathbf{E}_1 + \langle \mathbf{F}_2, \mathbf{E}_2 \rangle \mathbf{E}_2 = T_{21}\mathbf{E}_1 + T_{22}\mathbf{E}_2 \end{aligned}$$

Thus, using our two representations for \mathbf{A} , we have

$$\begin{aligned} \mathbf{A} &= c\mathbf{F}_1 + d\mathbf{F}_2 = c(T_{11}\mathbf{E}_1 + T_{12}\mathbf{E}_2) + d(T_{21}\mathbf{E}_1 + T_{22}\mathbf{E}_2) \\ &= (cT_{11} + dT_{21})\mathbf{E}_1 + (cT_{12} + dT_{22})\mathbf{E}_2 = a\mathbf{E}_1 + b\mathbf{E}_2 \end{aligned}$$

This tells us

$$\begin{aligned} a &= cT_{11} + dT_{21} \\ b &= cT_{12} + dT_{22} \end{aligned}$$

This implies

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} T_{11} & T_{21} \\ T_{12} & T_{22} \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} \implies [\mathbf{A}]_{\mathbf{E}} = \begin{bmatrix} T_{11} & T_{21} \\ T_{12} & T_{22} \end{bmatrix} [\mathbf{A}]_{\mathbf{F}}$$

or

$$[\mathbf{A}]_{\mathbf{E}} = \begin{bmatrix} \langle \mathbf{F}_1, \mathbf{E}_1 \rangle & \langle \mathbf{F}_2, \mathbf{E}_1 \rangle \\ \langle \mathbf{F}_1, \mathbf{E}_2 \rangle & \langle \mathbf{F}_2, \mathbf{E}_2 \rangle \end{bmatrix} [\mathbf{A}]_{\mathbf{F}}$$

We can do the same sort of analysis and interchange the role of \mathbf{E} and \mathbf{F} to get

$$[\mathbf{A}]_{\mathbf{F}} = \begin{bmatrix} \langle \mathbf{E}_1, \mathbf{F}_1 \rangle & \langle \mathbf{E}_2, \mathbf{F}_1 \rangle \\ \langle \mathbf{E}_1, \mathbf{F}_2 \rangle & \langle \mathbf{E}_2, \mathbf{F}_2 \rangle \end{bmatrix} [\mathbf{A}]_{\mathbf{E}}$$

The coefficients (T_{ij}) define a matrix which we will call $T_{\mathbf{F}\mathbf{E}}$ as these coefficients are completely determined by the basis \mathbf{E} and \mathbf{F} and we are transforming \mathbf{F} components into \mathbf{E} components. Hence,

we let

$$\mathbf{T}_{FE} = \begin{bmatrix} < \mathbf{F}_1, \mathbf{E}_1 > & < \mathbf{F}_2, \mathbf{E}_1 > \\ < \mathbf{F}_1, \mathbf{E}_2 > & < \mathbf{F}_2, \mathbf{E}_2 > \end{bmatrix} = \begin{bmatrix} T_{11} & T_{21} \\ T_{12} & T_{22} \end{bmatrix} = \mathbf{T}_{EF}^T$$

and

$$\mathbf{T}_{EF} = \begin{bmatrix} < \mathbf{E}_1, \mathbf{F}_1 > & < \mathbf{E}_2, \mathbf{F}_1 > \\ < \mathbf{E}_1, \mathbf{F}_2 > & < \mathbf{E}_2, \mathbf{F}_2 > \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \mathbf{T}_{FE}^T$$

So \mathbf{T}_{FE} and \mathbf{T}_{EF} are transposes of each other. We can easily show \mathbf{T}_{FE}^T is the inverse of \mathbf{T}_{FE} . We note for all vectors in \Re^2 ,

$$\mathbf{A}_E = \mathbf{T}_{FE} \mathbf{A}_F = \mathbf{T}_{FE} \mathbf{T}_{EF} \mathbf{A}_E = \mathbf{T}_{FE} \mathbf{T}_{FE}^T \mathbf{A}_E$$

and

$$\mathbf{A}_F = \mathbf{T}_{EF} \mathbf{A}_E = \mathbf{T}_{FE}^T \mathbf{T}_{FE} \mathbf{A}_F = \mathbf{T}_{FE}^T \mathbf{T}_{FE} \mathbf{A}_E$$

Hence $\mathbf{T}_{FE} \mathbf{T}_{FE}^T = \mathbf{T}_{FE}^T \mathbf{T}_{FE} = \mathbf{I}$ and \mathbf{T}_{FE}^T is the inverse of \mathbf{T}_{FE} .

3.4.2.1 Homework

Exercise 3.4.6

Exercise 3.4.7

Exercise 3.4.8

Exercise 3.4.9

Exercise 3.4.10

3.4.3 The Invariance of The Inner Product

For two orthonormal basis \mathbf{E} and \mathbf{F} , we now know the vector \mathbf{A} transforms as follows:

$$[\mathbf{A}]_F = [\mathbf{T}_1, \mathbf{T}_2]^T [\mathbf{A}]_E$$

Thus, for two vectors \mathbf{A} and \mathbf{B} , we can calculate

$$< [\mathbf{A}]_F, [\mathbf{B}]_F > = < [\mathbf{T}_1, \mathbf{T}_2]^T [\mathbf{A}]_E, [\mathbf{T}_1, \mathbf{T}_2]^T [\mathbf{B}]_E >$$

It is easy to see that using vector representations, $< \mathbf{A}, \mathbf{B} > = \mathbf{A}^T \mathbf{B} = \mathbf{B}^T \mathbf{A}$ where these are interpreted using the usual matrix multiplication routines we know. Thus,

$$\begin{aligned} < [\mathbf{A}]_F, [\mathbf{B}]_F > &= (< [\mathbf{T}_1, \mathbf{T}_2]^T [\mathbf{A}]_E)^T [\mathbf{T}_1, \mathbf{T}_2]^T [\mathbf{B}]_E > \\ &= [\mathbf{A}^T]_E ([\mathbf{T}_1, \mathbf{T}_2]) [\mathbf{T}_1, \mathbf{T}_2]^T [\mathbf{B}]_E \end{aligned}$$

But $([\mathbf{T}_1, \mathbf{T}_2])^T$ is the inverse of $[\mathbf{T}_1, \mathbf{T}_2]$. Thus, we find

$$< [\mathbf{A}]_F, [\mathbf{B}]_F > = [\mathbf{A}^T]_E [\mathbf{B}]_E = < [\mathbf{A}]_E, [\mathbf{B}]_E >$$

3.4. EXAMPLES

37

which tells us the inner product calculation is invariant under a transformation from one orthonormal basis to another. Of course, if one or both of these new basis are not orthonormal, the matrix \mathbf{T} here is not its own inverse and the value of the inner product could change.

So in general, not writing down all the subscripts about components with respect to various bases, we find unique scalars using the $\{\mathbf{E}_1, \mathbf{E}_2\}$ basis so that

$$\mathbf{A} = a\mathbf{E}_1 + b\mathbf{E}_2, \quad \mathbf{B} = c\mathbf{E}_1 + d\mathbf{E}_2$$

Thus,

$$\langle \mathbf{A}, \mathbf{B} \rangle = \langle a\mathbf{E}_1 + b\mathbf{E}_2, c\mathbf{E}_1 + d\mathbf{E}_2 \rangle = ac + bd$$

Using the other basis, we have

$$\begin{aligned}\mathbf{A} &= \alpha\mathbf{F}_1 + \beta\mathbf{F}_2 \\ \mathbf{B} &= \gamma\mathbf{F}_1 + \delta\mathbf{F}_2\end{aligned}$$

and so

$$\langle \mathbf{A}, \mathbf{B} \rangle = \langle \alpha\mathbf{F}_1 + \beta\mathbf{F}_2, \gamma\mathbf{F}_1 + \delta\mathbf{F}_2 \rangle = \alpha\gamma + \beta\delta$$

And we now know these two values will be the same!

3.4.3.1 Homework

Exercise 3.4.11

Exercise 3.4.12

Exercise 3.4.13

Exercise 3.4.14

Exercise 3.4.15

3.4.4 Two Dimensional Vectors As Functions

Let's review Linear Systems of first order ODE. These have the form

$$\begin{aligned}x'(t) &= a x(t) + b y(t) \\ y'(t) &= c x(t) + d y(t) \\ x(0) &= x_0, \quad y(0) = y_0\end{aligned}$$

for any numbers a, b, c and d and *initial conditions* x_0 and y_0 . The full problem is called, as usual, an *Initial Value Problem* or **IVP** for short. The two initial conditions are just called the **IC**'s for the problem to save writing.

- For example, we might be interested in the system

$$\begin{aligned}x'(t) &= -2 x(t) + 3 y(t) \\ y'(t) &= 4 x(t) + 5 y(t) \\ x(0) &= 5, \quad y(0) = -3\end{aligned}$$

Here the **IC**'s are $x(0) = 5$ and $y(0) = -3$.

- Another sample problem might be the one below.

$$\begin{aligned}x'(t) &= 14x(t) + 5y(t) \\y'(t) &= -4x(t) + 8y(t) \\x(0) &= 2, \quad y(0) = 7\end{aligned}$$

3.4.4.1 The Characteristic Equation

For linear first order problems like $u' = 3u$ and so forth, we find the solution has the form $u(t) = Ae^{3t}$ for some number A . We then determine the value of A to use by looking at the initial condition. To find the solutions here, we begin by rewriting the model in matrix - vector notation.

$$\begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}.$$

The 2×2 matrix above is called the **coefficient matrix** of this model and is usually denoted by \mathbf{A} . The initial conditions can then be redone in vector form as

$$\begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

Now it seems reasonable to believe that if a constant times e^{rt} solves a first order linear problem like $u' = ru$, perhaps a *vector* times e^{rt} will work here. Let's make this formal. So let's look at the problem below

$$\begin{aligned}x'(t) &= 3x(t) + 2y(t) \\y'(t) &= -4x(t) + 5y(t) \\x(0) &= 2 \\y(0) &= -3\end{aligned}$$

Assume the solution has the form $\mathbf{V} e^{rt}$. Let's denote the components of \mathbf{V} as follows:

$$\mathbf{V} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}.$$

We assume the solution is

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \mathbf{V} e^{rt}.$$

Then the derivative of $\mathbf{V} e^{rt}$ is

$$(\mathbf{V} e^{rt})' = r \mathbf{V} e^{rt} \implies \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} = r \mathbf{V} e^{rt}.$$

When we plug these terms into the matrix - vector form of the problem, we find

$$r \mathbf{V} e^{rt} = \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \mathbf{V} e^{rt}$$

Rewrite using the identity matrix \mathbf{I} as

$$r \mathbf{V} e^{rt} - \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \mathbf{V} e^{rt} = r \mathbf{I} \mathbf{V} e^{rt} - \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \mathbf{V} e^{rt}$$

3.4. EXAMPLES

39

$$\begin{aligned}
&= \left(r\mathbf{I} - \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \right) \mathbf{V} e^{rt} \\
&= \left(\begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix} - \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \right) \mathbf{V} e^{rt} \\
&= \begin{bmatrix} r-3 & -2 \\ -(-4) & r-5 \end{bmatrix} \mathbf{V} e^{rt}
\end{aligned}$$

Plugging this into our model, we find

$$\begin{bmatrix} r-3 & -2 \\ 4 & r-5 \end{bmatrix} \mathbf{V} e^{rt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

But e^{rt} is never 0, so we want r satisfying

$$\begin{bmatrix} r-3 & -2 \\ 4 & r-5 \end{bmatrix} \mathbf{V} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

For each r , we get two equations in V_1 and V_2 :

$$(r-3)V_1 - 2V_2 = 0, \quad 4V_1 + (r-5)V_2 = 0.$$

Let \mathbf{A}_r be this matrix. Any r for which the $\det \mathbf{A}_r \neq 0$ tells us these two lines have different slopes and so cross at the origin implying $V_1 = 0$ and $V_2 = 0$. Thus

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} e^{rt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

which will not satisfy **nonzero initial conditions**. So **reject** these r . Any value of r for which $\det \mathbf{A}_r = 0$ gives an infinite number of solutions which allows us to pick one that matches the initial conditions we have. The equation

$$\det(r\mathbf{I} - \mathbf{A}) = \det \begin{bmatrix} r-3 & -2 \\ 4 & r-5 \end{bmatrix} = 0.$$

is called the **characteristic equation** of this linear system. The **characteristic equation** is a quadratic, so there are three possibilities: two distinct roots, the real roots are the same and the roots are a complex conjugate pair.

Example 3.4.1 Derive the characteristic equation for the system below

$$\begin{aligned}
x'(t) &= 8x(t) + 9y(t) \\
y'(t) &= 3x(t) - 2y(t) \\
x(0) &= 12 \\
y(0) &= 4
\end{aligned}$$

Solution The matrix - vector form is

$$\begin{aligned}
\begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} &= \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \\
\begin{bmatrix} x(0) \\ y(0) \end{bmatrix} &= \begin{bmatrix} 12 \\ 4 \end{bmatrix}
\end{aligned}$$

The coefficient matrix \mathbf{A} is thus

$$\mathbf{A} = \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix}$$

Assume the solution has the form $\mathbf{V} e^{rt}$. Plug this into the system.

$$r \mathbf{V} e^{rt} - \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix} \mathbf{V} e^{rt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Rewrite using the identity matrix I and factor

$$\left(r \mathbf{I} - \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix} \right) \mathbf{V} e^{rt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Since $e^{rt} \neq 0$ ever, we find r and \mathbf{V} satisfy

$$\left(r \mathbf{I} - \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix} \right) \mathbf{V} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

If r is chosen so that $\det(r\mathbf{I} - \mathbf{A}) \neq 0$, the only solution to this system of two linear equations in the two unknowns V_1 and V_2 is $V_1 = 0$ and $V_2 = 0$. This leads to $x(t) = 0$ and $y(t) = 0$ always and this solution does not satisfy the initial conditions. Hence, we must find r which give $\det(r\mathbf{I} - \mathbf{A}) = 0$. The **characteristic equation** is thus

$$\begin{aligned} \det \begin{bmatrix} r - 8 & -9 \\ -3 & r + 2 \end{bmatrix} &= (r - 8)(r + 2) - 27 \\ &= r^2 - 6r - 43 = 0 \end{aligned}$$

Homework

Exercise 3.4.16

Exercise 3.4.17

Exercise 3.4.18

Exercise 3.4.19

Exercise 3.4.20

3.4.4.2 Finding The General Solution

The roots to the characteristic equation are of course the **eigenvalues** of the coefficient matrix \mathbf{A} . We usually organize the **eigenvalues** as with the largest one first although we don't have to. In fact, in the examples, we organize from small to large!

- Example: The eigenvalues are -2 and -1 . So $r_1 = -1$ and $r_2 = -2$. Since e^{-2t} decays faster than e^{-t} , we say the root $r_1 = -1$ is the **dominant** part of the solution.
- Example: The eigenvalues are -2 and 3 . So $r_1 = 3$ and $r_2 = -2$. Since e^{-2t} decays and e^{3t} grows, we say the root $r_1 = 3$ is the **dominant** part of the solution.
- Example: The eigenvalues are 2 and 3 . So $r_1 = 3$ and $r_2 = 2$. Since e^{2t} grows slower than e^{3t} , we say the root $r_1 = 3$ is the **dominant** part of the solution.

3.4. EXAMPLES

41

For each eigenvalue r we want to find nonzero vectors \mathbf{V} so that $(r\mathbf{I} - \mathbf{A})\mathbf{V} = \mathbf{0}$ where to help with our writing we let $\mathbf{0}$ be the two dimensional zero vector. The nonzero \mathbf{V} are called the **eigenvectors** for **eigenvalue** r and satisfy $\mathbf{AV} = r\mathbf{V}$.

For eigenvalue r_1 , find \mathbf{V} so that $(r_1\mathbf{I} - \mathbf{A})\mathbf{V} = \mathbf{0}$. There will be an infinite number of \mathbf{V} 's that solve this; we pick one and call it **eigenvector** \mathbf{E}_1 .

For eigenvalue r_2 , find \mathbf{V} so that $(r_2\mathbf{I} - \mathbf{A})\mathbf{V} = \mathbf{0}$. There will again be an infinite number of \mathbf{V} 's that solve this; we pick one and call it **eigenvector** \mathbf{E}_2 .

The general solution to our model will be

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = A\mathbf{E}_1 e^{r_1 t} + B\mathbf{E}_2 e^{r_2 t}.$$

where A and B are arbitrary. We use the ICs to find A and B . It is best to show all this with some examples.

Example 3.4.2 For the system below

$$\begin{aligned} \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} &= \begin{bmatrix} -20 & 12 \\ -13 & 5 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \\ \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} &= \begin{bmatrix} -1 \\ 2 \end{bmatrix} \end{aligned}$$

- Find the characteristic equation
- Find the general solution
- Solve the IVP

Solution The characteristic equation is

$$\det \left(r \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} -20 & 12 \\ -13 & 5 \end{bmatrix} \right) = 0$$

$$\begin{aligned} 0 &= \det \left(\begin{bmatrix} r+20 & -12 \\ 13 & r-5 \end{bmatrix} \right) \\ &= (r+20)(r-5) + 156 \\ &= r^2 + 15r + 56 \\ &= (r+8)(r+7) \end{aligned}$$

Hence, **eigenvalues or roots** of the characteristic equation are $r_1 = -8$ and $r_2 = -7$. Note since this is just a calculation, we are not following our labeling scheme.

For eigenvalue $r_1 = -8$, substitute the value into

$$\begin{bmatrix} r+20 & -12 \\ 13 & r-5 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \Rightarrow \begin{bmatrix} 12 & -12 \\ 13 & -13 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

This system of equations should be collinear: i.e. the rows should be multiples; i.e. both give rise to the same line. Our rows are multiples, so we can pick any row to find V_2 in terms of V_1 . Picking the top row, we get $12V_1 - 12V_2 = 0$ implying $V_2 = V_1$. Letting $V_1 = a$, we find $V_1 = a$ and $V_2 = a$: so

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = a \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Choose \mathbf{E}_1 :
The vector

$$\mathbf{E}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

is our choice for an **eigenvector** corresponding to eigenvalue $r_1 = -8$. So one of the solutions is

$$\begin{bmatrix} x_1(t) \\ y_1(t) \end{bmatrix} = \mathbf{E}_1 e^{-8t} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-8t}.$$

For eigenvalue $r_2 = -7$, substitute the value into

$$\begin{bmatrix} r+20 & -12 \\ 13 & r-5 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \Rightarrow \begin{bmatrix} 13 & -12 \\ 13 & -12 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

This system of equations should be collinear: i.e. the rows should be multiples; i.e. both give rise to the same line. Our rows are multiples, so we can pick any row to find V_2 in terms of V_1 . Picking the top row, we get $13V_1 - 12V_2 = 0$ implying $V_2 = (13/12)V_1$. Letting $V_1 = b$, we find $V_1 = b$ and $V_2 = (13/12)b$: so

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = b \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix}$$

Choose \mathbf{E}_2 to be

$$\mathbf{E}_2 = \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix}$$

as our choice for an **eigenvector** corresponding to eigenvalue $r_2 = -7$. So one of the solutions is

$$\begin{bmatrix} x_2(t) \\ y_2(t) \end{bmatrix} = \mathbf{E}_2 e^{-7t} = \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} e^{-7t}.$$

Note, it is easy to see \mathbf{E}_1 and \mathbf{E}_2 are linearly independent vectors in \mathbb{R}^2 . The general solution is then

$$\begin{aligned} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} &= A \mathbf{E}_1 e^{-8t} + B \mathbf{E}_2 e^{-7t} \\ &= A \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-8t} + B \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} e^{-7t} \end{aligned}$$

Now let's solve the initial value problem: we find A and B .

$$\begin{aligned} \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} &= \begin{bmatrix} -1 \\ 2 \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^0 + B \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} e^0 \\ &= A \begin{bmatrix} 1 \\ 1 \end{bmatrix} + B \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} \end{aligned}$$

So

$$\begin{aligned} A + B &= -1 \\ A + \frac{13}{12}B &= 2 \end{aligned}$$

3.4. EXAMPLES

43

Subtracting the bottom equation from the top equation, we get $-\frac{1}{12}B = -3$ or $B = 36$. Thus, $A = -1 - B = -37$. So

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = -37 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-8t} + 36 \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} e^{-7t}$$

Homework

Exercise 3.4.21

Exercise 3.4.22

Exercise 3.4.23

Exercise 3.4.24

Exercise 3.4.25

Let $C^1([0, 10])$ be the set of all functions x on $[0, 10]$ which are continuously differentiable. This is clearly a vector space with scalar multiplication \odot defined by $c \odot y$ which is the function whose value at t is

$$(c \odot y)(t) = cx(t)$$

To define addition of vectors \oplus , we let $x \oplus y$ be the function whose value at t is

$$(x \odot y)(t) = x(t) + y(t)$$

With these operations $C^1([0, 10])$ is a vector space over \mathfrak{R} . Now let $\mathbf{X} = C^1([0, 10]) \times C^1([0, 10])$ which is the collection of all ordered pairs of continuously differentiable functions over \mathfrak{R} . It is easy to see we make this into a vector space by defining \odot and \oplus to act on pairs of objects in the usual way. We can define an inner product on this space by

$$\langle \mathbf{f}, \mathbf{g} \rangle = \int_0^{10} \left\langle \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \right\rangle dt$$

where $\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$ and $\mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$ are arbitrary members of \mathbf{X} . The two solution $y_1(t) = \mathbf{E}_1 e^{-8t}$ and $y_2(t) = \mathbf{E}_2 e^{-7t}$ are members of \mathbf{X} which are linearly independent as when we write down the linearly dependent check, we have

$$a \odot \mathbf{E}_1 y_1 \oplus b \odot \mathbf{E}_2 y_2 = \mathbf{0}$$

This means

$$a \mathbf{E}_1 e^{-8t} + b \mathbf{E}_2 e^{-7t} = \mathbf{0}, : \forall t \geq 0$$

In particular at $t = 0$, we have

$$a \mathbf{E}_1 + b \mathbf{E}_2 = \mathbf{0}, : \forall t \geq 0$$

which is easily seen to force $a = b = 0$ since the eigenvectors \mathbf{E}_1 and \mathbf{E}_2 are linearly independent vectors in \mathfrak{R}^2 . Hence, these functions are linearly independent objects in this vector space. They play the same role as the two dimensional vectors in the plane \mathfrak{R}^2 in our first section called \mathbf{E} and \mathbf{F} . Their span \mathbf{S} is all linear combinations of y_1 and y_2 which is the solution space of this ODE.

Thus, $\{\mathbf{E}_1\mathbf{y}_1, \mathbf{E}_2\mathbf{y}_2\}$ is a basis for a two dimensional subspace inside \mathbf{X} . We can then use Graham Schmidt Orthogonalization to create an orthonormal basis for this solution space.

Listing 3.1: **Computing the Orthonormal Basis**

```
%Convert the eigenvectors to unit vectors
A = [-1;2];
E_1 = A/norm(A)
B = [1; 13/12];
E_2 = B/norm(B);
% Now do GSO
G_1 = E_1;
V = E_2 - dot(E_2,G_1)*G_1;
G_2 = V/norm(V);
>> dot(E_1,E_2)
ans = 0.35389
>> dot(G_1,G_2)
ans = 0
```

We check in this code fragment that \mathbf{E}_1 and \mathbf{E}_2 are not orthogonal but the new basis \mathbf{G} is. Then, the functions $\mathbf{u}_1 = \mathbf{G}_1 e^{-8t}$ and $\mathbf{u}_2 = \mathbf{G}_2 e^{-7t}$ also solve the ODE and we have

$$\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = \int_0^{10} \langle \mathbf{G}_1 e^{-8t}, \mathbf{G}_2 e^{-7t} \rangle dt = \int_0^{10} \langle \mathbf{G}_1, \mathbf{G}_2 \rangle e^{-15t} dt = 0$$

So we have found an orthonormal basis for the solution space.

Note everything we did before in the two dimensional example in \Re^2 carries through exactly. In fact, if we define the mapping $\phi : \text{span}(\{\mathbf{G}_1, \mathbf{G}_2\}) \rightarrow \text{span}(\{\mathbf{G}_1 e^{-8t}, \mathbf{G}_2 e^{-7t}\})$ by

$$\phi(\mathbf{E}) = \mathbf{G}_1 e^{-8t}, \quad \phi(\mathbf{G}_2) = \mathbf{G}_2 e^{-7t}$$

and extend linearly, then ϕ is a 1 – 1 and onto mapping of the span of $\{\mathbf{G}_1, \mathbf{G}_2\}$ to the span of $\{\mathbf{G}_1 e^{-8t}, \mathbf{G}_2 e^{-7t}\}$. Hence, we can identify the solution space of this ODE with the two dimensional space \Re^2 . We can see the set of solutions to a second order linear ODE gives us a two dimensional vector space.

3.4.4.3 Homework

Exercise 3.4.26

Exercise 3.4.27

Exercise 3.4.28

Exercise 3.4.29

Exercise 3.4.30

3.4.5 Three Dimensional Vectors in Space

Now we will use four pencils to motivate what is happening in three dimensions. Pick any two and hold them in your hand. These two pencils determine a plane as long the the pencils are not collinear. This plane does not have to be oriented parallel to our standard planes. Based on what we have said

3.4. EXAMPLES

45

already, the standard basis for \mathbb{R}^3 is

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

where we use these labels for the standard basis instead of the i , j and k you may have seen before. If we want to work in \mathbb{R}^4 or higher dimensions, we need the e_i notation. Pick any axis to orient the \mathbf{e}_1 along. The starting point of the \mathbf{e}_1 vector determines the origin of the three dimensional coordinate system we are constructing. The direction of \mathbf{e}_1 determines the $+x$ axis. The \mathbf{e}_2 vector then starts at the same origin and is positioned so it is orthogonal to \mathbf{e}_1 . The direction of \mathbf{e}_2 fixes the $+y$ axis. We then use the right hand rule to determine the placement of \mathbf{e}_3 . This determines the $+z$ axis. We now have a standard \mathbb{R}^3 coordinate system that uses the standard orthonormal basis. The first two pencils fit in this system. Their common origin doesn't have to match the origin of the \mathbb{R}^3 system we have just constructed but we want our pencils to determine a subspace in \mathbb{R}^3 . So think of their common origin as the same as the one we use for \mathbb{R}^3 as constructed. These two pencils may not have the same length and may not be orthogonal but they determine a plane. If the third pencil lies in this plane, then it is not independent from the first two. However, if it lies out of the plane the three pencils are linearly independent vectors whose span in all three dimensional vectors. Call the first pencil \mathbf{E}_1 , the second \mathbf{E}_2 and the third \mathbf{E}_3 . Then if the fourth pencil is \mathbf{W} , we can find the unique numbers which give its linear combination in terms of the basis $\{\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3\}$ by solving

$$a\mathbf{E}_1 + b\mathbf{E}_2 + c\mathbf{E}_3 = \mathbf{W} \implies [\mathbf{E}_1 \ \mathbf{E}_2 \ \mathbf{E}_3] \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathbf{W}$$

Since these three vectors are linearly independent,

$$a\mathbf{E}_1 + b\mathbf{E}_2 + c\mathbf{E}_3 = \mathbf{0} \implies [\mathbf{E}_1 \ \mathbf{E}_2 \ \mathbf{E}_3] \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathbf{0}$$

only has the solution $a = b = c = 0$. Hence, the kernel of this matrix $\mathbf{T} = [\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3]$ is $\mathbf{0}$ and so \mathbf{T} is invertible. The vectors \mathbf{E}_1 , \mathbf{E}_2 and \mathbf{E}_3 need not be orthogonal, so it is not so easy to find this inverse but what counts here is that the equation

$$a\mathbf{E}_1 + b\mathbf{E}_2 + c\mathbf{E}_3 = \mathbf{W}$$

has a unique solution for each \mathbf{W} .

Linear dependence of the vectors \mathbf{E}_1 , \mathbf{E}_2 and \mathbf{E}_3 is more interesting in \mathbb{R}^3 . If you pick any two vectors, linear dependence of the third on the other two means this:

- \mathbf{E}_3 is in the plane determined by \mathbf{E}_1 and \mathbf{E}_2
- \mathbf{E}_1 is in the plane determined by \mathbf{E}_2 and \mathbf{E}_3
- \mathbf{E}_2 is in the plane determined by \mathbf{E}_1 and \mathbf{E}_3

And if some of the vectors are collinear, the situation degrades

- \mathbf{E}_1 and \mathbf{E}_2 are collinear
- \mathbf{E}_2 and \mathbf{E}_3 are collinear
- \mathbf{E}_1 and \mathbf{E}_3 are collinear

So the span of \mathbf{E}_1 , \mathbf{E}_2 and \mathbf{E}_3 can be one dimensional (all vectors are collinear), two dimensional (two of the three are linearly independent) and three dimensional (all three vectors are linearly independent). In the last case, the vectors form a linearly independent spanning set and so are a basis.

If the vectors are not mutually orthogonal, we can use Graham Schmidt Orthogonalization to create an orthonormal basis as follows:

Listing 3.2: **Computing the 3D Orthonormal Basis**

```

E1 = [ -1;2;3];
E2 = [ 4;5;6];
E3 = [ -3;1;7];
% Now do GSO
G1 = E1/norm(E1);
V = E2 - dot(E2,G1)*G1;
G2 = V/norm(V);
V = E3-dot(E3,G1)*G1-dot(E3,G2)*G2;
G3 = V/norm(V);

```

If $\mathbf{E} = \{\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3\}$ is an orthonormal basis and $\mathbf{F} = \{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3\}$ is another orthonormal basis, the representation of a vector \mathbf{A} can be given with respect to the standard basis, the basis \mathbf{E} or the basis \mathbf{F} . Hence, we have the unique representation

$$\mathbf{A}_{\mathbf{E}} = a_1 \mathbf{E}_1 + a_2 \mathbf{E}_2 + a_3 \mathbf{E}_3 \implies [\mathbf{E}_1 \quad \mathbf{E}_2 \quad \mathbf{E}_3] \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \mathbf{A}_{\mathbf{E}}$$

It is easy to see

$$\begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \\ \mathbf{E}_3^T \end{bmatrix} [\mathbf{E}_1 \quad \mathbf{E}_2 \quad \mathbf{E}_3] = \mathbf{I}$$

Now let's consider

$$[\mathbf{E}_1 \quad \mathbf{E}_2 \quad \mathbf{E}_3] \begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \\ \mathbf{E}_3^T \end{bmatrix} \mathbf{E}_j = [\mathbf{E}_1 \quad \mathbf{E}_2 \quad \mathbf{E}_3] \mathbf{e}_j = \mathbf{E}_j$$

Thus, we see for any vector $a\mathbf{E}_1 + b\mathbf{E}_2 + c\mathbf{E}_3$, we have

$$[\mathbf{E}_1 \quad \mathbf{E}_2 \quad \mathbf{E}_3] \begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \\ \mathbf{E}_3^T \end{bmatrix} (a\mathbf{E}_1 + b\mathbf{E}_2 + c\mathbf{E}_3) = a\mathbf{E}_1 + b\mathbf{E}_2 + c\mathbf{E}_3$$

Hence, the inverse of $[\mathbf{E}_1 \quad \mathbf{E}_2 \quad \mathbf{E}_3]$ is its own transpose.

Finally, since \mathbf{A} has a representation with respect to the basis \mathbf{F} also, We have

$$\begin{aligned} \mathbf{F}_1 &= \langle \mathbf{F}_1, \mathbf{E}_1 \rangle \mathbf{E}_1 + \langle \mathbf{F}_1, \mathbf{E}_2 \rangle \mathbf{E}_2 + \langle \mathbf{F}_1, \mathbf{E}_3 \rangle \mathbf{E}_3 = T_{11}\mathbf{E}_1 + T_{12}\mathbf{E}_2 + T_{13}\mathbf{E}_3 \\ \mathbf{F}_2 &= \langle \mathbf{F}_2, \mathbf{E}_1 \rangle \mathbf{E}_1 + \langle \mathbf{F}_2, \mathbf{E}_2 \rangle \mathbf{E}_2 + \langle \mathbf{F}_2, \mathbf{E}_3 \rangle \mathbf{E}_3 = T_{21}\mathbf{E}_1 + T_{22}\mathbf{E}_2 + T_{23}\mathbf{E}_3 \\ \mathbf{F}_3 &= \langle \mathbf{F}_3, \mathbf{E}_1 \rangle \mathbf{E}_1 + \langle \mathbf{F}_3, \mathbf{E}_2 \rangle \mathbf{E}_2 + \langle \mathbf{F}_3, \mathbf{E}_3 \rangle \mathbf{E}_3 = T_{31}\mathbf{E}_1 + T_{32}\mathbf{E}_2 + T_{33}\mathbf{E}_3 \end{aligned}$$

3.4. EXAMPLES

47

Thus, using our two representations for \mathbf{A} , we have

$$\begin{aligned}\mathbf{A} &= b_1\mathbf{F}_1 + b_2\mathbf{F}_2 + b_3\mathbf{F}_3 = b_1(T_{11}\mathbf{E}_1 + T_{12}\mathbf{E}_2 + T_{13}\mathbf{E}_3) + b_2(T_{21}\mathbf{E}_1 + T_{22}\mathbf{E}_2 + T_{23}\mathbf{E}_3) \\ &\quad + b_3(T_{31}\mathbf{E}_1 + T_{32}\mathbf{E}_2 + T_{33}\mathbf{E}_3)\end{aligned}$$

This tells us

$$\begin{aligned}a_1 &= b_1T_{11} + b_2T_{21} + b_3T_{31} \\ a_2 &= b_1T_{12} + b_2T_{22} + b_3T_{32} \\ a_3 &= b_1T_{13} + b_2T_{23} + b_3T_{33}\end{aligned}$$

This implies

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{21} & T_{31} \\ T_{12} & T_{22} & T_{32} \\ T_{13} & T_{23} & T_{33} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \Rightarrow [\mathbf{A}]_{\mathbf{E}} = \begin{bmatrix} T_{11} & T_{21} & T_{31} \\ T_{12} & T_{22} & T_{32} \\ T_{13} & T_{23} & T_{33} \end{bmatrix} [\mathbf{A}]_{\mathbf{F}}$$

or

$$[\mathbf{A}]_{\mathbf{E}} = \begin{bmatrix} \langle \mathbf{F}_1, \mathbf{E}_1 \rangle & \langle \mathbf{F}_2, \mathbf{E}_1 \rangle & \langle \mathbf{F}_3, \mathbf{E}_1 \rangle \\ \langle \mathbf{F}_1, \mathbf{E}_2 \rangle & \langle \mathbf{F}_2, \mathbf{E}_2 \rangle & \langle \mathbf{F}_3, \mathbf{E}_2 \rangle \\ \langle \mathbf{F}_1, \mathbf{E}_3 \rangle & \langle \mathbf{F}_2, \mathbf{E}_3 \rangle & \langle \mathbf{F}_3, \mathbf{E}_3 \rangle \end{bmatrix} [\mathbf{A}]_{\mathbf{F}}$$

We can do the same sort of analysis and interchange the role of \mathbf{E} and \mathbf{F} to get

$$[\mathbf{A}]_{\mathbf{F}} = \begin{bmatrix} \langle \mathbf{E}_1, \mathbf{F}_1 \rangle & \langle \mathbf{E}_2, \mathbf{F}_1 \rangle & \langle \mathbf{E}_3, \mathbf{F}_1 \rangle \\ \langle \mathbf{E}_1, \mathbf{F}_2 \rangle & \langle \mathbf{E}_2, \mathbf{F}_2 \rangle & \langle \mathbf{E}_3, \mathbf{F}_2 \rangle \\ \langle \mathbf{E}_1, \mathbf{F}_3 \rangle & \langle \mathbf{E}_2, \mathbf{F}_3 \rangle & \langle \mathbf{E}_3, \mathbf{F}_3 \rangle \end{bmatrix} [\mathbf{A}]_{\mathbf{E}}$$

As before, the coefficients (T_{ij}) define a matrix which we will call $\mathbf{T}_{\mathbf{F}\mathbf{E}}$ as these coefficients are completely determined by the basis \mathbf{E} and \mathbf{F} and we are transforming \mathbf{F} components into \mathbf{E} components. Hence, we let

$$\mathbf{T}_{\mathbf{F}\mathbf{E}} = \begin{bmatrix} \langle \mathbf{F}_1, \mathbf{E}_1 \rangle & \langle \mathbf{F}_2, \mathbf{E}_1 \rangle & \langle \mathbf{F}_3, \mathbf{E}_1 \rangle \\ \langle \mathbf{F}_1, \mathbf{E}_2 \rangle & \langle \mathbf{F}_2, \mathbf{E}_2 \rangle & \langle \mathbf{F}_3, \mathbf{E}_2 \rangle \\ \langle \mathbf{F}_1, \mathbf{E}_3 \rangle & \langle \mathbf{F}_2, \mathbf{E}_3 \rangle & \langle \mathbf{F}_3, \mathbf{E}_3 \rangle \end{bmatrix} = \begin{bmatrix} T_{11} & T_{21} & T_{31} \\ T_{12} & T_{22} & T_{23} \\ T_{13} & T_{23} & T_{33} \end{bmatrix} = \mathbf{T}_{\mathbf{E}\mathbf{F}}^T$$

and

$$\mathbf{T}_{\mathbf{E}\mathbf{F}} = \begin{bmatrix} \langle \mathbf{E}_1, \mathbf{F}_1 \rangle & \langle \mathbf{E}_2, \mathbf{F}_1 \rangle & \langle \mathbf{E}_3, \mathbf{F}_1 \rangle \\ \langle \mathbf{E}_1, \mathbf{F}_2 \rangle & \langle \mathbf{E}_2, \mathbf{F}_2 \rangle & \langle \mathbf{E}_3, \mathbf{F}_2 \rangle \\ \langle \mathbf{E}_1, \mathbf{F}_3 \rangle & \langle \mathbf{E}_2, \mathbf{F}_3 \rangle & \langle \mathbf{E}_3, \mathbf{F}_3 \rangle \end{bmatrix} = \begin{bmatrix} T_{11} & T_{21} & T_{31} \\ T_{12} & T_{22} & T_{23} \\ T_{13} & T_{23} & T_{33} \end{bmatrix} = \mathbf{T}_{\mathbf{F}\mathbf{E}}^T$$

So $\mathbf{T}_{\mathbf{F}\mathbf{E}}$ and $\mathbf{T}_{\mathbf{E}\mathbf{F}}$ are transposes of each other. We can again show $\mathbf{T}_{\mathbf{F}\mathbf{E}}^T$ is the inverse of $\mathbf{T}_{\mathbf{F}\mathbf{E}}$ by a calculation similar to the one we did for \mathbb{R}^2 . In fact, since the dimension of the underlying space is not shown, the argument is identical. We note for all vectors in \mathbb{R}^3 ,

$$\mathbf{A}_{\mathbf{E}} = \mathbf{T}_{\mathbf{F}\mathbf{E}} \mathbf{A}_{\mathbf{F}} = \mathbf{T}_{\mathbf{F}\mathbf{E}} \mathbf{T}_{\mathbf{E}\mathbf{F}} \mathbf{A}_{\mathbf{E}} = \mathbf{T}_{\mathbf{F}\mathbf{E}} \mathbf{T}_{\mathbf{F}\mathbf{E}}^T \mathbf{A}_{\mathbf{E}}$$

and

$$\mathbf{A}_{\mathbf{F}} = \mathbf{T}_{\mathbf{E}\mathbf{F}} \mathbf{A}_{\mathbf{E}} = \mathbf{T}_{\mathbf{F}\mathbf{E}}^T \mathbf{T}_{\mathbf{F}\mathbf{E}} \mathbf{A}_{\mathbf{F}} = \mathbf{T}_{\mathbf{F}\mathbf{E}}^T \mathbf{T}_{\mathbf{F}\mathbf{E}} \mathbf{A}_{\mathbf{E}}$$

Hence $\mathbf{T}_{FE} \mathbf{T}_{FE}^T = \mathbf{T}_{FE}^T \mathbf{T}_{FE} = \mathbf{I}$ and \mathbf{T}_{FE}^T is the inverse of \mathbf{T}_{FE} .

The calculation that in inner product in \Re^3 in invariant under a change of orthonormal basis is then the same. Also, the arguments we use here can be adapted with little change to the \Re^n setting.

3.4.5.1 Homework

Exercise 3.4.31

Exercise 3.4.32

Exercise 3.4.33

Exercise 3.4.34

Exercise 3.4.35

3.4.6 The Solution Space of Higher Dimensional ODE Systems

Let's do a final example to show you how \Re^6 can be identified with the solution space of a linear system of ODEs. A general model of enhanced cytokine signalling production based on LRRK2 mutation is as follows. The altered cytokine response starts with the activation of an allele called A , in a small compartment of cells. Initially, all cells have a correct version of the **LRRK2** gene. We will denote this by $A^{-/-}$ where the superscript “ $-/-$ ” indicates there are no LRRK2 mutations. One of the mutant alleles becomes activated at mutation rate u_1 to generate a cell type denoted by $A^{+/-}$. The superscript $+/-$ tells us one allele is activated. The second allele becomes activated at rate \hat{u}_2 to become the cell type $A^{+/+}$. In addition, $A^{-/-}$ cells can also receive mutations that trigger **CIN**. This happens at the rate u_c resulting in the cell type $A^{-/- CIN}$. This kind of a cell can activate the first allele of the LRRK2 gene with normal mutation rate u_1 to produce a cell with one activated allele (i.e. a $+/-$) which started from a **CIN** state. We denote these cells as $A^{+/- CIN}$. We can also get a cell of type $A^{+/- CIN}$ when a cell of type $A^{+/-}$ receives a mutation which triggers **CIN**. We will assume this happens at the same rate u_c as before. The $A^{+/- CIN}$ cell then rapidly undergoes **LOH** at rate \hat{u}_3 to produce cells having the second allele of LRRK2 which is of type $A^{+/+ CIN}$. Finally, $A^{+/+}$ cells can experience **CIN** at rate u_c to generate $A^{+/+ CIN}$ cells. We show this information in Figure 3.1.

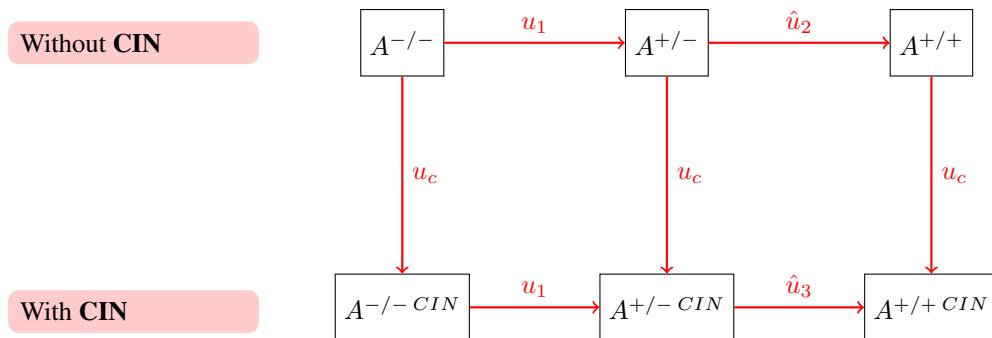


Figure 3.1: The pathways for the **LRRK2** allele gains.

Let N be the population size within which the LRRK2 mutations occur. We will assume a typical value of N is 10^3 to 10^4 . The first allele is activated by a point mutation. The rate at which this occurs is modeled by the rate u_1 as shown in Figure 3.1. We make the following assumptions:

3.4. EXAMPLES

49

- the mutations governed by the rates u_1 and u_c are **neutral**. This means that these rates do not depend on the size of the population N .
- The events governed by \hat{u}_2 and \hat{u}_3 give what is called **selective advantage**. This means that the size of the population size does matter.

Using these assumptions, we will model \hat{u}_2 and \hat{u}_3 as $\hat{u}_2 = N u_2$ and $\hat{u}_3 = N u_3$, where u_2 and u_3 are neutral rates. We can thus redraw our figure as Figure 3.2.

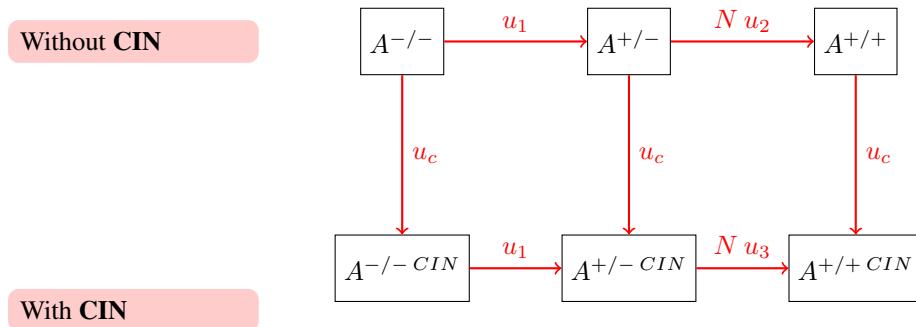


Figure 3.2: The pathways for the **LRRK2** allele gains rewritten using selective advantage.

The mathematical model is then setup as follows. Let

$X_0(t)$ is the probability a cell in in cell type $A^{-/-}$ at time t .

$X_1(t)$ is the probability a cell in in cell type $A^{+/-}$ at time t .

$X_2(t)$ is the probability a cell in in cell type $A^{+/+}$ at time t .

$Y_0(t)$ is the probability a cell in in cell type $A^{-/-} CIN$ at time t .

$Y_1(t)$ is the probability a cell in in cell type $A^{+/-} CIN$ at time t .

$Y_2(t)$ is the probability a cell in in cell type $A^{+/+} CIN$ at time t .

Looking at Figure 3.2, we can generate rate equations. First, let's rewrite Figure 3.2 using our variables as Figure 3.3. To generate the equations we need, note each box has arrows coming into

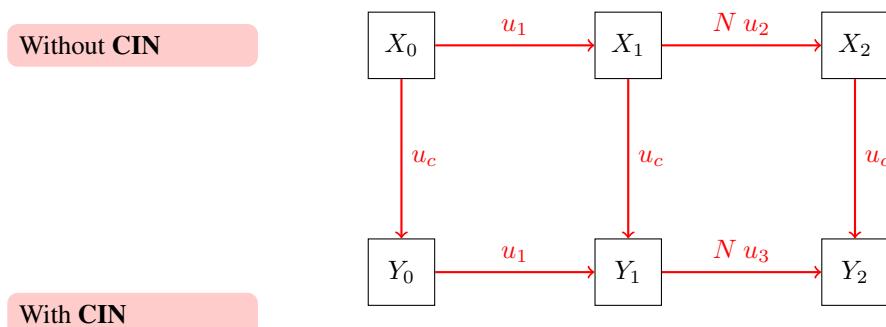


Figure 3.3: The pathways for the **LRRK2** allele gains rewritten using mathematical variables.

it and arrows coming out of it. The **arrows in** are **growth** terms for the net change of the variable in the box and the **arrows out** are the **decay or loss** terms. We model **growth** as **exponential growth** and **loss** as **exponential decay**. So X_0 only has arrows going out which tells us it only has **loss** terms. So we would say $(X'_0)_{loss} = -u_1 X_0 - u_c X_0$ which implies $X'_0 = -(u_1 + u_c) X_0$. Further, X_1 has arrows going in and out which tells us it has **growth** and **loss** terms. So we would say $(X'_1)_{loss} = -Nu_2 X_1 - u_c X_1$ and $(X'_1)_{growth} = u_1 X_0$ which implies $X'_1 = u_1 X_0 - (Nu_2 + u_c) X_1$. We can continue in this way to find all the model equations. We can then see the rate equations are

$$X'_0 = -(u_1 + u_c) X_0 \quad (3.1)$$

$$X'_1 = u_1 X_0 - (u_c + Nu_2) X_1 \quad (3.2)$$

$$X'_2 = Nu_2 X_1 - u_c X_2 \quad (3.3)$$

$$Y'_0 = u_c X_0 - u_1 Y_0 \quad (3.4)$$

$$Y'_1 = u_c X_1 + u_1 Y_0 - Nu_3 Y_1 \quad (3.5)$$

$$Y'_2 = Nu_3 Y_1 + u_c X_2 \quad (3.6)$$

Initially, at time 0, all the cells are in the state X_0 , so we have

$$X_0(0) = 1, \quad X_1(0) = 0, \quad X_2(0) = 0 \quad (3.7)$$

$$Y_0(0) = 0, \quad Y_1(0) = 0, \quad Y_2(0) = 0. \quad (3.8)$$

Now under what circumstances is the CIN pathway to LRRK2 mutations the dominant one? In order to answer this, we need to analyze the trajectories of this model. We can rewrite this as a matrix - vector system:

$$\begin{bmatrix} X'_0 \\ X'_1 \\ X'_2 \\ Y'_0 \\ Y'_1 \\ Y'_2 \end{bmatrix} = \begin{bmatrix} -(u_1 + u_c) & 0 & 0 & 0 & 0 & 0 \\ u_1 & -(u_c + Nu_2) & 0 & 0 & 0 & 0 \\ 0 & Nu_2 & -u_c & 0 & 0 & 0 \\ u_c & 0 & 0 & -u_1 & 0 & 0 \\ 0 & u_c & 0 & u_1 & -Nu_3 & 0 \\ 0 & 0 & u_c & 0 & Nu_3 & 0 \end{bmatrix} \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ Y_0 \\ Y_1 \\ Y_2 \end{bmatrix}$$

The eigenvalues of this system are the roots of the polynomial $p(\lambda) = \det(\lambda I - A)$ where A is the 6×6 coefficient matrix above. We find

$$p(\lambda) = \det \begin{bmatrix} \lambda + (u_1 + u_c) & 0 & 0 & 0 & 0 & 0 \\ -u_1 & \lambda + (u_c + Nu_2) & 0 & 0 & 0 & 0 \\ 0 & -Nu_2 & \lambda + u_c & 0 & 0 & 0 \\ -u_c & 0 & 0 & \lambda + u_1 & 0 & 0 \\ 0 & -u_c & 0 & -u_1 & \lambda + Nu_3 & 0 \\ 0 & 0 & -u_c & 0 & -Nu_3 & \lambda \end{bmatrix}$$

This is easily expanded using the properties of determinants (which we will discuss later!) to give

$$p(\lambda) = (\lambda + (u_1 + u_c)) \begin{bmatrix} \lambda + (u_c + Nu_2) & 0 & 0 & 0 & 0 \\ -Nu_2 & \lambda + u_c & 0 & 0 & 0 \\ 0 & 0 & \lambda + u_1 & 0 & 0 \\ -u_c & 0 & -u_1 & \lambda + Nu_3 & 0 \\ 0 & -u_c & 0 & -Nu_3 & \lambda \end{bmatrix}$$

3.4. EXAMPLES

51

$$\begin{aligned}
&= (\lambda + (u_1 + u_c)) (\lambda + (u_c + Nu_2)) \begin{bmatrix} \lambda + u_c & 0 & 0 & 0 \\ 0 & \lambda + u_1 & 0 & 0 \\ 0 & -u_1 & \lambda + Nu_3 & 0 \\ -u_c & 0 & -Nu_3 & \lambda \end{bmatrix} \\
&= (\lambda + (u_1 + u_c)) (\lambda + (u_c + Nu_2)) (\lambda + u_c) \begin{bmatrix} \lambda + u_1 & 0 & 0 \\ -u_1 & \lambda + Nu_3 & 0 \\ 0 & -Nu_3 & \lambda \end{bmatrix} \\
&= (\lambda + (u_1 + u_c)) (\lambda + (u_c + Nu_2)) (\lambda + u_c)(\lambda + u_1) (\lambda + Nu_3) (\lambda)
\end{aligned}$$

Hence, the eigenvalues are

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \end{bmatrix} = \begin{bmatrix} -(u_1 + u_c) \\ -(u_c + Nu_2) \\ -u_c \\ -u_1 \\ -Nu_3 \\ 0 \end{bmatrix}$$

Since this is a biological model, we get more insight into finding a solution with the critical parameters in this form rather than substituting numerical values and using a tool like MatLab. It is possible to calculate the six needed eigenvectors by hand for this system which we enjoyed doing but we understand not all share this interest. We find the eigenvectors are

$$E_1 = \begin{bmatrix} Nu_2 - u_1 \\ u_1 \\ -Nu_2 \\ -(Nu_2 - u_1) \\ -u_1 \frac{Nu_2 - u_1 - u_c}{Nu_3 - u_1 - u_c} \\ \frac{Nu_3(Nu_2 - u_1) - Nu_2 u_c}{Nu_3 - u_1 - u_c} \end{bmatrix}, \quad E_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \\ \frac{-u_c}{Nu_2 + u_c - Nu_3} \\ \frac{u_c}{Nu_2 + u_c - Nu_3} \end{bmatrix}, \quad E_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$E_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \frac{u_1}{Nu_3 - u_1} \\ \frac{Nu_3 - u_1}{Nu_3} \end{bmatrix}, \quad E_5 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}, \quad E_6 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

The general solution is thus any linear combination of the form

$$c_1 E_1 e^{-(u_1+u_c)t} + c_2 E_2 e^{-(u_c+N u_2)t} + c_3 E_3 e^{-u_c t} + c_4 E_4 e^{-u_1 t} + c_5 E_5 e^{-N u_3 t} + c_6 E_6 1$$

where $\mathbf{1}$ denotes the constant function $e^{0t} = 1$. If we let these six solutions be denoted by $\mathbf{y}_i(t) = \mathbf{E}_i e^{\lambda_i t}$, we see the general solution is a member of the span of $\{\mathbf{y}_1, \dots, \mathbf{y}_6\}$ and since these functions are linearly independent in the space $\mathbf{X} = C^1([0, T] \times C^1([0, T] \times$ for any appropriate T , we know this solution space is a six dimensional subspace of \mathbf{X} with the basis $\{\mathbf{y}_1, \dots, \mathbf{y}_6\}$. If \mathbf{F} and \mathbf{G} are two elements in \mathbf{X} , it is easy to show we can define an inner product on \mathbf{X} to be

$$\int_0^T \langle F(t), G(t) \rangle dt$$

We can construct an orthonormal basis for the solution space by applying GSO to the eigenvectors E_i to create the orthonormal basis G . Then the new functions $w_i = G_i e^{\lambda_i t}$ are solutions to the

ODE system which are mutually orthogonal. So it is straightforward to find an orthonormal basis here.

Note this solution space can be identified with any six dimensional subspace of any vector space by simply mapping one orthonormal basis to another and extending linearly.

It is not our intent here to pursue this problem further. It turns out looking at the solution space this way is not so helpful to understand what is going on. Finding approximate solutions using standard Taylor series expansions is better. You can read about this in other places if you are interested.

3.4.6.1 Homework

Exercise 3.4.36

Exercise 3.4.37

Exercise 3.4.38

Exercise 3.4.39

Exercise 3.4.40

3.5 Best Approximation in a Vector Space With Inner Product

An important problem is that of the idea of finding the best object in an subspace \mathcal{W} that approximates a given object . This is an easy theorem to prove.

Theorem 3.5.1 The Finite Dimensional Approximation Theorem

Let p be any object in the inner product space \mathcal{V} with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$. Let \mathcal{W} be a finite dimensional subspace with an orthonormal basis $\mathbf{E} = \{\mathbf{E}_1, \dots, \mathbf{E}_n\}$ where n is the dimension of the subspace. Then there is an unique object p^* in \mathcal{W} which satisfies

$$\|p - p^*\| = \min_{u \in \mathcal{W}} \|u - p\|$$

with

$$p^* = \sum_{i=1}^N \langle p, \mathbf{E}_i \rangle \mathbf{E}_i.$$

Further, $p - p^*$ is orthogonal to the subspace \mathcal{W} ; i.e. $\langle p^*, u \rangle = 0$ for all $u \in \mathcal{W}$.

Proof 3.5.1

Any object in the subspace has the representation $\sum_{i=1}^N a_i \mathbf{E}_i$ for some scalars a_i . Consider the function of N variables

$$\begin{aligned} E(a_1, \dots, a_N) &= \left\langle p - \sum_{i=1}^N a_i \mathbf{E}_i, p - \sum_{j=1}^N a_j \mathbf{E}_j \right\rangle \\ &= \langle p, p \rangle - 2 \sum_{i=1}^N a_i \langle p, \mathbf{E}_i \rangle \\ &\quad + \sum_{i=1}^N \sum_{j=1}^N a_i a_j \langle \mathbf{E}_i, \mathbf{E}_j \rangle. \end{aligned}$$

3.5. BEST APPROXIMATION IN A VECTOR SPACE WITH INNER PRODUCT

53

Simplifying using the orthonormality of the basis, we find

$$E(a_1, \dots, a_N) = \langle \mathbf{p}, \mathbf{p} \rangle - 2 \sum_{i=1}^N a_i \langle \mathbf{p}, \mathbf{E}_i \rangle + \sum_{i=1}^N a_i^2.$$

This is a quadratic expression and setting the gradient of E to zero, we find the critical points $a_j = \langle \mathbf{p}, \mathbf{E}_j \rangle$. This is a global minimum for the function E . Hence, the optimal \mathbf{p}^* has the form

$$\mathbf{p}^* = \sum_{i=1}^N \langle \mathbf{p}, \mathbf{E}_i \rangle \mathbf{E}_i.$$

Finally, we see

$$\begin{aligned} \langle \mathbf{p} - \mathbf{p}^*, \mathbf{E}_j \rangle &= \langle \mathbf{p}, \mathbf{E}_j \rangle - \sum_{k=1}^N \langle \mathbf{p}, \mathbf{E}_k \rangle \langle \mathbf{E}_k, \mathbf{E}_j \rangle \\ &= \langle \mathbf{p}, \mathbf{E}_j \rangle - \langle \mathbf{p}, \mathbf{E}_j \rangle = 0, \end{aligned}$$

and hence, $\mathbf{p} - \mathbf{p}^*$ is orthogonal of \mathcal{W} . ■

This is an extension of the Cauchy - Schwartz Theorem in a way. It was easy to see how to handle the best approximation idea when only two vectors were involved so we didn't need all the machinery above. But this is a powerful idea. We can put this into perspective by recalling some basic facts from Fourier Series.

Letting $u_n(x) = \sin(\frac{n\pi}{L}x)$ and using the standard inner product on $C([0, L])$, $\langle f, g \rangle = \int_0^L f(t)g(t)dt$, we know

$$\langle u_i, u_j \rangle = \begin{cases} \frac{L}{2}, & i = j \\ 0, & i \neq j. \end{cases}$$

We define the new functions \hat{u}_n by $\sqrt{\frac{2}{L}}u_n$. Then, $\langle \hat{u}_i, \hat{u}_j \rangle = \delta_i^j$ and the sequence of functions (\hat{u}_n) are all mutually orthogonal. It is clear $\|\hat{u}_n\|_2 = 1$ always. So the sequence of functions (\hat{u}_n) are all mutually orthogonal and length one. Letting $v_n(x) = \cos(\frac{n\pi}{L}x)$, we also know

$$\begin{aligned} \langle v_0, v_0 \rangle &= \frac{1}{L} \\ \langle v_i, v_j \rangle &= \begin{cases} \frac{L}{2}, & i = j \\ 0, & i \neq j. \end{cases} \end{aligned}$$

Hence, we can define the new functions

$$\begin{aligned} \hat{v}_0(x) &= \sqrt{\frac{1}{L}} \\ \hat{v}_n(x) &= \sqrt{\frac{2}{L}} \cos\left(\frac{n\pi}{L}x\right), n \geq 1, \end{aligned}$$

Then, $\langle \hat{v}_i, \hat{v}_j \rangle = \delta_i^j$ and the sequence of functions (\hat{v}_n) are all mutually orthogonal with length $\|\hat{v}_n\| = 1$.

If we start with a function f which is continuous on the interval $[0, L]$, we can define the trigonometric series associated with f as follows

$$\begin{aligned} S(x) &= \frac{1}{L} \langle f, \mathbf{1} \rangle \\ &\quad + \sum_{i=1}^{\infty} \left(\frac{2}{L} \left\langle f(x), \sin\left(\frac{i\pi}{L}x\right) \right\rangle \sin\left(\frac{i\pi}{L}x\right) + \frac{2}{L} \left\langle f(x), \cos\left(\frac{i\pi}{L}x\right) \right\rangle \cos\left(\frac{i\pi}{L}x\right) \right). \end{aligned}$$

In terms of the orthonormal families we have just defined, this can be rewritten as

$$S(x) = \lim_{N \rightarrow \infty} \left(\langle f, \hat{v}_0 \mathbf{1} \rangle \hat{v}_0 + \sum_{n=1}^N \langle f(x), \hat{v}_n \rangle \hat{v}_n + \sum_{n=1}^N \langle f(x), \hat{u}_n \rangle \hat{u}_n \right)$$

Note the projection of the data function f , $P_N(f)$, onto the subspace spanned by $\{\hat{u}_1, \dots, \hat{u}_N\}$ is the minimal norm solution to the problem $\inf_{u \in V_n} \|f - u\|_2$ where V_n is the span of $\{\hat{u}_1, \dots, \hat{u}_N\}$.

Further, projection of the data function f , $Q_N(f)$, onto the subspace spanned by $\{1, \hat{v}_1, \dots, \hat{v}_N\}$ is the minimal norm solution to the problem $\inf_{v \in W_n} \|f - v\|_2$ where W_n is the span of $\{\hat{v}_0, \dots, \hat{v}_N\}$.

The Fourier Series of $f : [0, 2L]$ has partial sums that can be written as

$$S_N = P_N(f) + Q_N(f) = \sum_{n=1}^N \langle f, \hat{u}_n \rangle \hat{u}_n + \langle f, \hat{v}_0 \rangle \hat{v}_0 + \sum_{n=1}^N \langle f, \hat{v}_n \rangle \hat{v}_n$$

and so the convergence of the Fourier Series is all about the conditions under which the projections of f to these subspaces converge pointwise to f .

3.5.1 Homework

Exercise 3.5.1

Exercise 3.5.2

Exercise 3.5.3

Exercise 3.5.4

Exercise 3.5.5

Chapter 4

Linear Transformations

We are now going to study matrices and learn to interpret them in a more general way.

4.1 Organizing Point Cloud Data

The first place to start is what is called a **point cloud** of data. This is a collection in which each data point is itself a list of real numbers. So if \mathbf{x} is such a data point, the list of numbers associated with it is $\{x_1, \dots, x_n\}$ where n is the number of components in the data. For example, there is a complicated system of inferring what individual molecules are doing in fluid flow such as in an artery called flow cytometry. Roughly speaking, lots of molecules are separated as they *flow* down a column filled with liquid. The separation can be due to mass, charge and other things. Then laser light of various frequencies is shown through the liquid with the moving molecules and how the molecules absorb the light is collected and stored as lists of numbers. So a particular molecular complex that is of interest might have a list of 14 numbers associated with it and a list like this is obtained at regular time points. So if you collect data every millisecond, you have 1000 sets of these lists per molecular complex per second. A typical complex we might wish to track in the context of how signals are used for information processing in biology would be one of several cytokine molecules. A given flow cytometry experiment might collect such data on 10 different cytokines for 3 seconds. If we label the cytokines as C_1, \dots, C_{10} , we have data of the form

$$\{\mathbf{X}_{i1}, \dots, \mathbf{X}_{i,3000}\}$$

for each cytokine of type C_i for time points 1 millisecond to 3000 milliseconds. This vast collection of data is a set of lists with 14 components. Hence, we know $\mathbf{X}_{ij} \in \mathbb{R}^{14}$, the collection of real numbers in 14 dimensions. Note, we can collect the data and store it as tables of numbers without referring to \mathbb{R}^{14} as a vector space with a specific orthonormal basis. The easiest way to organize this is to think of the initial data collection phase as generating vectors in \mathbb{R}^{14} using the standard orthonormal basis

$$\mathbf{e}_1 = \begin{bmatrix} 0, & i \neq 1 \\ 1, & i = 1 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0, & i \neq 2 \\ 1, & i = 2 \end{bmatrix}, \dots, \mathbf{e}_{14} = \begin{bmatrix} 0, & i \neq 14 \\ 1, & i = 14 \end{bmatrix},$$

To make sense of this collection of data, we want to find ways to **transform** the given **point cloud** into a new version of itself which we can use to understand the data better. We usually do this by applying what are called **Linear Transformations** which are mappings from \mathbb{R}^{14} into itself which satisfy a property called **linearity**. We find the use of a change of basis from one orthonormal basis to another and the use of specialized orthonormal basis constructed from eigenvalues and eigenvectors to be very useful. So let's look at this transformation problem in general.

4.1.1 Homework

Exercise 4.1.1

Exercise 4.1.2

Exercise 4.1.3

Exercise 4.1.4

Exercise 4.1.5

4.2 Linear Transformations

Given a basis E of \mathbb{R}^n , $E = \{e_1, \dots, e_n\}$, we know any vector x in \mathbb{R}^n has a decomposition $[x]_E$ where the components of x with respect to the basis E are x_i with

$$x = x_1 e_1 + \dots + x_n e_n$$

We often wish to **measure** the magnitude or size of a vector x . We have talked about doing this before and have used the term **norm** to denote the mapping that does the job. Let's be specific now.

Definition 4.2.1 Norm on a Vector Space

Let V be a vector space over the reals and let $\rho : V \rightarrow \mathbb{R}$ satisfy

N1: $\rho(x) \geq 0$ for all $x \in V$.

N2: $\rho(x) = 0$ if and only if $x = 0$

N3: $\rho(\alpha x) = |\alpha| \rho(x)$ for all real numbers α and $x \in V$

N4: $\rho(x + y) \leq \rho(x) + \rho(y)$ for all $x, y \in V$

Comment 4.2.1 We usually denote the norm ρ by the symbol $\|\cdot\|$ and sometimes add a subscript to remind us where the norm comes from. We will see many examples of this soon.

Comment 4.2.2 As we have mentioned before, if the vector space V has an inner product, the mapping $\rho(x) = \langle x, x \rangle$ defines a norm on V . The vector space V plus an inner product $\langle \cdot, \cdot \rangle$ is denoted as the pair $(V, \langle \cdot, \cdot \rangle)$ and is called an **inner product space**. The vector space V plus a norm ρ is denoted as the pair (V, ρ) and is called an **normed linear space**. Hence, an inner product space induces a norm on the vector space. It is known the **not all norms are induced by some inner product**. Infinite dimensional vector spaces are known where this is not true, but that is a story for another time.

Comment 4.2.3 For the vector space $C([a, b])$,

- the inner product is $\langle f, g \rangle = \int_a^b f(t)g(t)dt$ with induced norm $\|f\|_2 = \sqrt{\int_a^b f^2(t)dt}$.
- we can also use the norm $\|f\|_1 = \int_a^b |f(t)|dt$ which is not induced from an inner product.
- we can also use the norm $\|f\|_1 = \max_{a \leq t \leq b} |f(t)|$.

4.2.1 Homework

Exercise 4.2.1

Exercise 4.2.2**Exercise 4.2.3****Exercise 4.2.4****Exercise 4.2.5**

4.3 Sequence Spaces Revisited

Let's recall some ideas about sequence spaces. Given a sequence of real numbers (a_n) , which we assume has indexing starting at $n = 1$ for convenience, the set of all sequences has a variety of interesting subsets which might even be subspaces. It is clear the set of all sequences is a vector space over the reals but whether or not a subset is a subspace depends on whether the subset is closed under scalar multiplication and vector addition. Now we say the positive numbers p and q are **conjugate exponents** if $p > 1$ and $1/p + 1/q = 1$. If $p = 1$, we define its conjugate exponent to be $q = \infty$. Conjugate exponents satisfy some fundamental identities. Clearly, if $p > 1$, $\frac{1}{p} + \frac{1}{q} \implies 1 = \frac{p+q}{pq}$ and also $pq = p + q$ and $(p - 1)(q - 1) = 1$. We will use these identities quite a bit. We quote the following lemma whose proof is in many texts such as (Peterson (8) 2019).

Lemma 4.3.1 The $\alpha - \beta$ Lemma

Let α and β be positive real numbers and p and q be conjugate exponents. Then $\alpha \beta \leq \frac{\alpha^p}{p} + \frac{\beta^q}{q}$.

Proof 4.3.1

You should look at the proof in (Peterson (8) 2019) to refresh your memory of how this is done. ■

Important subsets of the set of all sequences can then be defined using conjugate exponents.

Definition 4.3.1 The ℓ^p Sequence Space

Let $p \geq 1$. The collection of all sequences, $(a_n)_{n=1}^\infty$ for which $\sum_{n=1}^\infty |a_n|^p$ converges is denoted by the symbol ℓ^p .

- (1) $\ell^1 = \{(a_n)_{n=1}^\infty : \sum_{n=1}^\infty |a_n| \text{ converges.}\}$
- (2) $\ell^2 = \{(a_n)_{n=1}^\infty : \sum_{n=1}^\infty |a_n|^2 \text{ converges.}\}$

We also define $\ell^\infty = \{(a_n)_{n=1}^\infty : \sup_{n \geq 1} |a_n| < \infty\}$.

There is a fundamental inequality connecting sequences in ℓ^p and ℓ^q when p and q are conjugate exponents called **Hölder's Inequality**. Its proof is straightforward but has a few tricks.

Theorem 4.3.2 Hölder's Inequality

Let $p > 1$ and p and q be conjugate exponents. If $x \in \ell^p$ and $y \in \ell^q$, then

$$\sum_{n=1}^\infty |x_n y_n| \leq \left(\sum_{n=1}^\infty |x_n|^p \right)^{1/p} \left(\sum_{n=1}^\infty |y_n|^q \right)^{1/q}$$

where $x = (x_n)$ and $y = (y_n)$.

Proof 4.3.2

This inequality is clearly true if either of the two sequences x and y are the zero sequence. So we can

assume both x and y have some nonzero terms in them. Then $x \in \ell^p$, we know

$$0 < u = \left(\sum_{n=1}^{\infty} |x_n|^p \right)^{1/p} < \infty, \quad 0 < v = \left(\sum_{n=1}^{\infty} |y_n|^q \right)^{1/q} < \infty$$

Now define new sequences, \hat{x} and \hat{y} by $\hat{x}_n = x_n/u$ and $\hat{y}_n = y_n/v$. Then, we have

$$\begin{aligned} \sum_{n=1}^{\infty} |\hat{x}_n|^p &= \sum_{n=1}^{\infty} \frac{|x_n|^p}{u^p} = \frac{1}{u^p} \sum_{n=1}^{\infty} |x_n|^p = \frac{u^p}{u^p} = 1. \\ \sum_{n=1}^{\infty} |\hat{y}_n|^q &= \sum_{n=1}^{\infty} \frac{|y_n|^q}{v^q} = \frac{1}{v^q} \sum_{n=1}^{\infty} |y_n|^q = \frac{v^q}{v^q} = 1. \end{aligned}$$

Now apply the $\alpha - \beta$ Lemma to $\alpha = |\hat{x}_n|$ and $\beta = |\hat{y}_n|$ for any nonzero terms \hat{x}_n and \hat{y}_n . Then $|\hat{x}_n \hat{y}_n| \leq |\hat{x}_n|^p/p + |\hat{y}_n|^q/q$.

This is also true, of course, if either \hat{x}_n or \hat{y}_n are zero although the $\alpha - \beta$ lemma does not apply!

Now sum over N terms to get

$$\sum_{n=1}^N |\hat{x}_n \hat{y}_n| \leq \frac{1}{p} \sum_{n=1}^N |\hat{x}_n|^p + \frac{1}{q} \sum_{n=1}^N |\hat{y}_n|^q$$

Since we know $x \in \ell^p$ and $y \in \ell^q$, we know

$$\begin{aligned} \sum_{n=1}^N |\hat{x}_n|^p &\leq \sum_{n=1}^{\infty} |\hat{x}_n|^p = 1 \\ \sum_{n=1}^N |\hat{y}_n|^q &\leq \sum_{n=1}^{\infty} |\hat{y}_n|^q = 1 \end{aligned}$$

So we have

$$\sum_{n=1}^N |\hat{x}_n \hat{y}_n| \leq \frac{1}{p} + \frac{1}{q} = 1$$

This is true for all N so the partial sums $\sum_{n=1}^N |\hat{x}_n \hat{y}_n|$ are bounded above. Hence, the partial sums converge to this supremum which is denoted by $\sum_{n=1}^{\infty} |\hat{x}_n \hat{y}_n|$. We conclude $\sum_{n=1}^{\infty} |\hat{x}_n \hat{y}_n| \leq 1$. But $\hat{x}_n \hat{y}_n = 1/(u v) x_n y_n$ and so we have $\frac{1}{u v} \sum_{n=1}^{\infty} |x_n y_n| \leq 1$ which implies the result as

$$\sum_{n=1}^{\infty} |x_n y_n| \leq u v = \left(\sum_{n=1}^{\infty} |x_n|^p \right)^{1/p} \left(\sum_{n=1}^{\infty} |y_n|^q \right)^{1/q}$$

■

We can also do this inequality for the case $p = 1$ and $q = \infty$.

Theorem 4.3.3 Hölder’s Theorem for $p = 1$ and $q = \infty$

If $x \in \ell^1$ and $y \in \ell^{\infty}$, then $\sum_{n=1}^{\infty} |x_n y_n| \leq \left(\sum_{n=1}^{\infty} |x_n| \right) \sup_{n \geq 1} |y_n|$.

Proof 4.3.3

We know since $y \in \ell^\infty$, $|y_n| \leq \sup_{k \geq 1} |y_k|$. Thus,

$$\sum_{n=1}^N |x_n y_n| \leq \left(\sum_{n=1}^{\infty} |x_n| \right) \sup_{k \geq 1} |y_k|$$

Thus the sequence of partial sums $\sum_{n=1}^N |x_n y_n|$ is bounded above by $\left(\sum_{n=1}^{\infty} |x_n| \right) \sup_{k \geq 1} |y_k|$. This gives us our result. \blacksquare

Now we want to apply these ideas to \Re^n which is a collection of numbers organized as vectors with n components. Here, we don't really care what the underlying orthonormal basis E which gives us these components is. Note the vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

can be identified with the sequence

$$(x_j) = \{x_1, \dots, x_n, 0, \dots\}$$

It is easy to see that since (x_j) has only n components, this sequence is in any ℓ^p . Hölder's Inequality specialized to \Re^n gives

Theorem 4.3.4 Hölder's Inequality in \Re^n

Let $p > 1$ and p and q be conjugate exponents. If $\mathbf{x} \in \Re^n$, then the associated sequence $(x_j) = \{x_1, \dots, x_n, 0, \dots\}$ is in ℓ^p for all p . We have for any two such sequences (x_j) and (y_j)

$$\sum_{i=1}^n |x_i y_i| \leq \left(\sum_{j=1}^n |x_j|^p \right)^{1/p} \left(\sum_{j=1}^n |y_j|^q \right)^{1/q}$$

$$\text{and } \sum_{j=1}^n |x_j y_j| \leq \left(\sum_{j=1}^n |x_j| \right) \max_{1 \leq j \leq n} |y_j|.$$

There is an associated inequality called Minkowski's Inequality which is also useful. The general version is

Theorem 4.3.5 Minkowski's Inequality

Let $p \geq 1$ and let x and y be in ℓ^p . Then,

$$\left(\sum_{n=1}^{\infty} |x_n + y_n|^p \right)^{\frac{1}{p}} \leq \left(\sum_{n=1}^{\infty} |x_n|^p \right)^{\frac{1}{p}} + \left(\sum_{n=1}^{\infty} |y_n|^p \right)^{\frac{1}{p}}$$

and for x and y in ℓ^∞ ,

$$\sup_{n \geq 1} |x_n + y_n| \leq \sup_{n \geq 1} |x_n| + \sup_{n \geq 1} |y_n|$$

Proof 4.3.4

(1): $p = \infty$

We know $|x_n + y_n| \leq |x_n| + |y_n|$ by the triangle inequality. So we have $|x_n + y_n| \leq \sup_{n \geq 1} |x_n| + \sup_{n \geq 1} |y_n|$. Thus, the right hand side is an upper bound for all the terms of the left side. We then can say $\sup_{n \geq 1} |x_n + y_n| \leq \sup_{n \geq 1} |x_n| + \sup_{n \geq 1} |y_n|$ which is the result for $p = \infty$.

(2): $p = 1$

Again, we know $|x_n + y_n| \leq |x_n| + |y_n|$ by the triangle inequality. Sum the first N terms on both sides to get

$$\sum_{n=1}^N |x_n + y_n| \leq \sum_{n=1}^N |x_n| + \sum_{n=1}^N |y_n| \leq \sum_{n=1}^{\infty} |x_n| + \sum_{n=1}^{\infty} |y_n|$$

The right hand side is an upper bound for the partial sums on the left. Hence, we have

$$\sum_{n=1}^{\infty} |x_n + y_n| \leq \sum_{n=1}^{\infty} |x_n| + \sum_{n=1}^{\infty} |y_n|$$

(3) $1 < p < \infty$

We have

$$\begin{aligned} |x_n + y_n|^p &= |x_n + y_n| |x_n + y_n|^{p-1} \leq |x_n| |x_n + y_n|^{p-1} + |y_n| |x_n + y_n|^{p-1} \\ \sum_{n=1}^N |x_n + y_n|^p &\leq \sum_{n=1}^N |x_n| |x_n + y_n|^{p-1} + \sum_{n=1}^N |y_n| |x_n + y_n|^{p-1} \end{aligned}$$

Let $a_n = |x_n|$, $b_n = |x_n + y_n|^{p-1}$, $c_n = |y_n|$ and $d_n = |x_n + y_n|^{p-1}$. Hölder's Inequality applies just fine to finite sequences: i.e. sequences in \mathbb{R}^N . So we have

$$\sum_{n=1}^N a_n b_n \leq \left(\sum_{n=1}^N a_n^p \right)^{\frac{1}{p}} \left(\sum_{n=1}^N b_n^q \right)^{\frac{1}{q}}$$

But $b_n^q = |x_n + y_n|^{q(p-1)} = |x_n + y_n|^p$ using the conjugate exponents identities we established. So we have found

$$\sum_{n=1}^N |x_n| |x_n + y_n|^{p-1} \leq \left(\sum_{n=1}^N |x_n|^p \right)^{\frac{1}{p}} \left(\sum_{n=1}^N |x_n + y_n|^p \right)^{\frac{1}{q}}$$

We can apply the same reasoning to the terms c_n and d_n to find

$$\sum_{n=1}^N |y_n| |x_n + y_n|^{p-1} \leq \left(\sum_{n=1}^N |y_n|^p \right)^{\frac{1}{p}} \left(\sum_{n=1}^N |x_n + y_n|^p \right)^{\frac{1}{q}}$$

We can use the inequalities we just figured out to get the next estimate

$$\begin{aligned} \sum_{n=1}^N |x_n + y_n|^p &\leq \left(\sum_{n=1}^N |x_n|^p \right)^{\frac{1}{p}} \left(\sum_{n=1}^N |x_n + y_n|^p \right)^{\frac{1}{q}} \\ &\quad + \left(\sum_{n=1}^N |y_n|^p \right)^{\frac{1}{p}} \left(\sum_{n=1}^N |x_n + y_n|^p \right)^{\frac{1}{q}} \end{aligned}$$

4.3. SEQUENCE SPACES REVISITED

61

Now factor out the common term to get

$$\sum_{n=1}^N |x_n + y_n|^p \leq \left(\left(\sum_{n=1}^N |x_n|^p \right)^{\frac{1}{p}} + \left(\sum_{n=1}^N |y_n|^p \right)^{\frac{1}{p}} \right) \left(\sum_{n=1}^N |x_n + y_n|^p \right)^{\frac{1}{q}}$$

Rewrite again as

$$\left(\sum_{n=1}^N |x_n + y_n|^p \right)^{1-\frac{1}{q}} \leq \left(\sum_{n=1}^N |x_n|^p \right)^{\frac{1}{p}} + \left(\sum_{n=1}^N |y_n|^p \right)^{\frac{1}{p}}$$

But $1 - 1/q = 1/p$, so we have $\left(\sum_{n=1}^N |x_n + y_n|^p \right)^{\frac{1}{p}} \leq \left(\sum_{n=1}^N |x_n|^p \right)^{\frac{1}{p}} + \left(\sum_{n=1}^N |y_n|^p \right)^{\frac{1}{p}}$.

Now apply the final estimate to find $\left(\sum_{n=1}^N |x_n + y_n|^p \right)^{\frac{1}{p}} \leq \left(\sum_{n=1}^{\infty} |x_n|^p \right)^{\frac{1}{p}} + \left(\sum_{n=1}^{\infty} |y_n|^p \right)^{\frac{1}{p}}$. This says the right hand side is an upper bound for the partial sums on the left side. Hence, we know

$$\left(\sum_{n=1}^{\infty} |x_n + y_n|^p \right)^{\frac{1}{p}} \leq \left(\sum_{n=1}^{\infty} |x_n|^p \right)^{\frac{1}{p}} + \left(\sum_{n=1}^{\infty} |y_n|^p \right)^{\frac{1}{p}}$$

■

Homework

Exercise 4.3.1

Exercise 4.3.2

Exercise 4.3.3

Exercise 4.3.4

Exercise 4.3.5

When we specialize to \mathbb{R}^n we obtain

Theorem 4.3.6 Minkowski’s Inequality in \mathbb{R}^n

Let $p \geq 1$ and let \mathbf{x} and \mathbf{y} be in \mathbb{R}^n with associated sequences (x_j) and (y_j) . Then,

$$\left(\sum_{j=1}^n |x_j + y_j|^p \right)^{\frac{1}{p}} \leq \left(\sum_{j=1}^n |x_j|^p \right)^{\frac{1}{p}} + \left(\sum_{j=1}^n |y_j|^p \right)^{\frac{1}{p}}$$

and for the case $p = \infty$, we have

$$\sup_{1 \leq j \leq n} |x_j + y_j| \leq \sup_{1 \leq j \leq n} |x_j| + \sup_{1 \leq j \leq n} |y_j|$$

In \mathbb{R}^n , define

$$1. \|\mathbf{x}\|_1 = \sum_{j=1}^n |x_j|$$

$$2. \|\mathbf{x}\|_2 = \sqrt{\sum_{j=1}^n |x_j|^2}$$

$$3. \|\mathbf{x}\|_\infty = \max_{1 \leq j \leq n} |x_j|$$

and in general $\|\mathbf{x}\|_p = (\sum_{j=1}^n |x_j|^p)^{1/p}$ for any $p \geq 1$. Minkowski's Inequality tells in all cases

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

This shows a number of things. Let \mathbf{V} be the set of all sequences (x_j) identified with vectors from \mathbb{R}^n . Then we can add them and scalar multiply them, so this is a vector space over \mathbb{R} . If we look at $\|\mathbf{x}\|_p$, it clearly is a mapping from \mathbf{V} to the reals which satisfies properties N1 to N3 for a norm on \mathbf{V} . Minkowski's Inequality then proves property N4 for a norm. Hence, each $\|\cdot\|_p$ is a norm on \mathbb{R}^n and can speak of the distinct normed linear spaces $(\mathbb{R}^n, \|\cdot\|_1)$, $(\mathbb{R}^n, \|\cdot\|_2)$, $(\mathbb{R}^n, \|\cdot\|_\infty)$ and in general $(\mathbb{R}^n, \|\cdot\|_p)$.

Convergence with respect to these norms is then defined in the usual way. We say \mathbf{x}_n converges in $\|\cdot\|_p$ norm to \mathbf{x} if

$$\forall \epsilon > 0, \exists N \ni n > N \implies \|\mathbf{x}_n - \mathbf{x}\|_p < \epsilon$$

All of these normed linear spaces are complete. We proved this in the more general ℓ^p setting in (Peterson (8) 2019), but let's specialize these proofs to \mathbb{R}^n here. They are a bit simpler since we do not have to worry about the convergence of series.

Homework

Exercise 4.3.6

Exercise 4.3.7

Exercise 4.3.8

Exercise 4.3.9

Exercise 4.3.10

Theorem 4.3.7 The Completeness of $(\mathbb{R}^n, \|\cdot\|_\infty)$

$(\mathbb{R}^n, \|\cdot\|_\infty)$ is complete; i.e. if (\mathbf{x}_k) is a Cauchy Sequence in this normed linear space, there is an \mathbf{x} in $(\mathbb{R}^n, \|\cdot\|_\infty)$ so that \mathbf{x}_k converges in $\|\cdot\|_\infty$ norm to \mathbf{x} .

Proof 4.3.5

Assume (\mathbf{x}_k) is a Cauchy Sequence. The element \mathbf{x}_k is a sequence itself which we denote by $(x_{k,j})$ for $1 \leq j \leq n$. Then for a given $\epsilon > 0$, there is an N so that

$$\max_{1 \leq j \leq n} |x_{k,j} - x_{m,j}| < \epsilon/2 \text{ when } m > k > N_\epsilon$$

So for each fixed index j , we have

$$|x_{k,j} - x_{m,j}| < \epsilon/2 \text{ when } m > k > N_\epsilon$$

This says for fixed j , the sequence $(x_{k,j})$ is a Cauchy sequence of real numbers and hence must converge to a real number we will call a_j . This defines a new vector $\mathbf{a} \in \mathbb{R}^n$. Does $\mathbf{x}_n \rightarrow \mathbf{a}$ in the $\|\cdot\|_\infty$ norm? We use the continuity of the function $|\cdot|$ to see for any $k > N_\epsilon$, we have

$$\lim_{m \rightarrow \infty} |x_{k,j} - x_{m,j}| \leq \epsilon/2 \implies |x_{k,j} - \lim_{m \rightarrow \infty} x_{m,j}| \leq \epsilon/2$$

This argument works for all j and so $|x_{k,j} - a_j| \leq \epsilon/2$ when $k > N_\epsilon$ for all j which implies $\max_{1 \leq j \leq n} |x_{k,j} - a_j| \leq \epsilon/2 < \epsilon$ or $\|\mathbf{x}_k - \mathbf{a}\|_\infty < \epsilon$ when $k > N_\epsilon$. So $\mathbf{x}_n \rightarrow \mathbf{a}$ in $\|\cdot\|_\infty$. ■

Next, let's look at vector convergence using the $\|\cdot\|_p$ norm for $p \geq 1$.

Theorem 4.3.8 $(\mathbb{R}^n, \|\cdot\|_p)$ is Complete

$(\mathbb{R}^n, \|\cdot\|_p)$ is complete; i.e. if (x_k) is a Cauchy Sequence in this normed linear space, there is an x in $(\mathbb{R}^n, \|\cdot\|_p)$ so that x_k converges in $\|\cdot\|_p$ norm to x .

Proof 4.3.6

Let x_k be a Cauchy Sequence in $\|\cdot\|_p$. Then given $\epsilon > 0$, there is an N_ϵ so that $m > k > N_\epsilon$ implies

$$\|x_k - x_m\|_p = \left(\sum_{j=1}^n |x_{k,j} - x_{m,j}|^p \right)^{\frac{1}{p}} < \epsilon/2.$$

Thus, if $m > k > N_\epsilon$, $\sum_{j=1}^n |x_{k,j} - x_{m,j}|^p < (\epsilon/2)^p$. Since this is a sum on non-negative terms, each term must be less than $(\epsilon/2)^p$. So we must have $|x_{k,j} - x_{m,j}|^p < (\epsilon/2)^p$ or $|x_{k,j} - x_{m,j}| < (\epsilon/2)$ when $m > k > N_\epsilon$. This tells us immediately the sequence of real numbers $(x_{k,j})$ is a Cauchy sequence of real numbers and so must converge to a number we will call a_j . This defines the vector $a \in \mathbb{R}^n$. Since $|\cdot|^p$ is continuous, we can say

$$\begin{aligned} (\epsilon/2)^p &\geq \lim_{m \rightarrow \infty} \left(\sum_{j=1}^n |x_{k,j} - x_{m,j}|^p \right) = \left(\sum_{j=1}^n |x_{k,j} - \lim_{m \rightarrow \infty} x_{m,j}|^p \right) \\ &= \sum_{j=1}^n |x_{k,j} - a_j|^p \end{aligned}$$

This says immediately that $\|x_k - a\|_p < \epsilon$ when $k > N_\epsilon$ so $x_k \rightarrow a$ in $\|\cdot\|_p$. ■

Comment 4.3.1 It is easy to see \mathbb{R}^n has an inner product given by

$$\langle x, y \rangle = \sum_{j=1}^n x_j y_j$$

Hence, $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ is an inner product space whose norm induces $\|\cdot\|_2$. Moreover Hölder's Inequality tells us

$$|\langle x, y \rangle| \leq \sum_{j=1}^n |x_j y_j| \leq \|x\|_2 \|y\|_2$$

which is our standard Cauchy - Schwartz Inequality in \mathbb{R}^n which we use to define the angle between vectors in \mathbb{R}^n as usual.

4.3.1 Homework

Exercise 4.3.11

Exercise 4.3.12

Exercise 4.3.13

Exercise 4.3.14

Exercise 4.3.15

4.4 Linear Transformations Between Normed Linear Spaces

Let's start by defining linear transformations between vector spaces carefully

Definition 4.4.1 Linear Transformations Between Vector Spaces

Let X and Y be two vector spaces over \mathbb{R} . We say $T : X \rightarrow Y$ is a linear transformation if

$$T(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha T(\mathbf{x}) + \beta T(\mathbf{y})$$

To discuss linear transformations between finite dimensional vector spaces, we need to have a formal definition of matrices. We have used them quite a bit in our explanations, but it is time to be formal. Note, we have already thought carefully about symmetric matrices in Chapter 5 but our interests are much more general here.

Definition 4.4.2 Real Matrices

Let M be a collection of real numbers organized in a table of m rows and n columns. The number M_{ij} refers to the entry in the i^{th} row and j^{th} column of this table. The table is organized like this

$$\begin{matrix} M_{11} & \dots & M_{1n} \\ M_{21} & \dots & M_{2n} \\ \vdots & \vdots & \vdots \\ M_{m1} & \dots & M_{mn} \end{matrix}$$

We identify the matrix M with this table and write

$$M = \begin{bmatrix} M_{11} & \dots & M_{1n} \\ \vdots & \vdots & \vdots \\ M_{m1} & \dots & M_{mn} \end{bmatrix}$$

The set of all matrices with m rows and n columns is denoted by M_{mn} . We also say M is a $m \times n$ matrix.

Comment 4.4.1 The n columns of M are clearly vectors in \mathbb{R}^m and the m rows of M are vectors in \mathbb{R}^n . Thus, we have some additional identifications:

$$M = \begin{bmatrix} M_{11} & \dots & M_{1n} \\ \vdots & \vdots & \vdots \\ M_{m1} & \dots & M_{mn} \end{bmatrix} = [C_1 \ \dots \ C_n] = \begin{bmatrix} R_1 \\ \vdots \\ R_m \end{bmatrix}$$

where column i is denoted by $C_j \in \mathbb{R}^m$ and row j is denoted by $R_j \in \mathbb{R}^n$. Note the rows are vectors displayed as the transpose of the usual column vector notation we use.

Comment 4.4.2 So for example the data pairs of time and temperature we measure in a cooling model experiment might turn out to be

Listing 4.1: Sample Data

1	201
3	198
6	190
9	185
12	182
15	179
20	168
25	161
30	154
40	146
50	135
60	123
80	109
100	87
120	82
140	79
150	77

This table defines a matrix of 18 rows and 2 columns and so defines a matrix in $M_{18,2}$.

4.4.0.1 Homework

Exercise 4.4.1

Exercise 4.4.2

Exercise 4.4.3

Exercise 4.4.4

Exercise 4.4.5

4.4.1 Basic Properties

Consider a matrix M in M_{mn} . We define the action of M on the vector space \mathbb{R}^n by

$$M(x) = ([R_1 \dots R_m])(x) = \begin{bmatrix} < R_1, x > \\ \vdots \\ < R_m, x > \end{bmatrix}$$

We generally just write $Mx = M(x)$ to save a few parenthesis. It is easy to see this definition of the action of M on \mathbb{R}^n defines a linear transformation from \mathbb{R}^n to \mathbb{R}^m .

The set of x in \mathbb{R}^n with $Mx = \mathbf{0} \in \mathbb{R}^m$ is called the **kernel** or **nullspace** of M . The span of the columns of M is called the column space of M . We can prove a fundamental result.

Theorem 4.4.1 If $M \in M_{mn}$, then $\dim(\text{kernel}) + \dim(\text{column space}) = n$

If M is a $m \times n$ matrix, if $p = \dim(\ker(M))$ and q is the dimension of the span of the columns of M , then $p + q = n$.

Proof 4.4.1

It is easy to see that $\ker(M)$ is a subspace of \mathbb{R}^n with dimension $p \leq n$. Let $\{K_1, \dots, K_p\}$ be an orthonormal basis of $\ker(M)$. We can use GSO to construct an additional $n - p$ vectors in \mathbb{R}^n , $\{L_1, \dots, L_{n-p}\}$ that are all orthogonal to $\ker(M)$ with length one. The subspace spanned by

$\{L_1, \dots, L_{n-p}\}$ is perpendicular to the subspace $\ker(M)$ and is called $(\ker(m))^\perp$ to denote it is what is called an orthogonal complement. We note $\mathfrak{R}^n = \ker(M) \oplus \text{span}\{L_1, \dots, L_{n-p}\}$.

Thus, $\{K_1, \dots, K_p, L_1, \dots, L_{n-p}\}$ is an orthonormal basis for \mathfrak{R}^n and hence $x = a_1 K_1 + \dots + a_p K_p + b_1 L_1 + \dots + b_{n-p} L_{n-p}$. It follows using the linearity of M that

$$Mx = b_1 M L_1 + \dots + b_{n-p} M L_{n-p}$$

Now if $b_1 M L_1 + \dots + b_{n-p} M L_{n-p} = 0$, this means $M(b_1 L_1 + \dots + b_{n-p} L_{n-p}) = 0$ too. This says $b_1 L_1 + \dots + b_{n-p} L_{n-p}$ is in both $\ker(M)$ and the $\text{span}\{L_1, \dots, L_{n-p}\}$ which forces $b_1 L_1 + \dots + b_{n-p} L_{n-p} = 0$. But $\{L_1, \dots, L_{n-p}\}$ is a linearly independent set and so all coefficients $b_i = 0$. This says the set $\{M L_1, \dots, M L_{n-p}\}$ is a linearly independent set in \mathfrak{R}^m .

Finally, note the column space of M is defined to be the $\text{span}\{C_1, \dots, C_n\}$. A vector is this span has the look $y = \sum_{i=1}^m a_i C_i$. This is the same as Ma where $a \in \mathfrak{R}^n$. However, we know \mathfrak{R}^n has the orthonormal basis $\{K_1, \dots, K_p, L_1, \dots, L_{n-p}\}$ and so $a = \sum_{i=1}^p b_i K_i + \sum_{i=1}^{n-p} c_i L_i$. Applying M we find

$$Ma = c_1 M L_1 + \dots + c_{n-p} M L_{n-p}$$

Since $\{M L_1, \dots, M L_{n-p}\}$ is a linearly independent, we see the range of M , the column space of M has dimension $n - p$. Thus, we have shown

$$\dim(\ker(M)) + \dim(\text{span}\{C_1, \dots, C_n\}) = n$$

where, of course, we also have $n - p \leq m$. ■

Comment 4.4.3 Thus, if $n - p = m$, the kernel of M must be $n - m$ dimensional.

4.4.1.1 Homework

Exercise 4.4.6

Exercise 4.4.7

Exercise 4.4.8

Exercise 4.4.9

Exercise 4.4.10

4.4.2 Mappings between Finite Dimensional Vector Spaces

Let X and Y be two finite dimensional vector spaces over \mathfrak{R} . Assume X has dimension n and Y has dimension m . Let E be a basis for X and F , a basis for Y . Before we go further, let's be clear about some of our notation.

- We use x to denote an object in the finite dimensional vector space X .
- For a given basis E , we can write $x = x_1 E_1 + \dots + E_n$ and the coefficients $\{x_1, \dots, x_n\}$ are n real numbers and so give a vector in \mathfrak{R}^n . We let $[x]_E$ indicate this vector of real numbers. If G was another basis, there would be another set of n numbers given by $x = u_1 G_1 + \dots + u_n G_n$ and this vector would be denoted by $[x]_G$. Then, there will be a transformation law which converts $[x]_E$ into $[x]_G$ and we will talk about that soon. The point is x is the same object in X which we can represent in different ways with a mechanism to map one representation into another.

Theorem 4.4.2 The Change of Basis Mapping

Let \mathbf{E} and \mathbf{F} be basis for the finite dimensional vector space \mathbf{X} which has dimension n . Then there is an invertible matrix $\mathbf{A}_{\mathbf{EF}}$ so that $[\mathbf{x}]_{\mathbf{G}} = \mathbf{A}_{\mathbf{EG}} [\mathbf{x}]_{\mathbf{E}}$. We read the matrix $\mathbf{A}_{\mathbf{EG}}$ as mapping \mathbf{E} into \mathbf{G} .

Proof 4.4.2

We know

$$\mathbf{x} = c_1 \mathbf{E}_1 + \dots + c_n \mathbf{E}_n$$

We also know each \mathbf{E}_j has an expansion in the \mathbf{G} basis; i.e. $\mathbf{E}_i = \sum_{j=1}^n A_{ji} \mathbf{G}_j$. Using this, we find

$$\mathbf{x} = \sum_{i=1}^n c_i \left(\sum_{j=1}^n A_{ji} \mathbf{G}_j \right) = \sum_{j=1}^n \left(\sum_{i=1}^n c_i A_{ji} \right) \mathbf{G}_j$$

But \mathbf{x} also has an expansion with respect to \mathbf{G} : $\mathbf{x} = \sum_{j=1}^n d_j \mathbf{G}_j$. Hence, equating coefficients, we have

$$d_j = \sum_{i=1}^n A_{ji} c_i$$

which can be organized into a familiar set of equations.

$$\begin{aligned} d_1 &= A_{11}c_1 + A_{12}c_2 + \dots + A_{1n}c_n \\ &\vdots \\ d_j &= A_{j1}c_1 + A_{j2}c_2 + \dots + A_{jn}c_n \\ &\vdots \\ d_n &= A_{n1}c_1 + A_{n2}c_2 + \dots + A_{nn}c_n \end{aligned}$$

or

$$\begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} = \begin{bmatrix} A_{11} & \dots & A_{1n} \\ \vdots & & \vdots \\ A_{n1} & \dots & A_{nn} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$$

We then further rewrite as $[\mathbf{x}]_{\mathbf{G}} = \mathbf{A}_{\mathbf{EG}} [\mathbf{x}]_{\mathbf{E}}$. A similar argument shows there is a matrix $\mathbf{B}_{\mathbf{GE}}$ so that $[\mathbf{x}]_{\mathbf{E}} = \mathbf{B}_{\mathbf{GE}} [\mathbf{x}]_{\mathbf{G}}$. Thus

$$\begin{aligned} [\mathbf{x}]_{\mathbf{G}} &= \mathbf{A}_{\mathbf{EG}} [\mathbf{x}]_{\mathbf{E}} = \mathbf{A}_{\mathbf{EG}} \mathbf{B}_{\mathbf{GE}} [\mathbf{x}]_{\mathbf{G}} \\ [\mathbf{x}]_{\mathbf{E}} &= \mathbf{B}_{\mathbf{GE}} [\mathbf{x}]_{\mathbf{G}} = \mathbf{B}_{\mathbf{GE}} \mathbf{A}_{\mathbf{EG}} [\mathbf{x}]_{\mathbf{E}} \end{aligned}$$

This tells us $\mathbf{A}_{\mathbf{EG}} \mathbf{B}_{\mathbf{GE}} = \mathbf{B}_{\mathbf{GE}} \mathbf{A}_{\mathbf{EG}} = \mathbf{I}$. Hence, $\mathbf{A}_{\mathbf{EG}}^{-1} = \mathbf{B}_{\mathbf{GE}}$. ■

Theorem 4.4.3 The Change of Basis Mapping In a Finite Dimensional Inner Product Space

Let \mathbf{E} and \mathbf{F} be basis for the finite dimensional vector space \mathbf{X} which has dimension n . We assume \mathbf{X} has an inner product $\langle \cdot, \cdot \rangle$. Then there is an invertible matrix $\mathbf{A}_{\mathbf{EF}}$ so that $[\mathbf{x}]_{\mathbf{G}} = \mathbf{A}_{\mathbf{EG}} [\mathbf{x}]_{\mathbf{E}}$ which has the following form:

$$\mathbf{A}_{\mathbf{EG}} = \begin{bmatrix} \mathbf{G}_1^T \\ \vdots \\ \mathbf{G}_n^T \end{bmatrix} [\mathbf{E}_1, \dots, \mathbf{E}_n]$$

with

$$\mathbf{A}_{\mathbf{EG}}^{-1} = \mathbf{A}_{\mathbf{EG}}^T = \begin{bmatrix} \mathbf{E}_1^T \\ \vdots \\ \mathbf{E}_n^T \end{bmatrix} [\mathbf{G}_1, \dots, \mathbf{G}_n]$$

Proof 4.4.3

Let \mathbf{G} be another orthonormal basis for \mathbf{X} . Then since $\mathbf{x} = c_1 \mathbf{E}_1 + \dots + c_n \mathbf{E}_n$, we have

$$\begin{aligned} \langle \mathbf{x}, \mathbf{G}_1 \rangle &= c_1 \langle \mathbf{G}_1, \mathbf{E}_1 \rangle + \dots + c_n \langle \mathbf{G}_1, \mathbf{E}_n \rangle \\ &\vdots = \\ \langle \mathbf{x}, \mathbf{G}_n \rangle &= c_1 \langle \mathbf{G}_n, \mathbf{E}_1 \rangle + \dots + c_n \langle \mathbf{G}_n, \mathbf{E}_n \rangle \end{aligned}$$

which in matrix - vector form is

$$[\mathbf{x}]_{\mathbf{G}} = \begin{bmatrix} \mathbf{G}_1^T \\ \vdots \\ \mathbf{G}_n^T \end{bmatrix} [\mathbf{E}_1, \dots, \mathbf{E}_n] [\mathbf{x}]_{\mathbf{E}}$$

The orthonormality of \mathbf{E} and \mathbf{G} then tell us immediately

$$\mathbf{A}_{\mathbf{EG}}^{-1} = \mathbf{A}_{\mathbf{EG}}^T = \begin{bmatrix} \mathbf{E}_1^T \\ \vdots \\ \mathbf{E}_n^T \end{bmatrix} [\mathbf{G}_1, \dots, \mathbf{G}_n]$$

■

Since there is an inner product on \mathbf{X} , we know its value does not depend on the choice of orthonormal basis for \mathbf{X} .

Homework

Exercise 4.4.11

Exercise 4.4.12

Exercise 4.4.13

Exercise 4.4.14

Exercise 4.4.15

Theorem 4.4.4 The Inner Product on a Finite Dimensional Vector Space is Independent of Choice of Orthonormal Basis

Let \mathbf{X} be an n dimensional vector space with inner product $\langle \cdot, \cdot \rangle$. Then its value is independent of the choice of orthonormal basis \mathbf{E} .

Proof 4.4.4

We know $\mathbf{x}_E = x_E^1 \mathbf{E}_1 + \dots + x_E^n \mathbf{E}_n$ and $\mathbf{x}_G = x_G^1 \mathbf{G}_1 + \dots + x_G^n \mathbf{G}_n$ with a similar notation for the representations of \mathbf{y} . We have

$$[\mathbf{x}]_E = \begin{bmatrix} x_E^1 \\ \vdots \\ x_E^n \end{bmatrix}, \quad [\mathbf{x}]_G = \begin{bmatrix} x_G^1 \\ \vdots \\ x_G^n \end{bmatrix}$$

and we know

$$[\mathbf{x}]_G = \begin{bmatrix} \mathbf{G}_1^T \\ \vdots \\ \mathbf{G}_n^T \end{bmatrix} [\mathbf{E}_1, \dots, \mathbf{E}_n] [\mathbf{x}]_E$$

Thus, by orthonormality of \mathbf{E} and \mathbf{G} , we have

$$\begin{aligned} \langle \mathbf{x}_G, \mathbf{y}_G \rangle &= \sum_{i=1}^n x_G^i y_G^i = \langle [\mathbf{x}]_G, [\mathbf{y}]_G \rangle \\ &= \left\langle \begin{bmatrix} \mathbf{G}_1^T \\ \vdots \\ \mathbf{G}_n^T \end{bmatrix} [\mathbf{E}_1, \dots, \mathbf{E}_n] [\mathbf{x}]_E, \begin{bmatrix} \mathbf{G}_1^T \\ \vdots \\ \mathbf{G}_n^T \end{bmatrix} [\mathbf{E}_1, \dots, \mathbf{E}_n] [\mathbf{y}]_E \right\rangle \\ &= [\mathbf{x}]_E^T [\mathbf{E}_1, \dots, \mathbf{E}_n]^T [\mathbf{G}_1, \dots, \mathbf{G}_n] \begin{bmatrix} \mathbf{G}_1^T \\ \vdots \\ \mathbf{G}_n^T \end{bmatrix} [\mathbf{E}_1, \dots, \mathbf{E}_n] [\mathbf{y}]_E \\ &= \langle [\mathbf{x}]_E, [\mathbf{y}]_E \rangle = \sum_{i=1}^n x_E^i y_E^i \end{aligned}$$

which shows the value is independent of the choice of orthonormal basis. ■

Homework

Exercise 4.4.16

Exercise 4.4.17

Exercise 4.4.18

Exercise 4.4.19

Exercise 4.4.20

We define linear transformations the same really.

Theorem 4.4.5 Linear Transformations Between Finite Dimensional Vector Spaces

Let \mathbf{X} and \mathbf{Y} be two finite dimensional vector spaces over \mathbb{R} . Assume \mathbf{X} has dimension n and \mathbf{Y} has dimension m . Let \mathbf{T} be a linear transformation between the spaces. Given a basis \mathbf{E} for \mathbf{X} and a basis \mathbf{F} for \mathbf{Y} , we can identify \mathbb{R}^n with the span of \mathbf{E} and \mathbb{R}^m with the span of \mathbf{F} . Then there is an $n \times m$ matrix $\mathbf{T}_{\mathbf{EF}} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ so that

$$[\mathbf{T}\mathbf{x}]_{\mathbf{F}} = \begin{bmatrix} T_{11} & \dots & T_{1n} \\ \vdots & \vdots & \vdots \\ T_{m1} & \dots & T_{mn} \end{bmatrix}_{\mathbf{EF}} [\mathbf{x}]_{\mathbf{E}}$$

Note $\mathbf{T}\mathbf{x}$ is the same object in \mathbf{Y} no matter its representation via a choice of basis \mathbf{F} for \mathbf{Y} .

Proof 4.4.5

Given \mathbf{x} in \mathbf{X} , we can write

$$\mathbf{x} = x_1 \mathbf{E}_1 + \dots + x_n \mathbf{E}_n$$

By the linearity of \mathbf{T} , we then have

$$\mathbf{T}\mathbf{x} = x_1 \mathbf{T}\mathbf{E}_1 + \dots + x_n \mathbf{T}\mathbf{E}_n$$

But each $\mathbf{T}\mathbf{E}_i$ is in \mathbf{F} , so we have the expansions

$$\begin{aligned} \mathbf{T}\mathbf{E}_1 &= \beta_{11} \mathbf{F}_1 + \dots + \beta_{m1} \mathbf{F}_m \\ \mathbf{T}\mathbf{E}_2 &= \beta_{12} \mathbf{F}_1 + \dots + \beta_{m2} \mathbf{F}_m \\ &\vdots = \vdots \\ \mathbf{T}\mathbf{E}_n &= \beta_{1n} \mathbf{F}_1 + \dots + \beta_{mn} \mathbf{F}_m \end{aligned}$$

which implies since $\mathbf{T}\mathbf{x}$ is also in \mathbf{Y} that

$$\mathbf{T}\mathbf{x} = x_1 \left(\sum_{j=1}^m \beta_{j1} \mathbf{F}_j \right) + \dots + x_n \left(\sum_{j=1}^m \beta_{jn} \mathbf{F}_j \right) = \sum_{j=1}^m \gamma_j \mathbf{F}_j$$

Now reorganize these sums to obtain

$$\sum_{j=1}^m \left(\sum_{i=1}^n \beta_{ji} x_i \right) \mathbf{F}_j = \sum_{j=1}^m \gamma_j \mathbf{F}_j$$

But expansions with respect to the \mathbf{F} basis are unique, so we have

$$\begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_m \end{bmatrix}_{\mathbf{F}} = \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \dots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \beta_{23} & \dots & \beta_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \beta_{m1} & \beta_{m2} & \beta_{m3} & \dots & \beta_{mn} \end{bmatrix}_{\mathbf{EF}} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_{\mathbf{E}}$$

Let $T_{ij} = \beta_{ij}$ and we have found the matrix \mathbf{T} so that

$$[\mathbf{T}\mathbf{x}]_{\mathbf{F}} = \mathbf{T}_{\mathbf{EF}} [\mathbf{x}]_{\mathbf{E}}$$

■

Theorem 4.4.6 A Linear Transformation between Two Finite Dimensional Vector Spaces Determines an Equivalence Class of Matrices

Let \mathbf{X} and \mathbf{Y} be two finite dimensional vector spaces over \mathbb{R} . Assume \mathbf{X} has dimension n and \mathbf{Y} has dimension m . Let \mathbf{T} be a linear transformation between the spaces. Then \mathbf{T} is associated with an equivalence class in the set of $m \times n$ matrices under the equivalence relation \sim where $\mathbf{A} \sim \mathbf{B}$ if and only if there are invertible matrices \mathbf{P} and \mathbf{Q} so that $\mathbf{A} = \mathbf{P}^{-1}\mathbf{BQ}$.

Proof 4.4.6

First, it is easy to see \sim defines an equivalence relation on the set of all $m \times n$ matrices. We leave that to you. Any such linear transformation \mathbf{T} is associated with an $m \times n$ matrix once the bases \mathbf{E} and \mathbf{F} are chosen. What happens if we change to a new basis \mathbf{G} for \mathbf{X} and a new basis \mathbf{H} for \mathbf{Y} ? We know from earlier calculations that there are invertible matrices \mathbf{A}_{GE} and \mathbf{B}_{HF} so that

$$[\mathbf{x}]_{\mathbf{E}} = \mathbf{A}_{GE} [\mathbf{x}]_{\mathbf{G}}, \quad [\mathbf{y}]_{\mathbf{F}} = \mathbf{B}_{HF} [\mathbf{y}]_{\mathbf{H}}$$

Using all this, we have

$$[\mathbf{T}\mathbf{x}]_{\mathbf{F}} = \mathbf{T}_{EF} [\mathbf{x}]_{\mathbf{E}} \implies \mathbf{B}_{HF} [\mathbf{T}\mathbf{x}]_{\mathbf{H}} = \mathbf{T}_{EF} \mathbf{A}_{GE} [\mathbf{x}]_{\mathbf{G}}$$

Since, \mathbf{B}_{HF} is invertible, this tells us

$$[\mathbf{T}\mathbf{x}]_{\mathbf{H}} = \mathbf{B}_{HF}^{-1} \mathbf{T}_{EF} \mathbf{A}_{GE} [\mathbf{x}]_{\mathbf{G}}$$

We also know $[\mathbf{T}\mathbf{x}]_{\mathbf{H}} = \mathbf{T}_{GH} [\mathbf{x}]_{\mathbf{G}}$ and thus we must have $\mathbf{T}_{GH} = \mathbf{B}_{HF}^{-1} \mathbf{T}_{EF} \mathbf{A}_{GE}$. This tells us $\mathbf{T}_{GH} \sim \mathbf{T}_{EF}$. ■

Comment 4.4.4 Thus any linear transformation between two finite dimensional vector spaces is identified with an equivalence class of $m \times n$ matrices. When we pick basis \mathbf{E} and \mathbf{F} for \mathbf{X} and \mathbf{Y} respectively, we choose a particular representative from this equivalence class we denote by \mathbf{T}_{EF} . Note we can say more about the structure of \mathbf{T}_{EF} if these bases are orthonormal.

Homework

Exercise 4.4.21

Exercise 4.4.22

Exercise 4.4.23

Exercise 4.4.24

Exercise 4.4.25

4.5 Magnitudes of Linear Transformations

For the linear transformation \mathbf{T} mapping the n dimensional vector space \mathbf{X} to the m dimensional vector space \mathbf{Y} , we know given a basis \mathbf{E} for \mathbf{X} and a basis \mathbf{F} for \mathbf{Y} $[\mathbf{T}\mathbf{x}]_{\mathbf{F}} = [\mathbf{T}]_{EF} [\mathbf{x}]_{\mathbf{E}}$. Let's add a norm to the vector spaces \mathbb{R}^n and \mathbb{R}^m . For now, let's assume we now use the normed linear spaces $(\mathbb{R}^m, \|\cdot\|_2)$ and $(\mathbb{R}^n, \|\cdot\|_2)$. Of course, we could use a different norm on the domain space and the range space, but we won't do that as it adds unnecessary confusion. Let the entries of $[\mathbf{T}]_{EF}$ be called T_{ij} where we will not label this entries with a EF in order to avoid too much clutter. But

of course, they **do** depend on \mathbf{E} and \mathbf{F} ; just keep that in the back of your mind. Note

$$\|[\mathbf{T}\mathbf{x}]_{\mathbf{F}}\|_2^2 = \sum_{j=1}^m ([\mathbf{T}\mathbf{x}]_{\mathbf{F}})_j^2 = \sum_{i=1}^m \left(\sum_{j=1}^n T_{ij}x_j \right)$$

Apply the Cauchy - Schwartz Inequality here:

$$\left(\sum_{j=1}^n T_{ij}x_j \right)^2 \leq \left(\sum_{j=1}^n T_{ij}^2 \right) \left(\sum_{j=1}^n x_j^2 \right) = \left(\sum_{j=1}^n T_{ij}^2 \right) \|\mathbf{x}_{\mathbf{E}}\|_2^2$$

Thus,

$$\|[\mathbf{T}\mathbf{x}]_{\mathbf{F}}\|_2^2 \leq \left(\sum_{i=1}^m \sum_{j=1}^n T_{ij}^2 \right) \|\mathbf{x}_{\mathbf{E}}\|_2^2$$

which implies

$$\|[\mathbf{T}\mathbf{x}]_{\mathbf{F}}\|_2 \leq \sqrt{\sum_{i=1}^m \sum_{j=1}^n T_{ij}^2} \|\mathbf{x}_{\mathbf{E}}\|_2$$

This leads to the following definition:

Definition 4.5.1 The Frobenius Norm of a Linear Transformation

For the linear transformation \mathbf{T} mapping the n dimensional vector space \mathbf{X} to the m dimensional vector space \mathbf{Y} , we know given a basis \mathbf{E} for \mathbf{X} and a basis \mathbf{F} for \mathbf{Y} $[\mathbf{T}\mathbf{x}]_{\mathbf{F}} = [\mathbf{T}]_{\mathbf{EF}} [\mathbf{x}]_{\mathbf{E}}$. The **Frobenius Norm** of \mathbf{T} is denoted $\|[\mathbf{T}]_{\mathbf{EF}}\|_{Fr}$ and is defined by

$$\|[\mathbf{T}]_{\mathbf{EF}}\|_{Fr} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n T_{ij}^2}$$

where T_{ij} are the entries in the matrix $[\mathbf{T}]_{\mathbf{EF}}$.

We have just proven the following result:

Theorem 4.5.1 The Frobenius Norm Fundamental Inequality

For the linear transformation \mathbf{T} mapping the n dimensional vector space \mathbf{X} to the m dimensional vector space \mathbf{Y} , we know given a basis \mathbf{E} for \mathbf{X} and a basis \mathbf{F} for \mathbf{Y} $[\mathbf{T}\mathbf{x}]_{\mathbf{F}} = [\mathbf{T}]_{\mathbf{EF}} [\mathbf{x}]_{\mathbf{E}}$. The Frobenius Norm of this matrix satisfies

$$\|[\mathbf{T}\mathbf{x}]_{\mathbf{F}}\|_2 \leq \|[\mathbf{T}]_{\mathbf{EF}}\|_{Fr} \|\mathbf{x}_{\mathbf{E}}\|_2$$

Proof 4.5.1

We have just gone over this argument. ■

4.5.0.1 Homework

Exercise 4.5.1

Exercise 4.5.2

Exercise 4.5.3

Exercise 4.5.4

Exercise 4.5.5

Chapter 5

Symmetric Matrices

Let's specialize to symmetric matrices now as they are of great interest to us in many applications in applied analysis.

5.1 The General Two by Two Symmetric Matrix

Let's start with a general 2×2 symmetric matrix A given by

$$A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

where a, b and d are arbitrary non zero numbers. The characteristic equation here is $r^2 - (a + d)r + ad - b^2$. Note that the term $ad - b^2$ is the determinant of A . The roots are given by

$$\begin{aligned} r &= \frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad-b^2)}}{2} \\ &= \frac{(a+d) \pm \sqrt{a^2 + 2ad + d^2 - 4ad + 4b^2}}{2} \\ &= \frac{(a+d) \pm \sqrt{a^2 - 2ad + d^2 + 4b^2}}{2} \\ &= \frac{(a+d) \pm \sqrt{(a-d)^2 + 4b^2}}{2} \end{aligned}$$

It is easy to see the term in the square root here is always positive and so we have two real roots. Hence, for a general symmetric 2×2 matrix, the eigenvalues are always real. Note we can find the eigenvectors with a standard calculation. For eigenvalue $\lambda_1 = \frac{(a+d)+\sqrt{(a-d)^2+4b^2}}{2}$, we must find the vectors V so that

$$\begin{bmatrix} \frac{(a+d)+\sqrt{(a-d)^2+4b^2}}{2} - a & -b \\ -b & \frac{(a+d)+\sqrt{(a-d)^2+4b^2}}{2} - d \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We can use the top equation to find the needed relationship between V_1 and V_2 . We have

$$\left(\frac{(a+d) + \sqrt{(a-d)^2 + 4b^2}}{2} - a \right) V_1 - bV_2 = 0.$$

Thus, we have for $V_2 = \frac{d-a+\sqrt{(a-d)^2+4b^2}}{2}$, $V_1 = b$. Thus, the first eigenvector is

$$\mathbf{E}_1 = \begin{bmatrix} b \\ \frac{d-a+\sqrt{(a-d)^2+4b^2}}{2} \end{bmatrix}$$

The second eigenvector is a similar calculation. We must find the vector \mathbf{V} so that

$$\begin{bmatrix} \frac{(a+d)-\sqrt{(a-d)^2+4b^2}}{2} - a & -b \\ -b & \frac{(a+d)-\sqrt{(a-d)^2+4b^2}}{2} - d \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We find

$$\left(\frac{(a+d)-\sqrt{(a-d)^2+4b^2}}{2} - a \right) V_1 - b V_2 = 0.$$

Thus, we have for $V_2 = d + \frac{(a+d)-\sqrt{(a-d)^2+4b^2}}{2}$, $V_1 = b$. Thus, the second eigenvector is

$$\mathbf{E}_2 = \begin{bmatrix} b \\ \frac{d-a-\sqrt{(a-d)^2+4b^2}}{2} \end{bmatrix}$$

Note that $\langle \mathbf{E}_1, \mathbf{E}_2 \rangle$ is

$$\begin{aligned} \langle \mathbf{E}_1, \mathbf{E}_2 \rangle &= b^2 + \left(\frac{d-a+\sqrt{(a-d)^2+4b^2}}{2} \right) \left(\frac{d-a-\sqrt{(a-d)^2+4b^2}}{2} \right) \\ &= b^2 + \frac{(d-a)^2}{4} - \frac{(a-d)^2+4b^2}{4} b^2 + \frac{(d-a)^2-(a-d)^2}{4} - b^2 = 0. \end{aligned}$$

Hence, these two eigenvectors are **orthogonal** to each other. Note, the two eigenvalues are

$$\begin{aligned} \lambda_1 &= \frac{(a+d)+\sqrt{(a-d)^2+4b^2}}{2} \\ \lambda_2 &= \frac{(a+d)-\sqrt{(a-d)^2+4b^2}}{2} \end{aligned}$$

5.1.1 Examples

The only way both eigenvalues can be zero is if both $a+d=0$ and $4(a+d)^2+4b^2=0$. That only happens if $a=b=d=0$ which we explicitly ruled out at the beginning of our discussion because we said a, b and d were nonzero. However, both eigenvalues can be negative, both can be positive or they can be of mixed sign as our examples show.

Example 5.1.1 For

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$$

the eigenvalues are $(9 \pm 5)/2$ and both are positive.

Example 5.1.2 For

$$\mathbf{A} = \begin{bmatrix} 3 & 5 \\ 5 & 6 \end{bmatrix}$$

5.1. THE GENERAL TWO BY TWO SYMMETRIC MATRIX

77

the eigenvalues are $(9 \pm \sqrt{9 + 100})/2$ giving $\lambda_1 = 9.72$ and $\lambda_2 = -5.22$ and the eigenvalues have mixed sign.

Example 5.1.3 For

$$\mathbf{A} = \begin{bmatrix} 3 & 3\sqrt{2} \\ 3\sqrt{2} & 6 \end{bmatrix}$$

the eigenvalues are $(9 \pm 9)/2$ giving $\lambda_1 = 9$ and $\lambda_2 = 0$.

Example 5.1.4 For

$$\mathbf{A} = \begin{bmatrix} 3 & 5 \\ 5 & 6 \end{bmatrix}$$

the eigenvalues are $(9 \pm \sqrt{9 + 100})/2$ giving $\lambda_1 = 9.72$ and $\lambda_2 = -5.22$ and the eigenvalues have mixed sign.

Example 5.1.5 For

$$\mathbf{A} = \begin{bmatrix} -3 & 5 \\ 5 & -6 \end{bmatrix}$$

the eigenvalues are $(-9 \pm \sqrt{34})/2$ giving $\lambda_1 = -7.42$ and $\lambda_2 = -1.58$. So here, both eigenvalues are negative.

However, in all cases the eigenvectors are still orthogonal.

5.1.1.1 Homework

Exercise 5.1.1

Exercise 5.1.2

Exercise 5.1.3

Exercise 5.1.4

Exercise 5.1.5

5.1.2 A Canonical Form

Now let's look at 2×2 symmetric matrices more abstractly. Don't worry, there is a payoff here in understanding! Let \mathbf{A} be a general 2×2 symmetric matrix. Then it has two distinct eigenvalues λ_1 and another one λ_2 . Consider the matrix \mathbf{P} given by

$$\mathbf{P} = [\mathbf{E}_1 \ \mathbf{E}_2]$$

whose transpose is then

$$\mathbf{P}^T = \begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \end{bmatrix}$$

It is easy to find that $\mathbf{P}^T \mathbf{P} = \mathbf{P} \mathbf{P}^T = \mathbf{I}$. Hence, $\mathbf{P}^{-1} = \mathbf{P}^T$. Thus,

$$\mathbf{P}^T \mathbf{A} \mathbf{P} = \begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{E}_1 & \mathbf{E}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{A}\mathbf{E}_1 & \mathbf{A}\mathbf{E}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \end{bmatrix} \begin{bmatrix} \lambda_1 \mathbf{E}_1 & \lambda_2 \mathbf{E}_2 \end{bmatrix}$$

After we do the final multiplications, we have

$$\mathbf{P}^T \mathbf{A} \mathbf{P} = \begin{bmatrix} \lambda_1 \langle \mathbf{E}_1, \mathbf{E}_1 \rangle & \lambda_2 \langle \mathbf{E}_1, \mathbf{E}_2 \rangle \\ \lambda_1 \langle \mathbf{E}_2, \mathbf{E}_1 \rangle & \lambda_2 \langle \mathbf{E}_2, \mathbf{E}_2 \rangle \end{bmatrix}$$

We know the eigenvectors are orthogonal, so we must have

$$\mathbf{P}^T \mathbf{A} \mathbf{P} = \begin{bmatrix} \lambda_1 \langle \mathbf{E}_1, \mathbf{E}_1 \rangle & 0 \\ 0 & \lambda_2 \langle \mathbf{E}_2, \mathbf{E}_2 \rangle \end{bmatrix}$$

Once last step and we are done! There is no reason, we can't choose as our eigenvectors, vectors of length one: here just replace \mathbf{E}_1 by the new vector $\mathbf{E}_1/\|\mathbf{E}_1\|$ where $\|\mathbf{E}_1\|$ is the usual Euclidean length of the vector. Similarly, replace \mathbf{E}_2 by $\mathbf{E}_2/\|\mathbf{E}_2\|$. Assuming this is done, we have $\langle \mathbf{E}_1, \mathbf{E}_1 \rangle = 1$ and $\langle \mathbf{E}_2, \mathbf{E}_2 \rangle = 1$. We are left with the identity

$$\mathbf{P}^T \mathbf{A} \mathbf{P} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

which can be rewritten as

$$\mathbf{A} = \mathbf{P} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{P}^T$$

This is an important thing. We have shown the 2×2 matrix \mathbf{A} can be decomposed into the product $\mathbf{A} = \mathbf{P} \Lambda \mathbf{P}^T$ where Λ is the diagonal matrix whose entries are the eigenvalues of \mathbf{A} which the most positive one in the (1, 1) position.

It is now clear how we solve an equation like $\mathbf{A} \mathbf{X} = \mathbf{b}$. We rewrite as $\mathbf{P} \Lambda \mathbf{P}^T \mathbf{X} = \mathbf{b}$ which leads to the solution

$$\mathbf{X} = \mathbf{P} \Lambda^{-1} \mathbf{P}^T \mathbf{b}$$

and we see the reciprocal eigenvalue sizes determine how large the solution can get. Another way to look at this is that the two eigenvectors can be used to find a representation of the data vector \mathbf{b} and the solution vector \mathbf{X} as follows:

$$\begin{aligned} \mathbf{b} &= \langle \mathbf{b}, \mathbf{E}_1 \rangle \mathbf{E}_1 + \langle \mathbf{b}, \mathbf{E}_2 \rangle \mathbf{E}_2 = b_1 \mathbf{E}_1 + b_2 \mathbf{E}_2 \\ \mathbf{X} &= \langle \mathbf{X}, \mathbf{E}_1 \rangle \mathbf{E}_1 + \langle \mathbf{X}, \mathbf{E}_2 \rangle \mathbf{E}_2 = X_1 \mathbf{E}_1 + X_2 \mathbf{E}_2 \end{aligned}$$

and so $\mathbf{A} \mathbf{X} = \mathbf{b}$ becomes

$$\begin{aligned} \mathbf{A} \left(X_1 \mathbf{E}_1 + X_2 \mathbf{E}_2 \right) &= b_1 \mathbf{E}_1 + b_2 \mathbf{E}_2 \\ \lambda_1 X_1 \mathbf{E}_1 + \lambda_2 X_2 \mathbf{E}_2 &= b_1 \mathbf{E}_1 + b_2 \mathbf{E}_2. \end{aligned}$$

The only way this equation works is if the coefficients on the eigenvectors match. So we have $X_1 = \lambda_1^{-1} b_1$ and $X_2 = \lambda_2^{-1} b_2$. This shows very clearly how the solution depends on the size of the reciprocal eigenvalues. Thus, if our problem has a very small eigenvalue, we would expect our solution vector to be unstable. Also, if one of the eigenvalues is 0, we would have real problems! We

5.1. THE GENERAL TWO BY TWO SYMMETRIC MATRIX

79

can address this somewhat by finding a way to force all the eigenvalues to be positive.

The eigenvectors here are **independent vectors** in \mathbb{R}^2 and since they **span** \mathbb{R}^2 , they form a **basis** which is an **orthonormal basis** because the vectors are orthogonal. Hence, any vector \mathbf{V} in \mathbb{R}^2 can be written as

$$\mathbf{V} = V_1 \mathbf{E}_1 + V_2 \mathbf{E}_2$$

and the components V_1 and V_2 are known as the components of \mathbf{V} relative to the basis $\{\mathbf{E}_1, \mathbf{E}_2\}$. We often just refer to this basis as \mathbf{E} . Hence, a vector \mathbf{V} has many possible representations. To refresh your mind here, the one you are most used to is the one which uses the basis vectors

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ and } \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

which is called the **standard basis**. When we write

$$\mathbf{V} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$$

unless it is otherwise stated, we assume these are the components of \mathbf{V} with respect to the standard basis. Now let's go back to our general vector \mathbf{V} . Since the vectors \mathbf{E}_1 and \mathbf{E}_2 are orthogonal, we can take inner products on both sides of the representation of \mathbf{V} with respect to the basis \mathbf{E} to get

$$\langle \mathbf{V}, \mathbf{E}_1 \rangle = \langle V_1 \mathbf{E}_1, \mathbf{E}_1 \rangle + \langle V_2 \mathbf{E}_2, \mathbf{E}_1 \rangle = V_1 \langle \mathbf{E}_1, \mathbf{E}_1 \rangle + V_2 \langle \mathbf{E}_2, \mathbf{E}_1 \rangle$$

But $\langle \mathbf{E}_2, \mathbf{E}_1 \rangle = 0$ as the vectors are perpendicular and $\langle \mathbf{E}_2, \mathbf{E}_1 \rangle = E_{11}^2 + E_{12}^2$ where E_{11} and E_{12} are the components of \mathbf{E}_1 in the standard basis. This is one though as we have chosen our eigenvectors to have length one. So we have $V_1 = \langle \mathbf{V}, \mathbf{E}_1 \rangle$. A similar calculation with inner products gives $V_2 = \langle \mathbf{V}, \mathbf{E}_2 \rangle$. So we could also have written

$$\mathbf{V} = \langle \mathbf{V}, \mathbf{E}_1 \rangle \mathbf{E}_1 + \langle \mathbf{V}, \mathbf{E}_2 \rangle \mathbf{E}_2$$

From our discussions here, it should be easy to see that while we can do all of the needed calculations in this 2×2 case, we would have a lot of trouble if the symmetric matrix was 3×3 , 4×4 or larger. Hence, let's explore another way. And yes it is more abstract and yes, you will probably be lying belly up on the floor with your legs and arms pointing straight up soon enough screaming “enough”! But if you keep at it, you'll see that when computation becomes too hard, the abstract approach is quite nice!

5.1.2.1 Homework

Exercise 5.1.6**Exercise 5.1.7****Exercise 5.1.8****Exercise 5.1.9****Exercise 5.1.10**

5.1.3 Two Dimensional Rotations

Let's look closer at the matrices we have found using the eigenvectors of the symmetric matrix \mathbf{A} . We know the rows and columns of the matrix

$$\mathbf{P} = [E_1 \ E_2]$$

are mutually orthogonal and the rows and columns are length one. Thus,

$$\mathbf{P} = \begin{bmatrix} E_{11} & E_{21} \\ E_{12} & E_{22} \end{bmatrix}$$

Since the eigenvectors are length one, we know $\sqrt{E_{11}^2 + E_{12}^2} = 1$ and $\sqrt{E_{21}^2 + E_{22}^2} = 1$. We also know $E_{11}E_{21} + E_{12}E_{22} = 0$ and $E_{11}E_{12} + E_{21}E_{22} = 0$. Hence, the first column defines an angle θ by $\cos(\theta) = E_{11}$ and $\sin(\theta) = E_{12}$. The second column defines the angle ψ by $\cos(\psi) = E_{21}$ and $\sin(\psi) = E_{22}$. Since $E_{11}E_{21} + E_{12}E_{22} = 0$ and $E_{11}E_{12} + E_{21}E_{22} = 0$ we also find θ and ψ are not independent as $\cos(\theta)\sin(\theta) + \cos(\psi)\sin(\psi) = 0$ and $\cos(\theta)\cos(\psi) + \sin(\theta)\sin(\psi) = 0$. The second equation is useful as it tells us $\cos(\theta - \psi) = 0$. This means $\psi = \theta \pm \pi/2, \pm 3\pi/2, \dots$. For these choices of ψ , we have

- $\cos(\psi) = \cos(\theta + \pi/2) = -\sin(\theta)\sin(\pi/2) = -\sin(\theta)$ and $\sin(\psi) = \sin(\theta + \pi/2) = \cos(\theta)\sin(\pi/2) = \cos(\theta)$.
- $\cos(\psi) = \cos(\theta - \pi/2) = +\sin(\theta)\sin(\pi/2) = \sin(\theta)$ and $\sin(\psi) = \sin(\theta - \pi/2) = -\cos(\theta)\sin(\pi/2) = -\cos(\theta)$.

Both of these solutions work. You can check other choices such as $\psi = \theta \pm 3\pi/2$ do not give anything new. Hence, \mathbf{P} can be written two way

$$\mathbf{P}_1 = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \text{ and } \mathbf{P}_2 = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

The matrix \mathbf{P}_1 is a classical rotation matrix. The standard basis vectors \mathbf{i} and \mathbf{j} when rotated in a counterclockwise direction of angle θ moves to new basis vectors

$$\mathbf{I}_1 = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \text{ and } \mathbf{I}_2 = \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix}$$

You should draw this in \mathbb{R}^2 ! The matrix \mathbf{P}_2 is a rotation with a reflection. The basic vector \mathbf{i} is rotated to the new basis vector \mathbf{I}_1 , but instead of getting the new basis vector \mathbf{I}_2 , we get the reflection of it through the x axis giving

$$\mathbf{I}_1^{\text{reflection}} = \mathbf{I}_1 = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \text{ and } \mathbf{I}_2^{\text{reflection}} = -\mathbf{I}_2 = \begin{bmatrix} \sin(\theta) \\ -\cos(\theta) \end{bmatrix}$$

Note $\det \mathbf{P}_1 = \cos(\theta)^2 + \sin(\theta)^2 = 1$ and $\det \mathbf{P}_2 = -\cos(\theta)^2 - \sin(\theta)^2 = -1$ for any θ . Clearly $\mathbf{P}_1^{-1} = \mathbf{P}_1^T$ and $\mathbf{P}_2^{-1} = \mathbf{P}_2^T$. You should draw this in \mathbb{R}^2 also as it is instructive!

Thus, any 2×2 symmetric matrix \mathbf{A} has the decomposition

$$\mathbf{A} = \mathbf{P}_1^T \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{P}_1$$

or

$$\mathbf{A} = \mathbf{P}_2^T \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{P}_2 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}^T \mathbf{P}_1^T \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{P}_1 \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

5.1.4 Homework

Exercise 5.1.11

Exercise 5.1.12

Exercise 5.1.13

Exercise 5.1.14

Exercise 5.1.15

5.2 Rotating Surfaces

Here is a nice example of these ideas. We will do this both by hand and using MatLab so you can see how both approaches work. Consider the surface $z = 3x^2 + 2xy + 4y^2$ which is a rotated paraboloid with elliptical cross sections. Let's find the angle θ so that the change of variable $R_\theta \begin{bmatrix} x \\ y \end{bmatrix}$ to the new coordinates \bar{x} and \bar{y} leads to a surface equation with no cross terms $\bar{x}\bar{y}$. First note

$$3x^2 + 2xy + 4y^2 = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

where we let

$$\mathbf{H} = \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix}$$

The matrix \mathbf{H} here is symmetric so using the eigenvectors and eigenvalues of \mathbf{H} , we can write

$$[\mathbf{E}_1 \quad \mathbf{E}_2]^T \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} [\mathbf{E}_1 \quad \mathbf{E}_2] = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

The orthonormal basis $\{\mathbf{E}_1, \mathbf{E}_2\}$ can be identified with either a rotation R_θ or a reflection R_θ^r . We will choose the rotation. Then using the change of variables

$$\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} = R_\theta \begin{bmatrix} x \\ y \end{bmatrix}$$

we find

$$z = \begin{bmatrix} x \\ y \end{bmatrix}^T R_\theta^T \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} R_\theta \begin{bmatrix} x \\ y \end{bmatrix} = \lambda_1 \bar{x}^2 + \lambda_2 \bar{y}^2.$$

The characteristic equation for \mathbf{H} is

$$\det(\lambda \mathbf{I} - \mathbf{H}) = \lambda^2 - 7\lambda + 11 = 0$$

with roots

$$\lambda_1 = \frac{7 + \sqrt{5}}{2} = 4.618, \quad \lambda_2 = \frac{7 - \sqrt{5}}{2} = 2.382$$

It is straightforward to find the corresponding eigenvectors

$$\mathbf{W}_1 = \begin{bmatrix} 1 \\ \frac{1+\sqrt{5}}{2} \end{bmatrix}, \quad \mathbf{W}_2 = \begin{bmatrix} 1 \\ \frac{1-\sqrt{5}}{2} \end{bmatrix},$$

Then

$$\begin{aligned} \|\mathbf{W}_1\|_2^2 &= \sqrt{1 + \left(\frac{1+\sqrt{5}}{2}\right)^2} \implies \|\mathbf{W}_1\|_2 = \sqrt{3.618} = 1.902 \\ \|\mathbf{W}_2\|_2^2 &= \sqrt{1 + \left(\frac{1-\sqrt{5}}{2}\right)^2} \implies \|\mathbf{W}_2\|_2 = \sqrt{1.382} = 1.176 \end{aligned}$$

Then the orthonormal basis we need is

$$\begin{aligned} \mathbf{E}_1 &= \frac{\mathbf{W}_1}{\|\mathbf{W}_1\|_2} = \begin{bmatrix} .526 \\ .8505 \end{bmatrix} \\ \mathbf{E}_2 &= \pm \frac{\mathbf{W}_2}{\|\mathbf{W}_2\|_2} = \pm \begin{bmatrix} .8505 \\ -.526 \end{bmatrix} \end{aligned}$$

If we choose the minus choice, we get

$$\mathbf{E}_2 = \begin{bmatrix} -.8505 \\ .526 \end{bmatrix}$$

and $[\mathbf{E}_1, \mathbf{E}_2]$ is the rotation matrix \mathbf{R}_θ . If we choose the plus choice, we get

$$\mathbf{E}_2 = \begin{bmatrix} .8505 \\ -.526 \end{bmatrix}$$

and $[\mathbf{E}_1, \mathbf{E}_2]$ is the reflected rotation matrix \mathbf{R}_θ^r . Our rotation matrix is then

$$\mathbf{R}_\theta = \begin{bmatrix} .526 & -.8505 \\ .8505 & .526 \end{bmatrix} = \begin{bmatrix} \cos(1.017) & -\sin(1.017) \\ \sin(1.017) & \cos(1.017) \end{bmatrix}$$

where $\theta = 1.017$ radians. As a check,

$$\begin{bmatrix} .526 & .8505 \\ -.8505 & .526 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} .526 & -.8505 \\ .8505 & .526 \end{bmatrix} = \begin{bmatrix} .526 & .8505 \\ -.8505 & .526 \end{bmatrix} \begin{bmatrix} 2.4285 & -2.0255 \\ 3.928 & 1.2535 \end{bmatrix} = \begin{bmatrix} 4.618 & 0 \\ 0 & 2.382 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

The change of variables is

$$\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} = \mathbf{R}_\theta \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} .526 & -.8505 \\ .8505 & .526 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} .526x - .8505y \\ .8505x + .526y \end{bmatrix}$$

Now let's do the same computations in MatLab.

Listing 5.1: **Surface Rotation: Eigenvalues and Eigenvectors**

```
>> H = [3,1;1,4]
```

```

H =
3   1
1   4

>> [W,D] = eig(H)
W =
-0.85065  0.52573
0.52573  0.85065

D =
Diagonal Matrix

2.3820      0
0      4.6180

```

The command `eig` returns the unit eigenvectors as columns of V and the corresponding eigenvalues as the diagonal entries of D . You can see by looking at the matrix V the columns are not set up correctly to be a rotation matrix. We pick our orthonormal basis from V and hence our rotation matrix R as follows:

Listing 5.2: **Surface Rotation: Setting the Rotation Matrix**

```

>> E1 = W(:,2)
E1 =
0.52573
0.85065

>> E2 = W(:,1)
E2 =
-0.85065
0.52573

>> R = [E1,E2]
R =
0.52573 -0.85065
0.85065  0.52573

```

We then check the decomposition.

Listing 5.3: **Surface Rotation: Checking the decomposition**

```

>> R'*H*R
ans =
4.61803  0.00000
-0.00000  2.38197

```

5.2.1 Homework

Exercise 5.2.1

Exercise 5.2.2

Exercise 5.2.3

Exercise 5.2.4

Exercise 5.2.5

5.3 An Complex ODE System Example

Let's look at a system of ODEs with complex roots to give an nontrivial example of how to use rotation matrices and these matrix decompositions to understand a solution. First, let's do a quick review as it has probably been a long time since you thought about this material. Let's begin with a theoretical analysis. If the real valued matrix A has a complex eigenvalue $r = \alpha + i\beta$, then there is a nonzero vector G so that

$$AG = (\alpha + i\beta)G.$$

Now take the complex conjugate of both sides to find

$$\bar{A} \bar{G} = \overline{(\alpha + i\beta)} \bar{G}.$$

However, since A has real entries, its complex conjugate is simply A again. Thus, after taking complex conjugates, we find

$$A \bar{G} = (\alpha - i\beta) \bar{G}$$

and we conclude that if $\alpha + i\beta$ is an eigenvalue of A with eigenvector G , then the eigenvalue $\alpha - i\beta$ has eigenvector \bar{G} . Hence, letting E be the real part of G and F be the imaginary part, we see $E + iF$ is the eigenvector for $\alpha + i\beta$ and $E - iF$ is the eigenvector for $\alpha - i\beta$.

5.3.1 The General Real and Complex Solution

We can write down the general complex solution immediately.

$$\begin{bmatrix} \phi(t) \\ \psi(t) \end{bmatrix} = c_1 (E + iF) e^{(\alpha+i\beta)t} + c_2 (E - iF) e^{(\alpha-i\beta)t}$$

for arbitrary complex numbers c_1 and c_2 . We can reorganize this solution into a more convenient form as follows.

$$\begin{aligned} \begin{bmatrix} \phi(t) \\ \psi(t) \end{bmatrix} &= e^{\alpha t} \left(c_1 (E + iF) e^{(i\beta)t} + c_2 (E - iF) e^{(-i\beta)t} \right) \\ &= e^{\alpha t} \left((c_1 e^{(i\beta)t} + c_2 e^{(-i\beta)t}) E + i(c_1 e^{(i\beta)t} - c_2 e^{(-i\beta)t}) F \right). \end{aligned}$$

5.3. AN COMPLEX ODE SYSTEM EXAMPLE

85

The first real solution is found by choosing $c_1 = 1/2$ and $c_2 = 1/2$. This give

$$\begin{bmatrix} x_1(t) \\ y_1(t) \end{bmatrix} = e^{\alpha t} \left(\left((1/2)(e^{(i\beta)t} + e^{(-i\beta)t}) \right) \mathbf{E} + i \left((1/2)(e^{(i\beta)t} - e^{(-i\beta)t}) \right) \mathbf{F} \right).$$

However, we know that $(1/2)(e^{(i\beta)t} + e^{(-i\beta)t}) = \cos(\beta t)$ and $(1/2)(e^{(i\beta)t} - e^{(-i\beta)t}) = i \sin(\beta t)$. Thus, we have

$$\begin{bmatrix} x_1(t) \\ y_1(t) \end{bmatrix} = e^{\alpha t} \left(\mathbf{E} \cos(\beta t) - \mathbf{F} \sin(\beta t) \right).$$

The second real solution is found by setting $c_1 = 1/2i$ and $c_2 = -1/2i$ which gives

$$\begin{aligned} \begin{bmatrix} x_2(t) \\ y_2(t) \end{bmatrix} &= e^{\alpha t} \left(\left((1/2i)(e^{(i\beta)t} - e^{(-i\beta)t}) \right) \mathbf{E} + i \left((1/2i)(e^{(i\beta)t} + e^{(-i\beta)t}) \right) \mathbf{F} \right) \\ &= e^{\alpha t} \left(\mathbf{E} \sin(\beta t) + \mathbf{F} \cos(\beta t) \right). \end{aligned}$$

The general real solution is therefore

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = e^{\alpha t} \left(a (\mathbf{E} \cos(\beta t) - \mathbf{F} \sin(\beta t)) + b (\mathbf{E} \sin(\beta t) + \mathbf{F} \cos(\beta t)) \right)$$

for arbitrary real numbers a and b .

Example 5.3.1

$$\begin{aligned} x'(t) &= 2x(t) + 5y(t) \\ y'(t) &= -x(t) + 4y(t) \\ x(0) &= 6 \\ y(0) &= -1 \end{aligned}$$

Solution The characteristic is $r^2 - 6r + 13 = 0$ which gives the eigenvalues $3 \pm 2i$. The eigenvalue equation for the first root, $3 + 2i$ leads to this system to solve for nonzero \mathbf{V} .

$$\begin{bmatrix} (3+2i)-2 & -5 \\ 1 & (3+2i)-4 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This reduces to the system

$$\begin{bmatrix} 1+2i & -5 \\ 1 & -1+2i \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Although it is not immediately apparent, the second row is a multiple of row one. Multiply row one by $-1 - 2i$. This gives the row $[-1 - 2i, 5]$. Now multiple this new row by -1 to get $[1 + 2i, -5]$ which is row one. So even though it is harder to see, these two rows are equivalent and hence we only need to choose one to solve for V_2 in terms of V_1 . The first row gives $(1 + 2i)V_1 + 5V_2 = 0$. Letting $V_1 = a$, we find $V_2 = (-1 - 2i)/5 a$. Hence, the eigenvectors have the form

$$\mathbf{G} = a \begin{bmatrix} 1 \\ -\frac{1+2i}{5} \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{1}{5} \end{bmatrix} + i \begin{bmatrix} 0 \\ -\frac{2}{5} \end{bmatrix}$$

Hence,

$$\mathbf{E} = \begin{bmatrix} 1 \\ -\frac{1}{5} \end{bmatrix} \text{ and } \mathbf{F} = \begin{bmatrix} 0 \\ -\frac{2}{5} \end{bmatrix}$$

The general real solution is therefore

$$\begin{aligned} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} &= e^{3t} \left(a \left(\mathbf{E} \cos(2t) - \mathbf{F} \sin(2t) \right) + b \left(\mathbf{E} \sin(2t) + \mathbf{F} \cos(2t) \right) \right) \\ &= e^{3t} \left(a \left(\begin{bmatrix} 1 \\ -\frac{1}{5} \end{bmatrix} \cos(2t) - \begin{bmatrix} 0 \\ -\frac{2}{5} \end{bmatrix} \sin(2t) \right) + b \left(\begin{bmatrix} 1 \\ -\frac{1}{5} \end{bmatrix} \sin(2t) + \begin{bmatrix} 0 \\ -\frac{2}{5} \end{bmatrix} \cos(2t) \right) \right) \\ &= e^{3t} \left[\begin{array}{c} a \cos(2t) + b \sin(2t) \\ (-\frac{1}{5}a - \frac{2}{5}b) \cos(2t) (\frac{2}{5}a + \frac{1}{5}b) \sin(2t) \end{array} \right] \end{aligned}$$

Now apply the initial conditions to obtain

$$\begin{bmatrix} 6 \\ -1 \end{bmatrix} = e^{3t} \begin{bmatrix} a \\ (-\frac{1}{5}a - \frac{2}{5}b) \end{bmatrix}$$

Thus, $a = 6$ and $b = -1/2$. The solution is therefore

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = e^{3t} \begin{bmatrix} 6 \cos(2t) - \frac{1}{2} \sin(2t) \\ -\cos(2t) - \frac{23}{10} \sin(2t) \end{bmatrix}$$

5.3.1.1 Homework

Exercise 5.3.1

Exercise 5.3.2

Exercise 5.3.3

Exercise 5.3.4

Exercise 5.3.5

5.3.2 Rewriting The Real Solution

We know the real solution is written as

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = e^{\alpha t} \left(\left(a \mathbf{E} + b \mathbf{F} \right) \cos(\beta t) + \left(b \mathbf{E} - a \mathbf{F} \right) \sin(\beta t) \right)$$

Now rewrite again in terms of the components of \mathbf{E} and \mathbf{F} to obtain

$$\begin{aligned} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} &= e^{\alpha t} \left(\left(a \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} + b \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \right) \cos(\beta t) + \left(b \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} - a \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \right) \sin(\beta t) \right) \\ &= e^{\alpha t} \begin{bmatrix} aE_1 + bF_1 & bE_1 - aF_1 \\ aE_2 + bF_2 & bE_2 - aF_2 \end{bmatrix} \begin{bmatrix} \cos(\beta t) \\ \sin(\beta t) \end{bmatrix} \end{aligned}$$

Finally, we can move back to the vector form and write

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = e^{\alpha t} [a\mathbf{E} + b\mathbf{F}, \quad b\mathbf{E} - a\mathbf{F}] \begin{bmatrix} \cos(\beta t) \\ \sin(\beta t) \end{bmatrix}.$$

5.3. AN COMPLEX ODE SYSTEM EXAMPLE

87

Now we want nonzero solutions, so our initial conditions will give us a and b so that $\sqrt{a^2 + b^2} \neq 0$. We have so far

$$\begin{aligned} \begin{bmatrix} \frac{x(t)}{e^{\alpha t}} \\ \frac{y(t)}{e^{\alpha t}} \end{bmatrix} &= \begin{bmatrix} aE_1 + bF_1 & bE_1 - aF_1 \\ aE_2 + bF_2 & bE_2 - aF_2 \end{bmatrix} \begin{bmatrix} \cos(\beta t) \\ \sin(\beta t) \end{bmatrix} \\ &= [\mathbf{E} \quad \mathbf{F}] \begin{bmatrix} a & b \\ b & -a \end{bmatrix} \begin{bmatrix} \cos(\beta t) \\ \sin(\beta t) \end{bmatrix} \\ &= \sqrt{a^2 + b^2} [\mathbf{E} \quad \mathbf{F}] \begin{bmatrix} \frac{a}{\sqrt{a^2 + b^2}} & \frac{b}{\sqrt{a^2 + b^2}} \\ \frac{b}{\sqrt{a^2 + b^2}} & -\frac{a}{\sqrt{a^2 + b^2}} \end{bmatrix} \begin{bmatrix} \cos(\beta t) \\ \sin(\beta t) \end{bmatrix} \end{aligned}$$

Now define the angle θ by $\cos(\theta) = a/\sqrt{a^2 + b^2}$ and so $\sin(\theta) = b/\sqrt{a^2 + b^2}$. Further let $L = \sqrt{a^2 + b^2}$. Then, we can rewrite the solution again as

$$\begin{bmatrix} \frac{x(t)}{Le^{\alpha t}} \\ \frac{y(t)}{Le^{\alpha t}} \end{bmatrix} = [\mathbf{E} \quad \mathbf{F}] \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix} \begin{bmatrix} \cos(\beta t) \\ \sin(\beta t) \end{bmatrix}$$

The matrix involving θ is a reflection matrix with angle θ which we will denote by \mathbb{R}_θ^r . Then we have

$$\begin{bmatrix} \frac{x(t)}{Le^{\alpha t}} \\ \frac{y(t)}{Le^{\alpha t}} \end{bmatrix} = [\mathbf{E} \quad \mathbf{F}] \mathbb{R}_\theta^r \begin{bmatrix} \cos(\beta t) \\ \sin(\beta t) \end{bmatrix}$$

which we can rewrite in terms of the rotation matrix \mathbf{R}_θ as

$$\begin{bmatrix} \frac{x(t)}{Le^{\alpha t}} \\ \frac{y(t)}{Le^{\alpha t}} \end{bmatrix} = [\mathbf{E} \quad \mathbf{F}] \mathbf{R}_\theta \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \cos(\beta t) \\ \sin(\beta t) \end{bmatrix} = [\mathbf{E} \quad \mathbf{F}] \mathbf{R}_\theta \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix}$$

Let $u(t) = x(t)/(Le^{\alpha t})$ and $v(t) = y(t)/(Le^{\alpha t})$. Then

$$\begin{aligned} u^2(t) + v^2(t) &= \left\langle [\mathbf{E} \quad \mathbf{F}] \mathbf{R}_\theta \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix}, [\mathbf{E} \quad \mathbf{F}] \mathbf{R}_\theta \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix} \right\rangle \\ &= \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix}^T \mathbf{R}_\theta^T [\mathbf{E} \quad \mathbf{F}]^T [\mathbf{E} \quad \mathbf{F}]^T \mathbf{R}_\theta \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix} \end{aligned}$$

Now $[\mathbf{E} \quad \mathbf{F}]^T [\mathbf{E} \quad \mathbf{F}]$ is a 2×2 symmetric matrix. Hence, we have a representation of it in terms of its real eigenvalues λ_1 and λ_2 and associated eigenvectors \mathbf{G}_1 and \mathbf{G}_2 .

$$[\mathbf{E} \quad \mathbf{F}]^T [\mathbf{E} \quad \mathbf{F}] = [\mathbf{G}_1, \mathbf{G}_2]^T \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} [\mathbf{G}_1, \mathbf{G}_2]$$

We also know from our earlier discussions, that $[\mathbf{G}_1, \mathbf{G}_2]$ is either a rotation matrix \mathbf{R}_ψ or a reflection $\mathbf{R}_\psi \mathbf{J}$ where

$$\mathbf{J} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

Thus, we have, in the case it is a rotation matrix

$$u^2(t) + v^2(t) = \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix}^T \mathbf{R}_\theta^T \mathbf{R}_\psi^T \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{R}_\psi \mathbf{R}_\theta \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix}$$

It is a straightforward calculation to show $\mathbf{R}_\psi \mathbf{R}_\theta = \mathbf{R}_{\psi+\theta}$. Thus, for a rotation matrix

$$u^2(t) + v^2(t) = \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix}^T \mathbf{R}_{\psi+\theta}^T \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{R}_{\psi+\theta} \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix}$$

Now note we can rewrite this in terms of new coordinates based on new unit vectors.

- The reflection

$$\mathbf{J} \begin{bmatrix} \cos(\beta t) \\ \sin(\beta t) \end{bmatrix} = \begin{bmatrix} \cos(\beta t) \\ -\sin(\beta t) \end{bmatrix}$$

corresponds to $i \rightarrow i' = i$ and $j \rightarrow j' = -j$. Call this the $x' - y'$ coordinate system. Note these vectors live on the unit circle in the $x' - y'$ coordinate system.

- Now apply the rotation $\mathbf{R}_{\psi+\theta}$ which rotates the reflected system above counterclockwise an angle of $\psi + \theta$. The coordinates in this new system are (\bar{x}, \bar{y}) . Call this the $x'' - y''$ coordinate system. Note these vectors live on the unit circle in the $x'' - y''$ coordinate system. The coordinates in this new system are (\bar{x}, \bar{y}) .

So we have

$$u^2(t) + v^2(t) = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}^T \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} = \lambda_1(\bar{x})^2 + \lambda_2(\bar{y})^2$$

which is an ellipse if both eigenvalues are positive, a circle if they are positive and equal and a hyperbola if they differ in algebraic sign. We get various degenerate possibilities if one of the eigenvalues is zero.

The analysis in the case of the reflected case is quite similar and the matrix \mathbf{J} does not alter the result qualitatively. Rewriting again, we find

$$x^2(t) + y^2(t) = L^2 e^{2\alpha t} (\lambda_1(\bar{x})^2 + \lambda_2(\bar{y})^2)$$

which shows we have spiral out, constant or spiral in trajectories depending on the sign of α .

5.3.2.1 Homework

Exercise 5.3.6

Exercise 5.3.7

Exercise 5.3.8

Exercise 5.3.9

Exercise 5.3.10

5.3.3 Signed Definite Matrices

A 2×2 matrix is said to be a **positive definite** matrix if $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ for all vectors \mathbf{x} . If we multiply this out, we find the inequality below

$$ax_1^2 + 2bx_1x_2 + dx_2^2 > 0,$$

If we complete the square, we find

$$a \left(\left(x_1 + (b/a)x_2 \right)^2 + \left((ad - b^2)/a^2 \right) x_2^2 \right) > 0$$

Now the leading term $a > 0$, and if the determinant of $\mathbf{A} = ad - b^2 > 0$, we would have the quadratic $\left(x_1 + (b/a)x_2 \right)^2 + \left((ad - b^2)/a^2 \right) x_2^2$ is always positive. Note that since this determinant is positive $ad > b^2$ which forces d to be positive as well. So in this case, a and d and $ad - b^2 > 0$. And the expression $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ in this case. Now recall what we found about the eigenvalues here. We had the eigenvalues were

$$r = \frac{(a+d) \pm \sqrt{(a-d)^2 + 4b^2}}{2}$$

Since $ad - b^2 > 0$, the term

$$(a-d)^2 + 4b^2 = a^2 - 2ad + d^2 + 4b^2 < a^2 - 2ad + 4ad + d^2 = (a+d)^2.$$

Thus, the square root is smaller than $a+d$ as a and d are positive. the first root is always positive and the second root is too as $(a+d) - \sqrt{(a-d)^2 + 4b^2} > a+d - (a+d) = 0$. So both eigenvalues are positive if a and d are positive and $ad - b^2 > 0$. Note the argument can go the other way. If we assume the matrix is positive definite, the we are forced to have $a > 0$ and $ad - b^2 > 0$ which gives the same result. We conclude our 2×2 symmetric matrix \mathbf{A} is positive definite if and only if $a > 0$, $d > 0$ and the determinant of $\mathbf{A} > 0$ too. Note a positive definite matrix has positive eigenvalues. A similar argument holds if we have determinant of $\mathbf{A} > 0$ but $a < 0$. The determinant condition will then force $d < 0$ too. We find that $\mathbf{x}^T \mathbf{A} \mathbf{x} < 0$. In this case, we say the matrix is **negative definite**. The eigenvalues are still

$$r = \frac{(a+d) \pm \sqrt{(a-d)^2 + 4b^2}}{2}.$$

But now, since $ad - b^2 > 0$, the term

$$(a-d)^2 + 4b^2 = a^2 - 2ad + d^2 + 4b^2 < a^2 - 2ad + 4ad + d^2 = |a+d|^2.$$

Since a and d are negative, $a+d < 0$ and so the second root is always negative. The first root's sign is determined by $(a+d) + \sqrt{(a-d)^2 + 4b^2} < (a+d) + |a+d| = 0$. So both eigenvalues are negative. We have found the matrix \mathbf{A} is negative definite if a and d are negative and the determinant of $\mathbf{A} > 0$. Note a negative definite matrix has negative eigenvalues.

5.3.4 Summarizing

Let put all the information we have together. We have been studying a special type of 2×2 matrix which is symmetric. This matrix \mathbf{A} has the form

$$\mathbf{A} = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

where a, b and d are arbitrary non zero numbers. The characteristic equation here is $r^2 - (a+d)r + ad - b^2$. Note that the term $ad - b^2$ is the determinant of \mathbf{A} . The roots are given by

$$r = \frac{(a+d) \pm \sqrt{(a-d)^2 + 4b^2}}{2}$$

- This matrix always has two distinct real eigenvalues and their respective eigenvectors \mathbf{E}_1 and \mathbf{E}_2 are mutually orthogonal. Hence, we usually normalize them so they form an orthonormal basis for \mathbb{R}^2 .
- We can use these eigenvectors to write \mathbf{A} in a canonical form.

$$\mathbf{A} = \mathbf{P} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{P}^T$$

where

$$\mathbf{P} = [\mathbf{E}_1 \quad \mathbf{E}_2]$$

giving

$$\mathbf{P}^T = \begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \end{bmatrix}$$

- If we solve $\mathbf{A} \mathbf{X} = \mathbf{b}$, it is most illuminating to rewrite both the unknown \mathbf{X} and the data \mathbf{b} with respect to the orthonormal basis determined by the eigenvectors of \mathbf{A} . When we do this we find the solution is expressed quite nicely $X_1 = \lambda_1^{-1} b_1$ and $X_2 = \lambda_2^{-1} b_2$ where b_1 and b_2 are the components of \mathbf{b} in the eigenvector basis. This shows very clearly how the solution depends on the size of the reciprocal eigenvalues.
- If $a > 0$ and the determinant of $\mathbf{A} = ad - b^2 > 0$, then $d > 0$ too and \mathbf{A} is positive definite. In this case, both eigenvalues are positive. If we started by assuming \mathbf{A} was positive definite, our chain of inference would lead us to the same conclusions about a , b and d .
- If $a < 0$ and the determinant of $\mathbf{A} = ad - b^2 > 0$, then $d < 0$ also and \mathbf{A} is negative definite. In this case, both eigenvalues are negative. And assuming \mathbf{A} is negative definite leads us backwards to $a < 0$, $d < -0$ and $ad - b^2 > 0$.

It seems obvious to ask if we can say similar things about symmetric matrices that are 3×3 , 4×4 and so on. Clearly, we can't use our determinant route to find answers as analyzing cubics and quartics is quite impossible. So the next section tries another approach. The dreaded road we call **let's go abstract**.

5.3.4.1 Homework

Exercise 5.3.11

Exercise 5.3.12

Exercise 5.3.13

Exercise 5.3.14

Exercise 5.3.15

Chapter 6

Matrix Sequences and Ordinary Differential Equations

Here is a nice application of symmetric matrices to systems of two linear differential equations. Let's do some review first, as the ODE material seems to always slip away right when you need it.

6.1 Linear Systems of ODE

Let's review Linear Systems of first order ODE. These have the form

$$\begin{aligned}x'(t) &= a x(t) + b y(t) \\y'(t) &= c x(t) + d y(t) \\x(0) &= x_0 \\y(0) &= y_0\end{aligned}$$

for any numbers a, b, c and d and *initial conditions* x_0 and y_0 . The full problem is called, as usual, an *Initial Value Problem* or **IVP** for short. The two initial conditions are just called the **IC**'s for the problem to save writing.

- For example, we might be interested in the system

$$\begin{aligned}x'(t) &= -2 x(t) + 3 y(t) \\y'(t) &= 4 x(t) + 5 y(t) \\x(0) &= 5 \\y(0) &= -3\end{aligned}$$

Here the **IC**'s are $x(0) = 5$ and $y(0) = -3$.

- Another sample problem might be the one below.

$$\begin{aligned}x'(t) &= 14 x(t) + 5 y(t) \\y'(t) &= -4 x(t) + 8 y(t) \\x(0) &= 2 \\y(0) &= 7\end{aligned}$$

6.1.1 The Characteristic Equation

For linear first order problems like $u' = 3u$ and so forth, we find the solution has the form $u(t) = Ae^{3t}$ for some number A . We then determine the value of A to use by looking at the initial condition. To find the solutions here, we begin by rewriting the model in matrix - vector notation.

$$\begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}.$$

The 2×2 matrix is called the **coefficient matrix** of this model. The initial conditions can then be redone in vector form as

$$\begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}.$$

Now it seems reasonable to believe that if a constant times e^{rt} solves a first order linear problem like $u' = ru$, perhaps a *vector* times e^{rt} will work here. Let's make this formal. So let's look at the problem below

$$\begin{aligned} x'(t) &= 3x(t) + 2y(t) \\ y'(t) &= -4x(t) + 5y(t) \\ x(0) &= 2 \\ y(0) &= -3 \end{aligned}$$

Assume the solution has the form $\mathbf{V} e^{rt}$. Let's denote the components of \mathbf{V} as follows:

$$\mathbf{V} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}.$$

We assume the solution is

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \mathbf{V} e^{rt}.$$

Then the derivative of $\mathbf{V} e^{rt}$ is

$$\begin{aligned} (\mathbf{V} e^{rt})' &= \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} e^{rt} \right)' = \left(\begin{bmatrix} V_1 e^{rt} \\ V_2 e^{rt} \end{bmatrix} \right)' \\ &= \begin{bmatrix} V_1 (e^{rt})' \\ V_2 (e^{rt})' \end{bmatrix} = \begin{bmatrix} V_1 r e^{rt} \\ V_2 r e^{rt} \end{bmatrix} \\ &= \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} r e^{rt} = r \mathbf{V} e^{rt} \end{aligned}$$

Hence,

$$\begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} = r \mathbf{V} e^{rt}.$$

When we plug these terms into the matrix - vector form of the problem, we find

$$r \mathbf{V} e^{rt} = \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \mathbf{V} e^{rt}$$

6.1. LINEAR SYSTEMS OF ODE

93

Rewrite as

$$r \mathbf{V} e^{rt} - \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \mathbf{V} e^{rt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Recall that the 2×2 identity matrix \mathbf{I} has the form

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \Rightarrow \mathbf{I} \mathbf{V} = \mathbf{V}.$$

So

$$\begin{aligned} r \mathbf{V} e^{rt} - \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \mathbf{V} e^{rt} &= r \mathbf{I} \mathbf{V} e^{rt} - \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \mathbf{V} e^{rt} \\ &= \left(r \mathbf{I} - \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \right) \mathbf{V} e^{rt} \\ &= \left(\begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix} - \begin{bmatrix} 3 & 2 \\ -4 & 5 \end{bmatrix} \right) \mathbf{V} e^{rt} \\ &= \begin{bmatrix} r-3 & -2 \\ -(-4) & r-5 \end{bmatrix} \mathbf{V} e^{rt} \end{aligned}$$

Plugging this into our model, we find

$$\begin{bmatrix} r-3 & -2 \\ 4 & r-5 \end{bmatrix} \mathbf{V} e^{rt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

But e^{rt} is never 0, so we want r satisfying

$$\begin{bmatrix} r-3 & -2 \\ 4 & r-5 \end{bmatrix} \mathbf{V} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

For each r , we get two equations in V_1 and V_2 :

$$(r-3)V_1 - 2V_2 = 0, \quad 4V_1 + (r-5)V_2 = 0.$$

Let \mathbf{A}_r be this matrix. Any r for which $\det \mathbf{A}_r \neq 0$ tells us these two lines have different slopes and so cross at the origin implying $V_1 = 0$ and $V_2 = 0$. Thus

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} e^{rt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

which will not satisfy **nonzero initial conditions**. So **reject** these r . Any value of r for which $\det \mathbf{A}_r = 0$ gives an infinite number of solutions which allows us to pick one that matches the initial conditions we have. The equation

$$\det(r\mathbf{I} - \mathbf{A}) = \det \begin{bmatrix} r-3 & -2 \\ 4 & r-5 \end{bmatrix} = 0.$$

is called the **characteristic equation** of this linear system. The **characteristic equation** is a quadratic, so there are three possibilities: two distinct roots, two real roots that match and a complex conjugate pair of roots.

Example 6.1.1 Derive the characteristic equation for the system below

$$\begin{aligned}x'(t) &= 8x(t) + 9y(t) \\y'(t) &= 3x(t) - 2y(t) \\x(0) &= 12 \\y(0) &= 4\end{aligned}$$

Solution The matrix - vector form is

$$\begin{aligned}\begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} &= \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \\ \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} &= \begin{bmatrix} 12 \\ 4 \end{bmatrix}\end{aligned}$$

The coefficient matrix \mathbf{A} is thus

$$\mathbf{A} = \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix}$$

Assume the solution has the form $\mathbf{V} e^{rt}$. Plug this into the system.

$$r \mathbf{V} e^{rt} - \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix} \mathbf{V} e^{rt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Rewrite using the identity matrix I and factor

$$\left(r \mathbf{I} - \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix}\right) \mathbf{V} e^{rt} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Since $e^{rt} \neq 0$ ever, we find r and \mathbf{V} satisfy

$$\left(r \mathbf{I} - \begin{bmatrix} 8 & 9 \\ 3 & -2 \end{bmatrix}\right) \mathbf{V} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

If r is chosen so that $\det(rI - A) \neq 0$, the only solution to this system of two linear equations in the two unknowns V_1 and V_2 is $V_1 = 0$ and $V_2 = 0$. This leads to $x(t) = 0$ and $y(t) = 0$ always and this solution does not satisfy the initial conditions. Hence, we must find r which give $\det(rI - A) = 0$. The characteristic equation is thus

$$\begin{aligned}\det \begin{bmatrix} r-8 & -9 \\ -3 & r+2 \end{bmatrix} &= (r-8)(r+2) - 27 \\ &= r^2 - 6r - 43 = 0\end{aligned}$$

6.1.2 Finding The General Solution

The roots to the characteristic equation are called **eigenvalues**. For each eigenvalue r we want to find nonzero vectors \mathbf{V} so that $(r \mathbf{I} - \mathbf{A}) \mathbf{V} = \mathbf{0}$ where to help with our writing we let $\mathbf{0}$ be the two dimensional zero vector. These nonzero \mathbf{V} are called the **eigenvectors** for **eigenvalue** r and satisfy $\mathbf{AV} = r\mathbf{V}$.

For eigenvalue r_1 , find \mathbf{V} so that $(r_1 \mathbf{I} - \mathbf{A}) \mathbf{V} = \mathbf{0}$. There will be an infinite number of \mathbf{V} 's that solve this; we pick one and call it **eigenvector** \mathbf{E}_1 .

For eigenvalue r_2 , find \mathbf{V} so that $(r_2 \mathbf{I} - \mathbf{A}) \mathbf{V} = \mathbf{0}$. There will again be an infinite number of \mathbf{V} 's that solve this; we pick one and call it **eigenvector** \mathbf{E}_2 .

The general solution to our model will be

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = A\mathbf{E}_1 e^{r_1 t} + B\mathbf{E}_2 e^{r_2 t}.$$

where A and B are arbitrary. We use the ICs to find A and B . It is best to show all this with some examples.

Example 6.1.2 For the system below

$$\begin{aligned} \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} &= \begin{bmatrix} -20 & 12 \\ -13 & 5 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \\ \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} &= \begin{bmatrix} -1 \\ 2 \end{bmatrix} \end{aligned}$$

- Find the characteristic equation
- Find the general solution
- Solve the IVP

Solution The characteristic equation is

$$\det \left(r \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} -20 & 12 \\ -13 & 5 \end{bmatrix} \right) = 0$$

$$\begin{aligned} 0 &= \det \left(\begin{bmatrix} r+20 & -12 \\ 13 & r-5 \end{bmatrix} \right) \\ &= (r+20)(r-5) + 156 \\ &= r^2 + 15r + 56 \\ &= (r+8)(r+7) \end{aligned}$$

Hence, **eigenvalues or roots** of the characteristic equation are $r_1 = -8$ and $r_2 = -7$. Note since this is just a calculation, we are not following our labeling scheme.

For eigenvalue $r_1 = -8$, substitute the value into

$$\begin{bmatrix} r+20 & -12 \\ 13 & r-5 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \Rightarrow \begin{bmatrix} 12 & -12 \\ 13 & -13 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

This system of equations should be collinear: i.e. the rows should be multiples; i.e. both give rise to the same line. Our rows are multiples, so we can pick any row to find V_2 in terms of V_1 . Picking the top row, we get $12V_1 - 12V_2 = 0$ implying $V_2 = V_1$. Letting $V_1 = a$, we find $V_1 = a$ and $V_2 = a$: so

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = a \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

96 CHAPTER 6. MATRIX SEQUENCES AND ORDINARY DIFFERENTIAL EQUATIONS

Choose \mathbf{E}_1 :
The vector

$$\mathbf{E}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

is our choice for an eigenvector corresponding to eigenvalue $r_1 = -8$. So one of the solutions is

$$\begin{bmatrix} x_1(t) \\ y_1(t) \end{bmatrix} = \mathbf{E}_1 e^{-8t} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-8t}.$$

For eigenvalue $r_2 = -7$, substitute the value into

$$\begin{bmatrix} r+20 & -12 \\ 13 & r-5 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \Rightarrow \begin{bmatrix} 13 & -12 \\ 13 & -12 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

This system of equations should be collinear: i.e. the rows should be multiples; i.e. both give rise to the same line. Our rows are multiples, so we can pick any row to find V_2 in terms of V_1 . Picking the top row, we get $13V_1 - 12V_2 = 0$ implying $V_2 = (13/12)V_1$. Letting $V_1 = b$, we find $V_1 = b$ and $V_2 = (13/12)b$: so

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = b \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix}$$

Choose \mathbf{E}_2 :
The vector

$$\mathbf{E}_2 = \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix}$$

is our choice for an eigenvector corresponding to eigenvalue $r_2 = -7$. So one of the solutions is

$$\begin{bmatrix} x_2(t) \\ y_2(t) \end{bmatrix} = \mathbf{E}_2 e^{-7t} = \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} e^{-7t}.$$

The general solution is then

$$\begin{aligned} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} &= A \mathbf{E}_1 e^{-8t} + B \mathbf{E}_2 e^{-7t} \\ &= A \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-8t} + B \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} e^{-7t} \end{aligned}$$

Now let's solve the initial value problem: we find A and B.

$$\begin{aligned} \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} &= \begin{bmatrix} -1 \\ 2 \end{bmatrix} = A \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^0 + B \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} e^0 \\ &= A \begin{bmatrix} 1 \\ 1 \end{bmatrix} + B \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} \end{aligned}$$

So

$$\begin{aligned} A + B &= -1 \\ A + \frac{13}{12}B &= 2 \end{aligned}$$

Subtracting the bottom equation from the top equation, we get $-\frac{1}{12}B = -3$ or $B = 36$. Thus, $A = -1 - B = -37$. So

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = -37 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-8t} + 36 \begin{bmatrix} 1 \\ \frac{13}{12} \end{bmatrix} e^{-7t}$$

6.1.3 Homework

For these problems

- Write matrix, vector form.
- Derive the characteristic equation.
- Find the two eigenvalues. Label the largest one as r_1 and the other as r_2
- Find the two associated eigenvectors as unit vectors
- Define

$$\mathbf{P} = [\mathbf{E}_1 \quad \mathbf{E}_2]$$

- Compute $\mathbf{P}^T \mathbf{A} \mathbf{P}$ where \mathbf{A} is the coefficient matrix of the ODE system.
- Show $\mathbf{A} = \mathbf{P} \Lambda \mathbf{P}^T$ for an appropriate Λ .
- Write general solution.
- Solve the IVP.

Exercise 6.1.1

$$\begin{aligned} x' &= x + 2y \\ y' &= 2x - 6y \\ x(0) &= 4 \\ y(0) &= -6. \end{aligned}$$

Exercise 6.1.2

$$\begin{aligned} x' &= 2x + 3y \\ y' &= 3x + 7y \\ x(0) &= 2 \\ y(0) &= -3. \end{aligned}$$

6.2 Symmetric Systems of ODEs

Let's look at a specific symmetric ODE system and find its solution.

Example 6.2.1 For the symmetric system below

$$\begin{aligned} \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} &= \begin{bmatrix} -20 & 12 \\ 12 & 5 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \\ \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} &= \begin{bmatrix} -1 \\ 2 \end{bmatrix} \end{aligned}$$

98 CHAPTER 6. MATRIX SEQUENCES AND ORDINARY DIFFERENTIAL EQUATIONS

- Find the characteristic equation
- Find the general solution
- Solve the IVP

Solution The characteristic equation is

$$\det \left(r \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} -20 & 12 \\ 12 & 5 \end{bmatrix} \right) = 0$$

Thus

$$\begin{aligned} 0 &= \det \left(\begin{bmatrix} r+20 & -12 \\ -12 & r-5 \end{bmatrix} \right) \\ &= (r+20)(r-5) - 144 \\ &= r^2 + 15r - 244 \end{aligned}$$

Hence, **eigenvalues or roots** of the characteristic equation are $r_1 = 9.83$ and $r_2 = -24.83$.

For eigenvalue $r_1 = 9.83$, substitute the value into

$$\begin{bmatrix} r+20 & -12 \\ -12 & r-5 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \Rightarrow \begin{bmatrix} 29.83 & -12 \\ -12 & 4.83 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

Picking the top row, we get $29.83V_1 - 12V_2 = 0$ implying $V_2 = 6.18V_1$. Letting $V_1 = a$, we find $V_1 = a$ and $V_2 = 6.18a$: so

$$\mathbf{V} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = a \begin{bmatrix} 1 \\ 6.18 \end{bmatrix}$$

Let $a = 1/\| [1 \ 6.18]^T \| = 1/6.26 = .16$ and choose the unit eigenvector

$$\mathbf{E}_1 = 0.16 \begin{bmatrix} 1 \\ 6.18 \end{bmatrix} = \begin{bmatrix} .16 \\ .99 \end{bmatrix}$$

So one of the solutions is

$$\begin{bmatrix} x_1(t) \\ y_1(t) \end{bmatrix} = \mathbf{E}_1 e^{9.83t} = \begin{bmatrix} .16 \\ .99 \end{bmatrix} e^{9.83t}.$$

For eigenvalue $r_2 = -24.83$, substitute the value into

$$\begin{bmatrix} r+20 & -12 \\ -12 & r-5 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \Rightarrow \begin{bmatrix} -4.83 & -12 \\ -12 & -29.83 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

Picking the top row, we get $-4.83V_1 - 29.83V_2 = 0$ implying $V_2 = -.16V_1$. Letting $V_1 = b$, we find $V_1 = b$ and $V_2 = -.16b$: so

$$\mathbf{V} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = b \begin{bmatrix} 1 \\ -.16 \end{bmatrix}$$

Let $b = 1/\| [1 \ -.16]^T \| = 1/1.01 = .99$ and choose the unit eigenvector

$$\mathbf{E}_2 = .99 \begin{bmatrix} 1 \\ -.16 \end{bmatrix} = \begin{bmatrix} .99 \\ -.16 \end{bmatrix}$$

6.2. SYMMETRIC SYSTEMS OF ODES

99

Note $\langle \mathbf{E}_1, \mathbf{E}_2 \rangle = (.16)(.99) + (.99)(-.16) = 0$. So the other solution is

$$\begin{bmatrix} x_2(t) \\ y_2(t) \end{bmatrix} = \mathbf{E}_2 e^{-29.83t} = \begin{bmatrix} .99 \\ -.16 \end{bmatrix} e^{-29.83t}.$$

The general solution

$$\begin{aligned} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} &= A \mathbf{E}_1 e^{9.83t} + B \mathbf{E}_2 e^{-29.83t} \\ &= A \begin{bmatrix} .16 \\ .99 \end{bmatrix} e^{9.83t} + B \begin{bmatrix} .99 \\ -.16 \end{bmatrix} e^{-29.83t} \end{aligned}$$

Finally, we find A and B using the initial conditions.

$$\begin{aligned} \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} &= \begin{bmatrix} -1 \\ 2 \end{bmatrix} = A \begin{bmatrix} .16 \\ .99 \end{bmatrix} e^0 + B \begin{bmatrix} .99 \\ -.16 \end{bmatrix} e^0 \\ &= A \begin{bmatrix} .16 \\ .99 \end{bmatrix} + B \begin{bmatrix} .99 \\ -.16 \end{bmatrix} \end{aligned}$$

Thus,

$$\begin{aligned} [\mathbf{E}_1 \quad \mathbf{E}_2] \begin{bmatrix} A \\ B \end{bmatrix} &= \begin{bmatrix} -1 \\ 2 \end{bmatrix} \Rightarrow [\mathbf{E}_1^T \quad \mathbf{E}_2^T] [\mathbf{E}_1 \quad \mathbf{E}_2] \begin{bmatrix} A \\ B \end{bmatrix} = [\mathbf{E}_1^T \quad \mathbf{E}_2^T] \begin{bmatrix} -1 \\ 2 \end{bmatrix} \\ \begin{bmatrix} A \\ B \end{bmatrix} &= [\mathbf{E}_1^T \quad \mathbf{E}_2^T] \begin{bmatrix} -1 \\ 2 \end{bmatrix} \end{aligned}$$

Letting $\mathbf{D} = [-1 \quad 2]^T$, $A = \langle \mathbf{E}_1, \mathbf{D} \rangle$ and $B = \langle \mathbf{E}_2, \mathbf{D} \rangle$.

6.2.1 Writing The Solution Another Way

The coefficient matrix is

$$\mathbf{A} = \begin{bmatrix} -20 & 12 \\ 12 & 5 \end{bmatrix}$$

and the eigenvalues of \mathbf{A} are $\lambda_1 = 9.83$ and $\lambda_2 = -29.83$. Then we know if

$$\mathbf{P} = [\mathbf{E}_1 \quad \mathbf{E}_2]$$

then

$$\mathbf{P}^T \mathbf{A} \mathbf{P} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \implies \mathbf{A} = \mathbf{P} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{P}^T$$

Let

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \implies \mathbf{A} = \mathbf{P} \boldsymbol{\Lambda} \mathbf{P}^T$$

Now let's calculate the powers of \mathbf{A} :

•

$$\mathbf{A}^2 = (\mathbf{P} \boldsymbol{\Lambda} \mathbf{P}^T) (\mathbf{P} \boldsymbol{\Lambda} \mathbf{P}^T) = \mathbf{P} \boldsymbol{\Lambda} (\mathbf{P}^T \mathbf{P}) \boldsymbol{\Lambda} \mathbf{P}^T$$

100 CHAPTER 6. MATRIX SEQUENCES AND ORDINARY DIFFERENTIAL EQUATIONS

$$\begin{aligned} A^3 &= P\Lambda^2P^T \\ &= (P\Lambda P^T)(A^2) = (P\Lambda P^T)(P\Lambda^2P^T) = P\Lambda(P^TP)\Lambda^2P^T \\ &= P\Lambda^3P^T \end{aligned}$$

- It is easy to see by POMI that $A^n = P\Lambda^n P^T$.

Recall for the scalar t , $At = tA$ simply multiplies each entry of A by the number t . Hence, from the above we have $A^n t = P\Lambda^n P^T t$. Consider

$$At = P \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} P^T t = P \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} E_{11} t & E_{12} t \\ E_{21} t & E_{22} t \end{bmatrix} = P \begin{bmatrix} \lambda_1 t & 0 \\ 0 & \lambda_2 t \end{bmatrix} P^T$$

Thus,

$$At = P \begin{bmatrix} \lambda_1 t & 0 \\ 0 & \lambda_2 t \end{bmatrix} \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$$

A similar calculation shows

$$A^2 t^2 = P \begin{bmatrix} \lambda_1^2 t^2 & 0 \\ 0 & \lambda_2^2 t^2 \end{bmatrix} \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} = P \begin{bmatrix} \lambda_1^2 t^2 & 0 \\ 0 & \lambda_2^2 t^2 \end{bmatrix} P^T$$

And the function

$$\begin{aligned} W_3(t) &= I + At + A^2 t^2 / 2! + A^3 t^3 / 3! \\ &= P \begin{bmatrix} 1 + \lambda_1 t + \lambda_1^2 t^2 / 2 + \lambda_1^3 t^3 / 6 & 0 \\ 0 & 1 + \lambda_2 t + \lambda_2^2 t^2 / 2 + \lambda_2^3 t^3 / 6 \end{bmatrix} P^T \\ &= P \begin{bmatrix} \sum_{k=0}^3 (\lambda_1^k t^k) / k! & 0 \\ 0 & \sum_{k=0}^3 (\lambda_2^k t^k) / k! \end{bmatrix} P^T \end{aligned}$$

In general

$$W_n(t) = P \begin{bmatrix} \sum_{k=0}^n (\lambda_1^k t^k) / k! & 0 \\ 0 & \sum_{k=0}^n (\lambda_2^k t^k) / k! \end{bmatrix} P^T$$

We know $e^{\lambda_1 t} = \lim_{n \rightarrow \infty} \sum_{k=0}^n (\lambda_1^k t^k) / k!$ and $e^{\lambda_2 t} = \lim_{n \rightarrow \infty} \sum_{k=0}^n (\lambda_2^k t^k) / k!$ This suggests

$$\begin{aligned} \lim_{n \rightarrow \infty} W_n(t) &= P \begin{bmatrix} \lim_{n \rightarrow \infty} \sum_{k=0}^n (\lambda_1^k t^k) / k! & 0 \\ 0 & \lim_{n \rightarrow \infty} \sum_{k=0}^n (\lambda_2^k t^k) / k! \end{bmatrix} P^T \\ &= P \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} P^T \end{aligned}$$

Define the matrix

$$e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix}$$

The system of ODEs $\mathbf{X}' = A\mathbf{X}$ can be written as $\mathbf{X}' = P\Lambda P^T \mathbf{X}$. Let $\mathbf{Y} = P^T \mathbf{X}$. Then the system becomes $\mathbf{Y}' = \Lambda \mathbf{Y}$, with general solution

$$\mathbf{Y}(t) = \begin{bmatrix} \alpha e^{\lambda_1 t} & 0 \\ 0 & \beta e^{\lambda_2 t} \end{bmatrix} = \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$$

for arbitrary α and β . Define

$$\Theta = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$$

Then the general solution is $\mathbf{Y}(t) = \mathbf{P}^T \mathbf{X} = e^{\Lambda t} \Theta$. This gives $\mathbf{X}(t) = \mathbf{P} e^{\Lambda t} \Theta$. Now define the matrix $e^{\mathbf{A}t} = \mathbf{P} e^{\Lambda t} \mathbf{P}^T$. Then $e^{\mathbf{A}t} \mathbf{P} = \mathbf{P} e^{\Lambda t}$ and so $\mathbf{X}(t) = e^{\mathbf{A}t} \mathbf{P} \Theta$. Note

$$\mathbf{P} \Theta = [\mathbf{E}_1 \quad \mathbf{E}_2] \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} = [\alpha \mathbf{E}_1 \quad \beta \mathbf{E}_2]$$

So for initial data, $\mathbf{X}_0 = [x_0^1 \quad x_0^2]^T$, we have

$$\begin{bmatrix} x_0^1 \\ x_0^2 \end{bmatrix} = [\alpha \mathbf{E}_1 \quad \beta \mathbf{E}_2]$$

implying $\alpha = \langle \mathbf{X}_0, \mathbf{E}_1 \rangle = x_0^1$ and $\beta = \langle \mathbf{X}_0, \mathbf{E}_2 \rangle = x_0^2$. Thus, the solution to the initial value problem is $\mathbf{X}(t) = e^{\mathbf{A}t} \mathbf{X}_0$ and the general solution to the dynamics has the form $\mathbf{X}(t) = e^{\mathbf{A}t} \mathbf{C}$ where \mathbf{C} is an arbitrary vector.

The general solution to the scalar ODE $x' = \lambda x$ is $x(t) = e^{\lambda t} c$ and we now know we can write the general solution to the vector system $\mathbf{X}' = \mathbf{A}\mathbf{X}$ is $\mathbf{X}(t) = e^{\mathbf{A}t} \mathbf{C}$ also as long as we interpret the exponential matrix $e^{\mathbf{A}t}$ right. Our argument was for 2×2 symmetric matrices, but essentially the same argument is used in the general $n \times n$ case but the canonical form of \mathbf{A} we need is called the **Jordan Canonical Form** which is discussed in more advanced classes.

6.2.2 Homework

Exercise 6.2.1 Prove that $\mathbf{A}^n = \mathbf{P} \Lambda^n \mathbf{P}^T$ by POMI.

Exercise 6.2.2 Show via POMI

$$\mathbf{A}^n t^n = \mathbf{P} \begin{bmatrix} \lambda_1^n t^n & 0 \\ 0 & \lambda_2^n t^n \end{bmatrix} \mathbf{P}^T$$

Exercise 6.2.3 Calculate $e^{\mathbf{A}t}$ for

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix}$$

Chapter 7

Continuity and Topology

We went over some ideas about topology in \mathbb{R}^2 in Section 2.2 but we need to do this in \mathbb{R}^n now.

7.1 Topology in n Dimensions

Most of this should be familiar although it is set in \mathbb{R}^n .

Definition 7.1.1 Balls in \mathbb{R}^n

- The ball about p in \mathbb{R}^n of radius $r > 0$ is

$$B(p; r) = \{x \in \mathbb{R}^n \mid \|x - p\| < r\}$$

where $\|\cdot\|$ is any norm on \mathbb{R}^n . This does not have to be the usual Euclidean norm.

- The closed ball about p in \mathbb{R}^n of radius $r > 0$ is

$$\overline{B(p; r)} = \{x \in \mathbb{R}^n \mid \|x - p\| \leq r\}$$

- The punctured ball about p in \mathbb{R}^n of radius $r > 0$ is

$$\hat{B}(p; r) = \{x \in \mathbb{R}^n \mid 0 < \|x - p\| \leq r\}$$

We can draw these sets in \mathbb{R} , \mathbb{R}^2 and \mathbb{R}^3 but, of course, we can not do that for $n > 3$. Hence, we **must** begin to think about these ideas **abstractly**. There is also the idea of a boundary point.

Definition 7.1.2 Boundary Points of a Subset in \mathbb{R}^n

Let D be a subset of \mathbb{R}^n . We say p is a **boundary point** of D if for all positive radii r , $B(p; r)$ contains a point of D and a point of its complement D^C . The set of all boundary points of D is denoted by ∂D .

Comment 7.1.1 If $D = \{x_1, x_2\}$, then D consists of two vectors only. It is clear that every ball of radius r smaller about x_1 that $\|x_1, x_2\|$ contains only x_1 from D . We can say the same for x_2 . So each point in D is a boundary point. In fact, we would call these points **isolated** points as there is a ball about each which contains only them.

Next, we define **open** and **closed** subsets.

Definition 7.1.3 Open and Closed Sets in \mathbb{R}^n

Let D be a subset of \mathbb{R}^n . If p is in D and there is a positive r so that $B(p; r) \subset D$, we say p is an **interior point** of D .

- If all p in D are interior points, we say D is an **open subset** of \mathbb{R}^n .
- The set of points in \mathbb{R}^n not in D is called the **complement** of D . We say D is **closed** if its complement is open. We denote the complement of D by D^C .

Note the complement of a closed set is open.

Now, as usual, we can talk about the convergence of sequences $\{x_n\}$ of points in \mathbb{R}^n .

Definition 7.1.4 Norm Convergence in n Dimensions

We say the sequence $\{x_n\}$ in \mathbb{R}^n converges in norm $\|\cdot\|$ to x in \mathbb{R}^n if given $\epsilon > 0$, there is a positive integer N so that

$$n > N \implies \|x_n - x\| < \epsilon$$

We usually just write $x_n \rightarrow x$.

Comment 7.1.2 All of the usual things about scalar convergence of sequences still hold for norm convergence in \mathbb{R}^n but we will not clutter up the narrative by going through these proofs.

- If a sequence converges all of its subsequences converge to the same limit.
- the usual algebra of limits theorem except no products and quotients.
- convergence sequences are bounded

There are more and we will leave them as exercises for you.

Homework:

Exercise 7.1.1 Prove if a sequence converges all of its subsequences converge to the same limit.

Exercise 7.1.2 Prove the usual algebra of limits theorem. Note we do not have products and quotients though.

Exercise 7.1.3 Prove convergent sequences are bounded.

There are then special points in a set D which can be accessed via sequences of points in D . We need to define several possibilities.

Definition 7.1.5 Limit Points, Cluster Points and Accumulation Points

Let D be a subset of \mathbb{R}^n .

- We say p is a **limit point** of D if there is a sequence $\{x_n\}$ in D so that $[x_n \rightarrow p]$. Note p need not be in D .
- We say p is a **cluster point** of D if there is a sequence $\{x_n \neq p\}$ in D so that $[x_n \rightarrow p]$.
- We say p is an **accumulation point** of D , for any positive radius r , $\hat{B}(p; r)$ contains points y in D . Note this clearly implies there is a sequence $\{x_n \neq p\}$ in D so that $x_n \rightarrow p$.

Comment 7.1.3 If p is an isolated point of D , then the constant sequence $\{x_n = p\}$ in D converges to p and so p is a limit point of D . However, an isolated point is **not** an accumulation point or cluster point.

We can prove a nice theorem about this now.

Theorem 7.1.1 D in \mathbb{R}^n is closed if and only if it contains all its limit points.

Let D be a subset of \mathbb{R}^n . Let D' be the set of limit points of D . Then D is closed $\iff D' \subset D$.

Proof 7.1.1

(\implies):

Assume D is a closed set. Let p be a limit point of D which is not in D . Then there is a sequence $\{x_n\}$ in D with $x_n \rightarrow p$. Since p is not in D , it is in D^C which is open. Thus p is an interior point of D^C and there is a $r > 0$ so $B(r, p) \subset D^C$. However, for $\epsilon = r/2$, there is an N so that $n > N \implies \|x_n - p\| < \epsilon = r/2$. Thus, for $n > N$, x_n is in D^C . This is a contradiction. Hence, p must be in D . This shows $D' \subset D$.

(\impliedby):

Assume D contains D' and consider D^C . If D^C is not open, then there is a p in D^C which is not an interior point of D^C . Hence for the sequence $\{r_n = 1/n\}$, there is a point x_n in $B(p, 1/n)$ with x_n in D . Clearly, $x_n \rightarrow p$ and hence p is a limit point of D and so must be in D by assumption. But this is not possible by our assumption. So our assumption must be wrong and D^C must be open. This tells us D is closed. \blacksquare

This leads to the definition of the **closure** of a set.

Definition 7.1.6 The Closure of a Set

The closure of a set D in \mathbb{R}^n is denoted by $\bar{D} = D \cup D'$.

We can then prove a useful characterization of the closure of a set.

Theorem 7.1.2 A D Set is Closure if and only if $D = \bar{D}$

The set D is closed if and only if $D = \bar{D}$.

Proof 7.1.2

(\implies):

If D is closed, $D' \subset D$ and so $D \cup D' = D$. So $\overline{D} = D$.

(\Leftarrow):

If $D = \overline{D} = D \cup D'$, Thus, D' is in D which tells us D is closed. ■

There is a lot more \mathbb{R}^n topology we could discuss but this is enough to get you started.

7.1.1 Homework

Exercise 7.1.4

Exercise 7.1.5

Exercise 7.1.6

Exercise 7.1.7

Exercise 7.1.8

7.2 Cauchy Sequences

We have seen Cauchy sequences frequently in (Peterson (8) 2019) and the same idea is easy to move to the \mathbb{R}^n setting.

Definition 7.2.1 Cauchy Sequences in \mathbb{R}^n

A sequence $\{\mathbf{x}_n\}$ in \mathbb{R}^n satisfies

$$\forall \epsilon > 0, \exists N \ni n, m > N \implies \|\mathbf{x}_n - \mathbf{x}_m\| < \epsilon$$

We know from the \mathbb{R} setting that Cauchy sequences of real numbers must converge to a real number due to the **Completeness Axiom**. We also know from discussions in (Peterson (8) 2019) that Cauchy Sequences in general normed spaces need not converge to an element in the space. Normed spaces in which Cauchy sequences converge to an element of the space are called **Complete Spaces**. We can easily prove \mathbb{R}^n is a complete space.

Theorem 7.2.1 \mathbb{R}^n is Complete

\mathbb{R}^n is a complete normed space.

Proof 7.2.1

Let $\{\mathbf{x}_n\}$ in \mathbb{R}^n be a Cauchy Sequence in \mathbb{R}^n . Pick $\epsilon > 0$. Then there is an N so that

$$n, m > N \implies \|\mathbf{x}_n - \mathbf{x}_m\| < \epsilon/2$$

or

$$n, m > N \implies \sum_{i=1}^n |x_{ni} - x_{mi}|^2 < \epsilon^2/4$$

Thus each piece of this sum satisfies

$$n, m > N \implies |x_{ni} - x_{mi}| < \epsilon/2$$

This says $\{x_{ni}\}$ is a Cauchy sequence in \mathbb{R} and so must converge to a number x_i . Let \mathbf{x} be the vector whose components are x_i . Then, since $|\cdot|$ is continuous, for $n, m > N$, we have

$$\lim_{m \rightarrow \infty} \left(\sum_{i=1}^n |x_{ni} - x_{mi}|^2 \right) \leq \epsilon^2/4$$

So

$$\|\mathbf{x}_n - \mathbf{x}\|^2 = \sum_{i=1}^n |x_{ni} - x_i|^2 \leq \epsilon^2/4$$

which implies $\|\mathbf{x}_n - \mathbf{x}\| < \epsilon$ when $n > N$. This shows \mathbb{R}^n is complete. ■

7.3 Compactness

Theorem 7.3.1 Bolzano - Weierstrass Theorem in \mathbb{R}^n

Every bounded sequence in \mathbb{R}^n has at least one convergent subsequence.

Proof 7.3.1

Let's assume the range of this sequence is infinite. If it were finite, there would be subsequences of it that converge to each of the values in the finite range. We assume the sequences start at $n = 1$ for convenience and by assumption, there is a positive number B so that $\|\mathbf{a}_n\| \leq B/2$ for all $n \geq 1$. Hence, all the components of this sequence live in what could be called a hyper rectangle $\mathcal{B} = [-B/2, B/2] \times \dots \times [-B/2, B/2] \subset \mathbb{R}^n$. We can denote this by $\mathcal{B} = \prod_{i=1}^n [-B/2, B/2]$. Let $J_0 = \prod_{i=1}^n I_0^i$ be the hyper rectangle $[\alpha_{01}, \beta_{01}] \times \dots \times [\alpha_{0n}, \beta_{0n}]$. Here, $\alpha_{0i} = -B/2$ and $\beta_{0i} = B/2$ for all i ; i.e. on each axis. The n dimensional area of J_0 , denoted by ℓ_0 is B^n .

Let \mathcal{S} be the range of the sequence which has infinitely many points and for convenience, we will let the phrase infinitely many points be abbreviated to IMPs.

Step 1:

Bisect each axis interval of J_0 into two pieces giving 2^n subregions of J_0 all of which have area $B^2/4$. Now at least one of the subregions contains IMPs of \mathcal{S} as otherwise each subregion has only finitely many points and that contradicts our assumption that \mathcal{S} has IMPs. Now all may contain IMPs so select one such subregion containing IMPs and call it J_1 . Label the endpoints of the component intervals making up the cross product that define J_1 as $I_1^i = [\alpha_{1i}, \beta_{1i}]$. hence, $J_1 = \prod_{i=1}^n I_1^i$ with $\ell_1 = B^2/4$. We see $J_1 \subset J_0$ and on the j^{th} axis, the subinterval endpoints satisfy

$$-B/2 = \alpha_{0i} \leq \alpha_{1i} \leq \beta_{1i} \leq \beta_{0i} = B/2$$

Since J_1 contains IMPs, we can select a sequence vector \mathbf{a}_{n_1} from J_1 .

Step 2:

Now subdivide J_1 into 2^n subregions just as before. At least one of these subregions contain IMPs of \mathcal{S} .

Choose one such subregion and call it J_2 . Then J_2 is the cross product of the intervals $I_2^i = [\alpha_{2i}, \beta_{2i}]$. The area of this subregion is now $\ell_2 = B^2/16$. We see $J_2 \subseteq J_1 \subseteq J_0$ and

$$-B = \alpha_{0i} \leq \alpha_{1i} \leq \alpha_{2i} \leq \beta_{2i} \leq \beta_{1i} \leq \beta_{0i} = B$$

Since J_2 contains IMPs, we can select a sequence vector \mathbf{a}_{n_2} from J_2 . It is easy to see this value can be chosen different from \mathbf{a}_{n_1} , our previous choice.

You should be able to see that we can continue this argument using induction.

Proposition:

$\forall p \geq 1, \exists$ an interval $J_p = \prod_{i=1}^n I_p^i$ with $I_p^i = [\alpha_{pi}, \beta_{pi}]$ with the area of J_p , $\ell_p = B^2/(2^{2p})$ satisfying $J_p \subseteq J_{p-1}$, J_p contains IMPs of \mathcal{S} and

$$-B/2 \leq \alpha_{0i} \leq \dots \leq \alpha_{1,i} \leq \dots \leq \alpha_{p,i} \leq \beta_{pi} \leq \dots \leq \beta_{1i} \leq \beta_{0i} \leq B/2$$

Finally, there is a sequence vector \mathbf{a}_{n_p} in J_p , different from $\mathbf{a}_{n_1}, \dots, \mathbf{a}_{n_{p-1}}$.

Proof We have already established the proposition is true for the basis steps J_1 and J_2 .

Inductive: We assume the interval J_q exists with all the desired properties. Since by assumption, J_q contains IMPs, bisect J_q into 2^n subregions as we have done before. At least one of these subregions contains IMPs of \mathcal{S} . Choose one of the subregions and call it J_{q+1} and label $J_{q+1} = \prod_{i=1}^n I_{q+1}^i$ where $I_{q+1}^i = [\alpha_{q+1,i}, \beta_{q+1,i}]$ We see immediately $\ell_{q+1} = B^2/2^{2(q+1)}$ with

$$\alpha_{q,i} \leq \alpha_{q+1,i} \leq \beta_{q+1,i} \leq \beta_{qi}$$

This shows the nested inequality we want is satisfied. Finally, since J_{q+1} contains IMPs, we can choose $\mathbf{a}_{n_{q+1}}$ distinct from the other \mathbf{a}_{n_i} 's. So the inductive step is satisfied and by the POMI, the proposition is true for all n . \square

From our proposition, we have proven the existence of sequences on each axis: (α_{pi}) , (β_{pi}) and (ℓ_p) which have various properties. The sequence ℓ_p satisfies $\ell_p = (1/4)\ell_{p-1}$ for all $p \geq 1$. Since $\ell_0 = B^2$, this means $\ell_1 = B^2/4$, $\ell_2 = B^2/16$, $\ell_3 = B^2/(2^2)^3$ leading to $\ell_p = B^2/(2^2)^p$ for $p \geq 1$. Further, we have the inequality chain

$$\begin{aligned} -B/2 &= \alpha_{0i} \leq \alpha_{1i} \leq \alpha_{2i} \leq \dots \leq \alpha_{pi} \\ &\leq \dots \leq \\ \beta_{pi} &\leq \dots \leq \beta_{2i} \leq \dots \leq \beta_{0i} = B/2 \end{aligned}$$

The rest of this argument is very familiar. Note (α_{pi}) is bounded above by $B/2$ and (β_{pi}) is bounded below by $-B/2$. Hence, by the completeness axiom, $\inf(\beta_{pi})$ exists and equals the finite number β^i ; also $\sup(\alpha_{pi})$ exists and is the finite number α^i .

So if we fix p , it should be clear the number β_{pi} is an upper bound for all the α_{pi} values (look at our inequality chain again and think about this). Thus β_{pi} is an upper bound for (α_{pi}) and so by definition of a supremum, $\alpha^i \leq \beta_{pi}$ for all p . Of course, we also know since α^i is a supremum, that $\alpha_{pi} \leq \alpha^i$. Thus, $\alpha_{pi} \leq \alpha^i \leq \beta_{pi}$ for all p . A similar argument shows if we fix p , the number α_{pi} is an lower bound for all the β_{pi} values and so by definition of an infimum, $\alpha_{pi} \leq \beta^i \leq \beta_{pi}$ for all the α_{pi} values. This tells us α^i and β^i are in $[\alpha_{pi}, \beta_{pi}]$ for all p . Next we show $\alpha^i = \beta^i$.

Let $\epsilon > 0$ be arbitrary. Since α^i and β^i are in an interval whose length is $\ell_p = (1/2^{2p})B^2$, we have $|\alpha^i - \beta^i| \leq (1/2^{2p})B^2$. Pick P so that $1/(2^{2P}B^2) < \epsilon$. Then $|\alpha^i - \beta^i| < \epsilon$. But $\epsilon > 0$ is arbitrary. Hence, $\alpha^i - \beta^i = 0$ implying $\alpha^i = \beta^i$. Finally, define the vector

$$\boldsymbol{\theta} = \begin{bmatrix} \alpha^1 \\ \vdots \\ \alpha^n \end{bmatrix}$$

We now must show $a_{n_k} \rightarrow \theta$. This shows we have found a subsequence which converges to θ . We know $\alpha_{pi} \leq a_{n_p}^i \leq \beta_{pi}$ and $\alpha_{pi} \leq \alpha^i \leq \beta_p^i$ for all p . Pick $\epsilon > 0$ arbitrarily. Given any p , we have

$$\begin{aligned}|a_{n_p,1}^i - \alpha^i| &= |a_{n_p,i}^i - \alpha_{pi} + \alpha_{pi} - \alpha^i| \leq |a_{n_p,i}^i - \alpha_{pi}| + |\alpha_{pi} - \alpha^i| \\ &\leq |\beta_{pi} - \alpha_{pi}| + |\alpha_{pi} - \beta_{pi}| = 2|\beta_{pi} - \alpha_{pi}| = 2(1/2^{2p})B^2.\end{aligned}$$

Choose P so that $(1/2^{2P})B^2 < \epsilon/2$. Then, $p > P$ implies $|a_{n_p,1}^i - \alpha^i| < 2\epsilon/2 = \epsilon$. Thus, $a_{n_k,i} \rightarrow \alpha^i$. This shows the subsequence converges to θ . ■

Note this argument is messy but quite similar to the one and two dimensional case. It is more a problem of correct labeling than intellectual difficulty!

There are several notions of **compactness** for a subset of \mathbb{R}^n . There is **topological compactness**.

Definition 7.3.1 Topologically Compact Subsets of \mathbb{R}^n

Let D be a subset of \mathbb{R}^n . We say the collection of open sets $\{U_\alpha\}$ where $\alpha \in I$ and I is any index set, whether finite, countable or uncountable, is an open cover of D if $D \subset \cup_\alpha U_\alpha$. We say this collection has a finite subcover if there are a finite number of sets in the collection, $\{U_{\beta_1}, \dots, U_{\beta_p}\}$ so that $D \subset \cup_{i=1}^p U_{\beta_i}$. D is topologically compact if every open cover of D has a finite subcover.

and **sequential compactness**.

Definition 7.3.2 Sequentially Compact Subsets of \mathbb{R}^n

Let D be a subset of \mathbb{R}^n . We say D is sequentially compact if every sequence $\{x_n\}$ in D has at least one subsequence which converges to a point of D .

We need to prove their equivalence through a sequence of results. Compare how we prove these to their one dimensional analogues in (Peterson (8) 2019).

Theorem 7.3.2 D is Sequentially Compact if and only if it is closed and bounded

D in \mathbb{R}^n is sequentially compact if and only if it is bounded and closed.

Proof 7.3.3

(\Rightarrow)

We assume D is sequentially compact. Assume (x_n) in D converges to x . Since (x_n) is a sequence in the sequentially compact set D , it has a subsequence (x_n^1) which converges to y in D . Since limits of convergent sequences are unique, we must have $x = y$. So x is in D and hence D is closed.

To show D is bounded, we use contradiction. Assume it is not bounded. Then for each n , there is x_n in D so that $\|x_n\| > n$. But since D is sequentially compact there is a subsequence (x_{n_k}) which must converge to x in D . We know convergence sequences are bounded and so there is a positive number B so that $\|x_{n_k}\| < B$ for all elements of the subsequence. But eventually $\|x_{n_k}\| > n_k > B$ which is the contradiction. We conclude D must be bounded.

(\Leftarrow)

Now we assume D is closed and bounded. We argue that D is sequentially compact using essentially the same procedure we used to prove the Bolzano - Weierstrass Theorem for \mathbb{R}^n . First, if D is a finite set, it is sequentially compact. So we can assume D has infinitely points. Since D is bounded, there is a positive number B so that $D \subset \prod_{i=1}^n [-B/2, B/2]$. Pick any x_0 in D in $D_0 = \prod_{i=1}^n [-B/2, B/2]$. Now bisect each of the coordinate axis giving 2^n pieces. Since D is infinite, at least one piece

contains infinitely many points of D . Call this piece D_1 and pick $\mathbf{x}_1 \neq \mathbf{x}_0 \in D_1$. Since both points are in D_0 ,

$$\|\mathbf{x}_1 - \mathbf{x}_0\|^2 = \sum_{i=1}^n |x_{1i} - x_{0i}|^2 < nB^2 \implies \|\mathbf{x}_1 - \mathbf{x}_0\| < \sqrt{n}B$$

Now do this again. We bisect D_1 on each axis to create 2^n pieces. Choose one piece containing infinitely many points of D and call it D_2 . The lengths of the sides of this piece are now $B/4$. Pick \mathbf{x}_2 in D_2 not the same as the previous two. Then \mathbf{x}_2 and \mathbf{x}_1 are both in D_1 and so $\|\mathbf{x}_2 - \mathbf{x}_1\| < \sqrt{n}(B/2)$. Continuing this process, we obtain a sequence \mathbf{x}_n with $\mathbf{x}_n \in D_n$. The sides of D_n are length $B/2^n$ and both \mathbf{x}_{n-1} and \mathbf{x}_n are in D_{n-1} . So $\|\mathbf{x}_{n-1} - \mathbf{x}_n\| < \sqrt{n}(B/2^{n-1})$. We claim this sequence is a Cauchy Sequence in \mathbb{R}^n . For $k > m$,

$$\begin{aligned} \|\mathbf{x}_k - \mathbf{x}_m\| &\leq \sum_{j=m}^{k-1} \|\mathbf{x}_{j+1} - \mathbf{x}_j\| \leq \sum_{j=m}^{k-1} \sqrt{n} \frac{B}{2^j} \leq \sqrt{n}B \left(\sum_{j=0}^{k-1} \frac{1}{2^j} - \sum_{j=0}^{k-1} \frac{1}{2^j} \right) \\ &= \sqrt{n}B \left(\frac{1 - 1/2^k}{1 - 1/2} - \frac{1 - 1/2^m}{1 - 1/2} \right) \leq 2\sqrt{n}B \left(\frac{1}{2^m} - \frac{1}{2^k} \right) \leq \sqrt{n}B/2^{m-1} \end{aligned}$$

Then, given $\epsilon > 0$, there is an N so that $\sqrt{n}B/2^{m-1} < \epsilon$ if $m > N$. Thus, $\|\mathbf{x}_k - \mathbf{x}_m\| < \epsilon$ if $k > m > N$ which shows (\mathbf{x}_n) is a Cauchy Sequence. Since \mathbb{R}^n is complete, there is \mathbf{x} so that $\mathbf{x}_n \rightarrow \mathbf{x}$. Since D is closed, it follows \mathbf{x} is in D . Now this same argument applies to any infinite sequence (\mathbf{w}_n) in D . This sequence will have a subsequence (\mathbf{w}_n^1) which converges to a point $\mathbf{w} \in D$. Hence, D is sequentially compact. ■

Theorem 7.3.3 D is topologically compact implies it is closed and bounded

D in \mathbb{R}^n is topologically compact implies it is closed and bounded.

Proof 7.3.4

The collection $\mathcal{O} = \{B(r, \mathbf{x}) | \mathbf{x} \in D\}$ is an open cover of D . Since D is topologically compact, \mathcal{O} has a finite subcover $\mathcal{V} = \{B(r, \mathbf{x}_i) | 1 \leq i \leq N\}$ for some positive integer N . If $\mathbf{x} \in D$, there is an index i so that $\mathbf{x} \in B(r, \mathbf{x}_i)$. Thus $\|\mathbf{x} - \mathbf{x}_i\| < r$ which implies $\|\mathbf{x}\| \leq r + \|\mathbf{x}_i\|$. Let $B = \max_{1 \leq i \leq N} \|\mathbf{x}_i\|$. Then, $\|\mathbf{x}\| \leq r + B$ which shows D is bounded.

Let's show D^C is open. Let $\mathbf{x} \in D^C$. For any \mathbf{y} in D , let $d_{xy} = \|\mathbf{x} - \mathbf{y}\|$ which is not zero as \mathbf{x} is in the complement. The collection $\mathcal{O} = \{B((1/2)d_{xy}, \mathbf{y}) | \mathbf{y} \in D\}$ is an open cover of D and so there is a finite subcover $\mathcal{V} = \{B((1/2)d_{x,y_i}, \mathbf{x}_i) | 1 \leq i \leq N\}$. Let $W = \cap B((1/2)d_{x,y_i}, \mathbf{x})$. By construction, each $B((1/2)d_{x,y_i}, \mathbf{x})$ is disjoint from $B((1/2)d_{x,y_i}, \mathbf{y}_i)$. So if \mathbf{z} is in W , \mathbf{z} is not in all $B((1/2)d_{x,y_i}, \mathbf{y}_i)$. But since \mathcal{V} is an open cover for D , this says $\mathbf{z} \in D^C$. Hence, if $r = \min_{1 \leq i \leq N} (1/2)d_{x,y_i}$, then $B(r, \mathbf{x})$ is contained in D^C . This shows \mathbf{x} is an interior point of D^C . So D^C is open which implies D is closed. ■

Theorem 7.3.4 A Hyperrectangle is topologically compact

The hyperrectangle $D = \prod_{i=1}^n [a_i, b_i]$ is topologically compact when each segment $[a_i, b_i]$ is finite in length.

Proof 7.3.5

This proof is similar to how we proved the Bolzano Weierstrass Theorem (BWT). So look at the similarities! Assume this is false and let $J_0 = \prod_{i=1}^n [a_i, b_i]$. Label this interval as $J_0 = [\alpha_0, \beta_0]$; i.e.

$\alpha_0 = a$ and $\beta_0 = b$. The volume of this hyperrectangle is $\ell_0 = \prod_{i=1}^n (b_i - a_i)$. Assume there is an open cover \mathcal{U} which does not have a fsc.

Now divide J_0 into 2^n pieces by bisecting each segment $[a_i, b_i]$. At least one of these pieces can not be covered by a fsc as otherwise all of D has a fsc which we have assumed is not true. Call this piece J_1 and note the volume of J_1 is $\ell_1 = (1/2)\ell_0$. We can continue in this way by induction. Assume we have constructed the piece J_n in this fashion with volume $\ell_n = (1/2)\ell_{n-1} = \ell_0/2^n$ and there is no fsc of J_n . Now divide J_n into 2^n equal pieces as usual. At least one of them can not be covered by a fsc. Call this interval J_{n+1} which has volume $\ell_{n+1} = (1/2)\ell_n$ like required.

We can see each $J_{n+1} \subset J_n$ by the way we have chosen the pieces and their volumes satisfy $\ell_n \rightarrow 0$. Using the same sort of arguments that we used in the proof of the BWT for Sequences, we see there is a unique point z which lies in all the J_n ; i.e. $z \in \cap_{i=1}^{\infty} J_n$.

Since z is in J_1 , there is a \mathbb{O}_1 in \mathcal{U} with $z \in \mathbb{O}_1$ because \mathcal{U} covers D . Also, since \mathbb{O}_1 is open, there is a circle $B(r_1, z) \subset \mathbb{O}_1$. Now choose K so that $\ell_K < \ell_1$. Then, J_k is contained in \mathbb{O}_1 for all $k > K$. This says \mathbb{O}_1 is a fsc of J_k which contradicts our construction process. Hence, our assumption that there is an open cover with no fsc is wrong and we have D is topologically compact. ■

Theorem 7.3.5 Closed Subsets of Hyperrectangles are topologically compact

Let C be a closed subset of the hyperrectangle $D = \prod_{i=1}^n [a_i, b_i]$ when each segment $[a_i, b_i]$ is finite in length. Then D is topologically compact.

Proof 7.3.6

Let \mathcal{U} be an open cover of B . Then the collection of open sets $\mathcal{O} = \mathcal{U} \cup B^C$ is an open cover of D and hence has a fsc $\mathcal{V} = \{\mathbb{O}_1, \dots, \mathbb{O}_N, B^C\}$ for some N with each $\mathbb{O}_n \in \mathcal{U}$. Hence, $B \subset \cup_{i=1}^N \mathbb{O}_n$ and we have found a fsc of \mathcal{U} which covers B . This shows B is topologically compact. ■

Theorem 7.3.6 D is topologically compact if and only if it is closed and bounded

D in \mathbb{R}^n is topologically compact if and only if it is closed and bounded

Proof 7.3.7

(\Rightarrow)

If D is topologically compact, by Theorem 7.3.3, it is closed and bounded.

(\Leftarrow)

If D is bounded, D is inside a hyperrectangle with finite edge lengths. Since D is a closed subset of the hyperrectangle, D is topologically compact by Theorem 7.3.5. ■

Theorem 7.3.7 D is topologically compact if and only if sequentially compact if and only if closed and bounded

Let D be a subset of \mathbb{R}^n . Then D is topologically compact $\iff D$ is sequentially compact
 $\iff D$ is closed and bounded.

Proof 7.3.8

(Topologically Compact \iff Closed and Bounded)

This is Theorem 7.3.6.

(*Sequentially Compact* \iff *Closed and Bounded*)

This is Theorem 7.3.2.

Comment 7.3.1 Since in these finite dimensional settings, the notions of **sequential compactness**, **topological compactness** and **closed and bounded** are equivalent, we often just say a set is **compact** and then use whichever characterization is useful for our purposes.

7.4 Functions of Many Variables

In (Peterson (8) 2019), we developed a number of simple MatLab scripts and functions to help us visualize functions of two variables. Of course, these are of little use for functions of three or more variables. For example, if $y = f(x)$ we know how to see this graphically. We graph the pairs $(x, f(x))$ in the usual xy plane. If we had $z = f(x, y)$, we would graph the triples $(x, y, f(x, y))$ in three space. We would use the right hand rule for the positioning of the $+z$ axis like usual. However, for $w = f(x, y, z)$, the visualization would require us to draw in \mathbb{R}^4 which we can not do. So much of what we want to do now must be done abstractly. Now let's look at this in the n dimensional setting. Let D be a subset of \mathbb{R}^n . and assume $z = f(\mathbf{x}) \in \mathbb{R}^m$ is defined locally on $B(\mathbf{x}_0, r)$ for some positive r . This defines m **component** functions f_1, \dots, f_m by

$$f(\mathbf{x}) = f(x_1, \dots, x_n) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))$$

The set $D \subset \mathbb{R}^n$ is at least an open set and we usually want it to be **path connected** as well. We will talk about this later, but a path connected subset of \mathbb{R}^n is one in which there is a **path** connecting any two points in D that lies entirely in D . We discuss this much more carefully in Chapter 17 but for now we will be casual about it.

Example 7.4.1 Although \mathbf{x} here is in \mathbb{R}^n and so is usually thought of as a column vector, we typically write it as a row vector. Here is a typical function $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$;

$$\begin{aligned} f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = f(x_1, x_2, x_3) &= \begin{bmatrix} x_1^2 + 3x_3^2 \\ 2x_1 + 5x_2 - \sin(x_3^2) \\ 14x_2^4 x_3^2 + 20 \end{bmatrix} \\ &= [x_1^2 + 3x_3^2, 2x_1 + 5x_2 - \sin(x_3^2), 14x_2^4 x_3^2 + 20] \end{aligned}$$

These notational abuses are quite common. For example, it is very normal to use row vector notation for the domain of f and column vector notation for the range of f . Go figure. For this example, the component functions are

$$\begin{aligned} f_1(x_1, x_2, x_3) &= x_1^2 + 3x_3^2, & f_2(x_1, x_2, x_3) &= 2x_1 + 5x_2 - \sin(x_3^2) \\ f_3(x_1, x_2, x_3) &= 14x_2^4 x_3^2 + 20 \end{aligned}$$

7.4.1 Limits and Continuity for Functions of Many Variables

Now let's look at continuity in this n dimensional setting. Let D be a subset of \mathbb{R}^n . and assume $z = f(\mathbf{x})$ is defined locally on $B(\mathbf{x}_0, r)$ for some positive r . Here is the two dimensional extension of the idea of a limit of a function.

Definition 7.4.1 The \mathbb{R}^n to \mathbb{R}^m Limit

If $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} f(\mathbf{x})$ exists, this means there is a vector $\mathbf{L} \in \mathbb{R}^m$ so that

$$\forall \epsilon > 0 \exists 0 < \delta < r \ni 0 < \|\mathbf{x} - \mathbf{x}_0\| < \delta \Rightarrow \|f(\mathbf{x}) - \mathbf{L}\| < \epsilon$$

We say $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} f(\mathbf{x}) = \mathbf{L}$. We use the same notation $\|\cdot\|$ for the norm in both the domain and the range. Of course, these norms are defined separately and we need to use the correct norms on both parts as needed.

We can now define new versions of limits and continuity for dimensional functions.

Definition 7.4.2 \mathbb{R}^n to \mathbb{R}^m Continuity

If $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} f(\mathbf{x})$ exists and matches $f(\mathbf{x}_0)$, we say f is continuous at \mathbf{x}_0 . That is

$$\begin{aligned} \forall \epsilon > 0 \exists 0 < \delta < r \ni 0 < \|\mathbf{x} - \mathbf{x}_0\| < \delta \\ \Rightarrow \|f(\mathbf{x}) - f(\mathbf{x}_0)\| < \epsilon \end{aligned}$$

We say $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0)$.

Statements about limits and continuity of these functions are equivalent to statements about limits and continuity of the component functions.

Theorem 7.4.1 The behavior of $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ is equivalent to the behavior of its component functions

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$. Then

- f has a limit at \mathbf{x}_0 of value \mathbf{L} if and only if $f_i(\mathbf{x})$ has limit L_i
- f is continuous at \mathbf{x}_0 if and only if $f_i(\mathbf{x})$ is continuous at \mathbf{x}_0 .

Proof 7.4.1

We will leave this proof to you. ■

Example 7.4.2 Prove

$$f(x, y) = \begin{bmatrix} 2x^2 + 3y^2 \\ 3x^2 + 5y^2 \end{bmatrix}$$

is continuous at $(2, 3)$.

Solution We find

$$f(2, 3) = \begin{bmatrix} 8 + 27 = 35 \\ 12 + 45 = 57 \end{bmatrix}$$

We will use the Euclidean norm on \mathbb{R}^2 here. Note, if we choose $r < 1$, then $\|(x, y) - (2, 3)\| < r$ implies $|x - 2| < r$ and $|y - 3| < r$. Then

$$\begin{aligned} \left\| \begin{bmatrix} 2x^2 + 3y^2 - 35 \\ 3x^2 + 5y^2 - 57 \end{bmatrix} \right\| &= \left\| \begin{bmatrix} 2(x-2+2)^2 + 3(y-3+3)^2 - 35 \\ 3(x-2+2)^2 + 5(y-3+3)^2 - 57 \end{bmatrix} \right\| \\ &= \left\| \begin{bmatrix} 2(x-2)^2 + 8(x-2) + 3(y-3)^2 + 6(y-3) \\ 3(x-2)^2 + 12(x-2) + 5(y-2)^2 + 30(y-3) \end{bmatrix} \right\| \end{aligned}$$

$$= \left\| \begin{bmatrix} 2(x-2)^2 \\ 3(x-2)^2 \\ 12(x-2) \\ 3(y-3)^2 \\ 5(y-2)^2 \\ 30(y-3) \end{bmatrix} \right\|$$

From the triangle inequality for $\|\cdot\|$, since this is the Euclidean norm, we have

$$\begin{aligned} \left\| \begin{bmatrix} 2x^2 + 3y^2 - 35 \\ 3x^2 + 5y^2 - 57 \end{bmatrix} \right\| &\leq 3\sqrt{2}|x-2|^2 + 12\sqrt{2}|x-2| + 5\sqrt{2}|y-3|^2 + 30\sqrt{2}|y-3| \\ &\leq 8\sqrt{2}r^2 + 42\sqrt{2}r < 50\sqrt{2}r \end{aligned}$$

Hence, given $\epsilon > 0$, if we choose $r < \min(1, \epsilon/(50\sqrt{2}))$ we have $\delta < r$ implies $\|f(x, y) - f(2, 3)\| < \epsilon$ when $(x, y) \in B((2, 3), r)$.

Note it is probably easier in practice to show each component is continuous separately. We get two different critical values of r and we just choose the minimum of them to prove continuity.

Example 7.4.3 Show

$$f(x, y) = \begin{cases} \begin{bmatrix} \frac{2x}{\sqrt{x^2+y^2}} \\ x^2 + 3y^2 \end{bmatrix}, & \text{if } (x, y) \neq (0, 0) \\ 0, & \text{if } (x, y) = (0, 0). \end{cases}$$

is not continuous at $(0, 0)$.

Solution First, although it is difficult to graph surfaces with discontinuities, it is possible. Let's modify the code **DrawMesh** in (Peterson (8) 2019). The function **DrawMesh** starts at the base point (x_0, y_0) and moves outward from there generating surface values that include the ones involving x_0 and/or y_0 . So to draw the surface all we have to do is to add a conditional tests in the setup phase for the surface values like this

Listing 7.1: **Adding Test to Avoid Discontinuity**

```
for i=-nx:nx
    if( i != 0)
        u = [x0+i*delx];
        x = [x,u];
    end
end
```

The test to avoid $i = 0$ is all we need. The full code is below:

Listing 7.2: **Draw a Surface with a Discontinuity**

```
function DrawDiscon(f, delx, nx, dely, ny, x0, y0)
% f is the function defining the surface
% delx is the size of the x step
% nx is the number of steps left and right from x0
% dely is the size of the y step
% ny is the number of steps left and right from y0
% (x0, y0) is the center of the rectangular domain
%
% set up x and y stuff
% want delx and dely to divide xf evenly
10
```

```

% we avoid x0 and y0 which is where the problem is
x = [];
for i=-nx:nx
    if( i != 0)
        15      u = [x0+i*delx];
        x = [x,u];
    end
end
%
20 y = [];
for i=-ny:ny
    if ( i != 0)
        25      u = [y0+i*dely];
        y = [y,u];
    end
sx = 2*nx;
sy = 2*ny;
hold on
30
% plot the surface
% set up grid of x and y pairs (x(i),y(j))
[X,Y] = meshgrid(x,y);
% set up the surface
35 Z = f(X,Y);
mesh(X,Y,Z);
xlabel('x');
ylabel('y');
zlabel('z');
40 rotate3d on
axis on
grid on
hold off
end

```

If you look at the generated surface yourself, you can spin the plot around and really get a good feel for the surface discontinuity. At $y = 0$, the surface value is $+2$ if $x > 0$ and -2 if $x < 0$ and you can see the attempt the plotted image makes to capture that sudden discontinuity in the strip of uncharted points at $y = 0$ on the plot. We captured one version of it in Figure 7.1.

We know the component f_2 is continuous everywhere as the argument is similar to the last example. So we only have to show f_1 does not have a limit at $(0, 0)$ as this is enough to show f_1 is not continuous at $(0, 0)$ and so f is not continuous at $(0, 0)$. If this limit exists, we should get the same value for the limit no matter what path we take to reach $(0, 0)$.

Let the first path be given by $x(t) = t$ and $y(t) = 2t$. We have two paths really; one for $t > 0$ and one for $t < 0$. We find for $t > 0$, $f_1(t, 2t) = 2t/\sqrt{t^2 + 4t^2} = 2t/(|t|\sqrt{5}) = -2/\sqrt{5}$ and hence the limit along this path $-2/\sqrt{5}$. Since the limiting value differs on two paths, the limit can't exist. Hence, f_1 is not continuous at $(0, 0)$.

Comment 7.4.1 The problem with visualization here is that we can never use this to help us with functions of more than two variables because we can't generate the plot. So we have to see it in our mind, so to speak.

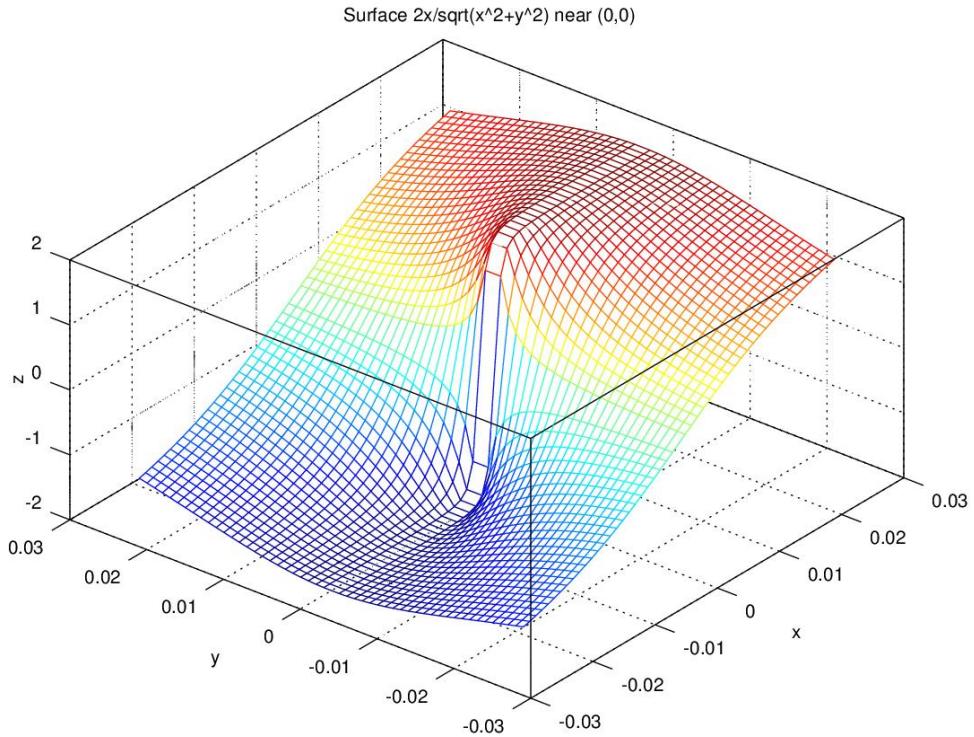


Figure 7.1: The surface $f(x, y) = 2x/\sqrt{x^2 + y^2}$ near the discontinuity at $(0, 0)$

7.4.2 Homework

Exercise 7.4.1 Let

$$f(x, y) = \begin{bmatrix} 6x^2 + 9y^2 \\ 3xy \end{bmatrix}$$

Prove f is continuous at $(2, 5)$.

Exercise 7.4.2 Let

$$f(x, y) = \begin{bmatrix} 2x^2 - 3y^2 \\ 7x + 8y \end{bmatrix}$$

Prove f is continuous at $(-1, 2)$.

Exercise 7.4.3 Let

$$f(x, y) = \begin{bmatrix} \frac{3x}{\sqrt{(x+1)^2 + 2(y-2)^2}} \\ 4x^2 + 2y^2 \end{bmatrix}$$

Prove f is not continuous at $(-1, 2)$. Draw the surface for f_1 here as well.

Theorem 7.4.2 If f is continuous on a compact domain, its range is compact

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be continuous. If D is compact, then $f(D)$ is also compact.

Proof 7.4.2

If $f(D)$ were not bounded, there would be a sequence (y_n) in $f(D)$ so that $|y_n| > n$ for all n . But since $y_n = f(\mathbf{x}_n)$ for some \mathbf{x}_n in D , this means there is a sequence (\mathbf{x}_n) in D which has a convergent subsequence (\mathbf{x}_{n_k}) which converges to some \mathbf{x} in D . Since convergent sequences are bounded, this means the sequence $(f(\mathbf{x}_{n_k}))$ must be bounded. Of course, it is not and this contradiction tells us our assumption $f(D)$ is not bounded is wrong. Hence $f(D)$ is bounded.

Next, let (y_n) be a convergent sequence in $f(D)$. So $y_n \rightarrow y$ for some y . There is then an associated sequence (\mathbf{x}_n) in D so that $y_n = f(\mathbf{x}_n)$. Since D is compact, there is a subsequence (\mathbf{x}_{n_k}) which converges to some \mathbf{x} in D . Since f is continuous on D , we have $f(\mathbf{x}_{n_k}) \rightarrow f(\mathbf{x}) \in f(D)$. But subsequences of convergent sequences converge to the same place, so we must have $y = f(\mathbf{x})$. Thus $y \in f(D)$ and $f(D)$ is closed.

Since $f(D)$ is closed and bounded, it is compact. ■

Theorem 7.4.3 If f continuous on a compact domain, it has global extrema

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be continuous on the compact set D . Then, f achieves a global maximum and global minimum on D .

Proof 7.4.3

Since $f(D)$ is compact, it is bounded. Thus there is a positive B so that $|f(\mathbf{x})| \leq B$. Hence, $\alpha = \inf_{\mathbf{x} \in D} f(\mathbf{x})$ and $\beta = \sup_{\mathbf{x} \in D} f(\mathbf{x})$ both exist and are finite by the Completeness Axiom. Using both the Infimum and Supremum Tolerance Lemma, we can find sequences (\mathbf{x}_n) and (\mathbf{y}_n) so that $f(\mathbf{x}_n) \rightarrow \alpha$ and $f(\mathbf{y}_n) \rightarrow \beta$. Since D is compact, there exist subsequences (\mathbf{x}_n^1) and (\mathbf{y}_n^1) with $\mathbf{x}_n^1 \rightarrow \mathbf{x}_m \in D$ and $\mathbf{y}_n^1 \rightarrow \mathbf{x}_M \in D$. Then $f(\mathbf{x}_n^1) \rightarrow \alpha$ and $f(\mathbf{y}_n^1) \rightarrow \beta$ too. By continuity of f , this tells us $f(\mathbf{x}_m) = \alpha$ and $f(\mathbf{x}_M) = \beta$. Thus α is the minimum of f on D and β is the maximum of f on D . ■

Chapter 8

Abstract Symmetric Matrices

Let's start with a few more ideas about matrices. We have already talked about magnitudes of linear transformations in Chapter 4.5 where the underlying spaces were n dimensional vector spaces. Now let's just look at these ideas for matrices whose underlying spaces are simply \Re^n .

8.1 Input - Output Ratios for Matrices

The **size** of a matrix can be measured in terms of its rows and columns, but a better way is to think about how it acts on vectors. We can think of a matrix as an **engine** which transforms inputs into outputs and it is a natural think to ask about the ratio of output to input. Such a ratio gives a measure of how the matrix alters the data. Since we can't divide by vectors, we typically use a measure of vector size for the output and input sides. Recall the **euclidean norm** of a vector x is $\|x\| = \sqrt{x_1^2 + \dots + x_n^2}$ where x_1 through x_n are the components of x in \Re^n with respect to the standard basis in \Re^n . This is also the $\|\cdot\|_2$ norm we have discussed before. Let A be a $n \times n$ matrix. Then A transforms vectors in \Re^n to new vectors in \Re^n . We measure the ratio of output to input by calculating $\|A(x)\|/\|x\|$ and, of course, this doesn't make sense if $x = 0$. We want to see how big this ratio can get, so we are interested in the maximum value of $\|A(x)\|/\|x\|$. Now the matrix A is linear; ie $A(cx) = cA(x)$. So in particular, if $y \neq 0$, we can say $A(y) = \|y\| A(y/\|y\|)$. Thus,

$$\frac{\|A(y)\|}{\|y\|} = \frac{\|y\| A\left(\frac{y}{\|y\|}\right)\|}{\|y\|} = \frac{\|A\left(\frac{y}{\|y\|}\right)\|}{\left\|\frac{y}{\|y\|}\right\|}$$

Now let $x = y/\|y\|$ and we have $\frac{\|A(y)\|}{\|y\|} = \frac{\|A(x)\|}{\|x\|}$ where x has norm 1. So

$$\max_{\|y\| \neq 0} \left(\frac{\|A(y)\|}{\|y\|} \right) = \max_{\|x\|=1} \left(\frac{\|A(x)\|}{\|x\|} \right)$$

In order to understand this, first note that the numerator $\|A(x)\| = \sqrt{\langle A(x), A(x) \rangle}$ and it is easy to see this is a continuous function of its arguments x_1 through x_n . We can't use just any arguments either; here we can only use those for which $x_1^2 + \dots + x_n^2 = 1$. The set of such x_i 's in \Re^n is a bounded set which contains all of its boundary points. For example, in \Re^2 , this set is the unit circle $x_1^2 + x_2^2 = 1$ and in \Re^3 , it is the surface of the sphere $x_1^2 + x_2^2 + x_3^2 = 1$. These sets are bounded and contain their boundary points and are therefore compact subsets. We know any continuous function

must have a global maximum and a global minimum over a compact domain. Hence, the problem of finding the maximum of $\|A(\mathbf{x})\|$ over the closed and bounded set $\|\mathbf{x}\| = 1$ has a solution. We define this maximum ratio to be the **norm** of the matrix A . We denote the matrix norm as usual by $\|A\|$ and so we have

Definition 8.1.1 The Norm of a Symmetric Matrix

*Let A be an $n \times n$ symmetric matrix. The **norm** of A is*

$$\|A\| = \max_{\|\mathbf{x}\|=1} \left(\frac{\|A(\mathbf{x})\|}{\|\mathbf{x}\|} \right)$$

8.1.1 Homework

Exercise 8.1.1

Exercise 8.1.2

Exercise 8.1.3

Exercise 8.1.4

Exercise 8.1.5

8.2 The Norm of a Symmetric Matrix

Now let's specialize to **symmetric** matrices. First, because of the symmetry, an easy calculation shows

$$\langle A(\mathbf{x}), \mathbf{y} \rangle = \mathbf{y}^T A \mathbf{x} = (A \mathbf{x})^T \mathbf{y} = \mathbf{x}^T A^T \mathbf{y}.$$

But this matrix is symmetric and so its transpose is the same as itself. We conclude

$$\langle A(\mathbf{x}), \mathbf{y} \rangle = \mathbf{x}^T A^T \mathbf{y} = \langle \mathbf{x}, A(\mathbf{y}) \rangle$$

We will use this identity in a bit. Now, when the matrix is symmetric, we claim we also have

$$\|A\| = \max_{\|\mathbf{x}\|=1} |\langle A(\mathbf{x}), \mathbf{x} \rangle|$$

Let the maximum on the right hand side be denoted by J for convenience. Now we know how inner products behave. We have for any vector \mathbf{y}

$$\langle A(\mathbf{y}), \mathbf{y} \rangle \leq \|A(\mathbf{y})\| \|\mathbf{y}\|$$

Next, from the way $\|A\|$ is defined, for any particular nonzero vector \mathbf{y} , we have $\|A(\mathbf{y})\| \leq \|A\| \|\mathbf{y}\|$. Thus, combining these ideas, we have for any vector \mathbf{x} with norm 1,

$$\langle A(\mathbf{x}), \mathbf{x} \rangle \leq \|A(\mathbf{x})\| \|\mathbf{x}\| \leq \|A\| \|\mathbf{x}\| \|\mathbf{x}\| = \|A\|.$$

8.2. THE NORM OF A SYMMETRIC MATRIX

121

Since this is true for all such vectors, we must have the maximum \mathbf{J} satisfies $\mathbf{J} \leq \|\mathbf{A}\|$. Next, we will show the reverse inequality and the two pieces together then show $\mathbf{J} = \|\mathbf{A}\|$.

Now for any nonzero \mathbf{y} we have

$$\langle \mathbf{A}\left(\mathbf{y}/\|\mathbf{y}\|\right), \left(\mathbf{y}/\|\mathbf{y}\|\right) \rangle \leq \mathbf{J}$$

which implies $\langle \mathbf{A}(\mathbf{y}), \mathbf{y} \rangle \leq \mathbf{J} / \|\mathbf{y}\|^2$. Now do the following calculations. These are just matrix multiplications and they are pretty straightforward, We have for any \mathbf{x} and \mathbf{y} that

$$\langle \mathbf{A}(\mathbf{x} + \mathbf{y}), (\mathbf{x} + \mathbf{y}) \rangle = \langle \mathbf{A}(\mathbf{x}), (\mathbf{x}) \rangle + \langle \mathbf{A}(\mathbf{y}), (\mathbf{y}) \rangle + 2 \langle \mathbf{A}(\mathbf{x}), (\mathbf{y}) \rangle \leq \mathbf{J} \|\mathbf{x} + \mathbf{y}\|^2$$

because the matrix is symmetric. We also have

$$\langle \mathbf{A}(\mathbf{x} - \mathbf{y}), (\mathbf{x} - \mathbf{y}) \rangle = \langle \mathbf{A}(\mathbf{x}), (\mathbf{x}) \rangle + \langle \mathbf{A}(\mathbf{y}), (\mathbf{y}) \rangle - 2 \langle \mathbf{A}(\mathbf{x}), (\mathbf{y}) \rangle \geq -\mathbf{J} \|\mathbf{x} - \mathbf{y}\|^2$$

Now subtract the second inequality from the first to get

$$4 \langle \mathbf{A}(\mathbf{x}), (\mathbf{y}) \rangle \leq \mathbf{J} \left(\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 \right)$$

Finally, we can calculate that

$$\left(\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 \right) = 2 \left(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 \right) = 4$$

as these vectors have length one. We conclude $\langle \mathbf{A}(\mathbf{x}), \mathbf{y} \rangle \leq \mathbf{J}$. Now consider the case where $\mathbf{y} = \mathbf{A}(\mathbf{x})$. If $\mathbf{A}(\mathbf{x}) = 0$, we would have

$$\|\mathbf{A}(\mathbf{x})\|^2 = \langle \mathbf{A}(\mathbf{x}), \mathbf{A}(\mathbf{x}) \rangle = \langle 0, 0 \rangle = 0,$$

and so $\|\mathbf{A}(\mathbf{x})\| \leq \mathbf{J}$. If $\mathbf{y} = \mathbf{A}(\mathbf{x}) \neq 0$, we can say

$$\langle \mathbf{A}(\mathbf{x}), \left(\mathbf{A}(\mathbf{x})/\|\mathbf{A}(\mathbf{x})\| \right) \rangle \leq \mathbf{J} \leq \mathbf{J}$$

which we can simplify to

$$\frac{\langle \mathbf{A}(\mathbf{x}), \mathbf{A}(\mathbf{x}) \rangle}{\|\mathbf{A}(\mathbf{x})\|} \leq \mathbf{J}.$$

But $\langle \mathbf{A}(\mathbf{x}), \mathbf{A}(\mathbf{x}) \rangle = \|\mathbf{A}(\mathbf{x})\|^2$ and so dividing, we find $\|\mathbf{A}(\mathbf{x})\| \leq \mathbf{J}$. This implies the maximum over all $\|\mathbf{x}\| = 1$, also satisfies this inequality and so

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}(\mathbf{x})\| \leq \mathbf{J}.$$

With the reverse inequality established, we have proven the result we want. Let's summarize our technical discussion. We have proven the following.

Theorem 8.2.1 Two Equivalent Ways To Calculate the Norm of a Symmetric Matrix

If \mathbf{A} is a symmetric $n \times n$ matrix, then

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}(\mathbf{x})\| = \max_{\|\mathbf{x}\|=1} |<\mathbf{A}(\mathbf{x}), \mathbf{x}>| = \mathbf{J}.$$

Proof 8.2.1

See the discussions above. ■

Homework

Exercise 8.2.1

Exercise 8.2.2

Exercise 8.2.3

Exercise 8.2.4

Exercise 8.2.5

8.2.1 Constructing Eigenvalues

For the symmetric matrix such, we can construct its eigenvalues using a procedure which also gives us useful estimates. Now let's find eigenvalues:

8.2.1.1 Eigenvalue One

Theorem 8.2.2 The First Eigenvalue

Either $\|\mathbf{A}\|$ or $-\|\mathbf{A}\|$ is an eigenvalue for \mathbf{A} . Letting the eigenvalue be λ_1 , then $|\lambda_1| = \|\mathbf{A}\|$ with an associated eigenvector of norm 1, \mathbf{E}_1 .

Proof 8.2.2

We know that the maximum value of $|<\mathbf{A}(\mathbf{x}), \mathbf{x}>|$ over all $\|\mathbf{x}\| = 1$ occurs at some unit vector. Call this unit vector \mathbf{E}_1 . For convenience, let $\alpha = \|\mathbf{A}\|$. Then we know $|<\mathbf{A}(\mathbf{E}_1), \mathbf{E}_1>| = \alpha$.

Case I: We assume $\lambda_1 = -\|\mathbf{A}\|$. Then,

$$\begin{aligned} <\mathbf{A}(\mathbf{E}_1) - (-\alpha)\mathbf{E}_1, \mathbf{A}(\mathbf{E}_1) - (-\alpha)\mathbf{E}_1> &= <\mathbf{A}(\mathbf{E}_1) + \alpha\mathbf{E}_1, \mathbf{A}(\mathbf{E}_1) + \alpha\mathbf{E}_1> \\ &= <\mathbf{A}(\mathbf{E}_1), \mathbf{A}(\mathbf{E}_1)> + 2\alpha <\mathbf{A}(\mathbf{E}_1), \mathbf{E}_1> \\ &\quad + \alpha^2 <\mathbf{E}_1, \mathbf{E}_1> \\ &= \|\mathbf{A}(\mathbf{E}_1)\|^2 + 2\alpha <\mathbf{A}(\mathbf{E}_1), \mathbf{E}_1> + \alpha^2 \\ &= <\mathbf{A}(\mathbf{E}_1), \mathbf{A}(\mathbf{E}_1)> - 2\alpha^2 + \alpha^2 \\ &= \|\mathbf{A}(\mathbf{E}_1)\|^2 - \alpha^2 \end{aligned}$$

since $\|\mathbf{E}_1\| = 1$. Then, overestimating, we have

$$<\mathbf{A}(\mathbf{E}_1) - (-\alpha)\mathbf{E}_1, \mathbf{A}(\mathbf{E}_1) - (-\alpha)\mathbf{E}_1> \leq \|\mathbf{A}\|^2 \|\mathbf{E}_1\|^2 - \alpha^2.$$

But $\alpha = \|\mathbf{A}\|$, so we have

$$\langle \mathbf{A}(\mathbf{E}_1)) - (-\alpha)\mathbf{E}_1, \mathbf{A}(\mathbf{E}_1)) - (-\alpha)\mathbf{E}_1 \rangle \leq \alpha^2 - \alpha^2 = 0.$$

We conclude $\|\mathbf{A}(\mathbf{E}_1)) - (-\alpha)\mathbf{E}_1\| = 0$ which tells us $\mathbf{A}(\mathbf{E}_1) = -\alpha\mathbf{E}_1$. Hence, $-\alpha$ is an eigenvalue of \mathbf{A} and we can pick \mathbf{E}_1 as the associated eigenvector of norm 1. **Case II:** We assume $\lambda_1 = \|\mathbf{A}\|$. The argument is then quite similar. We conclude $\|\mathbf{A}(\mathbf{E}_1) - \alpha\mathbf{E}_1\| = 0$ which tells us $\mathbf{A}(\mathbf{E}_1) = \alpha\mathbf{E}_1$. Hence, α is an eigenvalue of \mathbf{A} and we can pick \mathbf{E}_1 as the associated eigenvector of norm 1. ■

Homework

Exercise 8.2.6

Exercise 8.2.7

Exercise 8.2.8

Exercise 8.2.9

Exercise 8.2.10

8.2.1.2 Eigenvalue Two

We can now find the next eigenvalue using a similar argument.

Theorem 8.2.3 The Second Eigenvalue

There is a second eigenvalue λ_2 which satisfies $|\lambda_2| \leq |\lambda_1|$. Let the new symmetric matrix \mathbf{A}_2 be defined by

$$\mathbf{A}_2 = \mathbf{A} - \lambda_1 \mathbf{E}_1 \mathbf{E}_1^T,$$

Then, $|\lambda_2| = \|\mathbf{A}_2\|$ and the associated eigenvector \mathbf{E}_2 is orthogonal to \mathbf{E}_1 , i.e., $\langle \mathbf{E}_1, \mathbf{E}_2 \rangle = 0$.

Proof 8.2.3

The reasoning here is virtually identical to what we did for Theorem 8.2.2. First, note, if \mathbf{x} is a multiple of \mathbf{E}_1 , we have $\mathbf{x} = \mu\mathbf{E}_1$ for some μ giving

$$\begin{aligned} (\mathbf{A}_2(\mathbf{x})) &= \mathbf{A}(\mu\mathbf{E}_1) - \lambda_1\mu\mathbf{E}_1 \mathbf{E}_1^T \mathbf{E}_1 \\ &= \mu\lambda_1\mathbf{E}_1 - \mu\lambda_1\mathbf{E}_1 = 0. \end{aligned}$$

since \mathbf{E}_1 is an eigenvector with eigenvalue λ_1 . Hence, it is easy to see that

$$\max_{\|\mathbf{x}\|=1} |\langle \mathbf{A}_2(\mathbf{x}), \mathbf{x} \rangle| = \max_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbf{W}_1} |\langle \mathbf{A}_2(\mathbf{x}), \mathbf{x} \rangle|$$

where \mathbf{W}_1 is the set of vectors in that are orthogonal to eigenvector \mathbf{E}_1 . Since this is a smaller set of vectors, the maximum we obtain will, of course, be possibly smaller. Hence, we know $\max_{\|\mathbf{x}\|=1} |\langle \mathbf{A}_2(\mathbf{x}), \mathbf{x} \rangle| \geq \max_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbf{W}_1} |\langle \mathbf{A}_2(\mathbf{x}), \mathbf{x} \rangle|$.

The rest of the argument, applied to the new operator \mathbf{A}_2 is identical. Thus, we find

$$|\lambda_2| = \|\mathbf{A}_2\| = \max_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbf{W}_1} |<\mathbf{A}_2(\mathbf{x}), \mathbf{x}>|.$$

Further, $|\lambda_2| \leq |\lambda_1|$ and eigenvector \mathbf{E}_2 is orthogonal to \mathbf{E}_1 because it comes from \mathbf{W}_1 . ■

Homework

Exercise 8.2.11

Exercise 8.2.12

Exercise 8.2.13

Exercise 8.2.14

Exercise 8.2.15

8.2.1.3 Eigenvalue Three

We can now find the next eigenvalue.

Theorem 8.2.4 The Third Eigenvalue

There is a third eigenvalue λ_3 which satisfies $|\lambda_3| \leq |\lambda_2|$. Let the new symmetric matrix \mathbf{A}_3 be defined by

$$\mathbf{A}_3 = \mathbf{A} - \lambda_1 \mathbf{E}_1 \mathbf{E}_1^T - \lambda_2 \mathbf{E}_2 \mathbf{E}_2^T,$$

Then, $|\lambda_3| = \|\mathbf{A}_3\|$ and the associated eigenvector \mathbf{E}_3 is orthogonal to \mathbf{E}_1 and \mathbf{E}_2 .

Proof 8.2.4

The reasoning here is virtually identical what we did for the first two eigenvalues. First, note, if \mathbf{x} is in the plane determined by \mathbf{E}_1 and \mathbf{E}_2 , we have $\mathbf{x} = \mu_1 \mathbf{E}_1 + \mu_2 \mathbf{E}_2$ for some constants μ_1 and μ_2 giving

$$\begin{aligned} (\mathbf{A}_3(\mathbf{x})) &= \mathbf{A}(\mu_1 \mathbf{E}_1 + \mu_2 \mathbf{E}_2) - \lambda_1 \mu_1 \mathbf{E}_1 \mathbf{E}_1^T \mathbf{E}_1 - \lambda_2 \mu_2 \mathbf{E}_2 \mathbf{E}_2^T \mathbf{E}_2 \\ &= \mu_1 \lambda_1 \mathbf{E}_1 + \mu_2 \lambda_2 \mathbf{E}_2 - \mu_1 \lambda_1 \mathbf{E}_1 - \mu_2 \lambda_2 \mathbf{E}_2 = 0. \end{aligned}$$

since \mathbf{E}_1 and \mathbf{E}_2 are eigenvectors. Hence, it is easy to see that

$$\max_{\|\mathbf{x}\|=1} |<\mathbf{A}_3(\mathbf{x}), \mathbf{x}>| = \max_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbf{W}_2} |<\mathbf{A}_3(\mathbf{x}), \mathbf{x}>|$$

where \mathbf{W}_2 is the set of vectors in that are orthogonal to the plane determined by the first two eigenvectors. Since this is a smaller set of vectors, the maximum we obtain will, of course, be possibly smaller. Hence, we know $\max_{\|\mathbf{x}\|=1} |<\mathbf{A}_2(\mathbf{x}), \mathbf{x}>| \geq \max_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbf{W}_2} |<\mathbf{A}_3(\mathbf{x}), \mathbf{x}>|$.

The rest of the argument, applied to the new operator \mathbf{A}_3 is identical. Thus, we find

$$|\lambda_2| = \|\mathbf{A}_3\| = \max_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathbf{W}_2} |<\mathbf{A}_3(\mathbf{x}), \mathbf{x}>|.$$

8.3. WHAT DOES THIS MEAN?

125

Further, $|\lambda_3| \leq |\lambda_2|$ and eigenvector \mathbf{E}_3 is orthogonal to both \mathbf{E}_1 and \mathbf{E}_2 because it comes from W_2 . ■

Homework

Exercise 8.2.16**Exercise 8.2.17****Exercise 8.2.18****Exercise 8.2.19****Exercise 8.2.20**

Since \mathbf{A} is a $n \times n$ matrix, this process terminates after n steps. We can now state the full result.

Theorem 8.2.5 All Eigenvalues

There is a sequence of eigenvalues λ_i which satisfies $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$ with associated eigenvectors \mathbf{E}_i which form an orthonormal basis for \Re^n . Let the new matrix \mathbf{A}_i be defined by

$$\mathbf{A}_i = \mathbf{A} - \sum_{j=1}^i \lambda_j \mathbf{E}_j \mathbf{E}_j^T,$$

Then, $|\lambda_i| = \|\mathbf{A}_i\|$ and the associated eigenvector \mathbf{E}_i is orthogonal to \mathbf{E}_j for all $j \leq i$. Moreover, we can say

$$\mathbf{A} = \sum_{j=1}^n \lambda_j \mathbf{E}_j \mathbf{E}_j^T,$$

8.3 What Does This Mean?

Let's put all of this together. Our symmetric $n \times n$ matrix always has n real eigenvalues and their respective eigenvectors $\{\mathbf{E}_1, \dots, \mathbf{E}_n\}$ are mutually orthogonal and so define an orthonormal basis for \Re^n . Hence, we usually normalize them so they form an orthonormal basis for \Re^2 . We can use these eigenvectors to write \mathbf{A} in a canonical form just like we did for the 2×2 case. We define the two $n \times n$ matrices

$$\mathbf{P} = [\mathbf{E}_1 \ \mathbf{E}_2 \ \dots \ \mathbf{E}_n]$$

which has transpose

$$\mathbf{P}^T = \begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \\ \vdots \\ \mathbf{E}_n^T \end{bmatrix}$$

We know the eigenvectors are mutually orthogonal, so we must have $\mathbf{P}^T \mathbf{P} = \mathbf{I}$, $\mathbf{P} \mathbf{P}^T = \mathbf{I}$ and

$$\begin{aligned} \mathbf{P}^T \mathbf{A} \mathbf{P} &= \begin{bmatrix} \lambda_1 < \mathbf{E}_1, \mathbf{E}_1 > & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 < \mathbf{E}_2, \mathbf{E}_2 > & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \lambda_n < \mathbf{E}_n, \mathbf{E}_n > \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \lambda_n \end{bmatrix} \end{aligned}$$

which can be rewritten as

$$\mathbf{A} = \mathbf{P} \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \lambda_n \end{bmatrix} \mathbf{P}^T$$

And we have a nice representation for an arbitrary $n \times n$ symmetric matrix \mathbf{A} . Now at this point, we know all the eigenvalues are real and we can write them in descending order, but eigenvalues can be repeated and can even be zero. Also, we can represent the norm of the symmetric matrix \mathbf{A} in terms of its eigenvalues.

Theorem 8.3.1 The Eigenvalue Representation of the norm of a symmetric matrix

Let \mathbf{A} be an $n \times n$ symmetric matrix with eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$. Then

$$\|\mathbf{A}\| = |\lambda_1| = \max_{1 \leq i \leq n} |\lambda_i|$$

This also implies

$$\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\| \|\mathbf{x}\|$$

for all \mathbf{x} .

Proof 8.3.1

We know $\|\mathbf{A}\| \geq |<\mathbf{A}\mathbf{E}_1, \mathbf{E}_1>|$ where \mathbf{E}_1 is the unit norm eigenvector corresponding to eigenvalue λ_1 . Thus,

$$\|\mathbf{A}\| \geq |<\lambda_1 \mathbf{E}_1, \mathbf{E}_1>| = |\lambda_1|$$

Any \mathbf{x} has representation $\sum_{i=1}^n c_i \mathbf{E}_i$ in terms on the orthonormal basis of eigenvectors of \mathbf{A} . If $\|\mathbf{x}\| = 1$, this implies $\sum_{i=1}^n c_i^2 = 1$. Then

$$|<\mathbf{A}\mathbf{x}, \mathbf{x}>| = \left| \sum_{i=1}^n \sum_{j=1}^n c_i c_j \lambda_i <\mathbf{E}_i, \mathbf{E}_j> \right|$$

8.3. WHAT DOES THIS MEAN?

127

$$\leq \sum_{i=1}^n c_i^2 |\lambda_i| \leq (1) |\lambda_1|$$

Thus, $\|\mathbf{A}\| = \max \|\mathbf{x}\| = 1 |<\mathbf{Ax}, \mathbf{x}>| \leq |\lambda_1|$. Combining, we see $\|\mathbf{A}\| = \lambda_1$. ■

8.3.1 A Worked Out Example

Let \mathbf{A} be defined to be

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}$$

Then to find the first eigenvalue and eigenvector, we need to find a certain maximum. We let

$$f(x, y) = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x^2 + 6xy + 2y^2$$

and we want to find the maximum of $|f(x, y)|$ subject to $x^2 + y^2 = 1$. To do this, let $x = \cos(\theta)$ and $y = \sin(\theta)$. Then our constrained optimization problem becomes maximize $g(\theta)$ where

$$\begin{aligned} g(\theta) &= \cos^2(\theta) + 6\cos(\theta)\sin(\theta) + 2\sin^2(\theta) \\ &= 1 + 6\cos(\theta)\sin(\theta) + \sin^2(\theta) \\ &= 1 + 3\sin(2\theta) + (1/2)(1 - \cos(2\theta)) \\ &= (3/2) + 3\sin(2\theta) - (1/2)\cos(2\theta). \end{aligned}$$

over all $\theta \in [0, 2\pi]$. The critical points here at the two endpoint values $\theta = 0$ and $\theta = 2\pi$ and where $g'(\theta) = 0$. We have

$$g'(\theta) = 6\cos(2\theta) + \sin(2\theta).$$

Hence, the derivative is zero when $6\cos(2\theta) + \sin(2\theta) = 0$ or $\tan(2\theta) = -6$. Thus, $2\theta = 1.7359, 4.8776$ or 8.0192 which implies $\theta = 0.8680, 2.4388$ or 4.0096 . This is a quadratic expression, so there is no need to look at second order tests. We'll just evaluate the function at all the critical points. We have $g(0) = 1, g(2\pi) = 1, g(0.8680) = 4.5414$ and $g(2.4388) = 1.5414$. The maximum occurs at $\theta_1 = 0.8680$ and has value 4.5414. This is our eigenvalue λ_1 . The maximum occurs at

$$\mathbf{E}_1 = \begin{bmatrix} \cos(0.8680) \\ \sin(0.8680) \end{bmatrix} = \begin{bmatrix} 0.64635 \\ 0.76304 \end{bmatrix}$$

To find the second eigenvalue, we construct the new function $h(x, y)$ defined by

$$h(x, y) = \begin{bmatrix} x \\ y \end{bmatrix}^T \left(\begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} - \lambda_1 \mathbf{E}_1 \mathbf{E}_1^T \right) \begin{bmatrix} x \\ y \end{bmatrix}$$

and maximize its absolute value over the vectors of length one. From our theoretical discussion, the extreme values here are found by maximizing $f(x, y)$ over the vectors perpendicular to \mathbf{E}_1 . These are the multiples of \mathbf{F} given by

$$\mathbf{F} = \begin{bmatrix} \sin(\theta_1) \\ -\cos(\theta_1) \end{bmatrix}$$

Hence, the extreme values occurs at $\pm \mathbf{F}$. The value of $f(x, y)$ at $-\mathbf{F}$ is there is -1.5414 which in absolute value is 1.5414 and so we choose $\mathbf{E}_2 = \mathbf{F}$. These eigenvalues and their associated eigenvectors are exactly the same as the one we would find using the characteristic equation to find the eigenvalues and then solving the resulting eigenvector equations for a unit vector solution.

8.3.2 Homework

Exercise 8.3.1

Exercise 8.3.2

Exercise 8.3.3

Exercise 8.3.4

Exercise 8.3.5

8.4 Signed Definite Matrices Again

An $n \times n$ matrix is said to be a **positive definite** matrix if $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ for all vectors \mathbf{x} . We can't do our usual algebraic tricks now (thank goodness!) so let's sneak up on this a different way. Rewrite \mathbf{A} using our representation to get $\mathbf{x}^T \mathbf{P} \mathbf{D} \mathbf{P}^T \mathbf{x} > 0$ where \mathbf{D} is the diagonal matrix having the eigenvalues λ_i along the main diagonal. Next, let $\mathbf{y} = \mathbf{P}^T \mathbf{x}$. Then the inequality can be rewritten as

$$(\mathbf{P}^T \mathbf{x})^T \mathbf{D} \mathbf{P}^T \mathbf{x} = \mathbf{y}^T \mathbf{D} \mathbf{y} > 0.$$

We can easily multiply this out to get $\lambda_1 y_1^2 + \lambda_2 y_2^2 + \dots + \lambda_n y_n^2 > 0$. Since this is positive for all choices of y_i , set all but y_1 to zero. Then, we have $\lambda_1 y_1^2 > 0$ which implies $\lambda_1 > 0$. Similarly, setting all but y_2 to zero, we find $\lambda_2 > 0$. We can do this for all of the eigenvalues to conclude that \mathbf{A} positive definite forces all the eigenvalues to be positive. Going the other direction, if all the eigenvalues were positive, that would force the matrix to be positive definite.

A similar argument shows that \mathbf{A} is negative definite if and only if its eigenvalues are all negative. Note all, if we simply wanted $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$, we would say \mathbf{A} is positive semidefinite and our arguments would say this happens if and only if all the eigenvalues are nonnegative. And finally, if $\mathbf{x}^T \mathbf{A} \mathbf{x} \leq 0$, the matrix is negative semidefinite and all its eigenvalues are non positive.

8.4.1 Homework

Exercise 8.4.1

Exercise 8.4.2

Exercise 8.4.3

Exercise 8.4.4

Exercise 8.4.5

Chapter 9

Rotations and Orbital Mechanics

9.1 Introduction

It is easy to think that two dimensional vector spaces are somehow trivial and not worth a lot of our time. We think we can change your mind by looking at some problems in orbital mechanics. We are going to discuss how to write code to solve some problems in this field. This is a great example of how two dimensional vector spaces are used in practice. Our references for this material are

1. (Bate et al. (1) 1971): an old book on the basics behind interplanetary movement. It is a bit dated now.
2. (Thompson (15) 1961): a very nice book on the dynamics of space flight. It is old but most of this kind of information was worked out very early so don't be put off by the age of the text.
3. (Sutton and Biblarz (14) 2001): this is a standard senior level text in aerospace engineering undergraduate degree programs such as they have at the University of Illinois. It does not explain as much from the basics as (Bate et al. (1) 1971) and so it is a bit hard to read when you are starting out. But it has really interesting stuff on rocket design which is dated of course. However, the basic ways to do rocket propulsion and the technology to implement it have not changed all that much.
4. (Roy (13) 1982): this is a book we used when we worked at Aerospace Corporation back in the day. It is a nice complement to (Bate et al. (1) 1971).
5. (Curtis (3) 2014): a much newer book which replaces (Bate et al. (1) 1971).

There are a number of three dimensional coordinate systems we need to pay attention.

- The **Heliocentric - Ecliptic** coordinate system. The origin here is the center of the sun. The earth rotates around the sun and its position at any time is a point on a plane. The basis vectors for this plane are denoted by \mathbf{X}_{he} and \mathbf{Y}_{he} with the \mathbf{X}_{he} direction drawn from the center of the earth to the center of the sun on the first day of Spring. This is called the **vernal equinox** direction. The \mathbf{Y}_{he} direction is then 90° rotated from \mathbf{X}_{he} in the direction of the rotation of the earth and, of course, \mathbf{Y}_{he} is in the orbital plane determined by the movement of the earth around the sun. The right hand rule then determines the \mathbf{Z}_{he} direction. Hence, we can write a vector \mathbf{P} in the heliocentric - Ecliptic system as

$$\mathbf{P} = x_{he}\mathbf{X}_{he} + y_{he}\mathbf{Y}_{he} + z_{he}\mathbf{Z}_{he}$$

You can see this coordinate system in Figure 9.1.

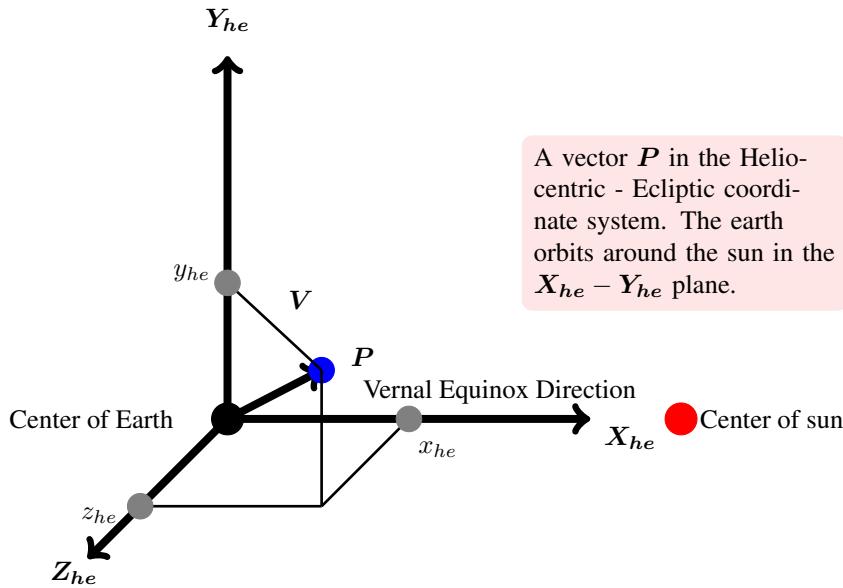


Figure 9.1: Heliocentric - Ecliptic coordinate system

- The **Geocentric - Equatorial** coordinate system. The origin is now at the center of the earth. The equatorial circle determines a plane and we choose the unit vector for the \mathbf{I} direction to be the vernal equinox direction we mentioned above. The \mathbf{J} unit vector is then 90° from \mathbf{I} and its direction is determined by the fact that the unit vector orthogonal to the equatorial plane is chosen to point toward the north pole. Hence, the right hand rule determines the orientation of \mathbf{J} . We see \mathbf{J} points east along the equator. A given vector \mathbf{P} can then be written as

$$\mathbf{P} = x_{ge}\mathbf{I} + y_{ge}\mathbf{J} + z_{ge}\mathbf{K}$$

You can see this coordinate system in Figure 9.2.

- The **Right Ascension-Declination** coordinate system. This is similar to a spherical coordinate system. For any point in three space, draw a line from the center of the earth to the point. Drop a perpendicular from that point to the equatorial plane or, if the point is below the equatorial plane, the perpendicular will point up until it hits the equatorial plane. This projection of the line connecting the center of the earth to the point will determine a line in the equatorial plane whose center is the center of the earth and whose terminal point is the point the vertical value of the point determines upon projection to the equatorial plane. The angle from the vernal equinox line which is the \mathbf{I} direction to this line is the angle α which is called the **right ascension**. Note this is exactly the same as the angle we call θ in the usual spherical coordinate system. We don't use the angle we call ϕ in the spherical coordinate system though. Recall ϕ is the angle from the positive z axis to the line connecting the center of the xyz coordinate system to the tip of the point \mathbf{P} determines in three dimensional space. There is another angle, which is measured up from the xy plane to this line which is $\pi/2 - \phi$. This angle is called the **declination** and is denoted by δ . To be clear, δ is measured from the line north to the \mathbf{K} axis. So $\delta = \pi$ would be a point on the line going to the south pole. The length of the line from the center of the earth to the point is given by R . Then a given vector \mathbf{P} can then be written as an ordered triple (R, α, δ) . Note we are not writing x in terms of new unit vectors specialized

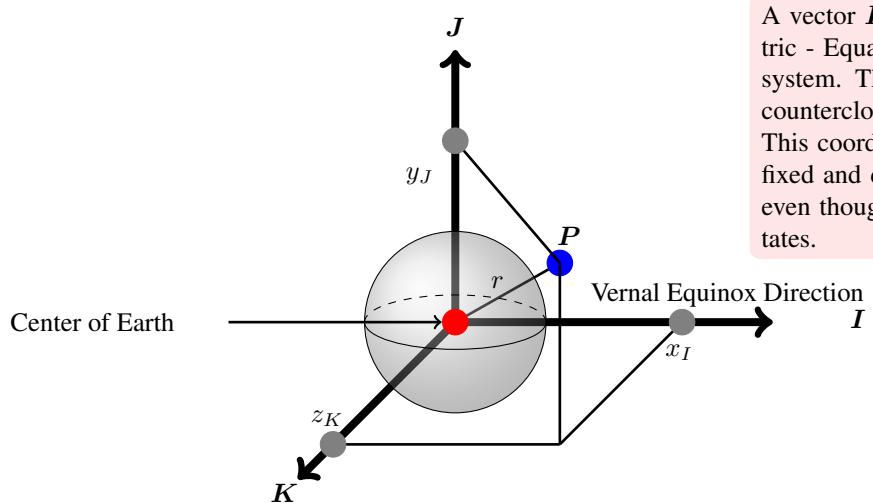


Figure 9.2: Geocentric - Equatorial coordinate system

to the right ascension - declination coordinate system. You can see this coordinate system in Figure 9.3.

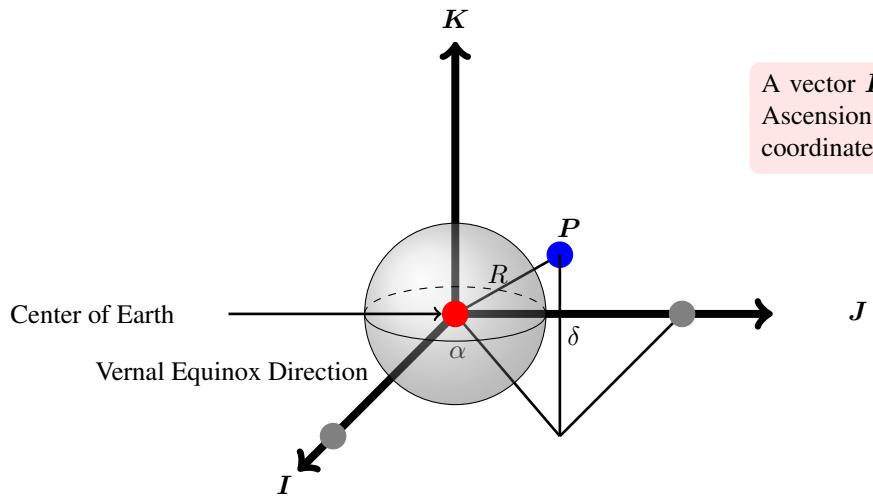


Figure 9.3: Right Ascension - Declination coordinate system

9.1.1 Homework

Exercise 9.1.1

Exercise 9.1.2

Exercise 9.1.3

Exercise 9.1.4

Exercise 9.1.5

9.2 Orbital Planes

The above coordinate systems are not specialized to the movement of a satellite in its orbit. The **Perifocal** coordinate system describes the position of points in space in terms of a coordinate system based on the orbital motion of the satellite.. Let M be the mass of the earth and m the mass of the satellite. We will think of position and velocity as numbers in \mathbb{R}^3 and so each position and associated velocity defines vectors

$$\mathbf{r}(t) = \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix}, \quad \mathbf{v}(t) = \mathbf{r}'(t) = \begin{bmatrix} x'(t) \\ y'(t) \\ z'(t) \end{bmatrix},$$

and we can decide to express this information in any choice of coordinate system we wish. The vector \mathbf{r} starts at the origin of the coordinate system and ends at the position of the satellite. So there is an obvious unit vector we can use here, $\mathbf{E}(t) = \mathbf{r}(t)/\|\mathbf{r}\|_2$. The gravitational force between the earth and the satellite is modeled as if all the mass is concentrated at their individual centers and we can write down the usual gravitational force equations you know from basic physics. The force is proportional to the acceleration and is directed along the line between the earth and the satellite. The equation for the vector acceleration is then

$$\frac{d^2}{dt^2} \mathbf{r}(t) = -\frac{G(m+M)}{\|\mathbf{r}(t)\|_2^2} \mathbf{E}(t) = -\frac{G(m+M)}{\|\mathbf{r}(t)\|_2^3} \mathbf{r}(t)$$

where G is the universal gravitational constant. Since m is negligible compared to M , we approximate the vector acceleration by

$$\frac{d^2}{dt^2} \mathbf{r}(t) = -\frac{G(M)}{\|\mathbf{r}(t)\|_2^3} \mathbf{r}(t) = -\frac{\mu}{\|\mathbf{r}(t)\|_2^3} \mathbf{r}(t)$$

where μ is called the gravitational parameter. To make our arguments a bit more concise, let's start using a common notation for $\|\mathbf{r}\|_2$. This is often just written as $|\mathbf{r}(t)|$ even though we know we are not just taking an absolute value. For the rest of our arguments here, we will use this more simple notation. Hence, we have $\mathbf{r}''(t) = -(\mu/|\mathbf{r}(t)|^3) \mathbf{r}(t)$. We will also drop the (t) now: just remember it is still there!

9.2.1 Orbital Constants

Now let's do some calculations: note $\mathbf{v} = \mathbf{r}'$ and $\mathbf{r}'' = \mathbf{v}'$.

$$\langle \mathbf{v}, \mathbf{v}' \rangle = \langle \mathbf{r}', \mathbf{r}'' \rangle = \langle \mathbf{r}', -(\mu/|\mathbf{r}|^3) \mathbf{r} \rangle = \langle \mathbf{v}, -(\mu/|\mathbf{r}|^3) \mathbf{r} \rangle$$

Thus

$$\langle \mathbf{v}, \mathbf{v}' \rangle + \langle \mathbf{v}, (\mu/|\mathbf{r}|^3) \mathbf{r} \rangle = 0$$

But we are using the $\|\cdot\|_2$ norm here which is the usual Euclidean norm in \mathbb{R}^3 . Now, it is easy to see

$$(|\mathbf{v}|)' = \frac{1}{2} \frac{2v_2v'_1 + 2v_2v'_2 + 2v_3v'_3}{\sqrt{v_1^2 + v_2^2 + v_3^2}} = \frac{\langle \mathbf{v}, \mathbf{v}' \rangle}{|\mathbf{v}|}$$

which implies $|\mathbf{v}| |\mathbf{v}'| = \langle \mathbf{v}, \mathbf{v}' \rangle$. It is usual to find a way to distinguish between the scalar $|\mathbf{v}|$ and the vector \mathbf{v} . We will let $\mathcal{V} = |\mathbf{v}|$. Hence, we have $\mathcal{V} \mathcal{V}' = \langle \mathbf{v}, \mathbf{v}' \rangle$. We will also write $\mathcal{R} = |\mathbf{r}|$ and a similar calculation shows $\mathcal{R} \mathcal{R}' = \langle \mathbf{r}, \mathbf{r}' \rangle$. Thus,

$$\langle \mathbf{v}, \mathbf{v}' \rangle + \langle \mathbf{v}, (\mu/|\mathbf{r}|^3) \mathbf{r} \rangle = 0 \implies \mathcal{V} \mathcal{V}' + (\mu/\mathcal{R}^2) \mathcal{V} = 0$$

But

$$(1/2) \frac{d}{dt} (\mathcal{V}^2)' = \mathcal{V} \mathcal{V}' \implies \frac{d}{dt} (-\mu/\mathcal{R}) = \mu \frac{1}{\mathcal{R}^2} \mathcal{R}'$$

Hence, we conclude

$$\frac{d}{dt} ((1/2)\mathcal{V}^2)' + \frac{d}{dt} \left(\frac{-\mu}{\mathcal{R}} \right) = 0$$

Our final equation is thus

$$\frac{d}{dt} \left(\frac{\mathcal{V}^2}{2} - \frac{\mu}{\mathcal{R}} \right) = 0$$

This says the term $\mathcal{V}/2 - \mu/\mathcal{R}$ is a constant which we call the **specific mechanical energy** and denote by \mathcal{E} .

Next, we since

$$\frac{d^2}{dt^2} \mathbf{r} + \frac{\mu}{|\mathbf{r}|^3} \mathbf{r} = 0$$

we can find the cross product of this with \mathbf{r} . We find

$$\mathbf{r} \times \mathbf{r}'' + \frac{\mu}{|\mathbf{r}|^3} \mathbf{r} \times \mathbf{r} = \mathbf{0}$$

Since $\mathbf{r} \times \mathbf{r} = \mathbf{0}$, we have $\mathbf{r} \times \mathbf{r}'' = \mathbf{0}$. By direct calculation, we can check

$$\frac{d}{dt} (\mathbf{r} \times \mathbf{r}') = \mathbf{r}' \times \mathbf{r}' + \mathbf{r} \times \mathbf{r}''$$

But $\mathbf{r}' \times \mathbf{r}' = \mathbf{0}$, so

$$\mathbf{r} \times \mathbf{r}'' = \frac{d}{dt} (\mathbf{r} \times \mathbf{r}')$$

We have shown

$$\mathbf{0} = \mathbf{r} \times \mathbf{r}'' = \frac{d}{dt} (\mathbf{r} \times \mathbf{r}') = \frac{d}{dt} (\mathbf{r} \times \mathbf{v})$$

The angular momentum here is $\mathbf{h} = \mathbf{r} \times \mathbf{v}$. So we have shown the angular momentum \mathbf{h} is constant. We let $\mathcal{H} = |\mathbf{h}|$. Thus, a satellite moves in an orbit around the center of the earth with constant angular momentum.

9.2.1.1 Homework

Exercise 9.2.1

Exercise 9.2.2

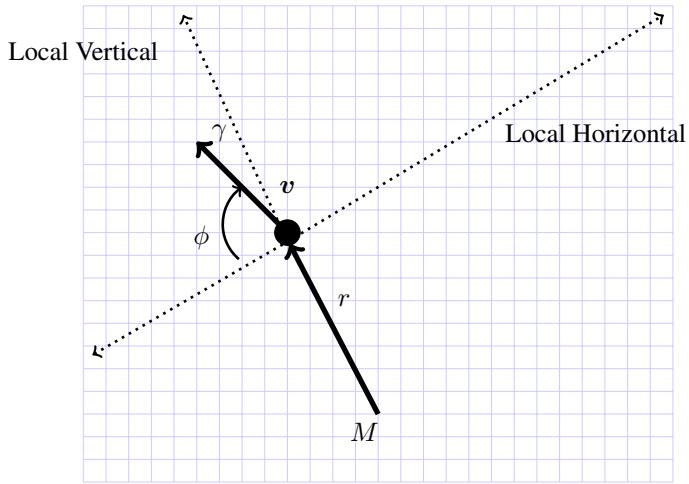
Exercise 9.2.3

Exercise 9.2.4

Exercise 9.2.5

9.2.2 The Orbital Motion is a Conic

Since \mathbf{r} and \mathbf{v} determine the orbital plane, it is convenient to setup some conventions as shown in Figure 9.4.



The radius vector from the center of the earth to the satellite determines a local vertical and local horizontal. The angle between \mathbf{v} and the local vertical is γ and the angle between \mathbf{v} and the local horizontal is ϕ .

Figure 9.4: The Orbital Plane

The **flight path angle**, ϕ , is the angle between the local horizontal and \mathbf{v} . The **zenith angle**, γ , is the angle between \mathbf{v} and the local vertical. We see $\gamma = \pi/2 - \phi$.

9.2.2.1 Homework

Exercise 9.2.6

Exercise 9.2.7

Exercise 9.2.8

Exercise 9.2.9

Exercise 9.2.10

9.2.3 The Constant \mathbf{B} Vector

We know the angular momentum \mathbf{h} is a constant vector and so

$$\mathcal{H} = |\mathbf{h}| = |\mathbf{r} \times \mathbf{v}| = \mathcal{R}\mathcal{V} \sin(\gamma) = \mathcal{R}\mathcal{V} \cos(\phi)$$

Now we know $\mathbf{r}'' = -(\mu/\mathcal{R}^3)\mathbf{r}$. Thus,

$$\mathbf{r}'' \times \mathbf{h} = -\frac{\mu}{\mathcal{R}^3} \mathbf{h} \times \mathbf{r}$$

9.2. ORBITAL PLANES

135

Then, consider

$$(\mathbf{r}' \times \mathbf{h})' = \mathbf{r}'' \times \mathbf{h} + \mathbf{r}' \times \mathbf{h}' = \mathbf{r}'' \times \mathbf{h}$$

as \mathbf{h} is a constant vector. Combining, we have

$$(\mathbf{r}' \times \mathbf{h})' = \frac{\mu}{\mathcal{R}^3} \mathbf{h} \times \mathbf{r}$$

Next, we look at

$$\frac{\mu}{\mathcal{R}^3} \mathbf{h} \times \mathbf{r} = \frac{\mu}{\mathcal{R}^3} ((\mathbf{r} \times \mathbf{v}) \times \mathbf{r})$$

You probably haven't worked much with all the identities associated with vector cross products, so let's go through this calculation.

$$\mathbf{r} \times \mathbf{v} = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ r_1 & r_2 & r_3 \\ v_1 & v_2 & v_3 \end{bmatrix} = (r_2 v_3 - v_2 r_3) \mathbf{i} + (r_3 v_1 - r_1 v_3) \mathbf{j} + (r_1 v_2 - v_1 r_2) \mathbf{k}$$

Thus, after some simplifying we leave to you

$$\begin{aligned} (\mathbf{r} \times \mathbf{v}) \times \mathbf{r} &= (-1) \cdot \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ r_1 & r_2 & r_3 \\ (r_2 v_3 - v_2 r_3) & (r_3 v_1 - r_1 v_3) & (r_1 v_2 - v_1 r_2) \end{bmatrix} \\ &= -(r_1 r_2 v_2 + r_1 r_3 v_3 - r_2^2 v_1 - r_3^2 v_1) \mathbf{i} + (-r_1 r_2 v_1 - r_2 r_3 v_3 + r_1^2 v_2 + r_3^2 v_2) \mathbf{j} \\ &\quad - (r_1 r_3 v_1 + r_2 r_3 v_2 - r_1^2 v_3 - r_2^2 v_3) \mathbf{k} \end{aligned}$$

We can rewrite this as

$$\begin{aligned} (\mathbf{r} \times \mathbf{v}) \times \mathbf{r} &= -(r_1(\langle \mathbf{r}, \mathbf{v} \rangle - r_1 v_1) - v_1(\langle \mathbf{r}, \mathbf{r} \rangle - r_1^2)) \mathbf{i} \\ &\quad + (-r_2(\langle \mathbf{r}, \mathbf{v} \rangle - r_2 v_2) + v_2(\langle \mathbf{r}, \mathbf{r} \rangle - r_2^2)) \mathbf{j} \\ &\quad - (r_3(\langle \mathbf{r}, \mathbf{v} \rangle - r_3 v_3) - v_3(\langle \mathbf{r}, \mathbf{r} \rangle - r_3^2)) \mathbf{k} \\ &= -(r_1 \langle \mathbf{r}, \mathbf{v} \rangle - v_1(\langle \mathbf{r}, \mathbf{r} \rangle)) \mathbf{i} + (-r_2 \langle \mathbf{r}, \mathbf{v} \rangle + v_2 \langle \mathbf{r}, \mathbf{r} \rangle) \mathbf{j} \\ &\quad - (r_3 \langle \mathbf{r}, \mathbf{v} \rangle - v_3 \langle \mathbf{r}, \mathbf{r} \rangle) \mathbf{k} \end{aligned}$$

We can reorganize the above to obtain our final formula:

$$(\mathbf{r} \times \mathbf{v}) \times \mathbf{r} = -\langle \mathbf{r}, \mathbf{v} \rangle \mathbf{r} + \langle \mathbf{r}, \mathbf{r} \rangle \mathbf{v}$$

Using this, we find

$$\begin{aligned} \frac{\mu}{\mathcal{R}^3} \mathbf{h} \times \mathbf{r} &= \frac{\mu}{\mathcal{R}^3} (-\langle \mathbf{r}, \mathbf{v} \rangle \mathbf{r} + \langle \mathbf{r}, \mathbf{r} \rangle \mathbf{v}) = \frac{\mu}{\mathcal{R}^3} (-\mathcal{R} \mathcal{R}' \mathbf{r} + \mathcal{R}^2 \mathbf{v}) \\ &= \frac{\mu}{\mathcal{R}} \mathbf{v} - \frac{\mu}{\mathcal{R}^2} \mathcal{R}' \mathbf{r} \end{aligned}$$

Recall, we know

$$(\mathbf{r}' \times \mathbf{h})' = \frac{\mu}{\mathcal{R}^3} \mathbf{h} \times \mathbf{r} = \frac{\mu}{\mathcal{R}} \mathbf{v} - \frac{\mu}{\mathcal{R}^2} \mathcal{R}' \mathbf{r}$$

But,

$$\mu \left(\frac{\mathbf{r}}{\mathcal{R}} \right)' = \mu \left(\frac{\mathbf{r}'}{\mathcal{R}} - \frac{\mathcal{R}'}{\mathcal{R}^2} \mathbf{r} \right) = \mu \left(\frac{\mathbf{v}}{\mathcal{R}} - \frac{\mathcal{R}'}{\mathcal{R}^2} \mathbf{r} \right)$$

This is what we had above! So we can say

$$(\mathbf{r}' \times \mathbf{h})' = \mu \left(\frac{\mathbf{r}}{\mathcal{R}} \right)'$$

We conclude the derivative of $\mathbf{r}' \times \mathbf{h} - \mu(\mathbf{r}/\mathcal{R})$ is a constant vector. Hence, there is a constant vector \mathbf{B} so that

$$\mathbf{r}' \times \mathbf{h} = \mu \frac{\mathbf{r}}{\mathcal{R}} + \mathbf{B}$$

9.2.3.1 Homework

Exercise 9.2.11

Exercise 9.2.12

Exercise 9.2.13

Exercise 9.2.14

Exercise 9.2.15

9.2.4 The Orbital Conic

We can use the equation giving us the constant \mathbf{B} vector to derive the equation of motion we associate with the satellite's motion. Consider

$$\langle \mathbf{r}, \mathbf{r}' \times \mathbf{h} \rangle = \left\langle \mathbf{r}, \mu \frac{\mathbf{r}}{\mathcal{R}} + \mathbf{B} \right\rangle = \mu \frac{\langle \mathbf{r}, \mathbf{r} \rangle}{\mathcal{R}} + \langle \mathbf{r}, \mathbf{B} \rangle = \mu \mathcal{R} + \langle \mathbf{r}, \mathbf{B} \rangle$$

Now, we need another vector computation you probably haven't seen. Let's look at $\langle \mathbf{A}, \mathbf{B} \times \mathbf{C} \rangle$ for any vectors \mathbf{A} , \mathbf{B} and \mathbf{C} .

$$\mathbf{B} \times \mathbf{C} = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ B_1 & B_2 & B_3 \\ C_1 & C_2 & C_3 \end{bmatrix} = (B_2 C_3 - B_3 C_2) \mathbf{i} + (B_3 C_1 - B_1 C_3) \mathbf{j} + (B_1 C_2 - B_2 C_1) \mathbf{k}$$

and so

$$\begin{aligned} \langle \mathbf{A}, \mathbf{B} \times \mathbf{C} \rangle &= A_1(B_2 C_3 - B_3 C_2) + A_2(B_3 C_1 - B_1 C_3) + A_3(B_1 C_2 - B_2 C_1) \\ &= C_1(A_2 B_3 - A_3 B_2) + C_2(A_3 B_1 - A_1 B_3) + C_3(A_1 B_2 - A_2 B_1) \\ &= \langle \mathbf{A} \times \mathbf{B}, \mathbf{C} \rangle \end{aligned}$$

So

$$\langle \mathbf{r}, \mathbf{r}' \times \mathbf{h} \rangle = \langle \mathbf{r} \times \mathbf{r}', \mathbf{h} \rangle = \langle \mathbf{h}, \mathbf{h} \rangle = \mathcal{H}^2$$

This means

$$\mathcal{H}^2 = \langle \mathbf{r}, \mathbf{r}' \times \mathbf{h} \rangle = \mu \mathcal{R} + \langle \mathbf{r}, \mathbf{B} \rangle = \mu \mathcal{R} + \mathcal{R} \mathcal{B} \cos(\nu)$$

where ν is the angle between \mathbf{r} and \mathbf{B} . Rewriting, we have

$$\mathcal{R} = \frac{\left(\frac{\mathcal{H}^2}{\mu}\right)}{1 + \left(\frac{\mathcal{B}}{\mu}\right) \cos(\nu)}$$

The scalar $\mathcal{E} = \mathcal{B}/\mu$ is called the **eccentricity** of the orbit and so the vector $e = \mathbf{B}/\mu$ is called the eccentricity vector. From what we know about the ellipse, the angle ν measures the angle between the vector pointing to the closest point on the ellipse to the current point. Hence, the constant vector \mathbf{B} must point in the direction of periapsis. Therefore e points in the direction of periapsis.

We know $e = \mathbf{B}/\mu u = \mathbf{r}' \times \mathbf{h} - \mu \frac{\mathbf{r}}{\mathcal{R}}$. Thus,

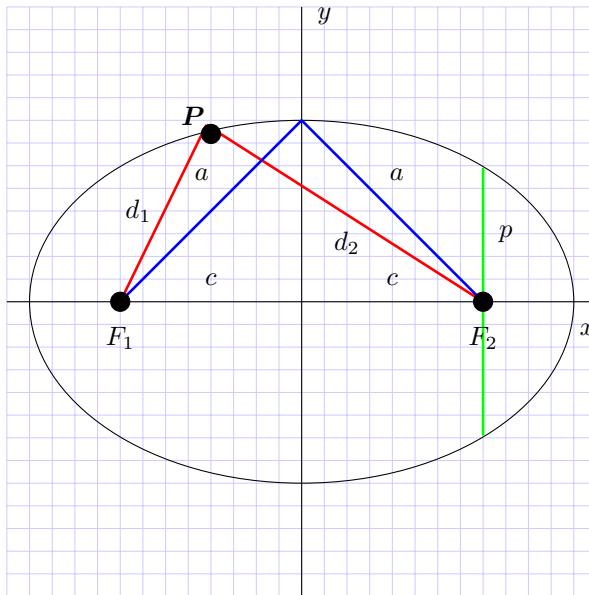
$$\mu e = \mathbf{v} \times (\mathbf{r} \times \mathbf{v}) - \mu \mathbf{r} / \mathcal{R} \quad (9.1)$$

$$= \langle \mathbf{v}, \mathbf{v} \rangle \mathbf{r} - \langle \mathbf{r}, \mathbf{v} \rangle \mathbf{v} - \mu \mathbf{r} / \mathcal{R} \quad (9.2)$$

$$= (\mathcal{V}^2 - \mu / \mathcal{R}) \mathbf{r} - \langle \mathbf{r}, \mathbf{v} \rangle \mathbf{v} \quad (9.3)$$

where we expand the triple cross product just like we did the previous one. This equation tells us the direction of periapsis for the conic!

You are probably more used to the Cartesian Coordinate version of an ellipse. Look at Figure 9.5.



The defining parameters for an ellipse are $d_1 + d_2 = 2a$ and $a^2 = b^2 + c^2$. The eccentricity is $e = c/a$.

Figure 9.5: An Ellipse

There are two points called foci that are used to define an ellipse. These points are labeled F_1 and F_2 and are on the x axis with each a distance c from the origin. Hence, $F_1 = (-c, 0)$ and $F_2 = (c, 0)$. Let d_1 be the distance from a point on the ellipse to focus F_1 and d_2 be the distance from a point on the ellipse to focus F_2 . The points on the ellipse must satisfy $d_1 + d_2 = 2a$ and we can use this to find certain relationships. At the top of the ellipse, we get $d_1 = d_2 = a$ which defined a right triangle as you can see in Figure 9.5. We let b be the distance from the center to the top of the ellipse: hence $A^2 = b^2 + c^2$. The ratio c/a is called the **eccentricity** and clearly for an ellipse, $0 > e < 1$. The length of the vertical line through the focus F_1 is $4p$ and half of this distance is called the latus rectum of the ellipse which is thus $2p$. The top part of the line defining the latus rectum gives us the

equation $d_1 + 2p = 2a$ and the pythagorean theorem for the triangle formed gives $d_1^2 = 4p^2 + 4c^2$. Thus $4(a-p)^2 = 4p^2 + 4c^2$. This implies $a^2 - 2ap = c^2$. Hence,

$$2ap = a^2 - c^2 = a^2(1 - e^2) \implies 2p = a(1 - e^2)$$

With some work (and it is not so trivial!) we can derive the corresponding equation for an ellipse in polar coordinates. We will leave that to you. We find

$$\mathcal{R} = \frac{2p}{1 + e \cos(\nu)}$$

Comparing the equation of motion we found for the satellite, we see $\mathcal{H}^2/\mu = 2p = a(1 - e^2)$ and $\mathcal{B}/\mu = e$. Now in the orbital plane

$$\begin{aligned} \mathbf{r}(\nu) &= \begin{bmatrix} \mathcal{R} \cos(\nu) \\ \mathcal{R} \sin(\nu) \end{bmatrix} = \begin{bmatrix} \frac{2p \cos(\nu)}{1+e \cos(\nu)} \\ \frac{2p \sin(\nu)}{1+e \cos(\nu)} \end{bmatrix} \\ \mathbf{v}(\nu) &= \begin{bmatrix} \frac{-2p \sin(\nu)}{1+e \cos(\nu)} - \frac{2p \cos(\nu)}{(1+e \cos(\nu))^2} (-e \sin(\nu)) \\ \frac{2p \cos(\nu)}{1+e \cos(\nu)} - \frac{2p \sin(\nu)}{(1+e \cos(\nu))^2} (-e \sin(\nu)) \end{bmatrix} \\ &= \begin{bmatrix} \frac{-2p \sin(\nu)(1+e \cos(\nu)) + 2pe \sin(\nu) \cos(\nu)}{(1+e \cos(\nu))^2} \\ \frac{2p \cos(\nu)(1+e \cos(\nu)) + 2pe \sin^2(\nu)}{(1+e \cos(\nu))^2} \end{bmatrix} \end{aligned}$$

Hence, at periapsis, $\nu = 0$ and

$$\mathbf{r}_p = \begin{bmatrix} \frac{2p}{1+e} \\ 0 \end{bmatrix}, \quad \mathbf{v}_p = \begin{bmatrix} 0 \\ \frac{2p}{1+e} \end{bmatrix}$$

and at apoapsis, $\nu = \pi$

$$\mathbf{r}_a = \begin{bmatrix} \frac{2p}{1-e} \\ 0 \end{bmatrix}, \quad \mathbf{v}_a = \begin{bmatrix} 0 \\ \frac{-2p}{1-e} \end{bmatrix}$$

Since angular momentum is constant on the orbit, we know

$$\mathcal{H} = \mathcal{R} \mathcal{V} \cos(\nu) = \mathcal{R}_p \mathcal{V}_p = \mathcal{R}_a \mathcal{V}_a$$

But this means

$$\mathcal{H} = \frac{2p}{1+e} \frac{2p}{1+e} = \frac{2p}{1-e} \frac{2p}{1-e}$$

9.2.4.1 Homework

Exercise 9.2.16

Exercise 9.2.17

Exercise 9.2.18

Exercise 9.2.19

Exercise 9.2.20

9.3 Three Dimensional Rotations

If we have a 3×3 symmetric matrix A , we know it has three mutually orthogonal eigenvectors corresponding to three real eigenvalues. Hence the unit eigenvectors form an orthonormal basis for \mathbb{R}^3 and we can write

$$\begin{bmatrix} \mathbf{E}_1^T \\ \mathbf{E}_2^T \\ \mathbf{E}_3^T \end{bmatrix} A \begin{bmatrix} \mathbf{E}_1 & \mathbf{E}_2 & \mathbf{E}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

where $\{\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3\}$ is an orthonormal basis for \mathbb{R}^3 with corresponding eigenvalues λ_1, λ_2 and λ_3 . We now the rows and columns of the matrix $P = [\mathbf{E}_1 \ \mathbf{E}_2 \ \mathbf{E}_3]$ are mutually orthogonal. Hence, like what happened in the two dimensional case, P is often a rotation matrix which transforms an orthonormal system of mutually orthogonal axis labeled x, y and z into a new orthogonal system labeled x', y' and z' . Let's study some of these rotation matrices. There are three simple building blocks:

- The axis of rotation is the z axis and we rotate the $x - y$ system about the z axis by θ degrees. The new system is thus x', y' and z .
- Now apply to this new system a rotation of ϕ degrees about the new y' axis. So the x' and z axis rotate into the new axis x'' and z' . This gives a new orthogonal system with axis x'', y' and z' .
- Now apply to this new system a rotation of δ degrees about the new x'' axis. This generates a new orthogonal system with axes x'', y'' and z'' .

Of course, we could organize three rotations in different ways but to explain this it helps to have a concrete set of choices. We want to draw some of this to help you see it, so we use MatLab. It is a bit complicated to draw things in any programming language. Let's begin with drawing a simple vector using `ThreeDTOPlot.m` shown below.

Listing 9.1: **ThreeDTOPlot.m**

```
function [x,y,z] = ThreeDTOPlot(A)
%
% A is the vector to plot using plot3
% x,y,z is prepared vector that plot3 uses
% N is the number of points to plot
%
t = linspace(0,1,2);
VA = @(t) t*A;
for i=1:2
    C = VA(t(i));
    x(i) = C(1);
    y(i) = C(2);
    z(i) = C(3);
end
end
```

This looks a little strange, but the `plot3` MatLab function draws the line between two vectors \mathbf{A} and \mathbf{B} like this. If

$$\mathbf{A} = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}$$

then to draw the line between \mathbf{A} and \mathbf{B} , we collect x , y and z points as

$$\mathbf{x} = \begin{bmatrix} A_1 \\ B_1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} A_2 \\ B_2 \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} A_3 \\ B_3 \end{bmatrix}$$

and then `plot3` plots the line between the point $(x(1), y(1), z(1))$ and the point $(x(2), y(2), z(2))$. Since this is a pain to do for all the vectors we want to plot, we've rolled this into a convenience function so we don't have to type so much.

In addition, we need our rotation matrices functions which we call `RotateX.m`, `RotateY.m` and `RotateZ.m`. Here is the code. The function `RotateX` rotates about the x axis by positive θ° . to give the new coordinate system (x, y', z') . This is a positive rotation which means your thumb points in the direction of the $+x$ axis and your fingers curl in the direction of the positive θ . Hence, this is a counterclockwise (ccw) rotation from the $+y$ axis giving the new coordinate system (x, y', z') . The rotation matrix here is $\mathbf{R}_{x,\theta}$ with

$$\mathbf{R}_{x,\theta} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_\theta \end{bmatrix}$$

where $\mathbf{0}$ here is whatever size of zero vector or its transpose we need and \mathbf{R}_θ is the usual 2×2 rotation matrix.

Listing 9.2: **RotateX.m**

```
function Rx = RotateX(theta)
% + theta means ccw which is what we usually want
% - theta means cw
ctheta = cos(theta);
s stheta = sin(theta);
%
Rx = [1,0,0;0,ctheta, -stheta; 0,stheta, ctheta];
end
```

The function `RotateY` rotates about the y axis by positive θ° . to give the new coordinate system (x', y, z') . The rotation matrix here is $\mathbf{R}_{y,\theta}$ with

$$\mathbf{R}_{y,\theta} = \begin{bmatrix} \cos(\theta) & 0 & -\sin(\theta) \\ 0 & 1 & 0 \\ \sin(\theta) & 0 & \cos(\theta) \end{bmatrix}$$

and you can clearly see \mathbf{R}_θ buried inside this matrix.

Listing 9.3: **RotateY.m**

```
function Ry = RotateY(theta)
```

```
% + theta means cw
% - theta means ccw which is what we usually want
ctheta = cos(theta);
stheta = sin(theta);
%
Ry = [ ctheta, 0, -stheta; 0, 1, 0; stheta, 0, ctheta];
end
```

The function **RotateZ** rotates about the z axis by positive θ° . to give the new coordinate system (x', y', z) . The rotation matrix here is $R_{z,\theta}$ with

$$R_{z,\theta} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and again you can clearly see R_θ buried inside this matrix.

Listing 9.4: **RotateZ.m**

```
function Rz = RotateZ(theta)
% + theta means ccw which is what we usually want
% - theta means cw
ctheta = cos(theta);
stheta = sin(theta);
Rz = [ ctheta, -stheta, 0; stheta, ctheta, 0; 0, 0, 1];
end
```

9.3.1 Homework

Exercise 9.3.1

Exercise 9.3.2

Exercise 9.3.3

Exercise 9.3.4

Exercise 9.3.5

9.3.2 Drawing Rotations

The easiest way to understand the rotations is to look at some graphs. An even better way is to build a model out of cardboard: you should try it! We can draw the two sets of orthogonal axis we get from the application of rotations using the following code. This code is designed to work with drawing elliptical orbits so we setup the size of the axis system based on the size of the ellipse we will want to draw. Since the apoapsis of the ellipse is $R/(1 - e)$ where e is the eccentricity and R is the scaling factor, we use that to set the axis dimensions. Then later, any ellipse we draw will fit into the coordinate system nicely. The rest of the code simply draws straight lines and determines the line style of the drawn line.

Listing 9.5: **DrawAxis.m**

```

function DrawAxis (R,e,N,E1,E2,E3,A,style ,color )
2 %
% Draw coordinate system
hold on;
Rmax = R/(1-e);
% Graph new z axis and new y axis
7 t = linspace(-1.5*Rmax,1.5*Rmax,N);
W1 = E1;
vW1 = @(t) t*W1;
for i=1:N
    s = t(i); D = vW1(s);
12 x1(i) = D(1) + A(1);
    y1(i) = D(2) + A(2);
    z1(i) = D(3) + A(3);
end
%
17 W2 = E2;
vW2 = @(t) t*W2;
for i=1:N
    s = t(i); D = vW2(t(i));
    x2(i) = D(1) + A(1);
22 y2(i) = D(2) + A(2);
    z2(i) = D(3) + A(3);
end
%
W3 = E3;
27 vW3 = @(t) t*W3;
for i=1:N
    s = t(i); D = vW3(t(i));
    x3(i) = D(1) + A(1);
    y3(i) = D(2) + A(2);
32 z3(i) = D(3) + A(3);
end
if (style == 1)
    plot3(x1,y1,z1,'-','LineWidth',2,'Color',color);
    plot3(x2,y2,z2,'-','LineWidth',2,'Color',color);
37 plot3(x3,y3,z3,'-','LineWidth',2,'Color',color);
elseif (style == 2)
    if (style == 2)
        plot3(x1,y1,z1,'o','LineWidth',2,'Color',color);
        plot3(x2,y2,z2,'o','LineWidth',2,'Color',color);
42 plot3(x3,y3,z3,'o','LineWidth',2,'Color',color);
    else
        disp ("not one or two");
    end
    hold off;
47 end

```

Let's look at a rotation about the $+z$ axis of angle θ . The code is simple.

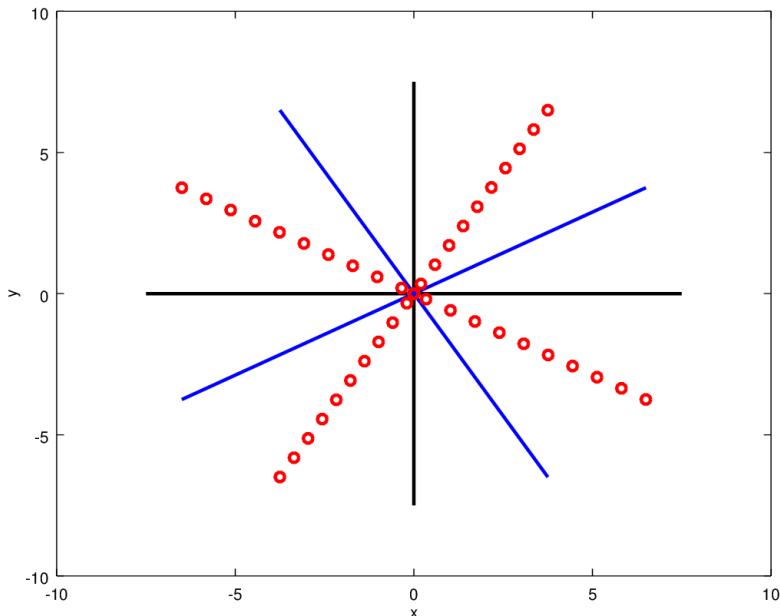
Listing 9.6: **Rotation about the z axis**

```

I = [1;0;0]; J = [0;1;0]; K = [0;0;1]; A = [0;0;0];
2 hold on;
DrawAxis (2,0.6,20,I,J,K,A,1,'black');
theta = pi/6;
Xp = RotateZ(theta)*I;
Yp = RotateZ(theta)*J;
7 Zp = RotateZ(theta)*K;
DrawAxis(2,0.6,20,Xp,Yp,Zp,A,1,'blue');
hold off
Xpn = RotateZ(-theta)*I;
Ypn = RotateZ(-theta)*J;
12 Zpn = RotateZ(-theta)*K;
DrawAxis(2,0.6,20,Xpn,Ypn,Zpn,A,2,'red');
 xlabel('x');
 ylabel('y')
 zlabel('z');

```

In Figure 9.6, we look down on the plot so that the $+z$ axis is pointing straight up and can not be seen. The positive θ rotation moves the x axis counter clockwise. On the other hand, the negative rotation by $-\theta$ is shown with axis plotted by circles and it moves the $+x$ axis clockwise. So a positive rotation about the $+z$ axis means the 2×2 rotation matrix seen inside should be R_θ . The standard way to pick the kind of rotation we want is to define a positive rotation about a positive axis as follows: line the thumb of your right hand up along the positive axis. Then your fingers will curl in the direction of what we call a positive rotation angle. So positive rotations along the $+z$ axis give the usual embedded R_θ part of the three dimensional rotation matrix.

Figure 9.6: A positive and negative rotation about the z axis

Homework

Exercise 9.3.6

Exercise 9.3.7

Exercise 9.3.8

Exercise 9.3.9

Exercise 9.3.10

Next, look at a rotation about the $+y$ axis of angle θ . The code is simple.

Listing 9.7: **Rotation about the y axis**

```
I = [1;0;0]; J = [0;1;0]; K = [0;0;1]; A = [0;0;0];
hold on;
DrawAxis (2,0.6,20,I,J,K,A,1,'black');
4 theta = pi/6;
Xp = RotateY(theta)*I;
Yp = RotateY(theta)*J;
Zp = RotateY(theta)*K;
DrawAxis(2,0.6,20,Xp,Yp,Zp,A,1,'blue');
9 hold off
Xpn = RotateY(-theta)*I;
Ypn = RotateY(-theta)*J;
Zpn = RotateY(-theta)*K;
DrawAxis(2,0.6,20,Xpn,Ypn,Zpn,A,2,'red');
14 xlabel('x');
ylabel('y')
zlabel('z');
```

In Figure 9.7, we look down on the plot so that the $+y$ axis is pointing straight up and can not be seen. The positive θ rotation moves the z axis **counter clockwise** and the $+x$ axis **clockwise**. This corresponds to a rotation by $-\theta$. So a positive rotation about the $+y$ axis means the 2×2 rotation matrix seen inside should be $R_{-\theta}$. Hence, if we want the $+x$ axis to rotate counter clockwise, we would apply R_θ .

Homework

Exercise 9.3.11

Exercise 9.3.12

Exercise 9.3.13

Exercise 9.3.14

Exercise 9.3.15

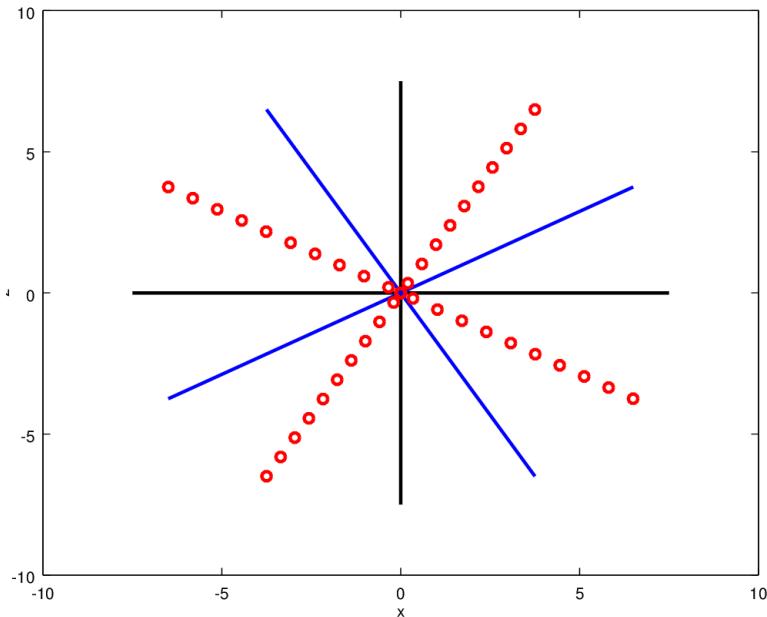
Finally, look at a rotation about the $+x$ axis of angle θ . The code is simple.

Listing 9.8: **Rotation about the x axis**

```
I = [1;0;0]; J = [0;1;0]; K = [0;0;1]; A = [0;0;0];
```

9.3. THREE DIMENSIONAL ROTATIONS

145

Figure 9.7: A positive and negative rotation about the y axis

```

hold on;
DrawAxis (2 ,0.6 ,20 ,I ,J ,K,A,1 ,’black’);
4 theta = pi/6;
Xp = RotateX(theta)*I;
Yp = RotateX(theta)*J;
Zp = RotateX(theta)*K;
DrawAxis(2 ,0.6 ,20 ,Xp,Yp,Zp,A,1 ,’blue’);
9 hold off
Xpn = RotateX(-theta)*I;
Ypn = RotateX(-theta)*J;
Zpn = RotateX(-theta)*K;
DrawAxis(2 ,0.6 ,20 ,Xpn,Ypn,Zpn,A,2 ,’red’ );
14 xlabel(’x’);
ylabel(’y’)
zlabel(’z’);

```

In Figure 9.8, we look down on the plot so that the $+x$ axis is pointing up at almost a vertical angle. The positive θ rotation moves the z axis **counter clockwise** and the $+x$ axis **counter clockwise**. This corresponds to a rotation by θ . So a positive rotation about the $+x$ axis means the 2×2 rotation matrix seen inside should be R_θ .

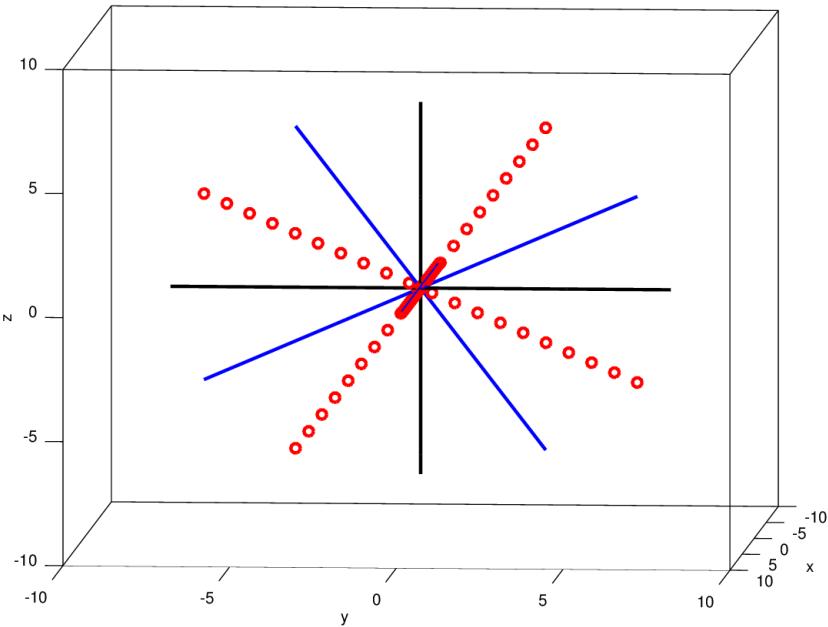
Homework

Exercise 9.3.16

Exercise 9.3.17

Exercise 9.3.18

Exercise 9.3.19

Figure 9.8: A positive and negative rotation about the x axis**Exercise 9.3.20****9.3.3 Rotated Ellipses**

For a given ellipse in the $x - y$ plane, we can rotate it into a given orbital plane using various rotations. The general formula for an ellipse is $r = \rho/(1 + e \cos(t))$ and we use that to set up the plot. In this code, we make sure we move in the counter clockwise direction for the $+x$ axis when we do a rotation with respect to the y axis by explicitly using `RotateY(-phi)`. We show the resulting plot in Figure 9.9 but be warned these plots are hard to see! We rotated it around until you can sort of see the rotations, but with three axis rotations, this gets hard fast.

Listing 9.9: **DrawEllipse.m**

```

function DrawEllipse(rho,e,theta,phi,delta)
% draw orbit in the new x-y coordinate system
Phi = linspace(0,2*pi,51);
R = @(t) rho/(1 + e*cos(t));
u = @(t) R(t)*cos(t);
v = @(t) R(t)*sin(t);
X = @(t) [u(t);v(t);0];
OP = @(t) RotateZ(theta)*X(t);
OPP = @(t) RotateY(-phi)*OP(t);
OPPP = @(t) RotateX(delta)*OPP(t);
x1 = zeros(3,51);
x2 = zeros(3,51);
x3 = zeros(3,51);

```

9.3. THREE DIMENSIONAL ROTATIONS

147

```

conic = zeros(3,51);
15 for i=1:51
    t = Phi(i);
    conic(:,i)= OPPP(t);
end
x1 = conic(1,:);
20 x2 = conic(2,:);
x3 = conic(3,:);
hold on;
plot3(x1,x2,x3,'-','LineWidth',3,'Color','green');
hold off;
25 end

```

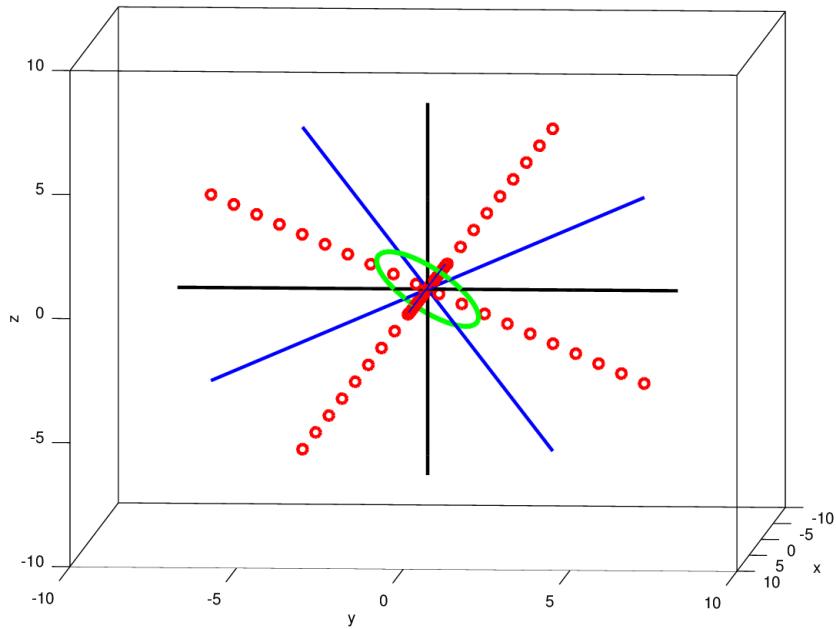


Figure 9.9: A Rotated Ellipse

9.3.3.1 Homework

Exercise 9.3.21**Exercise 9.3.22****Exercise 9.3.23****Exercise 9.3.24****Exercise 9.3.25**

9.4 Drawing the Orbital Plane

Now let's put it all together and draw an orbital plane somewhat carefully. We do this by gluing together the various pieces of code we have written and we include a few functions we have not discussed. However, the code is available for viewing and it is easy enough to figure it out after all we have discussed so far.

Listing 9.10: **DrawOrbit.m**

```

function DrawOrbit(R,e,N,theta,phi,delta,E1,E2,E3,A,style,color)
%
% Draw coordinate system
clf;
5 hold on;
% Graph the original axes
DrawAxisGrid(R,e,N);
DrawAxis(R,e,N,E1,E2,E3,A,1,color);
% Draw rotated system fixing z
10 Xp = RotateZ(theta)*E1;
Yp = RotateZ(theta)*E2;
Zp = E3;
%DDrawAxis(Xp,Yp,Zp,A,'blue');
Xpp = RotateY(-phi)*Xp;
15 Ypp = Yp;
Zpp = RotateY(-phi)*Zp;
%DDrawAxis(Xpp,Ypp,Zpp,A,'green');
Xppp = Xpp;
Yppp = RotateX(delta)*Ypp;
20 Zppp = RotateX(delta)*Zpp;
DrawAxis(R,e,N,Xppp,Yppp,Zppp,A,1,'red');
DrawEllipse(R,e,N,theta,phi,delta);
DrawAxisGrid(R,e,N);
 xlabel('x');
25 ylabel('y');
zlabel('z');
hold off;
end

```

To run this code, we use the command

Listing 9.11: **Plotting an Orbital Plane with Rotated Axes**

```
DrawOrbit(2,0.6,20,pi/6,3*pi/4,3*pi/4,I,J,K,A,1,'black');
```

The plot we generate can be manipulated to get the image in Figure 9.10.

If you look closely, you can see the orbital plane crosses the equatorial plane and the details of this crossing are important to understanding this orbit. If you comment out the code line in `DrawOrbit.m` which generates the rotated axes, you can see this a little more clearly as we show in Figure 9.11. The equatorial plane is drawn heavily crosshatched and the part of the orbital plane above the equatorial plane is reasonably easy to see.

Homework

9.4. DRAWING THE ORBITAL PLANE

149

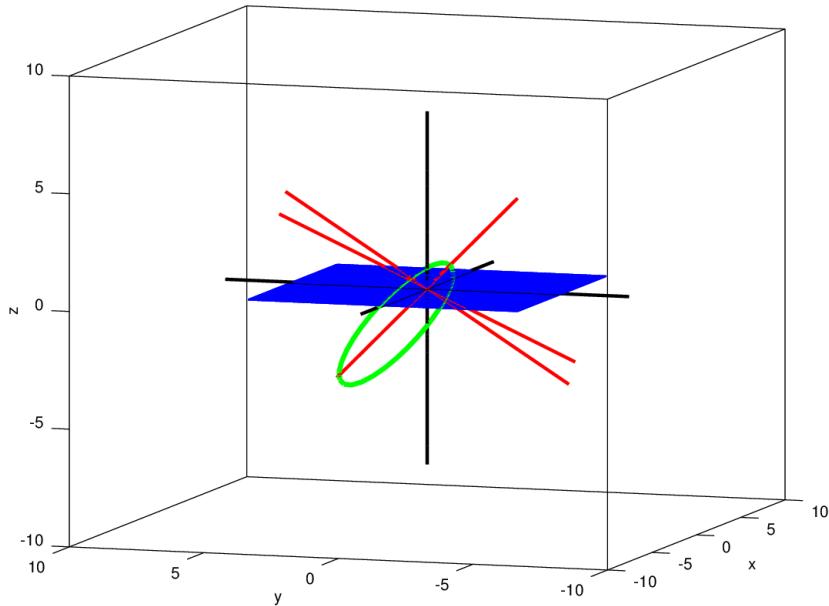


Figure 9.10: A Rotated Ellipse

Exercise 9.4.1**Exercise 9.4.2****Exercise 9.4.3****Exercise 9.4.4****Exercise 9.4.5****9.4.1 The Perifocal Coordinate System**

The position vector \mathbf{r} and its velocity vector \mathbf{v} determine a plane whose normal vector is \mathbf{h} . The closest point on the orbit to the center is called the **periapsis** while the point farthest away is called the **apoapsis**. We let \mathbf{x}_ω denote the axis pointing towards periapsis and \mathbf{y}_ω is the axis rotated 90° in the direction of the orbital motion. Hence, the position of the satellite in the orbit is given by an ordered pair (u, v) . We let \mathbf{P} and \mathbf{Q} be unit vectors along the \mathbf{x}_ω and \mathbf{y}_ω axes respectively. Of course, $\mathbf{P} = \mathbf{e}/\mathcal{E}$. Then, if the position of the satellite in the orbit is denoted by \mathbf{X} then \mathbf{X} has a unique representation in the two dimensional vector space determined by the orthonormal basis $\{\mathbf{P}, \mathbf{Q}\}$ given by $\mathbf{X} = x\mathbf{P} + y\mathbf{Q}$. The positive \mathbf{z}_ω axis is determined by the right hand rule and the unit vector along \mathbf{z}_ω is called \mathbf{W} . Of course, a vector $a\mathbf{P} + b\mathbf{Q} + c\mathbf{W}$ in this system needs to be converted to coordinates in other systems. You can see this coordinate system in Figure 9.12.

9.4.1.1 Homework**Exercise 9.4.6****Exercise 9.4.7****Exercise 9.4.8**

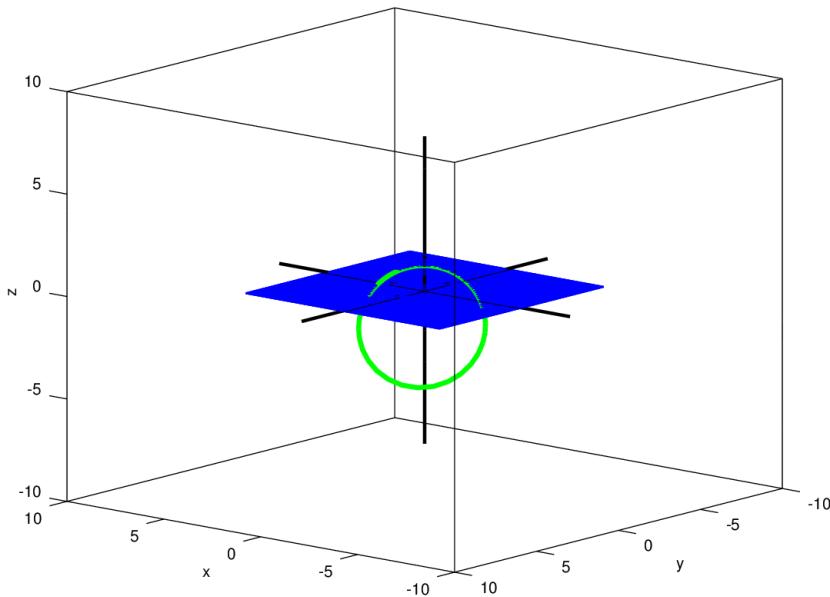


Figure 9.11: The Orbital Plane Crosses the Equatorial Plane

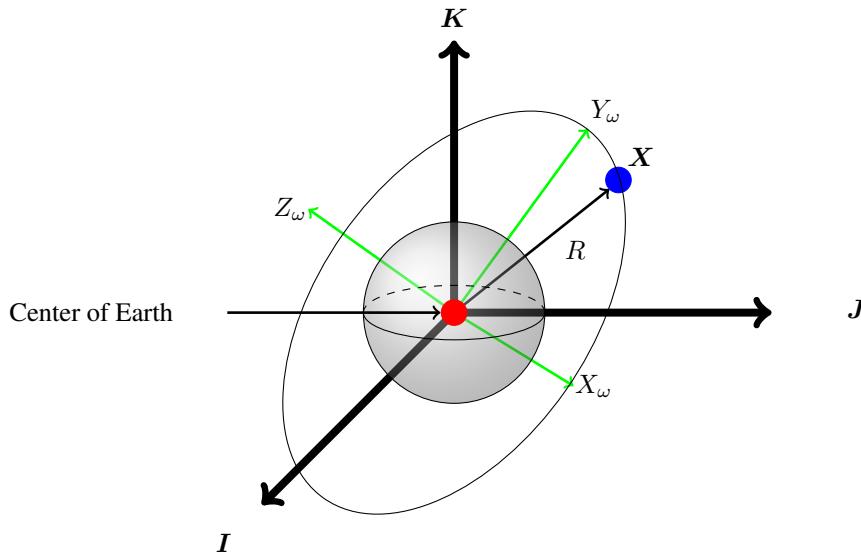
Exercise 9.4.9**Exercise 9.4.10****9.4.2 Orbital Elements**

The orbital elements of an orbit are enough to determine completely the look of the orbit. Let's describe them carefully.

- One half of the major axis of an ellipse is called a . You should look at Figure 9.5 to refresh your mind about this. From the origin of the ellipse at say $(0, 0)$, the point $(c, 0)$ is the focal point F_2 . The point closest to a foci is $(a, 0)$ and so the distance from F_2 to the periapsis is $a - c$.
- The **eccentricity** of the elliptical orbit is called \mathcal{E} . and often just e . We use a boldface e to denote the eccentricity vector and the length of $|e|$ which is traditionally just denoted by e looks a lot like e here. So to be clear, we will let the eccentricity be called \mathcal{E} although sometimes in code we just use e for convenience.
- The **inclination** angle is the angle between the unit vector \mathbf{K} and the angular momentum vector \mathbf{h} . Hence, if we know $\mathbf{R} = R_1\mathbf{I} + R_2\mathbf{J} + R_3\mathbf{K}$ and $\mathbf{r} = R_1\mathbf{I} + R_2\mathbf{J} + R_3\mathbf{K}$, we can calculate

$$\mathbf{h} = \mathbf{r} \times \mathbf{V} = \det \begin{bmatrix} \mathbf{I} & \mathbf{JK} \\ R_1 & R_2 & R_3 \\ V_1 & V_2 & V_3 \end{bmatrix}$$

Once, \mathbf{h} has been calculated, we can use it to find the **line of nodes** vector. The orbital plane crosses the equatorial plane except in degenerate cases such as an equatorial orbit. Now the



A vector X in the Perifocal coordinate system. X marks the position of the satellite in its orbit. The unit vector P is along the X_ω axis, the unit vector Q is along the Y_ω axis and the unit vector W lies on the Z_ω axis. The $I - J - K$ axis for the Geocentric - Equatorial and Right Ascension - Declination coordinate systems are shown for reference.

Figure 9.12: The Perifocal coordinate system

normal to the orbital plane is \mathbf{h} . When the orbital plane crosses the equatorial plane, at those points, it is in the equatorial plane and so must be perpendicular to the equatorial plane too. So the line of intersection of the orbital plane and the equatorial plane is orthogonal to both \mathbf{h} and \mathbf{K} . The cross product then determines a vector called the **node** vector which is the direction vector of the line of nodes. We let \mathbf{n} denote this vector and we see $\mathbf{n} = \mathbf{h} \times \mathbf{K}$.

- Since the line of nodes is in the equatorial plane, we can measure the angle between the line of nodes vector and the I vector. This angle is called the **longitude of the ascending node**, Ω and we see $\Omega = \cos^{-1}(\langle \mathbf{n}, \mathbf{I} \rangle)$.
- The **argument of periapsis** is the angle **in the orbital plane** measured from the line of nodes to the periapsis vector. We denote this by ω .
- The **time of periapsis passage** is the time, T , when the satellite was at periapsis.

In Figure 9.13, we try to show you these in an easy to follow diagram.

9.4.2.1 Homework

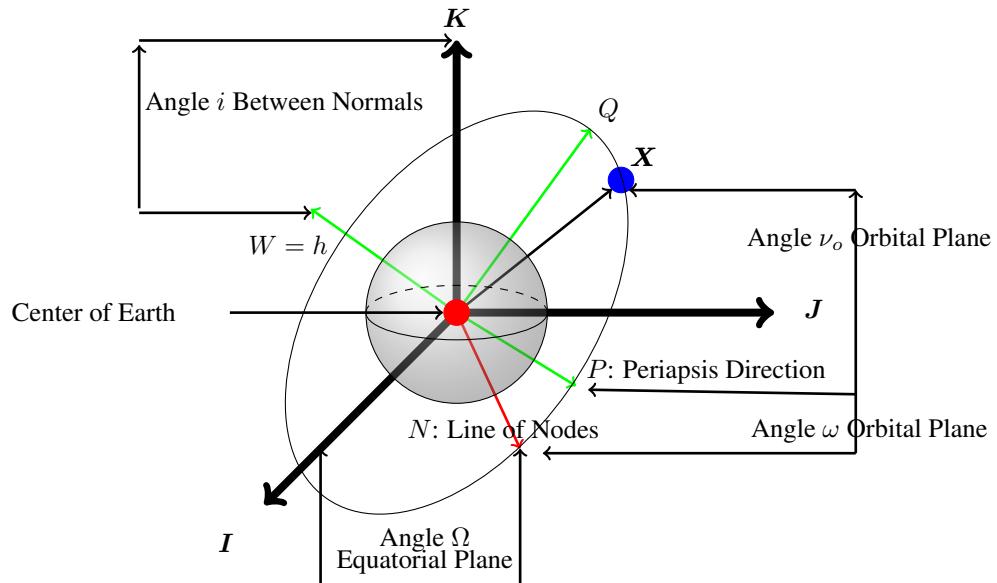
Exercise 9.4.11

Exercise 9.4.12

Exercise 9.4.13

Exercise 9.4.14

Exercise 9.4.15



The orbital elements for a given orbit. All of these are straightforward to calculate given a measurement of \mathbf{r} and \mathbf{v} in the IJK system.

Figure 9.13: The Orbital Elements

9.5 Drawing Orbital Plane In MatLab

Now let's find and graph the orbital plane for a specific example and try to use our knowledge of orbital mechanics along the way. We will explicitly calculate the P , Q and W orthonormal basis vectors of the perifocal coordinate system and the line of nodes vector n . Once that is done, we can plot these vectors as lines in 3D space. So we will not use rotation matrices to do the plots here. We assume we are given the position vector \mathbf{r}_0 and velocity vector \mathbf{v}_0 of the satellite at the initial time 0. These are given as vectors in the equatorial plane coordinate system so each has an I , J and K component. We need a cross product function, so for experience, let's build one.

Listing 9.12: CrossProduct.m

```

function C = CrossProduct(A,B)
% A is a 3D vector
% B is a 3D vector
% returns C, the cross product of A and B
%
% AxB = det(I,J,K;
C = [A(2)*B(3)-A(3)*B(2);A(3)*B(1)-A(1)*B(3);A(1)*B(2)-A(2)*B(1)];
end

```

First, we set up the standard equatorial frame unit vectors and find the angular momentum vector.

Listing 9.13: **Find Angular Momentum Vector**

```
% set standard I, J, K vectors
2 I = [1;0;0]; J = [0;1;0]; K = [0;0;1];
R0 = norm(r0);
V0 = norm(v0);
H = CrossProduct(r0,v0);
h = norm(H);
```

Now we can find \mathcal{H}^2/μ , the eccentricity vector and its norm.

Listing 9.14: **Find the eccentricity vector**

```
p = h^2;
% find the eccentricity vector
E = (1/mu)*(V0^2 - mu/R0)*r0 - dot(r0,v0)*v0;
4 % find the eccentricity
e = norm(E);
```

Then find the inclination angle, i , the line of nodes vector n and the longitude of the ascending node, Ω .

Listing 9.15: **Find the inclination, line of nodes and Ω**

```
% find the inclination
inc = acos( dot(H,K) / h );
% find the line of nodes vector
N = CrossProduct(K,H);
5 n = norm(N);
% find Omega
Omega = acos( dot(N,I) / n );
```

Then find the argument of periapsis, ω and the true anomaly, ν_0 .

Listing 9.16: **Find the argument of periapsis and true anomaly**

```
% find omega
omega = acos( dot(N,E) / (n*e) );
3 % find the true anomaly
nu0 = acos( dot(E,r0) / (e*R0) );
```

Now we can get the orbital plane.

Listing 9.17: **Find Perifocal Basis**

```
1 % unit normal to orbital plane
W = H/h;
```

```
% get perigee direction vector in orbital plane
P = E/e;
W = H/h;
6 Q = CrossProduct(W,P);
```

Now get the semimajor axis.

Listing 9.18: **Find the semimajor axis a**

```
rp = (h^2)/(1+e);
ra = (h^2)/(1-e);
a = (rp+ra)/2.0;
```

Next, plot the standard geocentric -equatorial and Perifocal coordinate axes. We start holding the plots now.

Listing 9.19: **Plot The Geocentric and Perifocal Axes**

```
hold on
2 % Draw IJK
[x,y,z] = ThreeDToPlot(ra*I);
plot3(x,y,z,'linewidth',2,'color','black');
[x,y,z] = ThreeDToPlot(ra*J);
plot3(x,y,z,'linewidth',2,'color','black');
7 [x,y,z] = ThreeDToPlot(ra*K);
plot3(x,y,z,'linewidth',2,'color','black');
% Draw PQW
[x,y,z] = ThreeDToPlot(ra*P);
plot3(x,y,z,'linewidth',2,'color','red');
12 [x,y,z] = ThreeDToPlot(-ra*P);
plot3(x,y,z,'linewidth',2,'color','red');
[x,y,z] = ThreeDToPlot(ra*Q);
plot3(x,y,z,'linewidth',2,'color','red');
[x,y,z] = ThreeDToPlot(-ra*Q);
17 plot3(x,y,z,'linewidth',2,'color','red');
[x,y,z] = ThreeDToPlot(ra*W);
plot3(x,y,z,'linewidth',2,'color','red');
```

Draw the line of nodes.

Listing 9.20: **Draw the line of nodes**

```
1 [x,y,z] = ThreeDToPlot(ra*N);
plot3(x,y,z,'linewidth',2,'color','green');
[x,y,z] = ThreeDToPlot(-ra*N);
```

Plot the orbit and stop the hold on plots.

Listing 9.21: **Plot the orbit**

```
% Draw orbit
2 OrbitSize = 80;
rorbitplane=zeros(OrbitSize,3);
nu = linspace(0,2*pi,OrbitSize);
for i = 1:OrbitSize
    c = cos(nu(i)); s = sin(nu(i));
7 rorbit = h^2/(1 + e*c);
rorbitplane(i,:)=rorbit*c*P+rorbit*s*Q;
T = rorbitplane(i,:);
x(i) = T(1);
y(i) = T(2);
12 z(i) = T(3);
end
plot3(x,y,z,'linewidth',2,'color','blue');
%
xlabel('x');
17 ylabel('y');
zlabel('z');
hold off
```

Print out the orbital elements.

Listing 9.22: **Print The Orbital Elements**

```
1 disp(sprintf('Omega: longitude of the ascending node = %8.4f',Omega
*180/pi));
disp(sprintf('omega: argument of periapsis = %8.4f',omega*180/pi));
disp(sprintf('nu0: the true anomaly = %8.4f',nu0*180/pi));
disp(sprintf('inclination = %8.4f',inc*180/pi));
disp(sprintf('a: semimajor axis = %8.4f',a));
6 disp(sprintf('p = h^2/mu = %8.4f',p));
disp(sprintf('e: the eccentricity = %8.4f',e));
disp(sprintf('h: the angular momentum = %8.4f',h));
```

The full code is below. With some practice, you can read code documented this way pretty easily. Notice, we try to add comment lines that are short but informative before most of the things we try to do.

Listing 9.23: **DrawOrbitalPlane.m**

```
function [omega,Omega,nu0,a,p,e,E,inc,P,Q,W,h,H,N] = DrawOrbitalPlane(
r0,v0)
2 %
% here mu is normalized to 1
% r0 is position vector at time 0
% v0 is velocity vector at time 0
% Omega is the longitude of the ascending node
7 % which determines the line of nodes direction
% omega is the argument of periapsis
```

```

%nu0 is the true anomaly
% p = h^2/mu = h^2
% E is the eccentricity vector
12 % N is the line of nodes vector
% P, Q, W is the perifocal basis
% H is angular momentum vector
% N is the line of nodes unit vector
% inc is the inclination
17 %
% set standard I, J, K vectors
clf;
mu = 1;
I = [1;0;0]; J = [0;1;0]; K = [0;0;1];
22 % Find the angular momentum vector
R0 = norm(r0);
V0 = norm(v0);
H = CrossProduct(r0,v0);
h = norm(H);
27 p = h^2;
% find the eccentricity vector
E = (1/mu)*( V0^2 - mu/R0)*r0 - dot(r0,v0)*v0;
% find the eccentricity
e = norm(E);
32 % find the inclination
inc = acos( dot(H,K)/h );
% find the line of nodes vector
N = CrossProduct(K,H);
n = norm(N);
37 % find Omega
Omega = acos( dot(N,I)/n );
% find omega
omega = acos( dot(N,E)/(n*e) );
% find the true anomaly
42 nu0 = acos( dot(E,r0)/(e*R0) );
% to get orbital plane:
% unit normal to orbital plane
W = H/h;
% get perigee direction vector in orbital plane
47 P = E/e;
W = H/h;
Q = CrossProduct(W,P);
rp = (h^2)/(1+e);
ra = (h^2)/(1-e);
52 a = (rp+ra)/2.0;
hold on
% Draw IJK
[x,y,z] = ThreeDToPlot(ra*I);
plot3(x,y,z,'linewidth',2,'color','black');
57 [x,y,z] = ThreeDToPlot(ra*J);
plot3(x,y,z,'linewidth',2,'color','black');
[x,y,z] = ThreeDToPlot(ra*K);
plot3(x,y,z,'linewidth',2,'color','black');
% Draw PQW
62 [x,y,z] = ThreeDToPlot(ra*P);
plot3(x,y,z,'linewidth',2,'color','red');
[x,y,z] = ThreeDToPlot(-ra*P);

```

```

plot3(x,y,z,'linewidth',2,'color','red');
[x,y,z] = ThreeDToPlot(ra*Q);
67 plot3(x,y,z,'linewidth',2,'color','red');
[x,y,z] = ThreeDToPlot(-ra*Q);
plot3(x,y,z,'linewidth',2,'color','red');
[x,y,z] = ThreeDToPlot(ra*W);
plot3(x,y,z,'linewidth',2,'color','red');
72 % Draw line of nodes
[x,y,z] = ThreeDToPlot(ra*N);
plot3(x,y,z,'linewidth',2,'color','green');
[x,y,z] = ThreeDToPlot(-ra*N);
plot3(x,y,z,'linewidth',2,'color','green');
77 % Draw orbit
OrbitSize = 80;
rorbitplane=zeros(OrbitSize,3);
nu = linspace(0,2*pi,OrbitSize);
for i = 1:OrbitSize
82 c = cos(nu(i)); s = sin(nu(i));
rorbit = h^2/(1 + e*c);
rorbitplane(i,:) = rorbit*c*P+rorbit*s*Q;
T = rorbitplane(i,:);
x(i) = T(1);
y(i) = T(2);
z(i) = T(3);
end
plot3(x,y,z,'linewidth',2,'color','blue');
%
92 xlabel('x');
ylabel('y');
zlabel('z');
axis on
grid on
97 box on;
hold off
disp(sprintf('Omega: longitude of the ascending node = %8.4f',Omega
*180/pi));
disp(sprintf('omega: argument of periapsis = %8.4f',omega*180/pi));
disp(sprintf('nu0: the true anomaly = %8.4f',nu0*180/pi));
102 disp(sprintf('inclination = %8.4f',inc*180/pi));
disp(sprintf('a: semimajor axis = %8.4f',a));
disp(sprintf('p = h^2/mu = %8.4f',p));
disp(sprintf('e: the eccentricity = %8.4f',e));
disp(sprintf('h: the angular momentum = %8.4f',h));
107 end

```

Running this code generates a graph which you can grab with your mouse and rotate around to get it where you want it. Let's try it.

Listing 9.24: **Runtime**

```

r0 = [1.29;0.75;0.0];
3 1.29900
    0.75000

```

```

0.00000

v0 = [.35;0.61;0.707]
8 v0 =

0.35000
0.61000
0.70700
13 [omega ,Omega,nu0,a,p,e,E,inc ,P,Q,W,h,H,N] = DrawOrbitalPlane (r0 ,v0 );
Omega: longitude of the ascending node = 30.0007
omega: argument of periaxis = 94.9983
nu0: the true anomaly = 94.9983
inclination = 63.4500
18 a: semimajor axis = 2.9506
p = h^2/mu = 1.4054
e: the eccentricity = 0.7237
h: the angular momentum = 1.1855

```

You can see our chosen snapshot of this graph in Figure 9.14. The line of nodes is plotted in green, the P Q and W perifocal coordinate system is plotted in red and the geocentric - equatorial I J K coordinate system in black. It is hard to get a nice plot of this stuff, but as you can see with a bit of work we can generate these automatically.

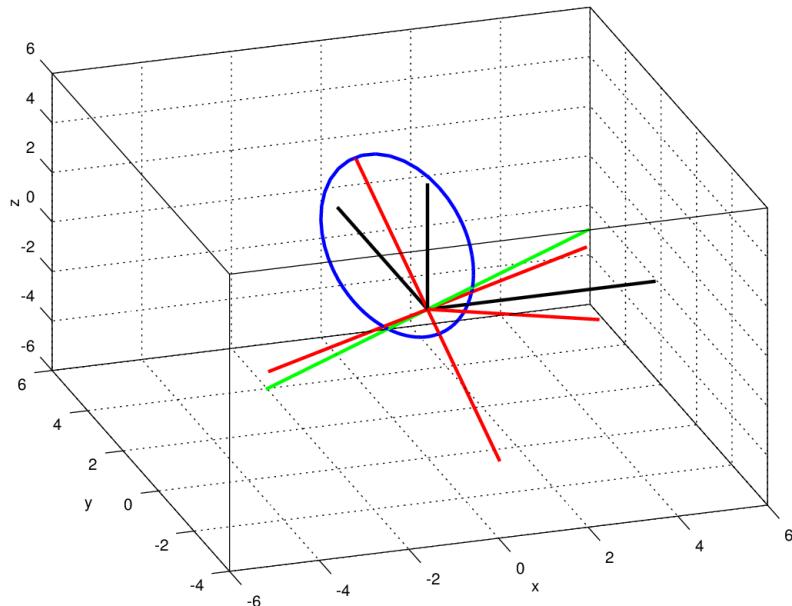


Figure 9.14: the orbit for $r_0 = [1.299; 0.75; 0.0]$, $v_0 = [0.35; 0.61; 0.707]$

Here is another one:

Listing 9.25: **Runtime**

```
r0 = [1.29;0.75;1.2];
```

9.5. DRAWING ORBITAL PLANE IN MATLAB

159

```

1.29900
4      0.75000
           1.20000

v0 = [.35;0.61;0.707]
v0 =
9      0.35000
           0.61000
           0.70700
[omega ,Omega ,nu0 ,a ,p ,e ,E ,inc ,P,Q,W,h,H,N] = DrawOrbitalPlane(r0 ,v0 );
14 Omega: longitude of the ascending node = 22.2954
omega: argument of periapsis = 74.2970
nu0: the true anomaly = 135.9555
inclination = 45.4007
a: semimajor axis = 19.9894
19 p = h^2/mu = 0.5578
e: the eccentricity = 0.9859
h: the angular momentum = 0.7469

```

The resulting orbit is seen in Figure 9.15. However, you can't see where the orbit crosses the equatorial plane very well because $r_p = 0.2827$ with $r_1 = 39.698$ and so the orbit just looks like it starts at the origin.

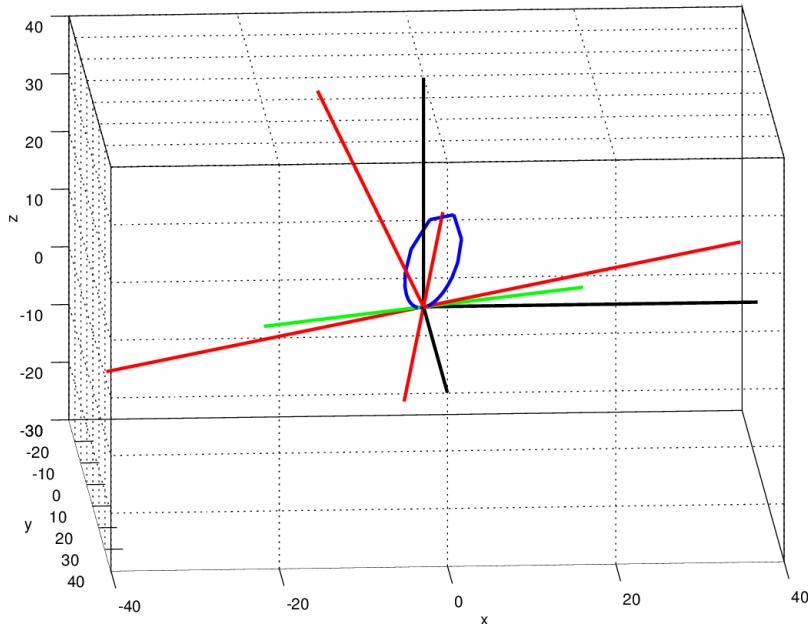


Figure 9.15: the orbit for $r_0 = [1.299; 0.75; 1.2]$, $v_0 = [0.35; 0.61; 0.707]$

9.5.1 Homework

Exercise 9.5.1

Exercise 9.5.2

Exercise 9.5.3

Exercise 9.5.4

Exercise 9.5.5

Chapter 10

Determinants and Matrix Manipulations

It is time for you to look at a familiar tool more carefully. Behold, determinants!

10.1 Determinants

We have already found the determinant of a 6×6 coefficient matrix that arose from a model of LRRK2 mutations. We know you have all done such calculations before so there was nothing new there. But you probably have not discussed the determinant very carefully which is what we will do now. We start with the determinant of a 3×3 matrix to set the stage. We let \mathbf{A} be defined by

$$\mathbf{A} = \begin{bmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{bmatrix}$$

Now think about the matrix \mathbf{A} as made up of the row vectors. Let these be denote by \mathbf{V}_1 , \mathbf{V}_2 and \mathbf{V}_3 . We want a determinant function that acts on these three arguments and gives us back a real number. For the moment, call this function Φ . Then $\Phi(\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3)$ is a number and we want Φ to satisfy the properties our usual 2×2 determinant function does. Let \mathbf{B} be a 2×2 matrix with row vectors \mathbf{W}_1 and \mathbf{W}_2 which also has the form

$$\mathbf{B} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

We know $\det(\mathbf{B}) = W_{11}W_{22} - W_{12}W_{21}$. So another way of saying this is

$$\det(\mathbf{W}_1, \mathbf{W}_2) = W_{11}W_{22} - W_{12}W_{21}$$

What basic properties does our familiar determinant have?

P1: det is homogeneous: that is, *constants come out* : What if we multiply a row by a constant c ?
Note

$$\det(\mathbf{W}_1, c\mathbf{W}_2) = W_{11}cW_{22} - W_{12}cW_{21} = c\det(\mathbf{W}_1, \mathbf{W}_2).$$

A similar calculation shows that $\det(c\mathbf{W}_1, \mathbf{W}_2) = c\det(\mathbf{W}_1, \mathbf{W}_2)$.

P2: **det doesn't change if we add rows** : What is we add row two to row one? We have

$$\begin{aligned}\det(\mathbf{W}_1 + \mathbf{W}_2, \mathbf{W}_2) &= (W_{11} + W_{21})W_{22} - (W_{12} + W_{22})W_{21} = W_{11}cW_{22} - W_{12}cW_{21} \\ &\det(\mathbf{W}_1, \mathbf{W}_2).\end{aligned}$$

A similar computation shows that adding row one to row two doesn't change the determinant.

P3: The determinant of the identity is one : this is an easy computation; $\det(\mathbf{I}) = 1 \times 1 = 1$.

Let's assume we can find a function Φ defined on $n \times n$ matrices which we will think of as a function defined on n row vectors of size n that satisfies Properties **P1**, **P2** and **P3**. Let the row vectors now be \mathbf{W}_1 through \mathbf{W}_n . We assume Φ satisfies

P1: Φ is homogeneous: that is, constants come out of slots :

$$\Phi(\mathbf{W}_1, \mathbf{W}_2, \dots, c\mathbf{W}_i, \dots, \mathbf{W}_n) = c\Phi(\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_n). \quad (10.1)$$

P2: Φ doesn't change if we add rows : As long as $j \neq i$,

$$\Phi(\mathbf{W}_1, \dots, \mathbf{W}_i + \mathbf{W}_j, \dots, \mathbf{W}_n) = \Phi(\mathbf{W}_1, \dots, \mathbf{W}_n). \quad (10.2)$$

P3 :

$$\Phi(\mathbf{I}) = 1. \quad (10.3)$$

10.1.1 Consequences One

Using these properties, we can figure out that others hold.

P4: If a row is zero, the Φ is zero : This follows from Property 10.1 by just setting $c = 0$.

If two rows are the same, the Φ is zero : This follows from Property 10.2. For convenience assume row one and row two are equal. We have

$$\begin{aligned}\Phi(\mathbf{W}_1, \mathbf{W}_1, \mathbf{W}_3, \dots, \mathbf{W}_n) &= \Phi(\mathbf{W}_1 - \mathbf{W}_1, \mathbf{W}_1, \mathbf{W}_3, \dots, \mathbf{W}_n) \\ &= \Phi(0, \mathbf{W}_1, \mathbf{W}_3, \dots, \mathbf{W}_n) = 0\end{aligned}$$

P5: Replacing a row by adding a multiple of another row doesn't change the determinant ; This uses our properties: the case $c = 0$ is obvious so let's just assume $c \neq 0$: using Property 10.1, we have for $j \neq i$,

$$\Phi(\mathbf{W}_1, \dots, \mathbf{W}_i + c\mathbf{W}_j, \dots, \mathbf{W}_n) = (1/c) \Phi(\mathbf{W}_1, \dots, \mathbf{W}_i + c\mathbf{W}_j, \dots, c\mathbf{W}_j, \dots, \mathbf{W}_n)$$

But we now have a matrix whose j^{th} row is $c\mathbf{W}_j$. Hence, using Property 10.2 adding that row to row i does not change the value of Φ . So we have

$$\Phi(\mathbf{W}_1, \dots, \mathbf{W}_i + c\mathbf{W}_j, \dots, \mathbf{W}_n) = (1/c) \Phi(\mathbf{W}_1, \dots, \mathbf{W}_i + c\mathbf{W}_j, \dots, c\mathbf{W}_j, \dots, \mathbf{W}_n)$$

$$\begin{aligned} &= (1/c) \Phi(\mathbf{W}_1, \dots, \mathbf{W}_i, \dots, c\mathbf{W}_j, \dots, \mathbf{W}_n) \\ &= \Phi(\mathbf{W}_1, \dots, \mathbf{W}_i, \dots, \mathbf{W}_j, \dots, \mathbf{W}_n) \end{aligned}$$

where we use Property 10.1 again.

P6: If the rows of B are linearly dependent, then Φ is zero : Let's do this for the case that \mathbf{W}_1 can be written in terms of the other $n - 1$ vectors. You'll get the gist of the argument and then you'll be able to see in your mind how to handle the other cases. In this case, we can say $\mathbf{W}_1 = \sum_{i=2}^n c_j \mathbf{W}_j$ for not all zero constants c_j . To save typing, we'll just write Λ_n for the arguments \mathbf{W}_2 through \mathbf{W}_n , Λ_{n-1} for the arguments \mathbf{W}_2 through \mathbf{W}_{n-1} and so forth. So using our properties, we find

$$\begin{aligned} \Phi(\mathbf{W}_1, \Lambda_n) &= \Phi\left(\sum_{i=2}^n c_j \mathbf{W}_j, \Lambda_n\right) \\ &= (1/c_n) \Phi\left(\sum_{i=2}^{n-1} c_j \mathbf{W}_j + c_n \mathbf{W}_n, \Lambda_{n-1}, c_n \mathbf{W}_n\right) \\ &= \Phi\left(\sum_{i=2}^{n-1} c_j \mathbf{W}_j, \Lambda_{n-1}, \mathbf{W}_n\right) \end{aligned}$$

We can do this over and over again, until we *peel* off all the terms of the summand in the first slot to obtain

$$\begin{aligned} \Phi(\mathbf{W}_1, \Lambda) &= \Phi(\mathbf{W}_2, \mathbf{W}_2, \dots, \mathbf{W}_n) \\ &= \Phi(\mathbf{W}_2 - \mathbf{W}_2, \mathbf{W}_2, \dots, \mathbf{W}_n) \\ &= \Phi(\mathbf{0}, \mathbf{W}_2, \dots, \mathbf{W}_n) = 0. \end{aligned}$$

P7: Φ is linear in each argument ; What if have the sum of two vectors in row one, $\alpha + \beta$? We will use the same Λ_n notation as before to save typing. There are lots of cases here, so we'll just do one and let you think about the others. We will assume the vectors \mathbf{W}_2 through \mathbf{W}_n are linearly independent. Add another vector, γ to them to make a basis and write down the expansions of both α and β in them. $\alpha = c\gamma + \sum_{j=2}^n c_j \mathbf{W}_j$ and $\beta = d\gamma + \sum_{j=2}^n d_j \mathbf{W}_j$. Then using our properties,

$$\begin{aligned} \Phi(\alpha + \beta, \Lambda_n) &= \Phi\left((c\gamma + \sum_{i=2}^n c_j \mathbf{W}_j) + (d\gamma + \sum_{j=2}^n d_j \mathbf{W}_j), \Lambda_n\right) \\ &= \Phi\left((c\gamma + \sum_{i=2}^n c_j \mathbf{W}_j) - c_n \mathbf{W}_n + (d\gamma + \sum_{j=2}^n d_j \mathbf{W}_j) - d_n \mathbf{W}_n, \Lambda_n\right) \end{aligned}$$

We then do this again and again; after doing this from $j = n - 1$ to $j = 2$, we have

$$\Phi(\alpha + \beta, \Lambda_n) = \Phi((c+d)\gamma, \Lambda_n) = (c+d) \Phi(\gamma, \Lambda_n)$$

Now working the other way, we have

$$\Phi(\alpha, \Lambda_n) + \Phi(\beta, \Lambda_n) = \Phi(c\gamma + \sum_{i=2}^n c_j \mathbf{W}_j, \Lambda_n) + \Phi(d\gamma + \sum_{j=2}^n d_j \mathbf{W}_j, \Lambda_n)$$

$$\begin{aligned} &= \Phi(c\gamma, \Lambda_n) + \Phi(d\gamma, \Lambda_n) \\ &= \Phi((c+d)\gamma, \Lambda_n) = (c+d) \Phi(\gamma, \Lambda_n) \end{aligned}$$

These match and so we are done with this case. The other cases are similar. If \mathbf{W}_2 through \mathbf{W}_n are dependent, the argument changes and so forth. We leave those details to your inquiring mind. We suggest a nice tall glass of stout will help with the pain.

P8: Interchanging two rows changes the sign of Φ : Let's do this for rows one and two to keep it simple. We have

$$\begin{aligned} \Phi(\mathbf{W}_1, \mathbf{W}_2, \Lambda_{n-1}) &= \Phi(\mathbf{W}_1, \mathbf{W}_1 + \mathbf{W}_2, \Lambda_{n-1}) \\ &= \Phi(\mathbf{W}_1 - (\mathbf{W}_1 + \mathbf{W}_2), \mathbf{W}_1 + \mathbf{W}_2, \Lambda_{n-1}) \\ &= \Phi(-\mathbf{W}_2, \mathbf{W}_1 + \mathbf{W}_2, \Lambda_{n-1}) \\ &= -\Phi(\mathbf{W}_2, \mathbf{W}_1 + \mathbf{W}_2, \Lambda_{n-1}) = -\Phi(\mathbf{W}_2, \mathbf{W}_1, \Lambda_{n-1}). \end{aligned}$$

Similar arguments, but messier notationally, work for the other rows. This implies a nice result. If we have a matrix whose rows are in the order (j_1, j_2, \dots, j_n) , then if it takes an even number of row swaps to get back to the order $(1, 2, \dots, n)$, then Φ doesn't change sign. But if it takes an odd number of interchanges, the sign flips.

10.1.2 Homework

Exercise 10.1.1

Exercise 10.1.2

Exercise 10.1.3

Exercise 10.1.4

Exercise 10.1.5

10.1.3 Consequences Two

Our three fundamental assumptions about Φ have additional consequences. We still don't know such a function Φ exists although we do know we have a nice candidate, \det , that works on 2×2 matrices. Let's work out some more consequences.

P9: $\Phi(\mathbf{AB} = \Phi(\mathbf{A}) \Phi(\mathbf{B})$: This is a really useful thing. Let \mathbf{A} have rows \mathbf{V}_1 through \mathbf{V}_n and \mathbf{B} have rows \mathbf{W}_1 through \mathbf{W}_n . Let \mathbf{Y}_i be the columns of \mathbf{B} . The usual matrix multiplication gives

$$\mathbf{AB} = \begin{bmatrix} <\mathbf{V}_1, \mathbf{Y}_1> & \dots & <\mathbf{V}_1, \mathbf{Y}_n> \\ \vdots & \vdots & \vdots \\ <\mathbf{V}_n, \mathbf{Y}_1> & \dots & <\mathbf{V}_n, \mathbf{Y}_n> \end{bmatrix}$$

Each row of the product \mathbf{AB} has a particular form. We'll show this for row one and you can figure out how to do it for the other rows from that. Do this on paper for a 3×3 and you'll see the pattern.

$$[<\mathbf{V}_1, \mathbf{Y}_1> \dots <\mathbf{V}_1, \mathbf{Y}_n>] = V_{11}\mathbf{W}_1 + \dots + V_{1n}\mathbf{W}_n.$$

Call this row vector \mathbf{C}_1 . We can do a similar expansion for the other rows of the product matrix. Hence, we want to calculate $\Phi(\mathbf{C}_1, \dots, \mathbf{C}_n)$. We have since Φ is additive

$$\begin{aligned}\Phi(\mathbf{C}_1, \dots, \mathbf{C}_n) &= \Phi(V_{11} \mathbf{W}_1 + \dots + V_{1n} \mathbf{W}_n, \mathbf{C}_2, \dots, \mathbf{C}_n) \\ &= \sum_{j_1=1}^n V_{1,j_1} \Phi(\mathbf{W}_{j_1}, \mathbf{C}_2, \dots, \mathbf{C}_n)\end{aligned}$$

Now do this again for \mathbf{C}_2 . We have

$$\begin{aligned}\Phi(\mathbf{W}_{j_1}, \mathbf{C}_2, \dots, \mathbf{C}_n) &= \Phi\left(\mathbf{W}_{j_1}, \sum_{j_2=1}^n V_{2,j_2} \mathbf{W}_{j_2}, \mathbf{C}_3, \dots, \mathbf{C}_n\right) \\ &= \sum_{j_2=1}^n V_{2,j_2} \Phi(\mathbf{W}_{j_1}, \mathbf{W}_{j_2}, \mathbf{C}_3, \dots, \mathbf{C}_n)\end{aligned}$$

So after two steps, we have

$$\Phi(\mathbf{C}_1, \dots, \mathbf{C}_n) = \sum_{j_1=1}^n V_{1,j_1} \sum_{j_2=1}^n V_{2,j_2} \Phi(\mathbf{W}_{j_1}, \mathbf{W}_{j_2}, \mathbf{C}_3, \dots, \mathbf{C}_n)$$

Whew! Now finish up by handling the other rows. We obtain

$$\begin{aligned}\Phi(\mathbf{C}_1, \dots, \mathbf{C}_n) &= \sum_{j_1=1}^n V_{1,j_1} \sum_{j_2=1}^n V_{2,j_2} \dots \sum_{j_n=1}^n V_{n,j_n} \Phi(\mathbf{W}_{j_1}, \mathbf{W}_{j_2}, \dots, \mathbf{W}_{j_n}) \\ &= \sum_{j_1=1}^n \dots \sum_{j_n=1}^n V_{1,j_1} \dots V_{n,j_n} \Phi(\mathbf{W}_{j_1}, \mathbf{W}_{j_2}, \dots, \mathbf{W}_{j_n}).\end{aligned}$$

Now many of these terms are zero as any time two rows match, Φ is zero. So we can say

$$\Phi(\mathbf{C}_1, \dots, \mathbf{C}_n) = \sum_{j_1, j_2, \dots, j_n=1; j_1 \neq j_2 \dots \neq j_n}^n V_{1,j_1} \dots V_{n,j_n} \Phi(\mathbf{W}_{j_1}, \mathbf{W}_{j_2}, \dots, \mathbf{W}_{j_n}).$$

Next, let $\epsilon_{j_1, \dots, j_n}$ denote the -1 or $+1$ associated with the row swapping necessary to move $\{j_1, \dots, j_n\}$ to $\{1, \dots, n\}$. Then, we can rewrite again as

$$\Phi(\mathbf{C}_1, \dots, \mathbf{C}_n) = \sum_{j_1, j_2, \dots, j_n=1; j_1 \neq j_2 \dots \neq j_n}^n \epsilon_{j_1, \dots, j_n} V_{1,j_1} \dots V_{n,j_n} \Phi(\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_n).$$

In particular, we can let $\mathbf{B} = \mathbf{I}$. Then we have $\mathbf{AB} = \mathbf{AI} = \mathbf{A}$ and

$$\begin{aligned}\Phi(\mathbf{V}_1, \dots, \mathbf{V}_n) &= \sum_{j_1, j_2, \dots, j_n=1; j_1 \neq j_2 \dots \neq j_n}^n \epsilon_{j_1, \dots, j_n} V_{1,j_1} \dots V_{n,j_n} \Phi(\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_n) \\ &= \sum_{j_1, j_2, \dots, j_n=1; j_1 \neq j_2 \dots \neq j_n}^n \epsilon_{j_1, \dots, j_n} V_{1,j_1} \dots V_{n,j_n}\end{aligned}$$

since $\Phi(\mathbf{I}) = 1$. We usually just refer to $\Phi(V_1, \dots, V_n)$ as $\Phi(A)$, so we have the identity

$$\Phi(A) = \sum_{j_1, j_2, \dots, j_n=1; j_1 \neq j_2 \dots \neq j_n}^n \epsilon_{j_1, \dots, j_n} V_{1,j_1} \dots V_{n,j_n}$$

Combining, we have shown $\Phi(AB) = \Phi(A)\Phi(B)$ since $\Phi(W_1, W_2, \dots, W_n) = \Phi(B)$.

We have gotten pretty far in our discussions. We now know that if a function Φ satisfies Properties **P1**, **P2** and **P3**, it must have the value

$$\Phi(A) = \sum_{j_1, j_2, \dots, j_n=1; j_1 \neq j_2 \dots \neq j_n}^n \epsilon_{j_1, \dots, j_n} V_{1,j_1} \dots V_{n,j_n}$$

This leads to our next consequence,

P10, $\Phi(A) = \Phi(A^T)$ ∵ Define a new function $\Psi(A) = \Phi(A^T)$. If we multiply a row in A^T by the constant c , this is the same as multiplying A^T by the matrix

$$R(c) = D(1, 0, \dots, c, 1, \dots, 1)$$

where D is the diagonal matrix with ones on the main diagonal except in the (i, i) position which has the value c . We can use Property **P9** now.

$$\Psi(R(c)A) = \Phi(R(c)A^T) = \Phi(R(c))\Phi(A^T) = c\Phi(A^T) = c\Psi(A)$$

and so Ψ satisfies **P1**. Next, adding row j to row i corresponds to multiplying A^T by

$$R = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 1 \text{ position (i,i)} & 0 & \dots & 0 & 1 \text{ position (i,j)} & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 1 \end{bmatrix}$$

Then

$$\Psi(RA) = \Phi(RA^T) = \Phi(R)\Phi(A^T) = 1\Phi(A^T) = \Psi(A)$$

and so Property **P2** holds. Finally, $\Psi(I) = \Phi(I^T) = \Phi(I) = 1$ and so Property **P3** holds as well. Since all three properties hold, we must also have

$$\Psi(A) = \sum_{j_1, j_2, \dots, j_n=1; j_1 \neq j_2 \dots \neq j_n}^n \epsilon_{j_1, \dots, j_n} V_{1,j_1} \dots V_{n,j_n} = \Phi(A).$$

We conclude $\Phi(A) = \Phi(A^T)$. This also implies another really important result: **everything we have said about row manipulations applies equally to column manipulations**.

We could use the formula above as our definition of how to find Φ . However, there is an easier way. The proper way to do this would be to define how to calculate a **determinant** using what are called cofactors in a recursive fashion. Let's go through how this works for a 3×3 . We have

$$\begin{aligned}\Phi(\mathbf{A}) &= \sum_{i,j,k=1; i \neq j \neq k}^3 \epsilon_{i,j,k} V_{1i} V_{2j} V_{3k} \\ &= \epsilon_{123} V_{11} V_{22} V_{33} + \epsilon_{132} V_{11} V_{23} V_{32} + \epsilon_{213} V_{12} V_{21} V_{33} \\ &\quad + \epsilon_{231} V_{12} V_{23} V_{31} + \epsilon_{312} V_{13} V_{21} V_{32} + \epsilon_{321} V_{13} V_{22} V_{31}.\end{aligned}$$

Now, the ϵ terms refer to the sign (either 1 or -1) associated with the rows being in that order. Remember an odd number of row interchanges changes Φ by -1 and an even number of row interchanges leaves the sign unchanged.

- $\epsilon_{123} = 1$, because rows are already in proper order.
- $\epsilon_{132} = -1$, because switching 3 with 2 gets 123 order.
- $\epsilon_{213} = -1$, because switching 2 with 1 gets 123 order.
- $\epsilon_{231} = 1$, because switching 3 with 1 gets 213 order; then 1 with 2 gives 123.
- $\epsilon_{312} = 1$, because switching 3 with 1 gets 132 order; then 3 with 2 gives 123.
- $\epsilon_{321} = -1$, because switching 3 with 1 gets 123 order.

Using these, we have

$$\begin{aligned}\Phi(\mathbf{A}) &= V_{11} V_{22} V_{33} - V_{11} V_{23} V_{32} - V_{12} V_{21} V_{33} \\ &\quad + V_{12} V_{23} V_{31} + V_{13} V_{21} V_{32} - V_{13} V_{22} V_{31}.\end{aligned}$$

Now choose any row or column you want: we will use column 3. We will organize these calculations around the entries of column 3. We write

$$\begin{aligned}\Phi(\mathbf{A}) &= V_{13} (V_{21} V_{32} - V_{22} V_{31}) + V_{23} (V_{12} V_{31} - V_{11} V_{32}) + V_{33} (V_{11} V_{22} - V_{12} V_{21}) \\ &= V_{13} \det \begin{pmatrix} V_{21} & V_{22} \\ V_{31} & V_{32} \end{pmatrix} - V_{23} \det \begin{pmatrix} V_{11} & V_{12} \\ V_{31} & V_{32} \end{pmatrix} + V_{33} \det \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \\ &= (-1)^{1+3} V_{13} \det \begin{pmatrix} V_{21} & V_{22} \\ V_{31} & V_{32} \end{pmatrix} + (-1)^{2+3} V_{23} \det \begin{pmatrix} V_{11} & V_{12} \\ V_{31} & V_{32} \end{pmatrix} \\ &\quad + (-1)^{3+3} V_{33} \det \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}.\end{aligned}$$

Now look at the matrix \mathbf{A} with column 3 singled out.

$$\mathbf{A} = \begin{bmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{bmatrix}$$

Look at the entry V_{13} and mentally cross out row 1 and column 3 from the matrix. This leaves a 2×2 submatrix which is identical to the one after the V_{13} entry above. Call this the **cofactor** C_{13} . Next, look at V_{23} and mentally cross out row 2 and column 3 from the matrix. This leaves a 2×2 submatrix which is identical to the one after the V_{23} entry above. Call this the **cofactor** C_{23} . Finally, look at V_{33} and mentally cross out row 3 and column 3 from the matrix. This leaves a 2×2 submatrix which is identical to the one after the V_{33} entry above. Call this the **cofactor** C_{33} . Hence, we can rewrite our computation as

$$\Phi(\mathbf{A}) = (-1)^{1+3} V_{13} \det \begin{pmatrix} V_{21} & V_{22} \\ V_{31} & V_{32} \end{pmatrix} + (-1)^{2+3} V_{23} \det \begin{pmatrix} V_{11} & V_{12} \\ V_{31} & V_{32} \end{pmatrix}$$

$$\begin{aligned}
 & +(-1)^{3+3} V_{33} \det \left(\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \right) \\
 & = (-1)^{1+3} V_{13} \det(\mathbf{C}_{13}) + (-1)^{2+3} V_{23} \det(\mathbf{C}_{23}) + (-1)^{3+3} V_{33} \det(\mathbf{C}_{33})
 \end{aligned}$$

What we have done here is to expand the computation of Φ using column 3 of the matrix. It is straightforward to choose a different row or column and do the same thing. The cofactors are defined similarly and we would find for a row expansion along row r

$$\Phi(\mathbf{A}) = (-1)^{r+1} V_{r1} \det(\mathbf{C}_{r1}) + (-1)^{r+2} V_{r2} \det(\mathbf{C}_{r2}) + (-1)^{r+3} V_{r3} \det(\mathbf{C}_{r3})$$

and for a column expansion along column c

$$\Phi(\mathbf{A}) = (-1)^{1+c} V_{1c} \det(\mathbf{C}_{1c}) + (-1)^{2+c} V_{2c} \det(\mathbf{C}_{2c}) + (-1)^{3+c} V_{3c} \det(\mathbf{C}_{3c})$$

This suggests Φ should be defined **recursively** starting with 3×3 matrices using the usual 2×2 determinant \det . If \mathbf{A} was 4×4 , we would define the determinant of a 3×3 matrix as above and define the determinant $\det(\mathbf{A})$ as follows: for a row expansion along row r

$$\begin{aligned}
 \det(\mathbf{A}) = & (-1)^{r+1} V_{r1} \det(\mathbf{C}_{r1}) + (-1)^{r+2} V_{r2} \det(\mathbf{C}_{r2}) \\
 & + (-1)^{r+3} V_{r3} \det(\mathbf{C}_{r3}) + (-1)^{r+4} V_{r4} \det(\mathbf{C}_{r4})
 \end{aligned}$$

where the \det 's are the 3×3 we defined in the previous step and the cofactors are now 3×3 submatrices. For a column expansion along column c , we have

$$\begin{aligned}
 \Phi(\mathbf{A}) = & (-1)^{1+c} V_{1c} \det(\mathbf{C}_{1c}) + (-1)^{2+c} V_{2c} \det(\mathbf{C}_{2c}) \\
 & + (-1)^{3+c} V_{3c} \det(\mathbf{C}_{3c}) + (-1)^{4+c} V_{4c} \det(\mathbf{C}_{4c}).
 \end{aligned}$$

It is clear we can continue to do this as the size of \mathbf{A} increases. Hence, instead of using the computational formula $\Phi(\mathbf{A}) = \sum_{j_1, j_2, \dots, j_n=1; j_1 \neq j_2 \dots \neq j_n}^n \epsilon_{j_1, \dots, j_n} V_{1, j_1} \dots V_{n, j_n}$ we define Φ recursively as the scheme above indicates. If you call the function defined by this recursive scheme Θ , we can prove it also satisfies Properties **P1**, **P2** and **P3** using what is called an induction argument. This argument works by assuming the stuff we have defined at step n (i.e. for $n \times n$ matrices) satisfies Properties **P1**, **P2** and **P3** and then showing the new version for matrices that are size $(n+1) \times (n+1)$ also satisfies the properties. Then Θ satisfies the usual Φ properties and so Θ and Φ must be the same function. As you can see, determinants of higher order matrices involve a lot of arithmetic! From this point on, we will simple call Φ by the name \det !! Let's state what we know for posterity.

Theorem 10.1.1 The Determinant

There is a unique real valued function \det defined on $n \times n$ matrices which satisfies the following properties:

P1: *The value of the \det is unchanged if a row or a column is multiplied by a constant c .*

P2: *The value of the \det is unchanged if row a is replaced by the new row: row a plus row b as long as $a \neq b$. This is also true for columns.*

P3: *The determinant of the identity matrix is 1.*

Proof 10.1.1

We had a very long discussion about this. We showed that there is only one function satisfying these three properties. ■

These three properties imply many others which we have discussed. Here they are in a more compact form.

Theorem 10.1.2 The Determinant Properties

The real valued function \det defined on $n \times n$ matrices which satisfies the following additional properties:

P4: *If a row or column is zero, then \det is zero.*

P5: *If row a is replaced by row $a + c$ row b , \det is unchanged. The same is true this is done to columns.*

P6: *If the rows or columns are linearly dependent, then \det is zero.*

P7: *\det is linear in each row or column slot.*

P8: *Interchanging two rows or columns changes the sign of \det .*

P9: $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$.

P10: $\det(\mathbf{A}) = \det(\mathbf{A}^T)$

Proof 10.1.2

We used the determinant's three properties to prove each of these additional ones in the arguments presented earlier. ■

Property **P9** is very important. From it, we derive the fundamental formula for the value of the \det . Recall this is $\Phi(\mathbf{A}) = \sum_{j_1, j_2, \dots, j_n=1; j_1 \neq j_2 \dots \neq j_n}^n \epsilon_{j_1, \dots, j_n} V_{1, j_1} \dots V_{n, j_n}$. From this, we know immediately some very important things which we state as another theorem.

Theorem 10.1.3 The Determinant Smoothness

The real valued function \det defined on $n \times n$ matrices is a continuous function of its n^2 parameters. Moreover, its partials of all orders exist and are continuous.

Proof 10.1.3

The value of the determinant is just a large polynomial in n^2 variables. Hence, the smoothness follows. ■

Finally, we have a recursive algorithm for calculating the determinant of a $n \times n$ matrix.

Theorem 10.1.4 The Determinant Algorithm

We can calculate the \det using a row expansion:

$$\det(\mathbf{A}) = \sum_{j=1}^n (-1)^{i+j} A_{rj} \det(\mathbf{C}_{rj}).$$

or via a column expansion:

$$\Phi(\mathbf{A}) = \sum_{j=1}^n (-1)^{i+j} A_{jc} \det(\mathbf{C}_{jc}).$$

Proof 10.1.4

We showed this explicitly for a 4×4 matrix earlier. This is just a generalization of that formula. ■

Comment 10.1.1 This calculation is expensive. A 2×2 takes 2 multiples and 2 adds. A 3×3 has 3 2×2 determinants in its computation. Hence, we have 6 multiplies and 6 adds for the determinants and then 3 more multiplies for the row or column entries and then 3 more for the -1 calculations. So, the total is 12 multiplies and 6 adds. What about a 4×4 matrix? We have 4 submatrix 3×3 determinants which gives 48 multiplies and 24 adds for them. Then the row or column entries give an addition 8 multiplies. The total is then 56 multiplies and 24 adds. So far, letting M denote multiplies and A denote adds:

$2 \times 2 : 2M$ and $2A$ which is larger than $2!$

$3 \times 3 : 12M$ and $6A$ which is larger than $3!$.

$4 \times 4 : 56M$ and $24A$ which is larger than $4!$.

Let's look at the 5×5 case. We have 5 4×4 submatrix determinants which gives $280M$ and $120A$. We have an addition 2×5 multiplies for the row or column entries. The total is thus $290M$ and $120A$ which is larger than $5!$. Thus, as the size of the matrix increases, the number of arithmetic operations to calculate a $n \times n$ determinant is larger than $n!$.

10.1.4 Homework

Exercise 10.1.6

Exercise 10.1.7

Exercise 10.1.8

Exercise 10.1.9

Exercise 10.1.10

10.2 Manipulating Matrices

To understand our problem with finding conditions to check to see if a matrix is positive and negative definite, we need to dig deeper into how a matrix and another matrix which we have been altering by row combinations are related. To begin, we focus on what are called **elementary row operations**.

10.2.1 Elementary Row Operations and Determinants

Let's try to solve the system

$$\begin{bmatrix} 1 & -2 & 3 \\ -2 & 5 & 7 \\ 4 & 3 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix}$$

We can manipulate this by adding multiples of one equation to another and by doing this sort of thing systematically, we can solve the problem. This is usually done much less formally when the code to perform an LU decomposition of the matrix A is discussed. But here we want to make some new points, so bear with us! These sorts of manipulations don't really need us to write down the x , y and z variables, so it is customary to make up a new matrix find taking the coefficient matrix A above and

adding an extra column to it which consists of the data \mathbf{b} . This new matrix is called the **augmented matrix**. Let's call it \mathbf{B} . Then, we have

$$\mathbf{B} = \left[\begin{bmatrix} 1 & -2 & 3 \\ -2 & 5 & 7 \\ 4 & 3 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix} \right]$$

Now if we multiple the original first equation by 2 and add it to the second equation, we get the new second equation

$$\begin{aligned} (2 - 2)x + (-4 + 5)y + (6 + 7)z &= 4 + 5 \\ y + 13z &= 9 \end{aligned}$$

This is the same as multiplying the coefficient matrix \mathbf{A} by the matrix $\mathbf{R}_{21}(2)$

$$\mathbf{R}_{21}(2) = \begin{bmatrix} 0 & 1 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which is read as replace row 2 of \mathbf{A} by 2 times row 1 + row 2. We see

$$\mathbf{R}_{21}(2) \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 & 3 \\ -2 & 5 & 7 \\ 4 & 3 & -1 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 3 \\ 0 & 1 & 13 \\ 4 & 3 & -1 \end{bmatrix}$$

We do this to the data also giving

$$\mathbf{R}_{21}(2) \mathbf{b} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 2 \\ 9 \\ 8 \end{bmatrix}$$

Hence, to save time, we can abuse notation and think of applying $\mathbf{R}_{21}(2)$ to the augmented matrix \mathbf{B} (just apply $\mathbf{R}_{21}(2)$ to the fourth column separately!) giving

$$\mathbf{R}_{21}(2) \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \left[\begin{bmatrix} 1 & -2 & 3 \\ -2 & 5 & 7 \\ 4 & 3 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix} \right] = \left[\begin{bmatrix} 1 & -2 & 3 \\ 0 & 1 & 13 \\ 4 & 3 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 9 \\ 8 \end{bmatrix} \right]$$

Now apply a new transformation, $\mathbf{R}_{31}(-4)$ to the new augmented matrix.

$$\mathbf{R}_{31}(-4) \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix} \left[\begin{bmatrix} 1 & -2 & 3 \\ 0 & 1 & 13 \\ 4 & 3 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 9 \\ 8 \end{bmatrix} \right] = \left[\begin{bmatrix} 1 & -2 & 3 \\ 0 & 1 & 13 \\ 0 & 11 & -13 \end{bmatrix} \begin{bmatrix} 2 \\ 9 \\ 0 \end{bmatrix} \right]$$

Note we have *zeroed* out all of the entries below the A_{11} position. Now let's work on column two. Apply the transformation $\mathbf{R}_{32}(-11)$. We have

$$\mathbf{R}_{32}(-11) \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -11 & 1 \end{bmatrix} \left[\begin{bmatrix} 1 & -2 & 3 \\ 0 & 1 & 13 \\ 0 & 11 & -13 \end{bmatrix} \begin{bmatrix} 2 \\ 9 \\ 0 \end{bmatrix} \right] = \left[\begin{bmatrix} 1 & -2 & 3 \\ 0 & 1 & 13 \\ 0 & 0 & -156 \end{bmatrix} \begin{bmatrix} 2 \\ 9 \\ -99 \end{bmatrix} \right]$$

Next, apply $R_{12}(2)$. We find

$$R_{12}(2) \mathbf{B} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 & 3 \\ 0 & 1 & 13 \\ 0 & 0 & -156 \end{bmatrix} \begin{bmatrix} 2 \\ 9 \\ -99 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 29 \\ 0 & 1 & 13 \\ 0 & 0 & -156 \end{bmatrix} \begin{bmatrix} 20 \\ 9 \\ -99 \end{bmatrix}$$

Now we have *zeroed* out all the entries above and below the original A_{22} . We have found that if we applied a sequence of elementary row operations we could transform \mathbf{A} into an upper triangular matrix. To summarize, we found constants c_{32} , c_{31} and c_{21} so that

$$R_{32}(c_{32})R_{31}(c_{31})R_{21}(c_{21}) \mathbf{A} = \begin{bmatrix} 1 & 0 & 29 \\ 0 & 1 & 13 \\ 0 & 0 & -156 \end{bmatrix}$$

Now each of the matrices $R_{32}(c_{32})$, $R_{31}(c_{31})$ and $R_{21}(c_{21})$ is formed by adding a multiple of some row of the identity to another row of the identity. Hence, all of these elementary row operation premultiplier matrices have determinant 1. Thus, we know

$$\begin{aligned} \det(\mathbf{A}) &= \det(R_{32}(c_{32})) \det(R_{31}(c_{31})) \det(R_{21}(c_{21})) \det(\mathbf{U}) \\ &= \det(\mathbf{U}) \end{aligned}$$

where \mathbf{U} is the upper triangular matrix we have found by applying these transformations. The determinant of \mathbf{U} is simple: applying our algorithm for the recursive computation of \det here, we find $\det(\mathbf{U})$ is just the product of its diagonals (just expand using column 1). Hence, just 3 multiplies here. Now the elementary row operations don't need to be applied as real matrix multiplications. Instead, we just perform the alterations to rows as needed. So $R_{21}(c_{21})$ needs 3 M and 3 adds but since we know the first column will be zeroed out, we don't really have to do that one. So just 2 M and 2 A. The next one, $R_{31}(c_{31})$ only needs 1 M and 1 A as we don't really have to do the column two entry. Then we are done. We have found $\det(\mathbf{A})$ in just 2 + 1 M and 2 + 1 A to get \mathbf{U} and then 3 M to get \det . The total is 6 M and 3 A. The recursive algorithm requires 12 M and 6 A so already we have saved 50% on the operation count! It gets better and better as the size of \mathbf{A} goes up. Hence, calculating \det using the recursive algorithm is very expensive and we prefer to use elementary row operations instead.

We have also converted our original system into the new system

$$\mathbf{U} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = R_{32}(c_{32})R_{31}(c_{31})R_{21}(c_{21})\mathbf{b}$$

which we can readily solve using backsubstitution.

10.2.1.1 Homework

Exercise 10.2.1

Exercise 10.2.2

Exercise 10.2.3

Exercise 10.2.4

Exercise 10.2.5

10.2.2 MatLab Implementations

Let's see if we can figure out how to find the multiplier matrices R using MatLab. In our code, we will call this matrix P although we can return it as the matrix R is we want! We need some new code. First, let introduce some needed utility functions. the first one, `LMult(i, j, c, n)` replaces row j in the matrix A by the new row j which is $c \text{ row}(i) + \text{row}(j)$.

Listing 10.1: **The Fundamental Row Operation**

```
function R = LMult(i,j,c,n);
%
% R is the left multiplier which replaces
% row j by c * row i + row j
%
R = eye(n,n);
R(j,i) = c;
end
```

We also need the code to do a row swap given by `SwapRows(i, j, n)` which swaps row i and row j of A .

Listing 10.2: **Swapping Rows**

```
function P = SwapRows(i,j,n)
%
row = n;
P = eye(row, row);
%
C = P(i,:);
D = P(j,:);
P(i,:) = D;
P(j,:) = C;
end
```

The code to find the multiplier formed by all the row operations is a modification of our old code `GePiv.m` so that we can keep track of all the multipliers. This is not very efficient, but it will help you see how it works.

Listing 10.3: **Finding the Multiplier**

```
function [P,piv] = GetP(A);
%
% A is nxn matrix
% piv is a nx1 integer permutation vector
%
[n,n] = size(A);
piv = 1:n;
R = eye(n,n);
RPrev = R;
%
Aorig = A;
%Mult = {};
W = {};
```

```

swapindex = 0;
A
15 for k=1:n-1
    % find maximum absolute value entry
    % in column k of rows k through n
    [maxc,r] = max(abs(A(k:n,k)));
    q = r+k-1;
20    % this line swaps the k and q positions in piv
    piv([k q]) = piv([q k]);
    % then manually swap row k and q in A
    swapindex = swapindex+1;
    W{swapindex} = SwapRows(k,q,n);
25    A = W{swapindex}*A;
    if A(k,k) ~=0
        % get multipliers
        c = -A(k+1:n,k)/A(k,k);
        % find multiplier matrix
30        RPrev = W{swapindex}*RPrev;
        for u = k+1:n
            % get multiplier
            S = LMult(k,u,c(u-k),n);
            % reset A. Note A has been possibly row swapped
35            A = S*A;
            R = S*RPrev;
            RPrev = R;
        end
    endA =
40
        1   2   3
        2   4   5
        3   5   6

45 end
P = R;
end

```

Let's try this out. We are going to print out a lot of the intermediary calculations, so it will be a bit verbose! First, let do column one.

Listing 10.4: **Doing Column 1**

```

>> A = [1,2,3;2,4,5;3,5,6];
>> [P,piv] = GetP(A);
3 % The original matrix
A =
    1   2   3
    2   4   5
    3   5   6
8 % we need to swap row 3 and row 1
W =
    0   0   1
    0   1   0
    1   0   0
13 % the swapped matrix is

```

10.2. MATRIX MANIPULATION

175

```

A =
    3   5   6
    2   4   5
    1   2   3
18 % now zero out the second entry in column one
% the row operation matrix is
S =
    1.00000  0.00000  0.00000
   -0.66667  1.00000  0.00000
23    0.00000  0.00000  1.00000
% the new matrix is then
A =
    3.00000  5.00000  6.00000
    0.00000  0.66667  1.00000
28    1.00000  2.00000  3.00000
% The combined row operation is then S*W
% which swaps column 1 and column 3
R =
    0.00000  0.00000  1.00000
33    0.00000  1.00000 -0.66667
    1.00000  0.00000  0.00000
% now zero out the third entry in column one
% the new row operation matrix is
S =
38    1.00000  0.00000  0.00000
    0.00000  1.00000  0.00000
   -0.33333  0.00000  1.00000
% the new matrix is
A =
43    3.00000  5.00000  6.00000
    0.00000  0.66667  1.00000
    0.00000  0.33333  1.00000
% the comined row operation matrix is
% the new one * the previous one giving
48 R =
    0.00000  0.00000  1.00000
    0.00000  1.00000 -0.66667
    1.00000  0.00000 -0.33333

```

Next, column two.

Listing 10.5: **Doing Column 2**

```

% We don't need to do another row swap, so
% the second swap matrix is just the identity
W =
    1   0   0
    0   1   0
    0   0   1
% The new A is unchanged as there is no swap
A =
9
    3.00000  5.00000  6.00000
    0.00000  0.66667  1.00000

```

```

0.00000  0.33333  1.00000
% now zero out the third entry in column 2
14 % the row operation matrix is
S =
1.00000  0.00000  0.00000
0.00000  1.00000  0.00000
0.00000  -0.50000  1.00000
19 5 the altered A is
A =
3.00000  5.00000  6.00000
0.00000  0.66667  1.00000
24 0.00000  0.00000  0.50000
% the combined row operation matrix
% is the new one times the previous one
R =
0.00000  0.00000  1.00000
29 0.00000  1.00000  -0.66667
1.00000  -0.50000  0.00000

```

Note just as a check, $\mathbf{P} * \mathbf{A}$ gives

Listing 10.6: **Checking PA**

```

>> P*A
ans =
3.00000  5.00000  6.00000
5 0.00000  0.66667  1.00000
0.00000  0.00000  0.50000

```

and our row operation matrix is

Listing 10.7: **The Multiplier P**

```

>> P
P =
0.00000  0.00000  1.00000
4 0.00000  1.00000  -0.66667
1.00000  -0.50000  0.00000

```

10.2.2.1 Homework

Exercise 10.2.6

Exercise 10.2.7

Exercise 10.2.8

Exercise 10.2.9

Exercise 10.2.10

10.2.3 Matrix Inverse Calculations

Now let's zero out the entries above the original A_{33} position with $\mathbf{R}_{13}(c)$ and $\mathbf{R}_{23}(d)$ for suitable c and d . Thus, applying $\mathbf{R}_{13}(29/156)$ we have

$$\mathbf{R}_{13}(29/156) \mathbf{B} = \begin{bmatrix} 1 & 0 & 29/156 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 29 \\ 0 & 1 & 13 \\ 0 & 0 & -156 \end{bmatrix} \begin{bmatrix} 20 \\ 9 \\ -99 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 13 \\ 0 & 0 & -156 \end{bmatrix} \begin{bmatrix} 1.5962 \\ 9 \\ -99 \end{bmatrix}$$

Finally, applying $\mathbf{R}_{23}(13/156)$, we have

$$\mathbf{R}_{23}(-1/13) \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 13/156 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 13 \\ 0 & 0 & -156 \end{bmatrix} \begin{bmatrix} 1.5962 \\ 9 \\ -99 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -156 \end{bmatrix} \begin{bmatrix} 1.5962 \\ 0.75 \\ -99 \end{bmatrix}$$

The last row then tells us $z = 99/156 = 0.6346$. We had already found $x = 1.5962$ and $y = 0.75$ so we have solved the system. This is what we did in code with the LU decomposition of \mathbf{A} and the use of upper and lower triangular solvers. You should go back and look at that again to get a better appreciation of what we have done here. Of course, not trusting our typing and arithmetic, we used our MatLab code to double check all of these calculations above. All we can say is this has been intense! But let's see what we have found. We wanted to solve the problem $\mathbf{AX} = \mathbf{b}$ where the data \mathbf{b} is given and \mathbf{X} denotes the triple $[x \ y \ z]^T$. We found that if we applied a sequence of elementary row operations we could transform \mathbf{A} into a diagonal matrix. To summarize, we found

$$\mathbf{R}_{23}(c_{23})\mathbf{R}_{13}(c_{13})\mathbf{R}_{12}(c_{12})\mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21}) \mathbf{A} = \begin{bmatrix} r_1 & 0 & 0 \\ 0 & r_2 & 0 \\ 0 & 0 & r_3 \end{bmatrix}$$

for appropriate constants c_{ij} and some nonzero constants r_1, r_2 and r_3 . Once this was done it is easy to find the unique x, y and z that solve the problem: we simply divide the entries of the transformed \mathbf{b} by the appropriate r_i . We can do this for any system of equations and, of course it fails if \mathbf{A} has problems.

Note, we can think of this a different way. Assume you are trying to solve three separate problems simultaneously: $\mathbf{AX} = \mathbf{e}_1$, $\mathbf{AX} = \mathbf{e}_2$ and $\mathbf{AX} = \mathbf{e}_3$ where $\mathbf{e}_1 = [1 \ 0 \ 0]^T$, $\mathbf{e}_2 = [0 \ 1 \ 0]^T$ and $\mathbf{e}_3 = [0 \ 0 \ 1]^T$. Form the new augmented matrix

$$\begin{bmatrix} [A_{11} \ A_{12} \ A_{13}] & [1 \ 0 \ 0] \\ [A_{21} \ A_{22} \ A_{23}] & [0 \ 1 \ 0] \\ [A_{31} \ A_{32} \ A_{33}] & [0 \ 0 \ 0] \end{bmatrix}$$

Then apply the elementary row operations like usual to convert \mathbf{A} into an upper triangular matrix \mathbf{U} . If this system is solvable, we can always do this. Thus, we find

$$\begin{aligned} \mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21}) \mathbf{A} &= \mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21}) \mathbf{I} \\ \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{21} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix} &= \mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21}) \mathbf{I} \end{aligned}$$

Now apply the other three elementary row operations.

$$\begin{aligned} \mathbf{R}_{23}(c_{23})\mathbf{R}_{13}(c_{13})\mathbf{R}_{12}(c_{12})\mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21}) \mathbf{A} \\ = \mathbf{R}_{23}(c_{23})\mathbf{R}_{13}(c_{13})\mathbf{R}_{12}(c_{12})\mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21}) \mathbf{I} \end{aligned}$$

$$\begin{bmatrix} r_1 & 0 & 0 \\ 0 & r_2 & 0 \\ 0 & 0 & r_3 \end{bmatrix} = \mathbf{R}_{23}(c_{23})\mathbf{R}_{13}(c_{13})\mathbf{R}_{12}(c_{12})\mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21}) \mathbf{I}$$

Now apply the diagonal matrix $\mathbf{D}(1/r_1), (1/r_2), (1/r_3)$ (whose main diagonal entries are $(1/r_1), (1/r_2), (1/r_3)$ with every other entry zero. Let \mathbf{B} be defined by

$$\mathbf{B} = \begin{bmatrix} 1/r_1 & 0 & 0 \\ 0 & 1/r_2 & 0 \\ 0 & 0 & 1/r_3 \end{bmatrix} \mathbf{R}_{23}(c_{23})\mathbf{R}_{13}(c_{13})\mathbf{R}_{12}(c_{12})\mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21})$$

Then, we have

$$\mathbf{I} = \mathbf{B} \mathbf{A}$$

We have just found something cool! For convenience, let \mathcal{R} be given by

$$\mathcal{R} = \mathbf{R}_{23}(c_{23})\mathbf{R}_{13}(c_{13})\mathbf{R}_{12}(c_{12})\mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21})$$

Then note

$$\mathbf{B} = \mathbf{D}(1/r_1), (1/r_2), (1/r_3) \mathcal{R}$$

and $\mathbf{B}\mathbf{A} = \mathbf{I}$ so that we have found the inverse of \mathbf{A} using elementary row operations. It is

$$\mathbf{A}^{-1} = \mathbf{D}(1/r_1), (1/r_2), (1/r_3) \mathcal{R}!$$

This is a computationally efficient way to find a matrix inverse, if it exists.

Homework

Exercise 10.2.11

Exercise 10.2.12

Exercise 10.2.13

Exercise 10.2.14

Exercise 10.2.15

10.2.3.1 MatLab Implementations

We can implement these ideas using our row operation matrices explicitly as follows. Again, we will implement the code using a matrix name different from the \mathbf{R} 's above. For this code we will use Q and the \mathbf{R} we want will end up being QP . Now this is not so efficient as normally, we would not store these matrices, but we wanted you to see how it works. We already know how to get the row operation matrix P that converts A into an upper triangular matrix. To finish the job, we don't have to worry about row swaps so the code is cleaner. We only apply this code to symmetric matrices.

Consider the code for `GetQ2(A)` given below.

Listing 10.8: **GetQ2.m: Diagonalize A**

```
function Q = GetQ2(A);
```

```

%
% A is nxn matrix in upper triangular form
%
5 [n,n] = size(A);
R = eye(n,n);
for k=2:n
    if A(k,k) ~=0
        % get multipliers
10    % look at rows k+1 to n and
        % divide them by the pivot value A(k,k)
        c = -A(1:k-1,k)/A(k,k);
        % find multiplier matrix
        for u = 1:k-1
            S = LMult(k,u,c(u),n);
            A = S*A;
            R = S*R;
        end
    end
20 end
Q = R;
end

```

This code finds the row operation matrices that convert PA into diagonal form. Let's do a 4×4 example.

Listing 10.9: **Upper Triangularizing A**

```

>> A = [4,1,3,2;1,5,6,3;3,6,10,5;2,3,5,1]
% original A
3 A =
    4     1     3     2
    1     5     6     3
    3     6    10     5
    2     3     5     1
8 >> [P,piv] = GetP(A);
% PA is upper triangular
>> P*A
ans =
    4.00000    1.00000    3.00000    2.00000
    0.00000    5.25000    7.75000    3.50000
    0.00000    0.00000   -1.76190   -0.66667
    0.00000    0.00000    0.00000   -1.59459
% apply further row operations to diagonalize PA
>> Q = GetQ2(P*A);
18 >> Q
Q =
    1.00000   -0.19048    0.86486    0.47458
    0.00000    1.00000    4.39865    0.35593
23    0.00000    0.00000    1.00000   -0.41808
    0.00000    0.00000    0.00000    1.00000
% note QP is the matrix which diagonalizes A
>> Q*P
ans =

```

```

28   1.42373  0.81356 -1.15254  0.47458
    1.06780  4.36017 -3.11441  0.35593
    0.50767  1.04520 -0.74657 -0.41808
    -0.18919 -0.10811 -0.37838  1.00000
    % Q*P*A gives us the diagonal matrix D
33 Q*P*A
ans =
    4.00000 -0.00000  0.00000 -0.00000
    -0.00000  5.25000 -0.00000 -0.00000
    0.00000  0.00000 -1.76190  0.00000
38   0.00000  0.00000  0.00000 -1.59459

```

Hence, we have $Q P A = D$ and so $D^{-1} Q P A = I$ which tells us $A^{-1} = D^{-1} Q P$! We can package this into a new piece of code `GetInvSymm(A)`.

Listing 10.10: **Simple Matrix Inverse Code**

```

function Ainv = GetInvSymm(A)
%
% A is an n x n symmetrix matrix
%
5 n = size(A);
[P,piv] = GetP(A);
Q = GetQ2(P*A);
D = Q*P*A;
Dinv = D;
10 for k=1:n
    Dinv(k,k) = 1/D(k,k);
end
Ainv = Dinv*Q*P;
end

```

For our example, we find

Listing 10.11: **Diagonalizing A**

```

1 >> Ainv = GetInvSymm(A);
% original matrix
A =
6
    4     1     3     2
    1     5     6     3
    3     6    10     5
    2     3     5     1
>> Ainv
Ainv =
11   0.355932  0.203390 -0.288136  0.118644
    0.203390  0.830508 -0.593220  0.067797
    -0.288136 -0.593220  0.423729  0.237288
    0.118644  0.067797  0.237288 -0.627119
% Checks
16 >> A*Ainv

```

```

ans =
  1.0000e+00  -2.2482e-15  3.5527e-15  -2.2204e-16
  -1.7708e-14  1.0000e+00  1.8430e-14  4.2188e-15
  -1.9318e-14  -1.4599e-14  1.0000e+00  4.8850e-15
21   -9.8949e-15  -7.8965e-15  9.5479e-15  1.0000e+00
>>> Ainv*A
ans =
  1.00000  -0.00000  0.00000  0.00000
-0.00000  1.00000  0.00000  0.00000
26  0.00000  0.00000  1.00000  0.00000
  0.00000  0.00000 -0.00000  1.00000

```

We see the inverse checks out! Normally, we just do elementary row operations more efficiently. For example, our earlier *LU* decomposition works by using row operations.

Homework

Exercise 10.2.16

Exercise 10.2.17

Exercise 10.2.18

Exercise 10.2.19

Exercise 10.2.20

10.3 Back To Definite Matrices

Our interests are in symmetric matrices that are positive or negative definite. Recall, a symmetric matrix \mathbf{A} can be written as

$$\mathbf{A} = \mathbf{P} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \mathbf{P}^T$$

where \mathbf{P} is the matrix of \mathbf{A} 's eigenvectors. This \mathbf{P} is different from the generic P we used in the upper triangularization code introduce above. So in this section, we have the matrix of eigenvectors \mathbf{P} and new matrices of the form \mathbf{R} which we will use to develop some ideas about minors. For our discussions here, we will ignore row swapping issues to keep our explanations from becoming too messy. However, you should be able to see adding the `piv` idea into the mixture is not terribly difficult. Note, letting Λ be the diagonal matrix $\mathbf{D}(\lambda_1, \lambda_2, \lambda_3)$, we have

$$\begin{aligned} \det(\mathbf{A}) &= \det(\mathbf{P}) \det(\Lambda) \det(\mathbf{P}^T) \\ &= \lambda_1 \lambda_2 \lambda_3 \end{aligned}$$

as $\mathbf{P}\mathbf{P}^T = \mathbf{I}$ and so $\det(\mathbf{P}) \det(\mathbf{P}^T) = 1$. We can apply the ideas above to the matrix \mathbf{A} to find a matrix \mathbf{R} given by

$$\mathbf{R} = \mathbf{R}_{32}(c_{32})\mathbf{R}_{31}(c_{31})\mathbf{R}_{21}(c_{21})$$

for appropriate constants c_{ij} so that

$$\mathcal{R}A = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix}$$

Now let \mathbf{U} denote the upper triangular matrix above. From our discussions above, we know $\det(\mathbf{U}) = U_{11}U_{22}U_{33}$. However, this determinant is also $\lambda_1\lambda_2\lambda_3$. Hence, we know $\lambda_1\lambda_2\lambda_3 = U_{11}U_{22}U_{33}$. Now the next step is also pretty intense so just pour a new cup of tea or coffee and settle down for some mind stretching.

10.3.1 Matrix Minors

Now let's study the transformed matrix $\mathcal{R}A\mathcal{R}^T$ which looks like

$$\mathcal{R}A\mathcal{R}^T = \left(\mathbf{R}_{32}\mathbf{R}_{31}\mathbf{R}_{21} \right) A \left(\mathbf{R}_{21}^T \mathbf{R}_{31}^T \mathbf{R}_{32}^T \right) = \Lambda$$

where we have dropped the c_{ij} terms for expositional clarity. You know they are there, right?

Now look at the innermost part, $\mathbf{R}_{21}A\mathbf{R}_{21}^T$. The matrix A has three important determinants associated with it called its **principle minors**. This transformation is applied with some constant c so we have

$$\mathbf{R}_{21}(c)A = \begin{bmatrix} 1 & 0 & 0 \\ c & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12} & A_{22} & A_{23} \\ A_{13} & A_{23} & A_{33} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ cA_{11} + A_{12} & cA_{12} + A_{22} & cA_{13} + A_{23} \\ A_{13} & A_{23} & A_{33} \end{bmatrix}$$

Next, calculate

$$\begin{aligned} \mathbf{R}_{21}(c)\mathbf{A}\mathbf{R}_{21}^T(c) &= \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ cA_{11} + A_{12} & cA_{12} + A_{22} & cA_{13} + A_{23} \\ A_{13} & A_{23} & A_{33} \end{bmatrix} \begin{bmatrix} 1 & c & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} A_{11} & cA_{11} + A_{12} & A_{13} \\ cA_{11} + A_{12} & c^2A_{11} + 2cA_{12} + A_{22} & cA_{13} + A_{23} \\ A_{13} & cA_{13} + A_{23} & A_{33} \end{bmatrix} \end{aligned}$$

The principle minors of our A as defined as follows:

$$\begin{aligned} (PM)_1 &= \det(A_{11}) = A_{11} \\ (PM)_2 &= \det \left(\begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix} \right) \\ (PM)_3 &= \det \left(\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12} & A_{22} & A_{23} \\ A_{13} & A_{23} & A_{33} \end{bmatrix} \right) \end{aligned}$$

Let $(PMT)_i$ denote the principle minors of the matrix A after the first pair of elementary row operations. From the above calculation, we see $(PMT)_1 = (PM)_1 = A_{11}$ and $(PMT)_2 = (PM)_2$. What about $(PMT)_3$? Let's do a sample calculation. We have

$$\mathbf{R}_{21}(c) \mathbf{A} \mathbf{R}_{21}^T(c) = \begin{bmatrix} A_{11} & cA_{11} + A_{12} & A_{13} \\ cA_{11} + A_{12} & c^2 A_{11} + 2cA_{12} + A_{22} & cA_{13} + A_{23} \\ A_{13} & cA_{13} + A_{23} & A_{33} \end{bmatrix}$$

Let's expand by column 3. We have

$$\begin{aligned} \det(\mathbf{R}_{21}(c) \mathbf{A} \mathbf{R}_{21}^T(c)) &= A_{13} \det \left(\begin{bmatrix} cA_{11} + A_{12} & c^2 A_{11} + 2cA_{12} + A_{22} \\ A_{13} & cA_{13} + A_{23} \end{bmatrix} \right) \\ &\quad - (cA_{13} + A_{23}) \det \left(\begin{bmatrix} A_{11} & cA_{11} + A_{12} \\ A_{13} & cA_{13} + A_{23} \end{bmatrix} \right) \\ &\quad + A_{33} \det \left(\begin{bmatrix} A_{11} & cA_{11} + A_{12} \\ cA_{11} + A_{12} & c^2 A_{11} + 2cA_{12} + A_{22} \end{bmatrix} \right) \end{aligned}$$

After calculating the determinants and some simplifications (yes, we expect you to use pen and paper and verify!), we find

$$\begin{aligned} \det(\mathbf{R}_{21}(c) \mathbf{A} \mathbf{R}_{21}^T(c)) &= A_{13} (A_{12}A_{23} - A_{13}A_{22} + cA_{11}A_{23} - cA_{12}A_{13}) \\ &\quad - (cA_{13} + A_{23}) (A_{11}A_{23} - A_{12}A_{13}) \\ &\quad + A_{33} (A_{11}A_{22} - A_{12}^2) \end{aligned}$$

The term $A_{11}A_{22} - A_{12}^2$ is the determinant of the cofactor C_{33} and buried in this messy expression are two other cofactors: $A_{12}A_{23} - A_{13}A_{22}$ is the determinant of the cofactor C_{31} and $A_{11}A_{23} - A_{11}A_{13}$ is the determinant of the cofactor C_{32} . Hence, we have

$$\begin{aligned} \det(\mathbf{R}_{21}(c) \mathbf{A} \mathbf{R}_{21}^T(c)) &= A_{13} (\det(C_{31} + cA_{11}A_{23} - cA_{12}A_{13}) - (cA_{13} + A_{23}) (\det(C_{32})) \\ &\quad + A_{33} \det(C_{33})) \end{aligned}$$

But we have a symmetric matrix and so $A_{31} = A_{13}$ and so expanding by the third row, we have

$$\begin{aligned} \det(\mathbf{A}) &= A_{31} \det(C_{31} - A_{32} \det(C_{32}) + A_{33} \det(C_{33})) \\ &= A_{13} \det(C_{31} - A_{23} \det(C_{32}) + A_{33} \det(C_{33})) \end{aligned}$$

We conclude

$$\begin{aligned} \det(\mathbf{R}_{21}(c) \mathbf{A} \mathbf{R}_{21}^T(c)) &= \det(\mathbf{A}) + cA_{11}A_{13}A_{23} - cA_{13}A_{12}A_{13} \\ &\quad - cA_{13} (A_{11}A_{23} - A_{12}A_{13}) \\ &= \det(\mathbf{A}) + cA_{11}A_{13}A_{23} - cA_{12}A_{13}^2 - cA_{13}A_{11}A_{23} + cA_{12}A_{13}^2 \\ &= \det(\mathbf{A}) \end{aligned}$$

Thus, for this one pair of elementary row operations applied to \mathbf{A} , we have the principle minors of \mathbf{A} and $\mathbf{R}_{21}(c) \mathbf{A} \mathbf{R}_{21}^T(c)$ are equal. Here, it is important we apply transformations that transform \mathbf{A} into a lower triangular matrix. If we don't the determinants of the principle minors will not match. the arguments for the other transformations are similar. So we can say the principle minors of $\mathbf{R}_{32}\mathbf{R}_{31}\mathbf{R}_{21} \mathbf{A} \mathbf{R}_{21}^T \mathbf{R}_{31}^T \mathbf{R}_{32}^T$ match the principle minors of \mathbf{A} .

We conclude the principle minors of $\mathcal{R}\mathbf{A}\mathcal{R}^T$ match the principle minors of \mathbf{A} . These principle minors are λ_1 , $\lambda_1\lambda_2$ and $\lambda_1\lambda_2\lambda_3$.

10.3.1.1 Homework

Exercise 10.3.1

Exercise 10.3.2

Exercise 10.3.3

Exercise 10.3.4

Exercise 10.3.5

Part III

Calculus of Many Variables

Chapter 11

Differentiability

We now consider the differentiability properties of mappings $f : D \subset \Re^n \rightarrow \Re^m$ in earnest. In (Peterson (8) 2019), we worked through a good introduction to functions of two variables. Now we want to look at higher dimensional analogues.

11.1 Partial Derivatives

If f is locally defined on $B(\mathbf{x}_0, r)$ in \Re^n , we can look at the **trace** corresponding to a \mathbf{x}_0 in the direction of the unit vector \mathbf{E} ; call this $f_{\mathbf{E}}(t) = f(\mathbf{x}_0 + t\mathbf{E})$ for sufficiently small t so the function values are well defined. This means

$$f_{\mathbf{E}}(t) = f(\mathbf{x}_0 + t\mathbf{E}) = f(x_{01} + tE_1, \dots, x_{0n} + tE_n)$$

If we calculate the difference between $f_{\mathbf{E}}(t)$ and $f_{\mathbf{E}}(0)$, we find for any $t \neq 0$

$$\frac{f_{\mathbf{E}}(t) - f_{\mathbf{E}}(0)}{t} = \frac{f(x_{01}, \dots, x_{0,i-1}, \mathbf{x}_{0i} + t, \dots, x_{0n}) - f(x_{01}, \dots, x_{0,i-1}, \mathbf{x}_{0i}, \dots, x_{0n})}{t}$$

where we have put in boldface the i^{th} slot where all the action is taking place. It is useful to determine if $\lim_{t \rightarrow 0} (1/t)(f(\mathbf{x}_0 + t\mathbf{E}) - f(\mathbf{x}_0))$ exists. In the case of two variables, this turns out to be the directional derivative of f in the direction of \mathbf{E} . But we need to do this carefully for n variables and we don't want to take advantage of two dimensional intuition. We start by looking at the standard basis $\{\mathbf{E}_1, \dots, \mathbf{E}_n\}$ for \Re^n where the components of $\mathbf{E}_i = \delta_{ij}$ as usual. Then, we can consider the limit

$$\lim_{t \rightarrow 0} \frac{f(\mathbf{x}_0 + t\mathbf{E}_i) - f(\mathbf{x}_0)}{t}$$

If this limit exists, it is called the **partial derivative** of f with respect to the i^{th} coordinate slot. There are many notations for this.

Definition 11.1.1

Let $z = f(\mathbf{x})$ be a function defined locally at \mathbf{x} in \Re^n . The partial derivative of f with respect to its i^{th} coordinate or slot at \mathbf{x} is defined by the limit

$$\frac{\partial f}{\partial x}(\mathbf{x}) = \lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t\mathbf{E}_i) - f(\mathbf{x})}{t} = \lim_{y \rightarrow x_i} \frac{f(\mathbf{x} + y\mathbf{E}_i) - f(\mathbf{x})}{y - x_i}$$

If this limit exists, it is denoted by a variety of symbols: for functions of two variables x and y , for the partial derivatives at the point (x_0, y_0) , we often use $f_x(x_0, y_0)$, $z_x(x_0, y_0)$, $\frac{\partial f}{\partial x}(x_0, y_0)$ and $\frac{\partial z}{\partial x}(x_0, y_0)$ and $f_y(x_0, y_0)$, $z_y(x_0, y_0)$, $\frac{\partial f}{\partial y}(x_0, y_0)$ and $\frac{\partial z}{\partial y}(x_0, y_0)$. However, this notation is not very useful when \mathbf{x} is in \Re^n . In the more general case, we use D_i to indicate the partial derivative with respect to the i^{th} slot instead of $\frac{\partial f}{\partial x_i}$.

Comment 11.1.1 It is easy to take partial derivatives. Just imagine the one variable held constant and take the derivative of the resulting function just like you did in your earlier calculus courses.

Comment 11.1.2 Notation can be really bad here. Consider the following partial we need to take in a problem with line integrals in Chapter 17. This is a calculation for functions of two variables and uses the chain rule which we covered in (Peterson (8) 2019). We need to find

$$(\mathbf{A}\mathbf{G}_{1v} + \mathbf{B}\mathbf{G}_{2v})_u$$

where \mathbf{A} and \mathbf{B} are functions of two variables and \mathbf{G}_1 and \mathbf{G}_2 are also. This is a change of variable problem with $x = \mathbf{G}_1(u, v)$ and $y = \mathbf{G}_2(u, v)$. In the chain rule, we need $\partial \mathbf{A}$ with respect to argument one and argument two. In the original \mathbf{A} function, these would be \mathbf{A}_x and \mathbf{A}_y but that is really confusing here. That would give

$$\begin{aligned} (\mathbf{A}\mathbf{G}_{1v} + \mathbf{B}\mathbf{G}_{2v})_u &= (\mathbf{A}_x \mathbf{G}_{1u} + \mathbf{A}_y \mathbf{G}_{2u}) \mathbf{G}_{1v} + \mathbf{A}\mathbf{G}_{1vu} \\ &\quad + (\mathbf{B}_x \mathbf{G}_{1u} + \mathbf{B}_y \mathbf{G}_{2u}) \mathbf{G}_{2v} + \mathbf{B}\mathbf{G}_{2vu} \end{aligned}$$

or we could use $\arg 1$ and $\arg 2$ which is ugly

$$\begin{aligned} (\mathbf{A}\mathbf{G}_{1v} + \mathbf{B}\mathbf{G}_{2v})_u &= (\mathbf{A}_{\arg 1} \mathbf{G}_{1u} + \mathbf{A}_{\arg 2} \mathbf{G}_{2u}) \mathbf{G}_{1v} + \mathbf{A}\mathbf{G}_{1vu} \\ &\quad + (\mathbf{B}_{\arg 1} \mathbf{G}_{1u} + \mathbf{B}_{\arg 2} \mathbf{G}_{2u}) \mathbf{G}_{2v} + \mathbf{B}\mathbf{G}_{2vu} \end{aligned}$$

But it is cleaner to just use \mathbf{A}_1 and \mathbf{A}_2 and so forth to indicate these partials with respect to slot one and slot two.

$$\begin{aligned} (\mathbf{A}\mathbf{G}_{1v} + \mathbf{B}\mathbf{G}_{2v})_u &= (\mathbf{A}_1 \mathbf{G}_{1u} + \mathbf{A}_2 \mathbf{G}_{2u}) \mathbf{G}_{1v} + \mathbf{A}\mathbf{G}_{1vu} \\ &\quad + (\mathbf{B}_1 \mathbf{G}_{1u} + \mathbf{B}_2 \mathbf{G}_{2u}) \mathbf{G}_{2v} + \mathbf{B}\mathbf{G}_{2vu} \end{aligned}$$

At any rate, don't be surprised if, in a complicated partial calculation, it all gets confusing!

11.1.1 Homework

Exercise 11.1.1

Exercise 11.1.2

Exercise 11.1.3

Exercise 11.1.4

Exercise 11.1.5

11.2 Tangent Planes

It is very useful to have an analytical way to discuss tangent planes and their relationship to the real surface to which they are attached even if we can't draw this.

Definition 11.2.1 Planes in n Dimensions

A plane in \mathbb{R}^n through the point x_0 is defined as the set of all vectors x so that the angle between the vectors D and N is 90° where D is the vector we get by connecting the point x_0 to the point x . Hence, for

$$D = x - x_0 = \begin{bmatrix} x_1 - x_{01} \\ \vdots \\ x_n - x_{0n} \end{bmatrix} \quad \text{and} \quad N = \begin{bmatrix} N_1 \\ \vdots \\ N_n \end{bmatrix}$$

The plane is the set of points x so that $\langle D, N \rangle = 0$. The vector N is called the **normal vector** to the plane. If $n > 2$, we usually call this a **hyperplane**. Also note hyperplanes through the origin of \mathbb{R}^n are $n - 1$ dimensional subspaces of \mathbb{R}^n .

If N is given, then the span of N , $sp(N)$, is a one dimensional subspace of \mathbb{R}^n and the orthogonal complement of it which is

$$sp(N)^\perp = \{x \in \mathbb{R}^n \mid \langle x, N \rangle = 0\}$$

is an $n - 1$ dimensional subspace of \mathbb{R}^n and we can construct an orthonormal basis for it as follows:

- Pick any nonzero w_1 not equal to a multiple of N . Then w_1 is not in $sp(N)$. Compute $F_1 = N/\|N\|$ and

$$v = w_1 - \langle w_1, F_1 \rangle F_1$$

Then v is orthogonal to F_1 . Let $F_2 = v/\|v\|$.

- If the $sp(F_1, F_2)$ is not \mathbb{R}^n , we know there is a nonzero vector w in $(sp(F_1, F_2))^\perp$ from which we can construct a third vector F_3 mutually orthogonal to F_1 and F_2 just we did before.
- If the $sp(F_1, F_2, F_3)$ is not \mathbb{R}^n , we do this again. If not, we are done.

This constructs $n - 1$ mutually orthonormal vectors F_1 to F_{n-1} which are a basis for $sp(N)^\perp$. Note any x in this subspace satisfies $\langle x, N \rangle = 0$.

On the other hand if A_1 through A_{n-1} is a linearly independent set in \mathbb{R}^n , the $sp(A_1, \dots, A_{n-1})$ is an $n - 1$ dimensional subspace. To find the normal vector, note we are looking for a vector N so that

$$\begin{bmatrix} A_{11} & \dots & A_{1,n} \\ \vdots & \vdots & \vdots \\ A_{11} & \vdots & A_{1,n} \end{bmatrix} \begin{bmatrix} N_1 \\ \vdots \\ N_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

This is the same as

$$N_1 \begin{bmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{n-1,1} \end{bmatrix} + N_1 \begin{bmatrix} A_{12} \\ A_{22} \\ \vdots \\ A_{n-1,2} \end{bmatrix} + \dots + N_n \begin{bmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{n-1,n} \end{bmatrix} = 0.$$

Letting

$$\mathbf{W}_i = \begin{bmatrix} A_{1i} \\ A_{2i} \\ \vdots \\ A_{n-1,i} \end{bmatrix}$$

we see $\mathbf{W}_1, \dots, \mathbf{W}_n$ is a set of n vectors in \Re^{n-1} and so must be linearly dependent. Thus, the equation

$$N_1 \mathbf{W}_1 + \dots + N_n \mathbf{W}_n = \mathbf{0}$$

must have a nonzero solution \mathbf{N} which will be the normal vector we seek.

Homework

Exercise 11.2.1

Exercise 11.2.2

Exercise 11.2.3

Exercise 11.2.4

Exercise 11.2.5

Now given our function of n variables, we calculate the n partial derivatives at a given point \mathbf{x}_0 giving $f_1(\mathbf{x}_0)$ to $f_n(\mathbf{x}_0)$. The vector

$$\mathbf{N} = \begin{bmatrix} f_1(\mathbf{x}_0) \\ \vdots \\ f_n(\mathbf{x}_0) \end{bmatrix}$$

therefore determines a hyperplane like we discussed above where the hyperplane is the set of all \mathbf{x} in \Re^n so that $\langle \mathbf{x}, \mathbf{N} \rangle = 0$.

Now what does this have to do with what we have called tangent planes for functions of two variables. Let's look at that case for some motivation. Recall the tangent plane to a surface $z = f(x, y)$ at the point (x_0, y_0) was the plane determined by the tangent lines $T(x, y_0)$ and $T(x_0, y)$. The $T(x, y_0)$ line was determined by the vector

$$\mathbf{A} = \begin{bmatrix} 1 \\ 0 \\ \frac{\partial f}{\partial x}(x_0, y_0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2x_0 \end{bmatrix}$$

and the $T(x_0, y)$ line was determined by the vector

$$\mathbf{B} = \begin{bmatrix} 0 \\ 1 \\ \frac{\partial f}{\partial y}(x_0, y_0) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 2y_0 \end{bmatrix}$$

The vector perpendicular to both \mathbf{A} and \mathbf{B} is given by $\mathbf{A} \times \mathbf{B}$ but we can't use the cross product in the dimension $n > 3$. The cross product gives

$$\mathbf{N} = \begin{bmatrix} -f_x(x_0, y_0) \\ -f_y(x_0, y_0) \\ 1 \end{bmatrix}$$

The tangent plane to $z = f(x, y)$ at $(x_0, y_0, f(x_0, y_0))$ is defined to be the set of all (x, y, z) in \mathbb{R}^3 so that

$$\left\langle \begin{bmatrix} -f_x(x_0, y_0) \\ -f_y(x_0, y_0) \\ 1 \end{bmatrix}, \begin{bmatrix} x - x_0 \\ y - y_0 \\ z - f(x_0, y_0) \end{bmatrix} \right\rangle = 0$$

or multiplying this out and rearranging

$$z = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$$

We can use this as the motivation for defining the tangent vector to $z = f(\mathbf{x})$ at \mathbf{x}_0 as follows:

Definition 11.2.2 The Tangent Plane to $z = f(\mathbf{x})$ at \mathbf{x}_0 in \mathbb{R}^n

The tangent plane in \mathbb{R}^n through the point \mathbf{x}_0 for the surface $z = f(\mathbf{x})$ is defined as the set of all vectors \mathbf{x} so that

$$\left\langle \begin{bmatrix} -f_1(\mathbf{x}_0) \\ -f_2(\mathbf{x}_0) \\ \vdots \\ -f_n(\mathbf{x}_0) \\ 1 \end{bmatrix}, \begin{bmatrix} x_1 - x_{01} \\ x_2 - x_{02} \\ \vdots \\ x_n - x_{0n} \\ z - f(\mathbf{x}_0) \end{bmatrix} \right\rangle = 0$$

or multiplying this out and rearranging

$$z = f(\mathbf{x}_0) + f_1(\mathbf{x}_0)(x_1 - x_{01}) + \dots + f_n(\mathbf{x}_0)(x_n - x_{0,n})$$

We can use another compact definition at this point. We can define the **gradient** of the function f to be the vector ∇f which is defined as follows.

Definition 11.2.3 The Gradient

The gradient of the function $z = f(\mathbf{x})$ at the point \mathbf{x}_0 is defined to be the vector ∇f where

$$\nabla f(\mathbf{x}_0) = \begin{bmatrix} f_1(\mathbf{x}_0) \\ f_2(\mathbf{x}_0) \\ \vdots \\ f_n(\mathbf{x}_0) \end{bmatrix}$$

Note the gradient takes a scalar function argument and returns a vector answer.

Using the gradient, the tangent plane equation can be rewritten as

$$\begin{aligned} z &= f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle \\ &= f(\mathbf{x}_0) + (\nabla f(\mathbf{x}_0))^T (\mathbf{x} - \mathbf{x}_0) \end{aligned}$$

It is also traditional to abbreviate $\nabla f(\mathbf{x}_0)$ by ∇f^0 . Then, the tangent plane equation becomes

$$z = f(\mathbf{x}_0) + (\nabla f^0)^T (\mathbf{x} - \mathbf{x}_0)$$

The obvious question to ask now is how much of a discrepancy is there between the value $f(\mathbf{x})$ and the value of the tangent plane?

Example 11.2.1 Find the gradient of $w = f(x, y, z) = x^2 + 4xyz + 9y^2 + 20z^4$ and the equation of the tangent plane to this surface at the point $(1, 2, 1)$.

Solution

$$\nabla f(x, y) = \begin{bmatrix} 2x + 4yz \\ 4xz + 18y \\ 4xy + 80z^3 \end{bmatrix} \implies \nabla f(1, 2, 1) = \begin{bmatrix} 2 + 8 = 10 \\ 4 + 36 = 40 \\ 4 + 80 = 84 \end{bmatrix}$$

The equation of the tangent plane at $(1, 2, 1)$ is then

$$\begin{aligned} z &= f(1, 2, 1) + \left\langle \begin{bmatrix} 10 \\ 40 \\ 84 \end{bmatrix}, \begin{bmatrix} x-1 \\ y-2 \\ z-1 \end{bmatrix} \right\rangle \\ &= 65 + 10(x-1) + 40(y-2) + 84(z-1) \end{aligned}$$

Homework

Exercise 11.2.6

Exercise 11.2.7

Exercise 11.2.8

Exercise 11.2.9

Exercise 11.2.10

11.3 Derivatives for Scalar Functions of n Variables

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be a function of n variables. As long as $f(\mathbf{x})$ is defined locally at \mathbf{x}_0 , the partials $f_1(\mathbf{x}_0)$ through $f_n(\mathbf{x}_0)$ exist if and only if there are error functions \mathcal{E}_1 through \mathcal{E}_N so that

$$\begin{aligned} f_1(\mathbf{x}_0 + h_1 \mathbf{E}_1) &= f(\mathbf{x}_0) + f_1(\mathbf{x}_0)(h_1(x_1 - x_{01})) + \mathcal{E}_1(\mathbf{x}_0, h_1) \\ f_2(\mathbf{x}_0 + h_2 \mathbf{E}_2) &= f(\mathbf{x}_0) + f_2(\mathbf{x}_0)(h_2(x_2 - x_{02})) + \mathcal{E}_2(\mathbf{x}_0, h_2) \\ &\vdots = \vdots \\ f_n(\mathbf{x}_0 + h_n \mathbf{E}_n) &= f(\mathbf{x}_0) + f_n(\mathbf{x}_0)(h_n(x_n - x_{0n})) + \mathcal{E}_n(\mathbf{x}_0, h_n) \end{aligned}$$

with $\mathcal{E}_j \rightarrow 0$ and $\mathcal{E}_j/h_j \rightarrow 0$ as $h_j \rightarrow 0$ for all indices j . This suggests the correct definition for the differentiability of a function of n variables.

Definition 11.3.1 Error Form of Differentiability For Scalar Function of n Variables

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be a function of n variables. If $f(\mathbf{x})$ is defined locally at \mathbf{x}_0 , then f is differentiable at \mathbf{x}_0 if there is a vector \mathbf{L} in \mathbb{R}^n so that the error function $\mathcal{E}(\mathbf{x}_0, \mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}_0) - \langle \mathbf{L}, \mathbf{x} - \mathbf{x}_0 \rangle$ satisfies $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \mathcal{E}(\mathbf{x}_0, \mathbf{x}) = 0$ and $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \mathcal{E}(\mathbf{x}_0, \mathbf{x}) / \|\mathbf{x} - \mathbf{x}_0\| = 0$. We define the derivative of f at \mathbf{x}_0 to be the linear map \mathbf{L} whose value at \mathbf{x} is $\mathbf{L}^T(\mathbf{x} - \mathbf{x}_0)$. We denote this map by $Df(\mathbf{x}_0) = \mathbf{L}^T$.

Note if f is differentiable at \mathbf{x}_0 , f must be continuous at \mathbf{x}_0 . The argument is simple:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \langle \mathbf{L}, \mathbf{x} - \mathbf{x}_0 \rangle + \mathcal{E}(\mathbf{x}_0, \mathbf{x})$$

and as $\mathbf{x} \rightarrow \mathbf{x}_0$, we have $f(\mathbf{x}) \rightarrow f(\mathbf{x}_0)$ which is the definition of f being continuous at (\mathbf{x}_0) . Hence, we can say

Theorem 11.3.1 Differentiable Implies Continuous: Scalar Function of n Variables

If f is differentiable at \mathbf{x}_0 then f is continuous at \mathbf{x}_0 .

Proof 11.3.1

We have sketched the argument already. ■

Comment 11.3.1 At each point \mathbf{x} where $Df(\mathbf{x})$ exists, we get a vector $\mathbf{L}(\mathbf{x})$. Hence, the derivative of a function of n variables defines a mapping from \mathbb{R}^n to \mathbb{R}^n by

$$Df(\mathbf{x}) = \mathbf{L}^T(\mathbf{x})$$

Comment 11.3.2 If f is differentiable at \mathbf{x}_0 , then there is a vector \mathbf{L} so that

$$f(\mathbf{x}) = f(\mathbf{x}_0 + \mathbf{L}^T(\mathbf{x} - \mathbf{x}_0) + \mathcal{E})(\mathbf{x}_0, \mathbf{x})$$

If we have chosen a basis for \mathbb{R}^n , then we have

$$f(\mathbf{x}) = f(\mathbf{x}_0 + [\mathbf{L}]_E^T ([\mathbf{x} - \mathbf{x}_0]_E) + \mathcal{E})([\mathbf{x}_0]_E, [\mathbf{x}]_E)$$

From this definition, we can show if the scalar function f is differentiable at the point \mathbf{x}_0 , then $L_i = f_i(\mathbf{x}_0)$ for all indices i . The argument goes like this: since f is differentiable at \mathbf{x}_0 , we can focus on what happens as we move from \mathbf{x}_0 to $\mathbf{x}_0 + h_i \mathbf{E}_i$. Hence, all the components of $\mathbf{x}_0 + h_i \mathbf{E}_i$ are the same as the ones in \mathbf{x}_0 . From the definition of differentiability we then have

$$\lim_{h_i \rightarrow 0} \frac{f(\mathbf{x} + h_i \mathbf{E}_i) - f(\mathbf{x}_0) - L_i h_i}{|h_i|} = 0.$$

as $\|\mathbf{x}_0 + h_i \mathbf{E}_i - \mathbf{x}_0\} = |h_i|$. For $h_i > 0$, we find

$$\lim_{h_i \rightarrow 0^+} \frac{f(\mathbf{x} + h_i \mathbf{E}_i) - f(\mathbf{x}_0) - L_i h_i}{h_i} = 0.$$

Hence $(f_i(\mathbf{x}_0))^+ = L_i$. Similarly, if $h_i < 0$, we have

$$\lim_{h_i \rightarrow 0^-} \frac{f(\mathbf{x} + h_i \mathbf{E}_i) - f(\mathbf{x}_0) - L_i h_i}{-h_i} = 0.$$

Hence $(f_i(\mathbf{x}_0))^- = L_i$ as well. Combining, we see $f_i(\mathbf{x}_0) = L_i$. The arguments for the other indices are much the same. Hence, we can say if the scalar function of n variables f is differentiable

at \mathbf{x}_0 then $f_i(\mathbf{x}_0)$ exist and

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \sum_{i=1}^n f_i(\mathbf{x}_0)(x_i - x_{0i}) + \mathcal{E}_f(\mathbf{x}_0, \mathbf{x})$$

where $\mathcal{E}_f(\mathbf{x}_0, \mathbf{x}) \rightarrow 0$ and $\mathcal{E}_f(\mathbf{x}_0, \mathbf{x})/||\mathbf{x} - \mathbf{x}_0|| \rightarrow 0$ as $\mathbf{x} \rightarrow \mathbf{x}_0$. Note this argument is a pointwise argument. It only tells us that differentiability at a point implies the existence of the partial derivatives at that point.

Theorem 11.3.2 The scalar function f is Differentiable implies $Df = (\nabla(f))^T$

If f is differentiable at \mathbf{x}_0 then $D\mathbf{F}(\mathbf{x}_0) = (\nabla(f))^T(\mathbf{x}_0)$.

Proof 11.3.2

We have just proven this. ■

Next, we look at the version of the chain rule for a scalar function of n variables.

11.3.1 The Chain Rule for Scalar Functions of n Variables

Let's review how you prove the chain rule. We assume there are two functions $u(x, y)$ and $v(x, y)$ defined locally about (x_0, y_0) and that there is a third function $f(u, v)$ which is defined locally around $(u_0 = u(x_0, y_0), v_0 = v(x_0, y_0))$. Now assume $f(u, v)$ is differentiable at (u_0, v_0) and $u(x, y)$ and $v(x, y)$ are differentiable at (x_0, y_0) . Then we can say

$$\begin{aligned} u(x, y) &= u(x_0, y_0) + u_x(x_0, y_0)(x - x_0) + u_y(x_0, y_0)(y - y_0) + \mathcal{E}_u(x_0, y_0, x, y) \\ v(x, y) &= v(x_0, y_0) + v_x(x_0, y_0)(x - x_0) + v_y(x_0, y_0)(y - y_0) + \mathcal{E}_v(x_0, y_0, x, y) \\ f(u, v) &= f(u_0, v_0) + f_u(u_0, v_0)(u - u_0) + f_v(u_0, v_0)(v - v_0) + \mathcal{E}_f(u_0, v_0, u, v) \end{aligned}$$

where all the error terms behave as usual as $(x, y) \rightarrow (x_0, y_0)$ and $(u, v) \rightarrow (u_0, v_0)$. Note that as $(x, y) \rightarrow (x_0, y_0)$, $u(x, y) \rightarrow u_0 = u(x_0, y_0)$ and $v(x, y) \rightarrow v_0 = v(x_0, y_0)$ as u and v are continuous at the (u_0, v_0) since they are differentiable there. Let's consider the partial of f with respect to x . Let $\Delta u = u(x_0 + \Delta x, y_0) - u(x_0, y_0)$ and $\Delta v = v(x_0 + \Delta x, y_0) - v(x_0, y_0)$. Thus, $u_0 + \Delta u = u(x_0 + \Delta x, y_0)$ and $v_0 + \Delta v = v(x_0 + \Delta x, y_0)$. Hence,

$$\begin{aligned} &\frac{f(u_0 + \Delta u, v_0 + \Delta v) - f(u_0, v_0)}{\Delta x} \\ &= \frac{f_u(u_0, v_0)(u - u_0) + f_v(u_0, v_0)(v - v_0) + \mathcal{E}_f(u_0, v_0, u, v)}{\Delta x} \\ &= f_u(u_0, v_0) \frac{u - u_0}{\Delta x} + f_v(u_0, v_0) \frac{v - v_0}{\Delta x} + \frac{\mathcal{E}_f(u_0, v_0, u, v)}{\Delta x} \end{aligned}$$

Continuing

$$\begin{aligned} &\frac{f(u_0 + \Delta u, v_0 + \Delta v) - f(u_0, v_0)}{\Delta x} = f_u(u_0, v_0) \frac{u_x(x_0, y_0)(x - x_0) + \mathcal{E}_u(x_0, y_0, x, y)}{\Delta x} \\ &+ f_v(u_0, v_0) \frac{v_x(x_0, y_0)(x - x_0) + \mathcal{E}_v(x_0, y_0, x, y)}{\Delta x} + \frac{\mathcal{E}_f(u_0, v_0, u, v)}{\Delta x} \\ &= f_u(u_0, v_0) u_x(x_0, y_0) + f_v(u_0, v_0) v_x(x_0, y_0) \\ &+ f_u(u_0, v_0) \frac{\mathcal{E}_u(x_0, y_0, x, y)}{\Delta x} + f_v(u_0, v_0) \frac{\mathcal{E}_v(x_0, y_0, x, y)}{\Delta x} + \frac{\mathcal{E}_f(u_0, v_0, u, v)}{\Delta x}. \end{aligned}$$

11.3. DERIVATIVES FOR SCALAR FUNCTIONS OF N VARIABLES

195

If f was locally constant, then

$$\mathcal{E}_f(u_0, v_0, u, v) = f(u, v) - f(u_0, v_0) - f_u(u_0, v_0)(u - u_0) - f_v((u_0, v_0)v - v_0) = 0$$

We know

$$\lim_{(a,b) \rightarrow (u_0, v_0)} \frac{\mathcal{E}_f(u_0, v_0, u, v)}{\|(a, b) - (u_0, v_0)\|} = 0$$

and this is true no matter what sequence (a_n, b_n) we choose that converges to $(u(x_0, y_0), v(x_0, y_0))$. If f is not locally constant, a little thought shows locally about $(u(x_0, y_0), v(x_0, y_0))$ there are sequences $(u_n, v_n) = (u(x_n, y_n), v(x_n, y_n)) \rightarrow (u_0, v_0) = (u(x_0, y_0), v(x_0, y_0))$ with $(u_n, v_n) \neq (u_0, v_0)$ for all n . Also, by continuity, as $\Delta x \rightarrow 0$, $(u_n, v_n) \rightarrow (u_0, v_0)$ too. For (u_n, v_n) from this sequence, we then have

$$\lim_{\Delta x \rightarrow 0} \frac{\mathcal{E}_f(u_0, v_0, u, v)}{\Delta x} = \left(\lim_{n \rightarrow \infty} \frac{\mathcal{E}_f(u_0, v_0, u, v)}{\|(u_n, v_n) - (u_0, v_0)\|} \right) \left(\lim_{\Delta x \rightarrow 0} \frac{\|(u_n, v_n) - (u_0, v_0)\|}{\Delta x} \right)$$

We know the first term goes to zero by the properties of \mathcal{E}_f . Expand the second term to get

$$\begin{aligned} & \left(\lim_{\Delta x \rightarrow 0} \frac{\|(u_n, v_n) - (u_0, v_0)\|}{\Delta x} \right) = \\ & \left(\lim_{\Delta x \rightarrow 0} \frac{\|(u_x(x_0, y_0)\Delta x + \mathcal{E}_u(x_0, y_0, x, y), v_x(x_0, y_0)\Delta x + \mathcal{E}_v(x_0, y_0, x, y))\|}{\Delta x} \right) \leq \\ & |u_x(x_0, y_0)| + |v_x(x_0, y_0)| \end{aligned}$$

Hence, the product limit is zero. The other error terms go to zero also as $(x, y) \rightarrow (x_0, y_0)$. Hence, we conclude

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial x}.$$

A similar argument shows

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial u} \frac{\partial u}{\partial y} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial y}.$$

This result is known as the **Chain Rule** and we will find there is a better way to package this. Now assume there are more variables. We assume there are three functions $u_1(x_1, x_2, x_3)$, $u_2(x_1, x_2, x_3)$ and $u_3(x_1, x_2, x_3)$ defined locally about (x_{10}, x_{20}, x_{30}) and that there is another function $f(u_1, u_2, u_3)$ which are defined locally around $(u_{i0} = u_i(x_{10}, x_{20}, x_{30}))$. Now assume $f(u_1, u_2, u_3)$ is differentiable at (u_{10}, u_{20}, u_{30}) and $u_i(x_1, x_2, x_3)$ are differentiable at (x_{10}, x_{20}, x_{30}) . Then we can say

$$\begin{aligned} u_i(x_1, x_2, x_3) &= u_i(x_{10}, x_{20}, x_{30}) + u_{i,x_1}(x_{10}, x_{20}, x_{30})(x_1 - x_{10}) \\ &\quad + u_{i,x_2}(x_{10}, x_{20}, x_{30})(x_2 - x_{20}) + u_{i,x_3}(x_{10}, x_{20}, x_{30})(x_3 - x_{30}) \\ &\quad + \mathcal{E}_{u_i}(x_{10}, x_{20}, x_{30}, x_1, x_2, x_3) \end{aligned}$$

where all the error terms behave as usual as $(x_1, x_2, x_3) \rightarrow (x_{10}, x_{20}, x_{30})$ and $(u_1, u_2, u_3) \rightarrow (u_{10}, u_{20}, u_{30})$. This is really messy to write down! To save space, let $\mathbf{x} = (x_1, x_2, x_3)$, $\mathbf{u} = (u_1, u_2, u_3)$ etc. Then we can rewrite the above as

$$u_i(\mathbf{x}) = u_i(\mathbf{x}_0) + u_{i,x_1}(\mathbf{x}_0)(x_1 - x_{10}) + u_{i,x_2}(\mathbf{x}_0)(x_2 - x_{20}) + u_{i,x_3}(\mathbf{x}_0)(x_3 - x_{30})$$

$$+\mathcal{E}_{u_i}(\mathbf{x}_0 \cdot \mathbf{x})$$

Again, note as $\mathbf{x} \rightarrow \mathbf{x}_0$, $u_i(\mathbf{x}) \rightarrow u_{i0} = u_i(\mathbf{x}_0)$ as u_i is continuous at \mathbf{x}_0 as it is differentiable there. Let's consider the partial of f with respect to x_j . Let $\Delta u_i = u_j(\mathbf{x}_0 + h_j \mathbf{E}_j) - u_j(\mathbf{x}_0)$. Thus, $u_{i0} + \Delta u_i = u_i(\mathbf{x}_0 + h_j \mathbf{E}_j)$. So we are using h_j to denote the changes in x_j which induce changes in u_i . We see

$$\begin{aligned} & \frac{f(u_{10} + \Delta u_1, u_{20} + \Delta u_2, u_{30} + \Delta u_3) - f(\mathbf{u}_0)}{h_j} = \\ & \frac{f_{u_1}(\mathbf{u}_0)(u_1 - u_{10}) + f_{u_2}(\mathbf{u}_0)(u_2 - u_{20}) + f_{u_3}(\mathbf{u}_0)(u_3 - u_{30})}{h_j} + \frac{\mathcal{E}_f(\mathbf{u}_0, \mathbf{u})}{h_j} = \\ & \sum_{i=1}^3 f_{u_i}(\mathbf{u}_0) \frac{(u_i - u_{i0})}{h_j} + \frac{\mathcal{E}_f(\mathbf{u}_0, \mathbf{u})}{h_j} \end{aligned}$$

Now replace $u_i - u_{i0}$ by its expansion:

$$\begin{aligned} u_i - u_{i0} &= \sum_{k=1}^3 u_{i,x_k}(\mathbf{x}_0)(x_k - x_{k0}) + \mathcal{E}_{u_i}(\mathbf{x}_0 \cdot \mathbf{x}) = u_{i,x_j}(\mathbf{x}_0)(x_j - x_{j0}) + \mathcal{E}_{u_i}(\mathbf{x}_0 \cdot \mathbf{x}) \\ &= u_{i,x_j}(\mathbf{x}_0) h_j + \mathcal{E}_{u_i}(\mathbf{x}_0 \cdot \mathbf{x}) \end{aligned}$$

This gives

$$\begin{aligned} & \frac{f(u_{10} + \Delta u_1, u_{20} + \Delta u_2, u_{30} + \Delta u_3) - f(\mathbf{u}_0)}{h_j} = \\ & \sum_{i=1}^3 f_{u_i}(\mathbf{u}_0) u_{i,x_i}(\mathbf{x}_0) + \sum_{i=1}^3 f_{u_i}(\mathbf{u}_0) \frac{\mathcal{E}_{u_i}(\mathbf{x}_0 \cdot \mathbf{x})}{h_j} + \frac{\mathcal{E}_f(\mathbf{u}_0, \mathbf{u})}{h_j} \end{aligned}$$

As $h_j \rightarrow 0$, $\Delta u_i \rightarrow 0$ and using the arguments we used in the two variable case, it is straightforward to show $\mathcal{E}_f(\mathbf{u}_0, \mathbf{u})/h_j \rightarrow 0$. The other two error terms go to zero also as $(x, y) \rightarrow (x_0, y_0)$. Hence, we conclude

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^3 \frac{\partial f}{\partial u_i} \frac{\partial u_i}{\partial x_j}$$

It is straightforward to see how to extend this result to n variables. Just a bit messy! The notation here is quite messy and confusing. Using the $\frac{\partial}{\partial(\arg:k)}$ doesn't work well as we have arg terms for both the x_j and u_i variables. Also, in our argument above we had $\mathbf{u} : D_u \subset \mathbb{R}^3 \rightarrow \mathbb{R}^3$ with u is defined locally about \mathbf{x}_0 . Hence, \mathbf{u} is a function of (x_1, \dots, x_3) with $\mathbf{u}(x_1, \dots, x_3) = (u_1, \dots, u_3) \in \mathbb{R}^3$. Further, we assumed there is another function $f : D_f \subset \mathbb{R}^3 \rightarrow \mathbb{R}$ which is defined locally around $(\mathbf{u}_0 = u(\mathbf{x}_0)) \in \mathbb{R}^3$. And we assume $f(\mathbf{u})$ is differentiable at \mathbf{u}_0 and $\mathbf{u}(\mathbf{x})$ is differentiable at \mathbf{x}_0 . However, we could be more general: we could assume $\mathbf{u} : D_u \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ with u is defined locally about \mathbf{x}_0 . Hence, \mathbf{u} is a function of (x_1, \dots, x_n) with $\mathbf{u}(x_1, \dots, x_n) = (u_1, \dots, u_m) \in \mathbb{R}^m$. Further, assume there is another function $f : D_f \subset \mathbb{R}^m \rightarrow \mathbb{R}$ which is defined locally around $(\mathbf{u}_0 = u(\mathbf{x}_0)) \in \mathbb{R}^m$. Further assume $f(\mathbf{u})$ is differentiable at \mathbf{u}_0 and $\mathbf{u}(\mathbf{x})$ is differentiable at \mathbf{x}_0 . The arguments are quite similar and we find

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^m \frac{\partial f}{\partial u_i} \frac{\partial u_i}{\partial x_j}$$

This leads to our theorem.

Theorem 11.3.3 The Chain Rule for Scalar Functions of n Variables

Assume $\mathbf{u} : D_u \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ with u is defined locally about \mathbf{x}_0 . Hence, \mathbf{u} is a function of (x_1, \dots, x_n) with $\mathbf{u}(x_1, \dots, x_n) = (u_1, \dots, u_m) \in \mathbb{R}^m$. Further, assume there is another function $f : D_f \subset \mathbb{R}^m \rightarrow \mathbb{R}$ which is defined locally around $(\mathbf{u}_0 = u(\mathbf{x}_0)) \in \mathbb{R}^m$. Further assume $f(\mathbf{u})$ is differentiable at \mathbf{u}_0 and $\mathbf{u}(\mathbf{x})$ is differentiable at \mathbf{x}_0 . Then f_i exists at \mathbf{x}_0 and is given by

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^m \frac{\partial f}{\partial u_i} \frac{\partial u_i}{\partial x_j}$$

Proof 11.3.3

We have just gone over the argument. ■

Comment 11.3.3 With a little bit of work, these arguments also prove the differentiability of the composition. In the two variable case, let $g(x, y) = f(u(x, y), v(x, y))$, i.e. $g = f \circ U$ where U is the vector function with components u and v , and define the error

$$\mathcal{E}_g = f_u(u_0, v_0) \mathcal{E}_u(x_0, y_0, x, y) + f_v(u_0, v_0) \mathcal{E}_v(x_0, y_0, x, y) + \mathcal{E}_f(u_0, v_0, u, v).$$

It is clear from the assumptions that $\mathcal{E}_g(x_0, y_0, x, y) \rightarrow 0$ as $\Delta x \rightarrow 0$ and $\mathcal{E}_g(x_0, y_0, x, y) \rightarrow 0$ as $\Delta y \rightarrow 0$. The arguments above show $\frac{\mathcal{E}_g(x_0, y_0, x, y)}{\Delta x} \rightarrow 0$ as $\Delta x \rightarrow 0$ and $\frac{\mathcal{E}_g(x_0, y_0, x, y)}{\Delta y} \rightarrow 0$ as $\Delta y \rightarrow 0$. This shows \mathcal{E}_g is the error function for the composition $g = f \circ U$. Hence $g = f \circ U$ is differentiable at (x_0, y_0) with

$$\begin{aligned} \mathbf{Dg}(x_0, y_0) &= \begin{bmatrix} f_u(u_0, v_0) \\ f_v(u_0, v_0) \end{bmatrix}^T \begin{bmatrix} u_x(x_0, y_0) & v_x(x_0, y_0) \\ u_y(x_0, y_0) & v_y(x_0, y_0) \end{bmatrix} \\ &= \mathbf{Df}(u_0, v_0) \begin{bmatrix} u_x(x_0, y_0) & v_x(x_0, y_0) \\ u_y(x_0, y_0) & v_y(x_0, y_0) \end{bmatrix} \end{aligned}$$

We will discuss this further.

Example 11.3.1 Let $f(x, y, z) = x^2 + 2x + 5y^4 + z^2$. Then if $x = r \sin(\phi) \cos(\theta)$, $y = r \sin(\phi) \sin(\theta)$ and $z = r \cos(\phi)$, $y = r \sin(\theta)$, using the chain rule, we find

$$\begin{aligned} \frac{\partial f}{\partial r} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial r} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial r} + \frac{\partial f}{\partial z} \frac{\partial z}{\partial r} \\ \frac{\partial f}{\partial \theta} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial \theta} + \frac{\partial f}{\partial z} \frac{\partial z}{\partial \theta} \\ \frac{\partial f}{\partial \phi} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial \phi} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial \phi} + \frac{\partial f}{\partial z} \frac{\partial z}{\partial \phi} \end{aligned}$$

This becomes

$$\begin{aligned} \frac{\partial f}{\partial r} &= \frac{\partial f}{\partial x}(\sin(\phi) \cos(\theta)) + \frac{\partial f}{\partial y}(-\sin(\phi) \sin(\theta)) + \frac{\partial f}{\partial z}(\cos(\phi)) \\ \frac{\partial f}{\partial \theta} &= \frac{\partial f}{\partial x}(-r \sin(\phi) \sin(\theta)) + \frac{\partial f}{\partial y}r \sin(\phi) \cos(\theta) + \frac{\partial f}{\partial z}(0) \end{aligned}$$

$$\frac{\partial f}{\partial \phi} = \frac{\partial f}{\partial x}(r \cos(\phi) \cos(\theta)) + \frac{\partial f}{\partial y}(r \cos(\phi) \sin(\theta)) + \frac{\partial f}{\partial z}(-r \sin(\phi))$$

You can then substitute in for x and y to get the final answer in terms of r , θ and ϕ .

Example 11.3.2 Let $u = x^2 + 2y^2 + z^2$ and $v = 4x^2 - 5y^2 + 7z$ and $f(u, v) = 10u^2v^4$. Then, by the chain rule

$$\begin{aligned}\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial x} \\ \frac{\partial f}{\partial y} &= \frac{\partial f}{\partial u} \frac{\partial u}{\partial y} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial y} \\ \frac{\partial f}{\partial z} &= \frac{\partial f}{\partial u} \frac{\partial u}{\partial z} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial z}\end{aligned}$$

This becomes

$$\begin{aligned}\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial u}(2x) + \frac{\partial f}{\partial v}(8x) \\ \frac{\partial f}{\partial y} &= \frac{\partial f}{\partial u}(4y) + \frac{\partial f}{\partial v}(-10y) \\ \frac{\partial f}{\partial z} &= \frac{\partial f}{\partial u}(2z) + \frac{\partial f}{\partial v}(7)\end{aligned}$$

You can then substitute in for u and v to get the final answer in terms of x , y and z .

11.3.1.1 Homework

Exercise 11.3.1

Exercise 11.3.2

Exercise 11.3.3

11.4 Partials and Differentiability

First, let's talk about some issues with smoothness for partials. We cover the following examples in (Peterson (8) 2019) also, but it is worth seeing them again. These examples are intense but you learn a lot by working through them.

11.4.1 Partials Can Exist but not be Continuous

We start with the function f defined by

$$f(x, y) = \begin{cases} \frac{2xy}{\sqrt{x^2+y^2}}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0) \end{cases}$$

This function has a removable discontinuity at $(0, 0)$ because

$$\begin{aligned}\left| \frac{2xy}{\sqrt{x^2+y^2}} - 0 \right| &= 2 \frac{|x| |y|}{\sqrt{x^2+y^2}} \leq 2 \frac{\sqrt{x^2+y^2}}{\sqrt{x^2+y^2}} \sqrt{x^2+y^2} \\ &= 2\sqrt{x^2+y^2}\end{aligned}$$

because $|x| \leq \sqrt{x^2 + y^2}$ and $|y| \leq \sqrt{x^2 + y^2}$. Pick $\epsilon > 0$ arbitrarily. Then if $\delta < \epsilon/2$, we have

$$\left| \frac{2xy}{\sqrt{x^2 + y^2}} - 0 \right| < 2 \cdot \frac{\epsilon}{2} = \epsilon$$

which proves that the limit exists and equals 0 as $(x, y) \rightarrow (0, 0)$. This tells us we can define $f(0, 0)$ to match this value. Thus, as said, this is a *removable discontinuity*.

The first order partials both exist at $(0, 0)$ as is seen by an easy limit.

$$\begin{aligned} f_x(0, 0) &= \lim_{x \rightarrow 0} \frac{f(x, 0) - f(0, 0)}{x} = \lim_{x \rightarrow 0} \frac{0 - 0}{x} = 0. \\ f_y(0, 0) &= \lim_{y \rightarrow 0} \frac{f(0, y) - f(0, 0)}{y} = \lim_{y \rightarrow 0} \frac{0 - 0}{y} = 0. \end{aligned}$$

We can also calculate the first order partials at any $(x, y) \neq (0, 0)$.

$$\begin{aligned} f_x(x, y) &= \frac{2y(x^2 + y^2)^{1/2} - 2xy(1/2)(x^2 + y^2)^{-1/2} 2x}{(x^2 + y^2)^1} \\ &= \frac{2y(x^2 + y^2) - 2x^2y}{(x^2 + y^2)^{3/2}} = \frac{2y^3}{(x^2 + y^2)^{3/2}} \\ f_y(x, y) &= \frac{2x(x^2 + y^2)^{1/2} - 2xy(1/2)(x^2 + y^2)^{-1/2} 2y}{(x^2 + y^2)^1} \\ &= \frac{2x(x^2 + y^2) - 2xy^2}{(x^2 + y^2)^{3/2}} = \frac{2x^3}{(x^2 + y^2)^{3/2}} \end{aligned}$$

We can show $\lim_{(x, y) \rightarrow (0, 0)} f_x(x, y)$ and $\lim_{(x, y) \rightarrow (0, 0)} f_y(x, y)$ do not exist as follows. We will find paths (x, mx) for $m \neq 0$ where we get different values for the limit depending on the value of m . We pick the path $(x, 5x)$ for $x > 0$. On this path, we find

$$\lim_{x \rightarrow 0^+} f_x(x, 5x) = \lim_{x \rightarrow 0^+} \frac{250x^3}{(26)^{3/2}|x^2|^{3/2}} = \lim_{x \rightarrow 0^+} \frac{250}{26\sqrt{26}}$$

Then, for example, using $m = -5$ for $x < 0$, we get

$$\lim_{x \rightarrow 0^+} f_x(x, -5x) = \lim_{x \rightarrow 0^+} \frac{-250x^3}{(26)^{3/2}|x^2|^{3/2}} = \lim_{x \rightarrow 0^+} \frac{-250}{26\sqrt{26}}$$

Since these are not the same, this limit can not exist. Hence f_x is not continuous at the origin. A similar argument shows $\lim_{(x, y) \rightarrow (0, 0)} f_y(x, y)$ does not exist either. So we have shown the first order partials exist locally around $(0, 0)$ but f_x and f_y are not continuous at $(0, 0)$.

Is f differentiable at $(0, 0)$? If so, there are numbers L_1 and L_2 so that two things happen:

$$\begin{aligned} \lim_{(x, y) \rightarrow (0, 0)} \left(\frac{2xy}{\sqrt{x^2 + y^2}} - L_1x - L_2y \right) &= 0 \\ \lim_{(x, y) \rightarrow (0, 0)} \frac{1}{\sqrt{x^2 + y^2}} \left(\frac{2xy}{\sqrt{x^2 + y^2}} - L_1x - L_2y \right) &= 0 \end{aligned}$$

The first limit is 0 because

$$\lim_{(x, y) \rightarrow (0, 0)} \left| \frac{2xy}{\sqrt{x^2 + y^2}} - L_1x - L_2y - 0 \right| \leq$$

$$\begin{aligned} & 2 \frac{|x||y|}{\sqrt{x^2 + y^2}} + |L_1| \sqrt{x^2 + y^2} + |L_2| \sqrt{x^2 + y^2} \\ & \leq 2 \frac{\sqrt{x^2 + y^2} \sqrt{x^2 + y^2}}{\sqrt{x^2 + y^2}} + |L_1| \sqrt{x^2 + y^2} + |L_2| \sqrt{x^2 + y^2} \\ & \leq (2 + |L_1| + |L_2|) \sqrt{x^2 + y^2} \end{aligned}$$

which goes to 0 as $(x, y) \rightarrow 0$. So that shows the first requirement is met.

However, the second requirement is not. Look at the paths (x, mx) and consider $\lim_{x \rightarrow 0^+}$ like we did before. We find, since $|x| = x$ here

$$\begin{aligned} \lim_{x \rightarrow 0^+} \frac{1}{x\sqrt{1+m^2}} \left(\frac{2mx^2}{x\sqrt{1+m^2}} - L_1x - mL_2x \right) &= \\ \frac{1}{\sqrt{1+m^2}} \left(\frac{2mx}{x\sqrt{1+m^2}} - L_1 - mL_2m \right) &= \frac{2m}{1+m^2} - \frac{L_1 + L_2m}{\sqrt{1+m^2}} \end{aligned}$$

Now look at path $(x, -mx)$ for $x > 0$. This is the limit from the other side. We have

$$\begin{aligned} \lim_{x \rightarrow 0^+} \frac{1}{x\sqrt{1+m^2}} \left(\frac{-2mx^2}{x\sqrt{1+m^2}} - L_1x + mL_2x \right) &= \\ \frac{1}{\sqrt{1+m^2}} \left(\frac{-2m}{\sqrt{1+m^2}} - L_1 + mL_2m \right) &= \frac{-2m}{1+m^2} + \frac{-L_1 + L_2m}{\sqrt{1+m^2}} \end{aligned}$$

If these limits were equal we would have

$$\frac{2m}{1+m^2} + \frac{-L_1 - L_2m}{\sqrt{1+m^2}} = \frac{-2m}{1+m^2} + \frac{-L_1 + L_2m}{\sqrt{1+m^2}}$$

This implies $L_2 = 2/\sqrt{1+m^2}$ which tells us L_2 depends on m . These values should be independent of the choice of path and so this function cannot be differentiable at $(0, 0)$.

We see that just because the first order partials exist locally in some $B_r(0, 0)$ does not necessarily tell us f is differentiable at $(0, 0)$. So **differentiable** at (x_0, y_0) implies the first order partials exist at (x_0, y_0) and $L_1 = f_x(x_0, y)$ and $L_2 = f_y(x_0, y)$. **But the converse is not true in general.**

We need to figure out the appropriate assumptions about the smoothness of f that will guarantee the existence of the derivative of f in this multivariable situations. We will do this for n dimensions as it is just as easy as doing the proof in \mathbb{R}^2 .

Theorem 11.4.1 Sufficient Conditions to Guarantee Differentiability

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ and assume $\nabla(f)$ exists and is continuous in a ball $B_r(\mathbf{p})$ of some radius $r > 0$ around the point $\mathbf{p} = [p_1, \dots, p_n]'$ in \mathbb{R}^n . Then f is differentiable at each point in $B_r(\mathbf{p})$.

Proof 11.4.1

By definition of $\frac{\partial f}{\partial x_i}$ at \mathbf{p} ,

$$f(\mathbf{p} + h\mathbf{E}_i) - f(\mathbf{p}) = \frac{\partial f}{\partial x_i}(\mathbf{p})h + \mathcal{E}_i(\mathbf{p}, \mathbf{p} + h\mathbf{E}_i)$$

Now if \mathbf{x} is in $B(r, \mathbf{p})$, all the first order partials of f are continuous. We have

$$\mathbf{x} = \mathbf{p} + (x_1 - p_1)\mathbf{E}_1 + (x_2 - p_2)\mathbf{E}_2 + \dots + (x_n - p_n)\mathbf{E}_n$$

Now let $r_i = x_i - p_i$. Then we can write (this is the telegraphing sum trick!)

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{p}) &= f(\mathbf{p} + \sum_{i=1}^n r_i \mathbf{E}_i) - f(\mathbf{p} + \sum_{i=1}^{n-1} r_i \mathbf{E}_i) + f(\mathbf{p} + \sum_{i=1}^{n-1} r_i \mathbf{E}_i) - f(\mathbf{p} + \sum_{i=1}^{n-2} r_i \mathbf{E}_i) \\ &\quad + f(\mathbf{p} + \sum_{i=1}^{n-2} r_i \mathbf{E}_i) - f(\mathbf{p} + \sum_{i=1}^{n-3} r_i \mathbf{E}_i) + \dots \\ &\quad + f(\mathbf{p} + \sum_{i=1}^2 r_i \mathbf{E}_i) - f(\mathbf{p} + \sum_{i=1}^1 r_i \mathbf{E}_i) + f(\mathbf{p} + \sum_{i=1}^1 r_i \mathbf{E}_i) - f(\mathbf{p}) \end{aligned}$$

To help you see this better, if $n = 3$, this expansion would look like

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{p}) &= f(\mathbf{p} + \sum_{i=1}^3 r_i \mathbf{E}_i) - f(\mathbf{p} + \sum_{i=1}^2 r_i \mathbf{E}_i) + f(\mathbf{p} + \sum_{i=1}^2 r_i \mathbf{E}_i) - f(\mathbf{p} + \sum_{i=1}^1 r_i \mathbf{E}_i) \\ &\quad + f(\mathbf{p} + \sum_{i=1}^1 r_i \mathbf{E}_i) - f(\mathbf{p}) \end{aligned}$$

or

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{p}) &= f(p_1 + r_1, p_2 + r_2, p_3 + r_3) - f(p_1 + r_1, p_2 + r_2, p_3) \Delta \text{ slot 3} \\ &\quad + f(p_1 + r_1, p_2 + r_2, p_3) - f(p_1 + r_1, p_2, p_3) \Delta \text{ slot 2} \\ &\quad + f(p_1 + r_1, p_2, p_3) - f(p_1, p_2, p_3) \Delta \text{ slot 1} \end{aligned}$$

So we have

$$f(\mathbf{x}) - f(\mathbf{p}) = \sum_{j=n}^1 (f(\mathbf{p} + \sum_{i=1}^j r_i \mathbf{E}_i) - f(\mathbf{p} + \sum_{i=1}^{j-1} r_i \mathbf{E}_i))$$

where we interpret $\sum_{j=1}^0$ as 0. To see this done without all this notation, you can look at the argument for the two variable version of this theorem in (Peterson (8) 2019). That argument is messy too though and in someways, this more abstract form is better!

Now apply the Mean Value Theorem in one variable to get

$$f(\mathbf{p} + \sum_{i=1}^j r_i \mathbf{E}_i) - f(\mathbf{p} + \sum_{i=1}^{j-1} r_i \mathbf{E}_i) = \frac{\partial f}{\partial x_i}(\mathbf{q}_j) r_j$$

where \mathbf{q}_j is between $\mathbf{p} + \sum_{i=1}^{j-1} r_i \mathbf{E}_i$ and $\mathbf{p} + \sum_{i=1}^j r_i \mathbf{E}_i$; i.e. $\mathbf{q}_j = \mathbf{p} + \sum_{i=1}^{j-1} r_i \mathbf{E}_i + s_j \mathbf{E}_j$ with $|s_j| < |r_j| < r$. So

$$f(\mathbf{x}) - f(\mathbf{p}) = \sum_{j=n}^1 \frac{\partial f}{\partial x_i}(\mathbf{q}_j) r_j$$

Since $\frac{\partial f}{\partial x_j}$ is continuous on $B_r(\mathbf{p})$, given $\epsilon > 0$ arbitrary, there are $\delta_j > 0$ so that for $1 \leq j \leq n$, $|\frac{\partial f}{\partial x_j}(\mathbf{x}) - \frac{\partial f}{\partial x_j}(\mathbf{p})| < \epsilon/n$. Thus, if $\|\mathbf{x} - \mathbf{p}\| < \hat{\delta} = \min\{\delta_1, \dots, \delta_n\}$, since the points \mathbf{q}_j are between $\mathbf{p} + \sum_{i=1}^{j-1} r_i \mathbf{E}_i$ and $\mathbf{p} + \sum_{i=1}^j r_i \mathbf{E}_i$, we have But, as we mentioned earlier, s_j is between p_j and $p_j + r_j$ which implies \mathbf{q}_j is in $B_r(\mathbf{p})$.

So we have $|\frac{\partial f}{\partial x_j}(\mathbf{q}_j) - \frac{\partial f}{\partial x_j}(\mathbf{p})| < \epsilon/n$. Thus, we find

$$\frac{\partial f}{\partial x_j}(\mathbf{p}) - \epsilon/n < \frac{\partial f}{\partial x_j}(\mathbf{q}_j) < \frac{\partial f}{\partial x_j}(\mathbf{p}) + \epsilon/n$$

Now $f(\mathbf{x}) - f(\mathbf{p}) = \sum_{j=1}^n \frac{\partial f}{\partial x_j}(\mathbf{q}_j) r_j$. Let I be the set of indices where $r_j > 0$ and J be the other indices. Then $\sum_{j=1}^n r_j = \sum_{j \in I} r_j + \sum_{j \in J} r_j$. For indices $j \in I$, multiplying the inequality $r_j > 0$ causes no changes:

$$\left(\sum_{j \in I} \frac{\partial f}{\partial x_j}(\mathbf{p}) r_j \right) - (\epsilon/n) \sum_{j \in I} r_j < \sum_{j \in I} \frac{\partial f}{\partial x_j}(\mathbf{q}_j) r_j < \left(\sum_{j \in I} \frac{\partial f}{\partial x_j}(\mathbf{p}) \right) + (\epsilon/n) \sum_{j \in I} r_j$$

If $j \in J$, $r_j < 0$ and multiplying changes the inequalities. We get

$$\left(\sum_{j \in J} \frac{\partial f}{\partial x_j}(\mathbf{p}) r_j \right) + (\epsilon/n) \sum_{j \in J} r_j < \sum_{j \in J} \frac{\partial f}{\partial x_j}(\mathbf{q}_j) r_j < \left(\sum_{j \in J} \frac{\partial f}{\partial x_j}(\mathbf{p}) \right) - (\epsilon/n) \sum_{j \in J} r_j$$

We can combine these two forms by noting that we can replace r_j by $\text{sign}(r_j)r_j$ in each inequality. This gives

$$\begin{aligned} & \left(\sum_{j=1}^n \frac{\partial f}{\partial x_j}(\mathbf{p}) r_j \right) - \sum_{j=1}^n \text{sign}(r_j) (\epsilon/n) r_j < \sum_{j=1}^n \frac{\partial f}{\partial x_j}(\mathbf{q}_j) r_j \\ &= f(\mathbf{x}) - f(\mathbf{p}) < \left(\sum_{j=1}^n \frac{\partial f}{\partial x_j}(\mathbf{p}) \right) + \sum_{j=1}^n \text{sign}(r_j)(\epsilon/n) r_j \end{aligned}$$

Thus,

$$|f(\mathbf{x}) - f(\mathbf{p}) - \sum_{j=1}^n \frac{\partial f}{\partial x_j}(\mathbf{p}) r_j| < (\epsilon/n) \sum_{j=1}^n |r_j|$$

But $(\epsilon/n) \sum_{j=1}^n |r_j| \leq \sum_{j=1}^n (\epsilon/n) \|\mathbf{x} - \mathbf{p}\| = \epsilon \|\mathbf{x} - \mathbf{p}\|$. We therefore have the estimate

$$|f(\mathbf{x}) - f(\mathbf{p}) - \sum_{j=1}^n \frac{\partial f}{\partial x_j}(\mathbf{p}) r_j| < \epsilon \|\mathbf{x} - \mathbf{p}\|$$

The term inside the absolute values is the error function $E(\mathbf{x}, \mathbf{p})$ and so we have $E(\mathbf{x}, \mathbf{p})/\|\mathbf{x} - \mathbf{p}\| < \epsilon$ if $\|\mathbf{x} - \mathbf{p}\| < \hat{\delta}$. This tells us both $E(\mathbf{x}, \mathbf{p})/\|\mathbf{x} - \mathbf{p}\|$ and $E(\mathbf{x}, \mathbf{p})$ go to 0 as $\mathbf{x} \rightarrow \mathbf{p}$. Hence, f is differentiable at \mathbf{p} . ■

11.4.2 Higher Order Partials

We define the second order partials of f as follows.

Definition 11.4.1 Second Order Partials

If $f(\mathbf{x})$, $f_{x_i}(\mathbf{x})$ are defined locally at \mathbf{x}_0 , we can attempt to find following limits:

$$\lim_{h \rightarrow 0} \frac{f_{x_i}(\mathbf{x}_0 + h\mathbf{E}_j) - f_{x_i}(\mathbf{x}_0)}{h}$$

If the limit exists, it is called the partial of f_{x_i} with respect to x_j and is denoted by $f_{x_i x_j}$. Note, the order is potentially important as

$$\lim_{h \rightarrow 0} \frac{f_{x_j}(\mathbf{x}_0 + h\mathbf{E}_i) - f_{x_j}(\mathbf{x}_0)}{h}$$

if it exists is the partial of f_{x_j} with respect to x_i which is denoted by $f_{x_j x_i}$. Of course, these do not have to match in general.

Comment 11.4.1 When these second order partials exist at \mathbf{x}_0 , we use the following notations interchangeably: $f_{ij} = \partial_{x_j}(f_{x_i})$, $f_{ji} = \partial_{x_i}(f_{x_j})$. As usual, these notations can get confusing.

The second order partials for a scalar function of n variables can be organized into a matrix called the **Hessian**.

Definition 11.4.2 The Hessian Matrix

If $f(x, y)$, f_x and f_y are defined locally at (x_0, y_0) and if the second order partials exist at (x_0, y_0) , we define the Hessian, $\mathbf{H}(\mathbf{x}_0)$ at (\mathbf{x}_0) to be the matrix

$$\mathbf{H}(\mathbf{x}_0) = \begin{bmatrix} f_{x_1 x_1}(\mathbf{x}_0) & \dots & f_{x_1 x_n}(\mathbf{x}_0) \\ \vdots & \vdots & \vdots \\ f_{x_n x_1}(\mathbf{x}_0) & \dots & f_{x_n x_n}(\mathbf{x}_0) \end{bmatrix} = \begin{bmatrix} f_{11}(\mathbf{x}_0) & \dots & f_{1n}(\mathbf{x}_0) \\ \vdots & \vdots & \vdots \\ f_{n1}(\mathbf{x}_0) & \dots & f_{nn}(\mathbf{x}_0) \end{bmatrix}$$

Hence for a scalar function of n variables, we have the gradient, ∇f which is the derivative of f , Df , in some circumstances and is an $n \times 1$ column vector, and the Hessian of f which is a $n \times n$ matrix. If we have a vector function, it is messier. The analogue of the gradient for $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the Jacobian of f , \mathbf{J}_f given by

$$\mathbf{J}_f(\mathbf{x}_0) = \begin{bmatrix} f_{1,x_1}(\mathbf{x}_0) & \dots & f_{1,x_n}(\mathbf{x}_0) \\ \vdots & \vdots & \vdots \\ f_{m,x_1}(\mathbf{x}_0) & \dots & f_{m,x_n}(\mathbf{x}_0) \end{bmatrix} = \begin{bmatrix} f_{11}(\mathbf{x}_0) & \dots & f_{1n}(\mathbf{x}_0) \\ \vdots & \vdots & \vdots \\ f_{m1}(\mathbf{x}_0) & \dots & f_{mn}(\mathbf{x}_0) \end{bmatrix}$$

where f_i are the component functions of f . Note the Jacobian of f is the derivative of f , Df , in some circumstances and is an $n \times m$ matrix. You can see why the notation f_{ij} is so confusing. If f is a scalar function this has a different meaning than if f is a vector valued function. So context is all. We can also define higher order derivatives such as $f_{x_i x_j x_k}$ which have more permutations. We will let you work through those notations as the need arises in your work.

Example 11.4.1 Let $f(x, y, z) = 2x - 8xy + 4yz^2$. Find the first and second order partials of f and its Hessian.

Solution The partials are

$$\begin{aligned} f_x(x, y, z) &= 2 - 8y, & f_y(x, y, z) &= -8x + 4z^2, & f_z(x, y, z) &= 8yz \\ f_{xx}(x, y, z) &= 0, & f_{xy}(x, y, z) &= -8, & f_{xz}(x, y, z) &= 0 \\ f_{yx}(x, y, z) &= -8, & f_{yy}(x, y, z) &= 0, & f_{yz}(x, y, z) &= 8z \end{aligned}$$

$$f_{zx}(x, y, z) = 0, \quad f_{zy}(x, y, z) = 8z, \quad f_{zz}(x, y, z) = 8y$$

and so the Hessian is

$$\mathbf{H}(x, y, z) = \begin{bmatrix} f_{xx}(x, y, z) & f_{xy}(x, y, z) & f_{xz}(x, y, z) \\ f_{yx}(x, y, z) & f_{yy}(x, y, z) & f_{yz}(x, y, z) \\ f_{zx}(x, y, z) & f_{zy}(x, y, z) & f_{zz}(x, y, z) \end{bmatrix} = \begin{bmatrix} 0 & -8 & 0 \\ -8 & 0 & 8z \\ 0 & 8z & 8y \end{bmatrix}$$

11.4.2.1 Homework

Exercise 11.4.1

Exercise 11.4.2

Exercise 11.4.3

Exercise 11.4.4

Exercise 11.4.5

11.4.3 When Do Mixed Partials Match?

Next, we want to know under what conditions the mixed order partials match. Here is a standard example with a function of two variables showing the two mixed partials don't agree at a point.

Example 11.4.2 Let

$$f(x, y) = \begin{cases} \frac{xy(x^2 - y^2)}{x^2 + y^2}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0) \end{cases}$$

Prove the mixed order partials do not match.

- Prove $\lim_{(x,y) \rightarrow (0,0)} f(x, y) = 0$ thereby showing f has a removable discontinuity at $(0, 0)$ and so the definition for f above makes sense.

Solution

$$\begin{aligned} \lim_{(x,y) \rightarrow (0,0)} \left| \frac{xy(x^2 - y^2)}{x^2 + y^2} \right| &= \lim_{(x,y) \rightarrow (0,0)} \left| \frac{xy(x+y)}{x^2 + y^2} \right| |x-y| = \lim_{(x,y) \rightarrow (0,0)} \left| \frac{x^2y + y^2x}{x^2 + y^2} \right| |x-y| \\ &\leq \lim_{(x,y) \rightarrow (0,0)} \left(\left| \frac{x^2y}{x^2 + y^2} \right| + \left| \frac{y^2x}{x^2 + y^2} \right| \right) |x-y| \\ &\leq \lim_{(x,y) \rightarrow (0,0)} (|y| + |x|) |x-y| \leq 4(x^2 + y^2) = 0 \end{aligned}$$

Thus, f has a removable discontinuity at $(0, 0)$ of value 0.

- Compute $f_x(0, 0)$ and $f_y(0, 0)$

Solution

$$f_x(0, 0) = \lim_{h \rightarrow 0} \frac{f(h, 0) - f(0, 0)}{h} = \frac{0}{h} = 0$$

$$f_y(0, 0) = \lim_{h \rightarrow 0} \frac{f(0, h) - f(0, 0)}{h} = \frac{0}{h} = 0$$

Also, at any $(x, y) \neq (0, 0)$, we have

$$\begin{aligned} f_x(x, y) &= \frac{x^4y + 4x^2y^3 - y^5}{(x^2 + y^2)^2} \\ f_y(x, y) &= \frac{-xy^4 + x^5 - 4x^3y^2}{(x^2 + y^2)^2} \end{aligned}$$

Then

$$\begin{aligned} &\lim_{(x,y) \rightarrow (0,0)} \left| \frac{yx^4 + 4x^2y^3 - y^5}{(x^2 + y^2)^2} \right| \\ &\leq \lim_{(x,y) \rightarrow (0,0)} \left(|y| \frac{x^4}{(x^2 + y^2)^2} + |y| \frac{y^4}{(x^2 + y^2)^2} + 4|y| \frac{x^2}{x^2 + y^2} \frac{y^2}{x^2 + y^2} \right) \\ &\leq \lim_{(x,y) \rightarrow (0,0)} (2|y| + 4|y|) \leq 5\sqrt{x^2 + y^2} = 0 \end{aligned}$$

Hence f_x has a removable discontinuity at $(0, 0)$. A similar calculation shows f_y has a removable discontinuity at $(0, 0)$.

- Compute $\partial_y(f_x)(0, 0)$ and $\partial_x(f_y)(0, 0)$. Are these equal?

Solution

$$\begin{aligned} f_{xy}(0, 0) &= \lim_{h \rightarrow 0} \frac{f_x(0, h) - f_x(0, 0)}{h} = \frac{-h^5/h^4}{h} = -1 \\ f_{yx}(0, 0) &= \lim_{h \rightarrow 0} \frac{f_y(0, h) - f_y(0, 0)}{h} = \frac{h^5/h^4}{h} = 1 \end{aligned}$$

Clearly, these mixed partials do not match.

- Compute $\partial_y(f_x)(x, y)$ and $\partial_x(f_y)(x, y)$ for $(x, y) \neq (0, 0)$.

Solution At any $(x, y) \neq (0, 0)$, we have

$$\begin{aligned} f_{xy}(x, y) &= \frac{x^6 - y^6 - 9y^4x^2 + 9y^2x^4}{(x^2 + y^2)^3} \\ f_{yx}(x, y) &= \frac{x^6 - y^6 + 9y^4x^2 - 9y^2x^4}{(x^2 + y^2)^3} \end{aligned}$$

- Determine if $\partial_y(f_x)(x, y)$ and $\partial_x(f_y)(x, y)$ are continuous at $(0, 0)$.

Solution If you look at the paths $(t, 2t)$ and $(t, 3t)$ for $t > 0$, you find you get different limiting values as $t \rightarrow 0^+$ for both f_{xy} and f_{yx} . Hence, these mixed partials are not continuous at $(0, 0)$.

We see the lack of continuity of the mixed partials in a ball about the origin is the problem.

This leads to our next theorem.

Theorem 11.4.2 Sufficient Conditions for the Mixed Order Partial Derivatives to Match

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ and fix $i \neq j$ for the set $\{1, \dots, n\}$. Assume all the first order partials $\frac{\partial f}{\partial x_i}$ and $\frac{\partial f}{\partial x_j}$ and the second order partials $\partial_{x_j} \left(\frac{\partial f}{\partial x_i} \right)$ and $\partial_{x_i} \left(\frac{\partial f}{\partial x_j} \right)$ all exist in $B_r(\mathbf{p})$ for some $r > 0$. Also assume $\partial_{x_j} \left(\frac{\partial f}{\partial x_i} \right)$ and $\partial_{x_i} \left(\frac{\partial f}{\partial x_j} \right)$ are continuous in $B_r(\mathbf{p})$. Then $\partial_{x_j} \left(\frac{\partial f}{\partial x_i} (\mathbf{p}) \right) = \partial_{x_i} \left(\frac{\partial f}{\partial x_j} \right)$ at \mathbf{p} .

Proof 11.4.2

Let's control this notational mess by noticing we are only looking at changes in x_i and x_j . Let $X_1 = x_i, X_2 = x_j$ and

$$\begin{aligned} F(X_1, X_2) &= f(x_1, \dots, x_{i-1}, \mathbf{x}_i = X_1, x_{i+1}, \dots, \\ &\quad x_{j-1}, \mathbf{x}_j = X_2, x_{j+1}, \dots, x_n) \end{aligned}$$

Also, let $F_1 = \partial_{x_i} f$, $F_2 = \partial_{x_j} f$ and $F_{12} = \partial_{x_j}(\partial_{x_i} f)$ and $F_{21} = \partial_{x_i}(\partial_{x_j} f)$. Now with these variables defined, the argument to prove the mixed partials match is essentially the same as the one in (Peterson (8) 2019). So even though there are n variables involved, we just have to focus our attention on the two that matter. Fix h small enough for that $\mathbf{p} + h\mathbf{e}_i + h\mathbf{e}_j$ is in $B_r(\mathbf{p})$. Note $\mathbf{p} + h\mathbf{e}_i = X_1 + h$ and $\mathbf{p} + h\mathbf{e}_j = X_2 + h$.

Step One:

Define

$$\begin{aligned} G(h) &= F(X_1 + h, X_2 + h) - F(X_1 + h, X_2) - F(X_1, X_2 + h) + F(X_1, X_2) \\ \phi(x) &= F(x, X_2 + h) - F(x, X_2), \quad \psi(y) = \phi(X_1 + y) - \phi(X_1) \end{aligned}$$

Now apply the Mean Value Theorem to ϕ . Note $\phi'(x) = \partial_{X_1} F(x, X_2 + h) - \partial_{X_1} F(x, X_2) = f_{x_i}(\mathbf{p} + x\mathbf{e}_i + h\mathbf{e}_j) - f_{x_i}(\mathbf{p} + x\mathbf{e}_i)$. Thus,

$$\begin{aligned} \phi(X_1 + h) - \phi(X_1) &= \phi'(X_1 + \theta_1 h) h, \quad \theta_1 \text{ between } X_1 \text{ and } X_1 + h \\ &= (f_{x_i}(\mathbf{p} + \theta_1 h\mathbf{e}_i + h\mathbf{e}_j) - f_{x_i}(\mathbf{p} + \theta_1 h\mathbf{e}_i))h \\ &= (F_{X_1}(X_1 + \theta_1 h, X_2 + h) - F_{X_1}(X_1 + \theta_1 h, X_2))h \end{aligned}$$

where we use the definition of F . Now define for y between X_2 and $X_2 + h$, the function ξ by

$$\begin{aligned} \xi(z) &= F_{X_1}(X_1 + \theta_1 h, z) \\ \xi'(z) &= \partial_{X_2} F_{X_1}(X_1 + \theta_1 h, z) \end{aligned}$$

Apply the Mean Value Theorem to ξ : there is a θ_2 so that $X_2 + \theta_2 h$ is between X_2 and $X_2 + h$ with

$$\xi(X_2 + h) - \xi(X_2) = \partial_{X_2} F_{X_1}(X_1 + \theta_1 h, X_2 + \theta_2 h) h$$

Thus,

$$\begin{aligned} h(\xi(X_2 + h) - \xi(X_2)) &= h(F_{X_1}(X_1 + \theta_1 h, X_2 + h) - F_{X_1}(X_1 + \theta_1 h, X_2)) \\ &= \partial_{X_2} F_{X_1}(X_1 + \theta_1 h, X_2 + \theta_2 h) h^2 \end{aligned}$$

But,

$$\begin{aligned} \phi(X_1 + h) - \phi(X_1) &= (F(X_1 + h, X_2 + h) - F(X_1 + h, X_2)) \\ &\quad - (F(X_1, X_2 + h) - F(X_1, X_2)) \end{aligned}$$

$$\begin{aligned}
&= G(h) \\
&= (F_{X_1}(X_1 + \theta_1 h, X_2 + h) - F_{X_1}(X_1 + \theta_1 h, X_2))h \\
&= (\xi(X_2 + h) - \xi(X_2))h \\
&= \partial_{X_2} F_{X_1}(X_1 + \theta_1 h, X_2 + \theta_2 h) h^2
\end{aligned}$$

So

$$G(h) = \partial_{X_2} F_{X_1}(X_1 + \theta_1 h, X_2 + \theta_2 h) h^2$$

Step Two:

Now start again with $L(w) = F(X_1 + h, w) - F(X_1, w)$, noting $G(h) = L(X_2 + h) - L(X_2)$. Apply the Mean Value Theorem to L like before to find

$$\begin{aligned}
G(h) &= L(X_2 + h) - L(X_2) = L'(X_2 + \theta_3 h)h \\
&= (\partial_{X_2} F(X_1 + h, X_2 + \theta_3 h) - \partial_{X_2} F(X_1, X_2 + \theta_3 h)) h
\end{aligned}$$

for an intermediate value of θ_3 with $X_2 + \theta_3 h$ between X_2 and $X_2 + h$.

Now define for z between X_2 and $X_2 + h$, the function γ by

$$\begin{aligned}
\gamma(z) &= F_{X_2}(z, X_2 + \theta_3 h) \\
\gamma'(z) &= \partial_{X_1} F_{X_2}(z, X_2 + \theta_3 h)
\end{aligned}$$

Now apply the Mean Value Theorem a fourth time: So there is a θ_4 with $X_1 + \theta_4 h$ between X_1 and $X_1 + h$ so that

$$\begin{aligned}
\gamma(X_1 + h) - \gamma(X_1) &= \gamma'(X_1 + \theta_4 h) h \\
&= \partial_{X_1} F_{X_2}(X_1 + \theta_4 h, X_2 + \theta_3 h) h
\end{aligned}$$

Using what we have found so far, we have

$$\begin{aligned}
G(h) &= (\partial_{X_2} F(X_1 + h, X_2 + \theta_3 h) - \partial_{X_2} F(X_1, X_2 + \theta_3 h)) h \\
&= (\gamma(X_1 + h) - \gamma(X_1)) h = (\partial_{X_1} F_{X_2}(X_1 + \theta_4 h, X_2 + \theta_3 h)) h^2
\end{aligned}$$

So we have two representations for $G(h)$:

$$G(h) = \partial_{X_2} F_{X_1}(X_1 + \theta_1 h, X_2 + \theta_2 h) h^2 = (\partial_{X_1} F_{X_2}(X_1 + \theta_4 h, X_2 + \theta_3 h)) h^2$$

Cancelling the common h^2 , we have

$$\partial_{X_2} F_{X_1}(X_1 + \theta_1 h, X_2 + \theta_2 h) = (\partial_{X_1} F_{X_2}(X_1 + \theta_4 h, X_2 + \theta_3 h))$$

Now let $h \rightarrow 0$ and since these second order partials are continuous at \mathbf{p} , we have $\partial_{X_2} F_{X_1}(X_1, X_2) = \partial_{X_1} F_{X_2}(X_1, X_2)$. This shows the result we wanted. ■

11.4.3.1 Homework

Let

$$f(x, y) = \begin{cases} \frac{4xy(x^2 - y^2)}{x^2 + 2y^2}, & (x, y) \neq (0, 0) \\ 0, & (x, y) = (0, 0) \end{cases}$$

Exercise 11.4.6 Prove $\lim_{(x,y) \rightarrow (0,0)} f(x, y) = 0$ thereby showing f has a removable discontinuity at $(0, 0)$ and so the definition for f above makes sense.

Exercise 11.4.7 Compute $f_x(0, 0)$ and $f_y(0, 0)$

Exercise 11.4.8 Compute $f_x(x, y)$ and $f_y(x, y)$ for $(x, y) \neq (0, 0)$. Simplify your answers to the best possible form.

Exercise 11.4.9 Compute $\partial_y(f_x)(0, 0)$ and $\partial_x(f_y)(0, 0)$. Are these equal?

Exercise 11.4.10 Compute $\partial_y(f_x)(x, y)$ and $\partial_x(f_y)(x, y)$ for $(x, y) \neq (0, 0)$.

Exercise 11.4.11 Determine if $\partial_y(f_x)(x, y)$ and $\partial_x(f_y)(x, y)$ are continuous at $(0, 0)$.

11.5 Derivatives for Vector Functions of n Variables

If $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$, then f determines m component functions $f_i : d \subset \mathbb{R}^n \rightarrow \mathbb{R}$; i.e. at any point \mathbf{x}_0) where f is locally defined, we have

$$f(\mathbf{x}_0) = \begin{bmatrix} f_1(\mathbf{x}_0) \\ \vdots \\ f_m(\mathbf{x}_0) \end{bmatrix}, \quad f(\mathbf{x}_0 + \Delta) = \begin{bmatrix} f_1(\mathbf{x}_0 + \Delta_1 \mathbf{E}_1) \\ \vdots \\ f_m(\mathbf{x}_0 + \Delta_n \mathbf{E}_m) \end{bmatrix}$$

for suitable Δ . If we assume each f_i is differentiable at \mathbf{x}_0 , we know

$$\begin{aligned} f_1(\mathbf{x}_0 + \Delta) &= f_1(\mathbf{x}_0) + \sum_{i=1}^n f_{1,x_i}(\mathbf{x}_0) (x_{0i} - \Delta_i) + \mathcal{E}_1(\mathbf{x}_0, \Delta) \\ &\vdots = \vdots \\ f_m(\mathbf{x}_0 + \Delta) &= f_m(\mathbf{x}_0) + \sum_{i=1}^n f_{m,x_i}(\mathbf{x}_0) (x_{0i} - \Delta_i) + \mathcal{E}_m(\mathbf{x}_0, \Delta) \end{aligned}$$

where

$$\lim_{\Delta \rightarrow 0} \mathcal{E}_i(\mathbf{x}_0, \Delta) = 0, \quad \lim_{\Delta \rightarrow 0} \frac{\mathcal{E}_i(\mathbf{x}_0, \Delta)}{\|\Delta\|} = 0$$

We can then write

$$f(\mathbf{x}_0 + \Delta) = \begin{bmatrix} f_1(\mathbf{x}_0) \\ \vdots \\ f_m(\mathbf{x}_0) \end{bmatrix} + \begin{bmatrix} f_{1,x_1}(\mathbf{x}_0) & \dots & f_{1,x_n}(\mathbf{x}_0) \\ \vdots & \ddots & \vdots \\ f_{m,x_1}(\mathbf{x}_0) & \dots & f_{m,x_n}(\mathbf{x}_0) \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \vdots \\ \Delta_n \end{bmatrix} + \begin{bmatrix} \mathcal{E}_1(\mathbf{x}_0, \Delta) \\ \vdots \\ \mathcal{E}_m(\mathbf{x}_0, \Delta) \end{bmatrix}$$

The matrix of partials above occurs frequently and is called the **Jacobian** of f .

Definition 11.5.1 The Jacobian

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ have partial derivatives at \mathbf{x}_0 . The Jacobian of f is the $m \times n$ matrix

$$\mathbf{J}_f(\mathbf{x}_0) = \begin{bmatrix} f_{1,x_1}(\mathbf{x}_0) & \dots & f_{1,x_n}(\mathbf{x}_0) \\ \vdots & \ddots & \vdots \\ f_{m,x_1}(\mathbf{x}_0) & \dots & f_{m,x_n}(\mathbf{x}_0) \end{bmatrix}$$

If f is differentiable at \mathbf{x}_0 , we can write

$$\mathbf{J}_f(\mathbf{x}_0) = \begin{bmatrix} (\nabla(f_1))^T(\mathbf{x}_0) \\ \vdots \\ (\nabla(f_m))^T(\mathbf{x}_0) \end{bmatrix} = \begin{bmatrix} \mathbf{D}\mathbf{f}_1(\mathbf{x}_0) \\ \vdots \\ \mathbf{D}\mathbf{f}_m(\mathbf{x}_0) \end{bmatrix}$$

Thus, we write

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \mathbf{J}_f(\mathbf{x}_0) \Delta + \mathcal{E}(\mathbf{x}_0, \Delta)$$

where $\mathcal{E}(\mathbf{x}_0, \Delta)$ is the vector error

$$\mathcal{E}(\mathbf{x}_0, \Delta) = \begin{bmatrix} \mathcal{E}_1(\mathbf{x}_0, \Delta) \\ \vdots \\ \mathcal{E}_m(\mathbf{x}_0, \Delta) \end{bmatrix}$$

and we see

$$\begin{aligned} \lim_{\Delta \rightarrow 0} \|\mathcal{E}(\mathbf{x}_0, \Delta)\| &= \lim_{\Delta \rightarrow 0} \sqrt{((\mathcal{E}_1(\mathbf{x}_0, \Delta))^2 + \dots + (\mathcal{E}_m(\mathbf{x}_0, \Delta))^2)} = 0 \\ \lim_{\Delta \rightarrow 0} \left\| \frac{\mathcal{E}(\mathbf{x}_0, \Delta)}{\|\Delta\|} \right\| &= \lim_{\Delta \rightarrow 0} \sqrt{\left(\left(\frac{\mathcal{E}_1(\mathbf{x}_0, \Delta)}{\|\Delta\|} \right)^2 + \dots + \left(\frac{\mathcal{E}_m(\mathbf{x}_0, \Delta)}{\|\Delta\|} \right)^2 \right)} = 0 \end{aligned}$$

The definition of differentiability for a vector function is thus

Definition 11.5.2 The Differentiability of a Vector Function of n variables

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ be defined locally at \mathbf{x}_0 . We say f is differentiable at \mathbf{x}_0 if there is a linear map $\mathbf{L}(\mathbf{x}_0) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and a error vector $\mathcal{E}(\mathbf{x}_0, \Delta)$ so that for given Δ ,

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \mathbf{L}(\mathbf{x}_0) \Delta + \mathcal{E}(\mathbf{x}_0, \Delta)$$

where

$$\lim_{\Delta \rightarrow 0} \mathcal{E}(\mathbf{x}_0, \Delta) = 0$$

$$\lim_{\Delta \rightarrow 0} \frac{\mathcal{E}(\mathbf{x}_0, \Delta)}{\|\Delta\|} = 0$$

where we use norm convergence with respect to the norm $\|\cdot\|$ used in \mathbb{R}^n and \mathbb{R}^m . For any given pair of orthonormal basis \mathbf{E} of \mathbb{R}^n and \mathbf{F} of \mathbb{R}^m , the linear map $\mathbf{L}(\mathbf{x}_0)$ has a representation $[L(\mathbf{x}_0)]_{\mathbf{EF}}$ so that

$$[f(\mathbf{x}_0 + \Delta)]_{\mathbf{F}} = [f(\mathbf{x}_0)]_{\mathbf{F}} + [L(\mathbf{x}_0)]_{\mathbf{EF}} [\Delta]_{\mathbf{E}} + [\mathcal{E}(\mathbf{x}_0, \Delta)]_{\mathbf{F}}$$

We then say the derivative of f at \mathbf{x}_0 is $\mathbf{L}(\mathbf{x}_0)$. The derivative of f at \mathbf{x}_0 is denoted by $Df(\mathbf{x}_0)$.

Comment 11.5.1 We usually are not so careful with the notation. It is understood that \mathbb{R}^n and \mathbb{R}^m have some choice of orthonormal basis \mathbf{E} and \mathbf{F} and so we identify

$$\begin{aligned} \mathbf{L}(\mathbf{x}_0) &\equiv [L(\mathbf{x}_0)]_{\mathbf{E}, \mathbf{F}}, & f(\mathbf{x}_0 + \Delta) &\equiv [f(\mathbf{x}_0 + \Delta)]_{\mathbf{F}}, & \Delta &\equiv [\Delta]_{\mathbf{E}} \\ \mathcal{E}(\mathbf{x}_0, \Delta) &\equiv [\mathcal{E}(\mathbf{x}_0, \Delta)]_{\mathbf{F}} \end{aligned}$$

It is not so hard to mentally make the conversion back and forth and you should get used to it.

It is easy to prove the following result. Note we are using the identification of the linear mapping to its matrix representation here.

Theorem 11.5.1 If the Vector Function f is Differentiable at \mathbf{x}_0 , $\mathbf{L}(\mathbf{x}_0) = J_f(\mathbf{x}_0)$

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ be defined locally at \mathbf{x}_0 . If f is differentiable at \mathbf{x}_0 , the $(\mathbf{L}(\mathbf{x}_0))_{ij} = f_{i,x_j}(\mathbf{x}_0)$; i.e. $\mathbf{L}(\mathbf{x}_0) = J_f(\mathbf{x}_0)$.

Proof 11.5.1

This argument is very similar to the one we did before. Just apply the argument to each component. We leave this to you. ■

11.5.0.2 Homework

Exercise 11.5.1

Exercise 11.5.2

Exercise 11.5.3

Exercise 11.5.4

Exercise 11.5.5

11.5.1 The Chain Rule For Vector Valued Functions

The next thing we want to do is to extend the chain rule to this setting. Note the proof here is different than the one we did in the scalar case. We are using matrix manipulations here instead of components.

Theorem 11.5.2 The Chain Rule for Vector Functions

Let $F_1 : D_1 \subset \mathbb{R}^n \rightarrow \mathcal{R}(F_1) \subset \mathbb{R}^m$ and $F_2 : D_2 \subset \mathbb{R}^m \rightarrow \mathcal{R}(F_2) \subset \mathbb{R}^r$. Here \mathcal{F} denotes the range of F_i . We assume there is overlap: $\mathcal{F} \cap D_2 = S \neq \emptyset$ and let $E = F^{-1}(S)$. Further, assume F_1 is differentiable at Q_0 in E and F_2 is differentiable at $P_0 = F_1(Q_0) \in S$. Then, $F_3 = F_2 \circ F_1$ is differentiable at Q_0 with matrix representation given by

$$[DF_3(Q_0)]_{r \times n} = [DF_2(P_0)]_{r \times m} [DF_1(Q_0)]_{m \times n}$$

or in linear mapping terms

$$DF_3(Q_0) = DF_2(P_0) \circ DF_1(Q_0)$$

Proof 11.5.2

You can see the rough flow of these mappings in Figure 11.1 which helps set the stage. We know

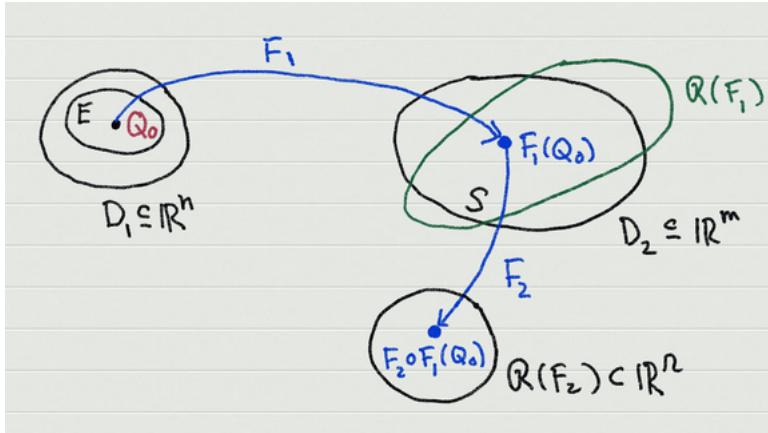


Figure 11.1: The Vector Chain Rule

$$F_2(\mathbf{P}) = F_2(\mathbf{P}_0) + \mathbf{J}_2(\mathbf{P}_0)(\mathbf{P} - \mathbf{P}_0) + \mathcal{E}_2(\mathbf{P}_0, \mathbf{P})$$

where $\mathbf{J}_2(\mathbf{P}_0)$ is a linear map from \mathbb{R}^m to \mathbb{R}^r and $\mathcal{E}_2(\mathbf{P}_0, \mathbf{P} \rightarrow \mathbf{0})$ as $\mathbf{P} \rightarrow \mathbf{P}_0$. Also,

$$F_1(\mathbf{Q}) = F_1(\mathbf{Q}_0) + \mathbf{J}_1(\mathbf{Q}_0)(\mathbf{Q} - \mathbf{Q}_0) + \mathcal{E}_1(\mathbf{Q}_0, \mathbf{Q})$$

where $\mathbf{J}_1(\mathbf{Q}_0)$ is a linear map from \mathbb{R}^n to \mathbb{R}^m and $\mathcal{E}_1(\mathbf{Q}_0, \mathbf{Q} \rightarrow \mathbf{0})$ as $\mathbf{Q} \rightarrow \mathbf{Q}_0$. Let $\mathbf{P} = F_1(\mathbf{Q})$ so that $\mathbf{P}_0 = F_1(\mathbf{Q}_0)$. Then

$$\begin{aligned} F_3(\mathbf{Q}) - F_3(\mathbf{Q}_0) &= F_2(F_1(\mathbf{Q})) - F_2(F_1(\mathbf{Q}_0)) \\ &= J_2(F_1(\mathbf{Q}_0))(F_1(\mathbf{Q}) - F_1(\mathbf{Q}_0)) + \mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q})) \end{aligned}$$

$$= J_2(F_1(\mathbf{Q}_0)) \left\{ J_1(\mathbf{Q}_0)(\mathbf{Q} - \mathbf{Q}_0) + \mathcal{E}_1(\mathbf{Q}_0, \mathbf{Q}) \right\} \\ + \mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))$$

Now reorganize:

$$F_3(\mathbf{Q}) - F_3(\mathbf{Q}_0) = J_2(F_1(\mathbf{Q}_0)) J_1(\mathbf{Q}_0)(\mathbf{Q} - \mathbf{Q}_0) + J_2(F_1(\mathbf{Q}_0)) \mathcal{E}_1(\mathbf{Q}_0, \mathbf{Q}) \\ + \mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))$$

or

$$F_3(\mathbf{Q}) - F_3(\mathbf{Q}_0) = J_2(F_1(\mathbf{Q}_0)) J_1(\mathbf{Q}_0)(\mathbf{Q} - \mathbf{Q}_0) + \mathcal{E}_3(\mathbf{Q}_0, \mathbf{Q})$$

where

$$\mathcal{E}_3(\mathbf{Q}_0, \mathbf{Q}) = J_2(F_1(\mathbf{Q}_0)) \mathcal{E}_1(\mathbf{Q}_0, \mathbf{Q}) + \mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))$$

Since $\mathcal{E}_1(\mathbf{Q}_0, \mathbf{Q}) \rightarrow 0$ as $\mathbf{Q} \rightarrow \mathbf{Q}_0$, the first term in \mathcal{E}_3 goes to zero. For the second term, we know since F_1 is differentiable at \mathbf{Q}_0 , F_1 is continuous at \mathbf{Q}_0 . Thus,

$$\lim_{\mathbf{Q} \rightarrow \mathbf{Q}_0} \mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q})) = \lim_{F_1(\mathbf{Q}) \rightarrow F_1(\mathbf{Q}_0)} \mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q})) = \mathbf{0}$$

Next, we need to look at

$$\frac{\mathcal{E}_3(\mathbf{Q}_0, \mathbf{Q})}{\|\mathbf{Q} - \mathbf{Q}_0\|} = \frac{J_2(F_1(\mathbf{Q}_0)) \mathcal{E}_1(\mathbf{Q}_0, \mathbf{Q})}{\|\mathbf{Q} - \mathbf{Q}_0\|} + \frac{\mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))}{\|\mathbf{Q} - \mathbf{Q}_0\|}$$

Again, the first term goes to zero as $\mathbf{Q} \rightarrow \mathbf{Q}_0$ since that is a property of \mathcal{E}_1 . The second term is more interesting. Now if F_1 was locally constant at \mathbf{Q}_0 , then $\mathbf{J}_1((\mathbf{Q}_0)) = \mathbf{0}$ and we know the error satisfies $\mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))/\|\mathbf{Q} - \mathbf{Q}_0\|$ is identically zero and we have shown what we want. If F_1 is not locally constant, we can say

$$\frac{\mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))}{\|\mathbf{Q} - \mathbf{Q}_0\|} = \begin{cases} 0, & \mathbf{Q} \neq \mathbf{Q}_0, F_1(\mathbf{Q}) = F_1(\mathbf{Q}_0) \\ \frac{\mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))}{\|F_1(\mathbf{Q}) - F_1(\mathbf{Q}_0)\|} \frac{\|F_1(\mathbf{Q}) - F_1(\mathbf{Q}_0)\|}{\|\mathbf{Q} - \mathbf{Q}_0\|}, & \mathbf{Q} \neq \mathbf{Q}_0, F_1(\mathbf{Q}) \neq F_1(\mathbf{Q}_0) \end{cases}$$

Hence, for all $\mathbf{Q} \neq \mathbf{Q}_0$ with for all $F_1(\mathbf{Q}) \neq F_1(\mathbf{Q}_0)$, we have

$$\frac{\mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))}{\|\mathbf{Q} - \mathbf{Q}_0\|} \leq \frac{\mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))}{\|F_1(\mathbf{Q}) - F_1(\mathbf{Q}_0)\|} \frac{\|F_1(\mathbf{Q}) - F_1(\mathbf{Q}_0)\|}{\|\mathbf{Q} - \mathbf{Q}_0\|} \\ = \frac{\mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))}{\|F_1(\mathbf{Q}) - F_1(\mathbf{Q}_0)\|} \frac{\|\mathbf{J}_1(\mathbf{Q}_0)\|_{Fr} \|\mathbf{Q} - \mathbf{Q}_0\| + \|\mathcal{E}_1(\mathbf{Q}_0, \mathbf{Q})\|}{\|\mathbf{Q} - \mathbf{Q}_0\|}$$

where we use the Frobenius norm inequalities from Chapter 4 in Theorem 4.5.1. So, simplifying a bit,

$$\frac{\mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))}{\|\mathbf{Q} - \mathbf{Q}_0\|} \leq \frac{\mathcal{E}_2(F_1(\mathbf{Q}_0), F_1(\mathbf{Q}))}{\|F_1(\mathbf{Q}) - F_1(\mathbf{Q}_0)\|} \left(\|\mathbf{J}_1(\mathbf{Q}_0)\|_{Fr} + \frac{\|\mathcal{E}_1(\mathbf{Q}_0, \mathbf{Q})\|}{\|\mathbf{Q} - \mathbf{Q}_0\|} \right)$$

As $\mathbf{Q} \rightarrow \mathbf{Q}_0$, $F_1(\mathbf{Q}) \rightarrow F_1(\mathbf{Q}_0)$ also by continuity. The case where all $F_1(\mathbf{Q}) = F_1(\mathbf{Q}_0)$ has an error bounded above by the case with $F_1(\mathbf{Q}) \neq F_1(\mathbf{Q}_0)$. Hence, from the estimates above, term one goes to zero and the second piece goes to $\mathbf{J}_1(\mathbf{Q}_0)$ as term two goes to zero. Hence, the limit is zero and we have shown the second part of the condition for \mathcal{E}_3 . We conclude F_3 is differentiable at \mathbf{Q}_0

and

$$DF_3(Q_0) = DF_2(P_0) DF_1(Q_0)$$

■

11.5.1.1 Homework

Exercise 11.5.6

Exercise 11.5.7

Exercise 11.5.8

Exercise 11.5.9

Exercise 11.5.10

11.6 Tangent Plane Error

Now that we have the chain rule, we can quickly develop other results such as how much error we make when we approximate our surface $f(x, y)$ using a tangent plane at a point (x_0, y_0) . We approximate nonlinear mappings for many purposes. We go through some examples in Chapter ?? and they are well worth studying. The first thing we need is to know when a scalar function of n variables is differentiable. Just because it's partials exist at a point is not enough to guarantee that! But we know if the partials are continuous around that point, then the derivative does exist. And that means we can write the function in terms of its tangent plane plus an error. The arguments to do this are not terribly hard, but they are a bit involved. For now, let's go back to the old idea of a tangent plane to a surface. For the surface $z = f(x, y)$ if its partials are continuous functions (they usually are for our work!) then f is differentiable and hence we know that

$$f(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) + E(x, y, x_0, y_0)$$

where $\mathcal{E}(x, y, x_0, y_0)/\sqrt{(x - x_0)^2 + (y - y_0)^2} \rightarrow 0$ and $E(x, y, x_0, y_0)$ go to 0 as $(x, y) \rightarrow (x_0, y_0)$. We can characterize the error must better if we have access to the second order partial derivatives of f . We have shown that if the second order partials are continuous locally near (x_0, y_0) then the mixed order partials f_{xy} and f_{yx} must match at the point (x_0, y_0) . Hence, for these **smooth** surfaces, the Hessian for a scalar function is a symmetric matrix!

11.6.1 The Mean Value Theorem

We can approximate a scalar function of many variables using their gradients. This is the first step in developing Taylor polynomial approximations of arbitrary order, but first step uses gradient information only. The approximations using second order partial information use both the gradient and the Hessian and will be covered next. The following theorem is called the Mean Value Theorem.

Theorem 11.6.1 The Mean Value Theorem

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable locally at \mathbf{x}_0 ; i.e. f is differentiable in $B(r, \mathbf{x}_0)$ for some $r > 0$. Then if \mathbf{x} and \mathbf{y} are in $B(r, \mathbf{x}_0)$, there is a point \mathbf{u} on the line connecting \mathbf{x} and \mathbf{y} so that

$$f(\mathbf{x}) - f(\mathbf{y}) = (\nabla(f))^T(\mathbf{u})(\mathbf{x} - \mathbf{y}) = \mathbf{D}\mathbf{f}(\mathbf{u})(\mathbf{x} - \mathbf{y})$$

The line connecting \mathbf{x} and \mathbf{y} is often denoted by $[\mathbf{x}, \mathbf{y}]$ even though \mathbf{x} and \mathbf{y} are vectors.

Proof 11.6.1

Let $\mathbf{v} = \mathbf{x} - \mathbf{y}$ and define $P : \mathbb{R} \rightarrow \mathbb{R}^n$ by $P(t) = \mathbf{y} + t(\mathbf{x} - \mathbf{y})$. Then

$$P'(t) = \lim_{h \rightarrow 0} \frac{P(t+h) - P(t)}{h} = \mathbf{v}$$

Let $F : \mathbb{R} \rightarrow \mathbb{R}$ be defined by $F(t) = f(P(t)) = f(\mathbf{y} + t\mathbf{v})$. By the Mean Value Theorem for functions of one variable, there is a c between 0 and 1 so that

$$F(1) - F(0) = F'(c)(1 - 0)$$

Thus, by the chain rule

$$F'(c) = \mathbf{D}\mathbf{f}(P(c)) P'(c) = (\nabla(f))^T(P(c)) \mathbf{v} = \sum_{i=1}^n f_{x_i}(P(c)) v_i$$

Now $F(1) = f(P(1)) = f(\mathbf{x})$ and $F(0) = f(P(0)) = f(\mathbf{y})$. Also, let $P(c) = \mathbf{y} + c(\mathbf{x} - \mathbf{y}) = \mathbf{u}$ which is a point on $[\mathbf{x}, \mathbf{y}]$. So we have shown

$$f(\mathbf{x}) - f(\mathbf{y}) = \mathbf{D}\mathbf{f}(\mathbf{u})(\mathbf{x} - \mathbf{y})$$

■

11.6.1.1 Homework

Exercise 11.6.1

Exercise 11.6.2

Exercise 11.6.3

Exercise 11.6.4

Exercise 11.6.5

11.6.2 Hessian Approximations

We can now explain the most common approximation result for tangent planes.

Two Variables:

Let $h(t) = f(x_0 + t\Delta x, y_0 + t\Delta y)$. Then we know we can write $h(t) = h(0) + h'(0)t + h''(c)\frac{t^2}{2}$. Using the chain rule, we find

$$h'(t) = f_x(x_0 + t\Delta x, y_0 + t\Delta y)\Delta x + f_y(x_0 + t\Delta x, y_0 + t\Delta y)\Delta y$$

and

$$\begin{aligned}
 h''(t) &= h''(t) \\
 &= \partial_x \left(f_x(x_0 + t\Delta x, y_0 + t\Delta y)\Delta x + f_y(x_0 + t\Delta x, y_0 + t\Delta y)\Delta y \right) \Delta x \\
 &\quad + \partial_y \left(f_x(x_0 + t\Delta x, y_0 + t\Delta y)\Delta x + f_y(x_0 + t\Delta x, y_0 + t\Delta y)\Delta y \right) \Delta y \\
 &= f_{xx}(x_0 + t\Delta x, y_0 + t\Delta y)(\Delta x)^2 + f_{yx}(x_0 + t\Delta x, y_0 + t\Delta y)(\Delta y)(\Delta x) \\
 &\quad + f_{xy}(x_0 + t\Delta x, y_0 + t\Delta y)(\Delta x)(\Delta y) + f_{yy}(x_0 + t\Delta x, y_0 + t\Delta y)(\Delta y)^2
 \end{aligned}$$

We can rewrite this in matrix - vector form as

$$h''(t) = [\Delta x \quad \Delta y] \begin{bmatrix} f_{xx}(x_0 + t\Delta x, y_0 + t\Delta y) & f_{yx}(x_0 + t\Delta x, y_0 + t\Delta y) \\ f_{xy}(x_0 + t\Delta x, y_0 + t\Delta y) & f_{yy}(x_0 + t\Delta x, y_0 + t\Delta y) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}$$

Of course, using the definition of \mathbf{H} , this can be rewritten as

$$h''(t) = \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}^T \mathbf{H}(x_0 + t\Delta x, y_0 + t\Delta y) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}$$

Thus, our tangent plane approximation can be written as

$$h(1) = h(0) + h'(0)(1 - 0) + h''(c) \frac{1}{2}$$

for some c between 0 and 1. Substituting for the h terms, we find

$$\begin{aligned}
 f(x_0 + \Delta x, y_0 + \Delta y) &= f(x_0, y_0) + f_x(x_0, y_0)\Delta x + f_y(x_0, y_0)\Delta y \\
 &\quad + \frac{1}{2} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}^T \mathbf{H}(x_0 + c\Delta x, y_0 + c\Delta y) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}
 \end{aligned}$$

Clearly, we have shown how to express the error in terms of second order partials. There is a point c between 0 and 1 so that

$$\mathcal{E}(x_0, y_0, \Delta x, \Delta y) = \frac{1}{2} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}^T \mathbf{H}(x_0 + c\Delta x, y_0 + c\Delta y) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}$$

Note the error is a quadratic expression in terms of the Δx and Δy . We also now know for a function of two variables, $f(x, y)$, we can estimate the error made in approximating using the gradient at the given point (x_0, y_0) as follows: We have

$$\begin{aligned}
 f(x, y) &= f(x_0, y_0) + \langle \nabla(f)(x_0, y_0), [x - x_0, y - y_0]^T \rangle \\
 &\quad + (1/2)[x - x_0, y - y_0]^T \mathbf{H}(x_0 + c(x - x_0), y_0 + c(y - y_0)) [x - x_0, y - y_0]^T
 \end{aligned}$$

Three Variables:

Let's shift to using (x_1, x_2, x_3) instead of (x, y, z) . Define let $h(t) = f(x_{01} + t\Delta_1, y_{01} + t\Delta_2, z_{01} + t\Delta_3)$. For convenience, let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} x_{01} \\ x_{02} \\ x_{03} \end{bmatrix}, \quad \mathbf{\Delta} = \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{bmatrix}$$

Then. $h(t) = f(\mathbf{x}_0 + t\Delta)$ and we know $h(t) = h(0) + h'(0)t + h''(c)\frac{t^2}{2}$. Using the chain rule, we find

$$h'(t) = \sum_{i=1}^3 f_{x_i}(\mathbf{x}_0 + t\Delta)\Delta_i$$

and

$$h''(t) = \sum_{i=1}^3 \left(\sum_{j=1}^3 f_{x_i, x_j}(\mathbf{x}_0 + t\Delta)\Delta_i \Delta_j \right)$$

We can rewrite this in matrix - vector form as

$$h''(t) = \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{bmatrix}^T \begin{bmatrix} f_{x_1, x_1}(\mathbf{x}_0 + t\Delta) & f_{x_1, x_2}(\mathbf{x}_0 + t\Delta) & f_{x_1, x_3}(\mathbf{x}_0 + t\Delta) \\ f_{x_2, x_1}(\mathbf{x}_0 + t\Delta) & f_{x_2, x_2}(\mathbf{x}_0 + t\Delta) & f_{x_2, x_3}(\mathbf{x}_0 + t\Delta) \\ f_{x_3, x_1}(\mathbf{x}_0 + t\Delta) & f_{x_3, x_2}(\mathbf{x}_0 + t\Delta) & f_{x_3, x_3}(\mathbf{x}_0 + t\Delta) \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{bmatrix}$$

Of course, using the definition of \mathbf{H} , this can be rewritten as

$$\begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{bmatrix}^T \mathbf{H}(\mathbf{x}_0 + t\Delta) \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{bmatrix} = \Delta^T \mathbf{H}(\mathbf{x}_0 + t\Delta) \Delta$$

Thus, our tangent plane approximation can be written as

$$h(1) = h(0) + h'(0)(1 - 0) + h''(c)\frac{1}{2}$$

for some c between 0 and 1. Substituting for the h terms, we find

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \sum_{i=1}^3 f_{x_i}(\mathbf{x}_0)\Delta_i + \frac{1}{2}\Delta^T \mathbf{H}(\mathbf{x}_0 + c\Delta) \Delta$$

Clearly, we have shown how to express the error in terms of second order partials. There is a point c between 0 and 1 so that

$$\mathcal{E}(\mathbf{x}_0, \Delta) = \Delta^T \mathbf{H}(\mathbf{x}_0 + c\Delta) \Delta$$

You can see how the analysis does not change much if we did the case of n variables. We'll leave that to you!

Vector Valued Functions:

We are now really doing a linear approximation to a vector valued function and this is not a tangent plane, but the basic idea is the same. We now have a vector function $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$, f would have m component functions. Now Δ and \mathbf{x}_0 have n components. A similar analysis uses

$$h(t) = h(t) = f(\mathbf{x}_0 + t\Delta) = \begin{bmatrix} f_1(\mathbf{x}_0 + t\Delta) \\ \vdots \\ f_m(\mathbf{x}_0 + t\Delta) \end{bmatrix}$$

Using the chain rule, we find

$$h'(t) = \begin{bmatrix} \sum_{i=1}^n f_{1,x_i}(\mathbf{x}_0 + t\Delta)\Delta_i \\ \vdots \\ \sum_{i=1}^n f_{m,x_i}(\mathbf{x}_0 + t\Delta)\Delta_i \end{bmatrix} = \mathbf{J}_f(\mathbf{x}_0 + t\Delta) \Delta$$

and

$$h''(t) = \begin{bmatrix} \sum_{i=1}^n \sum_{j=1}^n f_{1,x_i,x_j}(\mathbf{x}_0 + t\Delta)\Delta_i \Delta_j \\ \vdots \\ \sum_{i=1}^n \sum_{j=1}^n f_{1,x_i,x_j}(\mathbf{x}_0 + t\Delta)\Delta_i \Delta_j \end{bmatrix}$$

Each component in the vector above can be written in a matrix - vector form as

$$h''(t) = \begin{bmatrix} \Delta^T \mathbf{H}_1(\mathbf{x}_0 + t\Delta) \Delta \\ \vdots \\ \Delta^T \mathbf{H}_n(\mathbf{x}_0 + t\Delta) \Delta \end{bmatrix}$$

where \mathbf{H}_i is the Hessian for f_i . Thus, our linear approximation is

$$h(1) = h(0) + h'(0)(1 - 0) + h''(c)\frac{1}{2}$$

for some c between 0 and 1. Substituting for the h terms, we find

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \mathbf{J}_f(\mathbf{x}_0) \Delta + \frac{1}{2} \begin{bmatrix} \Delta^T \mathbf{H}_1(\mathbf{x}_0 + c\Delta) \Delta \\ \vdots \\ \Delta^T \mathbf{H}_n(\mathbf{x}_0 + c\Delta) \Delta \end{bmatrix}$$

Each of these entries corresponds to the error \mathcal{E}_i for component f_i and the vector error is

$$\mathcal{E} = \begin{bmatrix} \mathcal{E}_1 \\ \vdots \\ \mathcal{E}_n \end{bmatrix} \implies \|\mathcal{E}\| = \sqrt{(\mathcal{E}_1)^2 + \dots + (\mathcal{E}_n)^2}$$

Example 11.6.1 Let

$$f(x, y, z) = \begin{bmatrix} x^2y^4 + 2x + 3yz^2 + 10 \\ 4x^2 + 5z^3y^2 \end{bmatrix}$$

Estimate the linear approximation error about $(0, 0, 0)$.

Solution We have

$$\begin{aligned} f_{1,x} &= 2xy^4 + 2, & f_{1,y} &= 4x^2y^3 + 3z^2, & f_{1,z} &= 6yz \\ f_{1,xx} &= 2y^4, & f_{1,xy} &= 8xy^3, & f_{1,xz} &= 0 \\ f_{1,yx} &= 8xy^3, & f_{1,yy} &= 12x^2y^2, & f_{1,yz} &= 6z, \\ f_{1,zx} &= 0, & f_{1,zy} &= 6z, & f_{1,zz} &= 6y, \\ f_{2,x} &= 8x, & f_{2,y} &= 10z^3, & f_{2,z} &= 15y^2z^2 \\ f_{2,xx} &= 8, & f_{2,xy} &= 0, & f_{2,xz} &= 0 \\ f_{2,yx} &= 0, & f_{2,yy} &= 0, & f_{2,yz} &= 30z^2, \end{aligned}$$

$$f_{2,zx} = 0, \quad f_{2,zy} = 30yz^2, \quad f_{2,zz} = 30y^2z,$$

Hence,

$$\begin{aligned} \mathbf{J}_f(x, y, z) &= \begin{bmatrix} 2xy^4 + 2 & 4x^2y^3 + 3z^2 & 6yz \\ 8x & 10z^3 & 15y^2z^2 \end{bmatrix} \\ \mathbf{H}_1(x, y, z) &= \begin{bmatrix} 2y^4 & 8xy^3 & 0 \\ 8xy^3 & 12x^2y^2 & 6z \\ 0 & 6z & 6y \end{bmatrix}, \quad \mathbf{H}_2(x, y, z) = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 0 & 30z^2 \\ 0 & 20yz^2 & 30z^2 \end{bmatrix} \end{aligned}$$

This gives the errors

$$\begin{aligned} \mathcal{E}_1(0, 0, \Delta) &= \frac{1}{2} \Delta^T \mathbf{H}_1(c\Delta) \Delta \\ \mathcal{E}_2(0, 0, \Delta) &= \frac{1}{2} \Delta^T \mathbf{H}_2(c\Delta) \Delta \end{aligned}$$

for a c between 0 and 1. To find how much error is made for at $\mathbf{x}_0 + \Delta$, we would approximate each error separately, for example, as we detail in (Peterson (8) 2019) for functions of two variables. All of these calculations are terribly messy! So at $(0, 0, 0)$, letting (x^*, y^*, z^*) denote the inbetween point $\mathbf{x}_0 + c\Delta$, we have

$$\mathcal{E}_1(\mathbf{x}_0, \Delta) = [\Delta_1 \quad \Delta_2 \quad \Delta_3] \begin{bmatrix} 2(y^*)^4 & 8x^*(y^*)^3 & 0 \\ 8x^*(y^*)^3 & 12(x^*)^2(y^*)^2 & 6z^* \\ 0 & 6z^* & 6y^* \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{bmatrix}$$

If we restrict our attention to points in $B(r, (0, 0, 0))$, we see all the terms in Δ satisfy $\Delta_i | < r$ and $|x^*|, |y^*|$ and $|z^*|$ are all less than r . Thus

$$\begin{aligned} \mathcal{E}_1(\mathbf{x}_0, \Delta) &= [\Delta_1 \quad \Delta_2 \quad \Delta_3] \begin{bmatrix} 2(y^*)^4 \Delta_1 + 8x^*(y^*)^3 \Delta_2 \\ 8x^*(y^*)^3 \Delta_1 + 12(x^*)^2(y^*)^2 \Delta_2 + 6z^* \Delta_3 \\ 6z^* \Delta_2 + 6y^* \Delta_3 \end{bmatrix} \\ &= (2(y^*)^4 \Delta_1 + 8x^*(y^*)^3 \Delta_2) \Delta_1 \\ &\quad + (8x^*(y^*)^3 \Delta_1 + 12(x^*)^2(y^*)^2 \Delta_2 + 6z^* \Delta_3) \Delta_2 \\ &\quad + (6z^* \Delta_2 + 6y^* \Delta_3) \Delta_3 \end{aligned}$$

Thus,

$$\begin{aligned} |\mathcal{E}_1(\mathbf{x}_0, \Delta)| &\leq (2r^5 + 8r^5)r + (8r^5 + 12r^5 + 6r^2)r + (6r^2 + 6r^2)r \\ &= 30r^6 + 18r^3 \end{aligned}$$

If we further restrict our attention to $r < 1$, then $r^6 < r^2$ and $r^3 < r^2$. Thus, our overestimate of the error is $|\mathcal{E}_1(\mathbf{x}_0, \Delta)| < 48r^2$. A similar analysis will show an overestimate $|\mathcal{E}_1(\mathbf{x}_0, \Delta)| < Kr^2$ for some integer K . Hence, the total overestimate of the error is

$$\|\mathcal{E}\| = \sqrt{(\mathcal{E}_1(\mathbf{x}_0, \Delta)^2 + (\mathcal{E}_2(\mathbf{x}_0, \Delta)^2)} \leq \sqrt{(48 + K)r^2} = \sqrt{(48 + K)}r$$

We can therefore determine the value of r which will guarantee the error in making our linear approximation is as small as we like.

Comment 11.6.1 These calculations are horrible if the point involved is different from the origin. We do some for the two variable case in (Peterson (8) 2019) so, in principle, we know what to do. But

really intense!

11.6.2.1 Homework

Exercise 11.6.6 For the following function find its gradient and hessian and the tangent plane approximation plus error at the point $(0, 0, 0)$.

$$f(x, y, z) = x^2 + x^4 y^3 z^2 + 14x^2 y^3 z^4$$

Find the approximate error as a function of r .

Exercise 11.6.7 Find the linear approximation of

$$f(x, y, z) = \begin{bmatrix} 2x^2 + 3y^4 + 5z^4 \\ -2x + 5y - 8z^2 \end{bmatrix}$$

near $(0, 0, 0)$. Find the approximate error \mathcal{E}_1 as a function of r .

Exercise 11.6.8 Find the linear approximation of

$$f(x, y) = \begin{bmatrix} 2x^2 + 3y^4 \\ -2x^5 y^2 + 5y^6 \end{bmatrix}$$

near $(0, 0)$. Find the approximate errors \mathcal{E}_1 , \mathcal{E}_2 and the total error \mathcal{E} as a function of r .

11.7 A Specific Coordinate Transformation

Given a function $g(x, y)$, the Laplacian of g is

$$\nabla^2 g(x, y) = \frac{\partial^2 g}{\partial x^2}(x, y) + \frac{\partial^2 g}{\partial y^2}(x, y) = g_{xx}(x, y) + g_{yy}(x, y)$$

We usually leave off the (x, y) as it is understood and just write $\nabla^2 g = g_{xx} + g_{yy}$. Let's convert this to polar coordinates. We assume g has continuous first and second order partials and so the mixed order partials match. Define $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ by

$$T\left(\begin{bmatrix} r \\ \theta \end{bmatrix}\right) = \begin{bmatrix} T_1(r, \theta) \\ T_2(r, \theta) \end{bmatrix} = \begin{bmatrix} r \cos(\theta) \\ r \sin(\theta) \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$$

We casually abuse the notation here all the time. We usually write the column vector $\begin{bmatrix} r \\ \theta \end{bmatrix}$ as the ordered pair (r, θ) which is a lot like identifying the column vector with its transpose $\begin{bmatrix} r & \theta \end{bmatrix}$. There are similar confusions built into the notation for (x, y) . This is because the vector space \mathbb{R}^2 can be interpreted as ordered pairs, column vectors or row vectors and we generally just make these transformations without thinking about it. But, of course, they are always there. So we would normally write the statements above as

$$T(r, \theta) = \begin{bmatrix} T_1(r, \theta) \\ T_2(r, \theta) \end{bmatrix} = \begin{bmatrix} r \cos(\theta) \\ r \sin(\theta) \end{bmatrix} = (x, y)$$

Now define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$f(r, \theta) = (g \circ T)(r, \theta) = g(T_1(r, \theta), T_2(r, \theta)) = g(r \cos(\theta), r \sin(\theta)) = g(x, y)$$

We know

$$\begin{aligned}\mathbf{D}\mathbf{f}(r, \theta) &= (\nabla(f))^T(r, \theta) = [f_r(r, \theta) \quad f_\theta(r, \theta)] \\ \mathbf{D}\mathbf{g}(x, y) &= (\nabla(g))^T(x, y) = [g_x(x, y) \quad g_y(x, y)]\end{aligned}$$

and

$$\mathbf{DT}(r, \theta) = \mathbf{J}_T(r, \theta) = \begin{bmatrix} T_{1,r}(r, \theta) & T_{1,\theta}(r, \theta) \\ T_{2,r}(r, \theta) & T_{2,\theta}(r, \theta) \end{bmatrix} = \begin{bmatrix} x_r(r, \theta) & x_\theta(r, \theta) \\ y_r(r, \theta) & y_\theta(r, \theta) \end{bmatrix}$$

Hence, by the chain rule

$$\mathbf{D}\mathbf{f}(r, \theta) = \mathbf{D}\mathbf{g}(x, y) \mathbf{DT}(r, \theta)$$

or

$$[f_r(r, \theta) \quad f_\theta(r, \theta)] = [g_x(x, y) \quad g_y(x, y)] \begin{bmatrix} x_r(r, \theta) & x_\theta(r, \theta) \\ y_r(r, \theta) & y_\theta(r, \theta) \end{bmatrix}$$

Let's drop all of the (x, y) and (r, θ) terms as it just adds clutter. We have

$$[f_r \quad f_\theta] = [g_x \quad g_y] \begin{bmatrix} x_r & x_\theta \\ y_r & y_\theta \end{bmatrix}$$

which when multiplied out, gives our familiar expansions

$$\begin{aligned}f_r &= g_x x_r + g_y y_r \\ f_\theta &= g_x x_\theta + g_y y_\theta\end{aligned}$$

Thus, we have

$$\begin{aligned}f_r &= \cos(\theta)g_x + \sin(\theta)g_y \\ f_\theta &= -r \sin(\theta)g_x + r \cos(\theta)g_y\end{aligned}$$

This can be rewritten as

$$\begin{bmatrix} f_r \\ f_\theta \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -r \sin(\theta) & r \cos(\theta) \end{bmatrix} \begin{bmatrix} g_x \\ g_y \end{bmatrix}$$

Now we apply the chain rule again:

$$\begin{aligned}f_{\theta\theta} &= \frac{\partial}{\partial \theta}(-r \sin(\theta)g_x + r \cos(\theta)g_y) \\ &= (-r \sin(\theta)(g_{xx}x_\theta + g_{xy}y_\theta) + \frac{\partial}{\partial \theta}(-r \sin(\theta))g_x \\ &\quad + (r \cos(\theta))(g_{yx}x_\theta + g_{yy}y_\theta)) + \frac{\partial}{\partial \theta}(r \cos(\theta))g_y\end{aligned}$$

Thus,

$$\begin{aligned}f_{\theta\theta} &= (-r \sin(\theta)(-g_{xx}r \sin(\theta) + g_{xy}r \cos(\theta))) - r \cos(\theta)g_x \\ &\quad + (r \cos(\theta))(g_{yx}(-r \sin(\theta)) + g_{yy}(r \cos(\theta))) - r \sin(\theta)g_y \\ &= r^2(\sin^2(\theta)g_{xx} - \sin(\theta)\cos(\theta)g_{xy}) + r^2(-\sin(\theta)\cos(\theta)g_{yx} + \cos^2(\theta)g_{yy}) \\ &\quad - r(\cos(\theta)g_x + \sin(\theta)g_y)\end{aligned}$$

So, we now know

$$\begin{aligned}\frac{1}{r^2}f_{\theta\theta} &= \sin^2(\theta)g_{xx} - 2\sin(\theta)\cos(\theta)g_{xy} + \cos^2(\theta)g_{yy} - \frac{1}{r}(\cos(\theta)g_x + \sin(\theta)g_y) \\ &= \sin^2(\theta)g_{xx} - 2\sin(\theta)\cos(\theta)g_{xy} + \cos^2(\theta)g_{yy} - \frac{1}{r}f_r\end{aligned}$$

Hence, we have shown

$$\frac{1}{r^2}f_{\theta\theta} + \frac{1}{r}f_r = \sin^2(\theta)g_{xx} - 2\sin(\theta)\cos(\theta)g_{xy} + \cos^2(\theta)g_{yy}$$

Next, we need

$$\begin{aligned}f_{rr} &= \frac{\partial}{\partial r}(\cos(\theta)g_x + \sin(\theta)g_y) \\ &= \cos(\theta)(g_{xx}x_r + g_{xy}y_r) + \frac{\partial}{\partial r}(\cos(\theta))g_x + \sin(\theta)(g_{yx}x_r + g_{yy}y_r) \\ &\quad + \frac{\partial}{\partial r}(\sin(\theta))g_y \\ &= \cos(\theta)(g_{xx}\cos(\theta) + g_{xy}\sin(\theta)) + \sin(\theta)(g_{yx}\cos(\theta) + g_{yy}\sin(\theta))\end{aligned}$$

Thus,

$$f_{rr} = \cos^2(\theta)g_{xx} + 2\cos(\theta)\sin(\theta)g_{xy} + \sin^2(\theta)g_{yy}$$

Combining

$$\begin{aligned}f_{rr} + \frac{1}{r^2}f_{\theta\theta} + \frac{1}{r}f_r &= \sin^2(\theta)g_{xx} - 2\sin(\theta)\cos(\theta)g_{xy} + \cos^2(\theta)g_{yy} \\ &\quad + \cos^2(\theta)g_{xx} + 2\cos(\theta)\sin(\theta)g_{xy} + \sin^2(\theta)g_{yy} \\ &= g_{xx} + g_{yy}\end{aligned}$$

So the Laplacian in polar coordinates is

$$\nabla_{r\theta}^2 f = f_{rr} + \frac{1}{r^2}f_{\theta\theta} + \frac{1}{r}f_r = \nabla_{xy}^2 g$$

where $f(r, \theta) = g(T(r, \theta))$.

11.7.1 Homework

Converting partial differential equations in cartesian coordinates into their equivalent forms under a nonlinear coordinate transformation (these are called curvilinear transformations) is a really useful idea. We will leave much of the detail of the following problems to you as they are nicely detailed and intense. Remember, going through this sort of thing on your own is a very useful exercise and helps to build your personal skill sets. As a reference, any book on old fashioned advanced calculus and mathematical physics goes over this stuff. Stay away from the newer books and pick up an old text. A good choice is (Wrede and Speigel (16) 2010).

Exercise 11.7.1 Find the Laplacian in cylindrical coordinates.

Exercise 11.7.2 Find the Laplacian in spherical coordinates.

Exercise 11.7.3 Find the Laplacian in paraboloidal coordinates.

Exercise 11.7.4 *Find the Laplacian in ellipsoidal coordinates.*

Chapter 12

Multivariable Extremal Theory

Now let's look at the extremes of functions of many variables.

12.1 Differentiability and Extremals

We are now ready to connect the value of partial derivatives at an extremum such as a minimum or maximum of a scalar function of n variables. We know continuous functions on compact domains possess such extreme values, but we do not really know how to find them. Here is a start.

Theorem 12.1.1 At an Interior Extreme Point, The Partialials Vanish

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ have an extreme value at the interior point \mathbf{x}_0 in D . Then if all the partials $f_{x_i}(\mathbf{x}_0)$ exist, they must be zero.

Proof 12.1.1

Let \mathbf{x}_0 be an extreme value for f which is an interior point. For convenience of exposition, let's assume the extremum is a maximum locally. So there is a radius $r > 0$ so that if $\mathbf{y} \in B(r, \mathbf{x}_0)$, $f(\mathbf{y}) \leq f(\mathbf{x}_0)$. Then if each $f_{x_i}(\mathbf{x}_0)$ exists, for small enough $h > 0$, we must have

$$f(\mathbf{x}_0 + h\mathbf{E}_i) \leq f(\mathbf{x}_0)$$

Thus,

$$\frac{f(\mathbf{x}_0 + h\mathbf{E}_i) - f(\mathbf{x}_0)}{h} \leq 0 \implies \lim_{h \rightarrow 0^+} \frac{f(\mathbf{x}_0 + h\mathbf{E}_i) - f(\mathbf{x}_0)}{h} \leq 0$$

This tells us $(f_{x_i}(\mathbf{x}_0))^+ \leq 0$. On the other hand, for sufficiently small $h < 0$, we have

$$f(\mathbf{x}_0 + h\mathbf{E}_i) \leq f(\mathbf{x}_0)$$

Thus, because h is negative

$$\frac{f(\mathbf{x}_0 + h\mathbf{E}_i) - f(\mathbf{x}_0)}{h} \geq 0 \implies \lim_{h \rightarrow 0^-} \frac{f(\mathbf{x}_0 + h\mathbf{E}_i) - f(\mathbf{x}_0)}{h} \geq 0$$

This tells us $(f_{x_i}(\mathbf{x}_0))^- \geq 0$. Since we assume $f_{x_i}(\mathbf{x}_0)$ exists, we must have $(f_{x_i}(\mathbf{x}_0))^- = (f_{x_i}(\mathbf{x}_0))^+ = f_{x_i}(\mathbf{x}_0)$. Thus $0 \leq f_{x_i}(\mathbf{x}_0) \leq 0$ which tells us $f_{x_i}(\mathbf{x}_0) = 0$. We can do this for each index i . Thus, we have shown that if \mathbf{x}_0 is an interior point which corresponds to a local max-

imum, then if the partials $f_{x_i}(\mathbf{x}_0)$ exist they must be zero; i.e. $\nabla(f) = \mathbf{0}$. We can do a similar argument for a local minimum. ■

Comment 12.1.1 Now if $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable, then $Df = (\nabla(f))^T$ and so the result above tells that if f is differentiable at an extreme point which is an interior point \mathbf{x}_0 then $Df(\mathbf{x}_0) = \mathbf{0}$.

Points where the gradient is zero are thus important. Points where extreme behavior occurs are called **critical points** of f . Once we have found them, the next question is to classify them as minima, maxima or something else. For completeness, we state a formal definition.

Definition 12.1.1 Critical Points

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$. The point \mathbf{x}_0 is called a **critical point** of f if $\nabla(f)(\mathbf{x}_0) = \mathbf{0}$, the $\nabla(f)(\mathbf{x}_0)$ fails to exist or \mathbf{x}_0 is a boundary point of D . Recall this means $B(r, \mathbf{x}_0)$ has points of D and D^C for all choices of $r > 0$.

We have already discussed conditions for classifying critical points in the two variable case in (Peterson (8) 2019). For completeness, we will go over this again.

12.2 Second Order Extremal Conditions

At a place where the extremum of a function of n variables might occur with $\nabla(f)(\mathbf{x}_0) = \mathbf{0}$, it is clear the tangent plane will be **flat** as all of the partials will be zero. Of course, we can only see the flatness of this plane in three space but we will use that idea to help our intuition in higher dimensional spaces. From our discussion of differentiability, we know

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \sum_{i=1}^n f_{x_i}(\mathbf{x}_0) \Delta_i + \frac{1}{2} \Delta^T \mathbf{H}(\mathbf{x}_0 + c\Delta) \Delta$$

as long as all the second order partials of f all exist locally at \mathbf{x}_0 . Here,

$$\frac{1}{2} \Delta^T \mathbf{H}(\mathbf{x}_0 + c\Delta) \Delta = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n f_{x_i x_j}(\mathbf{x}_0 + c\Delta) \Delta_i \Delta_j$$

Now since \mathbf{x}_0 is a critical point with $\nabla(f)(\mathbf{x}_0) = \mathbf{0}$, we can write

$$\begin{aligned} f(\mathbf{x}_0 + \Delta) &= f(\mathbf{x}_0) + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n f_{x_i x_j}(\mathbf{x}_0 + c\Delta) \Delta_i \Delta_j \\ &= f(\mathbf{x}_0) + \frac{1}{2} \Delta^T \mathbf{H}(\mathbf{x}_0 + c\Delta) \Delta \end{aligned}$$

We can decide if we have a minimum or a maximum easily. The matrix $\mathbf{H}(\mathbf{x}_0 + c\Delta)$ is an $n \times n$ matrix. If we assume all the second order partials are continuous locally at \mathbf{x}_0 , then $\det \mathbf{H}(\mathbf{x}_0 + c\Delta)$ is also continuous locally at \mathbf{x}_0 . Hence, we can say there is a $B(r, \mathbf{x}_0)$ where both \mathbf{H} and $\det \mathbf{H}$ are continuous. We also know if a continuous function is positive at \mathbf{x}_0 , it is strictly positive locally at \mathbf{x}_0 . Similarly, if a continuous function is negative at \mathbf{x}_0 , it is strictly negative locally at \mathbf{x}_0 . Thus, there is a $B(r_1, \mathbf{x}_0)$ where $r_1 < r$ so that

- If $\det \mathbf{H}(\mathbf{x}_0) > 0$, then $\det \mathbf{H}(\mathbf{y}) > 0$ if $\mathbf{y} \in B(r_1, \mathbf{x}_0)$.
- If $\det \mathbf{H}(\mathbf{x}_0) < 0$, then $\det \mathbf{H}(\mathbf{y}) < 0$ if $\mathbf{y} \in B(r_1, \mathbf{x}_0)$.

So if we start with $\|\Delta\| < r_1$, the c we obtain above will give us $x_0 + c\Delta \in B(r_1, x_0)$. Thus, we will be able to say

- If $\det H(x_0) > 0$, then $\det H(x_0 + c\Delta) > 0$
- If $\det H(x_0) < 0$, then $\det H(x_0 + c\Delta) < 0$

Since all the second order partials are continuous locally, we know the mixed order partials must be the same. Hence the Hessian here is a symmetric matrix.

12.2.1 Positive and Negative Definite Hessians

We can get the algebraic sign we need for $\det H(x_0)$ using the notion of positive and negative definite matrices. The conditions above become

- If $H(x_0)$ is positive definite, then $\Delta^T H(x_0) \Delta > 0$ for any Δ . Since $\Delta^T H(x_0) \Delta$ is continuous at x_0 and $\Delta^T H(x_0) \Delta > 0$, this means there is a radius $r_2 < r_1$ so that $\Delta^T H(y) \Delta > 0$ in $B(r_2, x_0)$. By choosing $\|\Delta\| < r_2$, we guarantee c is small enough so that $x_0 + c\Delta \in B(r_2, x_0)$. Hence, $\Delta^T H(x_0 + c\Delta) \Delta > 0$. Thus, we have

$$f(x_0 + \Delta) = f(x_0) + \text{a positive number}$$

implying $f(x_0 + \Delta) > f(x_0)$ locally. Hence, f has a local minimum at x_0 .

We also know since $H(x_0)$ is positive definite, all of its eigenvalues are positive. We discussed the idea of the principle minors of a matrix in Chapter 10. You can go back there to review. Let λ_1 through λ_n be the eigenvalues of $H(x_0)$. We find for our $H(x_0)$ that

- The first minor is the determinant of 1×1 submatrix $H_{ij}(x_0)$ for $i = 1$ and $j = 2$. This is λ_1 . Call this determinant PM_1 .
- The second minor is the determinant of 2×2 submatrix $H_{ij}(x_0)$ for $1 \leq i, j \leq 2$. This is $\lambda_1 \lambda_2$. Call this determinant PM_2 .
- The third minor is the determinant of 3×3 submatrix $H_{ij}(x_0)$ for $1 \leq i, j \leq 2$. This is $\lambda_1 \lambda_2 \lambda_3$. Call this determinant PM_3 .
- The k^{th} minor is the determinant of $k \times k$ submatrix $H_{ij}(x_0)$ for $1 \leq i, j \leq k$. This is $\lambda_1 \dots \lambda_k$. Call this determinant PM_k .

Note we have a local minimum if the algebraic sign of the minors is $+ + + \dots +$. We can easily find the eigenvalues in MatLab and determine this. For $n > 2$, these minor calculations are intense. Note we are not phrasing the conditions for the local minimum in terms of the second order partials at x_0 although we can.

- If $H(x_0)$ is negative definite, then we can analyze just as we did above. We have $\Delta^T H(x_0) \Delta < 0$ for any Δ . Since $\Delta^T H(x_0) \Delta$ is continuous at x_0 and $\Delta^T H(x_0) \Delta < 0$, this means there is a radius $r_2 < r_1$ so that $\Delta^T H(y) \Delta < 0$ in $B(r_2, x_0)$. By choosing $\|\Delta\| < r_2$, we guarantee c is small enough so that $x_0 + c\Delta \in B(r_2, x_0)$. Hence, $\Delta^T H(x_0 + c\Delta) \Delta > 0$. Thus, we have

$$f(x_0 + \Delta) = f(x_0) + \text{a negative number}$$

implying $f(x_0 + \Delta) > f(x_0)$ locally. Hence, f has a local maximum at x_0 .

We also know since $H(x_0)$ is negative definite, all of its eigenvalues are negative. Let λ_1 through λ_n be the eigenvalues of $H(x_0)$. We find for our $H(x_0)$ that

$$- PM_1 = \lambda_1 < 0.$$

- $\mathbf{P}M_2 = \lambda_1 \lambda_2 > 0$.
- $\mathbf{P}M_3 = \lambda_1 \lambda_2 \lambda_3 < 0$.
- $\mathbf{P}M_k = \lambda_1 \dots \lambda_k$. The algebraic sign is determined by $(-1)^k$.

Note we have a local maximum if the algebraic sign of the minors is $- + - \dots (-1)^n$; i.e. the algebraic sign of the minor oscillates. We can easily find the eigenvalues in MatLab and determine this.

Since the Hessian $\mathbf{H}(\mathbf{y})$ is symmetric, its eigenvectors determine an orthonormal basis for \Re^n , $\{\mathbf{E}_1(\mathbf{y}), \dots, \mathbf{E}_n(\mathbf{y})\}$. Defining the matrix

$$\mathbf{P}(\mathbf{y}) = [\mathbf{E}_1(\mathbf{y}) \ \dots \ \mathbf{E}_n(\mathbf{y})]$$

whose i^{th} column is $\mathbf{E}_i(\mathbf{y})$, we have

$$\begin{aligned}\mathbf{H}(\mathbf{x}_0) &= \mathbf{P}(\mathbf{x}_0)\mathbf{D}(\mathbf{x}_0)\mathbf{P}^T(\mathbf{x}_0) \\ \mathbf{H}(\mathbf{x}_0 + c\Delta) &= \mathbf{P}(\mathbf{x}_0 + c\Delta)\mathbf{D}(\mathbf{x}_0 + c\Delta)\mathbf{P}^T(\mathbf{x}_0 + c\Delta)\end{aligned}$$

where $\mathbf{D}(\mathbf{y})$ is the diagonal matrix of the eigenvalues of $\mathbf{H}(\mathbf{y})$.

If $\mathbf{H}(\mathbf{x}_0)$ is positive or negative definite, its minors are either strictly positive or negative and hence, because $\det \mathbf{H}(\mathbf{x}_0)$ is continuous at \mathbf{x}_0 , there is a radius $r_2 < r_1$, where all the minors have the same algebraic sign as long as $\mathbf{x}_0 + c\Delta \in B(r_3, \mathbf{x}_0)$. By choosing $\|\Delta\|$ sufficiently small we can guarantee this. Hence, the remarks above concerning the minors and eigenvalues still hold. We have

$$\begin{aligned}f(\mathbf{x}_0 + \Delta) &= f(\mathbf{x}_0) + \frac{1}{2}\Delta^T \mathbf{H}(\mathbf{x}_0 + c\Delta)\Delta \\ &= f(\mathbf{x}_0) + \frac{1}{2}\Delta^T \mathbf{P}(\mathbf{x}_0 + c\Delta)\mathbf{D}(\mathbf{x}_0 + c\Delta)\mathbf{P}^T(\mathbf{x}_0 + c\Delta)\Delta \\ &= f(\mathbf{x}_0) + \frac{1}{2}\left(\mathbf{P}^T(\mathbf{x}_0 + c\Delta)\Delta\right)^T \mathbf{D}(\mathbf{x}_0 + c\Delta)\left(\mathbf{P}^T(\mathbf{x}_0 + c\Delta)\Delta\right)\end{aligned}$$

Let $\xi^c = \mathbf{P}^T(\mathbf{x}_0 + c\Delta)\Delta$. Then we have

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \frac{1}{2}\xi^{cT} \mathbf{D}(\mathbf{x}_0 + c\Delta)\xi^c$$

Let the eigenvalues of $\mathbf{H}(\mathbf{x}_0 + c\Delta)$ be denoted by λ_i^c for convenience. Then we can rewrite as

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \frac{1}{2}(\lambda_1^c(\xi_1^c)^2 + \dots + \lambda_n^c(\xi_n^c)^2)$$

and the signs of these eigenvalues are the same as the signs of the eigenvalues for $\mathbf{H}(\mathbf{x}_0)$ by our choice of $\|\Delta\| < r_3$. Note the following: the equation above makes it very clear that

- If $\mathbf{H}(\mathbf{x}_0)$ is positive definite, so is $\mathbf{H}(\mathbf{x}_0 + c\Delta)$ and so $\lambda_1^c(\xi_1^c)^2 + \dots + \lambda_n^c(\xi_n^c)^2 > 0$ gives us a local minimum.
- If $\mathbf{H}(\mathbf{x}_0)$ is negative definite, so is $\mathbf{H}(\mathbf{x}_0 + c\Delta)$ and so $\lambda_1^c(\xi_1^c)^2 + \dots + \lambda_n^c(\xi_n^c)^2 < 0$ gives us a local maximum.

12.2.2 Saddles

Since

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \frac{1}{2}(\lambda_1^c(\xi_1^c)^2 + \dots + \lambda_n^c(\xi_n^c)^2)$$

if the eigenvalues are not all the same sign, we can have what is called a saddle structure. Suppose λ_i^c and λ_j^c had different signs for some $i \neq j$. For concreteness, suppose $\lambda_i^c > 0$ and let $\lambda_i^c = \alpha > 0$. We also have $\lambda_j^c < 0$. Write $\lambda_j^c = -\beta$ where $\beta > 0$. Set all the other λ_k^c 's to zero. Then on that surface trace, we have

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \frac{1}{2} (\alpha(\xi_i^c)^2 - \beta(\xi_j^c)^2)$$

This shows the function has the behavior on the additional trace obtained by setting $\xi_i^c = 0$

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) - \frac{1}{2}\beta(\xi_j^c)^2$$

which shows f has a local maximum along this trace in \Re^n . On the other hand, if we set $\xi_j^c = 0$

$$f(\mathbf{x}_0 + \Delta) = f(\mathbf{x}_0) + \frac{1}{2}\alpha(\xi_i^c)^2$$

which shows f has a local minimum along this trace. This type of behavior, where f behaves like it has a minimum along some directions and a maximum along others is called **saddle** behavior. We conclude if $\mathbf{H}(\mathbf{x}_0)$ is not positive or negative definite, it will have eigenvalues of different sign. This will imply a saddle at \mathbf{x}_0 .

Of course, if $\det \mathbf{H}(\mathbf{x}_0) = 0$, we don't know anything. It might have a local minimum, a local maximum, a saddle or none of those things.

12.2.3 Expressing Conditions in Terms of Partial

The minor conditions can be expressed in terms of the partials. We have

$$\mathbf{PM}_1 = \det [f_{x_1 x_1}(\mathbf{x}_0)] = f_{x_1 x_1}(\mathbf{x}_0)$$

Then the second minor is

$$\mathbf{PM}_2 = \det \begin{bmatrix} f_{x_1 x_1}(\mathbf{x}_0) & f_{x_1 x_2}(\mathbf{x}_0) \\ f_{x_2 x_1}(\mathbf{x}_0) & f_{x_2 x_2}(\mathbf{x}_0) \end{bmatrix} = f_{x_1 x_1}(\mathbf{x}_0)f_{x_2 x_2}(\mathbf{x}_0) - (f_{x_1 x_2}(\mathbf{x}_0))^2$$

The third minor requires a 3×3 determinant

$$\mathbf{PM}_3 = \det \begin{bmatrix} f_{x_1 x_1}(\mathbf{x}_0) & f_{x_1 x_2}(\mathbf{x}_0) & f_{x_1 x_3}(\mathbf{x}_0) \\ f_{x_2 x_1}(\mathbf{x}_0) & f_{x_2 x_2}(\mathbf{x}_0) & f_{x_2 x_3}(\mathbf{x}_0) \\ f_{x_3 x_1}(\mathbf{x}_0) & f_{x_3 x_2}(\mathbf{x}_0) & f_{x_3 x_3}(\mathbf{x}_0) \end{bmatrix}$$

which is much more messy to write out and so we will not do that. We discuss the two variable case in (Peterson (8) 2019) using different strategies of proof that involve a completing the square. Clearly that technique is limited to the two variable case and the extra theory we use here allows us to get a better grasp on the extremal structure of f . In the two variable case, we find

- f has a local minimum at \mathbf{x}_0 with $\nabla(f)(\mathbf{x}_0) = \mathbf{0}$ when
 1. $f_{x_1 x_1}(\mathbf{x}_0) > 0$ (first minor condition).
 2. $f_{x_1 x_1}(\mathbf{x}_0)f_{x_2 x_2}(\mathbf{x}_0) - (f_{x_1 x_2}(\mathbf{x}_0))^2 > 0$ (second minor condition)
- f has a local maximum at \mathbf{x}_0 with $\nabla(f)(\mathbf{x}_0) = \mathbf{0}$ when
 1. $f_{x_1 x_1}(\mathbf{x}_0) < 0$ (first minor condition).

2. $f_{x_1 x_1}(\mathbf{x}_0) f_{x_2 x_2}(\mathbf{x}_0) - (f_{x_1 x_2}(\mathbf{x}_0))^2 > 0$ (second minor condition)

- f has a saddle at \mathbf{x}_0) if the two eigenvalues of $H(\mathbf{x}_0)$ have different signs. This implies $f_{x_1 x_1}(\mathbf{x}_0) f_{x_2 x_2}(\mathbf{x}_0) - (f_{x_1 x_2}(\mathbf{x}_0))^2 < 0$

The standard two variable theorem is thus:

Theorem 12.2.1 Extreme Test

If the partials of f are zero at the point (x_0, y_0) , we can determine if that point is a local minimum or local maximum of f using a second order test. We must assume the second order partials are continuous at the point (x_0, y_0) .

- If $f_{xx}^0 > 0$ and $f_{xx}^0 f_{yy}^0 - (f_{xy}^0)^2 > 0$ then $f(x_0, y_0)$ is a local minimum.
- If $f_{xx}^0 < 0$ and $f_{xx}^0 f_{yy}^0 - (f_{xy}^0)^2 > 0$ then $f(x_0, y_0)$ is a local maximum.
- If $f_{xx}^0 f_{yy}^0 - (f_{xy}^0)^2 < 0$ then $f(x_0, y_0)$ is a local saddle.

We just don't know anything if the test $f_{xx}^0 f_{yy}^0 - (f_{xy}^0)^2 = 0$.

Example 12.2.1 Use our tests to show $f(x, y) = x^2 + 3y^2$ has a minimum at $(0, 0)$.

Solution The partials here are $f_x = 2x$ and $f_y = 6y$. These are zero at $x = 0$ and $y = 0$. The Hessian at this critical point is

$$H(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix} = H(0, 0).$$

as H is constant here. Our second order test says the point $(0, 0)$ corresponds to a minimum because $f_{xx}(0, 0) = 2 > 0$ and $f_{xx}(0, 0) f_{yy}(0, 0) - (f_{xy}(0, 0))^2 = 12 > 0$.

Example 12.2.2 Use our tests to show $f(x, y) = x^2 + 6xy + 3y^2$ has a saddle at $(0, 0)$.

Solution The partials here are $f_x = 2x + 6y$ and $f_y = 6x + 6y$. These are zero at when $2x + 6y = 0$ and $6x + 6y = 0$ which has solution $x = 0$ and $y = 0$. The Hessian at this critical point is

$$H(x, y) = \begin{bmatrix} 2 & 6 \\ 6 & 6 \end{bmatrix} = H(0, 0).$$

as H is again constant here. Our second order test says the point $(0, 0)$ corresponds to a saddle because $f_{xx}(0, 0) = 2 > 0$ and $f_{xx}(0, 0) f_{yy}(0, 0) - (f_{xy}(0, 0))^2 = 12 - 36 < 0$.

Example 12.2.3 Show our tests fail on $f(x, y) = 2x^4 + 4y^6$ even though we know there is a minimum value at $(0, 0)$.

Solution For $f(x, y) = 2x^4 + 4y^6$, you find that the critical point is $(0, 0)$ and all the second order partials are 0 there. So all the tests fail. Of course, a little common sense tells you $(0, 0)$ is indeed the place where this function has a minimum value. Just think about how it's surface looks. But the tests just fail. This is much like the curve $f(x) = x^4$ which has a minimum at $x = 0$ but all the tests fail on it also.

Example 12.2.4 Show our tests fail on $f(x, y) = 2x^2 + 4y^3$ and the surface does not have a minimum or maximum at the critical point $(0, 0)$.

Solution For $f(x, y) = 2x^2 + 4y^3$, the critical point is again $(0, 0)$ and $f_{xx}(0, 0) = 4$, $f_{yy}(0, 0) = 0$ and $f_{xy}(0, 0) = f_{yx}(0, 0) = 0$. So $f_{xx}(0, 0) f_{yy}(0, 0) - (f_{xy}(0, 0))^2 = 0$ so the test fails. Note the $x = 0$ trace is $4y^3$ which is a cubic and so is negative below $y = 0$ and positive above $y = 0$. Not

much like a minimum or maximum behavior on this trace! But the trace for $y = 0$ is $2x^2$ which is a nice parabola which does reach its minimum at $x = 0$. So the behavior of the surface around $(0, 0)$ is not a maximum or a minimum.

Homework

Exercise 12.2.1

Exercise 12.2.2

Exercise 12.2.3

The theorem for n variables is this:

Theorem 12.2.2 Definiteness Extrema Test For n Variables

We assume the second order partials are continuous at the point \mathbf{x}_0 and $\nabla(f)(\mathbf{x}_0) = \mathbf{0}$.

- If the Hessian $\mathbf{H}(\mathbf{x}_0)$ is positive definite, the critical point is a local minimum.
- If the Hessian $\mathbf{H}(\mathbf{x}_0)$ is negative definite, the critical point is a local maximum.
- If the local Hessian $\mathbf{H}(\mathbf{x}_0)$ has eigenvalues of different sign, the the critical point is a saddle.

Moreover, the Hessian $\mathbf{H}(\mathbf{x}_0)$ is positive definite if its principle minors are all positive and has positive eigenvalues and $\mathbf{H}(\mathbf{x}_0, \mathbf{y}_0)$ is negative definite if its principle minors alternate in sign starting at $-$ and all of its eigenvalues are negative.

Proof 12.2.1

We have gone over the proof of this results in our earlier discussions. ■

Example 12.2.5 Let $f(x_1, x_2, x_3, x_4, x_5) = 2x_1^2 + 4x_2^2 + 9x_3^2 + 5x_4^2 + 8x_5^2$. Classify the critical points.

Solution The only critical point is $(0, 0, 0, 0, 0)$. The Hessian is

$$\mathbf{H} = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 18 & 0 & 0 \\ 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix}$$

\mathbf{H} is clearly positive definite with eigenvalues $4, 8, 18, 10, 6$ and the eigenvectors are the standard basic vectors. The minors are $4, 32, 576, 5760, 92160$ and are all positive. Hence, this critical point is a minimum.

Example 12.2.6 Let $f(x_1, x_2, x_3, x_4, x_5) = 2x_1^2 - 4x_2^2 + 9x_3^2 + 5x_4^2 + 8x_5^2$. Classify the critical points.

Solution The only critical point is $(0, 0, 0, 0, 0)$. The Hessian is

$$\mathbf{H} = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & -8 & 0 & 0 & 0 \\ 0 & 0 & 18 & 0 & 0 \\ 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 16 \end{bmatrix}$$

\mathbf{H} is clearly not positive or negative definite with eigenvalues 4, -8, 18, 10, 6 and the eigenvectors are the standard basic vectors. The minors are 4, -32, -576, 5760, -92160. Hence, this critical point is a saddle.

Example 12.2.7 Let $f(x_1, x_2, x_3, x_4, x_5) = 100 - 2x_1^2 - 4x_2^2 - 9x_3^2 - 5x_4^2 - 8x_5^2$. Classify the critical points.

Solution The only critical point is $(0, 0, 0, 0, 0)$. The Hessian is

$$\mathbf{H} = \begin{bmatrix} -4 & 0 & 0 & 0 & 0 \\ 0 & -8 & 0 & 0 & 0 \\ 0 & 0 & -18 & 0 & 0 \\ 0 & 0 & 0 & -10 & 0 \\ 0 & 0 & 0 & 0 & -16 \end{bmatrix}$$

\mathbf{H} is clearly not positive or negative definite with eigenvalues -4, -8, -18, -10, -6 and the eigenvectors are the standard basic vectors. The minors are -4, 32, -576, 5760, -92160. Hence, this critical point is a maximum.

Homework

The following 2×2 matrices give the Hessian \mathbf{H}^0 at the critical point $(1, 1)$ of an extremal value problem.

- Determine if \mathbf{H}^0 is positive or negative definite
- Determine if the critical point is a maximum or a minimum
- Find the two eigenvalues of \mathbf{H}^0 . Label the largest one as r_1 and the other as r_2
- Find the two associated eigenvectors as unit vectors
- Define

$$\mathbf{P} = [\mathbf{E}_1 \quad \mathbf{E}_2]$$

- Compute $\mathbf{P}^T \mathbf{H}^0 \mathbf{P}$
- Show $\mathbf{H}^0 = \mathbf{P} \Lambda \mathbf{P}^T$
- Find all the minors for an appropriate Λ .

Exercise 12.2.4

$$\mathbf{H}^0 = \begin{bmatrix} 1 & -3 \\ -3 & 12 \end{bmatrix}$$

Exercise 12.2.5

$$\mathbf{H}^0 = \begin{bmatrix} -2 & 7 \\ 7 & -40 \end{bmatrix}$$

The following 3×3 matrices give the Hessian \mathbf{H}^0 at the critical point $(1, 1, 2)$ of an extremal value problem.

- Determine if \mathbf{H}^0 is positive or negative definite
- Classify the critical point

- Find the three eigenvalues of \mathbf{H}^0 . Label the largest one as r_1 and the other as r_2
- Find the three associated eigenvectors as unit vectors
- Define the usual \mathbf{P} and compute $\mathbf{P}^T \mathbf{H}^0 \mathbf{P}$
- Show $\mathbf{H}^0 = \mathbf{P} \Lambda \mathbf{P}^T$ for an appropriate Λ .
- Find all the minors

Exercise 12.2.6

$$\mathbf{H}^0 = \begin{bmatrix} 1 & -3 & 4 \\ -3 & 12 & 5 \\ 4 & 5 & 4 \end{bmatrix}$$

Exercise 12.2.7

$$\mathbf{H}^0 = \begin{bmatrix} -2 & 7 & -1 \\ 7 & -4 & 5 \\ -1 & 5 & 12 \end{bmatrix}$$

The following 4×4 matrices give the Hessian \mathbf{H}^0 at the critical point $(1, 1, 2, -3)$ of an extremal value problem.

- Determine if \mathbf{H}^0 is positive or negative definite
- Classify the critical point
- Find the four eigenvalues of \mathbf{H}^0 . Label the largest one as r_1 and the other as r_2
- Find the three associated eigenvectors as unit vectors
- Define the usual \mathbf{P} and compute $\mathbf{P}^T \mathbf{H}^0 \mathbf{P}$
- Show $\mathbf{H}^0 = \mathbf{P} \Lambda \mathbf{P}^T$ for an appropriate Λ .
- Find all the minors

Exercise 12.2.8

$$\mathbf{H}^0 = \begin{bmatrix} 1 & -3 & 4 & 8 \\ -3 & 12 & 5 & 2 \\ 4 & 5 & 4 & -1 \\ 8 & 2 & -1 & 8 \end{bmatrix}$$

Exercise 12.2.9

$$\mathbf{H}^0 = \begin{bmatrix} -2 & 7 & -1 & 3 \\ 7 & -4 & 5 & 10 \\ -1 & 5 & 12 & 5 \\ 3 & 10 & 5 & 9 \end{bmatrix}$$

Chapter 13

The Inverse and Implicit Function Theorems

Let's look at mappings that transform one set of coordinates into another set.

13.1 Mappings

First, look at linear maps. This particular one could be replaced by any other with a nonzero determinant. So you can make up plenty of examples.

Example 13.1.1 Define $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ by $T(x, y) = (u, v)$ where $u = x - 3y$ and $v = 2x + 5y$. We can write this in matrix/vector form

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} T_1(x, y) \\ T_2(x, y) \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} x - 3y \\ 2x + 5y \end{bmatrix} = \begin{bmatrix} 1 & -3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

So there are many ways to choose to express this mapping. The matrix representation of T with respect to an orthonormal basis \mathbf{E} is

$$[T]_{\mathbf{E}, \mathbf{E}} = \begin{bmatrix} 1 & -3 \\ 2 & 5 \end{bmatrix}$$

We could also write the mapping simply as $\mathbf{U} = [T]_{\mathbf{E}, \mathbf{E}} \mathbf{X}$ where we let \mathbf{U} be the vector with components u and v and \mathbf{X} , the vector with components x and y . However, this is too cluttered and it is easy to identify T and its matrix representation $[T]_{\mathbf{E}, \mathbf{E}} \mathbf{X}$ which we will do from now on. By direct calculation, $\det(T) = 11 \neq 0$, so T^{-1} exists. Hence if $T(x_1, y_1) = T(x_2, y_2)$, we would have

$$\begin{bmatrix} 1 & -3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 & -3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \implies \begin{bmatrix} 1 & -3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies x_1 = x_2, y_1 = y_2$$

Thus, T is a 1–1 map. Further, if we chose a vector \mathbf{B} in \mathbb{R}^2 , then

$$\begin{bmatrix} 1 & -3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \implies \begin{bmatrix} x \\ y \end{bmatrix} = \left(\begin{bmatrix} 1 & -3 \\ 2 & 5 \end{bmatrix} \right)^{-1} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

and so we also know T is an onto map. Further, it is easy to see $J_T = T$, so we have since T is a linear map

$$DT = J_T = \begin{bmatrix} 1 & -3 \\ 2 & 5 \end{bmatrix}$$

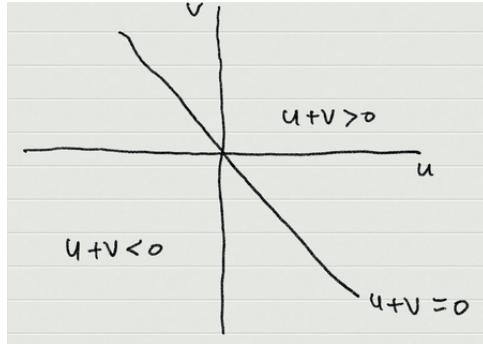
Therefore, in this case, $\det(DT(x_0, y_0)) = \det(T) \neq 0$ on \mathbb{R}^2 .

Now let's do a nonlinear mapping.

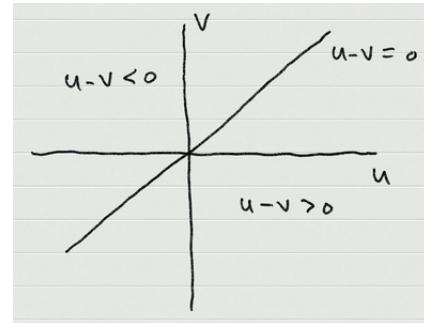
Example 13.1.2 We are now going to study $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $T(x, y) = (x^2 + y^2, 2xy)$. This mapping T is not linear and not 1–1 because $T(x, y) = T(-x, -y)$. What is the range of T ? If (u, v) is in the range, then $u = x^2 + y^2$ implying $u \geq 0$ always. Further, $v = 2xy$. Note

$$\begin{aligned} u + v &= x^2 + 2xy + y^2 = (x + y)^2 \geq 0 \\ u - v &= x^2 - 2xy + y^2 = (x - y)^2 \geq 0 \end{aligned}$$

Now $u + v$ divides the $u - v$ plane into three parts as seen in Figure 13.1(a): where $u + v > 0$, $u + v = 0$ and $u + v < 0$. Also, $u - v$ divides the plane into three parts also as seen in Figure 13.1(b). Hence, to be in the range, both $u + v$ and $u - v$ must be non negative and we already know



(a) The $u + v$ Plane Subdivisions



(b) The $u - v$ Plane Subdivisions

Figure 13.1: The $u + v$ and $u - v$ lines

$u \geq 0$. Thus, the range of T is the part of the $u - v$ plane shown in Figure 13.2(a); i.e

$$\mathcal{R}(T) = \{(u, v) | u \geq 0, u + v \geq 0, u - v \geq 0\} = \{(u, v) | u \geq 0, -u \leq v \leq u\}$$

The domain of T is \mathbb{R}^2 and it can be divided into four regions as shown in Figure 13.2(b). Thus, the four indicated regions are the disjoint and open sets

$$\begin{aligned} S_1 &= \{(x, y) | x > 0, -x < y < x\}, \quad S_2 = \{(x, y) | y > 0, -y < x < y\} \\ S_3 &= \{(x, y) | x < 0, -|x| < y < |x|\}, \quad S_4 = \{(x, y) | y < 0, -|y| < x < |y|\} \end{aligned}$$

Proposition 13.1.1 Finding Where $(x^2 + y^2, 2xy)$ is a Bijection

The mapping $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $T(x, y) = (x^2 + y^2, 2xy)$. is invertible on $\overline{S_2}$ and $T|_{\overline{S_2}}$ is onto $\mathcal{R}(T)$.

13.1. MAPPINGS

235

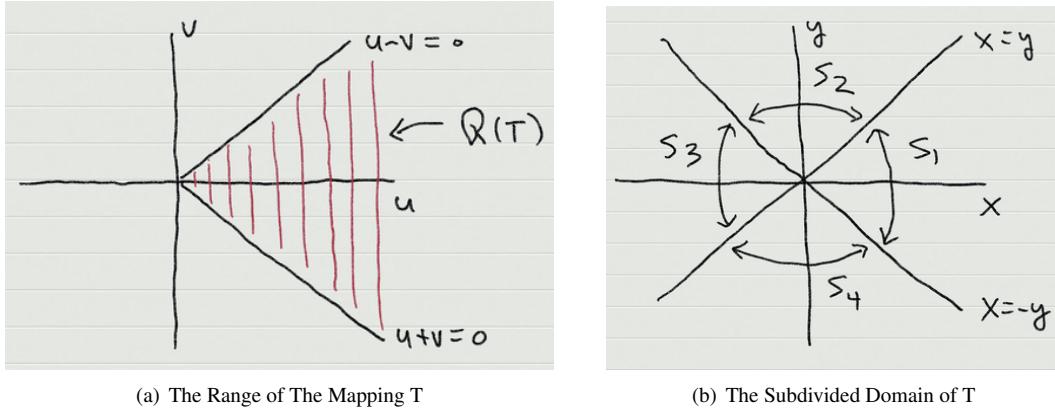


Figure 13.2: The Range and Domain of T

Proof 13.1.1

Assume $T(x_0, y_0) = T(x_1, y_1)$. Since (x_0, y_0) and (x_1, y_1) are in S_2 , we have $y_0 > 0, -y_0 < x_0 < y_0$ and $y_1 > 0, -y_1 < x_1 < y_1$. Now if $T(x_0, y_0) = (u_0, v_0)$ and $T(x_1, y_1) = (u_1, v_1)$,

$$\begin{aligned} u_0 + v_0 &= (x_0 + y_0)^2, & u_0 - v_0 &= (x_0 - y_0)^2 \\ u_1 + v_1 &= (x_1 + y_1)^2, & u_1 - v_1 &= (x_1 - y_1)^2 \end{aligned}$$

Since $T(x_0, y_0) = T(x_1, y_1)$, $(x_0 + y_0)^2 = (x_1 + y_1)^2$ and $(x_0 - y_0)^2 = (x_1 - y_1)^2$. Now in S_2 , $x_0 + y_0 > 0$ and $x_1 + y_1 > 0$. Thus, we can take the positive square root to find $x_0 + y_0 = x_1 + y_1$. Also, in S_2 , $x_0 - y_0 > 0$ and $x_1 - y_1 > 0$. Hence, in this case, we take the negative square root to get $y_0 - x_0 = y_1 - x_1$. We now know

$$\begin{aligned} x_0 + y_0 &= x_1 + y_1 \\ -x_0 + y_0 &= -x_1 + y_1 \end{aligned}$$

Adding, we have $y_0 = y_1$ which then implies $x_0 = x_1$ as well. Hence, T is 1–1 on S_2 .

In fact if $x \geq 0$ and $y = x$, $T(x, y) = T(x, x) = (2x^2, 2x^2)$ which is on the line $y = x$ in the range of T . Moreover, if $x \geq 0$ and $y = -x$, $T(x, y) = T(x, -x) = (2x^2, -2x^2)$ which is on the line $y = -x$ in the range of T . It is then easy to see this shows T is 1–1 on these lines too. So we can say T is 1–1 on $\overline{S_2}$. Hence, T^{-1} exists on $\overline{S_2}$.

To show $T : \overline{S_2} \rightarrow \mathcal{R}(T)$ is onto is next. We'll do this by constructing the inverse. If $T(x_0, y_0) = (u_0, v_0)$ with $(x_0, y_0) \in \overline{S_2}$, then

$$u_0 + v_0 = (x_0 + y_0)^2 \geq 0, \quad u_0 - v_0 = (x_0 - y_0)^2 \geq 0$$

In S_2 , $x_0 + y_0 > 0$ and $x_0 - y_0 < 0$. Taking square roots, we have

$$x_0 + y_0 = \sqrt{u_0 + v_0}, \quad x_0 - y_0 = -\sqrt{u_0 - v_0}$$

which implies

$$x_0 = \frac{1}{2}(\sqrt{u_0 + v_0} - \sqrt{u_0 - v_0})$$

$$y_0 = \frac{1}{2}(\sqrt{u_0 + v_0} + \sqrt{u_0 - v_0})$$

Hence,

$$T^{-1}(u_0, v_0) = \left(\frac{1}{2}(\sqrt{u_0 + v_0} - \sqrt{u_0 - v_0}), \frac{1}{2}(\sqrt{u_0 + v_0} + \sqrt{u_0 - v_0}) \right)$$

On the line $u = v$, if $(x, y) \in \overline{S_2}$, $y \geq 0$, $x + y \geq 0$ and $x - y \leq 0$. Thus

$$u + v = (x + y)^2, \quad u - v = (x - y)^2 \implies 2u = (x + y)^2, \quad 0 = (x - y)^2 \implies \frac{1}{2}\sqrt{2u} = x, y = x$$

Thus, if $u = v$, $T^{-1}(u, u) = (\frac{1}{2}\sqrt{2u}, \frac{1}{2}\sqrt{2u})$.

If $v = -u$, the argument is similar: we have

$$\begin{aligned} u + v &= (x + y)^2, \quad u - v = (x - y)^2 \implies 0 = (x + y)^2, \quad 2u = (x - y)^2 \\ &\implies y = -x, y > 0, -\frac{1}{2}\sqrt{2u} = x \end{aligned}$$

So if $u = -v$, $T^{-1}(u, -u) = (-\frac{1}{2}\sqrt{2u}, \frac{1}{2}\sqrt{2u})$. This completes the argument that $T : \overline{S_2} \rightarrow \mathcal{R}(T)$ is onto. ■

The map T is differentiable with

$$\mathbf{DT}(x_0, y_0) = \mathbf{J}_T(x_0, y_0) = \begin{bmatrix} 2x_0 & 2y_0 \\ 2y_0 & 2x_0 \end{bmatrix}$$

with $\det(\mathbf{DT}(x_0, y_0)) = 4(x_0^2 - x_1^2)$. Hence, if $(x_0, y_0) \in S_2$, $\det(\mathbf{DT}(x_0, y_0)) \neq 0$ and if (x_0, y_0) is on the boundary of S_2 , $\det(\mathbf{DT}(x_0, y_0)) = 0$. Thus, the determinant vanishes on the lines $y = \pm x$ in S_2 . Finally, in the interior of S_2 , $x + y > 0$ and $x - y < 0$ and so $\det(\mathbf{DT}(x_0, y_0)) = 4(x_0 - x_1)(x_0 + x_1) < 0$.

These two examples give us some insight.

- If $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ has $\det(\mathbf{DT}(x_0, y_0)) \neq 0$ does not necessarily force T^{-1} to exist on all of \mathbb{R}^2 . Our second example shows us we may need to restrict T to a suitable domain for the inverse to exist. This also shows us T^{-1} can exist even when $\det(\mathbf{DT}(x_0, y_0)) = 0$ as we found on the lines $y = \pm x$.
- Given $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ which is linear, we have a mapping **too nice** to be interesting!

Homework

Exercise 13.1.1

Exercise 13.1.2

Exercise 13.1.3

Exercise 13.1.4

Exercise 13.1.5

We have been using the determinant of the Jacobian of a mapping and we know $\mathbf{DF} = \mathbf{J}_F$ for our vector valued differentiable functions. In the literature, the determinant of the jacobian is very useful and some texts use the term **jacobian** for the **determinant of the Jacobian**. We won't do that here

and instead will always use $\det J_F$ to denote this computation. Let's look at our first **invertibility** theorem.

Theorem 13.1.2 The First Inverse Function Theorem

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ with D an open set. Assume f_{i,x_j} are all continuous on D and assume $Df(\mathbf{x}_0) \neq 0$ at $\mathbf{x}_0 \in D$. Then there is a radius $r > 0$ so that T is 1–1 on $B(r, \mathbf{x}_0)$; i.e. F^{-1} exists locally.

Proof 13.1.2

If $n = 1$, this is a simple scalar function of one variable. We have no partial derivatives and $Df(x_0) = f'(x_0)$. If $f'(x_0) \neq 0$, then given $\epsilon = |f'(x_0)|/2$, there is a $\delta > 0$ so that $|f'(x)| > |f'(x_0)|/2$ if $x \in B(\delta, x_0)$. Pick any two distinct points p and q in $B(\delta, x_0)$ and let $[p, q]$ be the line segment between p and q . By the Mean Value Theorem, there is point c in $[p, q]$ so that $f(p) - f(q) = f'(c)(q - p)$. Now if $f(p) = f(q)$, this would mean $0 = f'(c)(p - q)$. But $f'(c) \neq 0$ and $p \neq q$ and so this is not possible. Hence in $B(\delta, x_0)$, f must be 1–1 and so is invertible.

If $n > 1$, the argument gets more interesting. Before we do the general one, let $n = 2$. Define the function $M : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$M(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} f_{1,x_1}(\mathbf{x}) & f_{1,x_2}(\mathbf{x}) \\ f_{2,x_1}(\mathbf{y}) & f_{2,x_2}(\mathbf{y}) \end{bmatrix}$$

Then

$$\det M(\mathbf{x}, \mathbf{y}) = f_{1,x_1}(\mathbf{x}) f_{2,x_2}(\mathbf{y}) - f_{1,x_2}(\mathbf{x}) f_{2,x_1}(\mathbf{y})$$

Since f_{1,x_1} , f_{1,x_2} , f_{2,x_1} and f_{2,x_2} are continuous in D , so are the functions $M(\mathbf{x}, \mathbf{y})$ and $\det M(\mathbf{x}, \mathbf{y})$. We assume $\det J_f(\mathbf{x}_0) = M(\mathbf{x}_0, \mathbf{x}_0) \neq 0$ and hence, for $\epsilon = |\det M(\mathbf{x}_0, \mathbf{x}_0)|/2$, there is a positive δ so that if \mathbf{x} and \mathbf{y} are in $B(\delta, \mathbf{x}_0)$

$$|\det M(\mathbf{x}, \mathbf{y})| > |\det M(\mathbf{x}_0, \mathbf{x}_0)|/2 > 0$$

Now choose any two distinct points \mathbf{p} and \mathbf{q} in $B(\delta, \mathbf{x}_0)$. By the Mean Value Theorem, there are points \mathbf{c}_1 and \mathbf{c}_2 on the line segment $[\mathbf{p}, \mathbf{q}]$ between \mathbf{p} and \mathbf{q} so that

$$\begin{aligned} f_1(\mathbf{p}) - f_1(\mathbf{q}) &= <(\nabla(f_1))^T(\mathbf{c}_1, \mathbf{p} - \mathbf{q})> \\ f_2(\mathbf{p}) - f_2(\mathbf{q}) &= <(\nabla(f_2))^T(\mathbf{c}_2, \mathbf{p} - \mathbf{q})> \end{aligned}$$

But if $f(\mathbf{p}) = f(\mathbf{q})$, then

$$\begin{aligned} <(\nabla(f_1))^T(\mathbf{c}_1, \mathbf{p} - \mathbf{q})> &= 0 \\ <(\nabla(f_2))^T(\mathbf{c}_2, \mathbf{p} - \mathbf{q})> &= 0 \end{aligned}$$

But we know $\det M(\mathbf{c}_1, \mathbf{c}_2) \neq 0$ and so the only solution is $\mathbf{p} = \mathbf{q}$ which is not possible. So the assumption that $f(\mathbf{p}) = f(\mathbf{q})$ is not correct. We conclude f is 1–1 on $B(\delta, \mathbf{x}_0)$ and is invertible there.

To do the general theorem, we use the material on determinants developed in Chapter 10. By Theorem 10.1.3,

$$M(\mathbf{x}_1, \dots, \mathbf{x}_n) = \det \begin{bmatrix} (\nabla(f_1))^T(\mathbf{x}_1) \\ \vdots \\ (\nabla(f_n))^T(\mathbf{x}_n) \end{bmatrix}$$

is a continuous function of the variables $\mathbf{x}_1, \dots, \mathbf{x}_n$. Since $\det J_f(\mathbf{x}_0) = \det M(\mathbf{x}_0, \dots, \mathbf{x}_0) \neq 0$, for $\epsilon = |\det M(\mathbf{x}_0, \mathbf{x}_0)|/2$, there is a positive δ so that if \mathbf{x}_1 through \mathbf{x}_n are in $B(\delta, \mathbf{x}_0)$

$$|\det M(\mathbf{x}_1, \dots, \mathbf{x}_n)| > |\det M(\mathbf{x}_0, \mathbf{x}_0)|/2 > 0$$

Now choose any two distinct points \mathbf{p} and \mathbf{q} in $B(\delta, \mathbf{x}_0)$. By the Mean Value Theorem, there are points \mathbf{c}_1 through \mathbf{c}_n on the line segment $[\mathbf{p}, \mathbf{q}]$ between \mathbf{p} and \mathbf{q} so that

$$\begin{aligned} f_1(\mathbf{p}) - f_1(\mathbf{q}) &= <(\nabla(f_1))^T(\mathbf{c}_1, \mathbf{p} - \mathbf{q})> \\ &\vdots = \vdots \\ f_n(\mathbf{p}) - f_n(\mathbf{q}) &= <(\nabla(f_n))^T(\mathbf{c}_n, \mathbf{p} - \mathbf{q})> \end{aligned}$$

But $f(\mathbf{p}) = f(\mathbf{q})$ and so

$$\begin{aligned} <(\nabla(f_1))^T(\mathbf{c}_1, \mathbf{p} - \mathbf{q})> &= 0 \\ &\vdots = \vdots \\ <(\nabla(f_n))^T(\mathbf{c}_n, \mathbf{p} - \mathbf{q})> &= 0 \end{aligned}$$

The determinant is zero if its rows are linearly independent and it is non zero if its rows are linearly independent. Hence, the only solution to the above equation is $\mathbf{p} = \mathbf{q} = \mathbf{0}$. But this is not possible and hence f must be 1–1 on $B(\delta, \mathbf{x}_0)$ and it invertible there. ■

Homework

Exercise 13.1.6

Exercise 13.1.7

Exercise 13.1.8

Exercise 13.1.9

Exercise 13.1.10

13.2 Invertibility Results

We begin with a technical result.

Theorem 13.2.1 Interior Points of the Range of f

Let $D \subset \mathbb{R}^n$ be open and $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ with continuous partial derivatives on D . Assume \mathbf{x}_0 is in D and there is a positive r so that $\overline{B(r, \mathbf{x}_0)} \subset D$ with $\det J_f(\mathbf{x}) \neq 0$ for all \mathbf{x} in $\overline{B(r, \mathbf{x}_0)}$. Further, assume f is 1–1 on $\overline{B(r, \mathbf{x}_0)}$. then $f(\mathbf{x}_0)$ is an interior point of $f(B(r, \mathbf{x}_0))$.

Proof 13.2.1

For \mathbf{x} in $\overline{B(r, \mathbf{x}_0)}$, let $g(\mathbf{x}) = \|f(\mathbf{x} - f(\mathbf{x}_0)\|$. The g is a non negative real valued function whose domain in compact. We know the map $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous and f is continuous since it is differentiable. Note the continuous partials imply its differentiability. Hence, the composition $h(\mathbf{x}) = \|f(\mathbf{x} - f(\mathbf{x}_0)\|$ is continuous on the compact set $\overline{B(r, \mathbf{x}_0)}$. We see $g(\mathbf{x}) = 0$ implies $f(\mathbf{x} = f(\mathbf{x}_0)$. Since we assume f is 1-1 on this domain, this means $\mathbf{x} = \mathbf{x}_0$. Hence h is 1-1 on $\overline{B(r, \mathbf{x}_0)}$. Finally, note the boundary $S = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_0\| = r\}$ is a compact set also. Since f is 1-1 on this set, g can not be zero on S ; so $g(\mathbf{y}) > 0$ if $\mathbf{y} \in S$.

Since g is continuous on the compact set S , g attains a global minimum of value $m > 0$ for some $\mathbf{y} \in S$.

Show $B(m/2, f(\mathbf{x}_0)) \subset f(B(r, \mathbf{x}_0))$

Proof Let \mathbf{z} be in $B(m/2, f(\mathbf{x}_0))$. We must find a \mathbf{p} in $B(r, \mathbf{x}_0)$ with $\mathbf{z} = f(\mathbf{p})$. Let ϕ be the real valued function on $\overline{B(r, \mathbf{x}_0)}$ defined by

$$\phi(\mathbf{x}) = \|f(\mathbf{x}) - \mathbf{z}\| = \sqrt{\sum_{i=1}^n (f_i(\mathbf{x}) - z_i)^2}$$

We see ϕ is continuous on the compact set $\overline{B(r, \mathbf{x}_0)}$ and so attains a minimum value α at some point \mathbf{y}_0 in $\overline{B(r, \mathbf{x}_0)}$. Since \mathbf{z} is in $B(m/2, f(\mathbf{x}_0))$, $\|f(\mathbf{x}_0) - \mathbf{z}\| < m/2$ and so $\phi(\mathbf{x}_0) = \|f(\mathbf{x}_0) - \mathbf{z}\| < m/2$. This tells us

$$\alpha = \min_{\mathbf{x} \in \overline{B(r, \mathbf{x}_0)}} \phi(\mathbf{x}) = \min_{\mathbf{x} \in \overline{B(r, \mathbf{x}_0)}} \|f(\mathbf{x}) - \mathbf{z}\| = \phi(\mathbf{y}_0) \leq \phi(\mathbf{x}_0) < m/2.$$

Is it possible for $\mathbf{y}_0 \in S$? If so,

$$\begin{aligned} \phi(\mathbf{y}_0) &= \|f(\mathbf{y}_0) - \mathbf{z}\| = \|f(\mathbf{y}_0) - f(\mathbf{x}_0) + f(\mathbf{x}_0) - \mathbf{z}\| \\ &\geq \|f(\mathbf{y}_0) - f(\mathbf{x}_0)\| - \|f(\mathbf{x}_0) - \mathbf{z}\| \\ &= g(\mathbf{y}_0) - \|f(\mathbf{x}_0) - \mathbf{z}\| \end{aligned}$$

But $\|f(\mathbf{x}_0) - \mathbf{z}\| < m/2$ and $g(\mathbf{y}_0) \geq m > 0$ if \mathbf{y}_0 is in S . This implies

$$\phi(\mathbf{y}_0) \geq m - m/2 = m/2 > 0$$

But we know

$$\min_{\mathbf{x} \in \overline{B(r, \mathbf{x}_0)}} \phi(\mathbf{x}) = \alpha < m/2$$

This means $\phi(\mathbf{y}_0) \geq m/2$ is not possible. We conclude \mathbf{y}_0 is not in S . So the point where ϕ attains its minimum on $\overline{B(r, \mathbf{x}_0)}$ is not on the boundary S ; so it is an interior point and $\mathbf{y}_0 \in B(r, \mathbf{x}_0)$.

Since ϕ has a minimum at \mathbf{y}_0 , so does ϕ^2 . Since the minimum is an interior point, we must have $(\phi^2)_{x_i}(\mathbf{y}_0) = 0$ for all i . Now

$$\phi^2(\mathbf{x}) = \sum_{i=1}^n (f_i(\mathbf{x}) - z_i)^2 \implies (\phi^2)_{x_i} = 2 \sum_{i=1}^n (f_i(\mathbf{x}) - z_i) f_{i,x_i}$$

Hence at \mathbf{y}_0 ,

$$(\phi^2)_{x_i}(\mathbf{x}_0) = 2 \sum_{i=1}^n (f_i(\mathbf{x}_0) - z_i) f_{i,x_i}(\mathbf{x}_0) = 0$$

This is the same as $\mathbf{J}_f(\mathbf{y}_0)(f(\mathbf{y}_0) - \mathbf{z}) = \mathbf{0}$. Since $\det \mathbf{J}_f(\mathbf{y}_0) \neq 0$, this tells the unique solution is $f(\mathbf{y}_0) = \mathbf{z}$. We conclude $\mathbf{z} \in B(m/2, f(\mathbf{x}_0))$ or $f(\mathbf{y}_0) \in f(B(r, \mathbf{x}_0))$ as $\mathbf{y}_0 \in B(r, \mathbf{x}_0)$. Hence, $B(m/2, f(\mathbf{x}_0)) \subset f(B(r, \mathbf{x}_0))$ \square

This, of course, says $f(\mathbf{x}_0)$ is an interior point of $f(B(r, \mathbf{x}_0))$ and completes the proof. \blacksquare

We are now ready to tackle the **Inverse Function Theorem**. First, a result about inverse images of open sets.

Theorem 13.2.2 Inverse Images of open sets under continuous maps are open

Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a continuous mapping of the open set D . If V is open in the range of f , then $f^{-1}(V)$ is open in D .

Proof 13.2.3

Let $\mathbf{x}_0 \in f^{-1}(V)$. Then $f(\mathbf{x}_0) \in V$. Since V is open, $f(\mathbf{x}_0)$ is an interior point of V and so there is a radius $r > 0$ so that $B(r, f(\mathbf{x}_0)) \subset V$. Since f is continuous on D , for $\epsilon = r$, there is a $\delta > 0$ so that $\|f(\mathbf{x}) - f(\mathbf{x}_0)\| < r$ if $\|\mathbf{x} - \mathbf{x}_0\| < \delta$. Thus, $f(\mathbf{x}) \in B(r, f(\mathbf{x}_0))$ if $\mathbf{x} \in B(\delta, \mathbf{x}_0) \cap D$. Since D is open, $B(\delta, \mathbf{x}_0) \cap D$ is open and so there is a $\delta_1 > 0$ with $B(\delta_1, \mathbf{x}_0) \subset B(\delta, \mathbf{x}_0) \cap D$. This tells us $\mathbf{x} \in B(\delta_1, \mathbf{x}_0)$ implies $f(\mathbf{x}) \in B(r, f(\mathbf{x}_0)) \subset V$. We conclude $B(\delta_1, \mathbf{x}_0) \subset f^{-1}(V)$ and so \mathbf{x}_0 is an interior point of $f^{-1}(V)$. Since \mathbf{x}_0 is arbitrary, this shows $f^{-1}(V)$ is open. \blacksquare

Theorem 13.2.3 The Inverse Function Theorem

Let $D \subset \mathbb{R}^n$ be an open set and $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ has continuous first partial derivatives, f_{i,x_j} in D . Let $\mathbf{x}_0 \in D$ with $\det \mathbf{J}_f(\mathbf{x}_0) \neq 0$. Then there is an open set U containing \mathbf{x}_0 and an open set $V \subset f(U)$ and a map $g : V \subset \mathbb{R}^n \rightarrow U \subset \mathbb{R}^n$. so that

- $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ is 1 – 1 and onto V .
- $g : V \subset \mathbb{R}^n \rightarrow U \subset \mathbb{R}^n$ is 1 – 1 and onto U .
- $g \circ f(\mathbf{x}) = \mathbf{x}$ on U .
- g_{i,x_j} is continuous on V implying Dg exists on V .

Comment 13.2.1 If $\mathbf{y} \in V$, $g(\mathbf{y}) \in U$. So $f \circ g(\mathbf{y}) = f(g(\mathbf{y})) = \mathbf{z} \in V$. But there is an $\mathbf{x} \in U$ so that $f(\mathbf{x}) = \mathbf{y}$. Hence, $g(\mathbf{y}) = g(f(\mathbf{x})) = \mathbf{x}$ by the Theorem. We conclude $f \circ g(\mathbf{y}) = f(\mathbf{x}) = \mathbf{y}$. So we see both $f \circ g(\mathbf{x}) = \mathbf{x}$ and $g \circ f(\mathbf{y}) = \mathbf{y}$ on U and V respectively. So $g = f^{-1}$ and $f = g^{-1}$.

Proof 13.2.4

Since $\det \mathbf{J}_f(\mathbf{x}_0) \neq 0$ and f_{i,x_j} is continuous in D , the continuity of $\det \mathbf{J}_f(\mathbf{x}_0)$ implies we can find a positive r_0 so that $B(r_0, \mathbf{x}_0) \subset D$ and by Theorem 13.1.2, there is an $r_1 < r_0$, so that f is 1 – 1 on $B(r_1, \mathbf{x}_0)$. Choose any $r < r_1$. Then

$$\overline{B(r, \mathbf{x}_0)} \subset B(r_1, \mathbf{x}_0) \implies f[1] : -1 \text{ on } \overline{B(r_1, \mathbf{x}_0)}$$

13.2. INVERTIBILITY RESULTS

241

Now by Theorem 13.2.1, $f(\mathbf{x}_0)$ is an interior point of $f(B(r, \mathbf{x}_0))$. This tells us there is an $r_2 > 0$ so that $B(r_2, f(\mathbf{x}_0)) \subset f(B(r, \mathbf{x}_0))$. Let $V = B(r_2, f(\mathbf{x}_0))$ and $U = f^{-1}(V)$. Note $f^{-1}(V)$ is a subset of $B(r, \mathbf{x}_0)$. You can see these sets in Figure 13.3. Since V is open, $f^{-1}(V)$ is open also since

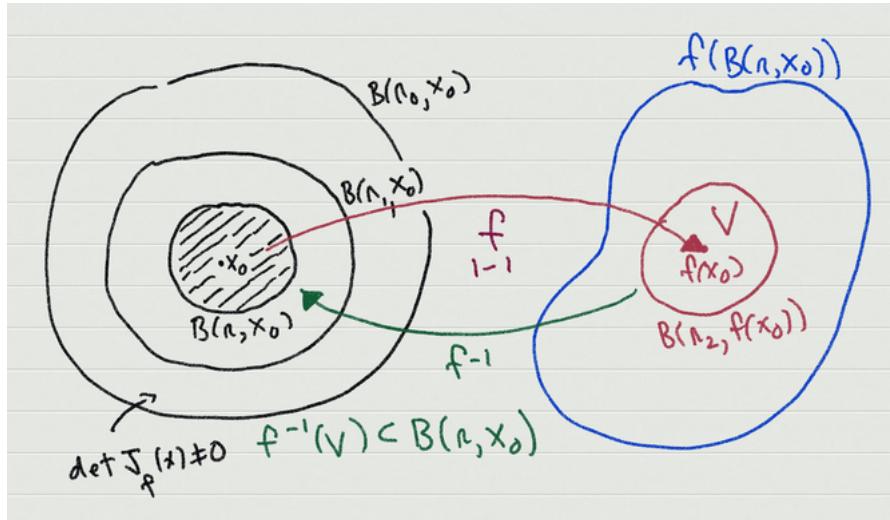


Figure 13.3: Inverse Function Mappings and Sets

f is continuous. We now know f is 1–1 on $B(r, \mathbf{x}_0)$ implying f is 1–1 on U with $U \subset B(r, \mathbf{x}_0)$. Thus, $g = f^{-1}$ does exist on $f(U) = V$ and $g \circ f(\mathbf{x}) = \mathbf{x}$ on U . This is part of what we need to prove. There is a lot more to do.

f is onto V :

Proof Let $\mathbf{y} \in V$. Then $\mathbf{y} \in B(r_2, f(\mathbf{x}_0)) \subset f(B(r, \mathbf{x}_0))$. Hence, there is $\mathbf{x} \in B(r, \mathbf{x}_0)$ with $\mathbf{y} = f(\mathbf{x})$ and so $\mathbf{x} \in f^{-1}(V) = U$. So f is onto V . \square

g is onto U :

Proof If $\mathbf{x} \in U = f^{-1}(V)$, then there is $\mathbf{y} \in V$ so that $f(\mathbf{x}) = \mathbf{y}$ because f is onto V . Thus, $g(f(\mathbf{x})) = g(\mathbf{y})$. However, we know g is f^{-1} here, so $g(\mathbf{y}) = \mathbf{x}$ and we have shown g is onto U . \square

g^{-1} is continuous on U :

Proof We know f is continuous and 1–1 on the compact set $\overline{B(r, \mathbf{x}_0)}$. Hence, $g = f^{-1}$ is continuous on $f(\overline{B(r, \mathbf{x}_0)})$. This implies g is continuous on V . What about g^{-1} 's continuity? Let (\mathbf{y}_n) in V converge to $\mathbf{y} \in V$. We know f is 1–1 on $\overline{B(r, \mathbf{x}_0)}$ to $f(\overline{B(r, \mathbf{x}_0)})$. So there is a sequence (\mathbf{x}_n) in $\overline{B(r, \mathbf{x}_0)}$ with $f(\mathbf{x}_n) = \mathbf{y}_n$. Since $\overline{B(r, \mathbf{x}_0)}$ is compact, there is a subsequence (\mathbf{x}_n^1) which converges to a point \mathbf{x} in $\overline{B(r, \mathbf{x}_0)}$. Since f is continuous there, we must have $f(\mathbf{x}_n^1) \rightarrow f(\mathbf{x})$. But $f(\mathbf{x}_n^1) = \mathbf{y}_n^1$ and since $\mathbf{y}_n \rightarrow \mathbf{y}$, we have $\mathbf{y}_n^1 \rightarrow \mathbf{y}$. Thus, $f(\mathbf{x}) = \mathbf{y}$.

If there was another subsequence (\mathbf{x}_n^2) which converged to \mathbf{x}^* in $\overline{B(r, \mathbf{x}_0)}$, by the continuity of f , we have $f(\mathbf{x}_n^2) \rightarrow f(\mathbf{x}^*)$. But $f(\mathbf{x}_n^2) = \mathbf{y}_n^2$ and so $\mathbf{y}_n^2 \rightarrow \mathbf{y}$. Thus, $f(\mathbf{x}^*) = \mathbf{y}$.

But $f(\mathbf{x}) = f(\mathbf{x}^*)$ means $\mathbf{x} = \mathbf{x}^*$ as f is 1–1. We have shown any convergent subsequence of \mathbf{x}_n must converge to the same point \mathbf{x} . These vector sequences have n component sequences. We have shown the set of subsequential limits for each component sequence consists of one number. Hence,

each component sequence of x_i converges. Thus, the set of subsequential limits of the sequence is \mathbf{x} and so $\mathbf{x}_n \rightarrow \mathbf{x}$. But this says $g^{-1}(\mathbf{y}_n) \rightarrow g^{-1}(\mathbf{y})$. \square

g_{x_i} exists and is continuous on V :

Proof Fix $\mathbf{y} \in V$ and choose indices j and k from $\{1, \dots, n\}$. We need to show

$$\lim_{h \rightarrow 0} \frac{g_j(\mathbf{y} + h\mathbf{E}_k) - g_j(\mathbf{y})}{h}$$

exists and that g_{y_j} is continuous. For small enough h , the line segment $[\mathbf{y}, \mathbf{y} + h\mathbf{E}_k]$ is in V . Let

$$\begin{aligned} \mathbf{x} &= g(\mathbf{y} = f^{-1}(\mathbf{y})) \in U \\ \mathbf{z} &= g(\mathbf{y} + h\mathbf{E}_k) = f^{-1}(\mathbf{y} + h\mathbf{E}_k) \in U \end{aligned}$$

Then, \mathbf{x} and \mathbf{z} are distinct as f^{-1} is 1–1. Note the line segment $[\mathbf{x}, \mathbf{z}]$ is in $B(r, \mathbf{x}_0)$. Since g is continuous on V , $\lim_{h \rightarrow 0} g(\mathbf{y} + h\mathbf{E}_k) = g(\mathbf{y})$. But this says $\lim_{h \rightarrow 0} \mathbf{z} = g(\mathbf{y}) = \mathbf{x}$

Now $f(\mathbf{z}) - f(\mathbf{x}) = \mathbf{y} + h\mathbf{E}_k - \mathbf{y} = h\mathbf{E}_k$, so

$$\begin{bmatrix} f_1(\mathbf{z}) - f_1(\mathbf{x}) \\ \vdots \\ f_{k-1}(\mathbf{z}) - f_{k-1}(\mathbf{x}) \\ f_k(\mathbf{z}) - f_k(\mathbf{x}) \\ f_{k+1}(\mathbf{z}) - f_{k+1}(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{z}) - f_n(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ h, \text{ component } k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Thus, $f_i(\mathbf{z}) = f_i(\mathbf{x})$ if $i \neq k$ and $f_k(\mathbf{z}) - f_k(\mathbf{x}) = h$. Now apply the Mean Value Theorem to each f_i . Then, there is a $\mathbf{w}_i \in [\mathbf{z}, \mathbf{x}]$ with

$$f_i(\mathbf{z}) - f_i(\mathbf{x}) = (\nabla(f_i))^T(\mathbf{z} - \mathbf{x}) = (\nabla(f_i))^T(\mathbf{w}_i)(g_i(\mathbf{y} + h\mathbf{E}_k) - g_i(\mathbf{y}))$$

This implies

$$\begin{bmatrix} (\nabla(f_1))^T(\mathbf{w}_1) \\ \vdots \\ (\nabla(f_{k-1}))^T(\mathbf{w}_{k-1}) \\ (\nabla(f_k))^T(\mathbf{w}_k) \\ (\nabla(f_{k+1}))^T(\mathbf{w}_{k+1}) \\ \vdots \\ (\nabla(f_n))^T(\mathbf{w}_n) \end{bmatrix} \begin{bmatrix} (g_1(\mathbf{y} + h\mathbf{E}_k) - g_1(\mathbf{y}))/h \\ \vdots \\ (g_{k-1}(\mathbf{y} + h\mathbf{E}_k) - g_{k-1}(\mathbf{y}))/h \\ (g_k(\mathbf{y} + h\mathbf{E}_k) - g_k(\mathbf{y}))/h \\ (g_{k+1}(\mathbf{y} + h\mathbf{E}_k) - g_{k+1}(\mathbf{y}))/h \\ \vdots \\ (g_n(\mathbf{y} + h\mathbf{E}_k) - g_n(\mathbf{y}))/h \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1, \text{ component } k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Since each \mathbf{w}_i is in the line segment $[z, x]$ which is in $B(r, \mathbf{x}_0) \subset B(r_0, \mathbf{x}_0)$, we know for

$$\Delta(\mathbf{w}_1, \dots, \mathbf{w}_n) = \begin{bmatrix} (\nabla(f_1))^T(\mathbf{w}_1) \\ \vdots \\ (\nabla(f_{k-1}))^T(\mathbf{w}_{k-1}) \\ (\nabla(f_k))^T(\mathbf{w}_k) \\ (\nabla(f_{k+1}))^T(\mathbf{w}_{k+1}) \\ \vdots \\ (\nabla(f_n))^T(\mathbf{w}_n) \end{bmatrix}$$

$\det \Delta(\mathbf{w}_1, \dots, \mathbf{w}_n) \neq 0$ and there is a unique solution for each $(g_k(\mathbf{y} + h\mathbf{E}_k) - g_k(\mathbf{y}))/h$. We can find this solution by Cramer's rule. Let Δ_i be the submatrix we get by replacing column i with the data vector \mathbf{E}_k .

$$\Delta_i(\mathbf{w}_1, \dots, \mathbf{w}_n) = \begin{bmatrix} f_{1,x_1}(\mathbf{w}_1) & \dots & f_{1,x_{i-1}}(\mathbf{w}_1) & 0 & f_{1,x_{i+1}}(\mathbf{w}_1) & \dots & f_{1,x_n}(\mathbf{w}_1) \\ \vdots & \vdots & \vdots & 0 & \vdots & \vdots & \vdots \\ f_{k-1,x_1}(\mathbf{w}_{k-1}) & \dots & f_{k-1,x_{i-1}}(\mathbf{w}_{k-1}) & 0 & f_{k-1,x_{i+1}}(\mathbf{w}_{k-1}) & \dots & f_{k-1,x_n}(\mathbf{w}_{k-1}) \\ f_{k,x_1}(\mathbf{w}_k) & \dots & f_{k,x_{i-1}}(\mathbf{w}_k) & 1 & f_{k,x_{i+1}}(\mathbf{w}_k) & \dots & f_{k,x_n}(\mathbf{w}_k) \\ f_{k+1,x_1}(\mathbf{w}_{k+1}) & \dots & f_{k+1,x_{i-1}}(\mathbf{w}_{k+1}) & 0 & f_{k+1,x_{i+1}}(\mathbf{w}_{k+1}) & \dots & f_{k+1,x_n}(\mathbf{w}_{k+1}) \\ \vdots & \vdots & \vdots & 0 & \vdots & \vdots & \vdots \\ f_{n,x_1}(\mathbf{w}_n) & \dots & f_{n,x_{i-1}}(\mathbf{w}_n) & 0 & f_{n,x_{i+1}}(\mathbf{w}_n) & \dots & f_{n,x_n}(\mathbf{w}_n) \end{bmatrix}$$

From Cramer's rule, we know

$$(g_i(\mathbf{y} + h\mathbf{E}_k) - g_i(\mathbf{y}))/h = \frac{\det \Delta_i(\mathbf{w}_1, \dots, \mathbf{w}_n)}{\det \Delta(\mathbf{w}_1, \dots, \mathbf{w}_n)}$$

We assume the continuity of the partials on D and so $\det \Delta_i(\mathbf{w}_1, \dots, \mathbf{w}_n)$ and $\det \Delta(\mathbf{w}_1, \dots, \mathbf{w}_n)$ are continuous functions of $\mathbf{w}_1, \dots, \mathbf{w}_n$. Therefore we know

$$\lim_{h \rightarrow 0} \frac{\det \Delta_i(\mathbf{w}_1, \dots, \mathbf{w}_n)}{\det \Delta(\mathbf{w}_1, \dots, \mathbf{w}_n)} = \frac{\det \Delta_i(z_1, \dots, z_n)}{\det \Delta(z_1, \dots, z_n)}$$

Thus, $g_{i,y_k}(\mathbf{y}) = \lim_{h \rightarrow 0} (g_i(\mathbf{y} + h\mathbf{E}_k) - g_i(\mathbf{y}))/h$ exists. If \mathbf{Y}_p were a sequence converging to \mathbf{y} , then we would have

$$g_{i,y_k}(\mathbf{Y}_p) = \frac{\det \Delta_i(z_1^p, \dots, z_n^p)}{\det \Delta(z_1^p, \dots, z_n^p)}$$

where the points obtained by the Mean Value Theorem now depend on p and so are labeled z_i^p . The determinants are still continuous on D and so

$$\lim_{p \rightarrow \infty} g_{i,y_k}(\mathbf{Y}_p) = \lim_{p \rightarrow \infty} \frac{\det \Delta_i(z_1^p, \dots, z_n^p)}{\det \Delta(z_1^p, \dots, z_n^p)}$$

When we apply the Mean Value Theorem, we find $\mathbf{w}_i^p \in [g(\mathbf{Y}_p), g(\mathbf{Y}_p + h\mathbf{E}_k)]$. As $p \rightarrow \infty$, $\mathbf{w}_i^p \rightarrow \mathbf{w}_i \in [g(\mathbf{y}), g(\mathbf{y} + h\mathbf{E}_k)]$ since g is continuous. Hence

$$\lim_{p \rightarrow \infty} \frac{\det \Delta_i(z_1^p, \dots, z_n^p)}{\det \Delta(z_1^p, \dots, z_n^p)} = \frac{\det \Delta_i(z_1, \dots, z_n)}{\det \Delta(z_1, \dots, z_n)} = g_{i,y_k}(\mathbf{y})$$

which tells us g_{i,y_k} is continuous. Note how useful Cramer’s Rule is here. Cramer’s rule is not very good computationally for $n > 2$ but it is a powerful theoretical tool for the study of smoothness of mappings. \square

This completes the proof. \blacksquare

Example 13.2.1 For $(x_1, x_2) = \mathbf{x} \in \mathbb{R}^2$, let $\mathbf{y} = (y_1, y_2) = (x_1^2 - x_2^2, 2x_1, x_2)$. This defined a map F . Then

$$\mathbf{J}_F(x_1, x_2) = \begin{bmatrix} 2x_1 & -2x_2 \\ 2x_2 & 2x_1 \end{bmatrix}$$

and $\det \mathbf{J}_F(x_1, x_2) = 4x_1^2 + 4x_2^2$. This is positive unless $(x_1, x_2) = (0, 0)$. Hence, at any $\mathbf{p} = (p_1, p_2) \neq \mathbf{0}$, we can invoke the Inverse Function Theorem to find an open set U containing \mathbf{p} and an open set V containing $F(\mathbf{p})$ so that $f : U \rightarrow V$ is 1–1 and onto and $f^{-1} : V \rightarrow U$ is also 1–1 and onto. Further f^{-1} has continuous partials on V . If you look back at Example 13.1.2, we figured out all of this for a slightly different map $G(x_1, x_2) = (x_1^2 + x_2^2, 2x_1x_2)$ in great detail. We could have done that here. Of course, these kinds of details are really needed if you want to know the explicit functional form of these mappings.

13.2.1 Homework

Exercise 13.2.1

Exercise 13.2.2

Exercise 13.2.3

Exercise 13.2.4

Exercise 13.2.5

13.3 Implicit Function Results

To motivate the implicit function theorem, we want to go over a long example which will illustrate the general approach to the problem. We often want to handle this kind of problem.

Example 13.3.1 We know $x^2 + y^2 + z^2 = 10$. Can we solve for z in terms of x and y ? Let $F(x, y, z) = x^2 + y^2 + z^2 - 10$. Then $f(x, y, z) = 0$. We know $z = \pm\sqrt{10 - x^2 - y^2}$ and for real solutions, we must have $\sqrt{x^2 + y^2} \leq \sqrt{10}$. $F_z(x, y, z) = 2z$ which is not zero as long as $z \neq 0$. Since $x^2 + y^2 = 10 - z^2$, we can’t pick just any $z_0 \neq 0$. We must choose $-\sqrt{10} \leq z_0 \leq \sqrt{10}$. So if $0 < z_0 < \sqrt{10}$, then the set S of viable choices of (x, y) is $S = \{(x, y) | x^2 + y^2 = 10 - z_0^2\}$ which is a circle about the origin in \mathbb{R}^2 of radius $\sqrt{10 - z_0^2}$. For concreteness, pick a point x_0, y_0 in Quadrant One with $x_0^2 + y_0^2 + z_0^2 - 10 = 0$. Then in the open set $V_0 = \{(x, y) : x^2 + y^2 < 10 - z_0^2\}$, we can solve for z in terms of (x, y) as $z = \sqrt{10 - x^2 - y^2}$. This defines a mapping $G(x, y) = \sqrt{10 - x^2 - y^2}$ which has continuous partials and $F(x, y, G(x, y)) = x^2 + y^2 + 10 - x^2 - y^2 - 10 = 0$. Finally, we have $G(x_0, y_0) = \sqrt{10 - x_0^2 - y_0^2} = z_0$. This is an example of finding a function defined implicitly by a constraint surface. Note at $z = 0$, $F_z(x_0, y_0, 0) = 0$ always and the constraint surface reduces to $x^2 + y^2 - 10 = 0$. For that constraint, there is no way to find z in terms of (x, y) . So it seems a restriction on $F_z \neq 0$ is important.

So this is the problem of given $x^2 + y^2 + z^2 = 10$, solve for z locally in terms of (x, y) . Note we could also ask to solve for x in terms of (y, z) or y in terms of (x, z) . We can do this explicitly here.

Example 13.3.2 The constraint surface is now $F(x, y, z) = \sin(x^2 + 2y^4z^2) + 5z^4 - 25 = 0$. This tells us $(24/5)^{0.25} < z < (26/5)^{0.25}$ or $1.48 < z < 1.51$. However there are no constraints on (x, y) at all. It is really hard to see how to solve for z in terms of (x, y) . But notice $F_z = 8y^4z \cos(x^2 + 2y^4z^2) + 20z^3$ which is not zero at many (x, y, z) . It turns out we can prove there is a way to solve for z in terms of (x, y) locally around a point (x_0, y_0) where $F(x_0, y_0, z_0) = 0$ with $F_z(x_0, y_0, z_0) \neq 0$. This result is called the implicit function theorem and it is our next task.

Example 13.3.3 Consider the following situation. We are given $U \subset \mathbb{R}^3$ which is an open set and a mapping $F(x, y, z)$ with domain U so that F has continuous first order partials in U . Assume there is a point (x_0, y_0, z_0) in U with $F(x_0, y_0, z_0) = 0$ with $F_z(x_0, y_0, z_0) \neq 0$. We claim we can find an open set V_0 containing (x_0, y_0) in \mathbb{R}^2 and a mapping G with continuous partials on V_0 so that $G(x_0, y_0) = z_0$ and $F(x, y, G(x, y)) = 0$ for all (x, y) in V_0 . This means the **constraint** surface $F(x, y, z) = 0$ can be used to provide a local solution for z in terms of x and y . This is something we do all the time as you can see from Example 13.3.1 and Example 13.3.2.

For any $(x, y, z) \in U$ define the projections $f_1(x, y, z) = x$, $f_2(x, y, z) = y$ and define $f_3(x, y, z) = F(x, y, z)$. Then letting $f = (f_1, f_2, f_3)$, we have

$$\mathbf{J}_f(x, y, z) = \begin{bmatrix} f_{1x} & f_{1y} & f_{1z} \\ f_{2x} & f_{2y} & f_{2z} \\ f_{3x} & f_{3y} & f_{3z} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ F_x & F_y & F_z \end{bmatrix}$$

Thus $\det \mathbf{J}_f(x, y, z) = F_z(x, y, z)$ and by our assumption $\det \mathbf{J}_f(x_0, y_0, z_0) = F_z(x_0, y_0, z_0) \neq 0$. Apply the Inverse Function Theorem to $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$. Then there is an open set $U_1 \subset U$ containing (x_0, y_0, z_0) , an open set $V_1 \subset f(U_1)$ containing $f(x_0, y_0, z_0)$ and $g = f^{-1}$ so that

$$\begin{aligned} f : U_1 &\rightarrow V_1, 1-1, \text{ onto} \\ g : V_1 &\rightarrow U_1, 1-1, \text{ onto} \end{aligned}$$

Let the components of $g = f^{-1}$ be denoted by (g_1, g_2, h) . Then, if $(X_1, X_2, X_3) \in V_1$,

$$g(X_1, Y_1, Z_1) = (g_1(X_1, Y_1, Z_1), g_2(X_1, Y_1, Z_1), h(X_1, Y_1, Z_1))$$

Since f maps U_1 to V_1 1-1 and onto, there exist a unique $(x, y, z) \in U_1$ with $f(x, y, z) = (X_1, Y_1, Z_1)$. But using the definition of f , this says $(x, y, F(x, y, z)) = (X_1, Y_1, Z_1)$. Thus, since $g = f^{-1}$,

$$\begin{aligned} (x, y, z) &= (g \circ f)(x, y, z) = g(f(x, y, z)) = g(X_1, Y_1, Z_1) \\ &= (g_1(X_1, Y_1, Z_1), g_2(X_1, Y_1, Z_1), h(X_1, Y_1, Z_1)) = (g_1(x_1, y_1, Z_1), g_2(x_1, y_1, Z_1), h(x_1, y_1, Z_1)) \end{aligned}$$

We see

$$x = g_1(x, y, Z_1), \quad y = g_2(x, y, Z_1), \quad z = h(x, y, Z_1)$$

and recall (x, y, z) is the unique point satisfying $f(x, y, z) = (X_1, Y_1, Z_1)$. Since $x = X_1$ and $y = Y_1$, this can be rewritten as $f(x, y, z) = (x, y, Z_1)$.

Let

$$V_0 = \{(x, y) \in \mathbb{R}^2 \mid (x, y, 0) \in V_1\}$$

We claim V_0 is an open set in \mathbb{R}^2 since V_1 is open in \mathbb{R}^3 . Let $(x, y) \in V_0$, Then $(x, y, 0) \in V_1$ and since V_1 is open, $(x, y, 0)$ is an interior point. So there is a radius $r > 0$ so that $B(r, (x, y, 0)) \subset V_1$.

If $(X, Y) \in B(r, x, y)$, then $\sqrt{(X - x)^2 + (Y - y)^2} < r$ which implies $(X, Y, 0) \in B(r, x, y, 0) \subset V_1$. Thus, $(X, Y, 0) \in V_1$ telling us $(X, Y) \in V_0$. We conclude $B(r, (x, y)) \subset V_0$ which says (x, y) is an interior point of V_0 . Thus, V_0 is open in \mathbb{R}^2 .

All of this works for any choice of (X_1, Y_1, Z_1) . In particular, if we have $Z_1 = 0$, we get the unique point z so that $f(x, y, z) = (x, y, 0)$. But that means $z = h(x, y, 0)$. Define the mapping G by $G(x, y) = h(x, y, 0) = z$. Now $g = f^{-1}$ has continuous partials in V_1 and since h is a component of g , h also has continuous partials in V_1 . We conclude G also have continuous partials on V_0 as $V_0 \subset V_1$.

Next, consider $G(x_0, y_0) = h(x_0, y_0, 0)$. By construction, $h(x_0, y_0, 0) = w$ where w is the unique point satisfying $f(x_0, y_0, w) = (x_0, y_0, 0)$. But $f(x_0, y_0, w) = (x_0, y_0, F(x_0, y_0, w))$, so we see $F(x_0, y_0, w) = 0$. There is only one point which has this property, so w must be z_0 . Finally, if $(x, y) \in V_0$, then $(x, y, 0) \in V_1$ and $h(x, y, 0) = z$ where (x, y, z) is the unique point with $f(x, y, z) = (x, y, 0)$. So

$$(x, y, z) = g \circ f(x, y, z) = g(f(x, y, z)) = g(x, y, 0) = (g_1(x, y, 0), g_2(x, y, 0), h(x, y, 0))$$

So $x = g_1(x, y, 0)$, $y = g_2(x, y, 0)$ and $z = h(x, y, 0)$. Thus, $g(x, y, 0) = (x, y, h(x, y, 0)) = (x, y, G(x, y))$. Note $G(x, y) = z$ is the unique point so that $(x, y, z) \in U_1$ and $f(x, y, z) = (x, y, 0)$. We also know that $f(x, y, G(x, y)) = (x, y, F(x, y, G(x, y))) = (x, y, 0)$. Thus, $F(x, y, G(x, y)) = 0$ for $(x, y) \in U_0$.

So you can see how we might go about proving a more general theorem. We will do two cases: first, we extend this argument of a scalar function of $n + 1$ variables and solve for the $(n + 1)^{st}$ variable in terms of the others. This is Theorem 13.3.1. Then, we extend to vector valued functions. We think of the domain as \mathbb{R}^{n+m} and we solve for the variables x_{n+1}, \dots, x_{n+m} in terms of x_1, \dots, x_n . This is Theorem 13.3.2. We have been trying to work through this slowly so that you can follow the arguments below. It is like the ones we have already done, but a bit more abstract.

Theorem 13.3.1 Implicit Function Theorem

Let $U \subset \mathbb{R}^{n+1}$ be an open set. Let $\mathbf{u} \in U$ be written as $\mathbf{u} = (\mathbb{X}, y)$ where $\mathbb{X} \in \mathbb{R}^n$. Assume $F : U \rightarrow \mathbb{R}$ has continuous first order partials in U and there is a point $(\mathbb{X}_0, y_0) \in U$ satisfying $F(\mathbb{X}_0, y_0) = 0$ with $F_y(\mathbb{X}_0, y_0) \neq 0$. Then there is an open set V_0 containing \mathbb{X}_0 and a function $G : V_0 \rightarrow \mathbb{R}$ with continuous first order partials so that $G(\mathbb{X}_0) = y_0$ and $F(\mathbb{X}, G(\mathbb{X})) = 0$ on V_0 .

Proof 13.3.1

The argument is very similar to the one we used in the example at the start of this section. Basically, we replace the use of (x_1, x_2) with \mathbb{X} and make some obvious notational changes. Let's get started.

Here we are using \mathbb{X} to denote the components (x_1, x_2, \dots, x_n) and we are setting the last component x_{n+1} to be the variable y . For any $(\mathbb{X}, y) \in U$ define the projections $f_i(\mathbb{X}, y) = x_i$ and define $f_{n+1}(\mathbb{X}, y) = F(\mathbb{X}, y)$. Then letting $f = (f_1, \dots, f_n, F)$, we have

$$\mathbf{J}_f(\mathbb{X}, y) = \begin{bmatrix} f_{1,x_1} & \dots & f_{1,x_n} & f_{1,y} \\ f_{2,x_1} & \dots & f_{2,x_n} & f_{2,y} \\ \vdots & \vdots & \vdots & \vdots \\ f_{n,x_1} & \dots & f_{n,x_n} & f_{n,y} \\ F_{x_1} & \dots & F_{x_n} & F_y \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & 0 \\ F_{x_1} & F_{x_2} & F_{x_3} & \dots & F_{x_n} & F_y & \end{bmatrix}$$

Thus $\det \mathbf{J}_F(\mathbb{X}, y) = F_y(\mathbb{X}, y)$ and by our assumption $\det \mathbf{J}_F(\mathbb{X}_0, y_0) = F_y(\mathbb{X}_0, y_0) \neq 0$. Apply the Inverse Function Theorem to $f : \mathfrak{R}^{n+1} \rightarrow \mathfrak{R}^{n+1}$. Then there is an open set $U_1 \subset U$ containing (\mathbb{X}_0, y_0) , an open set $V_1 \subset f(U_1)$ containing $f(\mathbb{X}_0, y_0)$ and $g = f^{-1}$ so that

$$\begin{aligned} f : U_1 &\rightarrow V_1, 1-1, \text{ onto} \\ g : V_1 &\rightarrow U_1, 1-1, \text{ onto} \end{aligned}$$

Let the components of $g = f^{-1}$ be denoted by (g_1, \dots, g_n, h) . Then, if $(\mathbb{Z}, u) \in V_1$,

$$g(\mathbb{Z}, u) = (g_1(\mathbb{Z}, u), \dots, g_n(\mathbb{Z}, u), h(\mathbb{Z}, u))$$

Since f maps U_1 to V_1 1-1 and onto, there exist a unique $(\mathbb{X}, y) \in U_1$ with $f(\mathbb{X}, y) = (\mathbb{Z}, u)$. But using the definition of f , this says $(\mathbb{X}, F(\mathbb{X}, y)) = (\mathbb{Z}, u)$. Thus, since $g = f^{-1}$, $\mathbb{X} = \mathbb{Z}$ and $F(\mathbb{X}, y) = u$. Further,

$$\begin{aligned} (\mathbb{X}, y) &= (g \circ f)(\mathbb{X}, y) = g(f(\mathbb{X}, y)) = g(\mathbb{Z}, u) = g(\mathbb{X}, u) \\ &= (g_1(\mathbb{X}, u), \dots, g_n(\mathbb{X}, u), h(\mathbb{X}, u)) \end{aligned}$$

We see $x_i = g_i(\mathbb{X}, u)$. Now recall (\mathbb{X}, y) is the unique point satisfying $f(\mathbb{X}, y) = (\mathbb{Z}, u)$. Since $\mathbb{X} = \mathbb{Z}$, this can be rewritten as $f(\mathbb{X}, y) = (\mathbb{X}, u)$.

Let

$$V_0 = \{\mathbb{X} \in \mathfrak{R}^n | (\mathbb{X}, 0) \in V_1\}$$

Using an argument like before, it is straightforward to show V_0 is an open set in \mathfrak{R}^n since V_1 is open in \mathfrak{R}^{n+1} .

All of this works for any choice of (\mathbb{X}, y) . In particular, if we have $y = 0$, we get the unique point u so that $f(\mathbb{X}, u) = (\mathbb{X}, 0)$. But that means $y = h(\mathbb{X}, 0)$. Define the mapping G by $G(\mathbb{X}, y) = h(\mathbb{X}, 0) = y$. Now $g = f^{-1}$ has continuous partials in V_1 and since h is a component of g , h also has continuous partials in V_1 . We conclude G also have continuous partials on V_0 as $V_0 \subset V_1$.

Next, consider $G(\mathbb{X}_0) = h(\mathbb{X}_0, 0)$. By construction, $h(\mathbb{X}_0, 0) = w$ where w is the unique point satisfying $f(\mathbb{X}_0, w) = (\mathbb{X}_0, 0)$. But $f(\mathbb{X}_0, w) = (\mathbb{X}_0, F(\mathbb{X}_0, w))$, so we see $F(\mathbb{X}_0, w) = 0$. There is only one point which has this property so w must be y_0 . Finally, if $(\mathbb{X}) \in V_0$, then $(\mathbb{X}, 0) \in V_1$ and $h(\mathbb{X}, 0) = y$ where (\mathbb{X}, y) is the unique point with $f(\mathbb{X}, y) = (\mathbb{X}, 0)$. So

$$(\mathbb{X}, y) = g \circ f(\mathbb{X}, y) = g(f(\mathbb{X}, y)) = g(\mathbb{X}, 0) = (g_1(\mathbb{X}, 0), \dots, g_n(\mathbb{X}, 0), h(\mathbb{X}, 0))$$

So $x_i = g_i(\mathbb{X}, 0)$ and $y = h(\mathbb{X}, 0)$. Thus, $g(\mathbb{X}, 0) = (\mathbb{X}, h(\mathbb{X}, 0)) = (\mathbb{X}, G(\mathbb{X}))$. Note $G(\mathbb{X}) = y$ is the unique point so that $(\mathbb{X}, y) \in U_1$ and $f(\mathbb{X}, y) = (\mathbb{X}, 0)$. We also know that $f(\mathbb{X}, G(\mathbb{X})) = (\mathbb{X}, F(\mathbb{X}, G(\mathbb{X}))) = (\mathbb{X}, 0)$. Thus, $F(\mathbb{X}, G(\mathbb{X})) = 0$ for $\mathbb{X} \in U_0$. ■

Example 13.3.4 Let $f(x, y, z) = \tanh(xyz) - xyz - 10 = 0$. Letting $u = xyz$, this is equivalent to $\tanh(u) - u - 10 = 0$ or $\tanh(u) = u + 10$. It is easy to see there is a solution by graphing $\tanh(u)$ and $u + 10$ and finding the intersection. So there is a value u^* where $\tanh(u^*) = u^* + 10$. In particular, for $z_0 = 1$, there is a value $x_0 y_0$ so that $\tanh(x_0 y_0) - x_0 y_0 - 10 = 0$; i.e. $f(x_0, z=0, 1) = 0$. Note

$$F_z(x, y, z) = \operatorname{sech}^2(xyz)xy - xy = xy(\operatorname{sech}^2(xyz) - 1)$$

From the graph where we found the intersection (you should sketch this yourself!), we can see

$x_0 y_0 \neq 0$ so $\operatorname{sech} 62(x_0 y_0) - 1 \neq 0$. We conclude $F_x(x_0, y_0, 1) \neq 0$.

Applying the Implicit Function Theorem, there is an open set $U \in \mathbb{R}^2$ containing $(x_0, y_0, 1)$ and a mapping G with continuous partials in U so that $G(x_0, y_0) = 1$ and

$$\tanh(xy G(x, y)) - xy G(x, y) - 10 = 0$$

for all $(x, y) \in U$.

Theorem 13.3.2 An Extension of the Implicit Function Theorem

Let $U \subset \mathbb{R}^{n+m}$ be an open set. Let $\mathbf{u} \in U$ be written as $\mathbf{u} = (\mathbb{X}, \mathbb{Y})$ where $\mathbb{X} \in \mathbb{R}^n$ and $\mathbb{Y} \in \mathbb{R}^m$. Assume $F : U \rightarrow \mathbb{R}^m$ has continuous first order partials in U and there is a point $(\mathbb{X}_0, \mathbb{Y}_0) \in U$ satisfying $F(\mathbb{X}_0, \mathbb{Y}_0) = 0$ with

$$\det \begin{bmatrix} F_{n+1,y_1}(\mathbb{X}_0, \mathbb{Y}_0) & \dots & F_{n+1,y_m}(\mathbb{X}_0, \mathbb{Y}_0) \\ \vdots & \vdots & \vdots \\ F_{n+m,y_1}(\mathbb{X}_0, \mathbb{Y}_0) & \dots & F_{n+m,y_m}(\mathbb{X}_0, \mathbb{Y}_0) \end{bmatrix} \neq 0$$

where the matrix above is a submatrix of the full \mathbf{J}_F . Then there is an open set V_0 containing $\mathbb{X}_0 \in \mathbb{R}^n$ and $G : V_0 \rightarrow \mathbb{R}^m$ with continuous partials so that $G(\mathbb{X}_0) = \mathbb{Y}_0$ and $F(\mathbb{X}, G(\mathbb{X})) = \mathbb{O}$ where \mathbb{O} is the zero vector in \mathbb{R}^m .

Proof 13.3.2

The argument is again very similar to the one we just used. This time instead of a single scalar variable y , we have a vector \mathbb{Y} . It is mostly a matter of notation. If you compare these proofs carefully, we mostly replace y by \mathbb{Y} , u by \mathbb{U} and so forth as part played by the variable x_{n+1} has now been expanded to m variables and hence we must use vector notation. However, the structure of the argument is very much the same.

Here we are using \mathbb{X} to denote the components (x_1, x_2, \dots, x_n) and we are setting the components x_{n+1}, \dots, x_{n+m} to be the variable \mathbb{Y} for the new variables y_1, \dots, y_m . For any $(\mathbb{X}, \mathbb{Y}) \in U$, define $f(\mathbb{X}, \mathbb{Y}) = (\mathbb{X}, F(\mathbb{X}, \mathbb{Y}))$. Thus, the first n components are simply the usual projections. Then, since F has m components F_1 through F_m , we have

$$\mathbf{J}_f(\mathbb{X}, \mathbb{Y}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ F_{1,x_1} & \dots & F_{1,x_n} & F_{1,y_1} & F_{1,y_2} & \dots & F_{1,y_m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ F_{m,x_1} & \dots & F_{m,x_n} & F_{m,y_1} & F_{m,y_2} & \dots & F_{m,y_m} \end{bmatrix}$$

where all of the partials are evaluated at (\mathbb{X}, \mathbb{Y}) although we have not put that evaluation point in in order to save space. Thus, at $(\mathbb{X}_0, \mathbb{Y}_0)$

$$\det \mathbf{J}_F(\mathbb{X}_0, \mathbb{Y}_0) = \det \begin{bmatrix} F_{1,y_1}(\mathbb{X}, \mathbb{Y}_0) & \dots & F_{1,y_m}(\mathbb{X}, \mathbb{Y}_0) \\ \vdots & \vdots & \vdots \\ F_{m,y_1}(\mathbb{X}, \mathbb{Y}_0) & \dots & F_{m,y_m}(\mathbb{X}, \mathbb{Y}_0) \end{bmatrix} \neq 0$$

by our assumption. Apply the Inverse Function Theorem to $f : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^{n+m}$. Then there is an open set $U_1 \subset U$ containing $(\mathbb{X}_0, \mathbb{Y}_0)$, an open set $V_1 \subset f(U_1)$ containing $f(\mathbb{X}_0, \mathbb{Y}_0)$ and $g = f^{-1}$

so that

$$\begin{aligned} f : U_1 &\rightarrow V_1, 1-1, \text{ onto} \\ g : V_1 &\rightarrow U_1, 1-1, \text{ onto} \end{aligned}$$

Let the components of $g = f^{-1}$ be denoted by $(g_1, \dots, g_n, h_1, \dots, h_m)$. Then, if $(\mathbb{Z}, \mathbb{U}) \in V_1$,

$$\begin{aligned} g(\mathbb{Z}, \mathbb{U}) &= (g_1(\mathbb{Z}, \mathbb{U}), \dots, g_n(\mathbb{Z}, \mathbb{U}), \\ &\quad h_1(\mathbb{Z}, \mathbb{U}), \dots, h_1(\mathbb{Z}, \mathbb{U})) \end{aligned}$$

Since f maps U_1 to V_1 1-1 and onto, there exist a unique $(\mathbb{X}, \mathbb{Y}) \in U_1$ with $f(\mathbb{X}, \mathbb{Y}) = (\mathbb{Z}, \mathbb{U})$. But using the definition of f , this says $(\mathbb{X}, F(\mathbb{X}, \mathbb{Y})) = (\mathbb{Z}, \mathbb{U})$. Thus, since $g = f^{-1}$, $\mathbb{X} = \mathbb{Z}$ and $F(\mathbb{X}, \mathbb{Y}) = \mathbb{U}$. Further,

$$\begin{aligned} (\mathbb{X}, \mathbb{Y}) &= (g \circ f)(\mathbb{X}, \mathbb{Y}) = g(f(\mathbb{X}, \mathbb{Y})) = g(\mathbb{Z}, \mathbb{U}) = g(\mathbb{X}, \mathbb{U}) \\ &= (g_1(\mathbb{X}, \mathbb{U}), \dots, g_n(\mathbb{X}, \mathbb{U}), \\ &\quad h_1(\mathbb{X}, \mathbb{U}), \dots, h_m(\mathbb{X}, \mathbb{U})) \end{aligned}$$

We see $x_i = g_i(\mathbb{X}, \mathbb{U})$. Now recall (\mathbb{X}, \mathbb{Y}) is the unique point satisfying $f(\mathbb{X}, \mathbb{Y}) = (\mathbb{Z}, \mathbb{U})$. Since $\mathbb{X} = \mathbb{Z}$, this can be rewritten as $f(\mathbb{X}, \mathbb{Y}) = (\mathbb{X}, \mathbb{U})$.

Let

$$V_0 = \{\mathbb{X} \in \mathfrak{R}^n | (\mathbb{X}, \mathbb{O}) \in V_1\}$$

Using an argument like before, it is straightforward to show V_0 is an open set in \mathfrak{R}^n since V_1 is open in \mathfrak{R}^{n+m} .

All of this works for any choice of (\mathbb{X}, \mathbb{Y}) . In particular, if we have $\mathbb{Y} = \mathbb{O}$, we get the unique point (\mathbb{U}) so that $f(\mathbb{X}, \mathbb{U}) = (\mathbb{X}, \mathbb{O})$. But that means $\mathbb{Y} = (h_1(\mathbb{X}, \mathbb{O}), \dots, h_m(\mathbb{X}, \mathbb{O}))$. For convenience, let

$$h(\mathbb{X}, \mathbb{Y}) = (h_1(\mathbb{X}, \mathbb{O}), \dots, h_m(\mathbb{X}, \mathbb{O}))$$

Then define the mapping G by $G(\mathbb{X}, \mathbb{Y}) = h(\mathbb{X}, \mathbb{O}) = \mathbb{Y}$. Now $g = f^{-1}$ has continuous partials in V_1 and since h_i are components of g , h also has continuous partials in V_1 . We conclude G also have continuous partials on V_0 as $V_0 \subset V_1$.

Next, consider $G(\mathbb{X}_0) = h(\mathbb{X}_0, \mathbb{O})$. By construction, $h(\mathbb{X}_0, \mathbb{O}) = \mathbb{W}$ where \mathbb{W} is the unique point satisfying $f(\mathbb{X}_0, \mathbb{W}) = (\mathbb{X}_0, \mathbb{O})$. But $f(\mathbb{X}_0, \mathbb{W}) = (\mathbb{X}_0, F(\mathbb{X}_0, \mathbb{W}))$, so we see $F(\mathbb{X}_0, \mathbb{W}) = \mathbb{O}$. There is only one point which has this property so \mathbb{W} must be \mathbb{Y}_0 . Finally, if $(\mathbb{X}) \in V_0$, then $(\mathbb{X}, \mathbb{O}) \in V_1$ and $h(\mathbb{X}, \mathbb{O}) = \mathbb{Y}$ where (\mathbb{X}, \mathbb{Y}) is the unique point with $f(\mathbb{X}, \mathbb{Y}) = (\mathbb{X}, \mathbb{O})$. So

$$(\mathbb{X}, \mathbb{Y}) = g \circ f(\mathbb{X}, \mathbb{Y}) = g(f(\mathbb{X}, \mathbb{Y})) = g(\mathbb{X}, \mathbb{O}) = (g_1(\mathbb{X}, \mathbb{O}), \dots, g_n(\mathbb{X}, \mathbb{O}), h(\mathbb{X}, \mathbb{O}))$$

So $x_i = g_i(\mathbb{X}, \mathbb{O})$ and $\mathbb{Y} = h(\mathbb{X}, \mathbb{O})$. Thus, $g(\mathbb{X}, \mathbb{O}) = (\mathbb{X}, h(\mathbb{X}, \mathbb{O})) = (\mathbb{X}, G(\mathbb{X}))$. Note $G(\mathbb{X}) = \mathbb{Y}$ is the unique point so that $(\mathbb{X}, \mathbb{Y}) \in U_1$ and $f(\mathbb{X}, \mathbb{Y}) = (\mathbb{X}, \mathbb{O})$. We also know that $f(\mathbb{X}, G(\mathbb{X})) = (\mathbb{X}, F(\mathbb{X}, G(\mathbb{X}))) = (\mathbb{X}, \mathbb{O})$. Thus, $F(\mathbb{X}, G(\mathbb{X})) = \mathbb{O}$ for $\mathbb{X} \in U_0$. ■

Example 13.3.5 Let $F(x, y, u, v) = (u^2 + 2xu - v^2 - 7y, u^3 + 3\sin(u) + 15x^2e^y + \cos(v) + 10)$. Then

$$\mathbf{J}_F(x, y, u, v) = \begin{bmatrix} 2u & -7 & 2u + 2x & -2v \\ 30xe^y & 15x^2e^y & 3u^2 + 3\cos(u) & -\sin(v) \end{bmatrix}$$

At $(0, 0, 0, 0)$,

$$\mathbf{J}_F(0, 0, 0, 0) = \begin{bmatrix} 0 & -7 & 0 & 0 \\ 0 & 0 & 3 & -0 \end{bmatrix}$$

and $F(0, 0, 0, 0) = (0, 11)$. The submatrix

$$A = \begin{bmatrix} -7 & 0 \\ 0 & 3 \end{bmatrix}$$

has a nonzero determinant. Hence, the Implicit Function Theorem, Theorem 13.3.2, tells us we can solve for (y, u) in terms of (x, v) to find $G(x, v)$ so that by relabeling the order of the variables, $F(x, v, G(x, v)) = 11$ in an open set U containing $(0, 0)$ with $F(0, 0, 0, 0) = (0, 11)$.

13.3.1 Homework

Exercise 13.3.1

Exercise 13.3.2

Exercise 13.3.3

Exercise 13.3.4

Exercise 13.3.5

13.4 Constrained Optimization

Let's look at the problem of finding the minimum or maximum of a function $f(x, y)$ subject to the constraint $g(x, y) = c$ where c is a constant. Let's see how far we can get without explicitly using the implicit function theorem. Let's suppose we have a point (x_0, y_0) where an extremum value occurs and assume f and g are differentiable at that point. Then, at another point (x, y) which satisfies the constraint, we have

$$g(x, y) = g(x_0, y_0) + g_x(x_0, y_0)(x - x_0) + g_y(x_0, y_0)(y - y_0) + E_g(x, y, x_0, y_0)$$

But $g(x, y) = g(x_0, y_0) = c$, so we have

$$0 = g_x^0(x - x_0) + g_y^0(y - y_0) + E_g(x, y, x_0, y_0)$$

where we let $g_x^0 = g_x(x_0, y_0)$ and $g_y^0 = g_y(x_0, y_0)$.

Now given an x , we assume we can find a y values so that $g(x, y) = c$ in $B_r(x)$ for some $r > 0$. Of course, this value need not be unique. Let $\phi(x)$ be the function defined by

$$\phi(x) = \{y | g(x, y) = c\}$$

For example, if $g(x, y) = c$ was the function $x^2 + y^2 = 25$, then $\phi(x) = \pm\sqrt{25 - x^2}$ for $-5 \leq x \leq 5$. Clearly, we can't find a full circle $B_r(x)$ when $x = -5$ or $x = 5$, so let's assume the

point (x_0, y_0) where the extremum value occurs does have such a local circle around it where we can find corresponding y values for all x values in $B_r(x_0)$. So locally, we have a function $\phi(x)$ defined around x_0 so that $g(x, \phi(x)) = c$ for $x \in B_r(x_0)$. Then we have

$$0 = g_x^0(x - x_0) + g_y^0(\phi(x) - \phi(x_0)) + E_g(x, \phi(x), x_0, \phi(x_0))$$

Now divide through by $x - x_0$ to get

$$0 = g_x^0 + g_y^0 \left(\frac{\phi(x) - \phi(x_0)}{x - x_0} \right) + \frac{E_g(x, \phi(x), x_0, \phi(x_0))}{x - x_0}$$

Thus,

$$g_y^0 \left(\frac{\phi(x) - \phi(x_0)}{x - x_0} \right) = -g_x^0 - \frac{E_g(x, \phi(x), x_0, \phi(x_0))}{x - x_0}$$

and assuming $g_x^0 \neq 0$ and $g_y^0 \neq 0$, we can solve to find

$$\frac{\phi(x) - \phi(x_0)}{x - x_0} = -\frac{g_x^0}{g_y^0} - \frac{1}{g_y^0} \frac{E_g(x, \phi(x), x_0, \phi(x_0))}{x - x_0}$$

Since g is differentiable at $(x_0, \phi(x_0))$, as $x \rightarrow x_0$, $\frac{E_g(x, \phi(x), x_0, \phi(x_0))}{x - x_0} \rightarrow 0$. Thus, we have ϕ is differentiable and therefore also continuous with

$$\phi'(x_0) = \lim_{x \rightarrow x_0} \frac{\phi(x) - \phi(x_0)}{x - x_0} = -\frac{g_x^0}{g_y^0}$$

Now since both $g_x^0 \neq 0$ and $g_y^0 \neq 0$, the fraction $-\frac{g_x^0}{g_y^0} \neq 0$ too. This means locally $\phi(x)$ is either strictly increasing or strictly decreasing; i.e. is a strictly monotone function.

Let's assume $\phi'(x_0) > 0$ and so ϕ is increasing on some interval $(x_0 - R, x_0 + R)$. Let's also assume the extreme value is a minimum, so we know on this interval $f(x, y) \geq f(x_0, y_0)$ with $g(x, y) = c$. This means $f(x, \phi(x)) - f(x_0, \phi(x_0)) \geq 0$ on $(x_0 - R, x_0 + R)$. Now do an expansion for f to get

$$f(x, \phi(x)) = f(x_0, \phi(x_0)) + f_x^0(x - x_0) + f_y^0(\phi(x) - \phi(x_0)) + E_f(x, \phi(x), x_0, \phi(x_0))$$

where $f_x^0 = f_x(x_0, \phi(x_0))$ and $f_y^0 = f_y(x_0, \phi(x_0))$. This implies

$$f(x, \phi(x)) - f(x_0, \phi(x_0)) = f_x^0(x - x_0) + f_y^0(\phi(x) - \phi(x_0)) + E_f(x, \phi(x), x_0, \phi(x_0))$$

Now we have assumed $f(x, \phi(x)) - f(x_0, \phi(x_0)) \geq 0$, so we have

$$f_x^0(x - x_0) + f_y^0(\phi(x) - \phi(x_0)) + E_f(x, \phi(x), x_0, \phi(x_0)) \geq 0$$

If $x - x_0 > 0$ we find

$$f_x^0 + f_y^0 \left(\frac{\phi(x) - \phi(x_0)}{x - x_0} \right) + \frac{E_f(x, \phi(x), x_0, \phi(x_0))}{x - x_0} \geq 0$$

Now take the limit as $x \rightarrow x_0^+$ to find

$$f_x^0 + f_y^0 \phi'(x_0) \geq 0$$

When $x - x_0 < 0$, we argue similarly to find

$$f_x^0 + f_y^0 \phi'(x_0) \leq 0$$

Combining, we have

$$0 \leq f_x^0 + f_y^0 \phi'(x_0) \leq 0$$

which tells us at this extreme value we have the equation $f_x^0 + f_y^0 \phi'(x_0) = 0$. This implies $f_y^0 = -\frac{f_x^0}{\phi'(x_0)}$. Next note

$$\begin{bmatrix} f_x^0 \\ f_y^0 \end{bmatrix} = \begin{bmatrix} f_x^0 \\ -\frac{f_x^0}{\phi'(x_0)} \end{bmatrix} = \left[-f_x^0 \left\{ -\left(\frac{g_y^0}{g_x^0} \right) \right\} \right] = \frac{f_x^0}{g_x^0} \begin{bmatrix} g_x^0 \\ g_y^0 \end{bmatrix}$$

This says there is a scalar $\lambda = \frac{f_x^0}{g_x^0}$ so that

$$\nabla(f)(x_0, y_0) = \lambda \nabla(g)(x_0, y_0)$$

where $g(x_0, y_0) = c$. The value $\lambda = \frac{f_x^0}{g_x^0}$ is called the **Lagrange Multiplier** for this extremal Problem. **This is the basis for the Lagrange Multiplier Technique for a constrained optimization problem.**

We can do a similar sort of analysis in the case the extremum is a maximum too. Our analysis assumes that the point (x_0, y_0) where the extremum occurs is like an interior point in the $\{(x, y) | g(x, y) = c\}$. That is, we assume for such an x_0 there is a interval $B_r(x_0)$ with any $x \in B_r(x_0)$ having a corresponding value $y = \phi(x)$ so that $g(x, \phi(x)) = c$. So the argument does not handle boundary points such as the ± 5 is our previous example.

To implement the Lagrange Multiplier technique, we define a new function

$$H(x, y, \lambda) = f(x, y) + \lambda(g(x, y) - c)$$

The critical points we see are the ones where the gradient of H is a multiple of the gradient of g for the reasons we discussed above. Note the λ we use here is the negative of the Lagrange Multiplier. To find them, we set the partials of H equal to zero.

$$\begin{aligned} \frac{\partial H}{\partial x} &= 0 \implies \frac{\partial f}{\partial x} = -\lambda \frac{\partial g}{\partial x} \\ \frac{\partial H}{\partial y} &= 0 \implies \frac{\partial f}{\partial y} = -\lambda \frac{\partial g}{\partial x} \\ \frac{\partial H}{\partial \lambda} &= 0 \implies g(x, y) = c \end{aligned}$$

The first two lines are the statement that the gradient of f is a multiple of the gradient of g and the third line is the statement that the constraint must be satisfied. As usual, solving these three nonlinear equations gives the interior point solutions and we always must also include the boundary points that solve the constraints as well.

Example 13.4.1 Extremize $2x + 3y$ subject to $x^2 + y^2 = 25$.

Solution Set $H(x, y, \lambda) = (2x + 3y) + \lambda(x^2 + y^2 - 25)$. The critical point equations are then

$$\begin{aligned}\frac{\partial H}{\partial x} &= 2 + \lambda(2x) = 0 \\ \frac{\partial H}{\partial y} &= 3 + \lambda(2y) = 0 \\ \frac{\partial H}{\partial \lambda} &= x^2 + y^2 - 25 = 0\end{aligned}$$

Solving for λ , we find $\lambda = -1/x = -3/(2y)$. Thus, $x = 2y/3$.

Using this relationship in the constraint, we find $(2y/3)^2 + y^2 = 25$ or $13y^2/9 = 25$. Thus, $y^2 = 225/13$ or $y = \pm 15/\sqrt{13}$. This means $x = \pm 10/\sqrt{13}$.

We find

$$\begin{aligned}f(10/\sqrt{13}, 15/\sqrt{13}) &= 45/\sqrt{13} = 12.48 \\ f(-10/\sqrt{13}, -15/\sqrt{13}) &= -45/\sqrt{13} = -12.48 \\ f(-5, 0) &= -10 \\ f(5, 0) &= 10 \\ f(0, -5) &= -15 \\ f(0, 5) &= 15.\end{aligned}$$

So the extreme values here occur at the boundary points $(0, \pm 5)$.

Example 13.4.2 Extremize $\sin(xy)$ subject to $x^2 + y^2 = 1$.

Solution Set $H(x, y, \lambda) = (\sin(xy)) + \lambda(x^2 + y^2 - 1)$. The critical point equations are then

$$\begin{aligned}\frac{\partial H}{\partial x} &= y \cos(xy) + \lambda(2x) = 0 \\ \frac{\partial H}{\partial y} &= x \cos(xy) + \lambda(2y) = 0 \\ \frac{\partial H}{\partial \lambda} &= x^2 + y^2 - 1 = 0\end{aligned}$$

Solving for λ in the first equation, we find $\lambda = -y \cos(xy)/2x$. Now use this in the second equation to find $x \cos(xy) - 2y^2 \cos(xy)/2x = 0$. Thus, $\cos(xy)(x - y^2/x) = 0$.

The first case is $\cos(xy) = 0$. This implies $xy = \pi/2 + 2n\pi$. These are rectangular hyperbolae and the closest one to the origin is $xy = \pi/2 = 1.57$ whose closest point to the origin is $(\sqrt{\pi/2}, \sqrt{\pi/2}) = (1, 25, 1, 25)$. This point is outside of the constraint set, so this case does not matter.

The second case is $x^2 = y^2$ which, using the constraint, implies $x^2 = 1/2$. Thus there are four possibilities: $(1/\sqrt{2}, 1/\sqrt{2})$, $(1/\sqrt{2}, -1/\sqrt{2})$, $(-1/\sqrt{2}, 1/\sqrt{2})$ and $(-1/\sqrt{2}, -1/\sqrt{2})$.

The other possibilities are the circle's boundary points $(0, \pm 1)$ and $(\pm 1, 0)$. We find

$$\begin{aligned}\sin(1/\sqrt{2}, 1/\sqrt{2}) &= \sin(-1/\sqrt{2}, -1/\sqrt{2}) = \sin(1/2) = .48 \\ \sin(1/\sqrt{2}, -1/\sqrt{2}) &= \sin(-1/\sqrt{2}, 1/\sqrt{2}) = -\sin(1/2) = -.48 \\ \sin(\pm 1, 0) &= 0, \sin(0, \pm 1) = 0\end{aligned}$$

So the extreme values occur here at the interior points of the constraint.

13.4.1 What Does the Lagrange Multiplier Mean?

Let's change our extremum problem by assuming the constraint constant is changing as a function of x ; i.e. the problem is now find the extreme values of $f(x, y)$ subject to $g(x, y) = C(x)$. We can use our tangent plane analysis like before. We assume the extreme value for constraint value $C(x_0)$ occurs at an interior point of the constraint. We have

$$g(x, \phi(x)) - C(x) = 0$$

Then

$$0 = g_x^0 + g_y^0 \left(\frac{\phi(x) - \phi(x_0)}{x - x_0} \right) + \left(\frac{E_g(x, \phi(x), x_0, \phi(x_0))}{x - x_0} \right) - \left(\frac{C(x) - C(x_0)}{x - x_0} \right)$$

We assume the constraint bound function C is differentiable and so letting $x \rightarrow x_0$, we find

$$g_x^0 + g_y^0 \phi'(x_0) = C'(x_0)$$

Now assume $(x_1, \phi(x_1))$ extremizes f for the constraint bound value $C(x_1)$. We have

$$\begin{aligned} f(x_1, \phi(x_1)) - f(x_0, \phi(x_0)) &= f_x^0(x_1 - x_0) + f_y^0(\phi(x_1) - \phi(x_0)) \\ &\quad + E_f(x_1, \phi(x_1), x_0, \phi(x_0)) \end{aligned}$$

or

$$\begin{aligned} \left(\frac{f(x_1, \phi(x_1)) - f(x_0, \phi(x_0))}{x_1 - x_0} \right) &= f_x^0 + f_y^0 \left(\frac{\phi(x_1) - \phi(x_0)}{x_1 - x_0} \right) \\ &\quad + \left(\frac{E_f(x_1, \phi(x_1), x_0, \phi(x_0))}{x_1 - x_0} \right) \end{aligned}$$

Now assuming $C(x) \neq 0$ locally around x_0 , we can write

$$\begin{aligned} \left(\frac{f(x_1, \phi(x_1)) - f(x_0, \phi(x_0))}{C(x_1) - C(x_0)} \right) \left(\frac{C(x_1) - C(x_0)}{x_1 - x_0} \right) &= \\ f_x^0 + f_y^0 \left(\frac{\phi(x_1) - \phi(x_0)}{x_1 - x_0} \right) + \left(\frac{E_f(x_1, \phi(x_1), x_0, \phi(x_0))}{x_1 - x_0} \right) & \end{aligned}$$

Now let $\Theta(x)$ be defined by $\Theta(x) = f(x, \phi(x))$ when $(x, \phi(x))$ is the extreme value for the problem of extremizing $f(x, y)$ subject to $g(x, y) = C(x)$. This is called the **Optimal Value Function** for this problem. Rewriting, we have

$$\begin{aligned} \left(\frac{\Theta(x_1) - \Theta(x_0)}{C(x_1) - C(x_0)} \right) \left(\frac{C(x_1) - C(x_0)}{x_1 - x_0} \right) &= \\ f_x^0 + f_y^0 \left(\frac{\phi(x_1) - \phi(x_0)}{x_1 - x_0} \right) + \left(\frac{E_f(x_1, \phi(x_1), x_0, \phi(x_0))}{x_1 - x_0} \right) & \end{aligned}$$

Now let $x_1 \rightarrow x_0$. We obtain

$$\lim_{x_1 \rightarrow x_0} \left(\frac{\Theta(x_1) - \Theta(x_0)}{C(x_1) - C(x_0)} \right) C'(x_0) = f_x^0 + f_y^0 \phi'(x_0)$$

Thus, the rate of change of the **optimal value** with respect to change in the constraint bound, $\frac{d\Theta}{dC}$ is well - defined and

$$\frac{d\Theta}{dC}(C_0) C'(x_0) = f_x^0 + f_y^0 \phi'(x_0)$$

where C_0 is the value $C(x_0)$. Assuming $C'(x_0) \neq 0$, we have

$$\frac{d\Theta}{dC}(C_0) = \frac{f_x^0}{C'(x_0)} + \frac{f_y^0}{C'(x_0)} \phi'(x_0)$$

But we know at the extreme value for x_0 that $f_y^0 = \frac{g_y^0}{g_x^0} f_x^0$. Thus,

$$\begin{aligned} \frac{d\Theta}{dC}(C_0) &= \frac{f_x^0}{C'(x_0)} + \frac{g_y^0}{g_x^0} \frac{f_x^0}{C'(x_0)} \phi'(x_0) = \frac{f_x^0}{C'(x_0)} \left(1 + \frac{g_y^0}{g_x^0} \phi'(x_0) \right) \\ &= \frac{f_x^0}{g_x^0} \frac{1}{C'(x_0)} \left(g_x^0 + g_y^0 \phi'(x_0) \right) \end{aligned}$$

But the Lagrange Multiplier λ_0 for extremal value for x_0 is $\lambda_0 = \frac{f_x^0}{g_x^0}$. So

$$\frac{d\Theta}{dC}(C_0) = \lambda_0 \left(g_x^0 + g_y^0 \phi'(x_0) \right) \frac{1}{C'(x_0)}$$

But $C'(x_0) = g_x^0 + g_y^0 \phi'(x_0)$. Hence, we find $\frac{d\Theta}{dC}(C_0) = \lambda_0$.

We see from our argument that the Lagrange Multiplier λ_0 is the rate of change of optimal value with respect to the constraint bound. There are cases:

- $sign(\lambda_0) = sign(\frac{f_x^0}{g_x^0}) = \pm = +$ which implies the optimal value goes up if the constraint bound is perturbed.
- $sign(\lambda_0) = sign(\frac{f_x^0}{g_x^0}) = \mp = +$ which implies the optimal value goes up if the constrained bound is perturbed.
- $sign(\lambda_0) = sign(\frac{f_x^0}{g_x^0}) = \pm = -$ which implies the optimal value goes down if the constrained bound is perturbed.
- $sign(\lambda_0) = sign(\frac{f_x^0}{g_x^0}) = \mp = -$ which implies the optimal value goes down if the constrained bound is perturbed.

Hence, we can interpret the Lagrange Multiplier as a **price**: the change in optimal value due to a change in constraint is essentially the *loss of value* experienced due to the constraint change. For example, if $|\lambda_0|$ is very small, it says the rate of change of the optimal value with respect to constraint modification is small too. This implies the optimal value is insensitive to constraint modification.

13.4.2 Homework

Exercise 13.4.1 Find the extreme values of $\sin(xy)$ subject to $x^2 + 2y^2 = 1$ using the Lagrange Multiplier Technique.

Exercise 13.4.2 Find the extreme values of $\cos(xy)$ subject to $x^2 + 2y^2 = 1$ using the Lagrange Multiplier Technique.

Exercise 13.4.3 If $a_n \rightarrow 1$ and $b_n \rightarrow 3$, use an $\epsilon - N$ proof to show $5a_n/b_n \rightarrow 5/3$.

Exercise 13.4.4 If $a_n \rightarrow 10$ and $b_n \rightarrow 2$, use an $\epsilon - N$ proof to show $5a_nb_n \rightarrow 100$.

Chapter 14

Linear Approximation Applications

Let's look at how we use tangent plane approximations in models and computations.

14.1 Linear Approximations to Nonlinear ODE

Here, we look at several nonlinear ODE applications.

14.1.1 An Insulin Model

This is a model we have discussed in (Peterson (5) 2016) which was first presented in (Braun (2) 1978). This will show you how you can use the ideas of tangent plane approximations to build practical models on nonlinear phenomena.

In diabetes there is too much sugar in the blood and the urine. This is a metabolic disease and if a person has it, they are not able to use up all the sugars, starches and various carbohydrates because they don't have enough **insulin**. Diabetes can be diagnosed by a **glucose tolerance test** (GTT). If you are given this test, you do an overnight fast and then you are given a large dose of sugar in a form that appears in the bloodstream. This sugar is called **glucose**. Measurements are made over about five hours or so of the concentration of glucose in the blood. These measurements are then used in the diagnosis of diabetes. It has always been difficult to interpret these results as a means of diagnosing whether a person has diabetes or not. Hence, different physicians interpreting the same data can come up with a different diagnosis, which is a pretty unacceptable state of affairs!

In this chapter, we are going to discuss a criterion developed in the 1960's by doctors at the Mayo Clinic and the University of Minnesota that was fairly reliable. It showcases a lot of our modeling in this course and will give you another example of how we use our tools. We start with a simple model of the blood glucose regulatory system.

Glucose plays an important role in vertebrate metabolism because it is a source of energy. For each person, there is an optimal blood glucose concentration and large deviations from this leads to severe problems including death. Blood glucose levels are autoregulated via standard forward and backward interactions like we see in many biological systems. An example is the signal that is used to activate the creation of a protein which we discussed earlier. The signaling molecules are typically either bound to another molecule in the cell or are free. The equilibrium concentration of free signal is due to the fact that the rate at which signaling molecules bind equals the rate at which they split apart from their binding substrate. When an external message comes into the cell called a trigger, it induces a change in this careful balance which temporarily upgrades or degrades the equilibrium signal concentration. This then influence the protein concentration rate. Blood glucose concentrations work like this too, although the details differ. The blood glucose concentration is influenced by a variety of signaling molecules just like the protein creation rates can be. Here are

some of them. The hormone that **decreases** blood glucose concentration is **insulin**. **Insulin** is a hormone secreted by the β cells of the pancreas. After we eat carbohydrates, our gastrointestinal tract sends a signal to the pancreas to secrete insulin. Also, the glucose in our blood directly stimulates the β cells to secrete insulin. We think insulin helps cells pull in the glucose needed for metabolic activity by attaching itself to membrane walls that are normally impenetrable. This attachment increases the ability of glucose to pass through to the inside of the cell where it can be used as *fuel*. So, if there is not enough insulin, cells don't have enough energy for their needs. The other hormones we will focus on all tend to **change** blood glucose concentrations also.

- **Glucagon** is a hormone secreted by the α cells of the pancreas. Excess glucose is stored in the liver in the form of **Glycogen**. There is the usual equilibrium amount of storage caused by the rate of glycogen formation being equal to the rate of the reverse reaction that moves glycogen back to glucose. Hence the glycogen serves as a reservoir for glucose and when the body needs glucose, the rate balance is tipped towards conversion back to glucose to release needed glucose to the cells. The hormone **glucagon** increases the rate of the reaction that converts glycogen back to glucose and so serves an important regulatory function. **Hypoglycemia** (low blood sugar) and **fasting** tend to increase the secretion of the hormone glucagon. On the other hand, if the blood glucose levels increase, this tends to suppress glucagon secretion; i.e. we have another back and forth regulatory tool.
- **Epinephrine** also called **adrenalin** is a hormone secreted by the adrenal medulla. It is part of an emergency mechanism to quickly increase the blood glucose concentration in times of extremely low blood sugar levels. Hence, epinephrine also increases the rate at which glycogen converts to glucose. It also directly inhibits how much glucose is able to be pulled into muscle tissue because muscles use a lot of energy and this energy is needed elsewhere more urgently. It also acts on the pancreas directly to inhibit insulin production which keeps glucose in the blood. There is also another way to increase glucose by converting lactate into glucose in the liver. Epinephrine increases this rate also so the liver can pump this extra glucose back into the blood stream.
- **Glucocorticoids** are hormones like **cortisol** which are secreted by the adrenal cortex which influence how carbohydrates are metabolized which in turn increase glucose if the metabolic rate goes up.
- **Thyroxin** is a hormone secreted by the thyroid gland and it helps the liver form glucose from sources which are not carbohydrates such as glycerol, lactate and amino acids. So another way to up glucose!
- **Somatotrophin** is called the growth hormone and it is secreted by the anterior pituitary gland. This hormone directly affect blood glucose levels (i.e. an increase in Somatotrophin increases blood glucose levels and vice versa) but it also inhibits the effect of insulin on muscle and fat cell's permeability which diminishes insulin's ability to help those cells pull glucose out of the blood stream. These actions can therefore increase blood glucose levels.

Now net hormone concentration is the sum of insulin plus the others. Let H denote this net hormone concentration. At normal conditions, call this concentration H_0 . There have been studies performed that show that under close to normal conditions, the interaction of the one hormone **insulin** with blood glucose completely dominates the net hormonal activity. That is normal blood sugar levels primarily depend on insulin-glucose interactions.

So if insulin increases from normal levels, it increases net hormonal concentration to $H_0 + \Delta H$ and decreases glucose blood concentration. On the other hand, if other hormones such as cortisol increased from base levels, this will make blood glucose levels go up. Since insulin dominates all activity at normal conditions, we can think of this increase in cortisol as a decrease in insulin with

a resulting drop in blood glucose levels. A decrease in insulin from normal levels corresponds to a drop in net hormone concentration to $H_0 - \Delta H$. Now let G denote blood glucose level. Hence, in our model an increase in H means a drop in G and a decrease in H means an increase in G ! Note our lumping of all the hormone activity into a single net activity is very much like how we would model food fish and predator fish in the predator prey model.

Homework

Exercise 14.1.1

Exercise 14.1.2

Exercise 14.1.3

Exercise 14.1.4

Exercise 14.1.5

14.1.1.1 Model Details

The idea of our model for diagnosing diabetes from the GTT is to find a simple dynamical model of this complicated blood glucose regulatory system in which the values of two parameters would give a nice criterion for distinguishing normal individuals from those with mild diabetes or those who are pre diabetic. Here is what we will do. We describe the model as

$$\begin{aligned} G'(t) &= F_1(G, H) + J(t) \\ H'(t) &= F_2(G, H) \end{aligned}$$

where the function J is the external rate at which blood glucose concentration is being increased. There are two nonlinear interaction functions F_1 and F_2 because we know G and H have complicated interactions. Let's assume G and H have achieved optimal values G_0 and H_0 by the time the fasting patient has arrived at the hospital. Hence, we don't expect to have any contribution to $G'(0)$ and $H'(0)$; i.e. $F_1(G_0, H_0) = 0$ and $F_2(G_0, H_0) = 0$. We are interested in the deviation of G and H from their optimal values G_0 and H_0 , so let $g = G - G_0$ and $h = H - H_0$. We can then write $G = G_0 + g$ and $H = H_0 + h$. The model can then be rewritten as

$$\begin{aligned} (G_0 + g)'(t) &= F_1(G_0 + g, H_0 + h) + J(t) \\ (H_0 + h)'(t) &= F_2(G_0 + g, H_0 + h) \end{aligned}$$

or

$$\begin{aligned} g'(t) &= F_1(G_0 + g, H_0 + h) + J(t) \\ h'(t) &= F_2(G_0 + g, H_0 + h) \end{aligned}$$

Now find the tangent plane approximations.

$$\begin{aligned} F_1(G_0 + g, H_0 + h) &= F_1(G_0, H_0) + \frac{\partial F_1}{\partial g}(G_0, H_0) g + \frac{\partial F_1}{\partial h}(G_0, H_0) h + E_{F_1} \\ F_2(G_0 + g, H_0 + h) &= F_2(G_0, H_0) + \frac{\partial F_2}{\partial g}(G_0, H_0) g + \frac{\partial F_2}{\partial h}(G_0, H_0) h + E_{F_2} \end{aligned}$$

but the terms $F_1(\mathbf{G}_0, \mathbf{H}_0) = 0$ and $F_2(\mathbf{G}_0, \mathbf{H}_0) = 0$, so we can simplify to

$$\begin{aligned} F_1(\mathbf{G}_0 + \mathbf{g}, \mathbf{H}_0 + \mathbf{h}) &= \frac{\partial F_1}{\partial \mathbf{g}}(\mathbf{G}_0, \mathbf{H}_0) \mathbf{g} + \frac{\partial F_1}{\partial \mathbf{h}}(\mathbf{G}_0, \mathbf{H}_0) \mathbf{h} + E_{F_1} \\ F_2(\mathbf{G}_0 + \mathbf{g}, \mathbf{H}_0 + \mathbf{h}) &= \frac{\partial F_2}{\partial \mathbf{g}}(\mathbf{G}_0, \mathbf{H}_0) \mathbf{g} + \frac{\partial F_2}{\partial \mathbf{h}}(\mathbf{G}_0, \mathbf{H}_0) \mathbf{h} + E_{F_2} \end{aligned}$$

Homework

Exercise 14.1.6

Exercise 14.1.7

Exercise 14.1.8

Exercise 14.1.9

Exercise 14.1.10

It seems reasonable to assume that since we are so close to ordinary operating conditions, the errors E_{F_1} and E_{F_2} will be negligible. Thus our model approximation is

$$\begin{aligned} \mathbf{g}'(t) &= \frac{\partial F_1}{\partial \mathbf{g}}(\mathbf{G}_0, \mathbf{H}_0) \mathbf{g} + \frac{\partial F_1}{\partial \mathbf{h}}(\mathbf{G}_0, \mathbf{H}_0) \mathbf{h} + \mathbf{J}(t) \\ \mathbf{h}'(t) &= \frac{\partial F_2}{\partial \mathbf{g}}(\mathbf{G}_0, \mathbf{H}_0) \mathbf{g} + \frac{\partial F_2}{\partial \mathbf{h}}(\mathbf{G}_0, \mathbf{H}_0) \mathbf{h} \end{aligned}$$

We can reason out the algebraic signs of the four partial derivatives to be

$$\begin{aligned} \frac{\partial F_1}{\partial \mathbf{g}}(\mathbf{G}_0, \mathbf{H}_0) &= - \\ \frac{\partial F_1}{\partial \mathbf{h}}(\mathbf{G}_0, \mathbf{H}_0) &= - \\ \frac{\partial F_2}{\partial \mathbf{g}}(\mathbf{G}_0, \mathbf{H}_0) &= + \\ \frac{\partial F_2}{\partial \mathbf{h}}(\mathbf{G}_0, \mathbf{H}_0) &= - \end{aligned}$$

The arguments for these algebraic signs come from our understanding of the physiological processes that are going on here. Let's look at a small positive deviation from the optimal value \mathbf{G}_0 while letting the net hormone concentration be fixed at \mathbf{H}_0 . At this point, we are not adding an external input, so here $\mathbf{J}(t) = 0$. Then our model approximation is

$$\mathbf{g}'(t) = \frac{\partial F_1}{\partial \mathbf{g}}(\mathbf{G}_0, \mathbf{H}_0) \mathbf{g}$$

At a state where we have an increase in blood sugar levels over optimal, i.e. $\mathbf{g} > 0$, the other hormones such as cortisol and glucagon will try to regulate the blood sugar level down by increasing their concentrations and for example storing more sugar into glycogen. Hence, the term $\frac{\partial F_1}{\partial \mathbf{g}}(\mathbf{G}_0, \mathbf{H}_0)$ should be negative as here \mathbf{g}' is negative as \mathbf{g} should be decreasing. So we model this as $\frac{\partial F_1}{\partial \mathbf{g}}(\mathbf{G}_0, \mathbf{H}_0) = -m_1$ for some positive number m_1 . Now consider a positive change in \mathbf{h} from

the optimal level while keeping at the optimal level G_0 . Then the model is

$$g'(t) = \frac{\partial F_1}{\partial h}(G_0, H_0) h$$

and since $h > 0$, this means the net hormone concentration is up which we interpret as insulin above normal. This means blood sugar levels go down which implies g' is negative again. Thus, $\frac{\partial F_1}{\partial h}(G_0, H_0)$ must be negative which means we model it as $\frac{\partial F_1}{\partial h}(G_0, H_0) = -m_2$ for some positive m_2 .

Now look at the h' model in these two cases. If we have a small positive deviation g from the optimal value G_0 while letting the net hormone concentration be fixed at H_0 , we have

$$h'(t) = \frac{\partial F_2}{\partial g}(G_0, H_0) g.$$

Again, since g is positive, this means we are above normal blood sugar levels which implies mechanisms are activated to bring the level down. Hence $h' > 0$ as we have increasing net hormone levels. Thus, we must have $\frac{\partial F_2}{\partial g}(G_0, H_0) = m_3$ for some positive m_3 . Finally, if we have a positive deviation h from optimal while blood sugar levels are optimal, the model is

$$h'(t) = \frac{\partial F_2}{\partial h}(G_0, H_0) h.$$

Since h is positive, we have the concentrations of the hormones that pull glucose out of the blood stream are above optimal. This means that too much sugar is being removed as so the regulatory mechanisms will act to stop this action implying $h' < 0$. This tells us $\frac{\partial F_2}{\partial h}(G_0, H_0) = -m_4$ for some positive constant m_4 . Hence, the four partial derivatives at the optimal points can be defined by four positive numbers m_1, m_2, m_3 and m_4 as follows:

$$\begin{aligned} \frac{\partial F_1}{\partial g}(G_0, H_0) &= -m_1 \\ \frac{\partial F_1}{\partial h}(G_0, H_0) &= -m_2 \\ \frac{\partial F_2}{\partial g}(G_0, H_0) &= +m_3 \\ \frac{\partial F_2}{\partial h}(G_0, H_0) &= -m_4 \end{aligned}$$

Homework

Exercise 14.1.11

Exercise 14.1.12

Exercise 14.1.13

Exercise 14.1.14

Exercise 14.1.15

14.1.1.2 Model Analysis

Our model dynamics are thus approximated by

$$\begin{aligned}\mathbf{g}'(t) &= -m_1 \mathbf{g} - m_2 \mathbf{h} + \mathbf{J}(t) \\ \mathbf{h}'(t) &= m_3 \mathbf{g} - m_4 \mathbf{h}\end{aligned}$$

This implies

$$\mathbf{g}''(t) = -m_1 \mathbf{g}' - m_2 \mathbf{h}' + \mathbf{J}'(t)$$

Now plug in the formula for \mathbf{h}' to get

$$\begin{aligned}\mathbf{g}''(t) &= -m_1 \mathbf{g}' - m_2 (m_3 \mathbf{g} - m_4 \mathbf{h}) + \mathbf{J}'(t) \\ &= -m_1 \mathbf{g}' - m_2 m_3 \mathbf{g} + m_2 m_4 \mathbf{h} + \mathbf{J}'(t).\end{aligned}$$

But we can use the \mathbf{g}' equation to solve for \mathbf{h} . This gives

$$m_2 \mathbf{h} = -\mathbf{g}'(t) - m_1 \mathbf{g} + \mathbf{J}(t)$$

which leads to

$$\begin{aligned}\mathbf{g}''(t) &= -m_1 \mathbf{g}' - m_2 m_3 \mathbf{g} + m_4 (-\mathbf{g}'(t) - m_1 \mathbf{g} + \mathbf{J}(t)) + \mathbf{J}'(t) \\ &= -(m_1 + m_4) \mathbf{g}' - (m_1 m_4 + m_2 m_3) \mathbf{g} + m_4 \mathbf{J}(t) + \mathbf{J}'(t).\end{aligned}$$

So our final model is

$$\mathbf{g}''(t) + (m_1 + m_4) \mathbf{g}' + (m_1 m_4 + m_2 m_3) \mathbf{g} = m_4 \mathbf{J}(t) + \mathbf{J}'(t).$$

Let $\alpha = (m_1 + m_4)/2$ and $\omega^2 = m_1 m_4 + m_2 m_3$ and we can rewrite as

$$\mathbf{g}''(t) + 2\alpha \mathbf{g}' + \omega^2 \mathbf{g} = S(t).$$

where $S(t) = m_4 \mathbf{J}(t) + \mathbf{J}'(t)$. Now the right hand side here is zero except for the very short time interval when the glucose load is being ingested. Hence, we can simply search for the solution to the homogeneous model

$$\mathbf{g}''(t) + 2\alpha \mathbf{g}' + \omega^2 \mathbf{g} = 0.$$

The roots of the characteristic equation here are

$$r = \frac{-2\alpha \pm \sqrt{4\alpha^2 - 4\omega^2}}{2} = -\alpha \pm \sqrt{\alpha^2 - \omega^2}.$$

The most interesting case is if we have complex roots. In that case, $\alpha^2 - \omega^2 < 0$. Let $\Omega^2 = |\alpha^2 - \omega^2|$. Then, the general phase shifted solution has the form $\mathbf{g} = Re^{-\alpha t} \cos(\Omega t - \delta)$ which implies

$$\mathbf{G} = \mathbf{G}_0 + Re^{-\alpha t} \cos(\Omega t - \delta).$$

Hence, our model has five unknowns to find: \mathbf{G}_0 , R , α , Ω and δ . The easiest way to do this is to measure \mathbf{G}_0 , the patient's initial blood glucose concentration, when the patient arrives. Then measure the blood glucose concentration N more times giving the data pairs (t_1, \mathbf{G}_1) , (t_2, \mathbf{G}_2) and so on out to (t_N, \mathbf{G}_N) . Then form the least squares error function

$$E = \sum_{i=1}^N (\mathbf{G}_i - \mathbf{G}_0 - Re^{-\alpha t_i} \cos(\Omega t_i - \delta))^2 \quad (14.1)$$

and find the five parameter values that make this error a minimum. This can be done using standard MatLab tools. Numerous experiments have been done with this model and if we let $T_0 = 2\pi/\Omega$, it has been found that if $T_0 < 4$ hours, the patient is normal and if T_0 is much larger than that, the patient has mild diabetes.

Homework

Exercise 14.1.16

Exercise 14.1.17

Exercise 14.1.18

Exercise 14.1.19

Exercise 14.1.20

14.1.1.3 Fitting the Data

Here is some typical glucose versus time data.

Time	Glucose Level
0	95
1	180
2	155
3	140

We will now try to find the parameter values which minimize the nonlinear least squares problem Equation 14.1. This appears to be a simple problem, but you will see all numerical optimization problems are actually fairly difficult. Our problem is to find the free parameters $\mathbf{G}_0, R, \alpha, \Omega$ and δ which minimize

$$E(\mathbf{G}_0, R, \alpha, \Omega, \delta) = \sum_{i=1}^N (\mathbf{G}_i - \mathbf{G}_0 - Re^{-\alpha t_i} \cos(\Omega t_i - \delta))^2$$

For convenience, let $\mathbf{X} = [\mathbf{G}_0, R, \alpha, \Omega, \delta]'$ and $f_i(\mathbf{X}) = \mathbf{G}_i - \mathbf{G}_0 - Re^{-\alpha t_i} \cos(\Omega t_i - \delta)$; then we can rewrite the error function as $E(\mathbf{X}) = \sum_{i=1}^N f_i^2(\mathbf{X})$. Gradient descent requires the gradient of this error function. This is just a messy calculation;

$$\frac{\partial E}{\partial \mathbf{G}_0} = 2 \sum_{i=1}^N f_i(\mathbf{X}) \frac{\partial f_i}{\partial \mathbf{G}_0}$$

$$\begin{aligned}\frac{\partial E}{\partial R} &= 2 \sum_{i=1}^N f_i(\mathbf{X}) \frac{\partial f_i}{\partial R} \\ \frac{\partial E}{\partial \alpha} &= 2 \sum_{i=1}^N f_i(\mathbf{X}) \frac{\partial f_i}{\partial \alpha} \\ \frac{\partial E}{\partial \Omega} &= 2 \sum_{i=1}^N f_i(\mathbf{X}) \frac{\partial f_i}{\partial \Omega} \\ \frac{\partial E}{\partial \delta} &= 2 \sum_{i=1}^N f_i(\mathbf{X}) \frac{\partial f_i}{\partial \delta}\end{aligned}$$

where the f_i partials are given by

$$\begin{aligned}\frac{\partial f_i}{\partial \mathbf{G}_o} &= -1 \\ \frac{\partial f_i}{\partial R} &= -e^{-\alpha t_i} \cos(\Omega t_i - \delta) \\ \frac{\partial f_i}{\partial \alpha} &= t_i R e^{-\alpha t_i} \cos(\Omega t_i - \delta), \\ \frac{\partial f_i}{\partial \Omega} &= t_i R e^{-\alpha t_i} \sin(\Omega t_i - \delta) \\ \frac{\partial f_i}{\partial \delta} &= -R e^{-\alpha t_i} \sin(\Omega t_i - \delta)\end{aligned}$$

and so

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{G}_o} &= -2 \sum_{i=1}^N f_i(\mathbf{X}) \\ \frac{\partial E}{\partial R} &= -2 \sum_{i=1}^N f_i(\mathbf{X}) e^{-\alpha t_i} \cos(\Omega t_i - \delta) \\ \frac{\partial E}{\partial \alpha} &= 2 \sum_{i=1}^N f_i(\mathbf{X}) t_i R e^{-\alpha t_i} \cos(\Omega t_i - \delta), \\ \frac{\partial E}{\partial \Omega} &= 2 \sum_{i=1}^N f_i(\mathbf{X}) t_i R e^{-\alpha t_i} \sin(\Omega t_i - \delta) \\ \frac{\partial E}{\partial \delta} &= -2 \sum_{i=1}^N f_i(\mathbf{X}) R e^{-\alpha t_i} \sin(\Omega t_i - \delta)\end{aligned}$$

Now suppose we are at the point \mathbf{X}_0 and we want to know how much of the descent vector D to use. Note, if we use the amount ξ of the descent vector at \mathbf{X}_0 , we compute the new error value $E(\mathbf{X}_0 - \xi D(\mathbf{X}_0))$. Let $g(\xi) = E(\mathbf{X}_0 - \xi D(\mathbf{X}_0))$. We see $g(0) = E(\mathbf{X}_0)$ and given a first choice of $\xi = \lambda$, we have $g(\lambda) = E(\mathbf{X}_0 - \lambda D(\mathbf{X}_0))$. Next, let $\mathbf{Y} = \mathbf{X}_0 - \xi D(\mathbf{X}_0)$. Then, using the chain rule, we can calculate the directional derivative of g at 0. First, we have

$$g'(\xi) = -\langle \nabla(E), D(\mathbf{X}_0) \rangle$$

and using the normalized gradient of E as the descent vector, we find

$$g'(0) = -\|\nabla(E)\|.$$

Now let's approximate g using a quadratic model. Since we are trying for a minimum, in general we try to take a step in the direction of the negative gradient which makes the error function go down. Then, we have $g(0) = E(\mathbf{X}_0)$ is less than $g(\lambda) = E(\mathbf{X}_0 - \lambda D(\mathbf{X}_0))$ and the directional derivative gives $g'(\lambda) = -\|\nabla(E)\| < 0$. Hence, if we approximate g by a simple quadratic model, $g(\xi) = A + B\xi + C\xi^2$, this model will have a unique minimizer and we can use the value of ξ where the minimum occurs as our next choice of descent step. This technique is called a **Line Search Method** and it is quite useful. To summarize, we fit our g model and find

$$\begin{aligned} g(0) &= E(\mathbf{X}_0) \\ g'(0) &= -\|\nabla(E)\| \\ g(\lambda) &= A + B\lambda + C\lambda^2 \implies C = \frac{E(\mathbf{X}_0 - \lambda D(\mathbf{X}_0)) - E(\mathbf{X}_0) + \|\nabla(E)\|\lambda}{\lambda^2} \end{aligned}$$

The minimum of this quadratic occurs at $\lambda^* = \frac{B}{2C}$ and this will give us our next descent direction $\mathbf{X}_0 - \lambda^* D(\mathbf{X}_0)$.

Homework

Exercise 14.1.21

Exercise 14.1.22

Exercise 14.1.23

Exercise 14.1.24

Exercise 14.1.25

14.1.1.4 Gradient Descent Implementation

Let's get started on how to find the optimal parameters numerically. Along the way, we will show you how hard this is. We start with a minimal implementation. What is complicated here is that we have lots of functions that depend on the data we are trying to fit. So the number of functions depends on the size of the data set which makes it harder to set up.

Listing 14.1: **NonLinear LS For Diabetes Model: Version One**

```
function [ Error ,G0,R,alpha ,Omega, delta ,normgrad ,update ] =
    DiabetesGradOne( Initial ,lambda ,maxiter ,data )
%
% Initial guess for G0, R, alpha, Omega, delta
% data = collection of (time, glucose) pairs
% maxiter = maximum number of iterations to use
% lambda = how much of the descent vector to use
%
% setup least squares error function
N = length(data);
10 E = 0.0;
Time = data(:,1);
G = data(:,2);
```

```

% Initial = [initial G, initial R, initial alpha, initial Omega,
% initial delta]
g = Initial(1); r = Initial(2); a = Initial(3); o = Initial(4); d =
Initial(5);
15 f = @(t,equilG,r,a,o,d) equilG + r*exp(-a^2*t).*(cos(o*t-d));
E = DiabetesError(f,g,r,a,o,d,Time,G)
for i = 1:maxiter
    % calculate error
    E = DiabetesError(f,g,r,a,o,d,Time,G);
20 if (i == 1)
    Error(1) = E;
end
Error = [Error;E];
% find grad of E
[gradE, normgrad] = ErrorGradient(Time,G,g,r,a,o,d);
% find descent direction D
if(normgrad>1)
    Descent = gradE/normgrad;
else
30 Descent = gradE;
end
Del = [g;r;a;o;d]-lambda*Descent;
g = Del(1); r = Del(2); a = Del(3); o = Del(4); d = Del(5);
end
35 G0 = g; R = r; alpha = a; Omega = o; delta = d;
E = DiabetesError(f,g,r,a,o,d,Time,G)
update = [G0;R;alpha;Omega;delta];
end

```

Note inside this function, we call another function to calculate the gradient of the norm. This is given below and implements the formulae we presented earlier for these partial derivatives.

Listing 14.2: **The Error Gradient Function**

```

function [gradE, normgrad] = ErrorGradient(Time,G,g,r,a,o,d)
2 %
% Time is the time data
% G is the glucose data
% g = current equilG
% r = current R
7 % a = current alpha
% o = current Omega
% d = current delta
%
% Calculate Error gradient
12 ferror = @(t,G,g,r,a,o,d) G - g - r*exp(-a^2*t).*(cos(o*t-d));
gerror = @(t,r,a,o,d) r*exp(-a^2*t).*(cos(o*t-d));
herror = @(t,r,a,o,d) r*exp(-a^2*t).*(sin(o*t-d));
N = length(Time);
sum = 0;
17 for k=1:N
    sum = sum + ferror(Time(k),G(k),g,r,a,o,d);
end
pEpequilg = -2*sum;

```

14.1. LINEAR APPROXIMATIONS TO NONLINEAR ODE

267

```

22    sum = 0;
    for k=1:N
        sum = sum + ferror(Time(k),G(k),g,r,a,o,d)*gerror(Time(k),r,a,o,
                      d)/r;
    end
    pEpR = -2*sum;
    sum = 0;
27    for k=1:N
        sum = sum + ferror(Time(k),G(k),g,r,a,o,d)*Time(k)*2*a*Time(k)*
              gerror(Time(k),r,a,o,d);
    end
    pEpa = 2*sum;
    sum = 0;
32    for k=1:N
        sum = sum + ferror(Time(k),G(k),g,r,a,o,d)*Time(k)*herror(Time(k),
                      ),r,a,o,d);
    end
    pEpo = 2*sum;
    sum = 0;
37    for k=1:N
        sum = sum + ferror(Time(k),G(k),g,r,a,o,d)*herror(Time(k),r,a,o,
                      d);
    end
    pEpd = -2*sum;
    gradE = [ pEpequilg;pEpR;pEpa;pEpo;pEpd];
42    normgrad = norm(gradE);
    end

```

We also need code for the error calculations which is given here.

Listing 14.3: **Diabetes Error Calculation**

```

function E = DiabetesError(f,g,r,a,o,d,Time,G)
2 %
% T = Time
% G = Glucose values
% f = nonlinear insulin model
% g,a,r,o,d = parameters in diabetes nonlinear model
7 % N = size of data
N = length(G);
E = 0.0;
% calculate error function
for i = 1:N
12    E = E +( G(i) - f(Time(i),g,r,a,o,d) )^2;
end

```

Let's look at some run time results using this code. We will use **Octave** for the computations.

Listing 14.4: **Run time results for gradient descent on the original data**

```

Data = [0,95;1,180;2,155;3,140];
2 Time = Data(:,1);

```

```

G = Data(:,2);
f = @(t ,equilG ,r ,a,o,d) equilG + r*exp(-a^2*t).*(cos(o*t-d));
time = linspace(0,3,41);
RInitial = 53.64;
7 GOInitial = 95 + RInitial;
AInitial = sqrt(log(17/5));
OInitial = pi;
dInitial = -pi;
Initial = [GOInitial;RInitial;AInitial;OInitial;dInitial];
12 [Error ,G0,R,alpha ,Omega, delta ,normgrad ,update] = DiabetesGrad(Initial
    ,5.0e-4,20000,Data ,0);
Initiale = 463.94
E = 376.40
[Error ,G0,R,alpha ,Omega, delta ,normgrad ,update] = DiabetesGrad(Initial
    ,5.0e-4,40000,Data ,0);
Initiale = 463.94
17 E = 377.77
[Error ,G0,R,alpha ,Omega, delta ,normgrad ,update] = DiabetesGrad(Initial
    ,5.0e-4,100000,Data ,0);
Initiale = 463.94
E = 377.77

```

After 100,000 iterations we still do not have a good fit. Note we start with a small constant $\lambda = 5.0e - 4$ here. Try it yourself. If you let this value be larger, the optimization spins out of control. Also, we have not said how we chose our initial values. We actually looked at the data on a sheet of paper and did some rough calculations to try for some decent values. We will leave that to you to figure out. If the initial values are poorly chosen, gradient descent optimization is a great way to generate really bad values! So be warned. You will have to exercise due diligence to find a sweet starting spot.

We can see how we did by looking the resulting curve fit in Figure 14.1.

Homework

Exercise 14.1.26

Exercise 14.1.27

Exercise 14.1.28

Exercise 14.1.29

Exercise 14.1.30

14.1.1.5 Adding Line Search

Now let's add line search and see if it gets better. We will also try scaling the data so all the variables in question are roughly the same size. For us, a good choice is to scale the G_0 and the R value by 50, although we could try other choices. We have already discussed line search for our problem, but here it is again in a quick nutshell. If we are minimizing a function of M variables, say $f(\mathbf{X})$, then if we are at the point \mathbf{X}_0 , we can look at the *slice* of this function if we move out from the base point \mathbf{X}_0 in the direction of the negative gradient, $\nabla(f(\mathbf{X}_0)) = \nabla(f^0)$. Define a function of the single variable ξ as $g(\xi) = f(\mathbf{X}_0 + \xi \nabla(f^0))$. Then, we can try to approximate g as a quadratic, $g(\xi) \approx A + B\xi + C\xi^2$. Of course, the actual function might not be approximated nicely by such a quadratic, but it is worth a shot! Once we fit the parameters A , B and C , we see this quadratic model is minimized at $\lambda^* = -\frac{B}{2C}$. The code now adds the line search code which is contained in the block

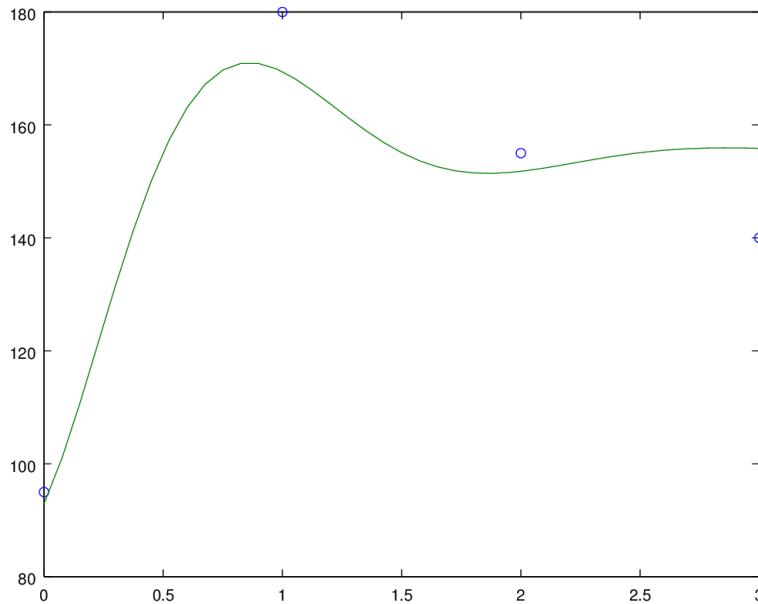


Figure 14.1: The Diabetes Model Curve Fit: No Line Search

Listing 14.5: **Line Search Code**

```

function [lambdastar ,DelOptimal ] = LineSearch(Del,EStart ,EFullStep ,
Base ,Descent ,normgrad ,lambda)
%
% Del is the update due to the step lambda
% Estart is E at the base value
% EFullStep is E at the given lambda value
% Base is the current parameter vector
% Descent is the current descent vector
% normgrad of grad of E at the base value
% lambda is the current step
%
% we have enough information to do a line search here
A = EStart;
BCheck = -normgrad;
C = ( EFullStep - A - BCheck*lambda )/(lambda^2);
lambdastar = -BCheck/(2*C);
if(C<0 || lambdastar < 0)
    % we are going to a maximum on the line search; reject
    DelOptimal = Del;
else
    % we have a minimum on the line search
    DelOptimal= Base-lambdastar*Descent;
end
end

```

Let's put it all together now.

Listing 14.6: **NonLinear LS For Diabetes Model**

```

function [Error ,G0,R,alpha ,Omega,delta ,normgrad ,update] = DiabetesGrad
    (Initial ,lambda ,maxiter ,data ,dolinesearch)
2 %
% Initial guess for G0, R, alpha, Omega, delta
% data = collection of (time, glucose) pairs
% tol = stop tolerance
% maxiter = maximum number of iterations to use
% lambda = how much of the descent vector to use
%
% setup least squares error function
N = length(data);
E = 0.0;
12 Time = data(:,1);
G = data(:,2);
% Initial = [initial G, initial R, initial alpha, initial Omega,
%             initial delta]
g = Initial(1);
r = Initial(2);
17 a = Initial(3);
o = Initial(4);
d = Initial(5);
f = @(t,equilG,r,a,o,d) equilG + r*exp(-a^2*t).*(cos(o*t-d));
Initiale = DiabetesError(f,g,r,a,o,d,Time,G)
22 for i = 1:maxiter
    % calculate error
    E = DiabetesError(f,g,r,a,o,d,Time,G);
    if (i == 1)
        Error(1) = E;
    end
    Error = [Error;E];
    % Calculate Error gradient
    [gradE, normgrad] = ErrorGradient(Time,G,g,r,a,o,d);
    % find descent direction
32 if(normgrad>1)
    Descent = gradE/normgrad;
    else
        Descent = gradE;
    end
    Base = [g;r;a;o;d];
    Del = Base-lambda*Descent;
    newg = Del(1); newr = Del(2); newa = Del(3); newo = Del(4); newd =
        Del(5);
    EStart = E;
    EFullStep = DiabetesError(f,newg,newr,newa,newo,newd,Time,G);
42 if(dolinesearch==1)
    [lambdastar ,DelOptimal] = LineSearch(Del,EStart ,EFullStep ,Base ,
        Descent ,normgrad ,lambda );
    g = DelOptimal(1); r = DelOptimal(2); a = DelOptimal(3); o =
        DelOptimal(4); d = DelOptimal(5);
    EOOptimalStep = DiabetesError(f,g,r,a,o,d,Time,G);
    else

```

```

47      g = Del(1); r = Del(2); a = Del(3); o = Del(4); d = Del(5);
end
G0 = g; R = r; alpha = a; Omega = o; delta = d;
E = DiabetesError(f,g,r,a,o,d,Time,G)
52 update = [G0;R;alpha;Omega;delta];
end

```

We can see how this is working by letting some of the temporary calculations print. Here are two iterations of line search printing out the A , B and C and the relevant energy values. Our initial values don't matter much here as we are just checking out the line search algorithm.

Listing 14.7: **Some Details of the Line Search**

```

Data = [0,95;1,180;2,155;3,140];
2 Time = Data(:,1);
G = Data(:,2);
f = @(t,equilG,r,a,o,d) equilG + r*exp(-a^2*t).*(cos(o*t-d));
time = linspace(0,3,41);
RInitial = 85*17/21;
7 GOInitial = 95 + RInitial;
AInitial = sqrt(log(17/4));
OInitial = pi;
dInitial = -pi;
Initial = [GOInitial;RInitial;AInitial;OInitial;dInitial];
12 [Error,G0,R,alpha,Omega,delta,normgrad,update] = DiabetesGrad(Initial
    ,5.0e-4,2,Data,1);
Initiale = 635.38
EFullStep = 635.31
A = 635.38
BCheck = -595.35
17 C = 9.1002e+05
lambdastar = 3.2711e-04
EFullStep = 635.27
A = 635.33
BCheck = -592.34
22 C = 9.0725e+05
lambdastar = 3.2645e-04
E = 635.29

```

Now let's remove those prints and let it run for awhile. We are using the original data here to try to find the fit.

Listing 14.8: **The Full Run with Line Search**

```

1 [Error,G0,R,alpha,Omega,delta,normgrad,update] = DiabetesGrad(Initial
    ,5.0e-4,20000,Data,1);
Initiale = 635.38
E = 389.07
[Error,G0,R,alpha,Omega,delta,normgrad,update] = DiabetesGrad(Initial
    ,5.0e-4,30000,Data,1);
Initiale = 635.38

```

6 E = 0.067171

We have success! The line search got the job done in 30,000 iterations while the attempt using just gradient descent without line search failed. but remember, we do additional processing at each step. We show the resulting curve fit in Figure 14.2.

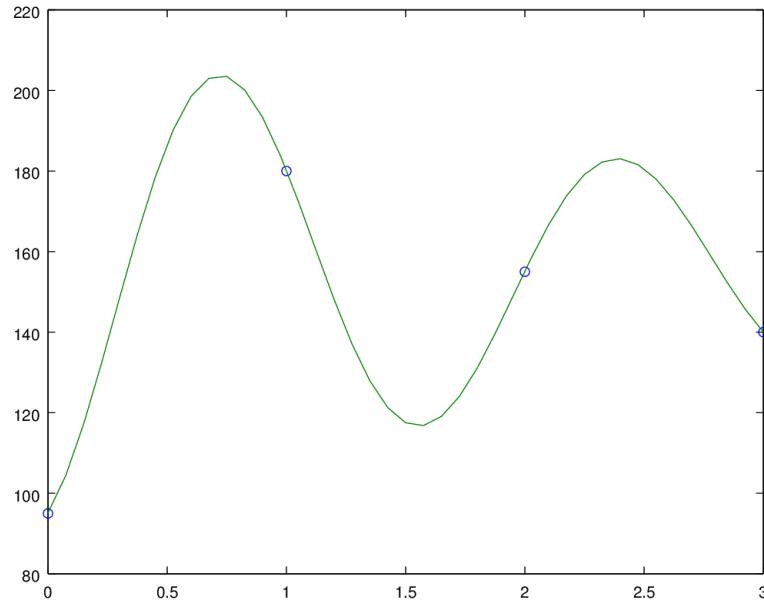


Figure 14.2: The Diabetes Model Curve Fit with Line Search on Unscaled Data

The qualitative look of this fit is a bit different. We leave it to you to think about how we are supposed to choose which fit is better; i.e. which fit is a better one to use for the biological reality we are trying to model? This is a really hard question. Finally, the optimal values of the parameters are

Listing 14.9: **Optimal Parameter Values**

```
update
update =
154.37240
4      62.73116
        0.57076
        3.77081
       -3.47707
```

The critical value of $2\pi/\Omega = 1.6663$ here which is less than 4 so this patient is normal!

Also, note these sorts of optimizations are very frustrating, If we use the scaled version of the first `Initial = [GOInitial; RInitial; AInitial; OInitial; dInitial];` we make no progress even though we run for 60,000 iterations and these iterations a bit more expensive because we use line search. So let's perturb the starting point a bit and see what happens.

Listing 14.10: **Runtime Results**

```
% use scaled data
Data = [0,95/50;1,180/50;2,155/50;3,140/50];
Time = Data(:,1);
G = Data(:,2);
5 RInitial = 53.64/50;
GOInitial = 95/50 + RInitial;
AInitial = sqrt(log(17/5));
OInitial = pi;
dInitial = -pi;
10 % perturb the start point a bit
Initial = [GOInitial;.7*RInitial;AInitial;1.1*OInitial;.9*dInitial];
[Error,G0,R,alpha,Omega,delta,normgrad,update] = DiabetesGrad(Initial
,5.0e-4,40000,Data,1);
Initiale = 0.36485
E = 1.5064e-06
```

We have success! We see the fit to the data in Figure 14.3.

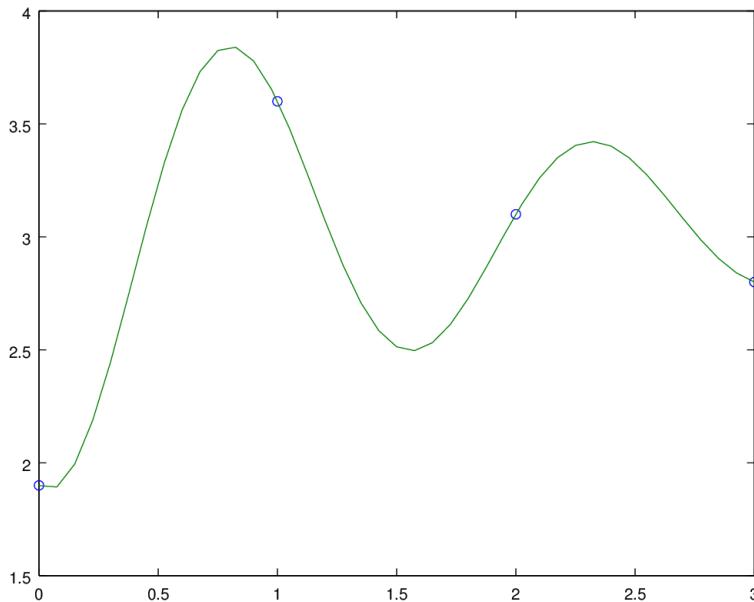


Figure 14.3: The Diabetes Model Curve Fit with Line Search on Scaled Data

The moral here is that it is quite difficult to automate our investigations. This is why truly professional optimization code is so complicated. In our problem here, we have a hard time finding a good starting point and we even find that scaling the data – which seems like a good idea – is not as helpful as we thought it would be. Breathe a deep sigh and accept this as our lot!

Homework

Exercise 14.1.31**Exercise 14.1.32****Exercise 14.1.33****Exercise 14.1.34****Exercise 14.1.35****14.1.2 An Autoimmune Model**

We assume we have a large population of cells \mathbf{T} which consists of cells which are infected or altered in two distinct ways by a trigger \mathbf{V} , based on signals \mathbf{I} , \mathbf{J} and \mathbf{K} . These two distinct populations of cells will be labeled \mathbf{M} and \mathbf{N} . There are also non infected cells, \mathbf{H} and non infected cells which will be removed due to auto immune action which we call \mathbf{C} , for collateral damage. We will be using the same approach to studying nonlinear interactions that was used in (Peterson et al. (11) June 21, 2017).

We assume the dynamics here are

$$\begin{aligned}\mathbf{C}'(t) &= F_1(\mathbf{C}, \mathbf{M}, \mathbf{N}) \\ \mathbf{M}'(t) &= F_2(\mathbf{C}, \mathbf{M}, \mathbf{N}) \\ \mathbf{N}'(t) &= F_3(\mathbf{C}, \mathbf{M}, \mathbf{N})\end{aligned}$$

There are then three nonlinear interaction functions F_1 , F_2 and F_3 because we know \mathbf{C} , \mathbf{M} and \mathbf{N} depend on each other's levels in very complicated ways. Usually, we assume the initial trigger dose \mathbf{V}_0 gives rise to some fraction of infected cells and the effect of the trigger will be different in the two cell populations \mathbf{M} and \mathbf{N} .

Assumption 14.1.1

We assume the number of infected cells is $p_0 \mathbf{V}_0$ which is split into $p_1 p_0 \mathbf{V}_0$ in population \mathbf{N} and $p_2 p_0 \mathbf{V}_0$ in \mathbf{M} , where $p_1 + p_2 = 1$.

For example, a reasonable choice is $p_1 = 0.99$ and $p_2 = 0.01$. Thus, the total amount of trigger that goes into altered cells is $p_0 \mathbf{V}_0$ and the amount of free trigger is therefore $(1 - p_0) \mathbf{V}_0$. Thus, we could expect $\mathbf{C}_0 = 0$, $\mathbf{M}_0 = p_2 p_0 \mathbf{V}_0$ and $\mathbf{N}_0 = \mathbf{M}_0 = p_1 p_0 \mathbf{V}_0$. However, we will explicitly assume we are starting from a point of equilibrium prior to the administration of the viral dose \mathbf{V}_0 . We could assume there is always some level of collateral damage, \mathbf{C}_0 in a host, but we will not do that. We will therefore assume \mathbf{C} , \mathbf{M} and \mathbf{N} have achieved these values $\mathbf{C}_0 = 0$, $\mathbf{M}_0 = 0$ and $\mathbf{N}_0 = 0$ right before the moment of alteration by the trigger. Hence, we don't expect to there to be initial contribution to $\mathbf{C}'(0)$, $\mathbf{M}'(0)$ and $\mathbf{N}'(0)$; i.e. $F_1(\mathbf{C}_0, \mathbf{M}_0, \mathbf{N}_0) = 0$, $F_2(\mathbf{C}_0, \mathbf{M}_0, \mathbf{N}_0) = 0$ and $F_3(\mathbf{C}_0, \mathbf{M}_0, \mathbf{N}_0) = 0$. We are interested in the deviation of \mathbf{C} , \mathbf{M} and \mathbf{N} from their optimal values \mathbf{C}_0 , \mathbf{M}_0 and \mathbf{N}_0 , so let $\mathbf{c} = \mathbf{C} - \mathbf{C}_0$, $\mathbf{m} = \mathbf{M} - \mathbf{M}_0$ and $\mathbf{n} = \mathbf{N} - \mathbf{N}_0$. We can then write $\mathbf{C} = \mathbf{C}_0 + \mathbf{c}$, $\mathbf{M} = \mathbf{M}_0 + \mathbf{m}$ and $\mathbf{N} = \mathbf{N}_0 + \mathbf{n}$. The model can then be rewritten as

$$\begin{aligned}(\mathbf{C}_0 + \mathbf{c})'(t) &= F_1(\mathbf{C}_0 + \mathbf{c}, \mathbf{M}_0 + \mathbf{m}, \mathbf{N}_0 + \mathbf{n}) \\ (\mathbf{M}_0 + \mathbf{m})'(t) &= F_2((\mathbf{C}_0 + \mathbf{c}, \mathbf{M}_0 + \mathbf{m}, \mathbf{N}_0 + \mathbf{n})) \\ (\mathbf{M}_0 + \mathbf{M})'(t) &= F_3((\mathbf{C}_0 + \mathbf{c}, \mathbf{M}_0 + \mathbf{m}, \mathbf{N}_0 + \mathbf{n}))\end{aligned}$$

or

$$\begin{aligned}\mathbf{c}'(t) &= F_1(\mathbf{C}_0 + \mathbf{c}, \mathbf{M}_0 + \mathbf{m}, \mathbf{N}_0 + \mathbf{n}) \\ \mathbf{m}'(t) &= F_2((\mathbf{C}_0 + \mathbf{c}, \mathbf{M}_0 + \mathbf{m}, \mathbf{N}_0 + \mathbf{n}))\end{aligned}$$

$$\mathbf{n}'(t) = F_3((\mathbf{C}_0 + \mathbf{c}, \mathbf{M}_0 + \mathbf{m}, \mathbf{N}_0 + \mathbf{n})$$

Next, we do a standard tangent plane approximation on the nonlinear dynamics functions F_1 , F_2 and F_3 to derive approximation dynamics. We find the approximate dynamics are

$$\begin{bmatrix} \mathbf{c}' \\ \mathbf{m}' \\ \mathbf{n}' \end{bmatrix} \approx \begin{bmatrix} F_{1c}^o & F_{1m}^o & F_{1n}^o \\ F_{2c}^o & F_{2m}^o & F_{2n}^o \\ F_{3c}^o & F_{3m}^o & F_{3n}^o \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{m} \\ \mathbf{n} \end{bmatrix}$$

where we now use a standard subscript scheme to indicate the partials. In (Peterson et al. (12) June 21, 2017), we show how to build additional models which include the signals IFN- γ (\mathbf{I}), \mathbf{J} and \mathbf{K} . However, you should be able to see how powerful the idea of linear approximation is even if we do not know the values of the partial derivatives at the equilibrium points.

14.1.2.1 Homework

Exercise 14.1.36

Exercise 14.1.37

Exercise 14.1.38

Exercise 14.1.39

Exercise 14.1.40

14.2 Finite Difference Approximations in PDE

Another way to approximate the solution of linear partial differential equations is to use what are called *finite difference techniques*. Essentially, in these methods, the various partial derivatives are replaced by tangent line like approximations. We first must discuss how we handle the resulting approximation error carefully. Let's review how to approximate a function of two variables using Taylor series. Let's assume that $u(x, t)$ is a nice, smooth function of the two variables x (our spatial variable) and t (our time variable). For ease of discussion, we will focus on the space interval $[0, L]$ and the time interval $[0, M]$ where L and M are both positive numbers. We assume that u is continuous on the rectangle $\mathcal{D} = [0, L] \times [0, M]$. This means that $\lim_{(x,t) \rightarrow (x_0,t_0)} f(x, t) = f(x_0, t_0)$ for all pairs (x_0, t_0) in \mathcal{D} . Note that since the rectangle is two dimensional, there are an infinite number of ways the pairs (x, t) can approach (x_0, t_0) . Continuity at the point (x_0, t_0) means that it does not matter how we do the approach; we always get the same answer. More precisely, if the positive tolerance ϵ is given, there is a positive number δ so that

$$\sqrt{(x - x_0)^2 + (t - t_0)^2} < \delta \text{ and } (x, t) \in \mathcal{D} \implies |f(x, t) - f(x_0, t_0)| < \epsilon.$$

We also assume the partial derivatives up to order 4 are continuous on \mathcal{D} . Thus, we assume the continuity of many orders of partials. These terms get hard to write down as there are so many of them, so let's define $D_{ij}^k u$ to be the k^{th} partial derivative of u with respect to x i times and t j times where $i + j$ must add up to k . More formally,

$$D_{ij}^k u = \frac{\partial^k u}{\partial x^i \partial t^j}, \text{ for all } i + j = k, \text{ with } i, j \geq 0.$$

So using this notation, $D_{10}^1 u = \frac{\partial u}{\partial x}$, $D_{11}^2 u = \frac{\partial^2 u}{\partial x \partial t}$, $D_{40}^4 u = \frac{\partial^4 u}{\partial x^4}$ and so forth. We will assume the partials $D_{ij}^k u$ are continuous on \mathcal{D} up to an including $k = 4$. It is known that continuous functions

on a rectangle of the form \mathcal{D} must be bounded. Hence, there are positive numbers B_{ij}^k so that for all (x, t) in \mathcal{D} , we have

$$|D_{ij}^k u(x, t)| < B_{ij}^k, \quad \text{for all } (x, t) \in \mathcal{D}.$$

Hence, the maximum of $|D_{ij}^k u(x, t)|$ over \mathcal{D} exists and letting $\|D_{ij}^k u\| = \max |D_{ij}^k u(x, t)|$ over all $(x, t) \in \mathcal{D}$, we have each $\|D_{ij}^k u\| \leq D_{ij}^k$. This implies we can take the maximum of all these individual bounds and state that there is a single constant C so that

$$\|D_{ij}^k u\| \leq C. \quad (14.2)$$

14.2.1 Approximating First Order Partial

Define $h(t) = u(x_0, t)$. The second order Taylor expansion of h is then

$$h(t) = h(t_0) + h'(t_0)(t - t_0) + \frac{1}{2}h''(c_t)(t - t_0)^2$$

where c_t is some point between t_0 and t . Using the chain rule for functions of two variables, it is easy to see $h(x_0) = u(x_0, y_0)$, $h'(t) = u_t(x_0, t)$ and $h''(t) = u_{tt}(x_0, t)$. Hence, we can rewrite the expansion as

$$u(x, t) = u(x_0, t_0) + u_t(x_0, t_0)(t - t_0) + \frac{1}{2}u_{tt}(x_0, c_t)(t - t_0)^2.$$

We can then write these another way by letting $\Delta t = t - t_0$. This gives

$$u(x_0, t_0 + \Delta t) = u(x_0, t_0) + u_t(x_0, t_0)\Delta t + \frac{1}{2}u_{tt}(x_0, c_t)(\Delta t)^2.$$

We can use these expansions to estimate the first order partials in a variety of ways.

Homework

Exercise 14.2.1

Exercise 14.2.2

Exercise 14.2.3

Exercise 14.2.4

Exercise 14.2.5

14.2.1.1 Central Differences

First, let's look at the central difference for the first order partial derivative with respect to t . We have

$$\begin{aligned} u(x_0, t_0 - \Delta t) &= u(x_0, t_0) - u_t(x_0, t_0)\Delta t + \frac{1}{2}u_{tt}(x_0, c_t^-)(\Delta t)^2 \\ u(x_0, t_0 + \Delta t) &= u(x_0, t_0) + u_t(x_0, t_0)\Delta t + \frac{1}{2}u_{tt}(x_0, c_t^+)(\Delta t)^2. \end{aligned}$$

Subtracting, we find

$$u(x_0, t_0 + \Delta t) - u(x_0, t_0 - \Delta t) = 2u_t(x_0, t_0)\Delta t + \frac{1}{2} \left(u_{tt}(x_0, c_t^+) - u_{tt}(x_0, c_t^-) \right) (\Delta t)^2.$$

Thus,

$$u_t(x_0, t_0) = \frac{u(x_0, t_0 + \Delta t) - u(x_0, t_0 - \Delta t)}{2\Delta t} - \frac{1}{2} \left(u_{tt}(x_0, c_t^+) - u_{tt}(x_0, c_t^-) \right) \Delta t$$

Homework

Exercise 14.2.6

Exercise 14.2.7

Exercise 14.2.8

Exercise 14.2.9

Exercise 14.2.10

14.2.1.2 Forward Differences

Next, let's look at the forward difference. We have

$$u(x_0, t_0 + \Delta t) = u(x_0, t_0) + u_t(x_0, t_0)\Delta t + \frac{1}{2}u_{tt}(x_0, c_t^+)(\Delta t)^2.$$

Subtracting, we find

$$u(x_0, t_0 + \Delta t) - u(x_0, t_0) = u_t(x_0, t_0)\Delta t + \frac{1}{2}u_{tt}(x_0, c_t^+)(\Delta t)^2.$$

Thus,

$$u_t(x_0, t_0) = \frac{u(x_0, t_0 + \Delta t) - u(x_0, t_0)}{\Delta t} - \frac{1}{2}u_{tt}(x_0, c_t^+)\Delta t$$

Homework

Exercise 14.2.11

Exercise 14.2.12

Exercise 14.2.13

Exercise 14.2.14

Exercise 14.2.15

14.2.2 Approximating Second Order Partialials

The approximation we wish to use for the second order partial u_{xx} is obtained like this. We fix the point (x_0, t_0) as usual and let $g(x) = u(x, t_0)$. The fourth order Taylor expansion of g is then

$$g(x) = g(x_0) + g'(x_0)(x - x_0) + \frac{1}{2}g''(x_0)(x - x_0)^2 + \frac{1}{6}g'''(x_0)(x - x_0)^3 + \frac{1}{24}g''''(c_x^+)(x - x_0)^4$$

where c_x^+ is some point between x_0 and x . It is easy to see $g(x_0) = u(x_0, y_0)$, $g'(x) = u_x(x, t_0)$ and $g''(x) = u_{xx}(x, t_0)$ and so forth. Hence, letting $\Delta x = x - x_0$, we can rewrite the expansion as

$$\begin{aligned} u(x_0 + \Delta x, t_0) &= u(x_0, t_0) + u_x(x_0, t_0)\Delta x + \frac{1}{2}u_{xx}(x_0, t_0)(\Delta x)^2 + \frac{1}{6}u_{xxx}(x_0, t_0)(\Delta x)^3 \\ &\quad + \frac{1}{24}u_{xxxx}(c_x^+, t_0)(\Delta x)^4 \end{aligned}$$

A similar expansion gives

$$\begin{aligned} u(x_0 - \Delta x, t_0) &= u(x_0, t_0) - u_x(x_0, t_0)\Delta x + \frac{1}{2}u_{xx}(x_0, t_0)(\Delta x)^2 - \frac{1}{6}u_{xxx}(x_0, t_0)(\Delta x)^3 \\ &\quad + \frac{1}{24}u_{xxxx}(c_x^-, t_0)(\Delta x)^4 \end{aligned}$$

Adding, we have

$$\begin{aligned} u(x_0 + \Delta x, t_0) + u(x_0 - \Delta x, t_0) &= 2u(x_0, t_0) + u_{xx}(x_0, t_0)(\Delta x)^2 + \frac{1}{24}(u_{xxxx}(c_x^+, t_0) \\ &\quad + u_{xxxx}(c_x^-, t_0))(\Delta x)^4 \end{aligned}$$

The intermediate value theorem tells us that a continuous function takes on every value between two points. Hence, there is a point c_x between c_x^+ and c_x^- so that

$$u_{xxxx}(c_x, t_0) = \frac{1}{2}(u_{xxxx}(c_x^+, t_0) + u_{xxxx}(c_x^-, t_0)).$$

We can then write the approximation as

$$u(x_0 + \Delta x, t) + u(x_0 - \Delta x, t) = 2u(x_0, t_0) + u_{xx}(x_0, t_0)(\Delta x)^2 + \frac{1}{12}u_{xxxx}(c_x, t_0)(\Delta x)^4$$

which tells us that

$$u_{xx}(x_0, t_0) = \frac{u(x_0 + \Delta x, t_0) + u(x_0 - \Delta x, t_0) - 2u(x_0, t_0)}{(\Delta x)^2} - \frac{1}{12}u_{xxxx}(c_x, t_0)(\Delta x)^2$$

14.2.3 Homework

Exercise 14.2.16

Exercise 14.2.17

Exercise 14.2.18

Exercise 14.2.19**Exercise 14.2.20**

14.3 Approximating the Diffusion Equation

Using these approximations, we can approximate the diffusion equation $u_t = Du_{xx}$ using a forward difference for the first order partial with respect to time as

$$u_t(x_0, t_0) - Du_{xx}(x_0, t_0) = \frac{u(x_0, t_0 + \Delta t) - u(x_0, t_0)}{\Delta t} - \frac{1}{2}u_{tt}(x_0, c_t^+) \Delta t \\ - D\left(\frac{u(x_0 + \Delta x, t_0) + u(x_0 - \Delta x, t_0) - 2u(x_0, t_0)}{(\Delta x)^2} - \frac{1}{12}u_{xxxx}(c_x, t_0)(\Delta x)^2\right)$$

This can be reorganized as

$$u_t(x_0, t_0) - Du_{xx}(x_0, t_0) = \frac{u(x_0, t_0 + \Delta t) - u(x_0, t_0)}{\Delta t} \\ - D\left(\frac{u(x_0 + \Delta x, t_0) + u(x_0 - \Delta x, t_0) - 2u(x_0, t_0)}{(\Delta x)^2}\right) \\ - \frac{1}{2}u_{tt}(x_0, c_t^+) \Delta t + \frac{D}{12}u_{xxxx}(c_x, t_0)(\Delta x)^2$$

Now consider the error

$$\left| u_t(x_0, t_0) - Du_{xx}(x_0, t_0) - \frac{u(x_0, t_0 + \Delta t) - u(x_0, t_0)}{\Delta t} \right. \\ \left. + D\left(\frac{u(x_0 + \Delta x, t_0) + u(x_0 - \Delta x, t_0) - 2u(x_0, t_0)}{(\Delta x)^2}\right) \right| \\ \leq \frac{1}{2}|u_{tt}(x_0, c_t^+)|\Delta t + \frac{D}{12}|u_{xxxx}(c_x, t_0)|(\Delta x)^2.$$

Then, using the estimates from Equation 14.2, we have

$$\left| u_t(x_0, t_0) - Du_{xx}(x_0, t_0) - \frac{u(x_0, t_0 + \Delta t) - u(x_0, t_0)}{\Delta t} \right. \\ \left. + D\left(\frac{u(x_0 + \Delta x, t_0) + u(x_0 - \Delta x, t_0) - 2u(x_0, t_0)}{(\Delta x)^2}\right) \right| \\ \leq \frac{1}{2}C\Delta t + \frac{D}{12}C(\Delta x)^2, \quad (14.3)$$

which clearly goes to zero as Δt and Δx go to zero. Hence, we know the replacement of the original partial derivatives in the diffusion equation by the finite difference approximations can be made as accurate as we wish by suitable choice of Δt and Δx . To see how we translate this into an approximate numerical method, divide the space interval $[0, L]$ into pieces of size Δx ; then there will be $N \approx \frac{L}{\Delta x}$ such pieces within the interval. Divide the time interval $[0, T]$ in a similar way by Δt into $M \approx \frac{T}{\Delta t}$ subintervals.

The true solution at the point $(n\Delta x, m\Delta t)$ is then $u((n\Delta x, m\Delta t))$ which we will denote by u_{mn} . Hence, we know that the pair $(n\Delta x, m\Delta t)$ satisfies

$$\begin{aligned} & \frac{u(n\Delta x, (m+1)\Delta t) - u(n\Delta x, m\Delta t)}{\Delta t} \\ & - D \left(\frac{u((n+1)\Delta x, m\Delta t) + u((n-1)\Delta x, m\Delta t) - 2u(n\Delta x, m\Delta t)}{(\Delta x)^2} \right) \\ & = -\frac{1}{2}u_{tt}(x_0, c_t^+) \Delta t + \frac{D}{12}u_{xxxx}(c_x, t_0)(\Delta x)^2 \end{aligned}$$

Letting $E_{n,m}^1 = -\frac{1}{2}u_{tt}(x_0, c_t^+) \Delta t$ and $E_{n,m}^2 = \frac{D}{12}u_{xxxx}(c_x, t_0)(\Delta x)^2$, this is then rewritten as

$$\frac{u_{n,m+1} - u_{n,m}}{\Delta t} - D \left(\frac{u_{n+1,m} + u_{n-1,m} - 2u_{n,m}}{(\Delta x)^2} \right) = E_{n,m}^1 \Delta t + E_{n,m}^2 (\Delta x)^2.$$

Now solve for the term $u_{n,m+1}$ to find

$$\begin{aligned} u_{n,m+1} &= u_{n,m} + D \frac{\Delta t}{(\Delta x)^2} \left(u_{n+1,m} + u_{n-1,m} - 2u_{n,m} \right) \\ &+ E_{n,m}^1 (\Delta t)^2 + E_{n,m}^2 \Delta t (\Delta x)^2. \end{aligned} \tag{14.4}$$

We therefore know the error we make in computing

$$u_{n,m+1} = u_{n,m} + D \frac{\Delta t}{(\Delta x)^2} \left(u_{n+1,m} + u_{n-1,m} - 2u_{n,m} \right)$$

using the true solution values is reasonably small when Δx and Δt are also sufficiently small due to the error estimates in Equation 14.3. At this point, there is no relationship between Δx and Δt that is required for the approximation to work. However, when we take Equation 14.4 and solve it iteratively as a recursive equation, problems arise. Consider a full diffusion equation model on the domain \mathcal{D} :

$$\begin{aligned} \frac{\partial u}{\partial t} - D \frac{\partial^2 u}{\partial x^2} &= 0 \\ u(0, t) &= 0, \text{ for } 0 \leq t \leq T \\ u(L, t) &= 0, \text{ for } 0 \leq t \leq T \\ u(x, 0) &= f(x), \text{ for } 0 \leq x \leq L \end{aligned}$$

The discrete boundary conditions at each *grid point* $(n\Delta x, m\Delta t)$ are then

$$u_{0,m} = u(0, m\Delta t) = 0, \quad 1 \leq m \leq M \tag{14.5}$$

$$u_{N,m} = u(N\Delta x, m\Delta t) = 0, \quad 1 \leq m \leq M \tag{14.6}$$

$$u_{n,0} = u(n\Delta x, 0) = f(n\Delta x) = f_n, \quad 0 \leq n \leq N \tag{14.7}$$

as $N\Delta x = L$ and $M\Delta t = T$. Using these boundary conditions coupled with the discrete dynamics of Equation 14.4, we obtain a full recursive system to solve. First, recall the true values satisfy Equation 14.4. Letting $r = \Delta t/(\Delta x)^2$ and adding the boundary conditions, we obtain the full system

$$\begin{aligned} u_{n,m+1} &= u_{n,m} + Dr(u_{n+1,m} + u_{n-1,m} - 2u_{n,m}) \\ &\quad + E_{n,m}^1(\Delta t)^2 + E_{n,m}^2 \Delta t (\Delta x)^2 \end{aligned} \quad (14.8)$$

$$u_{0,m} = 0, \quad 1 \leq m \leq M \quad (14.9)$$

$$u_{N,m} = 0, \quad 1 \leq m \leq M \quad (14.10)$$

$$u_{n,0} = f_n, \quad 0 \leq n \leq N \quad (14.11)$$

This gives us a recursive system we can solve of the form

$$v_{n,m+1} = v_{n,m} + Dr(v_{n+1,m} + v_{n-1,m} - 2v_{n,m}) \quad (14.12)$$

$$v_{0,m} = 0, \quad 1 \leq m \leq M \quad (14.13)$$

$$v_{N,m} = 0, \quad 1 \leq m \leq M \quad (14.14)$$

$$v_{n,0} = f_n, \quad 0 \leq n \leq N \quad (14.15)$$

where $v_{n,m}$ is the solution to this discrete recursion system at the grid points. To solve this system, we start at time 0, time index value $m = 0$, and use Equation 14.9 to get

$$v_{1,1} = v_{1,0} + Dr(v_{2,0} + v_{0,0} - 2v_{1,0}).$$

We know $v_{0,0} = f_0$, $v_{1,0} = f_1$ and $v_{2,0} = f_2$. Hence, we have

$$v_{1,1} = f_1 + Dr(f_2 + f_0 - 2f_1).$$

We can do this for any n to find

$$\begin{aligned} v_{n,1} &= v_{n,0} + Dr(v_{n+1,0} + v_{n-1,0} - 2v_{n,0}) \\ &= f_n + Dr(f_{n+1} + f_{n-1} - 2f_n) \end{aligned}$$

Once the values at $m = 1$ have been found, we use the recursion to find the values at the next time step $m = 2$ and so on.

14.3.0.1 Homework

Exercise 14.3.1

Exercise 14.3.2

Exercise 14.3.3

Exercise 14.3.4

Exercise 14.3.5

14.3.1 Error Analysis

Let's denote the difference between the true grid point values and the discrete solution values by $w_{n,m} = u_{n,m} - v_{n,m}$. A simple subtraction shows that $w_{n,m}$ satisfies the following discrete system:

$$\begin{aligned} w_{n,m+1} &= w_{n,m} + Dr(w_{n+1,m} + w_{n-1,m} - 2w_{n,m}) \\ &\quad + E_{n,m}^1(\Delta t)^2 + E_{n,m}^2\Delta t(\Delta x)^2 \end{aligned} \quad (14.16)$$

$$w_{0,m} = 0, \quad 1 \leq m \leq M \quad (14.17)$$

$$w_{N,m} = 0, \quad 1 \leq m \leq M \quad (14.18)$$

$$w_{n,0} = 0, \quad 0 \leq n \leq N \quad (14.19)$$

From our previous estimates, we know $E_{n,m}^1(\Delta t)^2 + E_{n,m}^2\Delta t(\Delta x)^2 \leq C(\Delta t)^2 + \Delta t(\Delta x)^2$ and so we can overestimate these terms to obtain

$$|w_{n,m+1}| \leq |1 - 2Dr||w_{n,m}| + 2Dr(|w_{n+1,m}| + |w_{n-1,m}|) + C((\Delta t)^2 + \Delta t(\Delta x)^2)$$

Now since this equation holds for x grid positions, letting $\|w^p\| = \max_{0 \leq N} |w_{n,p}|$, we see

$$|w_{n,m+1}| \leq |1 - 2Dr|\|w^m\| + 2Dr\|w^m\| + E(\Delta t, \Delta x)$$

where $E(\Delta t, \Delta x)$ is the error term $C((\Delta t)^2 + \Delta t(\Delta x)^2)$. This equation holds for all indices n on the left hand side which implies

$$\|w^{m+1}\| \leq (|1 - 2Dr| + 2Dr)\|w^m\| + E(\Delta t, \Delta x)$$

Letting $g = |1 - 2Dr| + 2Dr$, we find the estimate

$$\|w^1\| \leq g\|w^0\| + E(\Delta t, \Delta x) = E(\Delta t, \Delta x)$$

since $\|w^0\| = 0$ due to our error boundary conditions. The next steps are

$$\begin{aligned} \|w^2\| &\leq g\|w^1\| + E(\Delta t, \Delta x) \leq (1+g)E(\Delta t, \Delta x) \\ \|w^3\| &\leq g\|w^2\| + E(\Delta t, \Delta x) \leq (1+g+g^2)E(\Delta t, \Delta x) \\ &\dots \\ \|w^N\| &\leq g\|w^{N-1}\| + E(\Delta t, \Delta x) \leq (1+g+g^2+\dots+g^{N-1})E(\Delta t, \Delta x). \end{aligned}$$

If $g \leq 1$, we find $\|w^N\| \leq NE(\Delta t, \Delta x)$. However, since $g \leq 1$, we know $|1-2Dr|+2Dr \leq 1$. This says $|1-2Dr| \leq 1-2Dr$ which implies $Dr \leq 1/2$. Therefore, there is a relationship between Δt and Δx ; $\Delta t \leq (\Delta x)^2/D$. Thus, the equation $g = |1-2Dr|+2Dr$ becomes $g = 1-2Dr+2Dr = 1$ and g is never actually less than 1. Hence, since $N \approx L/\Delta x$ and , we have

$$\|w^N\| \leq \frac{L}{\Delta x}((\Delta t)^2 + \Delta t(\Delta x)^2) = \frac{LC}{\Delta x} \left(\frac{(\Delta x)^4}{D^2} + \frac{(\Delta x)^4}{D} \right) = \frac{LC(1+D)}{D^2}(\Delta x)^3$$

which goes to zero as Δx goes to zero. Hence, in this case, the recursive equation has a well behaved error and the solution to the recursion equation approaches the solution to the diffusion equation. which is finite. We will say our finite difference approximation is *stable* if the relationship $Dr \leq 1/2$ is satisfied.

Homework

Exercise 14.3.6**Exercise 14.3.7****Exercise 14.3.8****Exercise 14.3.9****Exercise 14.3.10**

Note we can arrive at this result a little faster by assuming ignoring the error $E(\Delta t, \Delta x)$ and looking at the solutions to the pure discrete recursion system only. We assume these have the form $w_{n,m} = g^m e^{in\beta}$ for a positive g , where recall $e^{i\theta} = \cos(\theta) + i \sin \theta$. Then, we find

$$\begin{aligned} g^{m+1} e^{in\beta} &= g^m e^{in\beta} + Dr \left(g^m e^{i(n+1)\beta} + g^m e^{i(n-1)\beta} - 2g^m e^{in\beta} \right) \\ &= g^m e^{in\beta} \left(1 + Dr \left(e^{i\beta} + e^{-i\beta} - 2 \right) \right) \end{aligned}$$

But $2 \cos(\beta) = e^{i\beta} + e^{-i\beta}$ and so we have

$$g^{m+1} e^{in\beta} = g^m e^{in\beta} \left(1 + 2Dr(\cos(\beta) - 1) \right)$$

Dividing, we have

$$g = 1 + 2Dr(\cos(\beta) - 1).$$

This is the *multiplier* that is applied at each time step. We can do the same analysis to show that solutions of the form $w_{n,m} = g^m e^{-in\beta}$ have the same multiplier. Then combinations of these two complex solutions lead to the usual two real solutions, $\phi_{n,m} = g^m \cos(n\beta)$ and $\psi_{n,m} = g^m \sin(n\beta)$. So the use of the complex form of the assumed solution is a useful way to obtain the multiplier g with minimal algebraic complications.

To ensure the error does not go to infinity, we must have $0 < g < 1$. We want

$$0 < 1 - 2Dr(1 - \cos(\beta)) < 1$$

Since $1 - 2Dr(1 - \cos(\beta)) < 1$ always, the top inequality is satisfied. We also know $1 - 2Dr(1 - \cos(\beta)) > 1 - 2Dr$, so we can satisfy the bottom inequality if we choose r so that $0 < 1 - 2Dr < 1 - 2Dr(1 - \cos(\beta))$. This implies $Dr < 1/2$. Thus, stability is not guaranteed unless

$$D\Delta t < (\Delta x)^2.$$

14.3.2 Homework**Exercise 14.3.11****Exercise 14.3.12**

Exercise 14.3.13**Exercise 14.3.14****Exercise 14.3.15**

14.4 Implementing The Diffusion Equation Finite Difference Approximation

We can now implement the finite difference scheme for a typical diffusion/ heat equation model. This is done in the function NumericalHeatDirichlet which uses the arguments dataf, the name of the data $f(x)$ on the bottom edge, gamma, the diffusion constant D , L and T and the dimension of the rectangle $\mathcal{D} = [0, L] \times [0, T]$. Stability implies that Δt can at most be $1/(2\gamma)(\delta x)^2$ and we allow the use of a fraction of that maximum by using the parameter deltfrac to determine Δt via the equation $\Delta t = \text{deltfrac } 1/(2\gamma)\delta x)^2$. Finally, we set the desired Δx using the parameter delx. Here are the details. First we calculate the desired Δt .

Listing 14.11: **Find Δt**

```
1 % calculate delt
delt = .5*deltfrac*delx^2/gamma
```

We then find the number of time and space steps and set up the x and t data points for the grid.

Listing 14.12: **Setup time and space grid data points**

```
% calculate N and M
N = round(L/delx);
M = round(T/delt);

% setup linspaces
x = linspace(0,L,N+1);
t = linspace(0,T,M+1);
```

We then implement the finite difference scheme using a double for loop construction.

Listing 14.13: **Implement Finite Difference Scheme**

```
% setup V
V = zeros(N+1,M+1);
%
% setup V(n,1) = data function (n)
for n = 1:N+1
    V(n,1) = dataf(x(n));
end
%
% setup r value
r = delt/(delx^2);

% find numerical solution to the heat equation
```

```

13 for m = 1:M
    for n = 2:N
        V(n,m+1) = gamma*r*V(n+1,m)+(1-2*gamma*r)*V(n,m) + gamma*r*V(n-1,m)
    end
end

```

Finally, we plot the solution surface.

Listing 14.14: **Plot Solution Surface**

```

% setup surface plot
[X,Time] = meshgrid(x,t);
3 mesh(X,Time,V,'EdgeColor','blue');
 xlabel('x axis');
 ylabel('t axis');
 zlabel('Solution');
 title('Heat Equation on Square');

```

The full code is below.

Listing 14.15: **NumericalHeatDirichlet.m**

```

function NumericalHeatDirichlet(dataf, gamma, L, T, deltfrac, delx)
%
3 % dataf is our data function
% gamma is the diffusion coefficient
% L is the x interval
% T is the time interval
% deltfrac is how much of the maximum del t to use
8 %
% delx is our chosen delta x
% we calculate delt = .5*gamma*(delx^2);
% N = round(L/delx)
% M = round(T/delt)
13 %
% calculate delt
delt = .5*deltfrac*delx^2/gamma

% calculate N and M
18 N = round(L/delx);
M = round(T/delt);

% setup linspaces
x = linspace(0,L,N+1);
23 t = linspace(0,T,M+1);

% setup V
V = zeros(N+1,M+1);

28 % setup V(n,1) = data function (n)
for n = 1:N+1

```

```

V(n,1) = dataf(x(n));
end

33 % setup r value
r = delt/(delx^2);
% find numerical solution to the heat equation
for m = 1:M
    for n = 2:N
        V(n,m+1) = gamma*r*V(n+1,m)+(1-2*gamma*r)*V(n,m) + gamma*r*V(n-1,m
            );
    end
end

% setup surface plot
43 [X,Time] = meshgrid(x,t);
mesh(X,Time,V,'EdgeColor','blue');
xlabel('x axis');
ylabel('t axis');
zlabel('Solution');
48 title('Heat Equation on Square');
end

```

Here is a typical use for a heat equation model with diffusion coefficient 0.09 for a space interval of $[0, 5]$ and a time interval of $[0, 10]$ using a space step size of 0.2.

Listing 14.16: **Generating a Finite Difference Heat Equation Solution**

```

1 pulse = @(x) pulsefunc(x,2,.2,100);
NumericalHeatDirichlet(pulse,.09,5,10,.9,.2);
delt = 0.20000

```

This generates the plot we see in Figure 14.4.

14.4.1 Homework

Exercise 14.4.1

Exercise 14.4.2

Exercise 14.4.3

Exercise 14.4.4

Exercise 14.4.5

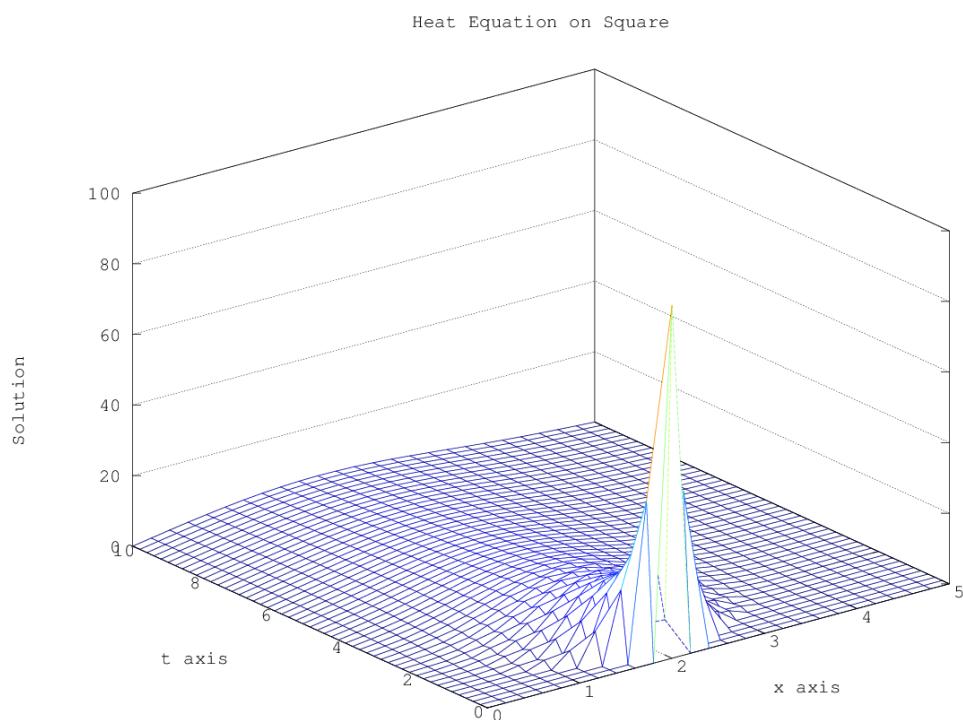
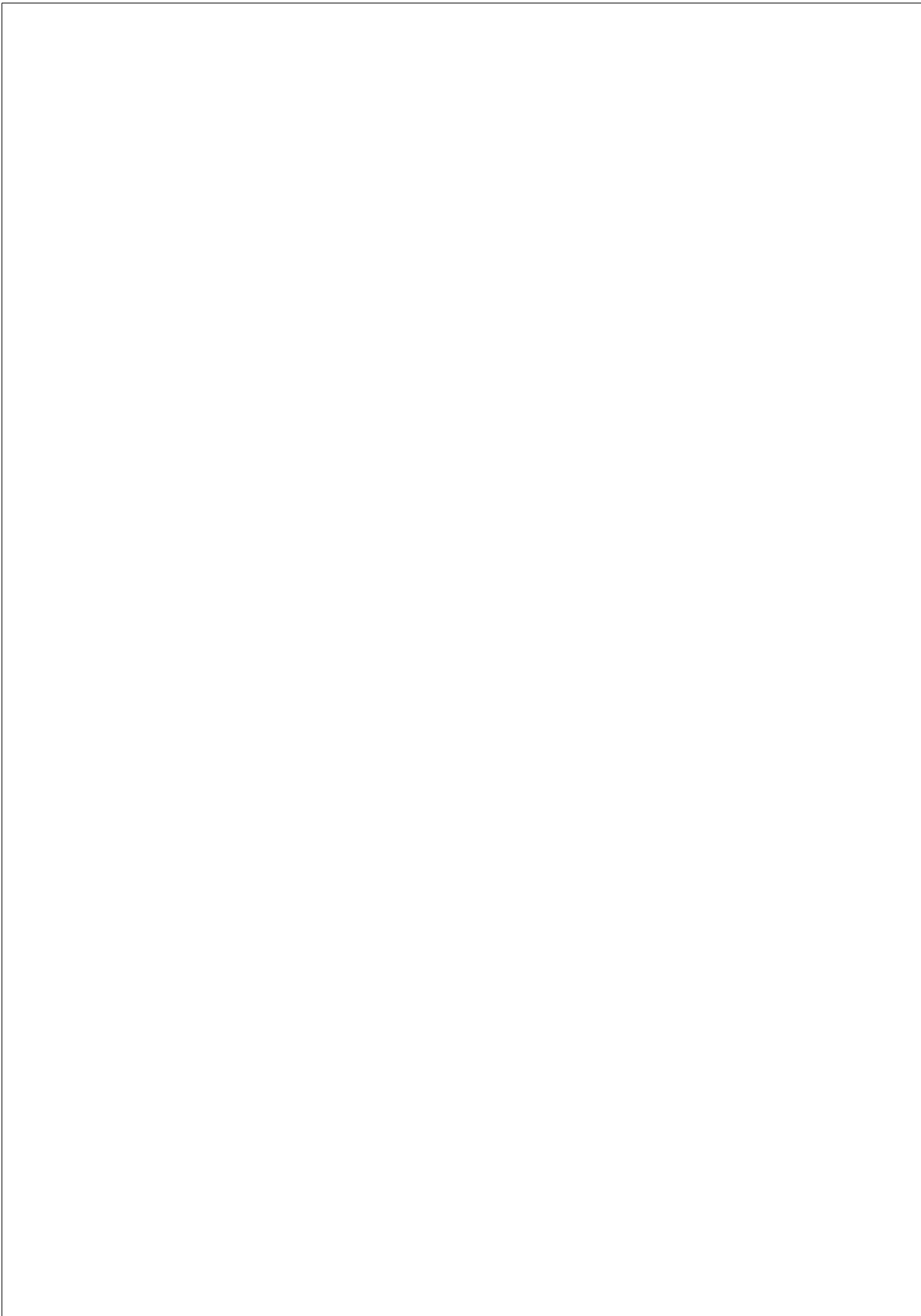


Figure 14.4: Finite Difference Solution To Heat Equation

Part IV

Integration



Chapter 15

Integration in Multiple Dimensions

In this chapter, we are going to discuss Riemann Integration in \mathbb{R}^n . This is a much more interesting thing than the development laid out in (Peterson (8) 2019) as the topology of \mathbb{R} is much simpler than the topology of \mathbb{R}^n for $n > 1$. We wish to do this from a more abstract perspective.

15.1 The Darboux Integral

Let $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be a bounded function on the bounded set A . Enclose the set A inside a bounded hyperrectangle

$$R = [a_1, b_1] \times [a_2, b_2] \times \dots \times [a_n, b_n] = \prod_{i=1}^n [a_i, b_i]$$

There is no need to try to find the tightest bounded box possible. We will show later the definition of integration we develop is independent of this choice. We extend f to R to \hat{f} on R as follows:

$$\hat{f}(x) = \begin{cases} f(x), & x \in A \\ 0, & x \in R \setminus A \end{cases}$$

In \mathbb{R}^2 , we would have what we see in Figure 15.1(a) and in \mathbb{R}^3 , it could look like the image in Figure 15.1(b). The extension \hat{f} allows us to define what we mean by integration without worrying about

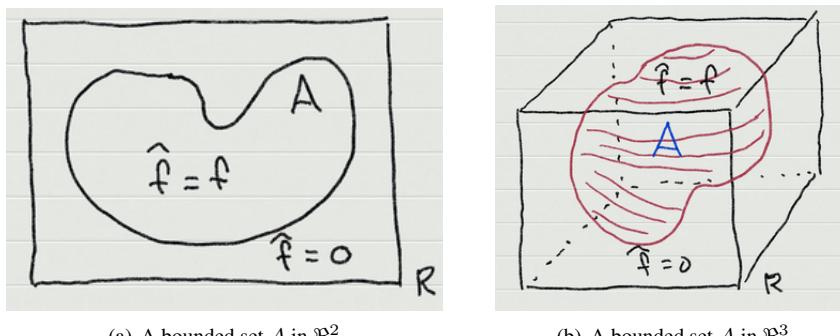


Figure 15.1: Bounded Integration Domains

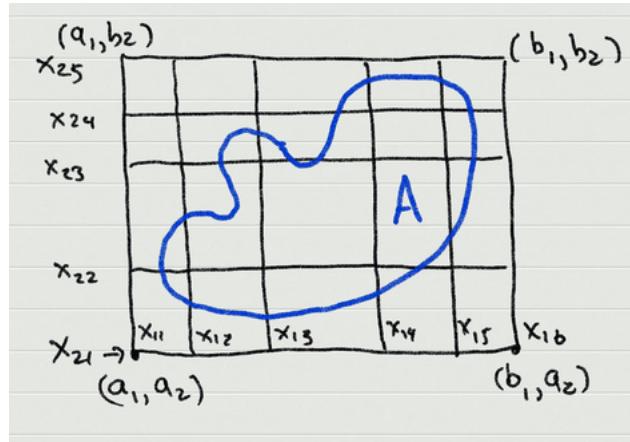


Figure 15.2: A bounded A and a partition P

the boundary of A , ∂A . We begin by slicing R up into a *rectangular* grid to form what are called **Partitions**.

Definition 15.1.1 Partitions of the bounded set A

A partition P of A is defined as follows: For any bounded rectangle R containing A , subdivide each axis using a finite collection of points this way:

$$\begin{aligned} P_1 &= \{x_{11} = a_1 < x_{12} < x_{13} < \dots < x_{1,p_1} = b_1\} \\ P_2 &= \{x_{21} = a_2 < x_{22} < x_{23} < \dots < x_{2,p_2} = b_2\} \\ &\vdots = \vdots \\ P_k &= \{x_{k1} = a_k < x_{k2} < x_{k3} < \dots < x_{k,p_k} = b_k\} \\ &\vdots = \vdots \\ P_n &= \{x_{n1} = a_n < x_{n2} < x_{n3} < \dots < x_{n,p_n} = b_n\} \end{aligned}$$

The partition P is the collection of points in \mathbb{R}^n defined by

$$P = P_1 \times P_2 \times \dots \times P_n$$

Note each P_k is a traditional partition as used in one dimensional integration theory. The sizes of each P_k are determined by the integers p_k and these need not be the same, of course.

Such a P defined a set of grid points in R and the lines of constant $x_{i,j}$ slice R and therefore also A into hyperrectangles. Figure 15.2 shows what this could look like in two dimensions. It is a lot more cluttered to try to draw this in three dimensions! And in higher than three dimensions, we can not come up visualizations at all, so it is best to be able to understand this abstractly. In the figure, we use the partition

$$P = P_1 \times P_2 = \{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\} \times \{x_{21}, x_{22}, x_{23}, x_{24}, x_{25}\}$$

If you are familiar with MatLab code, you should see that if P_1 is represented by one `linspace` command and P_2 , by another, the `meshgrid` command creates the partition P .

For simplicity, let's note a partition P determines a finite number of hyperrectangles of the form

$$S = [x_{1,j_1}; x_{1,j_1+1}] \times [x_{2,j_2}; x_{2,j_2+1}] \times \dots \times [x_{n,j_n}; x_{n,j_n+1}]$$

where we are separating points in each $[\cdot, \cdot]$ pair by semicolons as using a comma becomes very confusing. Here the integers j_1 through j_n range over all the possible choices each P_i allows; i.e. $1 \leq j_i < p_i$. There are a large number of these rectangles, but still finitely many. The partition P will determine $p_1 p_2 \dots p_n$ such rectangles and each rectangle will have a hypervolume

$$V(S) = \frac{(x_{1,j_1+1} - x_{1,j_1})}{\text{length on axis 1}} \cdot \frac{(x_{2,j_2+1} - x_{2,j_2})}{\text{length on axis 2}} \cdots \frac{(x_{n,j_n+1} - x_{n,j_n})}{\text{length on axis } n}$$

In general, it is not important how we label these finitely many rectangles, so we will just say the partition P determines a finite number of rectangles N and label them as S_1, \dots, S_N or $(S_i)_{i=1}^n$. Note a rectangle S_j need not intersect A ! For each rectangle S_i define

$$\begin{aligned} m_i(f) &= \inf_{(x_1, \dots, x_n) \in S_i} \hat{f}(x_1, \dots, x_n) \\ M_i(f) &= \sup_{(x_1, \dots, x_n) \in S_i} \hat{f}(x_1, \dots, x_n) \end{aligned}$$

These numbers are finite as we assume f is bounded on A . Since it is cumbersome use the n -tuple notation for the points in S_i , we will use \mathbf{x} to indicate this instead. Thus

$$m_i(f) = \inf_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x}), \quad M_i(f) = \sup_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x})$$

We can now define the Darboux lower and upper sums for f on A .

Definition 15.1.2 Darboux Lower and Upper Sums for f on A

Given a partition P of A with a bounding hyperrectangle R , P determines N hyperrectangles and the numbers $m_i(f)$ and $M_i(f)$ for $1 \leq i \leq N$. The Darboux Lower Sum associated with f , P , A and R is

$$L(f, P, A, R) = \sum_{i=1}^N m_i(f) V(S_i)$$

and the Darboux Upper Sum associated with f , P , A and R is

$$U(f, P, A, R) = \sum_{i=1}^N M_i(f) V(S_i)$$

The A is often understood from context and so we would usually write $L(f, P, R)$ and $U(f, P, R)$ to indicate these sums might depend on the choice of bounding hyperrectangle R . Since N depends on P , we generally write these as

$$L(f, P, R) = \sum_{i \in P} m_i(f) V(S_i), \quad U(f, P, R) = \sum_{i \in P} M_i(f) V(S_i)$$

Comment 15.1.1 For a given partition P , it is always true that $L(f, P, R) \leq U(f, P, R)$ from the definition of m_i and M_i .

Comment 15.1.2 Since each partition P is dependent on the choice of R , if we had two bounding rectangles with partitions $P(R_1)$ and $P(R_2)$, we would know $R_1 = (R_1 \cap R_2) \cup (R_1 \cap R_2^C) \cup (R_2 \cap R_1^C)$. The partition $P(R_1)$ corresponds to possibly non zero values of m_i and M_i only on $R_1 \cap R_2$. A similar argument shows the partition $Q(R_2)$ corresponds to possibly non zero values of m_i and M_i only on $R_1 \cap R_2$. You can see $P(R_1)$ induces a partition of $R_1 \cap R_2$, $P(R_1 \cap R_2)$ and also $Q(R_2)$ induces a partition of $R_1 \cap R_2$, $Q(R_1 \cap R_2)$. Hence,

$$\begin{aligned} L(f, P(R_1), R_1) &= L(f, P(R_1 \cap R_2), R_1 \cap R_2) \\ L(f, Q(R_2), R_2) &= L(f, Q(R_1 \cap R_2), R_1 \cap R_2) \end{aligned}$$

These sums could be different, of course.

However, by thinking about this just a little different, we can establish the following result.

Theorem 15.1.1 Lower and Upper Darboux Sums are Independent of the choice of Bounding Rectangle

$$L(f, P, R_1) = L(f, P, R_2), \quad U(f, P, R_1) = U(f, P, R_2)$$

Proof 15.1.1

First, note $R_1 = (R_1 \cap R_2) \cup (R_1 \cap R_2^C) \cup (R_2 \cap R_1^C)$. So a partition $P(R_1)$ induces a partition on $(R_1 \cap R_2)$ and there infinitely many ways to extend $P(R_1)$ to the piece $R_2 \cap R_1^C$. Let $Q(R_2)$ be any such partition of R_2 which retains the partition of $P(R_1)$ on $(R_1 \cap R_2)$ and extends it on $(R_2 \cap R_1^C)$. Both of the partition $P(R_1)$ and $Q(R_2)$ corresponds to possibly non zero values of m_i and M_i only on $R_1 \cap R_2$. Let $P(R_1 \cap R_2)$ denote the partition induced on $R_1 \cap R_2$ by $P(R_1)$. Then

$$L(f, P(R_1), R_1) = L(f, P(R_1 \cap R_2), R_1 \cap R_2) L(f, P(R_2), R_2)$$

Hence, the value of the lower sum doesn't depend on the choice of bounding rectangle. A similar argument shows the upper sums are independent of the choice of bounding rectangle also. ■

Comment 15.1.3 From Theorem 15.1.1, we know the lower and upper sums are independent of the choice of bounding rectangle and hence we no longer need to write $L(f, P, R)$ and $U(f, P, R)$. Hence, from now on we will simply write $L(f, P)$ and $U(f, P)$ where the partitions are determined from some choice of bounding rectangle R .

If we add more points to any of the P_i collections that comprise P , we generate a new partition P' called a **refinement** of P .

Definition 15.1.3 A Refinement of a Partition

Given a partition P of A , a refinement P' of P is any partition of A which contains all the original points of P . Hence, P is a refinement of itself. Of course, the interesting refinements are the ones which add at least one point to P . These extra points will then introduce new rectangles.

In Figure 15.3, we see a refinement P' that add one additional points. We also show the new rectangles generated. In this example, you can count the number of additional rectangle and see it is 18. It is very cluttered to draw what happens if you introduce two points and so on, so you have to learn how to understand this from an abstract point of view. The new partition in the figure is

$$P' = [x_{11}, x_{12}, x_{13}, x_{14}, \alpha, x_{15}, x_{16}] \times [x_{21}, x_{22}, x_{23}, \beta, x_{24}, x_{25}]$$

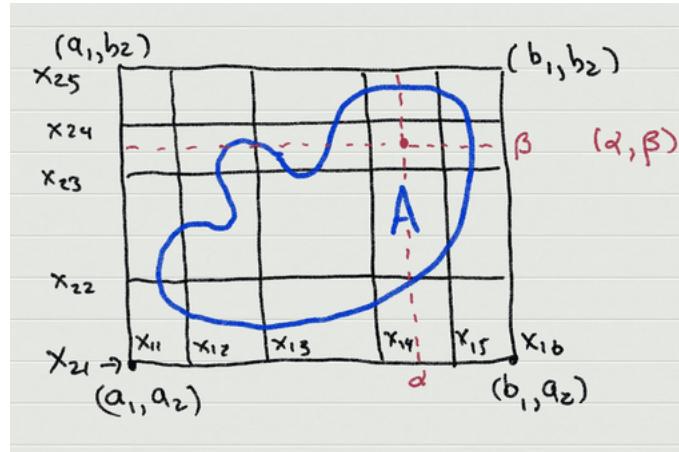


Figure 15.3: One Point (α, β) is added to the partition P

Given two partitions P and Q of A , we often want to merge them into one partition by counting only counting points in the union of the collection of points that comprise P and the points that comprise Q once.

Definition 15.1.4 The Common Refinement of Two Partitions

Let P and Q be two partitions of A . Let U be the union of the collection of points from P and Q . On the i^{th} axis, P and Q determine the collection of points

$$P_i = \{x_{i1}, \dots, x_{ip_i}\} \quad \text{and } Q_i = \{y_{i1}, \dots, y_{iq_i}\}$$

The union of P_i and Q_i gives the collection

$$P_i \cup Q_i = \{x_{i1}, \dots, x_{ip_i}, y_{i1}, \dots, y_{iq_i}\}$$

which determines a new ordering from low to high collection of points on the i^{th} axis $G_i = \{z_{i1}, \dots, z_{ig_i}\}$ which only uses unique entries as all duplicates are removed. The collection of all $G_1 \times \dots \times G_n$ gives the common refinement of P and Q denoted by $P \vee Q$.

Comment 15.1.4 Note $P \vee Q$ is a refinement of both P and Q .

Homework

Exercise 15.1.1 Modify the Riemann Sum MatLab code in (Peterson (8) 2019) for the two dimension case.

Exercise 15.1.2 Modify the Riemann Sum MatLab code in (Peterson (8) 2019) for the two dimension case with uniform partitions and modify the code to draw the Riemann sums as surfaces.

Exercise 15.1.3 For $P = \{-1, -0.7, 0, 0.5, 1.1, 1.8, 2.1, 2.7, 3.0\}$ and $Q = \{4, 4, 3, 5, 1, 5, 6, 6, 1, 6, 8, 7.3, 8\}$, draw the resulting partition $\pi = P \times Q$ and find the maximum area that a rectangle S_i in π can have.

Exercise 15.1.4 For $P = \{-1, -0.7, 0, 0.5, 1.1, 1.8, 2.1, 2.7, 3.0\}$ and $Q = \{4, 4, 3, 5, 1, 5, 6, 6, 1, 6, 8, 7.3, 8\}$, add the point 1.5 to P to create P' . Draw the resulting partition $\pi' = P' \times Q$ and find the maximum area that a rectangle S_i in π' can have.

Exercise 15.1.5 Let A be the circle of radius 2 centered at $(2, 6)$. For $P = \{-1, -0.7, 0, 0.5, 1.1, 1.8, 2.1, 2.7, 3.0\}$ and $Q = \{4, 4, 3, 5, 1, 5, 6, 6, 1, 6, 8, 7.3, 8\}$, $\pi = P \times Q$ is a partition of a bounding rectangle containing this circle. Let $f(x, y) = 2x^2 + 5y^2$. Find the lower and upper sums associated with this partition. Draw the bounding box, the partition and the circle and shade all the rectangles S_i that are outside the circle one color, the rectangles inside the circle another color and the rectangles that intersect the boundary of the circle another color.

Exercise 15.1.6 Let A be the circle of radius 2 centered at $(2, 6)$. For $P = \{-1, -0.7, 0, 0.5, 1.1, 1.8, 2.1, 2.7, 3.0\}$ and $Q = \{4, 4, 3, 5, 1, 5, 6, 6, 1, 6, 8, 7.3, 8\}$, $\pi = P \times Q$ is a partition of a bounding rectangle containing this circle. Let $f(x, y) = 2x^2 + 5y^2$. Add the point 1.5 to P to create P' giving the $\pi' = P' \times Q$. Find the lower and upper sums associated with this partition. Draw the bounding box, the partition and the circle and shade all the rectangles S_i that are outside the circle one color, the rectangles inside the circle another color and the rectangles that intersect the boundary of the circle another color. What effect did adding the extra point have here?

The next question is what happens to the Lower and Upper Darboux sums when refine a partition?

Theorem 15.1.2 Lower and Upper Darboux Sums and Partition Refinements

Let P' be a refinement of P for A with bounding hyperrectangle R . Then

$$L(f, P, R) \leq L(f, P', R), \quad U(f, P, R) \geq U(f, P', R)$$

Of course, these results are independent of the bounding rectangle and could be written

$$L(f, P) \leq L(f, P'), \quad U(f, P) \geq U(f, P')$$

Proof 15.1.2

From now on, we will just call hyperrectangles by the simpler phrase rectangles. We all know these are abstract things once $n > 3$! Let's start by adding one point to P . Let z be the added point. Then z will be in some rectangle S^* determined by P . Note it could be on the boundary of S^* and not necessarily in the interior. Let's organize the indices of P and P' . Look at Figure 15.3 to see additional rectangles in a specific case.

- The indices of P can be put into three disjoint sets:
 1. The singleton index j_* which is the index corresponding to S^* .
 2. The indices which correspond to rectangles due to P which are not broken into two pieces by the addition of the extra point z to P . The rectangles here are exactly the same as the rectangles generated by P' . Let I denote this set of indices.
 3. The indices which correspond to rectangles which are broken into two pieces. Each of these rectangles S_j splits into two rectangles for P' . Let J denote this set of indices.
- The indices of P' can be put into disjoint sets also.

1. For the index j_* from P , the additional point z causes the creation of 2^n new rectangles $T_{j_*, i}$ for $1 \leq i \leq 2^n$. So $\sum_{i=1}^{2^n} V(T_{j_*, i}) = V(S^*)$. Also, since

$$\inf_{\mathbf{x} \in T_{j_*, i}} \hat{f}(\mathbf{x}) \leq \inf_{\mathbf{x} \in S^*} \hat{f}(\mathbf{x})$$

we have

$$m_{\text{for } T_{j_*, i}}(f) \geq m_{\text{for } S^*}$$

which tells us

$$\sum_{i=1}^{2^n} m_{\text{for } T_{j_*,i}}(f) V(T_{j_*,i}) \geq m_{\text{for } S^*} \sum_{i=1}^{2^n} V(T_{j_*,i}) = m_{j_*} V(S^*).$$

2. The rectangles due to P' that are the same as the rectangles due to P that do not come from adding the extra point give a set of indices we call U . Let the infimum values here be called m'_i to distinguish them from infimum values m_j due to P and let the rectangles be called S'_i for the same reason. Then we have

$$\sum_{i \in U} m'_i(f) V(S'_i) = \sum_{j \in I} m_j(f) V(S_j)$$

3. The remaining indices correspond to the pairs of rectangles that come from the splitting due to the addition of the extra point for all rectangles from P except S^* . Call these indices V . For each rectangle, S_j from the index set J from P , we get two new rectangles: $S_j = H_{j,1} \cup H_{j,2}$ and $V(S_j) = V(H_{j,1}) + V(H_{j,2})$. Also,

$$m_{\text{for } H_{j,i}}(f) \geq m_{\text{for } S_j}(f)$$

and so

$$\begin{aligned} & \sum_{j \in J} (m_{\text{for } H_{j,1}}(f) V(H_{j,1}) + m_{\text{for } H_{j,2}}(f) V(H_{j,2})) \\ & \geq \sum_{j \in J} m_{\text{for } S_j}(V(H_{j,1}) + V(H_{j,2})) = \sum_{j \in J} m_j V(S_j). \end{aligned}$$

So the lower sums are

$$\begin{aligned} L(f, P, R) &= \sum_{j \in P} m_j(f) V(S_j) = m_*(f) V(S^*) + \sum_{j \in I} m_j(f) V(S_j) + \sum_{j \in J} m_j(f) V(S_j) \\ L(f, P', R) &= \sum_{j \in P'} m'_j(f) V(S'_j) \geq m_*(f) V(S^*) + \sum_{j \in I} m_j(f) V(S_j) + \sum_{j \in J} m_j(f) V(S_j) \\ &= L(f, P, R) \end{aligned}$$

This shows the result is true for the addition of one extra point into one rectangle due to P .

The next step is an induction argument on the number of points added to one rectangle of P . Assume we have added k points to the rectangle S^* and that the result holds. Now add one more point to S^* . The partition we get by adding n points to S^* is what we call the partition Q and the partition we get by adding the $k+1$ point is the partition Q' . The argument above still works and so $L(f, Q', R) \geq L(f, Q, R)$.

Now we use induction again but this time in the case where we add a finite number of points to more than one rectangle. We assume we have added a finite number of points to k rectangles of P . Now add a finite number of points to a $k+1$ rectangle. Call the partition we get from adding a finite number of points to k rectangles S and the new partition we get by adding a finite number of points to the $k+1$ rectangle S' . Then the argument from the previous step works and we have $L(f, S', R) \geq L(f, S, R)$.

Of course, a very similar argument works for the other case and we prove $U(f, P', R) \leq U(f, P, R)$.

■

There are many consequences of this result and it allows us to define Darboux Integration.

Theorem 15.1.3 Lower Darboux Sums are always less than Upper Darboux Sums

Let P and Q be partitions of A for any choice of bounding rectangle R . Then $L(f, P) \leq U(f, Q)$.

Proof 15.1.3

Let $P \vee Q$ be the common refinement of P and Q . Then, we have

$$L(f, P) \leq L(f, P \vee Q) \leq U(f, P \vee Q) \leq U(f, Q)$$

■

Theorem 15.1.3 immediately implies that for any fixed partition Q

$$\sup_P L(f, P) \leq U(f, Q)$$

Further, since $\sup_P L(f, P)$ is a lower bound for all $U(f, Q)$ not matter what Q is, we see

$$\sup_P L(f, P) \leq \inf_Q U(f, Q)$$

This leads to the lower and upper Darboux integrals.

Definition 15.1.5 The Lower and Upper Darboux Integral

For a bounded $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}$ where A is bounded the Lower Darboux Integral of f over A is

$$\underline{DI}(f, A) = \sup_P L(f, P)$$

and the Upper Darboux Integral of f over A is

$$\overline{DI}(f, A) = \inf_P U(f, P)$$

Comment 15.1.5 From the definitions, it is clear $\underline{DI}(f, A) \leq \overline{DI}(f, A)$.

When the lower and upper Darboux integral match, we obtain the Darboux integral.

Definition 15.1.6 The Darboux Integral

We say the bounded function $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}$ where A is bounded is Darboux Integrable on A if $\underline{DI}(f, A) = \overline{DI}(f, A)$. We denote this common value by $DI(f, A)$.

We eventually can tie this idea to our usual Riemann Integral but first we need to establish some results.

Theorem 15.1.4 The Riemann Criterion For Darboux Integrability

f is Darboux Integrable on A if and only if for all $\epsilon > 0$, there is a partition P_0 so that $U(f, P) - L(f, P) < \epsilon$ for all refinements P of P_0 .

Proof 15.1.4 \implies

Note the choice of bounding rectangle is not important here so we just assume there is one in the background that is used to determine the partitions. Assume f is Darboux Integrable on A . Then $\underline{DI}(f, A) = \overline{DI}(f, A)$. Using the Infimum and Supremum Tolerance Lemma, given $\epsilon > 0$, there are partitions P_ϵ and Q_ϵ so that

$$\begin{aligned}\overline{DI}(f, A) &\leq U(f, Q_\epsilon) < \overline{DI}(f, A) + \epsilon/2 \\ \underline{DI}(f, A) - \epsilon/2 &< L(f, P_\epsilon) \leq \underline{DI}(f, A)\end{aligned}$$

Let $P_0 = P_\epsilon \vee Q_\epsilon$. Then,

$$U(f, P_0) - L(f, P_0) \leq U(f, Q_\epsilon) - L(f, P_\epsilon) < \overline{DI}(f, A) + \epsilon/2 - \underline{DI}(f, A) + \epsilon/2$$

But f is Darboux Integrable on A and so $\overline{DI}(f, A) - \underline{DI}(f, A) = 0$. Thus, $U(f, P_0) - L(f, P_0) < \epsilon$. Now if P is a refinement of P_0 ,

$$U(f, P) - L(f, P) \leq U(f, P_0) - L(f, P_0) < \epsilon$$

This completes the proof.

 \Leftarrow

Assume $\forall \epsilon > 0$, there is a partition P_0 of A so that $U(f, P) - L(f, P) < \epsilon$ for all refinements P of P_0 . Let $\epsilon > 0$ be given. Then there is a partition P_0^ϵ so that $U(f, P_0^\epsilon) < U(f, P_0) + \epsilon$. Hence,

$$\overline{DI}(f, A) = \inf_P U(f, P) \leq U(f, P_0^\epsilon) < L(f, P_0^\epsilon) + \epsilon$$

But then

$$\overline{DI}(f, A) - \epsilon < L(f, P_0^\epsilon) \leq \underline{DI}(f, A)$$

This implies $0 \leq \overline{DI}(f, A) - \underline{DI}(f, A) < \epsilon$. Since $\epsilon > 0$ is arbitrary, we have $\overline{DI}(f, A) = \underline{DI}(f, A)$. Hence f is Darboux Integrable on A . \blacksquare

Homework

Exercise 15.1.7 Let A be the circle of radius 1.5 centered at $(2, 6)$. For $P = \{-1, -0.7, 0, 0.5, 1.1, 1.8, 2.1, 2.7, 3.0\}$ and $Q = \{4, 4, 3, 5, 1, 5, 6, 6, 1, 6, 8, 7.3, 8\}$, $\pi = P \times Q$ is a partition of a bounding rectangle containing this circle. Let $f(x, y) = 3x^2 + 2y^2$. Add the point 1.5 to P to create P' giving the refinement $\pi' = P' \times Q$. Find the lower and upper sums associated with this partition. Draw the bounding box, the partition and the circle and shade all the rectangles S_i that are outside the circle one color, the rectangles inside the circle another color and the rectangles that intersect the boundary of the circle another color. What effect did adding the extra point have here?

Exercise 15.1.8 Let A be the circle of radius 1.5 centered at $(2, 6)$. For $P = \{-1, -0.7, 0, 0.5, 1.1, 1.8, 2.1, 2.7, 3.0\}$ and $Q = \{4, 4, 3, 5, 1, 5, 6, 6, 1, 6, 8, 7.3, 8\}$, $\pi = P \times Q$ is a partition of a bounding rectangle containing this circle. Let $f(x, y) = 3x^2 + 2y^2$. Add the point 1.5 to P to create P' and add the point 5.8 to Q to create Q' giving the refinement $\pi' = P' \times Q'$. Find the lower and upper sums associated with both the π and π' partition. Draw the bounding box, the partition and the circle and shade all the rectangles S_i that are outside the circle one color, the rectangles inside the circle another color and the rectangles that intersect the boundary of the circle another color for both partitions. What effect did adding the extra two points have here?

Exercise 15.1.9 Let $f(x, y, z) = c$ for some constant c . What are the lower and upper sums of f for any set A and partition P of a rectangle R containing A ?

Exercise 15.1.10 Let A be the sphere of radius 1.5 centered at $(2, 6, 0.5)$. For $P = \{-1, -0.7, 0, 0.5, 1.1, 1.8, 2.1, 2.7, 3.0\}$ and $Q = \{4, 4, 3, 5, 1, 5, 6, 6, 1, 6, 8, 7.3, 8\}$, and $S = \{-2, -1.5, -1, 0, 0.6, 1.7, 2\}$. $\pi = P \times Q \times S$ is a partition of a bounding cube containing this circle. Let $f(x, y) = 3x^2 + 2y^2 + 4z^2$. Add the point 1.6 to P to create P' and add the point 5.9 to Q to create Q' and the point 0.3 to S to create S' giving the refinement $\pi' = P' \times Q' \times S'$. Find the lower and upper sums associated with both the π and π' partition. Draw the bounding cube, the partition and the sphere and shade all the cubes S_i that are outside the sphere one color, the cubes inside the sphere another color and the cubes that intersect the boundary of the sphere another color for both partitions. What effect did adding the extra three points have here?

Next, we prove an approximation result. But first a definition.

Definition 15.1.7 The Norm of a Partition

For the partition P of A , P determine N rectangles of form $S_k = \prod_{i=1}^n [a_i^k, b_i^k]$. For each rectangle, we can compute $d_k = \max_{1 \leq i \leq n} |b_i^k - a_i^k|$. The norm of P is

$$\|P\| = \max_k (\max_{1 \leq i \leq n} |b_i^k - a_i^k|) = \max_k (d_k)$$

Note, for a given rectangle S_k ,

$$\begin{aligned} V(S) &= \prod_{1 \leq i \leq n} |b_i^k - a_i^k| \leq \prod_{1 \leq i \leq n} d_k \\ &\leq (\max_k d_k)^n = \|S\|^n \end{aligned}$$

Comment 15.1.6 Suppose you were in 2D and you fixed the partition on the x_2 axis but let the partitioning of the x_1 axis get finer and finer. Then we would have $V(S) \rightarrow 0$ even though $\|S\|$ would not go to zero as the maximum axis distance is determined by the fixed partition of the x_2 axis.

Theorem 15.1.5 Approximation of the Darboux Integral

If f is Darboux Integrable on A , then given a sequence of partitions (P_n) with $\|P_n\| \rightarrow 0$, we have

$$U(f, P_n) \downarrow DI(f, A), \quad L(f, P_n) \uparrow DI(f, A),$$

Proof 15.1.5

We will only do the upper sum case and leave the other one to you. Let $\epsilon > 0$ be given. Then there is a partition P_ϵ so that $U(f, P_\epsilon) \geq \overline{DI}(f, A)$ and $U(f, P_\epsilon) \leq \overline{DI}(f, A) + \epsilon/2$. But f is Darboux integrable on A so we have

$$DI(f, A) \geq U(f, P_\epsilon) < DI(f, A, P) + \epsilon/2$$

Let $Q_n = P_n \vee P_\epsilon$. P_ϵ determines rectangles S_k of the form $\prod_{i=1}^n [a_i^k, b_i^k]$. Let

$$e_k = \min_{1 \leq i \leq n} (b_i^k - a_i^k), \quad e = \min_k (e_k)$$

For the tolerance $\xi = e/2$, there is an N so that if $n > N$, $\|P_n\| < e/2$. Pick any P_n with $n > N$. Any point in a rectangle S determined by P_ϵ is a corner point. If the rectangle S was given by $\prod[a_i^k, b_i^k]$, then the minimum distance between two points given by S is the minimum edge distance: i.e. $\min_{1 \leq i \leq n} (b_i^k - a_i^k) = e_k \geq e$. For a rectangle T given by P_n for $n > N$, if it is given by $\prod[c_i, d_i]$, the maximal distance between two points in T is $\max_{1 \leq i \leq n} (d_i - c_i) = \|P_n\| < e/2$. So if T contained two points x and y of some rectangle S of P_ϵ , letting $d(x, y)$ denote the distance between the points, we would have

$$d(x, y) < e/2 \text{ and } d(x, y) > e$$

This is not possible. Hence, for $n > N$, the rectangles determined by $P_n \vee P_\epsilon$ contain at most one point of P_ϵ .

For a rectangle S containing one point of P_ϵ , we see this single point divides S into 2^n pieces T_k . Then we have a term of this form

$$\sup_{x \in S} f(x) V((S)) - \sum_{k=1}^{2^n} (\sup_{x \in T_k} f(x)) V(T_j)$$

We can overestimate this term as

$$\begin{aligned} & \sup_{x \in S} f(x) V((S)) - \sum_{k=1}^{2^n} (\sup_{x \in T_k} f(x)) V(T_j) \\ & \leq \sup_{x \in S} f(x) V((S)) + \sup_{x \in S} f(x) \sum_{k=1}^{2^n} V(T_j) = 2 \sup_{x \in S} f(x) V((S)) \end{aligned}$$

Let $B = \sup_{x \in A} |f(x)|$. Then, for this rectangle

$$\sup_{x \in S} f(x) V(S) - \sum_{k=1}^{2^n} (\sup_{x \in T_k} f(x)) V(T_j) \leq 2BV(S)$$

We can do this for each point in P_ϵ and its associated rectangle S . Let M be the number of points in P_ϵ . Overestimate each $V(S)$ by

$$V(S) \leq \max_{S \text{ determined by } P_n} V(S)$$

Now consider $U(f, P_n) - U(f, P_n \vee P_\epsilon)$. The rectangles from P_n that do not contain a point of P_ϵ occur in both sums and so these contributions zero out. What is left is the sum over the rectangles that contain points of P_ϵ which we have overestimated in the calculations above. Thus

$$U(f, P_n) - U(f, P_n \vee P_\epsilon) \leq BM \max_{S \text{ determined by } P_n} V(S)$$

We assume $\|P_n\| \rightarrow 0$ so we also know $\max_{S \text{ determined by } P_n} V(S) \rightarrow 0$. Hence, there is \hat{N} so that

$$\max_{S \text{ determined by } P_n} V(S) \leq \frac{\epsilon}{4M(B+1)}$$

We conclude if $n > \max(N, \hat{N})$, then $U(f, P_n) - U(f, P_n \vee P_\epsilon) < 2BM \frac{\epsilon}{4M(B+1)} \leq \epsilon/2$. We know

$$DI(f, A) \leq U(f, P_n \vee P_\epsilon) \leq U(f, P_\epsilon) < DI(f, A) + \epsilon/2$$

and so

$$\begin{aligned} U(f, P_n) - DI(f, A) &= U(f, P_n) - U(f, P_n \vee P_\epsilon) + U(f, P_n \vee P_\epsilon) - DI(f, A) \\ &< \epsilon/2 + \epsilon/2 < \epsilon \end{aligned}$$

if $n > \max(N, \hat{N})$. This shows $U(f, P_n) \downarrow DI(f, A)$. A similar argument shows $L(f, P_n) \uparrow DI(f, A)$. ■

15.1.1 Homework

Exercise 15.1.11

Exercise 15.1.12

Exercise 15.1.13

Exercise 15.1.14

Exercise 15.1.15

15.2 The Riemann Integral in n Dimensions

There is another way to develop integration in \Re^n . We have gone through this carefully in \Re in (Peterson (8) 2019) and what we do in \Re^n is quite similar.

Let $f : A \subset \Re^n \rightarrow \Re$ be a bounded function on the bounded set A . Let R be any bounding rectangle that contains A and let P be any partition of A based on R . As before, this choice of bounding rectangle is immaterial and we simply choose one that is useful. Let S_1, \dots, S_N be the rectangles determined by P . An **in between** or **evaluation** set for P is any collection of points z_1, \dots, z_N where $z_i \in S_i$ for the rectangle S_i determined by P . Such an **evaluation** set is denoted by σ .

Definition 15.2.1 The Riemann Sum

The Riemann sum of f over A for partition P and evaluation set σ is denoted by $S(f, \sigma, P)$ where

$$S(f, \sigma, P) = \sum_{i=1}^N \hat{f}(z_i) V(S_i)$$

where S_1, \dots, S_N are the rectangles determined by P and $z_i \in \sigma$. For convenience, we typically simply write $S(f, \sigma, P) = \sum_{S_i \in P} f(z_i) V(S_i)$. It is clear we have

$$L(f, P) \leq S(f, \sigma, P) \leq U(f, P)$$

The Riemann Integral is then defined like this:

Definition 15.2.2 The Riemann Integral

We say f is Riemann Integrable over A if there is a real number I so that for all $\epsilon > 0$, there is a partition P_ϵ so that

$$|S(f, \sigma, P) - I| < \epsilon$$

for all partitions P that refine P_ϵ and any evaluation set σ for P . We typically use a relaxed notation for this and say $\sigma \subset P$ for short. We denote the number I by the symbol $RI(f, A)$.

There is a wonderful connection between the Riemann integral and the Darboux Integral of f over A .

Theorem 15.2.1 The Equivalence of the Riemann and Darboux Integral

The following are true statements:

1. f is Riemann Integrable on A implies f is Darboux Integrable on A and $RI(f, A) = DI(f, A)$.
2. f is Darboux Integrable on A implies f is Riemann Integrable on A and $DI(f, A) = RI(f, A)$.

Proof 15.2.1

1: We assume f is Riemann integrable on A . Then there is a number I so that given $\epsilon > 0$, there is a partition P_ϵ with

$$I - \epsilon/6 < S(f, \sigma, P) < I + \epsilon/6$$

for all refinements P of P_ϵ and $\sigma \subset P$. Let R denote our arbitrary choice of bounding rectangle for A . Then given any rectangle S determined by P_ϵ , the infimum and supremum tolerance lemma tell us there are points y_S and z_S in S so that

$$\begin{aligned} m_S(f) &\leq f(y_S) < m_S(f) + \frac{\epsilon}{6V(R)} \\ M_S(f) - \frac{\epsilon}{6V(R)} &< f(z_S) \leq M_S(f) \end{aligned}$$

where m_S and M_S are the usual infimum and supremum over S of f .

$$L(f, P_\epsilon) \leq \sum_{S \in P_\epsilon} f(y_S)V(S) < \sum_{S \in P_\epsilon} \left(m_S(f) + \frac{\epsilon}{6V(R)} \right) V(S)$$

or

$$\begin{aligned} L(f, P_\epsilon) &\leq \sum_{S \in P_\epsilon} f(y_S)V(S) < L(f, P_\epsilon) + \frac{\epsilon}{6V(R)} \sum_{S \in P_\epsilon} V(S) \\ &= L(f, P_\epsilon) + \frac{\epsilon}{6V(R)} V(R) = L(f, P_\epsilon) + \frac{\epsilon}{6} \end{aligned}$$

Let $\sigma_{Lower} = \{y_S | S \in P_\epsilon\}$. Then we have

$$L(f, P_\epsilon) \leq S(f, \sigma_{Lower}, P_\epsilon) < L(f, P_\epsilon) + \frac{\epsilon}{6}$$

In a similar way, for $\sigma_{Upper} = \{Z_S | S \in P_\epsilon\}$, we find

$$U(f, P_\epsilon) - \frac{\epsilon}{6} < S(f, \sigma_{Upper}, P_\epsilon) \leq U(f, P_\epsilon)$$

Hence,

$$U(f, P_\epsilon) - L(f, P_\epsilon) < S(f, \sigma_{Upper}, P_\epsilon) - S(f, \sigma_{Lower}, P_\epsilon) + \epsilon/6 + \epsilon/6$$

But we know how close the Riemann sums are to I . We find

$$S(f, \sigma_{Upper}, P_\epsilon) - S(f, \sigma_{Lower}, P_\epsilon) < I + \epsilon/6 - I + \epsilon/6 = \epsilon/3$$

Combining, we conclude

$$U(f, P_\epsilon) - L(f, P_\epsilon) < 2\epsilon/3 < \epsilon$$

This hold for any refinement of P_ϵ . Hence f satisfies the Riemann Criterion and so f is Darboux Integrable on A .

It remains to show $DI(f, A) = RI(f, A)$. Here are the details. Let $\epsilon > 0$ be given. From the definition of $\underline{DI}(f, A)$ and $\overline{DI}(f, A)$ there are partitions U_ϵ and V_ϵ so that

$$\begin{aligned} U(f, U_\epsilon) &< \overline{DI}(f, A) + \epsilon/4 = DI(f, A) + \epsilon/4 \\ L(f, V_\epsilon) &> \underline{DI}(f, A) - \epsilon/4 = DI(f, A) - \epsilon/4 \end{aligned}$$

Thus for the refinement $U_\epsilon \vee V_\epsilon$, we have

$$\begin{aligned} DI(f, A) - \epsilon/4 &< L(f, V_\epsilon) \leq L(f, U_\epsilon \vee V_\epsilon) \leq U(f, U_\epsilon \vee V_\epsilon) \\ &\leq U(f, U_\epsilon) < DI(f, A) + \epsilon/4 \end{aligned}$$

Also, since f is Riemann Integrable, there is a partition P_ϵ so that

$$|S(f, \sigma, P) - RI(f, A)| < \epsilon/4$$

for all refinements P of P_ϵ and any $\sigma \subset P$. Let $Q_\epsilon = U_\epsilon \vee V_\epsilon \vee P_\epsilon$. Then $|S(f, \sigma, Q_\epsilon) - RI(f, A)| < \epsilon/4$ and

$$\begin{aligned} DI(f, A) - \epsilon/4 &< L(f, U_\epsilon \vee V_\epsilon) \leq L(f, Q_\epsilon) \leq S(f, \sigma, Q_\epsilon) \\ &\leq U(f, Q_\epsilon) \leq U(f, U_\epsilon \vee V_\epsilon) < DI(f, A) + \epsilon/4 \end{aligned}$$

So

$$\begin{aligned} |RI(f, A) - DI(f, A)| &\leq |RI(f, A) - S(f, \sigma, Q_\epsilon)| + |S(f, \sigma, Q_\epsilon) - DI(f, A)| \\ &< \epsilon/4 + \epsilon/4 < \epsilon \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, we see $RI(f, A) = DI(f, A)$.

2: We now assume f is Darboux Integrable on A . Then, just as we argued in the last part of the proof of **1**, we see there is a partition $U_\epsilon \vee V_\epsilon$ so that

$$\begin{aligned} DI(f, A) - \epsilon &< L(f, U_\epsilon \vee V_\epsilon) \leq L(f, Q) \\ &\leq S(f, \sigma, Q) \leq U(f, Q) \\ &\leq U(f, U_\epsilon \vee V_\epsilon) < DI(f, A) + \epsilon \end{aligned}$$

for any refinement Q of $U_\epsilon \vee V_\epsilon$ and $\sigma \subset Q$. We conclude $|S(f, \sigma, Q) - DI(f, A)| < \epsilon$ for any refinement Q of $U_\epsilon \vee V_\epsilon$ and $\sigma \subset Q$. This says f is Riemann Integrable on A with value $RI(f, A) = DI(f, A)$. \blacksquare

Next, we prove the approximation result for Riemann Integration.

Theorem 15.2.2 Approximation of the Riemann Integral

If f is Riemann Integrable on A , then given a sequence of partitions (P_n) with $\|P_n\| \rightarrow 0$, we have $S(f, \sigma_n, P_n) \rightarrow RI(f, A)$ for any sequence $\sigma_n \subset P_n$.

Proof 15.2.2

We know $L(f, P_n) \leq S(f, \sigma_n, P_n) \leq U(f, P_n)$. Since f is also Darboux Integrable by the equivalence theorem, we know $DI(f, A) = RI(f, A)$ and $U(f, P_n) \downarrow DI(f, A)$ and $L(f, P_n) \uparrow DI(f, A)$. So $S(f, \sigma_n, P_n) \rightarrow DI(f, A) = RI(f, A)$. \blacksquare

Example 15.2.1

Example 15.2.2

15.2.1 Homework

Exercise 15.2.1

Exercise 15.2.2

Exercise 15.2.3

Exercise 15.2.4

Exercise 15.2.5

15.3 Volume Zero and Measure Zero

Now that we have moved explicitly into \mathbb{R}^n , you should see the idea of length, area and volume and so forth seem hazily defined. In (Peterson (8) 2019), one of the projects is the construction of Cantor type sets and the size of these sets can only be approached abstractly through the content of a set. We also developed a version of Lebesgue’s Theorem which tells us a function is Riemann integrable if and only if its set of discontinuities is of content zero. Now we want to extend this discussion to more dimensions. Let’s go back to the beginning first.

15.3.1 Measure Zero

What is the length of a single point? Note given any positive integer n , $x \in \overline{B(1/n, x)} = [x - 1/n, x + 1/n]$ and this interval has length $2/n$. Since n is arbitrary, it seems reasonable to define the length of $\{x\}$ to be

$$\text{length } \{x\} = \lim_{n \rightarrow \infty} 2/n = 0$$

What if we had a finite number of distinct points $\{x_1, \dots, x_p\}$? Since the points are distinct, there is a N_0 so that

$$x_1 \in [x_1 - 1/n, x_1 + 1/n], \quad x_2 \in [x_2 - 1/n, x_2 + 1/n],$$

$$x_p \in [x_p - 1/n, x_p + 1/n]$$

with all these intervals disjoint when $n > N_0$. Hence,

$$\text{length } \{x_1, \dots, x_p\} \leq 2/n + \dots + 2/n = 2p/n$$

and as $n \rightarrow 0$, we have $\text{length } \{x_1, \dots, x_p\} = 0$. Let's try and make this more precise for sets in \mathbb{R} and for sets in \mathbb{R}^n also. Also, since the notion of *length* is not quite a normal length, let's start using the term **measure** instead.

Definition 15.3.1 Sets of Measure Zero

A set $A \subset \mathbb{R}^n$ is said to have measure zero if for all $\epsilon > 0$, there is a collection of rectangles (S_i) , countable or infinite, so that $A \subset \cup S_i$ with $\sum V(S_i) < \epsilon$.

Using this definition, we see $\text{measure } \{x\} = 0$ for any $x \in \mathbb{R}$. and $\text{measure } \{x_1, \dots, x_p\} = 0$ for any finite number of points in \mathbb{R} . Here are some more examples.

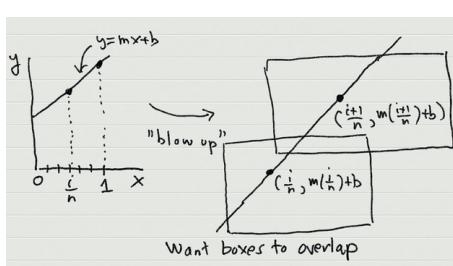
Example 15.3.1 \mathbb{Q} is countable, so it can be enumerated as $\mathbb{Q} = (r_i)_{1 \leq i < \infty}$. Then, $r_i \in [r_1 - \epsilon/2^i, r_1 + \epsilon/2^i]$ and $\mathbb{Q} \subset \cup[r_1 - \epsilon/2^i, r_1 + \epsilon/2^i]$ with $\sum_{i=1}^{\infty} \epsilon/2^{i-1} = \epsilon/2 < \epsilon$. Hence measure $\mathbb{Q} = 0$.

Example 15.3.2 The line $y = mx + b$ has measure zero in \mathbb{R}^2 . We will show this a bit indirectly. First consider the segment of the line for $0 \leq x \leq 1$. Divide $[0, 1]$ into n pieces. Pick n boxes as shown in Figure 15.4(a) so that they overlap. The distance between $(i/n, m(i/n) + b)$ and $((i+1)/n, m((i+1)/n) + b)$ is

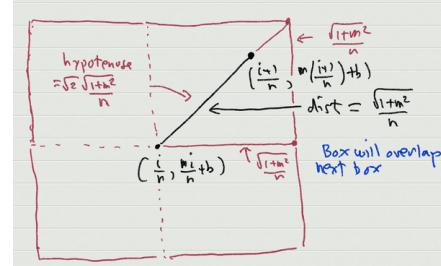
$$d((i/n, m(i/n) + b), ((i+1)/n, m((i+1)/n) + b)) = \sqrt{(1/n)^2 + (m/n)^2} = \sqrt{(m^2 + 1)/n^2}$$

So if we use boxes as shown in Figure 15.4(b), the boxes overlap. We see the line segment of $y = mx + b$ for $0 \leq x \leq 1$ is contained in $\cup_{i=0}^n \text{Box}_i$ and

$$\sum_{i=1}^n V(\text{Box}_i) = \sum_{i=1}^n \frac{4(1+m^2)}{n^2} = \frac{n+1}{n^2} 4(1+m^2)$$



(a) Covering Boxes for $y = mx + b$, $0 \leq x \leq 1$



(b) Overlapping Covering Boxes for $y = mx + b$, $0 \leq x \leq 1$

Figure 15.4: Showing the line $y = mx + b$ has measure zero in \mathbb{R}^2

This goes to zero as $n \rightarrow \infty$, So the line segment of $y = mx + b$ for $0 \leq x \leq 1$ has measure zero.

15.3. VOLUME ZERO AND MEASURE ZERO

307

We can do this for $y = mx + b$ for $1 \leq x \leq 2$, $y = mx + b$ for $2 \leq x \leq 3$ and so on and each of these line segments has measure zero. Now

$$(\text{line segment for } y = mx + b, x \geq 0) = \cup_{p=1}^{\infty} (\text{line segment for } y = mx + b, p - 1 \leq x \leq p)$$

Let $E_p = (\text{line segment for } y = mx + b, p - 1 \leq x \leq p)$. Then

$$(\text{line segment for } y = mx + b, x \geq 0) = \cup_{p=1}^{\infty} E_p$$

Since each E_p has measure zero, given $\epsilon > 0$, there is a collection of rectangles $\{S_{i,p}\}$ so that

$$E_p \subset \cup S_{i,p}, \quad \sum V(S_{i,p}) < \epsilon/2^p$$

Hence,

$$(\text{line segment for } y = mx + b, x \geq 0) \subset \cup_p \cup_i S_{ip}, \quad \sum_p \sum_i V(S_{ip}) < \sum_p \epsilon/2^p < \epsilon$$

This shows measure ($\text{line segment for } y = mx + b, x \geq 0$) = 0. A similar argument shows measure ($\text{line segment for } y = mx + b, x \leq 0$) = 0 also. From this, it follows measure ($y = mx + b$) = 0

The examples above indicate we can prove the following propositions.

Theorem 15.3.1 Finite Unions of Sets of Measure Zero Also Have Measure Zero

If $\{A_1, \dots, A_p\}$ is a finite collection of sets of measure zero, then $\cup_{i=1}^p A_i$ has measure zero also.

Proof 15.3.1

This is left to you. ■

Theorem 15.3.2 Countable Unions of Sets of Measure Zero are Also Measure Zero

If $\{A_i | i \geq 1\}$ is an infinite collection of sets of measure zero, then $\cup_{i=1}^{\infty} A_i$ has measure zero also.

Proof 15.3.2

Again, this is left for you. ■

It is time for a harder example.

Example 15.3.3 Consider the surface $z = x^2 + y^2$ in \mathbb{R}^3 . That is, $f(x, y) = x^2 + y^2$. We will show this is a set of measure zero in \mathbb{R}^3 . Let's consider only a finite piece of it in octant one. Note the projection of the surface to the $x - y$ plane is the circle $x^2 + y^2 = 1$ for $z = 1$. Hence, the square $[-1, 1] \times [-1, 1]$ contains the projection. Divide the square into n^2 uniform pieces of area $1/n^2$. These induce a corresponding subdivision of the disc $\{(x, y) | x^2 + y^2 \leq 1\}$. The corners of a patch P_{ij} in the $x - y$ plane are

$$P_{i,j} = \begin{bmatrix} (i/n, (j+1)/n) & \cdots & ((i+1)/n, (j+1)/n) \\ \vdots & \vdots & \vdots \\ (i/n, j/n) & \cdots & ((i+1)/n, j/n) \end{bmatrix}$$

$$= \begin{bmatrix} (x_i, y_{j+1}) & \cdots & (x_{i+1}, y_{j+1}) \\ \vdots & \vdots & \vdots \\ (x_i, y_j) & \cdots & (x_{i+1}, y_j) \end{bmatrix}$$

and this square is mapped to a patch with corners

$$f(P_{i,j}) = \begin{bmatrix} z_{i,j+1} & \cdots & z_{i+1,j+1} \\ \vdots & \vdots & \vdots \\ z_{i,j} & \cdots & z_{i+1,j} \end{bmatrix}$$

where $z_{pq} = x_p^2 + y_q^2$. Look at the square $[0, 1] \times [0, 1]$. Now $z = x^2 + y^2$ is a convex surface the maximum occurs at the largest i and j for the patch and the minimum is at the smallest i and j . So on the square $P_{i,j}$, the maximum and minimum surface values are

$$\begin{aligned} z_{M,i,j} &= x_{i+1}^2 + y_{j+1}^2 = \left(\frac{i+1}{n}\right)^2 + \left(\frac{j+1}{n}\right)^2 \\ z_{m,i,j} &= x_i^2 + y_j^2 = \left(\frac{i}{n}\right)^2 + \left(\frac{j}{n}\right)^2 \end{aligned}$$

Thus,

$$Z_{M,i,j} - z_{m,i,j} = \frac{2i+1}{n^2} + \frac{2j+1}{n^2}$$

This patch of surface is contained in the box $B_{i,j} = P_{i,j} \times [z_{i,j} - 1/n, z_{M,i,j} + 1/n]$ which has volume

$$V(B_{i,j}) = \frac{1}{n^2}(z_{M,i,j} - z_{m,i,j} + \frac{2}{n}) = \frac{2}{n^3} + \frac{2i+1+2j+1}{n^4}$$

The surface above $[0, 1] \times [0, 1]$ is contained in the union of the boxes $B_{i,j}$ and

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n V(B_{i,j}) &= \sum_{i=1}^n \sum_{j=1}^n \frac{2}{n^3} + \sum_{i=1}^n \sum_{j=1}^n \frac{2i+2j+2}{n^4} \\ &= \frac{2}{n} + \sum_{i=1}^n \left(\frac{2i}{n^3} + \frac{2n(n+1)}{2n^4} + \frac{2n}{n^4} \right) \\ &= \frac{2}{n} + \frac{2n(n+1)}{2n^3} + \frac{2n^2(n+1)}{2n^4} + \frac{2n^2}{n^4} \end{aligned}$$

We see $\sum_{i=1}^n \sum_{j=1}^n V(B_{i,j}) \rightarrow 0$ as $n \rightarrow \infty$. Thus, we conclude the measure of the surface above $[0, 1] \times [0, 1]$ is 0.

A similar argument shows the measure of the surface above $[0, 1] \times [-1, 0]$, above $[-1, 0] \times [0, 1]$ and above $[-1, 0] \times [-1, 0]$ are zero too. Hence the measure of the surface above $[-1, 1] \times [-1, 1]$ is zero. Let Ω_n be the surface above the annular region $\{(x, y) | n^2 \leq x^2 + y^2 \leq (n+1)^2\}$. Thus, $n^2 \leq z \leq (n+1)^2$. Using an argument similar to what we just did, we can show the measure of the surface above this annular region is zero. Note Ω_0 is contained in the surface above $[-1, 1] \times [-1, 1]$ and so the measure of Ω_0 is zero. The entire surface is the union of all the Ω_n for $n \geq 0$ and since each of these are measure zero, the entire surface in \mathbb{R}^2 is measure zero.

Homework

Exercise 15.3.1 Prove $y = x^2$ has measure zero in \mathbb{R}^2 .

Exercise 15.3.2 Prove the surface $z = 2x^2 + 3y^2$ has measure zero in \mathbb{R}^3 .

Exercise 15.3.3 Prove $y = 2x + 5$ has measure zero in \mathbb{R}^2 .

15.3.2 Volume Zero

Do all sets $A \in \mathbb{R}^n$ have associated with them a generalization of volume for a box? Let's see if we can define a suitable notion. The volume of a rectangle in \mathbb{R}^n , $R = \prod_{i=1}^n [a_i^k, b_i^k]$ is given by $V(R) = \prod_{i=1}^n (b_i^k - a_i^k)$. To generalize this idea, let's consider **characteristic functions**.

Definition 15.3.2 The Characteristic Function of a Set

If $A \subset \mathbb{R}^n$, the characterization of A is denoted by $\mathbf{1}_A$ and defined by

$$\mathbf{1}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

Let A be a rectangle in \mathbb{R}^n . Then A is its own bounding rectangle. Let P be any partition of A with associated rectangles S_i . Then, $\inf_{x \in S_i} \mathbf{1}_A(x) = 1$ and $\sup_{x \in S_i} \mathbf{1}_A(x) = 1$ also. Thus,

$$\begin{aligned} L(\mathbf{1}_A, P) &= \sum_{S_i \in P} 1 V(S_i) = V(A) \\ U(\mathbf{1}_A, P) &= \sum_{S_i \in P} 1 V(S_i) = V(A) \end{aligned}$$

Thus, $\mathbf{1}_A$ is Darboux Integrable implying Riemann Integrable over A with value $V(A)$.

It is time to use a different notation for the value of our integrals.

Definition 15.3.3 The Integration Symbol

If $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}$ for bounded A and f bounded is Riemann Integrable on A , the value of the Riemann Integral of f on A will be denoted by $\int_A f$. We sometimes want to remind ourselves explicitly that the integration is done over a subset of \mathbb{R}^n . Since lower and upper sums involves volumes of rectangles in \mathbb{R}^n , we will use the symbol dV_n to indicate this in the integration symbol. Hence, we can also write $RI(f, A) = \int_A f dV_n$.

Comment 15.3.1 Our traditional one calculus integral would be $\int_a^b f(x)dx$ for the Riemann Integrable of the bounded function f on $[a, b]$. We could also write this as $\int_{[a,b]} f dV_1$ which, of course, is a bit much. Just remember there are many ways to represent things. The new notation is better for an arbitrary number of dimensions n and not so great for $n = 1$.

Using this definition, we see $V(A) = \int_A \mathbf{1}_A$ when A is a rectangle in \mathbb{R}^n . This suggest a way to extend the idea of volume to some sets A of \mathbb{R}^n .

Definition 15.3.4 The Volume of a Set

If A is a bounded set in \Re^n and $\mathbf{1}_A$ is integrable, then the volume of A is defined to be $V(A) = \int_A \mathbf{1}_A$. As usual, the function $\mathbf{1}_A$ is extended to any bounded rectangle R by

$$\hat{\mathbf{1}}_A(x) = \begin{cases} \mathbf{1}_A(x), & x \in A \\ 0, & x \in R \setminus A \end{cases}$$

It is easy to show that if a set has volume zero, it is also has measure zero.

Theorem 15.3.3 Volume Zero Implies Measure Zero

If $A \subset \Re^n$ has volume zero, it is also measure zero.

Proof 15.3.3

Since A has volume zero, given $\epsilon > 0$, there is a partition P_ϵ so the $0 \leq U(\mathbf{1}_A, P_\epsilon) < \epsilon$ (remember all the lower sums here are zero!). But $U(\mathbf{1}_A, P_\epsilon) = \sum_{S_i \in P_\epsilon} V(S_i)$, so $A \subset \cup_{S_i \in P_\epsilon} S_i$ and $\sum_{S_i \in P_\epsilon} V(S_i) < \epsilon$. Since $\epsilon > 0$ is arbitrary, this shows A is measure zero. ■

The converse is not true. Let $A = \{(x, y, z) \in [0, 1] \times [0, 1] \times [0, 1] | x, y, z \in \mathbb{Q}\}$. Then

$$\mathbf{1}_A(x) = \begin{cases} 1, & x, y, z \in \mathbb{Q} \cap [0, 1] \\ 0, & \text{otherwise} \end{cases}$$

Let the bounding rectangle be $R = [0, 1] \times [0, 1] \times [0, 1]$. then for any partition P ,

$$\begin{aligned} L(\mathbf{1}_A, P) &= \sum_{S_i \in P} 0 V(S_i) = 0 \\ U(\mathbf{1}_A, P) &= \sum_{S_i \in P} 1 V(S_i) = V(R) = 1 \end{aligned}$$

This is because any $S_i \in P$ contains rational triples. Thus, $\underline{DI}(\mathbf{1}_A) = 0$ and $\overline{DI}(\mathbf{1}_A) = 1$. We conclude $\mathbf{1}_A$ is not integrable and so its volume is not defined. However, the rational triples here are countable and so have measure zero.

We are now ready to consider the question of when a function is Riemann Integrable on A .

15.4 When is a Function Riemann Integrable?

The result we need is called Lebesgue’s Theorem and we already have proven a version of this in (Peterson (8) 2019). But now we want to look at the situation in \Re^n for $n > 1$ in general. First, a new way to look at continuity.

Definition 15.4.1 The Oscillation of a Function

Let $f : A \subset \Re^n \rightarrow \Re$ where A is an open set. The oscillation of f at $x_0 \in A$ is defined to be

$$\omega(f, x_0) = \inf_{B(r, x_0), r > 0} \left(\sup_{x_1, x_2 \in B(r, x_0)} |f(x_1) - f(x_2)| \right)$$

We leave it to you to prove this result.

Theorem 15.4.1 Oscillation is Zero if and only Continuous

Let $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}$ where A is an open set. Then $\omega(f, x_0) = 0$ if and only if f is continuous at x_0 .

Proof 15.4.1

This one's for you! ■

Now on to Lebesgue's Theorem.

Theorem 15.4.2 Lebesgue's Theorem

Let $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}$, A bounded and f bounded on A . Then f is Riemann Integrable on A if and only the set of points where \hat{f} is not continuous is a set of measure zero.

Proof 15.4.2

⇒

We assume the set of discontinuities of \hat{f} has measure zero. Call this set D . For each $\epsilon > 0$, let $D_\epsilon = \{x \mid \omega(\hat{f}, x) \geq \epsilon\}$. Then $D = \cup_{\epsilon>0} D_\epsilon$ and each $D_\epsilon \subset D$ by Theorem 15.4.1. Assume y is a limit point of D_ϵ . If y is an isolated limit point, it is already in D_ϵ . If it is not isolated, there is a sequence $(y_n) \subset D_\epsilon$ with $y_n \neq y$ so that $y_n \rightarrow y$. For any $r > 0$, since $y_n \rightarrow y$, there is N_r so that $y_n \in B(r, y)$. Thus, there is an $s < r$ so that $B(s, y_n) \subset B(r, y)$, i.e. $y \in B(s, y_n)$ for any $n > N_r$. But then

$$\sup_{x_1, x_2 \in B(r, y)} |\hat{f}(x_1) - \hat{f}(x_2)| \geq \sup_{x_1, x_2 \in B(s, y_n)} |\hat{f}(x_1) - \hat{f}(x_2)| \geq \epsilon$$

because $y_n \in D_\epsilon$. Since this is true for all r , we have

$$\inf_{B(r, x_0), r > 0} \left(\sup_{x_1, x_2 \in B(r, y)} |\hat{f}(x_1) - \hat{f}(x_2)| \right) \geq \epsilon$$

which tells us $\omega(\hat{f}, y) \geq \epsilon$ or $y \in D_\epsilon$. Thus, D_ϵ is closed. Since it is inside the bounded set A , it is also bounded. We conclude D_ϵ is compact.

We know D has measure zero. thus, D_ϵ had measure zero as well. Then, there is a collection B_i of open rectangles (the definition uses closed rectangles, but it is easy to expand them a bit to make them open rectangles) so $D_\epsilon \subset \cup_i B_i$ and $\sum_i V(\overline{B_i}) < \epsilon$ where recall $\overline{B_i}$ is the closure of B_i . Since D_ϵ is topologically compact, the open cover (B_i) has a finite subcover $\{B_{i_1}, \dots, B_{i_p}\}$ which we will call $\{B_1, \dots, B_N\}$ by adding in any missing sets B_i as required with $N = i_p$. Hence, $D_\epsilon \subset \cup_{i=1}^N B_i$ and $\sum_{i=1}^N V(\overline{B_i}) < \epsilon$.

The collection $\{B_1, \dots, B_N\}$ of possibly overlapping rectangles determines a partition P_0 of the bounding rectangle we choose for A . Pick a partition P which refines P_0 . The new rectangles we form will either be inside the rectangles already made by P_0 or completely outside them. Thus, the rectangles S_i determined by P can be divided into two pieces with index sets I and J .

$$\begin{aligned} I &= \{i \in \{1, \dots, N\} \mid \exists j \ni S_i \subset B_j\} \\ J &= \{i \in \{1, \dots, N\} \mid S_i \subset (\cup_{j=1}^n B_j)^C\} \end{aligned}$$

Note, for each $j \in J$, S_j is disjoint from D_ϵ . Thus, if $\mathbf{z} \in S_j$, $j \in J$, $\omega(\hat{f}, \mathbf{z}) < \epsilon$. Since $\omega(\hat{f}, \mathbf{z})$ is defined using an infimum, there is $r_z > 0$ so that

$$\omega(\hat{f}, \mathbf{z}) \leq \sup_{\mathbf{x}_1, \mathbf{x}_2 \in B(r_z, \mathbf{z})} |\hat{f}(\mathbf{x}_1) - \hat{f}(\mathbf{x}_2)| < (1/2)(\omega(\hat{f}, \mathbf{z}) + \epsilon)$$

It is easier to use rectangles so let's switch to rectangles. Let $R^\circ(r_z, \mathbf{z})$ be an open rectangle contained in $B(r_z, \mathbf{z})$. Then we can say

$$\omega(\hat{f}, \mathbf{z}) \leq \sup_{\mathbf{x}_1, \mathbf{x}_2 \in R^\circ(r_z, \mathbf{z})} |\hat{f}(\mathbf{x}_1) - \hat{f}(\mathbf{x}_2)| < (1/2)(\omega(\hat{f}, \mathbf{z}) + \epsilon)$$

Let $\hat{\epsilon} = (1/2)(\omega(\hat{f}, \mathbf{z}) + \epsilon) < \epsilon$. Then, we have

$$-\hat{\epsilon} < \hat{f}(\mathbf{x}_1) - \hat{f}(\mathbf{x}_2) < \hat{\epsilon}, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in R^\circ(r_z, \mathbf{z})$$

or

$$\hat{f}(\mathbf{x}_2) - \hat{\epsilon} < \sup_{\mathbf{x}_1 \in R^\circ(r_z, \mathbf{z})} \hat{f}(\mathbf{x}_1) < \hat{f}(\mathbf{x}_2) + \hat{\epsilon}, \quad \forall \mathbf{x}_2 \in R^\circ(r_z, \mathbf{z})$$

implying

$$\sup_{\mathbf{x}_1 \in R^\circ(r_z, \mathbf{z})} \hat{f}(\mathbf{x}_1) \leq \inf_{\mathbf{x}_2 \in R^\circ(r_z, \mathbf{z})} \hat{f}(\mathbf{x}_2) + \hat{\epsilon}$$

We conclude

$$\sup_{\mathbf{x}_1 \in R^\circ(r_z, \mathbf{z})} \hat{f}(\mathbf{x}_1) - \inf_{\mathbf{x}_2 \in R^\circ(r_z, \mathbf{z})} \hat{f}(\mathbf{x}_2) \leq \hat{\epsilon} < \epsilon$$

We know $S_j \subset \cup_{\mathbf{z} \in S_j} R^\circ(r_z, \mathbf{z})$ and since S_j is compact, there is a finite subcover

$$\{R^\circ(r_{z_1}, \mathbf{z}_1), \dots, R^\circ(r_{z_q}, \mathbf{z}_q)\}, \quad S_j \subset \cup_{i=1}^q R^\circ(r_{z_i}, \mathbf{z}_i)$$

The rectangles in the subcover define a partitioning scheme inside S_j . Choose any additional refinement of P so that the new rectangles are all inside some $R^\circ(r_{x_i}, \mathbf{z}_i)$. Call this refinement P' . So for any rectangle E due to this new refinement of P , since it is contained in one of the sets in the subcover, we have

$$\sup_{\mathbf{x}_1 \in E} \hat{f}(\mathbf{x}_1) - \inf_{\mathbf{x}_2 \in E} \hat{f}(\mathbf{x}_2) \leq \hat{\epsilon} < \epsilon$$

Thus,

$$\sum_{E \in P'} \sup_{\mathbf{x} \in E} V(E) - \sum_{E \in P'} \inf_{\mathbf{x} \in E} V(E) \leq \hat{\epsilon} \sum_{E \in P'} V(E) = \hat{\epsilon} V(S_j) < \epsilon V(S_j)$$

We can carry out this process of refinement for each $j \in J$ to get a succession of refinements: $P \rightarrow P'_1 \rightarrow P'_2 \rightarrow \dots \rightarrow P'_M$ where M is the number of indices in J . This gives a partition Q . The indices of Q still split into two pieces I and J as described, but the additional refinements we have introduced allow us to make some estimates for each $S_j \in J$. We can say

$$U(\hat{f}, Q) - L(\hat{f}, P) = \sum_{E_i, i \in I} (\sup_{\mathbf{x} \in E_i} \hat{f}(\mathbf{x})) V(S_i) - \sum_{E_i, i \in I} (\inf_{\mathbf{x} \in E_i} \hat{f}(\mathbf{x})) V(S_i)$$

$$\begin{aligned}
 & + \sum_{E_i, i \in J} (\sup_{\mathbf{x} \in E_i} \hat{f}(\mathbf{x})) V(S_i) - \sum_{E_J, i \in I} (\inf_{\mathbf{x} \in E_i} \hat{f}(\mathbf{x})) |V(S_i)| \\
 & < \sum_{E_i, i \in I} (\sup_{\mathbf{x} \in E_i} \hat{f}(\mathbf{x})) V(S_i) - \sum_{E_i, i \in I} (\inf_{\mathbf{x} \in E_i} \hat{f}(\mathbf{x})) V(S_i) + \epsilon \sum_{E_i, i \in J} V(S_i) \\
 & = \sum_{E_i, i \in I} (\sup_{\mathbf{x} \in E_i} \hat{f}(\mathbf{x}) - \inf_{\mathbf{x} \in E_i} \hat{f}(\mathbf{x})) V(S_i) + \epsilon V(R)
 \end{aligned}$$

where R is the bounding rectangle for A . Now f is bounded on A so there is a constant M so that $|\hat{f}| < M$ on R . Thus,

$$U(\hat{f}, Q) - L(\hat{f}, Q) < 2M \sum_{E_i, i \in I} V(S_i) + \epsilon V(R)$$

But for each index $i \in I$, S_i is contained in some B_j and so $\cup_{i \in I} S_i \subset \cup_{j=1}^N B_j$ and $\sum_{i=1}^N V(B_j) < \epsilon$. We conclude

$$U(\hat{f}, Q) - L(\hat{f}, Q) < 2M\epsilon + \epsilon V(R)$$

This is enough to show f satisfies the Riemann Criterion and so f is integrable on A .

\implies

Now we assume f is integrable on A . The set of discontinuities of \hat{f} is the set

$$D = \{\mathbf{x} \in \mathbb{R}^n \mid \omega(\hat{f}, \mathbf{x}) \neq 0\}$$

We also know $D = \cup_n D_{1/n}$ where $D_{1/n}$ is defined as in the first part of the proof. Since f is integrable, given $\xi > 0$, there is a partition P_ξ so that

$$U(\hat{f}, P_\xi) - L(\hat{f}, P_\xi) = \sum_{S_i \in P_\xi} (\sup_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x}) - \inf_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x})) V(S_i) < \xi$$

for all refinements P of P_ξ . Earlier, we showed $D_{1/n}$ is compact. Let's decompose $D_{1/n}$ in this way.

$$\begin{aligned}
 D_{1/n} &= \{\mathbf{x} \in D_{1/n} \mid \exists i \ni \mathbf{x} \in \partial S_i \in P_\xi\} \iff Q_1^n \\
 &\quad \cup \{\mathbf{x} \in D_{1/n} \mid \exists i \ni \mathbf{x} \in \text{Int}(S_i^\circ) \in P_\xi\} \iff Q_2^n
 \end{aligned}$$

where ∂S_i is the usual notation for the boundary of S_i and S_i° denotes the interior of S_i . We know the measure of an edge of a rectangle is zero and we know countable unions of sets of measure zero are also zero. So we can conclude the measure of Q_1^n must be zero. It remains to see what the measure of Q_2^n is.

If S_i is a rectangle that intersects Q_2^n , then if $\mathbf{x} \in S_i$, there is a point $\mathbf{u} \in S$ and in $D_{1/n}$. Thus,

$$\omega(\hat{f}, \mathbf{u}) = \inf_{B(r, \mathbf{u}), r > 0} \sup_{\mathbf{x}_1, \mathbf{x}_2 \in B(r, \mathbf{u})} |\hat{f}(\mathbf{x}_1) - \hat{f}(\mathbf{x}_2)| \geq (1/n)$$

In particular, this can be rewritten in terms of open rectangles as

$$\omega(\hat{f}, \mathbf{u}) = \inf_{R^\circ(r, \mathbf{u}), r > 0} \sup_{\mathbf{x}_1, \mathbf{x}_2 \in R^\circ(r, \mathbf{u})} |\hat{f}(\mathbf{x}_1) - \hat{f}(\mathbf{x}_2)| \geq (1/n)$$

where $R^\circ(r, \mathbf{u})$ is an open square rectangle with center \mathbf{u} and axis dimensions r . Thus, for an open rectangle containing S , T° , we can say

$$\sup_{\mathbf{x}_1, \mathbf{x}_2 \in T^\circ(r, \mathbf{u})} |\hat{f}(\mathbf{x}_1) - \hat{f}(\mathbf{x}_2)| \geq (1/n)$$

and this implies

$$\sup_{\mathbf{x}_1, \mathbf{x}_2 \in S} |\hat{f}(\mathbf{x}_1) - \hat{f}(\mathbf{x}_2)| \geq (1/n)$$

Then, using the same (complicated!) arguments we used in the first part of this proof, we can say

$$\sup_{\mathbf{x} \in S} \hat{f}(\mathbf{x}) - \inf_{\mathbf{x} \in S} \hat{f}(\mathbf{x}) \geq (1/n)$$

Applying this argument to all such rectangles S that intersect Q_2^n , we find

$$\begin{aligned} (1/n) \sum_{S \cap Q_2^n \neq \emptyset} V(S) &\leq \sum_{S \cap Q_2^n \neq \emptyset} \left(\sup_{\mathbf{x} \in S} \hat{f}(\mathbf{x}) - \inf_{\mathbf{x} \in S} \hat{f}(\mathbf{x}) \right) V(S) \\ &\leq \sum_{S \in P_\xi} \left(\sup_{\mathbf{x} \in S} \hat{f}(\mathbf{x}) - \inf_{\mathbf{x} \in S} \hat{f}(\mathbf{x}) \right) V(S) < \xi \end{aligned}$$

Thus, the collection $V_\xi = \{S \cap Q_2^n \neq \emptyset\}$ is a set of rectangles that cover Q_2^n with $\sum_{S \cap Q_2^n \neq \emptyset} V(S) < n\xi$. Now since the measure of Q_1^n is zero, for $\xi > 0$, there is a collection of rectangles covering Q_1^n , $U_\xi = \{U_{1,\xi}, \dots, U_{p,\xi}\}$ so that $\sum_{i=1}^p V(U_{i,\xi}) < \xi$. The collection $\{S \cap Q_2^n \neq \emptyset\} \cup U_\xi$ is cover for $D_{1/n}$ with $\sum_{S \in U_\xi \cup V_\xi} V(S) < (n+1)\xi$. Finally, letting $\xi = \epsilon/(n+1)$, we have found a collection whose union contains $D_{1/n}$ with $\sum_{S \in U_\xi \cup V_\xi} V(S) < \epsilon$. Since $\epsilon > 0$ is arbitrary because ξ was arbitrary, we see $D_{1/n}$ has measure zero. This tells us D also has measure zero. ■

Whew! That is an intense proof! Compare the approach we have used here to the one we used in \Re in (Peterson (8) 2019). This one had to be able to handle rectangles in spaces of more than one dimension which is why it is much harder. Of course, this argument works just fine for \Re although the rectangles degenerate to just closed intervals.

Example 15.4.1 Consider the curve $x^2 + y^2 = 1$ in \Re^2 . Let's show it has measure zero in \Re^2 . We'll divide the curve into the four quadrants and just do the argument for quadrant one. In quadrant one, we have $y = \sqrt{1 - x^2}$ with $0 \leq x \leq 1$. Divide $[0, 1]$ into n uniform pieces and choose rectangles B_n centered at I/n as follows: Let $\mathbf{z}_i = (i/n, \sqrt{1 - (i/n)^2})$. Then the distance between \mathbf{z}_i and \mathbf{z}_{i+1} is

$$d(\mathbf{z}_i, \mathbf{z}_{i+1}) = \sqrt{((i/n)^2 + (\sqrt{1 - ((i+1)/n)^2} - \sqrt{1 - (i/n)^2})^2)}$$

Choose B_i to be the square centered at \mathbf{z}_i whose distance to \mathbf{z}_{i+1} is $\sqrt{2}d(\mathbf{z}_i, \mathbf{z}_{i+1})$. Then these squares overlap and their union covers the arc of $x^2 + y^2 = 1$ in quadrant one with summed area

$$\sum_{i=1}^n V(B_i) < \sum_{i=1}^n 4/n^2 = 4/n$$

Thus, for a given $\epsilon > 0$, there is N so that is $n > N$, the collection of rectangles covers the arc with summed volume less than ϵ . Hence the arc in quadrant one has measure zero in \Re^2 . A similar argument works in the other quadrants and thus this curve has measure zero in \Re^2 .

Example 15.4.2 Now look at the integral of $f(x, y) = 1 - x^2 - y^2$ on A the interior of the disc of radius one. Then let $R = [-1, 1] \times [-1, 1]$.

$$\hat{f}(\mathbf{x}) = \begin{cases} 1 - x^2 - y^2, & \mathbf{x} \in D = \{(x, y) | x^2 + y^2 < 1\} \\ 0, & \mathbf{x} \in R \setminus D \end{cases}$$

\hat{f} is clearly continuous on D and $\{(x, y) \in R | x^2 + y^2 > 1\}$. Also, the set of points where $x^2 + y^2 = 1$ is a set of measure zero in \mathbb{R}^2 . Hence, we know \hat{f} is continuous everywhere except a set of measure zero and so f is integrable on D . So $\int_D f dV$ exists and we can approximate it using any sequence of partitions whose norms go to zero. We usually would write $\int_D f dA$ as we are using area ideas in \mathbb{R}^2 . Finally, we would normally write $\int_{x^2+y^2<1} (x^2 + y^2) dA_{xy}$ to be more specific.

15.4.1 Homework

Exercise 15.4.1 Prove $x^2 + 4y^2$ is integrable on $[-1, 2] \times [3, 5]$.

Exercise 15.4.2 Prove $\sin(x^2 + 4xy + y^3)$ is integrable on the set A where $A = \{(x, y) | 0 \leq y \leq x^2\} \cap \{0 \leq y \leq 4\}$

Exercise 15.4.3 Prove $\sin(x^2 + 4xy + y^3)$ is integrable on the set A where $A = \{(x, y) | x^2 \leq y \leq 6\}$

Exercise 15.4.4 Prove $x^2 + y^2 + 4z^2$ is integrable on $D = \{(x, y, z) | z = x^2 + y^2 + z^2 \leq 25\}$

15.5 Integration and Sets of Measure Zero

Let $A \subset \mathbb{R}^n$ be a set of measure zero. Then if A contains a rectangle $E = \prod_{i=1}^n [a_i, b_i]$, we would know $V(E) > 0$. But we also know A has measure zero, so given $\epsilon = V(E)/2$, there is a collection B_i with $A \subset \cup_i B_i$ and $\sum_i V(B_i) < V(E)/2$. However, $E \subset A \subset \cup_i B_i$ implies $V(E) \leq \sum_i V(B_i)$ or $V(A) < V(A)/2$ which is not possible. We conclude if A is of measure zero, it can not contain a rectangle such as E .

Let S be a rectangle containing A and now let's assume A is a bounded set. Let $f : A \rightarrow \mathbb{R}$ be a bounded function which is integrable on A . Extend f to \hat{f} on s as usual. Let P be a partition of $S = \{S_1, \dots, S + M\}$. Since f is bounded, there are positives number m and M so that $m \leq f(\mathbf{x}) \leq M$ for $\mathbf{x} \in A$. Then

$$L(f, P) = \sum_{i \in P} \left(\inf_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x}) \right) V(S_i)$$

Now

$$\inf_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x}) = \inf \begin{cases} 0, & \mathbf{x} \in S_i \cap A^C \\ f(\mathbf{x}), & \mathbf{x} \in S_i \cap A \end{cases} \leq \inf \begin{cases} 0, & \mathbf{x} \in S_i \cap A^C \\ M, & \mathbf{x} \in S_i \cap A \end{cases} = M \inf_{\mathbf{x} \in S_i} \mathbf{1}_A$$

Thus,

$$L(f, P) \leq M \sum_{i \in P} \inf_{\mathbf{x} \in S_i} \mathbf{1}_A V(S_i)$$

If $\inf_{\mathbf{x} \in S_i} \mathbf{1}_A = 1$ for some index i , this would imply $S_i \cap A \neq \emptyset$. We know $A \subset \cup_i S_i$ though so this would force $S_i \subset A$. But by our argument at the start of this discussion, since A has measure zero, it is not possible for a nontrivial rectangle to be inside A . Thus, we must have $\inf_{\mathbf{x} \in S_i} \mathbf{1}_A = 0$

always. This tells us for all P

$$L(f, P) \leq M \sum_{i \in P} \inf_{\mathbf{x} \in S_i} \mathbf{1}_A V(S_i) \leq 0$$

We conclude $\underline{\int}_A f \leq 0$,

Next, we know $\sup_{\mathbf{x} \in S_i} \hat{f} = -\inf_{\mathbf{x} \in S_i} (-\hat{f})$, so

$$U(f, P) = \sum_{i \in P} \left(\sup_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x}) \right) V(S_i) = - \sum_{i \in P} \left(\inf_{\mathbf{x} \in S_i} (-\hat{f}(\mathbf{x})) \right) V(S_i) = -L(-f, P)$$

Then, by the same arguments we just used, we see $-f \leq -m$ implies

$$L(-f, P) \leq (-m) \sum_{i \in P} \inf_{\mathbf{x} \in S_i} \mathbf{1}_A V(S_i)$$

implying $L(-f, P) \leq 0$. Thus, $-L(-f, P) \geq 0$ for all P . Therefore $U(f, P) \geq 0$ for all P and so $\overline{\int}_A f \geq 0$. We conclude

$$\underline{\int}_A f \leq 0 \leq \overline{\int}_A f$$

But f is integrable on A , so we have $0 \leq \int_A f \leq 0$ or $\int_A f = 0$. We have proven an important theorem.

Theorem 15.5.1 If f is integrable on A and A has measure zero, then the integral of f on A is zero

Let $A \subset \mathbb{R}^n$ be bounded and have measure zero. If $f : A \rightarrow \mathbb{R}$ is integrable on A , then $\int_A f = 0$.

Proof 15.5.1

This is the argument we have just done. ■

Another interesting result is this:

Theorem 15.5.2 If the non negative f is integrable on A with value 0, then the measure of the set of points where $f > 0$ is measure zero

Let $A \subset \mathbb{R}^n$ be bounded and $f : A \rightarrow \mathbb{R}$ be integrable on A and non negative. Then the set of points where $f > 0$ is measure zero.

Proof 15.5.2

Let $A_m = \{\mathbf{x} \in A | f(\mathbf{x}) > 1/m\}$ for any positive integer m . Pick $\epsilon > 0$. Since $\int_A f = 0$, there is a partition P_m so that

$$\int_A f \leq U(f, P_m) < \int_A f + \epsilon/m \implies 0 \leq U(f, P_m) < \epsilon/m$$

as $\int_A f = 0$. Then,

$$\sum_{i \in P_m : S_i \cap A_m \neq \emptyset} \left(\sup_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x}) \right) V(S_i) \leq \sum_{i \in P_m} \left(\sup_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x}) \right) V(S_i) < \epsilon/m$$

If $S_i \cap A_m \neq \emptyset$, then there is an $\mathbf{x} \in S_i \cap A_m$ implying there is an \mathbf{x} with $\hat{f}(\mathbf{x}) > 1/m$. Hence,

$$\sum_{i \in P_m : S_i \cap A_m \neq \emptyset} 1/m V(S_i) \leq \sum_{i \in P_m} \left(\sup_{\mathbf{x} \in S_i} \hat{f}(\mathbf{x}) \right) V(S_i) < \epsilon/m$$

or $\sum_{i \in P_m : S_i \cap A_m \neq \emptyset} V(S_i) < \epsilon$. This tells us this collection of sets covers A_m with total summed volume smaller than ϵ . Since ϵ is arbitrary, we know A_m is measure zero which implies immediately the measure of the set of points where $f > 0$ is zero. ■

And, a result that seems natural.

Theorem 15.5.3 If f is integrable over A and f is zero except on a set of measure zero, this the integral is zero

Let $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}$ where A is bounded be integrable. Assume $B = \{\mathbf{x} \in A | f(\mathbf{x}) \neq 0\}$ is measure zero. Then $\int_A f = 0$.

Proof 15.5.3

Since f is integrable, for any sequence of partitions P_n with $\|P_n\| \rightarrow 0$, $S(f, \sigma_n, P_n) \rightarrow \int_A f$ where σ_n is any evaluation set from P_n . Now if S_{in} is a rectangle in P_n , if all points \mathbf{z} from S_{in} were in B , this would mean B contains a rectangle. But previous arguments tell us this is not possible since B has measure zero. Thus, each S_{in} contains a point not in B . Call this point \mathbf{z}_{in} and let $\sigma_n = \{\mathbf{z}_{in} | i \in P_n\}$. Then

$$S(f, \sigma_n, P_n) = \sum_{i \in P_n} f(\mathbf{z}_{in}) V(S_i) = 0$$

This shows $S(f, \sigma_n, P_n) \rightarrow 0 = \int_A f$. ■

Chapter 16

Change of Variables and Fubini’s Theorem

We now begin our discussion of how to make a change of variables in an integral. We start with linear maps.

16.1 Linear Maps

We start with the simplest case. We have this hazy notion of what it means for a set to have a volume: its indicator function must be integrable. But we really only know how to compute simple volume such as the volumes of rectangles. The next proof starts the process of understanding change of variable theorems by looking at what happens to a set with volume when it is transformed with a linear map.

Theorem 16.1.1 Change of Variable for a Linear Map

If $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a linear map and $D \subset \mathbb{R}^n$ is a set which has volume, i.e $\mathbf{1}_D$ is integrable and $V(D) = \int_D \mathbf{1}_D$, then $L(D)$ also has volume as

$$V(L(D)) = \int_{L(D)} \mathbf{1}_{L(D)} = |\det L| \int_D \mathbf{1}_D = \int_D |\det L| \mathbf{1}_D$$

Proof 16.1.1

Note if L is not invertible, the range of L is at most a $n - 1$ dimensional subspace of \mathbb{R}^n which has measure zero. Thus, $0 = \int_{L(D)} \mathbf{1}_{L(D)}$ and since $\det L = 0$, this matches $\det L \int_D \mathbf{1}_D = 0$. So this will work even if L is not invertible.

For any invertible matrix A there is a sequence of elementary matrices L_i so that

$$L_p L_{p-1} \cdots L_1 A = I$$

where each L_i is invertible. We discuss these things carefully in Chapter 10.2 but let’s review a bit here. These matrices L_i come in two types: T_1 and T_2 :

$$T_1 = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 1 & 1, (i,j) \text{ position} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 000 & 0 & 1 & & \end{bmatrix} = I + \Lambda_{ij}$$

where I is the usual identity matrix and Λ_{ij} is the matrix of all zeros except a 1 at the (i,j) position. Note $\det(T_1) = 1$ and so $\det(T_1^{-1}) = 1$ also. It is easy to see

$$T_1^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1, (i,j) \text{ position} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 000 & 0 & 1 & & \end{bmatrix} = I - \Lambda_{ij}$$

Next,

$$T_2 = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & c, (j,j) \text{ position} & \vdots \\ 000 & 0 & 1 & & \end{bmatrix}$$

It is easy to see

$$T_2 = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & 1/c, (j,j) \text{ position} & \vdots \\ 000 & 0 & 1 & & \end{bmatrix}$$

Note $\det(T_2) = c$ and so $\det(T_2^{-1}) = 1/c$ also. So $T_2 A$ creates a new matrix with scales row j by c and $T_1 A$ adds row j to row i . Thus, $T_1 T_2 A$ adds c times row j to row i to create a new row i . The same is true for the inverses. So $T_2^{-1} A$ creates a new matrix with scales row j by $1/c$ and $T_1^{-1} A$ subtracts row j to row i . Thus, $T_1 T_2 A$ subtracts $1/c$ times row j to row i to create a new row i .

The linear map L has a matrix representation A and so there are invertible linear maps λ_1 to λ_p so that $L = \lambda_1^{-1} \lambda_2^{-1} \cdots \lambda_p^{-1}$. Thus,

$$\begin{aligned} \det(L) &= \det(\lambda_1^{-1}) \cdots \det(\lambda_p^{-1}) \\ &= \det(L_1^{-1}) \cdots \det(L_p^{-1}) = \prod_{i \in I} (1/c_i) \end{aligned}$$

where I is the set of indices where L_i corresponds to a Type 2 matrix. Note this is independent of the matrix representations of these maps.

If λ_i^{-1} is of type T_2 , then $\lambda_i(x_1, x_2, \dots, x_n) = (x_1, x_2, \dots, 1/cx_j, \dots, x_n)$ for some j . and if S is a rectangle, we see $V(\lambda_i S) = |1/c|V(S)$.

If λ_i is of Type T_1 , we have $\lambda_i(x_1, x_2, \dots, x_n) = (x_1, x_2, \dots, x_i + x_j, \dots, x_n)$ for some i and j . Now apply λ_i to a rectangle S . The new rectangle $\lambda_i S$ preserves all the structure except in the $i - j$ plane. Then for $S = [a_1, b_1] \times \dots \times [a_n, b_n]$, all of S is preserved except in the $2 - 3$ plane. There we have

$$\begin{bmatrix} (a_2, b_3) & \dots & (b_2, b_3) \\ \vdots & \vdots & \vdots \\ (a_2, a_3) & \dots & (b_2, a_3) \end{bmatrix} \longrightarrow \begin{bmatrix} (a_2 + b_4, b_3) & \dots & (b_2 + b_4, b_3) \\ \vdots & \vdots & \vdots \\ (a_2 + a_4, a_3) & \dots & (b_2 + a_4, a_3) \end{bmatrix}$$

We see the area of the original rectangle in the $2 - 3$ plane is $(b_2 - a_2)(b_3 - a_3)$ which is the same as the area of the parallelogram in the transformed rectangle. This experiment can be repeated for other choices. We have found the T_1 applied to a rectangle S does not change its volume. A similar analysis shows T_1^{-1} applied to a rectangle S does not change its volume.

Now Pick a bounding rectangle R large enough to contain both D and $L(D)$. Since D has volume, by the integrable approximation theorem, given a sequence of partitions P_n with $\|P_n\| \rightarrow 0$, $U(\mathbf{1}_D, P_n) \downarrow \int_D \mathbf{1}_D = V(D)$ and $L(\mathbf{1}_D, P_n) \uparrow \int_D \mathbf{1}_D = V(D)$. Thus, for a given $\epsilon > 0$ there an N so that if $n > N$ $U(\mathbf{1}_D, P_n) < V(D) + \epsilon/|\det(L)|$ and $L(\mathbf{1}_D, P_n) > V(D) - \epsilon/|\det(L)|$. Hence, there are rectangles S_1, \dots, S_u in P_n , so that $\sum_{j=1}^u V(S_j) < V(D) + \epsilon/|\det(L)|$ and rectangles (possibly different) T_1, \dots, T_v in P_n so that $\sum_{j=1}^v V(T_j) > V(D) - \epsilon/|\det(L)|$. Then, from our arguments above

$$V(\lambda_1^{-1} \cdots \lambda_p^{-1}(S)) = |\det(L)|V(S)$$

Consider

$$U(\mathbf{1}_{L(D)}, P_n) = \sum_{S \in P_n} \sup_{\mathbf{x} \in S} ((\mathbf{1}_{L(D)}(\mathbf{x})V(S))$$

Now, if $\sup_{\mathbf{x} \in S} ((\mathbf{1}_{L(D)}(\mathbf{x})) = 1$. this means there is $\mathbf{x} \in S$ so that $L_{-1}(\mathbf{x}) \in D$. This means there is a rectangle $T \in P_n$ so that $\sup_{\mathbf{x} \in T} ((\mathbf{1}_D(\mathbf{x})) = 1$. Therefore, this 1 in the upper sum $U(\mathbf{1}_{L(D)}, P_n)$ also occurs in $U(\mathbf{1}_D, P_n)$. So every 1 in the upper sum $U(\mathbf{1}_{L(D)}, P_n)$ also occurs in $U(\mathbf{1}_D, P_n)$. We conclude

$$U(\mathbf{1}_{L(D)}, P_n) \leq U(\mathbf{1}_D, P_n)$$

Also,

$$L(\mathbf{1}_{L(D)}, P_n) = \sum_{S \in P_n} \inf_{\mathbf{x} \in S} ((\mathbf{1}_{L(D)}(\mathbf{x})V(S))$$

Now, if $\inf_{\mathbf{x} \in S} ((\mathbf{1}_{L(D)}(\mathbf{x})) = 1$. this means there is $S \subset L(D)$ so that $L_{-1}(\mathbf{x}) \in D$ for all $\mathbf{x} \in S$; i.e. $L^{-1}(S) \subset D$. This does not mean we can be sure we can find a $T \in P_n$ which is inside D . Thus, this 1 in $L(\mathbf{1}_{L(D)}, P_n)$ need not be repeated in $L(\mathbf{1}_D, P_n)$. We conclude

$$U(\mathbf{1}_{L(D)}, P_n) \geq U(\mathbf{1}_D, P_n)$$

Therefore,

$$U(\mathbf{1}_{L(D)}, P_n) - L(\mathbf{1}_{L(D)}, P_n) \leq U(\mathbf{1}_D, P_n) - L(\mathbf{1}_D, P_n)$$

We also know

$$U(\mathbf{1}_D, P_n) - L(\mathbf{1}_D, P_n) < \epsilon$$

Combining, we see $\mathbf{1}_{L(D)}$ is integrable since it satisfies the Riemann Criterion. Also

$$\begin{aligned} \sum_{j=1}^p (V(L(S_j)) / |\det(L)|) &< V(D) + \epsilon / |\det(L)| \\ \sum_{j=1}^q (V(L(T_j)) / |\det(L)|) &< V(D) + \epsilon / |\det(L)| \end{aligned}$$

or

$$\begin{aligned} U(\mathbf{1}_{L(D)}, P_n) \leq U(\mathbf{1}_D, P_n) &= \sum_{j=1}^u V(L(S_j)) < |\det(L)|V(D) + \epsilon \\ L(\mathbf{1}_{L(D)}, P_n) \geq L(\mathbf{1}_D, P_n) &= \sum_{j=1}^v V(L(T_j)) < |\det(L)|V(D) + \epsilon \end{aligned}$$

Since P_n is a sequence of partitions with $\|P_n\| \rightarrow 0$, we have

$$\begin{aligned} \int_{L(D)} \mathbf{1}_{L(D)} &= \lim_{P_n} U(\mathbf{1}_{L(D)}, P_n) \leq |\det(L)|V(D) + \epsilon \\ \int_{L(D)} \mathbf{1}_{L(D)} &= \lim_{P_n} L(\mathbf{1}_{L(D)}, P_n) \geq |\det(L)|V(D) - \epsilon \end{aligned}$$

giving for all $\epsilon > 0$

$$\left| \int_{L(D)} \mathbf{1}_{L(D)} - |\det(L)|V(D) \right| < \epsilon$$

Thus, $L(D)$ does have volume and $\int_{L(D)} \mathbf{1}_{L(D)} = |\det(L)|V(D)$. ■

It is easy to show that given two linearly independent vectors in \mathbb{R}^2 , say \mathbf{V} and \mathbf{W} , the parallelogram formed by these two vectors has an area that can be computed by looking at the projection of \mathbf{W} onto \mathbf{V} . If the angle between \mathbf{V} and \mathbf{W} is θ , then the area of the parallelogram is

$$\begin{aligned} A &= \|\mathbf{V}\| \|\mathbf{W}\| \sin(\theta) = \|\mathbf{V}\| \|\mathbf{W}\| \sqrt{\frac{\|\mathbf{V}\|^2 \|\mathbf{W}\|^2 - \langle \mathbf{V}, \mathbf{W} \rangle^2}{\|\mathbf{V}\|^2 \|\mathbf{W}\|^2}} \\ &= \sqrt{|\det(T)|^2} = |\det(T)| \end{aligned}$$

where T is the matrix formed using the \mathbf{V} and \mathbf{W} as its columns. It is a lot harder to work this out in \mathbb{R}^3 and trying to extend this idea to \mathbb{R}^n shows you there must be another way to reason this out. Theorem 16.1.1 is a reasonable way to attack this. It does require the large setup of a general integration theory in \mathbb{R}^n and the notion of extending volume in general, but it has the advantage of being quite clear logically. We suspect the **volume** of the parallelepiped in \mathbb{R}^n formed by n linearly independent vectors should be $|\det(T)|$ where T is the matrix formed using the vectors as columns. We can state this as follows:

Theorem 16.1.2 The Volume of a Rectangle in \mathbb{R}^n

If V_1, \dots, V_n are linearly independent in \mathbb{R}^n , the volume of the parallelepiped determined by them is $\det(T)$

Proof 16.1.2

First, note if D is a rectangle in \mathbb{R}^n , $\mathbf{1}_D(x) = 1$ on all of D and it lacks continuity only on ∂D which is a set of measure zero. Thus, it is integrable. Theorem 16.1.1 applied with $L = I$ tells us $\int_D \mathbf{1}_D = V(D)$ which is the usual $\prod(b_i - a_i)$ where $[a_i, b_i]$ is the i^{th} edge of the rectangle D . You can also prove this directly by looking at upper and lower sums without using this Theorem and you should try that. In this case, V_1, \dots, V_n determine an invertible linear map L on $D = [0, 1] \times \dots \times [0, 1]$ and so $\int_{L(D)} \mathbf{1}_{L(D)} = \det(L)V(D) = \det(L)$. ■

Example 16.1.1 Here is a sample volume calculation in \mathbb{R}^3 . We pick three independent vectors, set up the corresponding matrix and find its determinant.

Listing 16.1: **Sample Volume Calculation**

```
V1 = [1;2;-3];
V2 = [-3;4;10];
V3 = [11;22;1];
A = [V1,V2,V3]
A =
```

$$\begin{matrix} 1 & -3 & 11 \\ 2 & 4 & 22 \\ -3 & 10 & 1 \end{matrix}$$

```
det(A)
ans = 340.00
```

We see $V(A) = 340$.

Solution Yo

16.1.1 Homework

Exercise 16.1.1 Prove if D is a rectangle in \mathbb{R}^n , $V(D) = \prod(b_i - a_i)$ where $[a_i, b_i]$ is the i^{th} edge of the rectangle D from a lower and upper sum argument without using Theorem 16.1.1.

Exercise 16.1.2 Here are the vectors. Find the volume of the parallelepiped determined by them.

Listing 16.2: **Volume Calculation**

```
V1 = [-1;2;4;10];
V2 = [2;-3;5;1];
V3 = [8;-6;-7;1];
V4 = [9;1;2;4];
```

Exercise 16.1.3 Show the volume of a parallelepiped in \mathbb{R}^n is independent of the choice of orthonormal basis.

16.2 The Change of Variable Theorem

We start with an overestimate.

The Change of Variable result is a bit complicated to prove. So be prepared to follow some careful arguments!

Theorem 16.2.1 Subsets of Sets That have Volume

Let $D \subset \mathbb{R}^n$ be open and bounded. Assume $g(D)$ has volume. Then if $V \subset D$, $g(V)$ also have volume.

Proof 16.2.1

Since $g(V)$ has volume, for given $\epsilon > 0$, there is a partition P_ϵ of $g(D)$ so that

$$U(\hat{g}, P) - L(\hat{g}, P) < \epsilon$$

For any refinement P of P_ϵ . The partition P_ϵ is also a partition of $g(V)$ and if $h = g\mathbf{1}_V$, it is easy to see

$$U(\hat{h}, P) - L(\hat{h}, P) \leq U(\hat{g}, P) - L(\hat{g}, P) < \epsilon$$

which implies h satisfies the Riemann Criterion and so h is integrable and thus $g(V)$ has volume. ■

Our first change of variable theorem involves the integration of a simple constant.

Theorem 16.2.2 The Change of Variable Theorem for a Constant Map

Let $D \subset \mathbb{R}^n$ be open and bounded. Let $g : D \rightarrow \mathbb{R}^n$ have continuous partial derivatives on D and assume g is 1–1 on D with $J_g(x) \neq 0$ on D . If $g(D)$ has Volume, $\int_{g(D)} 1 = \int_D g|\det J_g|$.

Proof 16.2.2

Let $C(s, \mathbf{x}_0)$ be a cube of edge size s inside the open set D . The $C(s, \mathbf{x}_0)$ has the form

$$C(s, \mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\|_\infty < s\}$$

and $V(C(s, \mathbf{x}_0)) = (2s)^n$. Since g has continuous partial derivatives, using the Mean Value Theorem, we can say there are points \mathbf{u}_i on $[\mathbf{x}_0, \mathbf{x}]$ so that

$$g_i(\mathbf{x}) - g_i(\mathbf{x}_0) = \sum_{k=1}^n g_{i,x_k}(\mathbf{u}_i) (x_k - x_{0,k})$$

Note each $\mathbf{u}_i \in C(s, \mathbf{x}_0)$. We can then overestimate

$$\begin{aligned} |g_i(\mathbf{x}) - g_i(\mathbf{x}_0)| &\leq \sum_{k=1}^n |g_{i,x_k}(\mathbf{u}_i)| \|\mathbf{x} - \mathbf{x}_0\|_\infty \leq \sum_{k=1}^n |g_{i,x_k}(\mathbf{u}_i)| s \\ &\leq n \|J_g\|_{Fr}(\mathbf{u}_i) s \leq n \max_{\mathbf{y} \in C(s, \mathbf{x}_0)} \|J_g\|_{Fr}(\mathbf{y}) s \end{aligned}$$

as $\sqrt{n} < n$. Hence,

$$\|g(\mathbf{x}) - g(\mathbf{x}_0)\|_\infty \leq \max_{\mathbf{y} \in C(s, \mathbf{x}_0)} \|J_g\|_{Fr}(\mathbf{y}) ns$$

implying

$$V(g(C(s, \mathbf{x}_0))) \leq (\max_{\mathbf{y} \in C(s, \mathbf{x}_0)} \|\mathbf{J}_{\mathbf{g}}\|_{Fr}(\mathbf{y}))^n n^n V(C(s, \mathbf{x}_0))$$

Thus, $g(C(s, \mathbf{x}_0))$ is contained in a cube centered at $g(\mathbf{x}_0)$ defined by

$$E(s, \mathbf{x}_0) = \left\{ \mathbf{z} \mid \|\mathbf{z} - g(\mathbf{x}_0)\|_{\infty} \leq n \max_{\mathbf{y} \in C(s, \mathbf{x}_0)} \|\mathbf{J}_{\mathbf{g}}(\mathbf{y})\|_{Fr} \right\}$$

Now let L be an invertible linear map. Then, if S has volume, $L^{-1}(S)$ also has volume and

$$V(L^{-1}(S)) = \int_{L^{-1}(S)} 1 = \int_S (\det(L^{-1})) 1 = (\det(L^{-1})) V(S)$$

Let $S = g(C(s, \mathbf{x}_0))$. We know S has volume because of Theorem 16.2.1.

Thus, for any invertible map L ,

$$\det(L^{-1}) V(g(C(s, \mathbf{x}_0))) = V((L^{-1}g)(C(s, \mathbf{x}_0))) = \left(\max_{\mathbf{y} \in C(s, \mathbf{x}_0)} \|(L^{-1}\mathbf{J}_{\mathbf{g}})(\mathbf{y})\|_1 \right)^n n^n V(C(s, \mathbf{x}_0))$$

implying

$$V(g(C(s, \mathbf{x}_0))) = V((L^{-1}g)(C(s, \mathbf{x}_0))) = (\det L) \left(\max_{\mathbf{y} \in C(s, \mathbf{x}_0)} \|(L^{-1}\mathbf{J}_{\mathbf{g}})(\mathbf{y})\|_1 \right)^n n^n V(C(s, \mathbf{x}_0))$$

Now divide $C(s, \mathbf{x}_0)$ into M cubes C_1, \dots, C_M with centers $\mathbf{x}_1, \dots, \mathbf{x}_M$ and apply $L_i = \mathbf{J}_{\mathbf{g}}(\mathbf{x}_i)$ to each cube C_k . We find

$$V(g(C_k)) = (\det L_k) \left(\max_{\mathbf{y} \in C(s, \mathbf{x}_0)} \|(L_k^{-1}\mathbf{J}_{\mathbf{g}})(\mathbf{y})\|_1 \right)^n n^n V(C_k)$$

and thus

$$V(g(C(s, \mathbf{x}_0))) = \sum_{k=1}^M V(g(C_k)) = \sum_{k=1}^M (\det L_k) \left(\max_{\mathbf{y} \in C(s, \mathbf{x}_0)} \|(L_k^{-1}\mathbf{J}_{\mathbf{g}})(\mathbf{y})\|_1 \right)^n V(C_k)$$

The function $(\mathbf{J}_{\mathbf{g}}(\mathbf{y}))^{-1}$ is uniformly continuous on $C(s, \mathbf{x}_0)$ with respect to the Frobenius (i.e sup) norm on matrices (you should be able to check this easily) and so

$$\lim_{\mathbf{z} \rightarrow \mathbf{y}} \mathbf{J}_{\mathbf{g}}(\mathbf{z})^{-1} = \mathbf{J}_{\mathbf{g}}(\mathbf{y})^{-1}$$

which tells us

$$\lim_{\mathbf{z} \rightarrow \mathbf{y}} \mathbf{J}_{\mathbf{g}}(\mathbf{z})^{-1} \mathbf{J}_{\mathbf{g}}(\mathbf{y}) = \mathbf{J}_{\mathbf{g}}(\mathbf{y})^{-1} \mathbf{J}_{\mathbf{g}}(\mathbf{y}) = I$$

Thus, given $\epsilon > 0$, there is a $\delta > 0$ so that

$$\|\mathbf{z} - \mathbf{y}\| < \delta \implies \|\mathbf{J}_{\mathbf{g}}(\mathbf{z})^{-1} \mathbf{J}_{\mathbf{g}}(\mathbf{y}) - I\|_{\infty} < \epsilon$$

But this says

$$\|\mathbf{z} - \mathbf{y}\| < \delta \implies \|I\|_{\infty} - \epsilon < \|\mathbf{J}_{\mathbf{g}}(\mathbf{z})^{-1} \mathbf{J}_{\mathbf{g}}(\mathbf{y})\|_{\infty} < \|I\|_{\infty} + \epsilon$$

or

$$\|\mathbf{z} - \mathbf{y}\| < \delta \implies 1 - \epsilon < \|\mathbf{J}_g(\mathbf{z})^{-1} \mathbf{J}_g(\mathbf{y})\|_\infty < 1 + \epsilon$$

Now we find the subdivision of cubes we need. Let $\epsilon = 1/p$ and for the δ associated with $\epsilon = 1/p$, choose it so that $\delta_p < 1/p$. Choose the cubes C_1, \dots, C_{M_p} and centers $\mathbf{x}_1 < \dots, \mathbf{x}_{M_p}$ so that $\|\mathbf{x}_i - \mathbf{y}\| < \delta_p$. Then, we must have

$$\|\mathbf{z} - \mathbf{x}_i\| < \delta_p \implies 1 - 1/p < \|\mathbf{J}_g(\mathbf{x}_i)^{-1} \mathbf{J}_g(\mathbf{y})\|_\infty < 1 + 1/p$$

We now have the estimate

$$\begin{aligned} V(g(C(s, \mathbf{x}_0))) &\leq \sum_{i=1}^{M_p} \left(\max_{\mathbf{y} \in C_i} (1 + 1/p) \right)^n |\det(\mathbf{J}_g(\mathbf{x}_i))| n^n V(C_i) \\ &\leq \sum_{i=1}^{M_p} (1 + 1/p)^n \det(\mathbf{J}_g(\mathbf{x}_i)) n^n V(C_i) \end{aligned}$$

Since $(1 + 1/p) \rightarrow 1$ as $p \rightarrow \infty$, for a given ϵ , there is a p_0 so that $(1 + 1/p)^n n^n < 1 + \epsilon$. We conclude if $p > p_0$,

$$V(g(C(s, \mathbf{x}_0))) \sum_{i=1}^{M_p} (1 + \epsilon) |\det(\mathbf{J}_g(\mathbf{x}_i))| V(C_i)$$

Now, let’s connect this to Riemann sums. Let P_p be the partition of $g(C(s, \mathbf{x}_0))$ determined by these cubes. Then for the evaluation set $\sigma_p = \mathbf{x}_1, \dots, \mathbf{x}_{M_p}$, we have

$$\sum_{i \in P_p} \det(\mathbf{J}_g(\mathbf{x}_i)) V(C_i) = S(|\det(\mathbf{J}_g)|, \sigma_p, P_p)$$

implying for $p > p_0$,

$$V(g(C(s, \mathbf{x}_0))) \leq (1 + \epsilon) S(|\det(\mathbf{J}_g)|, \sigma_p, P_p)$$

We know ($|\det(\mathbf{J}_g|$ is integrable on $C(s, \mathbf{x}_0)$ and since $\|P_p\| \rightarrow 0$, we know $S(|\det(\mathbf{J}_g)|, \sigma_p, P_p) \rightarrow \int_{g(C(s, \mathbf{x}_0))} |\det(\mathbf{J}_g)|$. Since $\epsilon > 0$ is arbitrary, we know

$$V(g(C(s, \mathbf{x}_0))) \leq \int_{g(C(s, \mathbf{x}_0))} |\det(\mathbf{J}_g)|$$

To finish, we let P be a partition of $g(D)$ into cubes C_i . Then each C_i is contained in a cube W_i and $d(D) \subset \cup W_i$. Since $g(D)$ has volume, so does $g(C_i)$ and

$$\int_{g(C_i)} \mathbf{1}_{g(C_i)} = V(g(C_i))$$

Hence,

$$\begin{aligned} \int_{g(D)} \mathbf{1}_{g(D)} &= V(g(D)) = \sum \int_{g(C_i)} \mathbf{1}_{g(C_i)} = \sum V(g(C_i)) \\ &\leq \sum \int_{C_i} |\det(\mathbf{J}_g)| = \int_D |\det(\mathbf{J}_g)| \end{aligned}$$

16.2. THE CHANGE OF VARIABLE THEOREM

327

We have shown one part of the inequality we need: $\int_{g(D)} \mathbf{1}_{g(D)} \leq \int_D |\det(\mathbf{J}_g)|$.

The argument we just gave works just as well for another case. If $f : g(D) \rightarrow \mathbb{R}$ has continuous partial derivatives on $g(D)$, then the composite map, then $\int_{g(D)} f \leq \int_D f \circ g |\det(\mathbf{J}_g)|$. Let $\phi = (f \circ g) |\det(\mathbf{J}_g)|$ and $\psi = g^{-1}$. Then our last result applies and

$$\int_{\psi(g(D))} \phi \leq \int_{g(D)} \phi \circ \psi |\det(\mathbf{J}_\psi)|$$

or

$$\int_D (f \circ g) |\det(\mathbf{J}_g)| \leq \int_{g(D)} ((f \circ g) |\det(\mathbf{J}_g)|) \circ g^{-1} |\det(\mathbf{J}_g^{-1})|$$

At any particular point $\mathbf{y} \in g(D)$, we have

$$\phi \circ \psi(\mathbf{y}) = f(g(g^{-1}(\mathbf{y}))) |\det(\mathbf{J}_g)(g^{-1}(\mathbf{y}))| |\det(\mathbf{J}_g^{-1})(\mathbf{y})| = f(\mathbf{y})$$

as the last two terms are inverses of one another. Thus, $((f \circ g) |\det(\mathbf{J}_g)|) \circ g^{-1} |\det(\mathbf{J}_g^{-1})| = f$ and we have

$$\int_D f \circ (g |\det(\mathbf{J}_g)|) \leq \int_{g(D)} f$$

Then for the choice $f = 1$, since $1 \circ g = 1$, we find

$$\int_D |\det(\mathbf{J}_g)| \leq \int_{g(D)} 1 = \int_{g(D)} \mathbf{1}_{g(D)}$$

This provides the other half of the inequality we need. Combining, we have $\int_D |\det(\mathbf{J}_g)| = \int_{g(D)} \mathbf{1}_{g(D)}$. So we have shown the Change of Variable Theorem works for constant maps f . ■

Next, we extend from a constant map to a general map.

Theorem 16.2.3 The Change of Variable Theorem for a General Map

Let $D \subset \mathbb{R}^n$ be open and bounded. Let $g : D \rightarrow \mathbb{R}^n$ have continuous partial derivatives on D and assume g is 1–1 on D with $\mathbf{J}_g(\mathbf{x}) \neq 0$ on D . If $g(D)$ has Volume, then for $f : g(D) \rightarrow \mathbb{R}$ integrable, $f \circ g |\det \mathbf{J}_G|$ is also integrable and $\int_{g(D)} f = \int_D f \circ g |\det \mathbf{J}_G|$.

Proof 16.2.3

We assume f is integrable on $g(D)$. Let R be a rectangle containing $g(D)$ and P be a partition of $g(D)$ using this rectangle. For any rectangle S determined by P , define the function h_S by $h_S(\mathbf{x}) = \inf_{\mathbf{x} \in S} f(\mathbf{x})$. Then, h_S is a constant on S . Then

$$L(f, P) = \sum_{S \in P} \inf_{\mathbf{x} \in S} f(\mathbf{x}) V(S) = \sum_{S \in P} h_S V(S) = \sum_{S \in P} \int_{S \in P} h_S$$

Now apply the Change of Variable Theorem to the constant function h_S . We have

$$\int_S h_S = \int_{g^{-1}(S)} h_S \circ g |\det \mathbf{J}_g|$$

If $\mathbf{x} \in g^{-1}(S)$, then $h_S \circ g(\mathbf{x}) = h_S(g(\mathbf{x}))$. Also, if $\mathbf{x} \in g^{-1}(S)$, $g(\mathbf{x}) \in S$. We conclude

$$h_S \circ g(\mathbf{x}) = h_S(g(\mathbf{x})) = \inf_{g(\mathbf{x}) \in S} f(g(\mathbf{x})) = \inf_{\mathbf{x} \in g^{-1}(S)} f(g(\mathbf{x}))$$

So

$$\int_{g^{-1}(S)} h_S \circ g |det \mathbf{J}_g| = \int_{g^{-1}(S)} \left(\inf_{\mathbf{x} \in g^{-1}(S)} f(g(\mathbf{x})) \right) |det \mathbf{J}_g|$$

Combining,

$$\begin{aligned} L(f, P) &= \sum_{S \in P} \int_{S \in P} h_S = \sum_{S \in P} \int_{g^{-1}(S)} \left(\inf_{\mathbf{x} \in g^{-1}(S)} f(g(\mathbf{x})) \right) |det \mathbf{J}_g| \\ &\leq \sum_{S \in P} \int_{g^{-1}(S)} f(g(\mathbf{x})) |det \mathbf{J}_g| \end{aligned}$$

But $\sum_{S \in P} \int_{g^{-1}(S)} = \int_D$. Hence, we have shown $L(f, P) \leq \int_D f \circ g |det \mathbf{J}_g|$. This implies

$$\int_{g(D)} f = \sup_P L(f, P) \leq \int_D f \circ g |det \mathbf{J}_g|$$

Since we assume f is integrable over $g(D)$, we conclude $\int_{g(D)} f \leq \int_D f \circ g |det \mathbf{J}_g|$.

Since we now know the Change of Variable Theorem hold for constant f , we can use a similar argument to show

$$\inf_P U(f, P) \geq \int_D f \circ g |det \mathbf{J}_g|$$

Thus, $\overline{\int}_{g(D)} f \geq \int_D f \circ g |det \mathbf{J}_g|$. But we assume f is integrable on $g(D)$ and therefore, we have shown

$$\int_D f \circ g |det \mathbf{J}_g| \leq \int_{g(D)} f \leq \int_D f \circ g |det \mathbf{J}_g|$$

which shows $\int_{g(D)} f = \int_D f \circ g |det \mathbf{J}_g|$. We have shown the Change of Variable Theorem hold for a general f . ■

16.2.1 Homework

Exercise 16.2.1 Work out the Change of Variable results for polar coordinate transformations.

Exercise 16.2.2 Work out the Change of Variable results for cylindrical coordinate transformations.

Exercise 16.2.3 Work out the Change of Variable results for spherical coordinate transformations.

Exercise 16.2.4 Work out the Change of Variable results for ellipsoidal coordinate transformations.

Exercise 16.2.5 Work out the Change of Variable results for paraboloidal coordinate transformations.

16.3 Fubini Type Results

We all remember how to evaluate double integrals for the area of a region D by rewriting them like this

$$\int_D f = \int_a^b \int_{f(x)}^{g(x)} F(x, y) dy dx$$

where $\int_D f$ is the two dimensional integral we have been developing. Expressing this integral in terms of successive single variable integrals leads to what are called **iterated** integrals which is how we commonly evaluate them in our calculus classes. Thus, using the notation $\int dx$, $\int dy$ etc. to indicate these one dimensional integrals along one axis only, we want to know when

$$\begin{aligned} \int_D F &= \int \left(\int F(x, y) dy \right) dx = \int \int F(x, y) dy dx \\ \int_D F &= \int \left(\int F(x, y) dx \right) dy = \int \int F(x, y) dx dy \\ \int_{\Omega} G &= \int \left(\int \left\{ \int G(x, y, z) dy \right\} \right) dz = \int \int \int G(x, y, z) dy dx dz \\ \int_{\Omega} G &= \int \left(\int \left\{ \int G(x, y, z) dy \right\} \right) dx = \int \int \int G(x, y, z) dy dz dx \\ \int_{\Omega} G &= \int \left(\int \left\{ \int G(x, y, z) dx \right\} \right) dy = \int \int \int G(x, y, z) dx dy dz \\ \int_{\Omega} G &= \int \left(\int \left\{ \int G(x, y, z) dx \right\} \right) dz = \int \int \int G(x, y, z) dxdz dy \\ \int_{\Omega} G &= \int \left(\int \left\{ \int G(x, y, z) dz \right\} \right) dx = \int \int \int G(x, y, z) dz dx dy \\ \int_{\Omega} G &= \int \left(\int \left\{ \int G(x, y, z) dz \right\} \right) dy = \int \int \int G(x, y, z) dz dy dx \end{aligned}$$

where D is a bounded open subset of \mathbb{R}^2 and Ω is a bounded open subset of \mathbb{R}^3 . Note in the Ω integration there are many different ways to organize the integration. If we were doing an integration over \mathbb{R}^n there would in general be a lot of choices! The conditions under which we can do this give rise to what are called Fubini type theorems. If general, you think of the bounded open set V as being in $\mathbb{R}^{n+m} = \mathbb{R}^n \times \mathbb{R}^m$, label the volume elements as dV_n and dV_m and want to know when

$$\begin{aligned} \int_V F &= \int \left(\int F(x_1, \dots, x_n, x_{n+1}, \dots, x_{n+m}) dV_n \right) dV_m \\ &= \int \left(\int F dV_n \right) dV_m = \int \int F dV_n dV_m \end{aligned}$$

We are only going to work out a few of these type theorems so that you can get a taste for the ways we prove these results.

16.3.1 Fubini on a Rectangle

Let's specialize to the bounded set $D = [a, b] \times [c, d] \subset \mathbb{R}^2$. Assume $f : D \rightarrow \mathbb{R}$ is continuous. There are three integrals here:

- The integral over the two dimensional set D , $\int_D f dV_2$ where the subscript 2 reminds us that this is the two dimensional situation.

- The integral of functions like $g : [a, b] \rightarrow \mathbb{R}$. Assume g is continuous. Then the one dimensional integral would be represented as $\int_a^b g(x)dx$ in the integration treatment (Peterson (8) 2019) but here we will use $\int_{[a,b]} g dV_1$.
- The integral of functions like $h : [c, d] \rightarrow \mathbb{R}$. Assume h is continuous. Then the one dimensional integral $\int_c^d h(y)dy$ is called $\int_{[c,d]} h dV_2$. Note although the choice of integration variable y in $\int_c^d g(y)dy$ is completely arbitrary, it makes a lot of sense to use this notation instead of $\int_c^d g(x)dx$ because we want to cleanly separate what we are doing on each axis from the other axis. So we have to exercise a bit of discipline here to set up the variable names and notation to help us understand.

Now any partition P of D divides the bounding box R we choose to enclose D into rectangles. If $P = P_1 \times P_2$ where $P_1 = \{a = x_0, x_1, \dots, x_{n-1}, x_n = b\}$ and $P_2 = \{c = y_0, y_1, \dots, y_{m-1}, y_m = d\}$, the rectangles have the form

$$S_{ij} = [x_i, x_{i+1}] \times [y_j, y_{j+1}] = \Delta x_i \times \Delta y_j$$

where $\Delta x_i = x_{i+1} - x_i$ and $\Delta y_j = y_{j+1} - y_j$. Thus for an evaluation set σ with points $(z_i, w_j) \in S_{ij}$, the Riemann sum is

$$S(f, \sigma, P) = \sum_{(i,j) \in P} f(z_i, w_j) \Delta x_i \Delta y_j = \sum_{(i,j) \in P} f(z_i, w_j) \Delta y_j \Delta x_i$$

Note the order of the terms in the area calculation for each rectangle do not matter. Also, we could call $\Delta x_i = \Delta V_{1i}$ (for the first axis) and $\Delta y_j = \Delta V_{2j}$ (for the second axis) and rewrite as

$$S(f, \sigma, P) = \sum_{(i,j) \in P} f(z_i, w_j) \Delta V_{1i} \Delta V_{2j} = \sum_{(i,j) \in P} f(z_i, w_j) \Delta V_{2j} \Delta V_{1i}$$

We can also revert to writing $\sum_{(i,j) \in P}$ as a double sum to get

$$S(f, \sigma, P) = \sum_{i=1}^n \left(\sum_{j=1}^m f(z_i, w_j) \Delta V_{2j} \right) \Delta V_{1i} = \sum_{j=1}^m \left(\sum_{i=1}^n f(z_i, w_j) \Delta V_{1i} \right) \Delta V_{2j}$$

Then the idea is this

• **y first then x**

For $\sum_{i=1}^n \left(\sum_{j=1}^m f(z_i, w_j) \Delta V_{2j} \right) \Delta V_{1i}$, fix the variables on axis one and suppose

$$\lim_{\|P_2\| \rightarrow 0} \sum_{j=1}^m f(z_i, w_j) \Delta V_{2j} = \int_{z_i \times [c,d]} f(z_i, y) dy$$

To avoid using specific sizes like m for P_2 here, we would usually just say

$$\lim_{\|P_2\| \rightarrow 0} \sum_{j \in P_2} f(z_i, w_j) \Delta V_{2j} = \int_{z_i \times [c,d]} f(z_i, y) dy$$

Now for this to happen we need the function $f(\cdot, y)$ to be integrable for each choice of x for slot one. An easy way to do this is to assume f is continuous in $(x, y) \in D$ which will force $f(\cdot, y)$ to be a continuous function of y for each x . Let’s look at this integral more carefully.

16.3. FUBINI TYPE RESULTS

331

We have defined a new function $G(x)$ by

$$G(x) = \int_{[c,d]} f(x, y) dy = \int_c^d f(x, y) dy$$

If $G(x)$ is integrable on $[a, b]$, we can say

$$\begin{aligned} \lim_{\|P_1\| \rightarrow 0} \sum_{i=1}^n \left(\lim_{\|P_2\| \rightarrow 0} \sum_{j=1}^m f(z_i, w_j) \Delta V_{2j} \right) \Delta V_{1i} &= \lim_{\|P_1\| \rightarrow 0} \sum_{i=1}^n \left(\int_{z_i \times [c,d]} f(z_i, y) dy \right) \Delta V_{1i} \\ &= \lim_{\|P_1\| \rightarrow 0} \sum_{i=1}^n G(z_i) \Delta V_{1i} = \int_a^b G(x) dx \end{aligned}$$

We can say this again. Let ΔV_{ij} be the area $\Delta x_i \Delta y_j$. Then

$$\begin{aligned} \int_D f dV_2 &= \lim_{\|P_1 \times P_2\| \rightarrow 0} \sum_{(i,j) \in P_1 \times P_2} f(z_i, w_j) \Delta V_{ij} \\ &= \lim_{\|P_1\| \rightarrow 0} \sum_{i \in P_1} \left(\lim_{\|P_2\| \rightarrow 0} \sum_{j \in P_2} f(z_i, w_j) \Delta V_{2j} \right) \Delta V_{1i} \\ &= \lim_{\|P_1\| \rightarrow 0} \sum_{i \in P_1} \left(\int_c^d f(z_i, y) dy \right) \Delta V_{1i} \\ &= \int_a^b \left(\int_c^d f(x, y) dy \right) dx = \int_a^b \int_c^d f(x, y) dy dx \end{aligned}$$

- **x first then y**

For $\sum_{j=1}^m \left(\sum_{i=1}^n f(z_i, w_j) \Delta V_{1i} \right) \Delta V_{2j}$, fix the variables on axis two and suppose

$$\lim_{\|P_1\| \rightarrow 0} \sum_{i=1}^n f(z_i, w_j) \Delta V_{1i} = \int_{[a,b] \times w_j} f(x, w_j) dx$$

Of course, this is the same as

$$\lim_{\|P_1\| \rightarrow 0} \sum_{i \in P_1} f(z_i, w_j) \Delta V_{1i} = \int_{[a,b] \times w_j} f(x, w_j) dx$$

Now for this to happen we need the function $f(x, \cdot)$ to be integrable for each choice of y for slot two. If we assume f is continuous in $(x, y) \in D$ this will force $f(x, \cdot)$ to be a continuous function of x for each y . Let's look at this integral more carefully. We have defined a new function $H(y)$ by

$$H(y) = \int_{[a,b]} f(x, y) dx = \int_a^b f(x, y) dx$$

If $H(y)$ is integrable on $[c, d]$, we can say

$$\lim_{\|P_2\| \rightarrow 0} \sum_{j=1}^m \left(\lim_{\|P_1\| \rightarrow 0} \sum_{i=1}^n f(z_i, w_j) \Delta V_{1i} \right) \Delta V_{2j} = \lim_{\|P_2\| \rightarrow 0} \sum_{j=1}^m \left(\int_{[a,b] \times w_j} f(x, w_j) dx \right) \Delta V_{2j}$$

$$= \lim_{\|P_2\| \rightarrow 0} \sum_{j=1}^m H(w_j) \Delta V_{2j} = \int_c^d H(y) dy$$

or

$$\begin{aligned} \int_D f dV_2 &= \lim_{\|P_1 \times P_2\| \rightarrow 0} \sum_{(i,j) \in P_1 \times P_2} f(z_i, w_j) \Delta V_{ij} \\ &= \lim_{\|P_2\| \rightarrow 0} \sum_{j \in P_2} \left(\lim_{\|P_1\| \rightarrow 0} \sum_{i \in P_1} f(z_i, w_j) \Delta V_{1i} \right) \Delta V_{2j} \\ &= \lim_{\|P_2\| \rightarrow 0} \sum_{j \in P_2} \left(\int_a^b f(x, w_j) dx \right) \Delta V_{2j} \\ &= \int_c^d \left(\int_a^b f(x, y) dx \right) dy = \int_c^d \int_a^b f(x, y) dx dy \end{aligned}$$

We have laid out an approximate chain of reasoning to prove our first Fubini result. Let’s get to it.

Theorem 16.3.1 Fubini’s Theorem on a Rectangle: 2D

Let $f : [a, b] \times [c, d] \rightarrow \mathbb{R}$ be continuous. Then

$$\begin{aligned} \int_D f dV_2 &= \int_a^b \left(\int_c^d f(x, y) dy \right) dx = \int_a^b \int_c^d f(x, y) dy dx \\ \int_D f dV_2 &= \int_c^d \left(\int_a^b f(x, y) dx \right) dy = \int_c^d \int_a^b f(x, y) dx dy \end{aligned}$$

Proof 16.3.1

For $x \in [a, b]$, let G be defined by $G(x) = \int_c^d f(x, y) dy$. We need to show G is continuous for each x . Note since f is continuous on $[a, b] \times [c, d]$, given $\epsilon > 0$, there is a $\delta > 0$ so that

$$\sqrt{(x' - x)^2 + (y' - y)^2} < \delta \implies |f(x, y) - f(x', y')| < \epsilon$$

In particular, for fixed $y' = y$, there is a $\delta > 0$ so that

$$|x' - x| < \delta \implies |f(x', y) - f(x, y)| < \epsilon/(d - c)$$

Thus, if $|x' - x| < \delta$,

$$\begin{aligned} |G(x') - G(x)| &= \left| \int_c^d (f(x', y) - f(x, y)) dy \right| \leq \int_c^d |f(x', y) - f(x, y)| dy \\ &\leq (d - c) \epsilon / (d - c) = \epsilon \end{aligned}$$

and so we know G is continuous in x for each y . A similar argument shows the function H defined at each $y \in [c, d]$ by $H(y) = \int_a^b f(x, y) dx$ is also continuous in y for each x .

Since f is continuous on $[a, b] \times [c, d]$, and the boundary of D has measure zero, we see the set of discontinuities of f on D is a set of measure zero and so f is integrable on D . Let $P_n = P_{1n} \times P_{2n}$

be a sequence of partitions with $\|P_n\| \rightarrow 0$. Then, we know for any evaluation set σ of P , that

$$\int_D f = \lim_{\|P_n\| \rightarrow 0} \sum_{i \in P_{1n}} \sum_{j \in P_{2n}} f(z_i, w_j) \Delta x_i \Delta y_j$$

and following the reasoning we sketch out earlier, we can choose to organize this as **x first then y** or **y first then x**. We will do the case of **x first then y** only and leave the details of the other case to you. fix $\epsilon > 0$, Then there is N , so that $n > N$ implies

$$\left| \sum_{j \in P_{2n}} \left(\sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i \right) \Delta y_j - \int_D f \right| < \epsilon/2$$

Since $f(x, w_j)$ is continuous in x for each w_j , we know $\int_a^b f(x, w_j) dx$ exists and so for any sequence of partitions Q_n of $[a, b]$ with $\|Q_n\| \rightarrow 0$, using as the evaluation set the points z_i , we have

$$\sum_{i \in Q_n} f(z_i, w_j) \Delta x_i \rightarrow \int_a^b f(x, w_j) dx.$$

Since $\|P_n\| \rightarrow 0$, so does the sequence P_{1n} , there is $N_1 < N$ so that $n > N_1$ implies

$$\left| \sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i - \int_a^b f(x, w_j) dx \right| = \left| \sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i - H(w_j) \right| < \epsilon/(4(d-c))$$

Similarly, since $\|P_{2n}\| \rightarrow 0$ and H is integrable, there is $N_2 > N_1 > N$ so that

$$\left| \sum_{j \in P_{2n}} H(w_j) \Delta y_j - \int_c^d H(y) dy \right| < \epsilon/4$$

where the points w_j form the evaluation set we use in each P_{2n} . Thus,

$$\begin{aligned} & \left| \sum_{j \in P_{2n}} \left(\sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i \right) \Delta y_j - \int_c^d H(y) dy \right| \\ &= \left| \sum_{j \in P_{2n}} \left(\sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i - H(w_j) + H(w_j) \right) \Delta y_j - \int_c^d H(y) dy \right| \\ &\leq \left| \sum_{j \in P_{2n}} \left(\sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i - H(w_j) \right) \Delta y_j \right| + \left| \sum_{j \in P_{2n}} H(w_j) \Delta y_j - \int_c^d H(y) dy \right| \\ &\leq \sum_{j \in P_{2n}} \left| \sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i - H(w_j) \right| \Delta y_j + \left| \sum_{j \in P_{2n}} H(w_j) \Delta y_j - \int_c^d H(y) dy \right| \\ &< \sum_{j \in P_{2n}} \epsilon/(4(d-c)) \Delta y_j + \epsilon/4 = \epsilon/2 \end{aligned}$$

We can now complete the argument. For $n > N_2$,

$$\left| \int_c^d H(y) dy - \int_d f \right| \leq \left| \sum_{j \in P_{2n}} \left(\sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i \right) \Delta y_j - \int_D f \right|$$

$$+ \left| \sum_{j \in P_{2n}} \left(\sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i \right) \Delta y_j - \int_c^d H(y) dy \right| \\ < \epsilon/2 + \epsilon/2 = \epsilon$$

Thus

$$\int_D f = \int_c^d H(y) dy = \int_c^d \left(\int_a^b f(x, y) dx \right) dy = \int_c^d \int_a^b f(x, y) dx dy$$

A very similar argument handles the case **y first then x** giving

$$\int_D f = \int_a^b G(x) dx = \int_a^b \left(\int_c^x f(x, y) dy \right) dx = \int_a^b \int_c^x f(x, y) dy dx$$

■

As you can see, the argument to prove Theorem 16.3.1 is somewhat straightforward but you have to be careful to organize the inequality estimates. A similar approach can prove the more general theorem.

Theorem 16.3.2 Fubini’s Theorem for a Rectangle: n dimensional

Let $f : R_1 \times R_2 \rightarrow \mathbb{R}$ be continuous where R_1 is a rectangle in \mathbb{R}^n and $R_2 \in \mathbb{R}^m$. Then

$$\int_D f dV_{n+m} = \int_{R_1} \left(\int_{R_2} f(\mathbf{x}, \mathbf{y}) dV_m \right) dV_n = \int_{R_1} \int_{R_2} f(\mathbf{x}, \mathbf{y}) dV_m dV_n \\ \int_D f dV_{n+m} = \int_{R_2} \left(\int_{R_1} f(\mathbf{x}, \mathbf{y}) dV_n \right) dV_m = \int_{R_2} \int_{R_1} f(\mathbf{x}, \mathbf{y}) dV_n dV_m$$

where **x** are the variables x_1, \dots, x_n from \mathbb{R}^n and **y** are the variables x_{n+1}, \dots, x_{n+m} .

Proof 16.3.2

For example, $f : [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3] \times [a_4, b_4] \rightarrow \mathbb{R}$ would use $R_1 = [a_1, b_1] \times [a_2, b_2]$ and $R_2 = [a_3, b_3] \times [a_4, b_4]$. The proof here is quite similar. We prove $G(\mathbf{x}$ and $H(\mathbf{y}$ are continuous since f is continuous. Then, since f is integrable, given any sequence of partitions P_n with $\|P_n\| \rightarrow 0$, there is N so that $n > N$ implies

$$\left| \sum_{j \in P_{2n}} \left(\sum_{i \in P_{1n}} f(z_i, w_j) \Delta x_i \right) \Delta y_j - \int_D f \right| < \epsilon/2$$

where $P_n = P_{1n} \times P_{2n}$ like usual except P_{1n} is a partition of R_1 and P_{2n} is a partition of R_2 . The summations like $\sum_{j \in P_{2n}}$ are still interpreted in the same way. The rest of the argument is straightforward. ■

Now let’s specialize to some useful two dimensional situations. Consider the situation shown in Figure 16.1.

Theorem 16.3.3 Fubini’s Theorem: 2D: a top and bottom curve

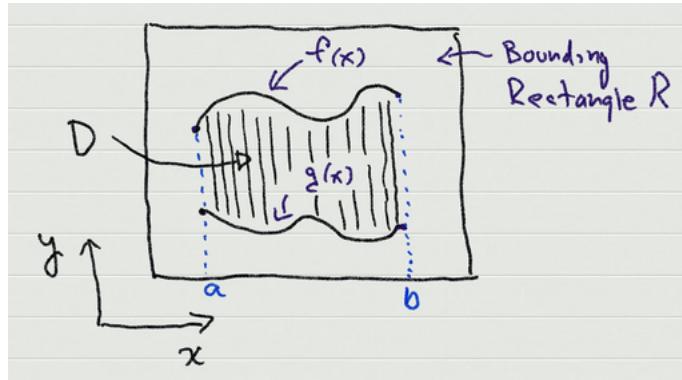


Figure 16.1: Fubini's Theorem in 2D: a top and a bottom curve

Let f and g be continuous real valued functions on $[a, b]$ and $F : D \rightarrow \mathbb{R}$ be continuous on D where $D = \{(x, y) : a \leq x \leq b, g(x) \leq y \leq f(x)\}$. Then

$$\int_D F dV_2 = \int_a^b \int_{g(x)}^{f(x)} dy dx$$

Proof 16.3.3

Let's apply Theorem 16.3.2 on the rectangle $[a, b] \times [g_m, f_M]$ where g_m is the minimum value of g on $[a, b]$ which we know exists because g is continuous on a compact domain. Similarly f_M is the maximum value of f on $[a, b]$ which also exists. Then the D showing in Figure 16.1 is contained in $[a, b] \times [g_m, f_M]$. We see \hat{F} is continuous on this rectangle except possibly on the curves given by f and g . We also know the set $S_f = \{(x, f(x)) | a \leq x \leq b\}$ and the set $S_g = \{(x, g(x)) | a \leq x \leq b\}$ both have measure zero in \mathbb{R}^2 . Thus \hat{F} has a discontinuity set of measure zero and so F is integrable on $[a, b] \times [g_m, f_M]$ which matches the integral of F on D .

Next, since F is continuous on the interior of D , for fixed x in $[a, b]$, $G(y) = F(x, y)$ is continuous on $[g(x), f(x)]$. Of course, this function need not be continuous at the endpoints but that is a set of measure zero in \mathbb{R}^1 anyway. Hence, \hat{G} is integrable on $[g_m, f_M]$ and hence G is integrable on $[g(x), f(x)]$. By Theorem 16.3.2, we conclude

$$\begin{aligned} \int_{[a,b] \times [g_m, f_M]} F dV_2 &= \int_D F dV_2 = \int_a^b \int_{g_m}^{f_M} G(y) dy dx \\ &= \int_a^b \int_{g(x)}^{f(x)} F(x, y) dy dx \end{aligned}$$

■

Next, look at a typical closed curve situation.

Theorem 16.3.4 Fubini's Theorem: 2D: a top and bottom closed curve

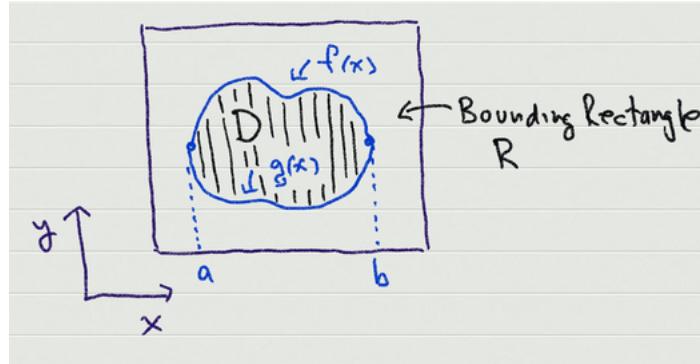


Figure 16.2: Fubini’s Theorem in 2D: a top and bottom closed curve

Let f and g be continuous real valued functions on $[a, b]$ which form the bottom and top halves of a closed curve and $F : D \rightarrow \mathbb{R}$ be continuous on D where $D = \{(x, y) : a \leq x \leq b, g(x) \leq y \leq f(x)\}$. Then

$$\int_D F dV_2 = \int_a^b \int_{g(x)}^{f(x)} dy dx$$

Proof 16.3.4

From Figure 16.2, we see the closed curve defines a a box $[a, b] \times [g_m, f_M]$ where g_m is the minimum value of g and f_M is the maximum value of f on $[a, b]$. These exist because f and g are continuous on the compact domain $[a, b]$. The rest of the argument is exactly the same as the proof of Theorem 16.3.3. We have

$$\begin{aligned} \int_{[a, b] \times [g_m, f_M]} F dV_2 &= \int_D F dV_2 = \int_a^b \int_{g_m}^{f_M} G(y) dy dx \\ &= \int_a^b \int_{g(x)}^{f(x)} F(x, y) dy dx \end{aligned}$$

■

It is easy to see we could prove similar theorems for a region $D = \{(x, y) | c \leq y \leq d, g(y) \leq x \leq f(y)\}$ for functions f and g continuous on $[c, d]$ and F continuous on D . We would find

$$\int_{[g_m, f_M] \times [c, d]} F dV_2 = \int_D F dV_2 = \int_c^d \int_{g(y)}^{f(y)} F(x, y) dx dy$$

Call the top - bottom region a TB region and the left -right region a LR region. Then you should be able to see how you can break a lot of regions into finite combinations of TB and LR regions and apply Fubini’s Theorem to each piece to complete the integration.

Comment 16.3.1 We can work out similar theorems in \mathbb{R}^3 , \mathbb{R}^4 and so forth. Naturally, this process gets complicated as the dimensions increases. Still, the principles of the arguments are the same. The boundary surfaces are all of measure zero as they are a strict subspace of \mathbb{R}^n . You should try a few to see how the arguments are constructed. This is how you learn how things work: by doing.

16.3.2 Homework

Exercise 16.3.1 Prove Fubini’s Theorem for rectangles in \mathbb{R}^3 . It is enough to prove it for one of the six cases.

Exercise 16.3.2 Since we can approximate these integrals using Riemann sums, can you estimate how computationally expensive this gets using uniform partitions? Would it be easy to compute a Riemann approximation to a 10 dimensional integral over a rectangle?

Exercise 16.3.3 This problem shows you how we can use these ideas to do something interesting!

1. Compute the two dimensional $I_R = \int_{D_R} e^{-x^2-y^2}$ where $D_R = B(R, (0, 0))$ for all R by using the polar coordinate transformation. Can you compute $\lim_{R \rightarrow \infty} I_R$?
2. let $J_R = \int_{S_R} e^{-x^2-y^2} dx dy$. Prove $\lim_{R \rightarrow \infty} J_R = \lim_{R \rightarrow \infty} I_R$.
3. What is the value of $\int_{-\infty}^{\infty} e^{-x^2} dx$?

Exercise 16.3.4 Draw a bounded 2D region in \mathbb{R}^2 which is constructed from two TB’s and one LR and convince yourself how Fubini’s theorem is applied to compute the integral.

Exercise 16.3.5 Draw an arbitrary bounded region in \mathbb{R}^2 and decompose it into appropriate TB and LR regions on which Fubini’s Theorem can be applied and convince yourself how the total integral is computed.

Chapter 17

Line Integrals

It is now time to set the stage for the connection between a type of one dimensional integral and an equivalent two dimensional integral. This is challenging material and it has deep meaning. However, we will start out slow in this chapter and show you a low level version of what it all means. Let's start with the idea of a **path** or **curve** in \mathbb{R}^2 .

17.1 Paths

Definition 17.1.1 Two Dimensional Curves

Let $[a, b]$ be a finite interval of real numbers. Let f and g be two continuously differentiable functions on $[a, b]$. The path defined by these functions is the set of ordered pairs

$$\mathcal{C} = \{(x, y) | x = f(t), y = g(t), a \leq t \leq b\}$$

*If the starting and ending point of the curve are the same, we say the curve is **closed**.*

Comment 17.1.1 We often abuse this notation and talk about the path \mathcal{C} being given by the pairs (x, y) with

$$x = x(t), \quad y = y(t), \quad a \leq t \leq b$$

even though the letters x and y are being used in two different ways. In practice, it is not that hard to get the two uses separate.

The curve \mathcal{C} has a tangent line that is defined at each point t_0 given by

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x(t_0) \\ y(t_0) \end{bmatrix} + c \begin{bmatrix} x'(t_0) \\ y'(t_0) \end{bmatrix}$$

and has a tangent vector $\mathbf{T}(t_0)$ given by

$$\mathbf{T}(t_0) = \begin{bmatrix} x'(t_0) \\ y'(t_0) \end{bmatrix}$$

and an associated vector perpendicular to $\mathbf{T}(t_0)$ given

$$\mathbf{N}(t_0) = \begin{bmatrix} -y'(t_0) \\ x'(t_0) \end{bmatrix}$$

You can see $\langle \mathbf{T}(t_0), \mathbf{N}(t_0) \rangle = 0$. Also note

$$\mathbf{N}^*(t_0) = \begin{bmatrix} y'(t_0) \\ -x'(t_0) \end{bmatrix}$$

is also perpendicular to $\mathbf{T}(t_0)$. Both choices for this perpendicular vector are called the **Normal** vector to the curve \mathcal{C} at t_0 . Using the right hand rule, we can calculate $\mathbf{T}(t_0) \times \mathbf{N}(t_0)$. This will either be \mathbf{k} or $-\mathbf{k}$ depending on whether to get to $\mathbf{N}(t_0)$ from $\mathbf{T}(t_0)$, we need to move our right hand counterclockwise (ccw) (this give $+\mathbf{k}$) or clockwise (this gives $-\mathbf{k}$). Look at Figure 17.1 to get a feel for this. We are deliberately showing you hand drawn figures as that is what we would do in a lecture and what you should learn how to do on your scratch paper as you read this.

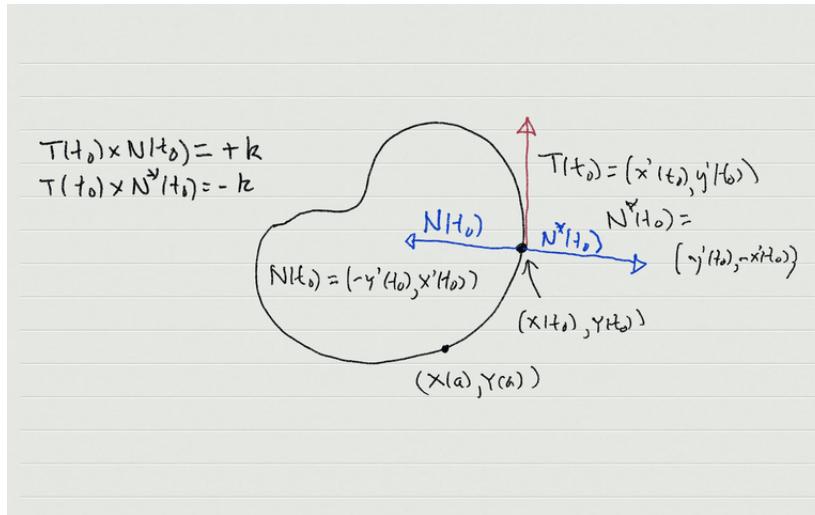


Figure 17.1: A Simple Closed Curve with Tangent and Normal Vectors

In Figure Figure 17.2, we show a self intersecting curve like a figure eight. Note in the right half, the tangent vector moves ccw and in the left half, the tangent vector moves cw. We indicate the direction of both \mathbf{N} and $\mathbf{N}^* = -\mathbf{N}$ also. Note, if \mathbf{T} is chosen, we can not decide which of \mathbf{N} or $-\mathbf{N}$. The point Q here indicates where the tangent vector switches from moving ccw to moving cw. Then in Figure 17.3, we show the same figure eight curve with what we would clearly decide was the inside and outside of the curve. Now the curve \mathcal{C} is defined by the two functions $x(t)$ and $y(t)$ which are the components of a vector field we can call γ . Thus, we can identify the curve \mathcal{C} and the range of the vector function γ . Hence, in Figure 17.3 instead of labeling the curve as \mathcal{C} we call it γ . We use the γ notation with some changes in definition in Chapter 18 where we discuss another way of looking at all of this using the language of differential forms. And, thinking more carefully about insides and outsides of curves will lead us into ideas from homology. So there is much to do! But for right now, we can already see it is not so easy to decide how to pick the normal – should it be \mathbf{N} or $-\mathbf{N}$? It appears it should be \mathbf{N} if \mathbf{T} is moving ccw and $-\mathbf{N}$ if \mathbf{T} is moving cw. Now, scribble out a picture of a curve with has two self intersections, four self intersections with concomitant changes in ccw and cw movement of \mathbf{T} as the parameter t increases. We want to believe, based on simple examples, that each closed curve γ divides \mathbb{R}^2 into three disjoint pieces: the **interior**, **exterior** and the curve itself.

Hence, if we want to point towards the interior of the closed curve, we should use $\mathbf{N}(t)$ if a point on the curve is moving ccw and otherwise, use $-\mathbf{N}(t)$. Also, note we can choose $\pm\mathbf{N}$ and \mathbf{T} to be unit vectors by simply dividing them by their length $\|\mathbf{T}(t)\|$ and $\|\mathbf{N}(t)\|$, respectively. The only time this becomes impossible is these norms are zero. This only occurs at a point t where both $x'(t)$ and

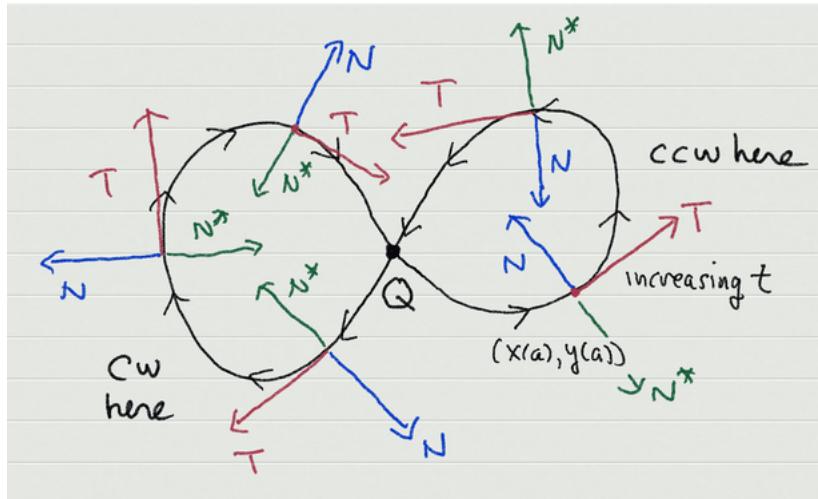


Figure 17.2: A Figure Eight Curve with Tangent and Normal Vectors

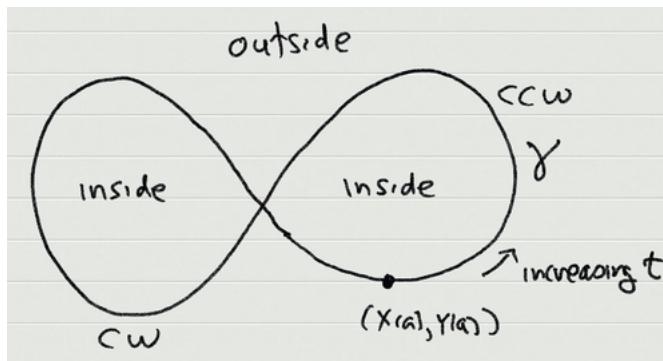


Figure 17.3: The Inside and Outside of a Curve

$y'(t) = 0$. We can usually figure out a way to define a nice tangent and normal anyway but we won't complicate the discussion here by that. Now as long as the functions $x(t)$ and $y(t)$ are continuously differentiable on $[a, b]$, Consider a vector field

$$\mathbf{F} = \begin{bmatrix} A(x, y) \\ B(x, y) \end{bmatrix}$$

defined on a open set U of \mathbb{R}^2 which contains the curve \mathcal{C} . We define the action of \mathbf{F} along \mathcal{C} to be the **line integral** of \mathbf{F} along \mathcal{C} from the starting point of the curve to its ending point using a very specific Riemann integral.

Definition 17.1.2 The Line Integral of a Force Field along a Curve

Let the vector field \mathbf{F} be defined on an open set U in \mathbb{R}^2 be defined by

$$\mathbf{F} = \begin{bmatrix} A(x, y) \\ B(x, y) \end{bmatrix}$$

where A and B are at least continuous at each point on \mathcal{C} . Then the line integral of F along \mathcal{C} from $\mathbf{P} = (x(a), y(a))$ to $\mathbf{q} = (x(b), y(b))$ is defined by

$$\begin{aligned} \int_{\mathcal{C}} \left\langle \mathbf{F}, \begin{bmatrix} x' \\ y' \end{bmatrix} \right\rangle dt &= \int_{\mathcal{C}} \left(\left\langle \begin{bmatrix} A(x, y) \\ B(x, y) \end{bmatrix}, \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} \right\rangle \right) dt \\ &= \int_a^b (A(x(t), y(t))x'(t) + B(x(t), y(t))y'(t)) dt \end{aligned}$$

Since $\Phi(t) = A(x(t), y(t))x'(t) + B(x(t), y(t))y'(t)$ is continuous on $[a, b]$, this Riemann integral exists and so the line integral is well-defined.

Example 17.1.1 Suppose each point on \mathcal{C} represents the electric field at the point t . Then how do we compute the total work done in moving a test charge from the beginning point of the curve to the ending point? First, note the function

$$\Phi(t) = \mathbf{A}(x(t), y(t))x'(t) + \mathbf{B}(x(t), y(t))y'(t)$$

is a continuous function on $[a, b]$ and so is Riemann Integrable. Hence, the value of this integral is given by

$$\lim_{n \rightarrow \infty} S(\Phi, \pi_n, \sigma_n)$$

for any sequence of partitions $\{\pi_n\}$ whose norms go to zero and σ_n is any evaluation set in π_n . Then, for a given partition, the charge in the portion of the path between partition point t_j and t_{j+1} can be approximated by

$$\begin{aligned} \mathbf{A}(x(z_j), y(z_j))(x(t_{j+1}) - x(t_j)) + \mathbf{B}(x(z_j), y(z_j))(y(t_{j+1}) - y(t_j)) &= \\ \left(\mathbf{A}(x(z_j), y(z_j)) \frac{x(t_{j+1}) - x(t_j)}{t_{j+1} - t_j} + \mathbf{B}(x(z_j), y(z_j)) \frac{y(t_{j+1}) - y(t_j)}{t_{j+1} - t_j} \right) (t_{j+1} - t_j) & \end{aligned}$$

for any choice of z_j in $[t_j, t_{j+1}]$. Now apply the Mean Value Theorem to see

$$\begin{aligned} \left(\mathbf{A}(x(z_j), y(z_j)) \frac{x(t_{j+1}) - x(t_j)}{t_{j+1} - t_j} + \mathbf{B}(x(z_j), y(z_j)) \frac{y(t_{j+1}) - y(t_j)}{t_{j+1} - t_j} \right) (t_{j+1} - t_j) &= \\ \left(\mathbf{A}(x(z_j), y(z_j))x'(s_j) + \mathbf{B}(x(z_j), y(z_j))y'(s_j) \right) (t_{j+1} - t_j) & \end{aligned}$$

The approximation to the mass on the portion $[t_j, t_{j+1}]$ can just as well be done setting $z_j = s_j$ in the \mathbf{A} and \mathbf{B} terms. Thus, the approximation can be

$$\Phi(s_j)(t_{j+1} - t_j) = \left(\mathbf{A}(x(s_j), y(s_j))x'(s_j) + \mathbf{B}(x(s_j), y(s_j))y'(s_j) \right) (t_{j+1} - t_j)$$

The Riemann Approximation Theorem then tells us

$$\lim_{n \rightarrow \infty} \sum_{\pi_n} \Phi(s_j) \Delta t_j = \int_a^b \Phi(t) dt$$

which is how we defined the line integral.

Comment 17.1.2 In the same setting as the previous example, we could find the work done along a closed path \mathcal{C} . The same line integral would represent this work.

Example 17.1.2 What if the mass at a point on the string represented by the curve \mathcal{C} was represented by the scalar $m(x(t), y(t))$? The mass of the piece of the string corresponding to the subinterval $[t_j, t_{j+1}]$ for a given partition π of $[a, b]$ can be approximated by

$$\begin{aligned} M_j &= m(x(s_j), y(s_j)) \sqrt{(x(t_{j+1}) - x(t_j))^2 + (y(t_{j+1}) - y(t_j))^2} \\ &= m(x(s_j), y(s_j)) \sqrt{\left(\frac{x(t_{j+1}) - x(t_j)}{t_{j+1} - t_j}\right)^2 + \left(\frac{y(t_{j+1}) - y(t_j)}{t_{j+1} - t_j}\right)^2} (t_{j+1} - t_j) \end{aligned}$$

From the Mean Value Theorem, we know there is an z_j between t_j and t_{j+1} so that

$$\begin{bmatrix} x(t_{j+1}) - x(t_j) \\ y(t_{j+1}) - y(t_j) \end{bmatrix} = (t_{j+1} - t_j) \begin{bmatrix} x'(z_j) \\ y'(z_j) \end{bmatrix}$$

Thus, we have

$$M_j = m(x(s_j), y(s_j)) \sqrt{x'(z_j))^2 + y'(z_j))^2} (t_{j+1} - t_j)$$

Choosing $s_j = z_j$ always, we find the Riemann sums here are

$$S(\Phi, \pi, \sigma) = \sum_{\pi} m(x(z_j), y(z_j)) \sqrt{x'(z_j))^2 + y'(z_j))^2} (t_{j+1} - t_j)$$

where $\Phi(t) = m(x(t), y(t))\sqrt{x'(t)^2 + y'(t)^2}$ which is a nice continuous function on $[a, b]$. Hence, the Riemann Integral $\int_a^b \Phi(t) dt$ exists and can be used to define the mass on the curve. Note, this argument, while similar to the one we used for the line integrals, is not the same as $m(x, y)$ is not a vector function.

17.1.1 Homework

Exercise 17.1.1

Exercise 17.1.2

Exercise 17.1.3

Exercise 17.1.4

Exercise 17.1.5

17.2 Conservative Force Fields

Let's restrict our attention to nice **closed** curves. We want them to be **simple** which means they do not have self intersections so the figure eight type curves are not allowed. We also want them to have finite length as we can calculate from the usual arc length formulae. Recall the length of the curve \mathcal{C} from a to b is

$$L_a^b = \int_a^b \sqrt{x'(t)^2 + y'(t)^2} dt$$

Since we assume $x(t)$ and $y(t)$ are continuously differentiable on $[a, b]$, we know the integrand is continuous on $[a, b]$ and hence it is bounded by some $B > 0$. Thus, $L_a^b \leq B(b - a)$. So our curves do have finite arc length. Such curves are called **rectifiable**. We also want the normal vector to be uniquely defined so it points towards the interior of the closed curve. We usually think of the tangent vector as moving ccw here, so the normal we pick is always the \mathbf{N} instead of the $-\mathbf{N}$. This kind of curve is called **orientable**. Together, using the first letters of these properties, we want to consider SCROCs: i.e. **simple**, **closed**, **rectifiable** and **orientable** curves. So what is a **conservative** force field \mathbf{F} ?

Definition 17.2.1 Conservative Force Field

Let the vector field \mathbf{F} be defined on an open set U in \mathbb{R}^2 by

$$\mathbf{F} = \begin{bmatrix} A(x, y) \\ B(x, y) \end{bmatrix}$$

where A and B are at least continuous at each point in U . Let \mathcal{C} be a SCROC corresponding to the vector function on $[a, b]$ (this domain is dependent on the choice of \mathcal{C}).

$$\gamma = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$$

We say \mathbf{F} is a **Conservative Force Field** if $\int_{\mathcal{C}} \langle \mathbf{F}, \gamma' \rangle dt = 0$ for all SCROCs \mathcal{C} .

Comment 17.2.1 Let the start point be \mathbf{P} and the endpoint be \mathbf{Q} . Let \mathcal{C}_1 be a path that goes from \mathbf{P} to \mathbf{Q} and \mathcal{C}_2 be a path that moves in the reverse. Let's also assume the the path \mathbf{C} obtained by traversing \mathcal{C}_1 followed by \mathcal{C}_2 . The paths are associated with vector functions γ_1 and γ_2 . We would write this as $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2$ for convenience. Then, if \mathbf{F} is conservative, we have

$$\int_{\mathcal{C}} \langle \mathbf{F}, \gamma' \rangle dt = \int_{\mathcal{C}_1} \langle \mathbf{F}, \gamma_1' \rangle + \int_{\mathcal{C}_2} \langle \mathbf{F}, \gamma_2' \rangle = 0$$

This implies

$$\int_{\mathcal{C}_1} \langle \mathbf{F}, \gamma_1' \rangle = \int_{-\mathcal{C}_2} \langle \mathbf{F}, \gamma_2' \rangle$$

where the integration over $-\mathcal{C}_2$ simply indicates we are moving on \mathcal{C}_2 backwards. This says the path we take from \mathbf{P} to \mathbf{Q} doesn't matter. In other words, the value we obtain by the line integral depends only on the start and final point. We call this property **path independence** of the line integral.

Let's assume \mathbf{F} is a conservative force field. Let the start point be $\mathbf{P} = (x_s, y_s)$ and the endpoint be some $\mathbf{Q} = (x_e = x_s + h, y_e = y_s + h)$. Since \mathbf{P} is an interior point of U , there is a ball of radius $r > 0$ about \mathbf{P} which is in U . So we can pick an endpoint \mathbf{Q} like this in this ball. You can look at Figure 17.4 to see the kind of paths we are going to use in the argument below. Since the value of the line integral is independent of path, look at the path

$$\begin{aligned} \mathcal{C}_1 : \gamma_1 &= \begin{bmatrix} x_1(t) = x_s + th \\ y_1(t) = y_s \end{bmatrix}, \quad 0 \leq t \leq 1 \\ \mathcal{C}_2 : \gamma_2 &= \begin{bmatrix} x_2(t) = x_s + h \\ y_2(t) = y_s + th \end{bmatrix}, \quad 0 \leq t \leq 1 \end{aligned}$$

17.2. CONSERVATIVE FORCE FIELDS

345

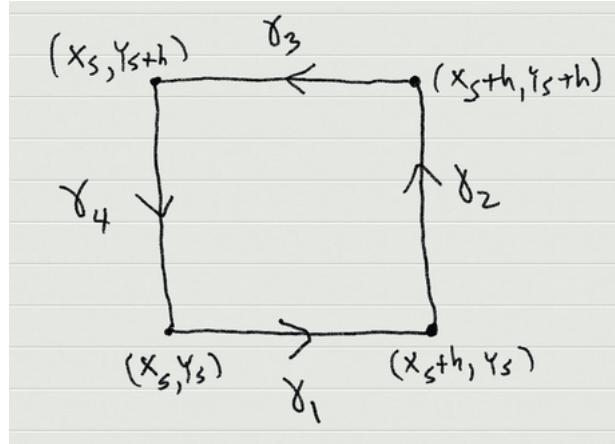


Figure 17.4: A Closed Path to prove $A_y = B_x$

Then, we can calculate the line integrals $\int_{\mathcal{C}_1}$ and $\int_{\mathcal{C}_2}$ to find

$$\begin{aligned} & \int_{\mathcal{C}_1} \langle \mathbf{F}, \gamma_1' \rangle + \int_{\mathcal{C}_2} \langle \mathbf{F}, \gamma_2' \rangle \\ &= \int_0^1 A(x_s + th, y_s) h dt + \int_0^1 B(x_s, y_s + th) h dt \end{aligned}$$

Now we can go back to P along the path

$$\begin{aligned} \mathcal{C}_3 : \gamma_3 &= \begin{bmatrix} x_3(t) = x_e - th = x_s + (1-t)h \\ y_3(t) = y_e \end{bmatrix}, \quad 0 \leq t \leq 1 \\ \mathcal{C}_4 : \gamma_4 &= \begin{bmatrix} x_4(t) = x_s \\ y_4(t) = y_e - th = y_s + (1-t)h \end{bmatrix}, \quad 0 \leq t \leq 1 \end{aligned}$$

Then, we can calculate the line integrals $\int_{\mathcal{C}_3}$ and $\int_{\mathcal{C}_4}$ to find

$$\begin{aligned} & \int_{\mathcal{C}_3} \langle \mathbf{F}, \gamma_3' \rangle + \int_{\mathcal{C}_4} \langle \mathbf{F}, \gamma_4' \rangle \\ &= \int_0^1 A(x_e - th, y_e) (-h) dt + \int_0^1 B(x_s, y_e - th) (-h) dt \end{aligned}$$

Because of path independence, we have $\int_{\mathcal{C}_1} + \int_{\mathcal{C}_2} = - \int_{\mathcal{C}_3} - \int_{\mathcal{C}_4}$ and so

$$\begin{aligned} & \int_0^1 A(x_s + th, y_s) h dt + \int_0^1 B(x_s, y_s + th) h dt \\ &= - \int_0^1 A(x_e - th, y_e) (-h) dt - \int_0^1 B(x_s, y_e - th) (-h) dt \end{aligned}$$

Thus,

$$h \int_0^1 \left(A(x_s + th, y_s) - A(x_e - th, y_e) \right) dt$$

$$= h \int_0^1 \left(B(x_s, y_e - th) - B(x_e, y_s + th) \right) dt$$

Hence, cancelling the common h ,

$$\int_0^1 \left(A(x_s + th, y_s) - A(x_e - th, y_e) \right) dt = \int_0^1 \left(B(x_s, y_e - th) - B(x_e, y_s + th) \right) dt$$

Next, let $h \rightarrow 0$:

$$\begin{aligned} & \lim_{h \rightarrow 0} \int_0^1 \left(A(x_s + th, y_s) - A(x_e - th, y_e) \right) dt \\ &= \lim_{h \rightarrow 0} \int_0^1 \left(B(x_s, y_e - th) - B(x_e, y_s + th) \right) dt \end{aligned}$$

The integrals are continuous with respect to h and so

$$\int_0^1 \left(A(x_s, y_s) - A(x_s, y_e) \right) dt = \int_0^1 \left(B(x_s, y_s) - B(x_e, y_s) \right) dt$$

If we assume A and B have first order partials, then

$$\int_0^1 \left(-A_y(x_s, y_s)h - E_A(x_s, y_s, h) \right) dt = \int_0^1 \left(-B_x(x_s, y_s)h - E_B(x_s, y_s, h) \right) dt$$

Thus,

$$A_y(x_s, y_s) + \int_0^1 \frac{E_A(x_s, y_s, h)}{h} dt = \left(B_x(x_s, y_s) + \int_0^1 \frac{E_B(x_s, y_s, h)}{h} dt \right)$$

Now let $h \rightarrow 0$ again to obtain the final result: $A_y(x_s, y_s) = B_x(x_s, y_s)$. We can state this as an important result.

Theorem 17.2.1 **Conservative Force Fields Imply** $A_y = B_x$

Let the conservative vector field \mathbf{F} be defined on an open set U in \mathbb{R}^2 by

$$\mathbf{F} = \begin{bmatrix} A(x, y) \\ B(x, y) \end{bmatrix}$$

where A and B have first order partials at each point in U . Then $A_y = B_x$ in U .

Proof 17.2.1

We have just gone through this argument. ■

17.2.1 Homework

Exercise 17.2.1

Exercise 17.2.2

Exercise 17.2.3

Exercise 17.2.4

Exercise 17.2.5

17.3 Potential Functions

To understand this at a deeper level, we need to look at connectedness more. Let's introduce the idea of a set U being **path connected**.

Definition 17.3.1 Path Connected Sets

The set V in \mathbb{R}^2 is path connected if given P and Q in V , there is a path \mathcal{C} connecting P to Q .

Comment 17.3.1 It is not true that a connected open set has to be path connected. There is a standard counterexample of this which you should see. We need to extend connectedness to closed sets.

Definition 17.3.2 Connected Closed Sets

The closed set C in \mathbb{R}^2 is connected if we can not find two open sets A and B so that $C = A' \cup B'$ with $A' \cap B = \emptyset$ and $B' \cap A = \emptyset$.

To find our counterexample, consider

$$S = \{(x, y) \in \mathbb{R}^2 \mid y = \sin(1/x), x > 0\} \cup (\{0\} \times [-1, 1])$$

Let $S_+ = \{(x, y) \in \mathbb{R}^2 \mid y = \sin(1/x), x > 0\}$ and $S_0 = \{0\} \times [-1, 1]$. It is easy to see any two points in S_+ can be connected by a path. Just parameterize the portion of the curve $y = \sin(1/x)$ that connects the two points: $x(t) = t$ and $y(t) = \sin(1/t)$ for suitable $[a, b]$. Now take a point in S_+ and a point in S_0 and try to find a path that connects the two points. There is no way a path that begins in S_+ can jump to S_0 . Hence, the set S is not path connected. If you think about it though, S itself is connected.

Since the set of cluster points of $\sin(1/x)$ at 0 in $[-1, 1]$, the closure of S_+ must be all of S . Now if S was not connected, we could write $S = A' \cup B'$ where A and B are open sets with $A' \cap B = \emptyset$ and $B' \cap A = \emptyset$. But then $A \cap S$ and $B \cap S$ is a decomposition of S and since S_+ is connected, we can assume without loss of generality $B \cap S_+ = \emptyset$. Hence, $S_+ = A \cap S_+ \subset A$ which implies $S \subset A'$. That tells us S does not intersect B' and so $S = A'$ implying S is connected. This argument works in the other case, $A \cap S = \emptyset$ as well. Thus, as S_+ is connected, its closure S must be connected also.

So S is a subset of \mathbb{R}^2 which is connected but not path connected.

We can now say much more about conservative force fields. Let's consider a force field \mathbf{F} in U and now assume \mathbf{A} and \mathbf{A} to have continuous first order partials with $A_y = -B_x$ in U . Let's also assume the open set U is path connected. Pick a point (x_0, y_0) in U and any other point (x, y) in U . Since $A_y = B_x$ in U , we know the line integrals over paths connecting these two points are independent of path. The point (x, y) is an interior point in U and so there is a radius $r > 0$ with $B_r(x, y)$ in U . For h sufficiently small, the lines connecting (x, y) to $(x + h, y)$ and $(x + h, y)$ to $(x, y + h)$ are in U too. Let a curve connecting (x_0, y_0) to (x, y) be chosen and call it \mathcal{C} with corresponding vector function γ . Define the function $\phi(x, y) = \int_{\mathcal{C}} \langle \mathbf{F}, \gamma' dt \rangle$ as usual. Then since we are free to choose any path to get from (x, y) to $(x + h, y + h)$ we wish, we can consider the new paths

$$\begin{aligned}\mathcal{C}_1 : \gamma_1 &= \begin{bmatrix} x_1(t) = x + th \\ y_1(t) = y \end{bmatrix}, \quad 0 \leq t \leq 1 \\ \mathcal{C}_2 : \gamma_2 &= \begin{bmatrix} x_2(t) = x + h \\ y_2(t) = y + th \end{bmatrix}, \quad 0 \leq t \leq 1\end{aligned}$$

We illustrate this in Figure 17.5. Then

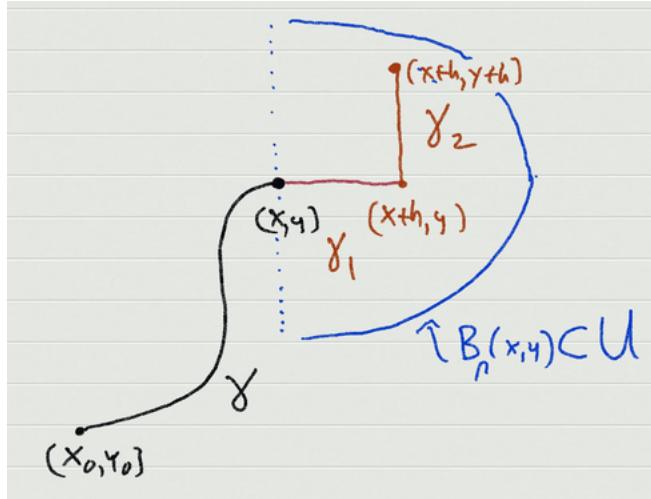


Figure 17.5: Defining the function ϕ on U

$$\lim_{h \rightarrow 0^+} \frac{\phi(x + h, y) - \phi(x, y)}{h} = \lim_{h \rightarrow 0^+} \frac{1}{h} \int_0^1 \mathbf{A}(x + th, y) dt = \lim_{h \rightarrow 0^+} \int_0^1 \mathbf{A}(x + th, y) dt$$

Now apply the Mean Value Theorem for integrals to find

$$\lim_{h \rightarrow 0^+} \frac{\phi(x + h, y) - \phi(x, y)}{h} = \lim_{h \rightarrow 0^+} \mathbf{A}(x + ch, y)$$

where c is between 0 and h . Hence, as $h \rightarrow 0^+$, we find $(\phi_x(x, y))^+ = \mathbf{A}(x, y)$. A similar argument, with a slightly different picture of course, shows $(\phi_x(x, y))^- = \mathbf{A}(x, y)$. Hence, $\phi_x = \mathbf{A}$.

We can argue in a very similar fashion to show $\phi_y = \mathbf{B}$. We have proven another result. This function ϕ is called a **potential** function corresponding to \mathbf{F} .

Theorem 17.3.1 The Potential Function for \mathbf{F}

Let the vector field \mathbf{F} be defined on an path connected open set U in \mathbb{R}^2 by

$$\mathbf{F} = \begin{bmatrix} A(x, y) \\ B(x, y) \end{bmatrix}$$

where A and B have continuous first order partials in U . Then \mathbf{F} is a conservative force field if and only if there is a function ϕ on U so that $\nabla\phi = \mathbf{F}$. Moreover, for any path \mathcal{C} with start point $\mathbf{P} = (x_s, y_s)$ and end point $\mathbf{Q} = (x_e, y_e)$,

$$\int_{\mathcal{C}} \langle \mathbf{F}, \gamma' \rangle dt = \phi(x_e, y_e) - \phi(x_s, y_s)$$

Proof 17.3.1

(\Rightarrow):

If \mathbf{F} is a conservative force field, we know the line integrations are path independent and by the

argument above, we can find the desired ϕ function.

(\Leftarrow):

If such a function ϕ exists, then $\mathbf{A}_x = \mathbf{B}_y$ and \mathbf{F} is a conservative force field.

Finally, if \mathbf{F} is conservative

$$\begin{aligned}\int_{\mathcal{C}} < \mathbf{F}, \gamma' > dt &= \int_a^b \left(\mathbf{A}(x(t), y(t))x'(t) + \mathbf{B}(x(t), y(t))y'(t) \right) dt \\ &= \int_a^b \left(\phi_x(x(t), y(t))x'(t) + \phi_y(x(t), y(t))y'(t) \right) dt \\ &= \int_a^b \phi'(t) dt = \phi(x_e, y_e) - \phi(x_s, y_s)\end{aligned}$$

■

17.3.1 Homework

Exercise 17.3.1

Exercise 17.3.2

Exercise 17.3.3

Exercise 17.3.4

Exercise 17.3.5

17.4 Finding Potential Functions

17.4.1 Homework

Exercise 17.4.1

Exercise 17.4.2

Exercise 17.4.3

Exercise 17.4.4

Exercise 17.4.5

17.5 Green’s Theorem

There is a connection between the line integral over a SCROC and a double integral over the area enclosed by the curve. From our earlier discussions, since curves can be very complicated with their interiors and exteriors not so clear, by restricting our interest to a SCROC, we can look carefully at the meat of the idea and not get lost in extraneous details. Remember this is what we always must do: look for the **core** of the thing. We start with a very simple SCROC as shown in Figure 17.6. The curve is the finite rectangle $[a, b] \times [c, d]$, the open set $U = (a, b) \times (c, d)$ and we let ∂U , the boundary of U be given by the curve which traverses the rectangle ccw. Thus, the normal vector always points in.

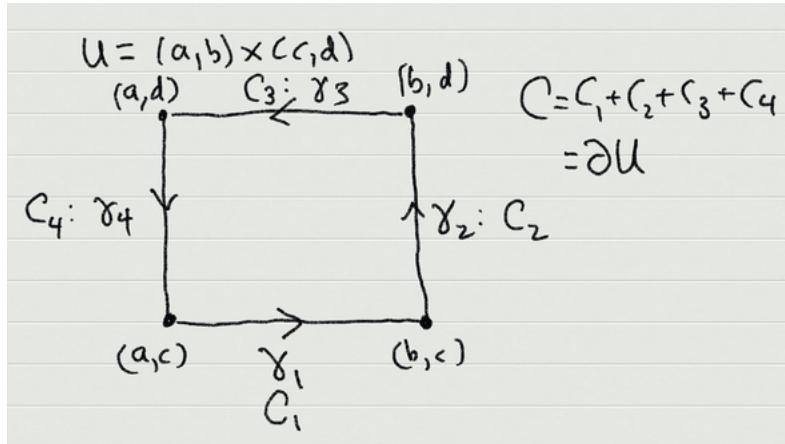


Figure 17.6: Green’s Theorem for a Rectangle

The curve $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3 + \mathcal{C}_4$ as shown in the Figure. We use very simple definitions for γ_1 through γ_4 here.

$$\begin{aligned}\mathcal{C}_1 : \gamma_1 &= \begin{bmatrix} x, & a \leq x \leq b \\ c & \end{bmatrix}, & \mathcal{C}_2 : \gamma_2 &= \begin{bmatrix} x+h \\ y, & c \leq y \leq d \end{bmatrix} \\ \mathcal{C}_3 : \gamma_3 &= \begin{bmatrix} x, & b \rightarrow a \\ d & \end{bmatrix}, & \mathcal{C}_4 : \gamma_4 &= \begin{bmatrix} a \\ y, & d \rightarrow c \end{bmatrix},\end{aligned}$$

The line integral over \mathcal{C} is then

$$\begin{aligned}&\int_a^b \mathbf{A}(x, c) dx + \int_c^d \mathbf{B}(b, y) dy + \int_b^a \mathbf{A}(x, d) dx + \int_d^c \mathbf{B}(a, y) dy = \\&\int_a^b \left(\mathbf{A}(x, c) - \mathbf{A}(x, d) \right) dx + \int_c^d \left(\mathbf{B}(b, y) - \mathbf{B}(a, y) \right) dy \\&\int_a^b \int_c^d -\mathbf{A}_y(x, y) dy dx + \int_c^d \int_a^b \mathbf{B}_x(x, y) dx dy\end{aligned}$$

This can be rewritten as

$$\int_{\mathcal{C}} \langle \mathbf{F}, \gamma' \rangle dt = \int_a^b \int_c^d (\mathbf{B}_x - \mathbf{A}_y) dx dy$$

This is our first version of this result

$$\int_{\partial U} \langle \mathbf{F}, \gamma' \rangle dt = \iint_U (\mathbf{B}_x - \mathbf{A}_y) dA$$

where dA is the usual area element notation in a double integral. The open set U here is particularly simple. Next, let’s look at a more interesting ∂U . This is a SCROC which encloses an area which can be described either with a left and right curve (so the area is $\int \int dx dy$) or with a bottom and top curve (so the area is $\int \int dy dx$).

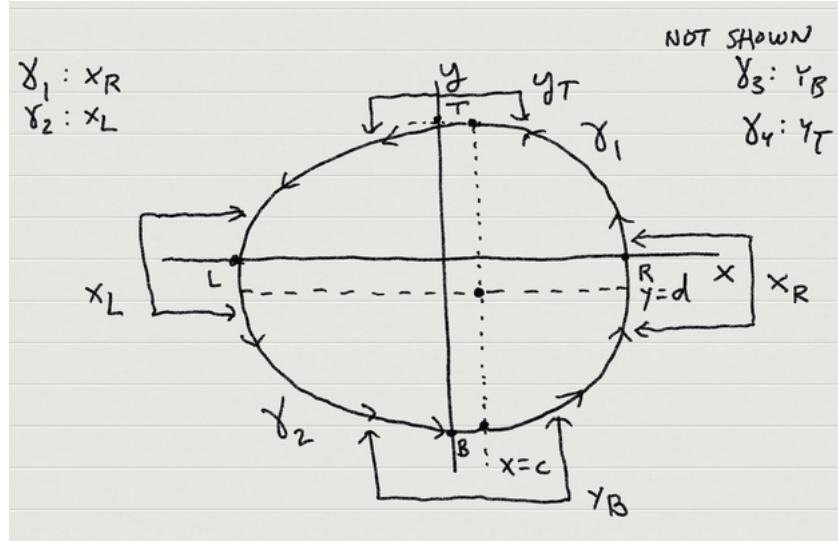


Figure 17.7: Green's Theorem for a SCROC: Right/ Left and Bottom/ Top

The curve $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2$ as shown in the Figure. We use

$$\mathcal{C}_1 : \gamma_1 = \begin{bmatrix} x = x_R(y), \\ y, B \leq y \leq T \end{bmatrix}, \quad \mathcal{C}_2 : \gamma_2 = \begin{bmatrix} x = x_L(y) \\ y, T \rightarrow B \end{bmatrix}$$

The line integral over \mathcal{C} is then

$$\begin{aligned} & \int_B^T \mathbf{A}(x_R(y), y) x'_R(y) dy + \int_B^T \mathbf{B}(x_R(y), y) dy - \int_B^T \mathbf{A}(x_L(y), y) x'_L(y) dy \\ & - \int_B^T \mathbf{B}(x_L(y), y) dy = \int_B^T \mathbf{A}(x_R(y), y) x'_R(y) dy - \int_B^T \mathbf{A}(x_L(y), y) x'_L(y) dy \\ & + \int_B^T (\mathbf{B}(x_R(y), y) - \mathbf{B}(x_L(y), y)) dy \end{aligned}$$

First, note

$$\int_B^T (\mathbf{B}(x_R(y), y) - \mathbf{B}(x_L(y), y)) dy = \int_B^T \int_{x_L(y)}^{x_R(y)} \mathbf{B}_x(x, y) dx dy$$

Next, consider

$$\begin{aligned} & \int_B^T \mathbf{A}(x_R(y), y) x'_R(y) dy - \int_B^T \mathbf{A}(x_L(y), y) x'_L(y) dy = \int_B^d \mathbf{A}(x_R(y), y) x'_R(y) dy \\ & + \int_d^T \mathbf{A}(x_R(y), y) x'_R(y) dy - \int_B^d \mathbf{A}(x_L(y), y) x'_L(y) dy - \int_d^T \mathbf{A}(x_L(y), y) x'_L(y) dy \end{aligned}$$

We can regroup this as follows:

$$\int_B^d \left(\mathbf{A}(x_R(y), y) x'_R(y) - \mathbf{A}(x_L(y), y) x'_L(y) \right) dy$$

$$+ \int_d^T \left(\mathbf{A}(x_R(y), y) x'_R(y) - \mathbf{A}(x_L(y), y) x'_L(y) \right) dy$$

Now when $x = x_R(y)$ on $[B, d]$, $y = y_B(x)$ on $[c, R]$; on $[d, T]$, $y = y_T(x)$ on $[R, c]$. Also, when $x = x_L(y)$ on $[B, d]$, $y = y_B(x)$ on $[c, L]$; on $[d, T]$, $y = y_T(x)$ on $[L, c]$. Making these changes of variables, we now have

$$\begin{aligned} & \int_B^T \mathbf{A}(x_R(y), y) x'_R(y) dy - \int_B^T \mathbf{A}(x_L(y), y) x'_L(y) dy = \\ & \int_c^R (\mathbf{A}(x, y_B(x)) - \mathbf{A}(x, y_T(x))) dx + \int_c^L (\mathbf{A}(x, y_B(x)) - \mathbf{A}(x, y_T(x))) dy \end{aligned}$$

or

$$\begin{aligned} & \int_B^T \mathbf{A}(x_R(y), y) x'_R(y) dy - \int_B^T \mathbf{A}(x_L(y), y) x'_L(y) dy = \\ & \int_L^R (\mathbf{A}(x, y_B(x)) dx - \mathbf{A}(x, y_T(x)) dx = \int_L^R \int_{y_B(x)}^{y_T(x)} -\mathbf{A}_y(x, y) dx dy \end{aligned}$$

Combining our results, we see we have shown

$$\int_{\mathcal{C}} \langle \mathbf{F}, \gamma' \rangle dt = \int_a^b \int_c^d (\mathbf{B}_x - \mathbf{A}_y) dx dy$$

This is our second version of this result

$$\int_{\partial U} \langle \mathbf{F}, \gamma' \rangle dt = \int_U \int (\mathbf{B}_x - \mathbf{A}_y) dA$$

Also, we would get the same result if we had chosen to move around the boundary of our region using the y_B and y_T curves which would have determined γ_3 and γ_4 . For our arguments to work, it was essential that we could invert the equations $x = x_R(y)$ and $x = x_L(y)$ on suitable intervals. If the region determined by the SCROC, doesn't allow us to do this, we have difficulties as we can see in the next example. We now examine a SCROC laid out bottom to top. We show this in Figure 17.8. In this figure, for a change of pace, we label the bottom and top curves f_B and f_T , respectively. You should get used to using a variety of notations for these kinds of arguments.

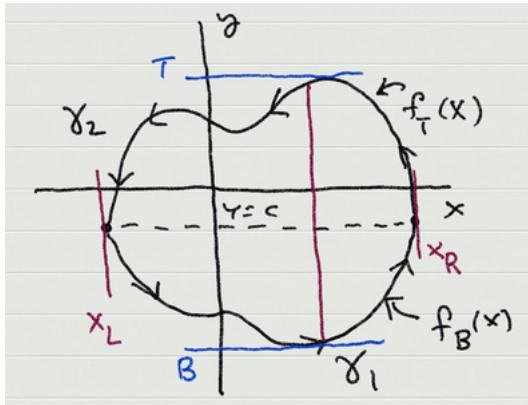


Figure 17.8: Green’s Theorem for a SCROC: Bottom Top

17.5. GREEN'S THEOREM

353

The curve $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2$ as shown in the Figure. We use

$$\mathcal{C}_1 : \gamma_1 = \begin{bmatrix} x, x_L \leq x \leq x_R \\ y = f_B(x), \end{bmatrix}, \quad \mathcal{C}_2 : \gamma_2 = \begin{bmatrix} x, x_R \rightarrow x_L \\ y = f_T(x), \end{bmatrix}$$

The line integral over \mathcal{C} is then

$$\begin{aligned} & \int_{x_L}^{x_R} \mathbf{A}(x, f_B(x)) dx + \int_{x_L}^{x_R} \mathbf{B}(x, f_B(x)) f'_B(x) dx + \int_{x_R}^{x_L} \mathbf{A}(x, f_T(x)) dx \\ & + \int_{x_R}^{x_L} \mathbf{B}(x, f_T(x)) f'_T(x) dx = \\ & \int_{x_L}^{x_R} \left(A(x, f_B(x)) - A(x, f_T(x)) \right) dx + \int_{x_L}^{x_R} \mathbf{B}(x, f_B(x)) f'_B(x) dx \\ & - \int_{x_L}^{x_R} \mathbf{B}(x, f_T(x)) f'_T(x) dx \end{aligned}$$

The first integral can be rewritten giving

$$\begin{aligned} & \int_{x_L}^{x_R} \mathbf{A}(x, f_B(x)) dx + \int_{x_L}^{x_R} \mathbf{B}(x, f_B(x)) f'_B(x) dx + \int_{x_R}^{x_L} \mathbf{A}(x, f_T(x)) dx \\ & + \int_{x_R}^{x_L} \mathbf{B}(x, f_T(x)) f'_T(x) dx = \int_{x_L}^{x_R} \int_{f_B(x)}^{f_T(x)} -\mathbf{A}_y(x, y) dy dx \\ & + \int_{x_L}^{x_R} \mathbf{B}(x, f_B(x)) f'_B(x) dx - \int_{x_L}^{x_R} \mathbf{B}(x, f_T(x)) f'_T(x) dx \end{aligned}$$

However, in the last two integrals above, we can not use the arguments from the last example as we don't know how to invert y_B and y_T on appropriate intervals. This is why the last example was structured the way it was. We need the inversion information. But, all is not lost. Consider the region shown in Figure 17.9.

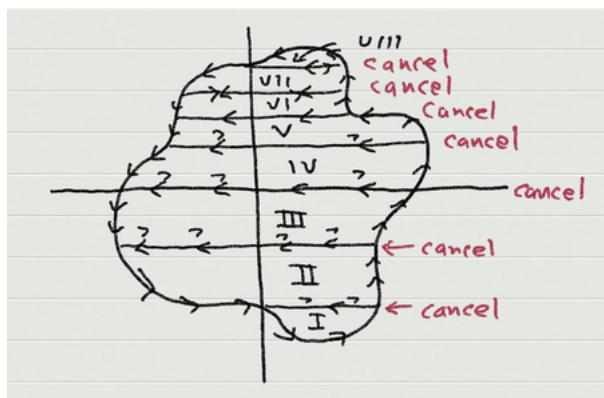


Figure 17.9: Green's Theorem for a more general SCROC

In this more general case, we divide the region into pieces which have the right structure (i.e. a nice x_L, x_R, y_B and y_T) so that the inversions can be done. In the figure, you see we define many closed paths which are set up so that the horizontal movements across the figure always cancel out. The lower path and the path above it move in opposite directions and so the net contribution to the line integral is zero. The part that is nonzero is the line integral we want. The second version of Green's

Theorem applies to each piece. So in general, we can prove Green’s Theorem for any SCROC although it does get quite involved!

So let’s state Green’s Theorem now.

Theorem 17.5.1 Green’s Theorem for a SCROC

Let \mathcal{C} be a SCROC enclosing the open set U in \mathbb{R}^2 . Let γ be the vector function used in \mathcal{C} and let ∂U denote the boundary of U which is traversed ccw by γ . Then, for any vector field \mathbf{F}

$$\int_{\partial U} \langle \mathbf{F}, \gamma' \rangle dt = \iint_U (\mathbf{B}_x - \mathbf{A}_y) dA$$

where \mathbf{A} and \mathbf{B} are the components of \mathbf{F} .

17.5.1 Homework

Exercise 17.5.1

Exercise 17.5.2

Exercise 17.5.3

Exercise 17.5.4

Exercise 17.5.5

17.6 Green’s Theorem for Images of the unit square

We can also prove a version of Green’s Theorem for mappings that smoothly take $[0, 1] \times [0, 1]$ to \mathbb{R}^2 in a one-to-one way.

Definition 17.6.1 Oriented Two Cells

We assume the coordinates of \mathbb{I}^2 are (u, v) and the coordinates of the image $\mathbf{F}(\mathbb{I}^2) = D$ are (x, y) . We say D in \mathbb{R}^2 is an oriented two cell if there is a 1-1 continuously differentiable map $\mathbf{G} : [0, 1] \times [0, 1] \rightarrow D$ where \mathbf{G} is actually defined on an open set U containing $[0, 1] \times [0, 1]$ satisfying $\mathbf{G}([0, 1] \times [0, 1]) = D$. In addition, we assume $\det \mathbf{J}_{\mathbf{G}}(u, v) > 0$ on U . For convenience, we let $\mathbb{I}^2 = [0, 1] \times [0, 1]$.

Consider the picture shown in Figure 17.10. We use the \mathbf{G}_1 and \mathbf{G}_2 as the component functions for \mathbf{F} . Thus,

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{bmatrix} \implies \mathbf{J}_{\mathbf{G}}(u, v) = \begin{bmatrix} \mathbf{G}_{1u}(u, v) & \mathbf{G}_{1v}(u, v) \\ \mathbf{G}_{2u}(u, v) & \mathbf{G}_{2v}(u, v) \end{bmatrix}$$

The edges of \mathbb{I}^2 are mapped by \mathbf{G} into curves in \mathbb{R}^2 as shown in the figure. The edge corresponding to γ_1 is the trace $\mathbf{G}(u, 0)$ and the tangent along this trace is

$$\mathbf{T}(u, 0) = \begin{bmatrix} \mathbf{G}_{1u}(u, 0) & \mathbf{G}_{1v}(u, 0) \\ \mathbf{G}_{2u}(u, 0) & \mathbf{G}_{2v}(u, 0) \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = [\mathbf{G}_{1u}(u, 0)\mathbf{G}_{2u}(u, 0))]$$

Hence, the limiting tangent vector as we approach $(1, 0)$ along $\mathbf{F}(\gamma_1)$ is

$$\mathbf{T}_{\gamma_1}(1, 0) = [\mathbf{G}_{1u}(1, 0)\mathbf{G}_{2u}(1, 0))]$$

17.6. GREEN'S THEOREM FOR IMAGES OF THE UNIT SQUARE

355

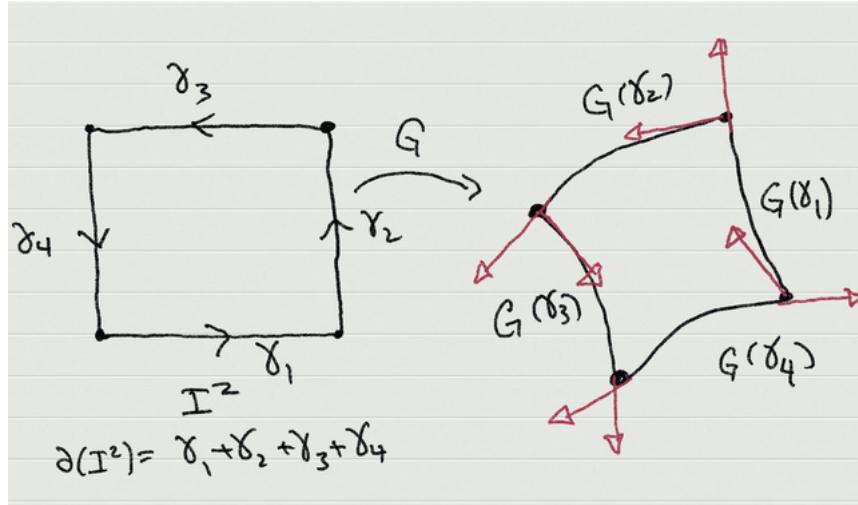


Figure 17.10: An Oriented Two Cell

On the other hand, if we look at the trace for the edge given by γ_2 , we find

$$\mathbf{T}(1, y) = \begin{bmatrix} \mathbf{G}_{1u}(1, v) & \mathbf{G}_{2v}(1, v) \\ \mathbf{G}_{2u}(1, v) & \mathbf{G}_{2v}(1, v) \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = [\mathbf{G}_{1v}(1, v)\mathbf{G} - \mathbf{G}_{2u}(1, v)]$$

At the point $(1, 0)$, on the curve $\mathbf{F}(\gamma_2)$, we find the limiting tangent vector

$$\mathbf{T}_{\gamma_2}(1, 0) = [\mathbf{G}_{1v}(1, 0)\mathbf{G}_{2v}(1, 0)]$$

Note

$$\mathbf{T}_{\gamma_1}(1, 0) \times \mathbf{T}_{\gamma_2} = \det J_{\mathbf{g}}(1, 0) \mathbf{k} = (\mathbf{G}_{1u}(1, 0)\mathbf{G}_{2v}(1, 0) - \mathbf{G}_{1v}(1, 0)\mathbf{G}_{2u}(1, 0)) > 0$$

by assumption. Hence as we rotate $\mathbf{T}_{\gamma_1}(1, 0)$ into $\mathbf{T}_{\gamma_2}(1, 0)$, by the right hand rule the cross product points up. Hence, the curve is moving ccw because of the assumption on the positivity of the determinant of the Jacobian. We can do this sort of analysis at all the vertices we get when the edges in \mathbb{R}^2 under the map \mathbf{F} are joined together. Since $\det J_{\mathbf{G}}(1, 0)$, it is not possible for the tangent vectors at the vertices to align as if they did, the determinant would be zero there. Thus, the mapping \mathbf{G} will always generate four curves with four vertices where the tangent vectors do not line up. We often call the angles between the limiting value tangent vectors at the vertices the **interior** angles of D . From what we have just said, since the curve must be moving ccw, these interior angles must be less than π . That rules out some D 's we could draw.

To connect the line integrals around $\partial\mathbb{I}^2$ to the line integrals around $\mathbf{G}(\partial\mathbb{I}^2)$, we define what is called the **pullback** of \mathbf{F} where \mathbf{F} is the vector field in the line integral. This *pulls* \mathbf{F} which acts on $\mathbf{G}(\partial\mathbb{I}^2)$ in (x, y) space *back* to a function defined on (u, v) space. The pullback of \mathbf{F} is denoted by \mathbf{F}^* and is defined by

$$(\mathbf{F}^*)(u, v) = [A(\mathbf{G}_1(u, v), \mathbf{G}_2(u, v)) \quad B(\mathbf{G}_1(u, v), \mathbf{G}_2(u, v))] \begin{bmatrix} \mathbf{G}_{1u} & \mathbf{G}_{1v} \\ \mathbf{G}_{2u} & \mathbf{G}_{2v} \end{bmatrix}$$

where, for convenience, we leave out the (u, v) in the Jacobian. So the line integral is

$$\int_{\partial\mathbb{I}^2} \langle \mathbf{F}^*, \gamma' \rangle dt = \int_{\partial\mathbb{I}^2} \left\langle \begin{bmatrix} \mathbf{A}(\mathbf{G}_1, \mathbf{G}_2) \\ \mathbf{B}(\mathbf{G}_1, \mathbf{G}_2) \end{bmatrix}, \begin{bmatrix} \mathbf{G}_{1u} & \mathbf{G}_{1v} \\ \mathbf{G}_{2u} & \mathbf{G}_{2v} \end{bmatrix} \gamma' \right\rangle dt$$

Now make the change of variables $x = \mathbf{G}_1(u, v)$ and $y = \mathbf{G}_2(u, v)$. Then

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{1u}(u, v) & \mathbf{G}_{1v}(u, v) \\ \mathbf{G}_{2u}(u, v) & \mathbf{G}_{2v}(u, v) \end{bmatrix} \begin{bmatrix} du \\ dv \end{bmatrix}$$

Thus, after the change of variables, we have

$$\int_{\partial\mathbb{I}^2} \langle \mathbf{F}^*, \gamma' \rangle dt = \int_{\mathbf{G}(\partial\mathbb{I}^2)} \left\langle \begin{bmatrix} \mathbf{A}(x, y) \\ \mathbf{B}(x, y) \end{bmatrix}, (\mathbf{G}(\gamma))' \right\rangle dt$$

Thus, we know how to define the line integral around the boundary of the image of the unit square under a nice mapping. Green’s Theorem applies to the left side, so we have

$$\begin{aligned} \iint_{\mathbb{I}^2} \left((\mathbf{A}\mathbf{G}_{1v} + \mathbf{B}\mathbf{G}_{2v})_u - (\mathbf{A}\mathbf{G}_{1u} + \mathbf{B}\mathbf{G}_{2u})_v \right) dA_{uv} = \\ \int_{\mathbf{G}(\partial\mathbb{I}^2)} \left\langle \begin{bmatrix} \mathbf{A}(x, y) & \mathbf{B}(x, y) \end{bmatrix}, (\mathbf{G}(\gamma))' \right\rangle dt \end{aligned}$$

Let’s simplify the integrand for \iint over \mathbb{I}^2 . We use $\mathbf{A}_1, \mathbf{A}_2$ etc. to indicate the partials of \mathbf{A} with respect to the first and second argument. Also, we must now assume more smoothness of \mathbf{G} as we want to take second order partials and we want the mixed order partials to match. So now \mathbf{G} has continuous second order partials.

$$\begin{aligned} (\mathbf{A}\mathbf{G}_{1v} + \mathbf{B}\mathbf{G}_{2v})_u - (\mathbf{A}\mathbf{G}_{1u} + \mathbf{B}\mathbf{G}_{2u})_v = \\ (\mathbf{A}_1\mathbf{G}_{1u} + \mathbf{A}_2\mathbf{G}_{2u})\mathbf{G}_{1v} + \mathbf{A}\mathbf{G}_{1vu} + (\mathbf{B}_1\mathbf{G}_{1u} + \mathbf{B}_2\mathbf{G}_{2u})\mathbf{G}_{2v} + \mathbf{B}\mathbf{G}_{2vu} \\ - (\mathbf{A}_1\mathbf{G}_{1v} + \mathbf{A}_2\mathbf{G}_{2v})\mathbf{G}_{1u} - \mathbf{A}\mathbf{G}_{1uv} - (\mathbf{B}_1(\mathbf{G}_{1v} + \mathbf{B}_2\mathbf{G}_{2v})\mathbf{G}_{2u} - \mathbf{B}\mathbf{G}_{2uv} \end{aligned}$$

Now cancel some terms to get

$$\begin{aligned} (\mathbf{A}\mathbf{G}_{1v} + \mathbf{B}\mathbf{G}_{2v})_u - (\mathbf{A}\mathbf{G}_{1u} + \mathbf{B}\mathbf{G}_{2u})_v = (\mathbf{A}_1\mathbf{G}_{1u} + \mathbf{A}_2\mathbf{G}_{2u})\mathbf{G}_{1v} \\ + (\mathbf{B}_1\mathbf{G}_{1u} + \mathbf{B}_2\mathbf{G}_{2u})\mathbf{G}_{2v} - (\mathbf{A}_1\mathbf{G}_{1v} + \mathbf{A}_2\mathbf{G}_{2v})\mathbf{G}_{1u} - (\mathbf{B}_1\mathbf{G}_{1v} + \mathbf{B}_2\mathbf{G}_{2v})\mathbf{G}_{2u} \end{aligned}$$

The terms with \mathbf{A}_1 and \mathbf{B}_2 also cancel giving

$$\begin{aligned} (\mathbf{A}\mathbf{G}_{1v} + \mathbf{B}\mathbf{G}_{2v})_u - (\mathbf{A}\mathbf{G}_{1u} + \mathbf{B}\mathbf{G}_{2u})_v &= (\mathbf{B}_1 - \mathbf{A}_2)(\mathbf{G}_{2u}\mathbf{G}_{1v} - \mathbf{G}_{2v}\mathbf{G}_{1u}) \\ &= (\mathbf{B}_1 - \mathbf{A}_2) \det \mathbf{J}_{\mathbf{G}} \end{aligned}$$

What have we shown? We have this chain of results:

$$\begin{aligned} \iint_{\mathbf{G}(\mathbb{I}^2)} (\mathbf{B}_1 - \mathbf{A}_2) dA_{xy} &= \iint_{\mathbb{I}^2} (\mathbf{B}_1 - \mathbf{A}_2) \det \mathbf{J}_{\mathbf{G}} dA_{uv} \\ &= \iint_{\mathbb{I}^2} \left((\mathbf{A}\mathbf{G}_{1v} + \mathbf{B}\mathbf{G}_{2v})_u - (\mathbf{A}\mathbf{G}_{1u} + \mathbf{B}\mathbf{G}_{2u})_v \right) dA_{uv} \end{aligned}$$

$$\begin{aligned}
 &= \int_{\partial\mathbb{I}^2} \langle \mathbf{F}^*, \gamma' \rangle dt \\
 &= \int_{G(\partial\mathbb{I}^2)} \left\langle \begin{bmatrix} \mathbf{A}(x, y) \\ \mathbf{B}(x, y) \end{bmatrix}, (\mathbf{G}(\gamma))' \right\rangle dt
 \end{aligned}$$

This is the proof for Green's Theorem for Oriented 2 - Cells. Let's state it formally.

Theorem 17.6.1 Green's Theorem for Oriented 2 - Cells

Assume \mathbf{G} is a 1-1 map from the oriented 2 cell \mathbb{I}^2 to $\mathbf{G}(\mathbb{I}^2) = D$ in \mathbb{R}^2 . Assume \mathbf{G} has continuous second order partials on an open set U which contains $[0, 1] \times [0, 1]$. Let \mathcal{C} be a path around $\partial\mathbb{I}^2$ with associated vector function γ which traverses the boundary ccw. Then,

$$\int_{G(\partial\mathbb{I}^2)} \left\langle \begin{bmatrix} \mathbf{A}(x, y) \\ \mathbf{B}(x, y) \end{bmatrix}, (\mathbf{G}(\gamma))' \right\rangle dt = \iint_{G(\mathbb{I}^2)} (\mathbf{B}_1 - \mathbf{A}_2) dA_{xy}$$

Proof 17.6.1

We have just gone through this argument. ■

Comment 17.6.1 You probably noticed how awkward these arguments sometimes were and there was a considerable amount of manipulation involved. Everything is considerably simplified when we begin using the language of differential forms which is discussed in Chapter 18. If ω is a 1 - form defined on $\mathbf{G}(\mathbb{I}^2)$ and $d\omega$ is the 2 - form obtained from ω , the above theorem looks like this:

$$\int_{G(\partial\mathbb{I}^2)} \omega = \iint_{G(\mathbb{I}^2)} d\omega$$

Further, the integral symbols are simplified a bit: it is understood \int over a two dimensional subset can be handled by $\int \int$. So context tells what to do and we just say

$$\int_{G(\partial\mathbb{I}^2)} \omega = \int_{G(\mathbb{I}^2)} d\omega$$

Comment 17.6.2 When you study one dimensional things, the work is much simpler. You can see how much more complicated our arguments become when we pass to differentiation and integration over portions of \mathbb{R}^2 . Make sure you think about this as it helps you begin to see how to free yourself from one dimensional bias and prepares you for higher dimensions than two. Some of the things we have begun to wrestle with are

- What exactly is a curve in \mathbb{R}^2 ? What do we mean by the interior and exterior of a closed curve?
- What do we mean by the smoothness of functions on subsets of \mathbb{R}^2 ?
- Clearly, some of what we do with line integrals is a way to generalize the Fundamental Theorem of Calculus. You should see this is both an interesting problem and a difficult one as well.

17.6.1 Homework

Exercise 17.6.1

Exercise 17.6.2

Exercise 17.6.3

Exercise 17.6.4

Exercise 17.6.5

17.7 Motivational Notation

Recall, we often use the notation $dx = x'(t)dt$ as a convenient way to handle substitution in an integration. We can use this idea to repackage the idea of a line integral of a force field along a curve. We know

$$\begin{aligned}\int_{\mathcal{C}} \left\langle \mathbf{F}, \begin{bmatrix} x' \\ y' \end{bmatrix} \right\rangle dt &= \int_{\mathcal{C}} \left(\left\langle \begin{bmatrix} A(x, y) \\ B(x, y) \end{bmatrix}, \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} \right\rangle \right) dt \\ &= \int_a^b (A(x(t), y(t))x'(t) + B(x(t), y(t))y'(t)) dt\end{aligned}$$

Let's rewrite using differential notation. We get

$$\int_{\mathcal{C}} \left\langle \mathbf{F}, \begin{bmatrix} x' \\ y' \end{bmatrix} \right\rangle dt = \int_a^b (A(x(t), y(t))dx + B(x(t), y(t))dy) dt$$

This suggests we combine the vector field and differential notation and define something we will call a 1 - form which has the look

$$\omega = Adx + Bdy$$

and define

$$\int_{\mathcal{C}} \omega = \int_a^b (A(x(t), y(t))dx + B(x(t), y(t))dy) dt$$

We define the **exterior derivative** (more on this later) of ω to be

$$d\omega = (B_1 - A_2) dx dy = (B_1 - A_2) dA_{xy}$$

Now go back to the oriented 2 - cell problem. Then,

$$\begin{aligned}\omega &= Adu + Bdv \\ \int_{G(\mathbb{I}^2)} \omega &= \int_{G(\gamma)} (A(u(t), v(t))du + B(u(t), v(t))dv) dt \\ d\omega &= (B_1(u, v) - A_2(u, v))dudv = (B_1 - A_2) dA_{uv}\end{aligned}$$

The **pullback** of G is then

$$\begin{aligned}G^*(\omega) &= A(G_1, G_2)(G_{1u}du + G_{1v}dv) + B(G_1, G_2)(G_{2u}du + G_{2v}dv) \\ &= (A(G_1, G_2)G_{1u} + B(G_1, G_2)G_{2u})du + (A(G_1, G_2)G_{1v} + B(G_1, G_2)G_{2v})dv\end{aligned}$$

and so the exterior derivative gives

$$\begin{aligned} d(G^*(\omega)) &= \left(-(A(G_1, G_2)G_{1u} + B(G_1, G_2)G_{2u})_2 + \right. \\ &\quad \left. (A(G_1, G_2)G_{1v} + B(G_1, G_2)G_{2v})_1 \right) dudv \end{aligned}$$

We also define the **pullback** of the exterior derivative

$$G^*(d\omega) = (B_1 - A_2) \det J_G$$

Using the same calculations as before, we then find

$$d(G^*(\omega)) = (B_1 - A_2) \det J_G dudv = G^*(d\omega)$$

Now go back to our arguments for Green’s Theorem for the oriented 2 - cells. Using this new notation, our proof now goes like this:

$$\int \int_{G(\mathbb{I}^2)} d\omega = \int \int_{\mathbb{I}^2} G^*(d\omega) = \int \int_{\mathbb{I}^2} d(G^*(\omega)) = \int_{\partial \mathbb{I}^2} G^*(\omega) = \int_{G(\partial \mathbb{I}^2)} \omega$$

We hide the details of the differential area $dxdy$ and so forth. We can also rewrite using context for whether it is a single or double integral and just write

$$\int_{G(\mathbb{I}^2)} d\omega = \int_{\mathbb{I}^2} G^*(d\omega) = \int_{\mathbb{I}^2} d(G^*(\omega)) = \int_{\partial \mathbb{I}^2} G^*(\omega) = \int_{G(\partial \mathbb{I}^2)} \omega$$

In the next chapter we will explore differential forms more carefully.

17.7.1 Homework

Exercise 17.7.1

Exercise 17.7.2

Exercise 17.7.3

Exercise 17.7.4

Exercise 17.7.5

Part V

Differential Forms

Chapter 18

Differential Forms

The idea of a **differential form** is very powerful and one that you are not exposed to in the first courses in analysis. So it is time to rectify that and introduce you to a new and useful way of looking at functions of many variables.

Definition 18.0.1 Tangent Vector Fields

A tangent vector field \mathbf{V} on \mathbb{R}^n is a function \mathbf{V} which assigns to each point $\mathbf{P} \in \mathbb{R}^n$ a vector \mathbf{v}_P . In addition, we let $\mathbb{R}_{\mathbf{P}}^n = T_{\mathbf{P}}(\mathbb{R}^n)$ denote the set of all vectors \mathbf{w} in \mathbb{R}^n of the form $\mathbf{w} = \mathbf{u} + \mathbf{P}$ where \mathbf{u} in \mathbb{R}^n is arbitrary.

Comment 18.0.1 The set $\mathbb{R}_{\mathbf{P}}^n = T_{\mathbf{P}}(\mathbb{R}^n)$ is clearly an n dimensional vector space over \mathbb{R} which consists of all the translations of vectors in \mathbb{R}^n by a fixed vector \mathbf{P} . We also call this the vector space rooted at \mathbf{P} .

We also want to look at **smooth** functions on an open set U in \mathbb{R}^n .

Definition 18.0.2 Smooth Functions on \mathbb{R}^n

A smooth function or C^∞ function on the open set U in \mathbb{R}^n is a mapping $f : U \rightarrow \mathbb{R}$ whose partial derivatives of all orders exist and are continuous.

18.1 One Forms

We define a differential 1 - form on U or simply a 1 - from on U as follows:

Definition 18.1.1 One Forms on an open set U

A differential 1-form or just 1-form ω on U consists of a pair of smooth functions \mathbf{P} and \mathbf{Q} on U . The one form ω is therefore associated with a vector field

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix}$$

which means $\mathbf{H} : U \rightarrow \mathbb{R}^2$ with smooth component functions. We also write this as

$$\mathbf{H} = \mathbf{A}\mathbf{i} + \mathbf{B}\mathbf{j}$$

where \mathbf{i} and \mathbf{j} are the standard orthonormal basis for \mathbb{R}^2 . We define the action of ω on a vector \mathbf{V} rooted at the point \mathbf{P} with components V_1 and V_2 and $P - 1$ and P_2 , respectively to be

$$\omega(\mathbf{V}) = \mathbf{A}(P_1, P_2)V_1 + \mathbf{B}(P_1, P_2)V_2$$

and it is easy to see ω is linear.

To understand this better, we need to look at the **dual vector space** to a vector space.

Definition 18.1.2 Algebraic Dual of a Vector Space

Let \mathcal{V} be a vector space over \mathbb{R} . The algebraic dual to \mathcal{V} is denoted by \mathcal{V}^* and is defined by

$$\mathcal{V}^* = \{f : \mathcal{V} \rightarrow \mathbb{R} \mid f \text{ is linear}\}$$

Such a function f is called a linear functional on \mathcal{V} .

Comment 18.1.1 If the vector space \mathcal{V} was also an inner product space and hence had an induced norm or if \mathcal{V} had a norm without an inner product, then we could ask if a linear functional was continuous. Let $\|\cdot\|$ be this norm. Recall f is continuous at $\mathbf{p} \in \mathcal{V}$ if for any $\epsilon > 0$, there is a ball $B_\delta(\mathbf{p}) = \{\mathbf{x} \in \mathcal{V} : \|\mathbf{x} - \mathbf{p}\| < \delta\}$ in $(\mathcal{V} \text{ so } |f(\mathbf{x}) - f(\mathbf{p})| < \epsilon)$. We can not prove this here but we will show you where the proof breaks down. We discuss this properly in (Peterson (9) 2019). Let \mathbf{E} be a basis for \mathcal{V} with norm $\|\cdot\|$. and we can assume each \mathbf{E}_j has norm one. If f is a linear functional on \mathcal{V} , let $\xi_j = f(\mathbf{E}_j)$. Then for any $\mathbf{x} \in \mathcal{V}$, there is a unique representation with respect to this basis

$$\mathbf{x} = c_1\mathbf{E}_1 + \dots + c_n\mathbf{E}_n$$

and so by the linearity of f

$$f(\mathbf{x}) = x_1f(\mathbf{E}_1) + \dots + x_n(\mathbf{E}_n) = \sum_{i=1}^n x_i\xi_i$$

Thus

$$|f(\mathbf{x}) - f(\mathbf{p})| \leq \sum_{i=1}^n |x_i - p_i| |\xi_i| \leq f_E \sum_{i=1}^n |x_i - p_i|$$

where $f_E = \max_{1 \leq i \leq n} |\xi_i|$ and p_i are the components of \mathbf{p} with respect to the basis \mathbf{E} . So given an $\epsilon > 0$, we can make $|f(\mathbf{x}) - f(\mathbf{p})| < \epsilon$ if we choose $\sum_{i=1}^n |x_i - p_i| < \epsilon/f_E$. However, this does not

help as

$$\|\mathbf{x} - \mathbf{p}\| = \left\| \sum_{i=1}^n (c_i - p_i) \mathbf{E}_i \right\| \leq \sum_{i=1}^n |c_i - p_i| \|\mathbf{E}_i\| = \sum_{i=1}^n |c_i - p_i|$$

But choosing $\|\mathbf{x} - \mathbf{p}\| < \delta = \epsilon/f_E$, does not guarantee that $\sum_{i=1}^n |c_i - p_i| < \epsilon/f_E$ which we need. So we are missing a way to handle the $\|\mathbf{x} - \mathbf{p}\|$ correctly here. To fill this in, we need more details about how normed linear spaces work which we go over in (Peterson (9) 2019). The set of continuous linear functionals on a vector space \mathcal{V} is denoted by \mathcal{V}' and it turns out $\mathcal{V}' = \mathcal{V}^*$ when \mathcal{V} is finite dimensional normed linear space.

The vector space $\mathbf{X} = (\mathbb{R}^2, \{\mathbf{i}, \mathbf{j}\})$ is traditionally thought of as column vectors and the dual space to $(\mathbb{R}^2, \{\mathbf{i}, \mathbf{j}\})$ is then \mathbf{X}^* the set of linear functionals on \mathbf{X} which is the same as \mathbf{X}' the set of continuous linear functionals on \mathbf{X} . The dual to $(\mathbb{R}^2, \{\mathbf{i}, \mathbf{j}\})$ is interpreted as the set of row vectors and a basis dual to any basis $\mathbf{E} = \{\mathbf{E}_1, \mathbf{E}_2\}$ is $\mathbf{F} = \{\mathbf{F}_1, \mathbf{F}_2\}$ where

$$\begin{aligned} \mathbf{F}_1(\mathbf{E}_1) &= 1, & \mathbf{F}_1(\mathbf{E}_2) &= 0 \\ \mathbf{F}_2(\mathbf{E}_1) &= 0, & \mathbf{F}_2(\mathbf{E}_2) &= 1 \end{aligned}$$

Hence,

$$\begin{aligned} \mathbf{F}_1 \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) &= V_1 \\ \mathbf{F}_2 \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) &= V_2 \end{aligned}$$

Hence, we can identify \mathbf{F}_1 and \mathbf{F}_2 as follows

$$\begin{aligned} \mathbf{F}_1 &= [1 \ 0] \implies \mathbf{F}_1 \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) = [1 \ 0] \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = V_1 \\ \mathbf{F}_2 &= [0 \ 1] \implies \mathbf{F}_2 \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) = [0 \ 1] \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = V_2 \end{aligned}$$

The usual notation for the dual basis to $\{\mathbf{i}, \mathbf{j}\}$ is not \mathbf{F}_1 and \mathbf{F}_2 . Instead, it is $d\mathbf{x} = \mathbf{F}_1$ and $d\mathbf{y} = \mathbf{F}_2$. Hence, we define the action of a 1-form ω on a vector \mathbf{V} rooted at \mathbf{P} as follows:

$$\begin{aligned} \omega \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) &= ([\mathbf{F}_1 \ \mathbf{F}_2] \mathbf{H}) \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) = \left([\mathbf{F}_1 \ \mathbf{F}_2] \begin{bmatrix} A \\ B \end{bmatrix} \right) \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) \\ &= (\mathbf{A}\mathbf{F}_1 + \mathbf{B}\mathbf{F}_2) \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) = \mathbf{A}(P_1, P_2)\mathbf{F}_1 \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) + \mathbf{B}(P_1, P_2)\mathbf{F}_2 \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) \\ &= \mathbf{A}(P_1, P_2)V_1 + \mathbf{B}(P_1, P_2)V_2 \end{aligned}$$

This is quite a mouthful. It is easier to follow if we use the $d\mathbf{x}$ and $d\mathbf{y}$ notation as defined

$$\begin{aligned} \omega \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) &= (\mathbf{A}d\mathbf{x} + \mathbf{B}d\mathbf{y}) \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) = \mathbf{A}(P_1, P_2)d\mathbf{x} \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) + \mathbf{B}(P_1, P_2)d\mathbf{y} \left(\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) \\ &= \mathbf{A}(P_1, P_2)V_1 + \mathbf{B}(P_1, P_2)V_2 \end{aligned}$$

To be even more succinct, for a vector \mathbf{V} with the usual components V_1 and V_2 rooted at \mathbf{P} :

$$\begin{aligned} \omega(\mathbf{V}) &= (\mathbf{A}d\mathbf{x} + \mathbf{B}d\mathbf{y})(\mathbf{V}) = \mathbf{A}(P_1, P_2)d\mathbf{x}(\mathbf{V}) + \mathbf{B}(P_1, P_2)d\mathbf{y}(\mathbf{V}) \\ &= \mathbf{A}(P_1, P_2)V_1 + \mathbf{B}(P_1, P_2)V_2 \end{aligned}$$

We can interpret these results a bit differently by thinking of the dual space to \mathfrak{R}_P^2 where we now think of this space as the space of all tangent vectors attached to a point P where P has coordinates (x_0, y_0) . Consider a curve \mathcal{C} in \mathbb{R}^2 given by the smooth curve $f(x, y) = c$ for some constant c . Then, the tangent plane at (x_0, y_0) is given by

$$z = f_x^0(x - x_0) + f_y^0(y - y_0)$$

and this plane has the normal vector $\mathbf{N} = f_x^0\mathbf{i} \times f_y^0\mathbf{j}$. Each curve f determines a plane of this type. The simplest plane is the one determined by $f(x, y) = x + y = c$. The tangent plane is then $z = \mathbf{1}(x - x_0) + \mathbf{1}(y - y_0)$. Any point in \mathfrak{R}_P^2 can thus be interpreted as the vector $\mathbf{Q} = f_x^0\mathbf{i} + f_y^0\mathbf{j}$ for some smooth function $f(x, y) = c$ where \mathbf{i} and \mathbf{j} are rooted at P . Another way to look at it is the \mathbf{i} is a unit vector in the direction of x rooted at x_0 and \mathbf{j} is a unit vector in the direction of y rooted at y_0 . A linear functional ϕ acting on \mathfrak{R}_P^2 thus gives

$$\phi(c_1(x - x_0)\mathbf{i} + c_2(y - y_0)\mathbf{j}) = c_1\phi((x - x_0)\mathbf{i}) + c_2\phi((y - y_0)\mathbf{j})$$

Now the simplest function, $x + y = c$ gives the tangent plane $z = \mathbf{1}(x - x_0) + \mathbf{1}(y - y_0)$ and we have

$$\phi(\mathbf{1}(x - x_0)\mathbf{i} + \mathbf{1}(y - y_0)\mathbf{j}) = \mathbf{1}\phi((x - x_0)\mathbf{i}) + \mathbf{1}\phi((y - y_0)\mathbf{j})$$

This suggests we rethink the dual basis. We define

$$\begin{aligned} \mathbf{F}_1\left(\begin{bmatrix} x - x_0 \\ 0 \end{bmatrix}\right) &= \frac{\partial}{\partial x}(x - x_0) = \frac{d}{dx}(x - x_0) = \mathbf{1} \\ \mathbf{F}_2\left(\begin{bmatrix} 0 \\ (y - y_0) \end{bmatrix}\right) &= \frac{\partial}{\partial y}(y - y_0) = \frac{d}{dy}(y - y_0) = \mathbf{1} \end{aligned}$$

With this reinterpretation, we have

$$\phi(c_1(x - x_0)\mathbf{i} + c_2(y - y_0)\mathbf{j}) = c_1\frac{\partial}{\partial x} + c_2\frac{\partial}{\partial y}$$

and so

$$\phi(f_x^0(x - x_0)\mathbf{i} + f_y^0(y - y_0)\mathbf{j}) = f_x^0\frac{\partial}{\partial x} + f_y^0\frac{\partial}{\partial y}$$

Thus, in terms of the dual basis, we would write $\phi = c_1\frac{\partial}{\partial x} + c_2\frac{\partial}{\partial y}$ and

$$\begin{aligned} \phi(f_x^0(x - x_0)\mathbf{i} + f_y^0(y - y_0)\mathbf{j}) &= \left(c_1\frac{\partial}{\partial x} + c_2\frac{\partial}{\partial y}\right) \left(f_x^0(x - x_0)\mathbf{i} + f_y^0(y - y_0)\mathbf{j}\right) \\ &= c_1f_x^0 + c_2f_y^0 \end{aligned}$$

The common symbols for this dual basis are $d\mathbf{x} = \frac{\partial}{\partial x}$ and $d\mathbf{y} = \frac{\partial}{\partial y}$. Thus, a linear functional is written $\phi = c_1d\mathbf{x} + c_2d\mathbf{y}$.

Next, we need to define a smooth path in U .

Definition 18.1.3 Smooth Paths

Let U be an open set in \mathbb{R}^2 . A smooth path in U is a mapping $\gamma : [a, b] \rightarrow U$ for some finite interval $[a, b]$ in \mathbb{R} which is continuous on $[a, b]$ and differentiable on (a, b) . We assume γ can be extended to a smooth C^∞ map in a neighborhood of $[a, b]$. Hence

$$\gamma(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}, \quad a \leq t \leq b$$

where x and y are C^∞ on $(a - \epsilon, b + \epsilon)$ for some positive ϵ .

Comment 18.1.2 The initial point of the path is $\gamma(a)$ and the final point is $\gamma(b)$. We say γ is a path from $\gamma(a)$ to $\gamma(b)$.

If $\gamma(t_0) \neq \mathbf{0}$, then at the point $\gamma(t_0)$, there is a two dimensional vector space rooted at $\gamma(t_0)$ we can call $T(t_0)$ defined by the span of the the vectors

$$\mathbf{E}_1(t_0) = \mathbf{i}, \quad \mathbf{E}_2(t_0) = \mathbf{j}$$

Note $T(t_0)$ is determined by the tangent vector to the curve $\gamma(t)$ at the point t_0 . Any vector in this vector space has the representation rooted at $\gamma(t_0)$ given by

$$\begin{bmatrix} p \\ q \end{bmatrix} = (t - t_0) \begin{bmatrix} \alpha x'(t_0) \\ \beta y'(t_0) \end{bmatrix} = \alpha x'(t_0) (t - t_0) \mathbf{E}_1(t_0) + \beta y'(t_0) (t - t_0) \mathbf{E}_2(t_0)$$

This shows $\{\mathbf{E}_1(t_0), \mathbf{E}_2(t_0)\}$ is a basis for $T(t_0)$. We define a dual basis $\{dx(t_0), dy(t_0)\}$ by

$$dx(t_0) = \mathbf{E}_1^T(t_0), \quad dy(t_0) = \mathbf{E}_2^T(t_0)$$

Then at each t , we can represent $\gamma(t)$ in $T(t_0)$ by

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} \alpha(t) x'(t_0) \\ \beta(t) y'(t_0) \end{bmatrix} = \gamma(t_0) + \alpha(t) x'(t_0) \mathbf{E}_1(t_0) + \beta(t) y'(t_0) \mathbf{E}_2(t_0)$$

Now if f is a linear functional on $T(t_0)$, then

$$\begin{aligned} f\left(\alpha(t) x'(t_0) \mathbf{E}_1(t_0) + \beta(t) y'(t_0) \mathbf{E}_2(t_0)\right) \\ = \alpha(t) x'(t_0) f(\mathbf{E}_1(t_0)) + \beta(t) y'(t_0) f(\mathbf{E}_2(t_0)) \\ = \alpha(t) x'(t_0) c_f + \beta(t) y'(t_0) d_f \end{aligned}$$

as the action of f is completely determined by how it acts on the basis for $T(t_0)$. Now consider

$$\begin{aligned} (c_1 dx + c_2 dy) \left(\gamma(t_0) + \alpha(t) x'(t_0) \mathbf{E}_1(t_0) + \beta(t) y'(t_0) \mathbf{E}_2(t_0) \right) \\ = c_1 \alpha(t) x'(t_0) + c_2 \beta(t) y'(t_0) \end{aligned}$$

For equality, we need

$$\begin{aligned} \alpha(t) x'(t_0) c_f &= c_1 \alpha(t) x'(t_0) \implies c_1 = c_f \\ c_2 \beta(t) y'(t_0) &= \beta(t) y'(t_0) d_f \implies c_2 = d_f \end{aligned}$$

We conclude the span of $\{dx, dy\}$ is the dual of $T(t_0)$. Thus, the action of dx and dy on the path γ should be defined to be

$$dx(\gamma(t)) = x'(t), \quad dy(\gamma(t)) = y'(t)$$

Note this is same as what we had before except now we have the chain rule involved. We have a linear functional

$$\phi = c_1 dx + c_2 dy = c_1 \frac{\partial}{\partial x} \frac{d}{dt} + c_2 \frac{\partial}{\partial y} \frac{d}{dt}$$

With all this done, we have an understanding of how to interpret the action of the 1 - form ω on the path γ . At each t , we define

$$\begin{aligned} \omega(\gamma) &= P(x(t_0), y(t_0)) dx(t_0)(\gamma(t) + Q(x(t_0), y(t_0)) dy(t_0)(\gamma(t)) \\ &= P(x(t_0), y(t_0)) x'(t_0) + Q(x(t_0), y(t_0)) y'(t_0) \end{aligned}$$

This is quite a notational morass, so we usually say this sloppier and let context be our guide. The action of the 1 - form ω on the smooth path γ is

$$\begin{aligned} \omega(\gamma(t)) &= P(\gamma(t)) dx(\gamma(t)) + Q(\gamma(t)) dy(\gamma(t)) \\ &= P(x(t), y(t)) x'(t) + Q(x(t), y(t)) y'(t) = \begin{bmatrix} P \\ Q \end{bmatrix} (\gamma(t)) \gamma'(t) \end{aligned}$$

We can then define the integral $\int_{\gamma} \omega$.

Definition 18.1.4 Integration of a 1 -form over a smooth path

Let γ be a smooth path in the open set U in \mathbb{R}^2 and let $\omega = P dx + Q dy$ be a 1 - form. Then

$$\int_{\gamma} \omega = \int_a^b \left(P(x(t), y(t)) x'(t) + Q(x(t), y(t)) y'(t) \right) dt$$

Note this is a well - defined Riemann Integrable as the integrand is continuous.

Recall, in Chapter 17, these types of integrals are called **Line Integrals**. Hence, what we went through above can be interpreted from another point of view. Suppose each point on the smooth path $\gamma(t)$ represents the charge of the path at the point t . Then how do we compute the total charge of the path? First, note the function

$$U(t) = P(x(t), y(t)) x'(t) + Q(x(t), y(t)) y'(t)$$

is a continuous function on $[a, b]$ and hence, it is Riemann Integrable. Hence, the value of this integral is given by

$$\lim_{n \rightarrow \infty} S(U, \pi_n, \sigma_n)$$

for any sequence of partitions $\{\pi_n\}$ whose norms go to zero and σ_n is any evaluation set in π_n . Then, for a given partition, the mass in the portion of the path between partition point t_j and t_{j+1} can be approximated by

$$P(x(t_j), y(t_j))(x(t_{j+1}) - x(t_j)) + Q(x(t_j), y(t_j))(y(t_{j+1}) - y(t_j)) =$$

$$\left(\mathbf{P}(x(t_j), y(t_j)) \frac{x(t_{j+1} - x(t_j)}{t_{j+1} - t_j} + \mathbf{Q}(x(t_j), y(t_j)) \frac{y(t_{j+1} - y(t_j)}{t_{j+1} - t_j} \right) (t_{j+1} - t_j)$$

Now apply the Mean Value Theorem to see

$$\begin{aligned} & \left(\mathbf{P}(x(t_j), y(t_j)) \frac{x(t_{j+1} - x(t_j)}{t_{j+1} - t_j} + \mathbf{Q}(x(t_j), y(t_j)) \frac{y(t_{j+1} - y(t_j)}{t_{j+1} - t_j} \right) (t_{j+1} - t_j) = \\ & \left(\mathbf{P}(x(t_j), y(t_j)) x'(s_j) + \mathbf{Q}(x(t_j), y(t_j)) y'(s_j) \right) (t_{j+1} - t_j) \end{aligned}$$

The approximation to the mass on the portion $[t_j, t_{j+1}]$ can just as well be done setting $t_j = s_j$ in the \mathbf{P} and \mathbf{Q} terms. Thus, the approximation can be

$$U(s_j) (t_{j+1} - t_j) = \left(\mathbf{P}(x(s_j), y(s_j)) x'(s_j) + \mathbf{Q}(x(s_j), y(s_j)) y'(s_j) \right) (t_{j+1} - t_j)$$

The Riemann Approximation Theorem then tells us

$$\lim_{n \rightarrow \infty} \sum_{\pi_n} U(s_j) \Delta t_j = \int_a^b U(t) dt$$

which is the same result as before. We see our interpretation of the integration of the 1 - form ω over the smooth path γ is the same as what we call the line integral of the vector field $\mathbf{P}i + \mathbf{Q}j$ on the path given by the smooth curve γ .

18.1.1 A Simple Example

Consider a simple circle in the plane parameterized by $x(t) = \cos(t)$, $y(t) = \sin(t)$ for $0 \leq t \leq 2\pi$. Hence, the smooth path here is

$$\gamma(t) = \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix}$$

for $0 \leq t \leq 2\pi$ where we start at the point $(1, 0)$ at $t = 0$ and end there as well at $t = 2\pi$. This is a circle in the plane of radius 1 centered at the origin. From the center of the circle, if you draw a vector to a point outside the circle, the variable t represents the angle from the positive x axis measured counterclockwise (ccw) to this line. You can imagine that the picture we draw would be very similar even if we chose an anchor point different from the center. We would still have a well defined angle t . If we choose an arbitrary starting angle t_0 , the angle we measure as we move ccw around the circle would be start at t_0 and would end right back at the start point with the new angle $t_0 + 2\pi$. Hence, there is no way around the fact that as we move ccw around the circle, the angle we measure has a discontinuity. To allow for more generality later, let this angle be denoted by $\theta(t)$ which in our simple example is just $\theta(t) = t$. Since $\tan(\theta(t)) = y(t)/x(t)$, we have the general equation for the rate of change of the angle:

$$\theta'(t) = \frac{-y(t)x'(t) + x(t)y'(t)}{x^2(t) + y^2(t)}$$

Of course, here this reduces to $\theta'(t) = 1$ as we really just have $\theta(t) = t$ which has a very simple derivative! Note we can use this to define a 1 - form ω by

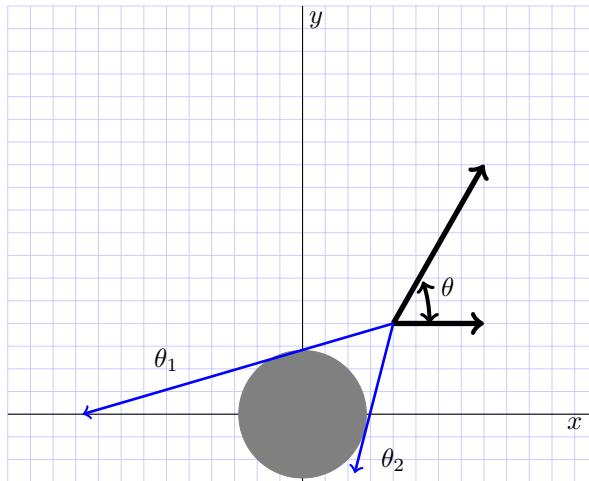
$$\omega = \frac{-y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy$$

Then, consider this scaled integral of ω on the path γ :

$$\begin{aligned}\frac{1}{2\pi} \int_{\gamma} \omega &= \frac{1}{2\pi} \int_0^{2\pi} \frac{-y(t)x'(t) + x(t)y'(t)}{x^2(t) + y^2(t)} dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} 1 dt = 1.\end{aligned}$$

On the other hand, suppose we put the anchor point of our angle measurement system outside the circle. So imagine the reference line for the angle $\theta(t)$ to be moved from a vector starting at the origin of the circle to a point outside the circle to a new vector starting outside the circle. For convenience, assume this new vector is in quadrant one. At the point this vector is rooted, it determines a local coordinate system for the plane whose positive x axis direction is given by the bottom of the usual reference triangle we draw at this point. In Figure 18.1, we show a typical setup. Note the angles measured start at θ_1 , increase to θ_2 and then decrease back to θ_1 . Then, since the angles are now measured from the base point (x_0, y_0) outside the circle, we set up the line integral a bit different. We find the change in angle is now zero and there is no discontinuous jump in the angle.

$$\begin{aligned}\frac{1}{2\pi} \int_{\theta_1}^{\theta_2} \frac{-(y(t) - y_0)x'(t) + (x(t) - x_0)y'(t)}{(x(t) - x_0)^2 + (y(t) - y_0)^2} dt + \\ \frac{1}{2\pi} \int_{\theta_2}^{\theta_1} \frac{-(y(t) - y_0)x'(t) + (x(t) - x_0)y'(t)}{(x(t) - x_0)^2 + (y(t) - y_0)^2} dt = 0.\end{aligned}$$



The angle reference system is outside the circle. The angle θ_1 is the first angle needed for the circle and the angle θ_2 is the last.

Figure 18.1: Angle Measurements From Outside the Circle

In general, if we try to measure the angle using a point inside the circle, we always get +1 for this integral and if we try to measure the angle from a reference point outside the circle we get 0. This integer we are calculating is called the *winding number* associated with the particular curve given by our circle and we will return to it later in more detail. Now, if we had a parameterized curve in the plane given by the pair of functions $(x(t), y(t))$, the line integrals we have just used still are a good way to define the change in angle as we move around the curve. We will define the *winding number* of the smooth path γ relative to a point P to mean we are measuring the change in angle as we move around the curve γ from its start point to its finish point where we assume the start and finish point are the same. Such smooth paths are called **closed**. So if functions that determine γ are defined on

$[a, b]$, a closed curve means $(x(a), y(a)) = (x(b), y(b))$. Hence, closed curves are nice analogues of the simplest possible closed curve - the circle at the origin. The winding number, $W(\gamma, P)$, is defined to be

$$W(\gamma, P) = \frac{1}{2\pi} \int_{\gamma} \omega$$

where we use as 1 - form

$$\omega = \frac{-(y - y_0)}{(x - x_0)^2 + (y - y_0)^2} dx + \frac{(x - x_0)}{(x - x_0)^2 + (y - y_0)^2} dy$$

where the coordinates of P are (x_0, y_0) .

Now let's back up a bit. In Figure 18.1, we found the change in angle as we moved around the circle was zero giving us a winding number of zero because we had placed the angle measurement systems reference outside the circle. Another way of saying this is that $W(\gamma, P) = 0$ for γ being the circle centered at the origin when P is outside the circle. A little thought tells you that this should be true for more arbitrary paths γ although, of course, there is lots of detail we are leaving out as well as subtleties. Now look at Figure 18.1 again and think of the circle and its interior as a collection of points we can not probe with our paths γ . We can form closed loops around any point P in the plane in general and we can *deform* these closed paths into new closed paths that let P pop out of the inside of γ as long as we are free to alter the path. Altering the path means we find new parameterizations $(x(t), y(t))$ which pass through new points in the plane. We are free to do this kind of alteration as freely as we want as long as we don't pick a P inside the gray circle. Think about it. If we had a closed path γ that enclosed a point P inside the gray circle, then we can not deform this γ into a new path that excludes P from its interior because we are not allowed to let the points $(x(t), y(t))$ on the new path enter the gray circle's interior. So we can probe for the existence of this collection of point which have been excluded by looking at the winding numbers of paths γ . Paths γ whose winding numbers relative to a point P inside the gray circle will all have nonzero winding numbers. Another way of saying this is that if P is **not** inside the gray circle, there are paths γ with P **not** in their interior and so their winding number is zero. But for P inside the gray circle, we can not find any closed paths γ whose interior excludes P and hence their winding numbers are always not zero (in our simple thought experiment, they are $+1$). Also, please note we keep talking about the **interior** of a smooth path γ as if it is always easy to figure out what that is. In general, it is not! So we still have much to discuss. The first books you can read in this area is the book on *Algebraic Topology* (Fulton (4) 1995) which requires you to be focused and ready to learn new ways to think. But these are sophisticated ideas and you should not be disheartened by their difficulty!

18.2 Exact and Closed 1 - Forms

Let U be an open set in \mathbb{R}^2 . We have have already defined what a smooth function on U is, but now we need to know more about the topology of the open set U itself. We need the idea of a connected set. There are several ways to do this. We will start with the simplest.

Definition 18.2.1 Connected Sets

An open set U in \mathbb{R}^n is connected if it is not possible to find two disjoint open sets A and B so that $U = A \cup B$. If U is not connected, the sets A and B are called the components of U .

An important fact about the relationship between connected open sets and smooth functions is this:

Theorem 18.2.1 All Smooth Functions with zero derivatives are constant if and only if the domain is connected

Let $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be a smooth function on the open set U . Then

$$\text{all smooth } f \ni \nabla f = \mathbf{0} \text{ are constant} \iff U \text{ is connected}$$

Proof 18.2.1

\implies :

If $\nabla f = \mathbf{0}$ in U , then it is easy to see the second order partials of f are identically 0. Given x_0 in U , it is an interior points and so there is a neighborhood of it, $B_r(x_0)$ in U . Hence, for any $x \in B_r(x_0)$, we know

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2} [x - x_0]^T \begin{bmatrix} f_{11}(c) & \dots & f_{1n}(c) \\ \vdots & \ddots & \vdots \\ f_{n1}(c) & \dots & f_{nn}(c) \end{bmatrix} [x - x_0]$$

where c is a point on the line between x_0 and x . This point is also in $B_r(x_0)$. However, since all the second order partials are zero, the Hessian here is the zero matrix always and we have $f(x) = f(x_0)$. Since these points were arbitrarily chosen, we see f is constant on $B_r(x_0)$; i.e. we have shown such functions f must be locally constant on U .

For the given x_0 and a given f , let $A = f^{-1}(d)$ where $d = f(x_0)$. If $y \in A$, then because y is an interior point, there is a neighborhood $B_s(y)$ contained in U . Hence, $B_s(y) \cap A$ is an open set with z in A for all z in $B_s(y) \cap A$. This shows y is an interior point of A and so A is open. Now let $B = U \setminus A$. Pick any y in B . Then $f(y) \neq d$ and since f is locally constant at y , we know there is a neighborhood about y with corresponding function values not in A . Hence, they are in B which shows y is an interior point of B . We conclude B is an open set and we have written $U = A \cup B$ with A and B disjoint and open. This shows U is not connected. This is a contradiction and so we must conclude B is empty and $U = f^{-1}(d)$. Thus any such f must be constant on all of U .

\iff :

If U were not connected, we can see using the argument in the first part that f is a constant on each component of U . This would mean f might not be constant on U . Thus, U must be connected in this case. \blacksquare

Note we could use any orthonormal basis E here, but we might as well just do the internal mapping from E to $\{i, j\}$ mentally and just think of everything in terms of a starting basis which is the traditional standard one.

Let's look at a few results in this area.

Theorem 18.2.2 A Simple Recapture Theorem in \mathbb{R}^2

Let f be a C^∞ function in the open set U in \mathbb{R}^2 . Let the 1-form $\omega = f_x dx + f_y dy$. Then $\int_\gamma \omega = f(\gamma(b)) - f(\gamma(a))$ where the notation $f(\gamma(t))$ means $f(x(t), y(t))$ where $x(t)$ and $y(t)$ are the component functions of the smooth path γ .

Proof 18.2.2

This is a simple application of the Fundamental Theorem of Calculus.

$$\begin{aligned} \left(f(\gamma(t)) \right)' &= f_x(x(t), y(t)) x'(t) + f_y(x(t), y(t)) y'(t) \\ \implies \int_{\gamma} \omega &= \int_a^b \left(f(\gamma(t)) \right)' dt = f(\gamma(b)) - f(\gamma(a)) \end{aligned}$$

■

Comment 18.2.1 The 1-form $\omega = f_x dx + f_y dy$ occurs very frequently and we use a special notation for it: $df = f_x dx + f_y dy$

An important idea is that of **exactness**.

Definition 18.2.2 Exact 1 - forms

Let f be a C^∞ function in the open set U in \mathbb{R}^2 . with corresponding 1-form $df = f_x dx + f_y dy$. A 1-form ω is a **differential** off if $\omega = df$. A 1-form ω is called **exact** when $\omega = df$ for some smooth f .

Theorem 18.2.3 The angle function is not exact on $\mathbb{R}^2 \setminus \{0,0\}$

For $f(x, y) = \tan^{-1}(y/x)$,

$$df = \frac{-y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy$$

Note f is not a smooth function on $\mathbb{R}^2 \setminus \{0,0\}$ but df is a smooth 1-form on $\mathbb{R}^2 \setminus \{0,0\}$. In fact, there is no smooth function g on $\mathbb{R}^2 \setminus \{0,0\}$ so that $df = dg$.

Proof 18.2.3

By direct calculation, for the smooth path

$$\gamma(t) = \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix}$$

for $0 \leq t \leq 2\pi$, we have $\int_{\gamma} df = 2\pi$ and $\gamma(0) = \gamma(2\pi)$. Note, f_x and f_y fail to exist at some points on the path, so the argument we used in Theorem 18.2.2 does not apply. However, using the argument we used in Theorem 18.2.2 does work for a smooth g ; we have

$$\int_{\gamma} \omega = \int_0^{2\pi} \left(f(\gamma(t)) \right)' dt = f(\gamma(2\pi)) - f(\gamma(0)) = 0$$

Hence, no such g can exist.

■

We can also glue smooth paths together to build **segmented paths**.

Definition 18.2.3 Smooth Segmented Paths

A smooth segmented path γ is a finite sequence of smooth paths $\{\gamma_1, \dots, \gamma_n\}$ where each γ_i is a smooth path and the final point of γ_i is the initial point of γ_{i+1} . We write

$$\gamma = \gamma_1 + \dots + \gamma_n$$

and for any 1-form ω , we define

$$\int_{\gamma} \omega = \int_{\gamma_1} \omega + \dots + \int_{\gamma_n} \omega$$

The first result is to extend Theorem 18.2.2 to segmented paths.

Theorem 18.2.4 The Recapture Theorem for Segmented Paths

Let γ be a smooth segmented path in the open set U in \mathbb{R}^2 . Then if $\omega = df$ in U

$$\int_{\gamma} \omega = f(Q) - f(P)$$

where $P = \gamma_1(a)$ and $Q = \gamma_n(b)$. P is the initial and Q is the final point of the path.

Proof 18.2.4

We can apply Theorem 18.2.2 to each piece of the segmented path.

$$\int_{\gamma} \omega = \sum_{i=1}^n \left(f(\gamma_i(b)) - f(\gamma_i(a)) \right)$$

But $\gamma_i(b) = \gamma_{i+1}(a)$. Hence,

$$\int_{\gamma} \omega = f(\gamma_n(b)) + \sum_{i=1}^{n-1} \left(f(\gamma_{i+1}(a)) - f(\gamma_i(a)) \right)$$

But the summation here is telescoping and gives

$$\int_{\gamma} \omega = f(\gamma_n(b)) - f(\gamma_1(a))$$

which is the result. ■

What about the converse?

Theorem 18.2.5 Equivalent Characterizations of Exactness

Let ω be a 1-form on the open set U in \mathbb{R}^2 . Then the following are equivalent:

1. $\int_{\gamma} \omega = \int_{\delta} \omega$ for any smooth segmented paths γ and δ with the same domain, $[a, b]$, and the same initial and final points.
2. $\int_{\gamma} \omega = 0$ for smooth segmented paths γ that are closed; i.e. their initial and final points are the same.
3. $\omega = df$ for a smooth function f on U ; i.e. ω is exact.

Proof 18.2.5

First, the inverse of the path γ will be denoted by γ^{-1} and is defined by $\gamma^{-1} = \gamma(b + a - t)$.

(2 \implies 1:)

Note $\int_{\gamma^{-1}} \omega = -\int_{\gamma} \omega$. So given paths γ and δ in U with the same initial and final points, let τ be the closed path

$$\tau = \gamma_1 + \dots + \gamma_n - \delta_1 - \dots - \delta_m$$

Then by (2), $\int_{\tau} \omega = 0$ which implies $\int_{\gamma} \omega = \int_{\delta} \omega$.

(1 \implies 2:)

Let γ be a closed path and let δ be a constant path, i.e. it never leaves the initial point. By (1), $\int_{\gamma} \omega = \int_{\delta} \omega = 0$.

(1 \implies 3:)

Let $\omega = U dx + V dy$ for concreteness. Assume U is connected. If U has more than one component, you can do the argument that follows on each component. Pick a fixed point P_0 in U . Let the smooth segmented path γ have initial point P_0 and final point $P = (x, y)$ in U . For the point $P' = (x + h, y)$, define the path σ by $\sigma(t) = (x + t, y)$ for $0 \leq t \leq h$. Then $\gamma + \sigma$ is a segmented path from P_0 to P' . For any Q in U , define f in U by $f(Q) = \int_{\tau} \omega$ for any smooth segmented path τ from P_0 to Q . Since P is in U , P is an interior point and so for sufficiently small h , P' is in U also. Thus $f(P')$ is well-defined and

$$\begin{aligned} \frac{f(x + h, y) - f(x, y)}{h} &= \frac{1}{h} \left(\int_{\gamma + \sigma} \omega - \int_{\gamma} \omega \right) = \frac{1}{h} \int_{\sigma} \omega \\ &= \frac{1}{h} \int_0^h U(x + t, y) dt \end{aligned}$$

Now apply the Mean Value Theorem to $g(h) = \int_0^h U(x + t, y) dt$ whose derivative is $g'(h) = U(x + h, y)$ by the Fundamental Theorem of Calculus. Thus, $g(h) - g(0) = g'(c) h$ for some c between 0 and h . Hence,

$$\frac{1}{h} \int_0^h U(x + t, y) dt = \frac{1}{h} \left(U(x + c, y) h \right) = U(x + c, y)$$

But U is continuous, so as $h \rightarrow 0$,

$$\frac{\partial f}{\partial x}(\mathbf{P}) = \lim_{h \rightarrow 0} \frac{f(x + h, y) - f(x, y)}{h} = \lim_{h \rightarrow 0} U(x + c, y) = U(x, y) = U(\mathbf{P})$$

A similar argument shows $\frac{\partial f}{\partial y}(\mathbf{P}) = V(\mathbf{P})$. This shows $\omega = df$.

(3 \implies 1:)

If $\omega = df$, then $\int_{\gamma} \omega = f(Q) - f(P)$ where P and Q are the initial and final point of γ . This is also true for another path δ with the same initial and final points. This shows (1) holds. ■

18.3 Two and Three Forms

A 1-form $\omega = A dx + B dy$ is therefore a mapping $\omega : \mathbb{R}_P^2 = T_P(\mathbb{R}^2) \rightarrow \mathbb{R}$ defined by

$$\omega(V_P) = A(P_1, P_2)V_1 + B(P_1, P_2)V_2$$

Now consider mappings of the form $\psi = Adx dy$. Then, we define the action of ψ as follows:

Definition 18.3.1 2 - Forms

Let U be an open set in \mathbb{R}^2 . a 2-form $\psi : \mathbb{R}_P^2 \times \mathbb{R}_P^2 \rightarrow \mathbb{R}$ of the form $A dx dy$ for a smooth function A on U has action defined by

$$\begin{aligned}\psi(V_P, W_P) &= A dx dy(V_P, W_P) \\ &= A(P_1, P_2) \left(dx(V_P)dy(W_P) - dx(W_P)dy(V_P) \right) \\ &= A(P_1, P_2)(V_1 W_2 - V_2 W_1)\end{aligned}$$

Note it is clear from this definition that $dx dy = -dy dx$. Further, $dx dx = dy dy = 0$.

Next, we can look at 3-forms. To do this, we have to figure out a way to alternate the algebraic sign of the terms in the expansion of $dxdydz$ as well as move the setting to open sets U in \mathbb{R}^3 . You probably have seen the permutation groups S_n on n symbols before. Let's refresh that memory. First, let's look at two symbols:

Definition 18.3.2 The Symmetric Group S_2

For concreteness, the symmetric groups S_2 will be phrased in terms of positive integers. S_2 is the group consisting of the two ways the symbols 1 and 2 can be shuffled. These are the ordering 12 and the ordering 21. The ordering 21 can be obtained from 12 by flipping the entries: i.e. $12 \rightarrow 21$. Since one flip is involved, we define the algebraic sign of the ordering 21 to be -1 while since there are an even number of flips (i.e zero to be exact) the algebraic sign of 12 is $+$. We let $\sigma(i_1 i_2)$ be the sign of the ordering.

Next, three symbols:

Definition 18.3.3 The Symmetric Group S_3

For concreteness, the symmetric groups S_3 will be phrased in terms of positive integers. S_3 is the group consisting of the ways the symbols 1, 2 and 3 can be shuffled. We let $\sigma(i_1 i_2 i_3)$ be the sign of the ordering.

- 123 does not need reordering so $\sigma = +1$.
- 213: $123 \rightarrow 213$; so one flip implying $\sigma = -1$.
- 321: $123 \rightarrow 132 \rightarrow 312 \rightarrow 321$: three flips implying $\sigma = (-1)^3 = -1$.
- 132: $123 \rightarrow 132$: so one flip implying $\sigma = -1$
- 312: $123 \rightarrow 132 \rightarrow 312$: so two flips implying $\sigma = (-1)^2 + +1$.
- 231: $123 \rightarrow 213 \rightarrow 231$: so two flips implying $\sigma = (-1)^2 = +1$.

Using the symmetric group S_3 , we can define what we mean by $dxdydz$.

Definition 18.3.4 3 - Forms

Let U be an open subset of \mathbb{R}^3 . Then a 3 - form $\psi : \mathbb{R}_P^3 \times \mathbb{R}_P^3 \times \mathbb{R}_P^3 \rightarrow \mathbb{R}$ of the form $\mathbf{A} dx^1 dx^2 dx^3$ for a smooth function \mathbf{A} on U has action defined by

$$\begin{aligned}\psi(X_P^1, X_P^2, X_P^3) &= \mathbf{A} dx^1 dx^2 dx^3(X_P^1, X_P^2, X_P^3) \\ &= \mathbf{A}(P_1, P_2, P_3) \sum_{(i_1 i_2 i_3) \in S_3} \sigma(i_1 i_2 i_3) dx^1(X_P^{i_1}) dx^2(X_P^{i_2}) dx^3(X_P^{i_3})\end{aligned}$$

Comment 18.3.1 It is easier to remember this using determinants. For a 2 -form

$$\begin{aligned}dx dy(V_P, W_P) &= dx(V_P) dy(W_P) - dx(W_P) dy(V_P) \\ &= (V_1 W_2 - W_1 V_2) = \det \begin{bmatrix} V_1 & V_2 \\ W_1 & W_2 \end{bmatrix} \\ &= \left\langle \det \begin{bmatrix} i & j & k \\ V_1 & V_2 & 0 \\ W_1 & W_2 & 0 \end{bmatrix}, k \right\rangle\end{aligned}$$

Comment 18.3.2 Hence, for a 3 - Form, we can use determinants also:

$$\begin{aligned}&dx^1 dx^2 dx^3(X_P^1, X_P^2, X_P^3) \\ &(+ dx^1 dx^2 dx^3(X_P^1, X_P^2, X_P^3)) (- dx^1 dx^2 dx^3(X_P^2, X_P^1, X_P^3)) \\ &(- dx^1 dx^2 dx^3(X_P^3, X_P^1, X_P^2)) (- dx^1 dx^2 dx^3(X_P^1, X_P^3, X_P^2)) \\ &(+ dx^1 dx^2 dx^3(X_P^3, X_P^1, X_P^2)) (+ dx^1 dx^2 dx^3(X_P^2, X_P^3, X_P^1))\end{aligned}$$

Thus, we have

$$\begin{aligned}dx^1 dx^2 dx^3(X_P^1, X_P^2, X_P^3) &= X^{11} X^{22} X^{33} - X^{21} X^{12} X^{33} - X^{31} X^{22} X^{13} \\ &\quad - X^{11} X^{32} X^{23} + X^{31} X^{12} X^{23} + X^{21} X^{32} X^{13}\end{aligned}$$

We can rewrite this as

$$\begin{aligned}dx^1 dx^2 dx^3(X_P^1, X_P^2, X_P^3) &= X^{11}(X^{22} X^{33} - X^{23} X^{32}) - X^{12}(X^{31} X^{23} - X^{21} X^{33}) \\ &\quad + X^{13}(X^{21} X^{32} - X^{22} X^{31})\end{aligned}$$

This can be reorganized like follows:

$$\begin{aligned}dx^1 dx^2 dx^3(X_P^1, X_P^2, X_P^3) &= X^{11} \det \begin{bmatrix} X^{22} & X^{23} \\ X^{32} & X^{33} \end{bmatrix} - X^{12} \det \begin{bmatrix} X^{21} & X^{23} \\ X^{31} & X^{33} \end{bmatrix} \\ &\quad + X^{13} \det \begin{bmatrix} X^{21} & X^{22} \\ X^{31} & X^{32} \end{bmatrix} \\ &= \det \begin{bmatrix} X^{11} & X^{12} & X^{13} \\ X^{21} & X^{22} & X^{23} \\ X^{31} & X^{32} & X^{33} \end{bmatrix}\end{aligned}$$

Comment 18.3.3 From the 3 - Form expansion, if you look at $dx^2 dx^2 dx^3$, we would find

$$dx^2 dx^2 dx^3(X_P^1, X_P^2, X_P^3) = \det \begin{bmatrix} X^{21} & X^{22} & X^{23} \\ X^{21} & X^{22} & X^{23} \\ X^{31} & X^{32} & X^{33} \end{bmatrix}$$

Since two rows in the determinant are the same, the determinant is zero. A similar argument shows $dx^i dx^j dx^k$ is zero if any two superscripts are the same. Thus,

$$dxdxdz = 0, \quad dxdydy = 0, \quad dxdzdz = 0$$

and so forth. Finally, changing order gives

$$dx^2 dx^1 dx^3 (X_P^1, X_P^2, X_P^3) = \det \begin{bmatrix} X^{21} & X^{22} & X^{23} \\ X^{11} & X^{12} & X^{13} \\ X^{31} & X^{32} & X^{33} \end{bmatrix}$$

which interchanges two rows in the matrix and hence the determinant changes sign. Thus, we know

$$dxdydz = -dydxdz, \quad dxdydy = -dxdzdy, \quad dxdydz = -dzdydx$$

and so forth.

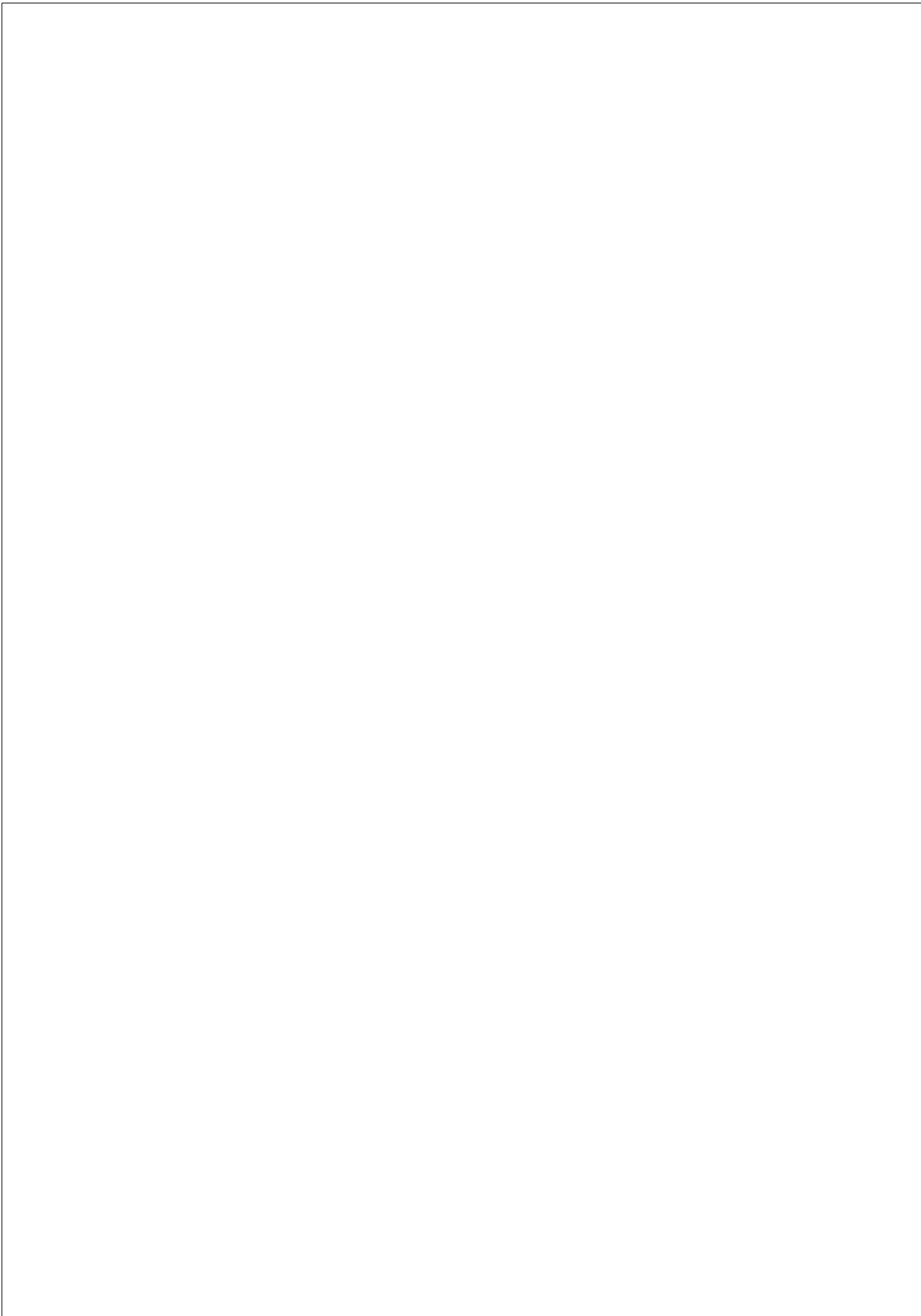
18.4 Exterior Derivatives

18.5 Integration of Forms

18.6 Green’s Theorem Revisited

Part VI

Applications



Chapter 19

The Exponential Matrix

Let A be a $n \times n$ matrix. We will let $\|A\|_{Fr}$ be the Frobenius norm of A ; hence, we know $\|Ax\| \leq \|A\|_{Fr} \|x\|$. For convenience, we will simply let $\|A\|_{Fr}$ be denoted by $\|A\|$. Before we begin, lets consider the product of two $n \times n$ matrices A and B . We have

$$C_{ij} = (AB)_{ij} = \sum_{k=1}^n A_{ik} B_{kj} \leq \sqrt{\sum_{k=1}^n A_{ik}^2} \sqrt{\sum_{k=1}^n B_{kj}^2}$$

Thus

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n C_{ij}^2 &\leq \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n A_{ik}^2 B_{kj}^2 \\ &= \left(\sum_{i=1}^n \sum_{k=1}^n A_{ik}^2 \right) \left(\sum_{j=1}^n \sum_{k=1}^n B_{kj}^2 \right) = \|A\|_{Fr}^2 \|B\|_{Fr}^2 \end{aligned}$$

Therefore $\|AB\|_{Fr} \leq \|A\|_{Fr} \|B\|_{Fr}$. We will use this in a bit.

19.1 The Exponential Matrix

Now consider the finite sum

$$P_n = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots + \frac{A^n}{n!}$$

Does $\lim_{n \rightarrow \infty} P_n$ exist? i.e. is there a matrix P such that $P = \lim_{n \rightarrow \infty} P_n$? Thus, we want to show

$$\forall \epsilon > 0, \exists N \ni \|P - P_n\| < \epsilon, \text{ if } n > N$$

Formally, letting $A^0 = I$, we suspect the matrix we seek is expressed as the following infinite series.

$$P = \sum_{j=0}^{\infty} \frac{A^j}{j!}$$

It is straightforward to show the set of all $n \times n$ real matrices form a complete normed vector space under the Frobenius norm. This is not surprising as if $M_{n,n}$ denotes this set of matrices, it can be

identified with \Re^{n^2} and this space is complete using the usual $\|\cdot\|_2$ norm. The Frobenius norm is essentially just the $\|\cdot\|_2$ norm applied to objects requiring a double summation. So if we can show the sequence of partial sums is a Cauchy Sequence with respect to the Frobenius norm, we know there is a unique matrix P which can be denoted by the series. For $n > m$, $P_n - P_m = \sum_{j=m+1}^n \frac{A^j}{j!}$ and

$$\|P_n - P_m\| \leq \sum_{j=m+1}^n \frac{\|A^j\|}{j!} \leq \sum_{j=m+1}^n \frac{\|A\|^j}{j!}$$

where we have used induction to show $\|A^n\| \leq \|A\|^n$. But $\|A\| = \alpha$ is a constant, so

$$\|P_n - P_m\| \leq \sum_{j=m+1}^n \frac{(\alpha)^j}{j!}$$

We know that

$$e^\alpha = \sum_{j=0}^{\infty} \frac{(\alpha)^j}{j!}$$

converges and so it is a Cauchy Sequence of real numbers. Therefore, given $\epsilon > 0$, there is a N so that $\sum_{j=m+1}^n \alpha^j / (j!) < \epsilon$ when $n > m > N$. We conclude if $n > m > N$,

$$\|P_n - P_m\| \leq \sum_{j=m+1}^n \frac{(\alpha)^j}{j!} < \epsilon$$

Thus the sequence of partial sums is a Cauchy Sequence and by the completeness of $M_{n,n}$ with the Frobenius norm, there is a unique matrix P so that $P = \lim_{n \rightarrow \infty} P_n$ in the Frobenius norm. The partial sums P_n are exactly the partial sums we would expect for a function like e^A even though A is a matrix. Hence, we use this sequence of partial sums to define what we mean by e^A which is called the **matrix exponential function**. Since this analysis works for any matrix A , in particular, given a value of t , it works for the matrix At to give the matrix e^{At} .

Homework

Exercise 19.1.1

Exercise 19.1.2

Exercise 19.1.3

Exercise 19.1.4

Exercise 19.1.5

Let's apply this to systems of linear differential equations Let $P(t) = e^{At}$ denote the **matrix exponential function**. consider the system

$$\begin{bmatrix} x'_1 \\ \vdots \\ x'_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \cdots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

i.e. the linear system of differential equations

$$\mathbf{x}' = A\mathbf{x}$$

where A is a constant matrix. We will show the columns of e^{At} give the basis of the set of solutions to this problem. Note in Chapter 6, we looked at this type of system for a symmetric 2×2 matrix A and found the general solution to that system $\mathbf{x}' = A\mathbf{x}$ is $\mathbf{x}(t) = e^{At}\mathbf{C}$ where

$$\begin{aligned} e^{At} &= \sum_{j=0}^{\infty} (At)^j / (j!) \\ &= [\mathbf{E}_1 \quad \mathbf{E}_2] \begin{bmatrix} \sum_{k=0}^n (\lambda_1^k t^k) / k! = e^{\lambda_1 t} & 0 \\ 0 & \sum_{k=0}^n (\lambda_2^k t^k) / k! = e^{\lambda_2 t} \end{bmatrix} [\mathbf{E}_1 \quad \mathbf{E}_2]^T \end{aligned}$$

where \mathbf{E}_1 and \mathbf{E}_2 were the orthogonal basis of eigenvectors to A known to exist because A was 2×2 symmetric. So in this case, the matrix exponential function was easy to compute. In general, the $n \times n$ matrix A does not have such a nice structure, but we still now how to build it using partial sums of powers of A .

19.1.1 Homework

Exercise 19.1.6

Exercise 19.1.7

Exercise 19.1.8

Exercise 19.1.9

Exercise 19.1.10

19.2 The Jordan Canonical Form

In an advanced course in linear algebra, you will learn about the **Jordan Canonical Form** of a matrix. We can prove every matrix A has one and it is similar to the decomposition we find for symmetric matrices in a way. A crucial part of the Jordan Canonical form are what are called **Jordan Blocks**. Consider the three 3×3 matrices below where λ is an eigenvalue. (These could possibly be submatrices of a larger matrix.)

$$\begin{array}{c} \left[\begin{array}{ccc} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{array} \right] \quad \left[\begin{array}{ccc} \lambda & 0 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{array} \right] \quad \left[\begin{array}{ccc} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{array} \right] \\ (1) \qquad \qquad \qquad (2) \qquad \qquad \qquad (3) \end{array}$$

How many eigenvectors do we get for each of these? We know that for (3) we will get 3 distinct eigenvectors, but what about for (1) and (2)? Recall, to get the eigenvalues of a matrix we look at

$$\det(A - \mu I) = 0 \qquad \qquad (\det(\mu I - A) = 0)$$

and then to find the eigenvectors we solve

$$(A - \mu I)(\mathbf{v}) = \mathbf{0}$$

(Here μ is used to denote the eigenvalues because λ is being used in the matrices.) So, for matrix (1) we get

$$\det(A - \mu I) = \begin{vmatrix} \lambda - \mu & 1 & 0 \\ 0 & \lambda - \mu & 1 \\ 0 & 0 & \lambda - \mu \end{vmatrix} = 0$$

$$(\lambda - \mu)^3 = 0$$

So $\mu = \lambda$ and λ is our eigenvalue. To get the eigenvectors of (1) we solve

$$\left[\begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} - \begin{pmatrix} \mu & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \mu \end{pmatrix} \right] \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \mathbf{0}$$

or

$$\begin{pmatrix} \lambda - \mu & 1 & 0 \\ 0 & \lambda - \mu & 1 \\ 0 & 0 & \lambda - \mu \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

We see that $v_1 = v_2 = 0$ and v_3 is arbitrary. Thus from (1) we only get one distinct eigenvector. The eigenvalue λ has algebraic multiplicity 3 as it occurs three times in the characteristic equation. It has geometric multiplicity 1 as it has only one eigenvector. Hence, the eigenspace for this eigenvalue of algebraic multiplicity three is one one dimensional. It is clear means in this case the eigenvectors associated with the eigenvalue do not give a subspace of the same dimension as the algebraic multiplicity. Note the eigenvalues are on the diagonal and the superdiagonal right above that is all 1's. This is the first of the Jordan Canonical Blocks possible for this eigenvalue of multiplicity 3. Note there are 2 1s in the superdiagonal and there are $(3 - 2)$ eigenvectors here.

In a similar fashion we can show that matrix (2) has 2 distinct eigenvectors. Note it can be written like

$$\left[\begin{array}{c|cc} \lambda & 0 & 0 \\ \hline 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{array} \right]$$

We interpret this as follows. Here the algebraic multiplicity is still 3 but now the geometric multiplicity is now 2. Note the structure of this matrix is different from (1). The $(1, 1)$ entry is just λ and this top position corresponds to a one dimensional eigenspace. The bottom 2×2 submatrix has 1's on the diagonal and a 1 on the superdiagonal about that. Since it is a 2×2 submatrix, the superdiagonal has only one entry and there are $(3 - 1)$ eigenvectors here. This is the second type of Jordan Canonical block associated with an eigenvalue of algebraic multiplicity 3.

In case (3), the eigenvalues form the diagonal and the superdiagonal about it is all 0's. Note there are 3 eigenvectors now which is the same as $(3 - 0)$ where the 0 is the number of 1s on the superdiagonal. Here the eigenvalue has algebraic multiplicity 3 and geometric multiplicity 3 also. This is the third type of Jordan Canonical block associated with an eigenvalue of algebraic multiplicity 3.

The Jordan Canonical Blocks for an eigenvalue λ of algebraic multiplicity 4 are then

$$\begin{array}{cccc} \left[\begin{array}{cccc} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{array} \right] & \left[\begin{array}{cccc} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{array} \right] & \left[\begin{array}{cccc} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{array} \right] & \left[\begin{array}{cccc} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{array} \right] \\ (1) & (2) & (3) & (4) \\ AM = 4 & AM = 4 & AM = 4 & AM = 4 \\ GM = 1 & GM = 2 & GM = 3 & GM = 4 \end{array}$$

where AM and GM are the algebraic and geometric multiplicity of the eigenvalue in each Jordan Canonical Block. From now on, let a Jordan Canonical Block be denoted by JCB for convenience.

Homework

Exercise 19.2.1

Exercise 19.2.2

Exercise 19.2.3

Exercise 19.2.4

Exercise 19.2.5

What happens if we multiply a JCB by itself many times?

$$\begin{aligned} A &= \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} \\ A^2 &= \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} = \begin{pmatrix} \lambda^2 & 2\lambda & 1 \\ 0 & \lambda^2 & 2\lambda \\ 0 & 0 & \lambda^2 \end{pmatrix} \\ A^3 &= \begin{pmatrix} \lambda^2 & 2\lambda & 1 \\ 0 & \lambda^2 & 2\lambda \\ 0 & 0 & \lambda^2 \end{pmatrix} \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} = \begin{pmatrix} \lambda^3 & 3\lambda^2 & 3\lambda \\ 0 & \lambda^3 & 3\lambda \\ 0 & 0 & \lambda^3 \end{pmatrix} \end{aligned}$$

We will get the same sort of formulas for A^4 , A^5 , etc. You can see if we could compute e^{At} for a JCB then we could compute it for a matrix written in a form containing only JCB's.

The general Jordan Canonical Block is thus

$$J = \underbrace{\left[\begin{array}{ccccc} \lambda & 1 & & & 0 \\ & \ddots & \ddots & & \\ & & \ddots & & \\ & & & \ddots & 1 \\ 0 & & & & \lambda \end{array} \right]}_n \Bigg\} n$$

We can rewrite this as

$$J = \lambda I + Z, \quad Z = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}$$

with ones on the super diagonal. It turns out that the matrix Z is nilpotent, i.e. $Z^n = 0$ for some n . Let's show this:

$$Z = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix}$$

$$Z^2 = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix}$$

$$z_{ij}^2 = \sum_{l=1}^n z_{il} z_{lj} = z_{i,i+1} z_{i+1,j} = \begin{cases} 1 & \text{if } j = i+2 \\ 0 & \text{otherwise} \end{cases}$$

So

$$Z^2 = \begin{bmatrix} 0 & 0 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & \ddots & 0 \\ 0 & & & & 0 \end{bmatrix}$$

$$Z^3 = Z^2 Z = ZZ^2$$

$$z_{ij}^3 = \sum_{l=1}^n z_{il}^2 z_{lj} = z_{i,i+2} z_{i+2,j} = \begin{cases} 1 & \text{if } j = i+3 \\ 0 & \text{otherwise} \end{cases}$$

19.3. EXPONENTIAL MATRIX CALCULATIONS

387

We find

$$Z^3 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ & \ddots & \ddots & \ddots & 1 \\ & & \ddots & \ddots & 0 \\ 0 & & & \ddots & 0 \end{bmatrix} \leftarrow \begin{array}{l} \text{main-3 } (i, i+3) \\ \text{main-2 } (i, i+2) \\ \text{main-1 } (i, i+1) \\ \text{main} \end{array}$$

We can only go so far though. After $n - 1$ iterations, we have

$$Z^{n-1} = \begin{bmatrix} 0 & & 1 \\ & \ddots & \\ 0 & & 0 \end{bmatrix}, \quad Z^n = 0$$

Therefore Z is a nilpotent matrix.

19.2.1 Homework

Exercise 19.2.6

Exercise 19.2.7

Exercise 19.2.8

Exercise 19.2.9

Exercise 19.2.10

19.3 Exponential Matrix Calculations

Theorem 19.3.1 If A and B commute, $e^{A+B} = e^A e^B$

Let A and B be $n \times n$ matrices which commute; i.e. $AB = BA$. Then if A and B commute then $e^{A+B} = e^A e^B$

Proof 19.3.1

We claim

$$\sum_{i=0}^n (A^i / (i!)) \sum_{j=0}^n (B^j / (j!)) = \sum_{k=0}^n \sum_{j=0}^k \frac{A^j B^{k-j}}{j!(k-j)!}$$

To see this, consider the array

$$\begin{bmatrix} (0,0) & (0,1) & (0,2) & \dots & (0,n) \\ (1,0) & (1,1) & (1,2) & \dots & (1,n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (n,0) & (n,1) & (n,2) & \dots & (n,n) \end{bmatrix}$$

Summing over all $(n+1)^2$ entries indexed by this array is the same as summing over the diagonals here

$$\begin{bmatrix} (i+j) = 0 & (i+j) = 1 & (i+j) = 2 & \dots & (i+j) = n \\ (i+j) = 1 & (i+j) = 2 & \vdots & \vdots & (i+j) = n+1 \\ (i+j) = 2 & \vdots & \vdots & \vdots & \vdots \\ \vdots & (i+j) = n & \vdots & \vdots & \vdots \\ (i+j) = n & (i+j) = n+1 & \dots & \dots & (i+j) = 2n \end{bmatrix}$$

Let Z_0 be the set of non negative integers. The scheme shown above defines a 1-1 and onto mapping g from $Z_0 \times Z_0$ to Z_0 . If we let $a_{ij} = A^i B^j / (i! j!)$, we are wondering if the $\sum_i \sum_j a_{ij}$ is the same as the $\sum a_{g(i,j)}$ where $a_{g(i,j)}$ is the particular index in the scheme above. Thus, we are wondering if

$$\begin{aligned} \sum_{i=0}^n \sum_{j=0}^n (A^i / (i!)) (B^j / (j!)) &= a_{g(0,0)} + (a_{g(1,0)} + a_{g(0,1)}) + (a_{g(2,0)} + a_{g(1,1)} + a_{g(0,2)}) + \dots \\ &= \sum_{k=0}^n \sum_{j=0}^k \frac{A^j B^{k-j}}{j!(k-j)!} \end{aligned}$$

The $\sum_{i=0}^n \sum_{j=0}^n (A^i / (i!)) (B^j / (j!))$ converges to $e^A e^B$. Hence, it is a Cauchy Sequence in the Frobenius norm. If you think about the indexing scheme, we can see

$$\left(\sum_{i+j=2n} + \sum_{i+j=3n} + \dots + \sum_{i+j=pn} \right) A^i B^j / (i! j!)$$

does not contain the terms $\sum_{i=0}^n \sum_{j=0}^n A^i B^j / (i! j!)$. Further all of these terms are contained in the difference

$$\begin{aligned} &\left(\sum_{i=0}^{pn} \sum_{j=0}^{pn} - \sum_{i=0}^n \sum_{j=0}^n \right) A^i / (i!) B^j / (j!) \\ &\left(\sum_{i=0}^n \sum_{j=n+1}^{pn} + \sum_{i=n+1}^{pn} \sum_{j=0}^n + \sum_{i=n+1}^{pn} \sum_{j=n+1}^{pn} \right) A^i / (i!) B^j / (j!) \end{aligned}$$

Thus,

$$\begin{aligned} &\left\| \left(\sum_{i+j=2n} + \sum_{i+j=3n} + \dots + \sum_{i+j=pn} \right) A^i B^j / (i! j!) \right\| \\ &\leq \left(\sum_{i+j=2n} + \sum_{i+j=3n} + \dots + \sum_{i+j=pn} \right) \|A\|^i \|B\|^j / (i! j!) \\ &\leq \left(\sum_{i=0}^n \sum_{j=n+1}^{pn} + \sum_{i=n+1}^{pn} \sum_{j=0}^n + \sum_{i=n+1}^{pn} \sum_{j=n+1}^{pn} \right) \|A\|^i \|B\|^j / (i! j!) \end{aligned}$$

Now we can simplify. We have

$$\sum_{i=0}^n \|A\|^i / (i!) \sum_{j=n+1}^{pn} \|B\|^j / (j!) + \sum_{i=n+1}^{pn} \|A\|^i / (i!) \sum_{j=0}^n \|B\|^j / (j!)$$

$$\begin{aligned}
& + \sum_{i=n+1}^{pn} \|A\|^i / (i!) \sum_{j=n+1}^{pn} \|B\|^j / (j!) \\
& \leq e^{\|A\|} \sum_{j=n+1}^{pn} \|B\|^j / (j!) + e^{\|B\|} \sum_{i=n+1}^{pn} \|A\|^i / (i!) + \sum_{i=n+1}^{pn} \sum_{j=n+1}^{pn} \|A\|^i \|B\|^j / (i!j!)
\end{aligned}$$

Hence, given $\epsilon > 0$, there are integers Q_1 , Q_2 and Q_3 so that

$$\begin{aligned}
pn > n + 1 > Q_1 & \implies e^{\|A\|} \sum_{j=n+1}^{pn} \|B\|^j / (j!) < \epsilon/3 \\
pn > n + 1 > Q_2 & \implies e^{\|B\|} \sum_{i=n+1}^{pn} \|A\|^i / (i!) < \epsilon/3 \\
pn > n + 1 > Q_3 & \implies \sum_{i=n+1}^{pn} \sum_{j=n+1}^{pn} \|A\|^i \|B\|^j / (i!j!) < \epsilon/3
\end{aligned}$$

because $\sum_{i=0}^n \|A\|^i / (i!) \rightarrow e^{\|A\|}$ and $\sum_{j=0}^n \|B\|^j / (j!) \rightarrow e^{\|B\|}$. Thus, for $pn > n + 1 > \max\{Q_1, Q_2, Q_3\}$, we have

$$\left\| \left(\sum_{i+j=2n} + \sum_{i+j=3n} + \dots + \sum_{i+j=pn} \right) A^i B^j / (i!j!) \right\| < \epsilon$$

which tells us the sequence $\sum_{g(i,j)=0}^{g(i,j)=n}$ is a Cauchy Sequence and so converges to matrix C . Finally, consider

$$\begin{aligned}
\|e^A e^B - C\| & \leq \left\| e^A e^B - \sum_{i=0}^n \sum_{j=0}^n A^i B^j / (i!j!) \right\| + \left\| C - \sum_{g(i,j)=0}^{g(i,j)=N} A^i B^j / (i!j!) \right\| \\
& + \left\| \sum_{i=0}^n \sum_{j=0}^n A^i B^j / (i!j!) - \sum_{g(i,j)=0}^{g(i,j)=N} A^i B^j / (i!j!) \right\|
\end{aligned}$$

For a given $\epsilon > 0$, the first and second terms can be made less than $\epsilon/3$ because of the convergence of the sums to $e^A e^B$ and C respectively. Thus, there is an Q_1 so that if $n > Q_1$,

$$\|e^A e^B - C\| < 2\epsilon/3 + \left\| \sum_{i=0}^n \sum_{j=0}^n A^i B^j / (i!j!) - \sum_{g(i,j)=0}^{g(i,j)=N} A^i B^j / (i!j!) \right\|$$

The last term is analyzed just like we did in the argument to show the sequence defining C was a Cauchy Sequence. We find there is a Q_2 so that is $n > N > Q_2$

$$\left\| \sum_{i=0}^n \sum_{j=0}^M A^i B^j / (i!j!) - \sum_{g(i,j)=0}^{g(i,j)=N} A^i B^j / (i!j!) \right\| < \epsilon/3$$

Thus, for large enough choices of n and N , we have $\|e^A e^B - C\| < \epsilon$. Since ϵ is arbitrary, this shows $C = e^A e^B$.

Thus,

$$\begin{aligned}
 e^A e^B &= \left(I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots + \frac{A^k}{k!} + \cdots \right) \\
 &\quad \left(I + B + \frac{B^2}{2!} + \frac{B^3}{3!} + \cdots + \frac{B^k}{k!} + \cdots \right) \\
 &= I + \\
 &\quad \frac{B}{1!} + \frac{A}{1!} + \\
 &\quad \frac{B^2}{2!} + \frac{AB}{1!1!} + \frac{A^2}{2!} + \\
 &\quad \frac{B^3}{3!} + \frac{AB^2}{1!2!} + \frac{A^2B}{2!1!} + \frac{A^3}{3!} + \\
 &\quad \vdots \\
 &\quad \frac{B^k}{k!} + \frac{AB^{k-1}}{1!(k-1)!} + \frac{A^2B^{k-2}}{2!(k-2)!} + \\
 &\quad \vdots
 \end{aligned}$$

Then, using the fact that A and B commute,

$$\begin{aligned}
 e^A e^B &= \sum_{n=0}^{\infty} \sum_{j=0}^n \frac{A^j B^{n-j}}{j!(n-j)!} \\
 &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{j=0}^n \frac{n!}{j!(n-j)!} A^j B^{n-j} \\
 &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{j=0}^n \binom{n}{j} A^j B^{n-j} \\
 &= \sum_{n=0}^{\infty} \frac{(A+B)^n}{n!} \quad \text{since } \sum_{j=0}^n \binom{n}{j} A^j B^{n-j} = (A+B)^n \\
 &= e^{A+B}
 \end{aligned}$$

■

Homework

Exercise 19.3.1

Exercise 19.3.2

Exercise 19.3.3

Exercise 19.3.4

Exercise 19.3.5

19.3.1 Jordan Block Matrices

In applying this to a Jordan Block we get

$$\begin{aligned} J &= \lambda I + Z \\ e^{Jt} &= e^{\lambda It + Zt} = e^{\lambda It} e^{Zt} \end{aligned}$$

where

$$\begin{aligned} e^{\lambda It} &= I + t\lambda I + \frac{(t\lambda I)^2}{2!} + \cdots + \frac{(t\lambda I)^n}{n!} + \cdots \\ &= I + t\lambda I + \frac{(t\lambda)^2 I}{2!} + \cdots + \frac{(t\lambda)^n I}{n!} + \cdots \\ &= \text{diag}[1 + \lambda t + \frac{(\lambda t)^2}{2!} + \cdots + \frac{(\lambda t)^n}{n!} + \cdots] \\ &= \text{diag}[e^{\lambda t}] = e^{\lambda t} I \end{aligned}$$

where **diag** denotes a diagonal matrix. We also find

$$\begin{aligned} e^{Zt} &= I + tZ + \frac{t^2 Z^2}{2!} + \cdots + \frac{t^{n-1} Z^{n-1}}{(n-1)!} + \cdots \\ &= \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix} + t \begin{bmatrix} 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & 1 & \\ & & & & 0 \end{bmatrix} + \cdots + \frac{t^{n-1}}{(n-1)!} \begin{bmatrix} 0 & 0 & & & 1 \\ & \ddots & \ddots & & \\ & & \ddots & 0 & \\ & & & & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & t & \frac{t^2}{2!} & & \frac{t^{n-1}}{(n-1)!} \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \frac{t^2}{2!} & \\ & & & t & \\ & & & & 1 \end{bmatrix} \leftarrow \begin{array}{l} \text{main-2} \\ \leftarrow \text{main-1} \\ \leftarrow \text{main} \end{array} \end{aligned}$$

So then

$$e^{Jt} = e^{\lambda It} e^{Zt} = e^{\lambda t} \begin{bmatrix} 1 & t & \frac{t^2}{2} & & \frac{t^{n-1}}{(n-1)!} \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \frac{t^2}{2} \\ & & & t & \\ & & & & 1 \end{bmatrix}$$

So if the Jordan form of the matrix A was

$$J = \begin{bmatrix} 2 & & & \\ & 2 & 1 & \\ & & 2 & \\ & & & -3 & 0 \\ & & & & -3 \\ & & & & -3 & 1 \\ & & & & & -3 \end{bmatrix}$$

then

$$e^{Jt} = \begin{bmatrix} e^{2t} & & & \\ & e^{2t} & te^{2t} & \\ & & e^{2t} & \\ & & & e^{-3t} \\ & & & & e^{-3t} \\ & & & & & e^{-3t} & te^{-3t} \\ & & & & & & e^{-3t} \end{bmatrix}$$

Homework

Exercise 19.3.6

Exercise 19.3.7

Exercise 19.3.8

Exercise 19.3.9

Exercise 19.3.10

19.3.2 General Matrices

In general, for an $n \times n$ matrix A there is an invertible matrix P such that

$$A = PJP^{-1}$$

where J is in Jordan canonical form. So then,

$$e^{At} = I + PJP^{-1}t + \overbrace{(PJP^{-1}PJP^{-1})}^{PJ^2P^{-1}} \frac{t^2}{2!} + \cdots + (PJ^n P^{-1}) \frac{t^n}{n!}$$

and we can easily prove we can factor out the P and P^{-1} to get

$$\begin{aligned} e^{At} &= P(I + J + \frac{J^2}{2!} + \cdots + \frac{J^n}{n!})P^{-1} \\ &= Pe^{Jt}P^{-1} \end{aligned}$$

But how easy are P and P^{-1} to find? Look at a 2×2 matrix A having eigenvalues λ_1 and λ_2 and eigenvectors v_1 and v_2 such that $v_1 \perp v_2$ and each has unit length. Then, as we know

$$\begin{aligned} \underbrace{\begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix}}_Q A \underbrace{\begin{pmatrix} v_1 & v_2 \end{pmatrix}}_{Q^T} &= \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \begin{pmatrix} Av_1 & Av_2 \end{pmatrix} \\ &= \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} \begin{pmatrix} \lambda_1 v_1 & \lambda_2 v_2 \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 v_1^T v_1 & \lambda_2 v_2^T v_2 \\ \lambda_1 v_2^T v_1 & \lambda_2 v_2^T v_2 \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \end{aligned}$$

So $QAQ^T = J$ where J is the above matrix in Jordan canonical form. Then

$$\begin{aligned} AQ^T &= Q^T J \\ A &= Q^T J Q \end{aligned}$$

and set

$$\begin{aligned} P &= Q^T = (v_1, v_2) \\ P^{-1} &= Q \end{aligned}$$

We can therefore do this fairly easily for symmetric matrices. From this we can see that certain matrices have some nice forms and things just fall out nicely.

19.3.3 Homework

Exercise 19.3.11

Exercise 19.3.12

Exercise 19.3.13

Exercise 19.3.14

Exercise 19.3.15

19.4 Applications to Linear ODE

Consider the following system:

$$\mathbf{x}' = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & \cdots & a_{1n} \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = A\mathbf{x}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 = \begin{bmatrix} x_{01} \\ x_{02} \\ \vdots \\ x_{0n} \end{bmatrix}$$

Theorem 19.4.1 The Derivative of the exponential matrix

Let A be an $n \times n$ matrix. Then $\frac{d}{dt}(e^{At}) = Ae^{At}$.

Proof 19.4.1

Since $e^{At} = \sum_{j=0}^{\infty} \frac{A^j}{j!} t^j$, to find its derivative, we look at

$$\frac{d}{dt}(e^{At}) = \frac{d}{dt}\left(\sum_{j=0}^{\infty} \frac{A^j}{j!} t^j\right)$$

We want to use an analogue of the derivative interchange theorem we proved in (Peterson (8) 2019) to show

$$\frac{d}{dt}\left(\sum_{j=0}^{\infty} \frac{A^j}{j!} t^j\right) = \sum_{j=0}^{\infty} \frac{d}{dt}\left(\frac{A^j}{j!} t^j\right)$$

The series we have here is a series of matrices and convergence is with respect to the Frobenius norm so it is not immediately clear the earlier results apply. For a series of real numbers, we would perform the following checks applied to the sequence of partial sums S_n . First, we fix a T . We need to check

1. S_n is differentiable on $[0, T]$. It is easy to see

$$\left(\sum_{j=0}^n \frac{A^j}{j!} t^j\right)' = \sum_{j=0}^n j \frac{A^j}{j!} t^{j-1}$$

as the derivative $Bf(t)$ for a matrix B and a differentiable function f is easily seen to be $Bf'(t)$ by looking at the derivative of each component.

2. S'_n is Riemann Integrable on $[0, t]$: This is true as the integral of $Bf(t)$ for any integrable f is $B \int_0^t f(s)ds$ for any matrix B by looking at components.
3. There is at least one point $t_0 \in [0, t]$ such that the sequence $(S_n(t_0))$ converges. This is true as the series here converges at all t .
4. $S'_n \xrightarrow{\text{unif}} y$ on $[0, T]$ and the limit function y is continuous. This one is tougher. What we have shown in our work so far is that S_n is a Cauchy Sequence with respect to the Frobenius norm. In fact, we know that if $T_n(t) = \sum_{i=0}^n \|A\|^i / (i!)$ is a partial sum of $e^{\|A\|t}$ which we know converges uniformly on any finite interval $[0, T]$, we use it to bound the partial sums S_n . Hence, for any $\epsilon > 0$, there is N so that $n > m > N$ implies $\|S_n(t) - S_m(t)\|_{Fr} \leq |T_n(t) - T_m(t)| < \epsilon$ on $[0, T]$. If you look at the arguments we use in proving these sorts of theorems in (Peterson (8) 2019), you can see how they can be modified to show S_n converges uniformly on $[0, T]$ to a continuous function e^{At} . Everything thing just said can be modified a bit to prove S'_n converges uniformly to some matrix function D on $[0, T]$ as well.

You can then modify the proof of the derivative interchange theorem to show there is a function W on $[0, T]$ so that $S_n \xrightarrow{\text{unif}} W$ on $[0, T]$ and $W' = D$. Since limits are unique, we then have $W = e^{At}$ with $(e^{At})' = D$. This is the same as saying the interchange works and

$$\frac{d}{dt}\left(\sum_{j=0}^{\infty} \frac{A^j}{j!} t^j\right) = \sum_{j=0}^{\infty} \frac{d}{dt}\left(\frac{A^j}{j!} t^j\right)$$

Now

$$\begin{aligned}\sum_{j=0}^{\infty} \frac{d}{dt} \left(\frac{A^j t^j}{j!} \right) &= \sum_{j=0}^{\infty} \frac{A^j j t^{j-1}}{j!} = \sum_{j=1}^{\infty} \frac{A^j j t^{j-1}}{j!} = \sum_{j=1}^{\infty} \frac{A^j t^{j-1}}{(j-1)!} \\ &= A(I + tA + \frac{t^2}{2}A^2 + \cdots + \frac{t^n}{n!}A^n + \cdots) = Ae^{At}\end{aligned}$$

where it is straightforward to prove the we can factor the A out of the series by a partial sum argument which we leave to you. So we get that $\frac{d}{dt}(e^{At}) = Ae^{At}$. ■

Homework

Exercise 19.4.1

Exercise 19.4.2

Exercise 19.4.3

Exercise 19.4.4

Exercise 19.4.5

19.4.1 The Homogeneous Solution

Theorem 19.4.2 The general solution to a linear ODE system

Let A be a $n \times n$ matrix. Then $\Phi(t) = e^{At} \cdot C$ is a solution to $\mathbf{x}' = A\mathbf{x}$.

Proof 19.4.2

We note

$$\Phi'(t) = \frac{d}{dt}(e^{At} \cdot C)$$

and if we write e^{At} using its column vectors

$$e^{At} = [\Phi_1(t), \dots, \Phi_n(t)]$$

then

$$\begin{aligned}\frac{d}{dt}(e^{At} \cdot C) &= \frac{d}{dt} \left(\begin{bmatrix} \Phi_1(t) & \cdots & \Phi_n(t) \end{bmatrix} \begin{bmatrix} C_1 \\ \vdots \\ C_n \end{bmatrix} \right) \\ &= \frac{d}{dt} \left(\sum_{i=1}^n \Phi_i(t) C_i \right) = \sum_{i=1}^n \frac{d}{dt} (\Phi_i(t) C_i) = \sum_{i=1}^n C_i \Phi'_i(t) \\ &= \begin{bmatrix} \Phi'_1(t) & \cdots & \Phi'_n(t) \end{bmatrix} \begin{bmatrix} C_1 \\ \vdots \\ C_n \end{bmatrix} \\ &= \frac{d}{dt}(e^{At}) \cdot C = Ae^{At} \cdot C = A\Phi(t)\end{aligned}$$

This shows Φ is a solution to the system. ■

We also know that

$$e^{At} = Pe^{Jt}P^{-1}C$$

where J is the Jordan canonical form of A . For the structure of the Jordan Canonical Form of A , you can see its columns are linearly independent functions. Thus, letting the columns of e^{Jt} be $\Psi(t)$, we see

$$[\Phi_1 \ \dots \ \Phi_n] = P [\Psi_1 \ \dots \ \Psi_n] P^{-1}$$

To see the columns of e^{At} are linearly independent, consider the usual linear dependence equation

$$\alpha_1\Phi_1 + \dots + \alpha_n\Phi_n = \mathbf{0} \implies [\Phi_1 \ \dots \ \Phi_n] \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \mathbf{0}$$

This implies

$$P [\Psi_1 \ \dots \ \Psi_n] P^{-1} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \mathbf{0}$$

or

$$[\Psi_1 \ \dots \ \Psi_n] P^{-1} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = P^{-1}\mathbf{0} = \mathbf{0}$$

Since e^{Jt} has independent columns, the only solution here is

$$P^{-1} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \mathbf{0} \implies \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \mathbf{0}$$

We conclude the functions Φ are independent. Since $\Phi'_i = A\Phi_i$, we also know each Φ is a solution to $\mathbf{x}' = A\mathbf{x}$. The solutions $\Phi = e^{At}C$ for arbitrary $C \in \mathbb{R}^n$ are the span of the linearly independent functions Φ_1 to Φ_n . So the functions Φ_1 to Φ_n form a basis for the set of solutions to $\mathbf{x}' = A\mathbf{x}$.

When we search for solutions to $\mathbf{x}' = A\mathbf{x}$, we look for solutions of the form $\mathbf{x} = Ve^{rt}$ and we find the only scalars r that work are the eigenvalues of A . Of course, if a root is repeated the eigenspace for that eigenvalue need not have the same geometric multiplicity as its algebraic multiplicity. The structure of the Jordan Canonical block here shows us what happens. Any solution satisfies

$$\mathbf{x}' = PJP^{-1}\mathbf{x}$$

and letting $\mathbf{y} = P^{-1}\mathbf{x}$, we see we are searching for solutions to $\mathbf{y}' = J\mathbf{y}$. Now focus on one JCB, say J_i . Then setting all variables in \mathbf{y}_i to be zero except for the ones that concern J_i , we look for solutions

$$\mathbf{y}'_i = J_i \mathbf{y}_i$$

and we know we either get $e^{\lambda_i t}$ solutions with corresponding eigenvectors or if the eigenspace is too

small we also get solutions such as $t^k e^{\lambda_i t}$ for appropriate k . So the only solutions we can get to this system are linear combinations of the columns of e^{At} . Hence, any solution is a linear combination of the columns of e^{At} . We conclude the solution space of $\mathbf{x}' = A\mathbf{x}$ is the n dimensional vector space with basis $\{\Phi_1, \dots, \Phi_{n-1}\}$ and any function in the span of this basis can be written as $e^{At}C$ for some $C \in \mathbb{R}^n$. Let's summarize:

Theorem 19.4.3 The Solution Space to a linear System of ODE

Let $\mathcal{S} = \{\mathbf{y} \mid \mathbf{y}' = A\mathbf{y}\}$ Then

1. \mathcal{S} is an n -dimensional vector space and any basis for \mathcal{S} consists of n linearly independent solutions to $\mathbf{x}' = A\mathbf{x}$.
2. The columns of e^{At} give a basis for \mathcal{S} .
3. e^{At} is invertible and we write its inverse as e^{-At} .
4. $\Phi(t) = e^{A(t-t_0)}\mathbf{x}_0$ is the solution to $\mathbf{x}' = A\mathbf{x}$, $\mathbf{x}(t_0) = \mathbf{x}_0$.

Proof 19.4.3

Most of this has already been proven. Note the columns of e^{At} are linearly independent, it does have an inverse. Thus

$$e^{(A-A)t} = I \implies e^A e^{-A} = I \implies e^{-A} = (e^A)^{-1}$$

Next, note $\mathbf{x}' = A\mathbf{x}$, $\mathbf{x}(t_0) = \mathbf{x}_0$ has general solution $\Phi(t) = e^{At}C$. Using the initial condition we can find

$$\Phi(t_0) = \mathbf{x}_0 = e^{At_0}C \implies C = e^{-At_0}\mathbf{x}_0$$

So

$$\Phi(t) = e^{At} (e^{-At_0}) \mathbf{x}_0 = e^{A(t-t_0)}\mathbf{x}_0$$

■

19.4.2 Homework

Exercise 19.4.6

Exercise 19.4.7

Exercise 19.4.8

Exercise 19.4.9

Exercise 19.4.10

19.5 The Non homogeneous Solution

Now let's look at what are called non homogeneous systems. Now consider the nonhomogeneous system

$$\mathbf{x}' = A\mathbf{x} + b(t)$$

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

We say Ψ is a particular solution if we have

$$\Psi' = A\Psi + b(t)$$

These are hard to find in practice, but we can find a general formula that is a good start. If we have a particular solution in hand, the general solution is then

$$\Phi(t) = e^{A(t-t_0)}C + \Psi(t)$$

Theorem 19.5.1 A Particular Solution to a linear ODE System

Let b be a continuous function. Then $\Psi(t) = e^{At} \int_{t_0}^t e^{-As} b(s)ds$ is a particular solution to

$$\mathbf{x}' = A\mathbf{x} + b(t)$$

Here, integration is defined on each component in the usual way. Hence, the solution to the system is

$$\Phi(t) = e^{A(t-t_0)}\mathbf{x}_0 + \int_{t_0}^t e^{-A(t-s)}b(s)ds$$

Proof 19.5.1

This is a straightforward calculation. The general solution to the nonhomogeneous system should be

$$\Phi(t) = e^{At}C + e^{At} \int_{t_0}^t e^{-As}b(s)ds$$

Using the initial condition we get that the actual solution to the above nonhomogeneous system is

$$\Phi(t) = e^{A(t-t_0)}\mathbf{x}_0 + \int_{t_0}^t e^{-A(t-s)}b(s)ds$$

where the second part holds since

$$e^{At} \int_{t_0}^t e^{-As}b(s)ds = \int_{t_0}^t e^{At}e^{-As}b(s)ds = \int_{t_0}^t e^{A(t-s)}b(s)ds$$

Note,

$$\begin{aligned} \Psi' &= Ae^{At} \int_{t_0}^t e^{-As}b(s)ds + e^{At} \left(\int_{t_0}^t e^{-As}b(s)ds \right)' \\ &= Ae^{At} \int_{t_0}^t e^{-As}b(s)ds + e^{At} (e^{-At}b(t)) \end{aligned}$$

as the Fundamental Theorem of Calculus holds for these vectors functions which is easily seen by looking at components. Therefore

$$\Psi' = Ae^{At} \int_{t_0}^t e^{-As}b(s)ds + b(t) = A\Psi + b(t)$$

The solution to the system is thus

$$\Phi(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^t e^{-A(t-s)}b(s)ds$$

■

Comment 19.5.1 We see then that if we were able to find e^{At} we would then be able to solve many of these linear systems!

19.5.1 Homework

Exercise 19.5.1

Exercise 19.5.2

Exercise 19.5.3

Exercise 19.5.4

Exercise 19.5.5

19.6 A Diagonalizable Test Problem

Take

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

Let's find e^{At} .

1. Find eigenvalues:

$$\begin{vmatrix} -r & 1 \\ -2 & 3-r \end{vmatrix} = 3r + r^2 - 2 = (r+2)(r+1) - 0$$

So eigenvalues are $r_1 = -2, r_2 = -1$.

2. The Cayley–Hamilton Theorem (which we have not discussed and that is covered in your advanced linear algebra course) tells us A satisfies the characteristic equation. Hence, we know $A^2 + 3A + 2I = \mathbf{0}$ So

$$\begin{aligned} A^2 &= -3A - 2I \\ A^3 &= A^2 A = -3A^2 - 2A = -3(-3A - 2I) - 2A = 7A - 6I \\ &\vdots \\ A^n &= a_n A + b_n I \end{aligned}$$

Now

$$\begin{aligned} e^A &= I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots + \frac{A^n}{n!} + \cdots \\ &= I + A + \left(\frac{a_2}{2!} A + \frac{b_2}{2!} I \right) + \cdots + \left(\frac{a_n}{n!} A + \frac{b_n}{n!} I \right) + \cdots \end{aligned}$$

$$= \left(1 + \frac{b_2}{2!} + \frac{b_3}{3!} + \dots \right) I + \left(1 + \frac{a_2}{2!} + \frac{a_3}{3!} + \dots \right) A$$

and we know that both of these sequences in parenthesis converge because e^A converges. So $e^A = \alpha I + \beta A$ for some α and β . Then e^{At} can be written as $e^{At} = \alpha(t)I + \beta(t)A$. Next,

$$\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} = P \overbrace{\begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}}^J P^{-1}$$

From before we know that $e^{At} = Pe^{Jt}P^{-1}$, so

$$\begin{aligned} Pe^{Jt}P^{-1} &= P\alpha(t)IP^{-1} + \beta(t)PJP^{-1} \\ P \begin{bmatrix} e^{-2t} & 0 \\ 0 & e^{-t} \end{bmatrix} P^{-1} &= P \left[\alpha(t)I + \beta(t) \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix} \right] P^{-1} \end{aligned}$$

i.e.

$$\begin{aligned} e^{-2t} &= \alpha(t) - 2\beta(t) & (1) \\ e^{-t} &= \alpha(t) - \beta(t) & (2) \end{aligned}$$

Taking the negative of (2) we get

$$\begin{aligned} e^{-2t} &= \alpha(t) - 2\beta(t) \\ -e^{-t} &= -\alpha(t) + \beta(t) \end{aligned}$$

Adding these we get

$$e^{-2t} - e^{-t} = -\beta(t) \quad \text{or} \quad \beta(t) = e^{-t} - e^{-2t}$$

So

$$\begin{aligned} \alpha(t) &= \beta(t) + e^{-t} \\ &= 2e^{-t} - e^{-2t} \end{aligned}$$

Therefore

$$\begin{aligned} e^{At} &= (2e^{-t} - e^{-2t}) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + (e^{-t} - e^{-2t}) \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix} \\ &= \begin{pmatrix} 2e^{-t} - e^{-2t} & e^{-t} - e^{-2t} \\ -2e^{-t} + 2e^{-2t} & -e^{-t} + 2e^{-2t} \end{pmatrix} \end{aligned}$$

and we know that each of the columns in e^{At} should be solutions to the linear system of equations associated with the matrix A .

Solve the system $\mathbf{x}' = A\mathbf{x}$ Now, as mentioned before, both columns of e^{At} should be solutions to the above system. Let's check it for

$$\Phi_1 = \begin{pmatrix} 2e^{-t} - e^{-2t} \\ -2e^{-t} + 2e^{-2t} \end{pmatrix}$$

19.7. A NON DIAGONALIZABLE TEST PROBLEM

401

First,

$$\Phi'_1 = \begin{pmatrix} -2e^{-t} + 2e^{-2t} \\ 2e^{-t} - 4e^{-2t} \end{pmatrix}$$

and second,

$$\begin{aligned} A\Phi_1 &= \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix} \begin{pmatrix} 2e^{-t} - e^{-2t} \\ -2e^{-t} + 2e^{-2t} \end{pmatrix} \\ &= \begin{pmatrix} -2e^{-t} + 2e^{-2t} \\ 2e^{-t} - 4e^{-2t} \end{pmatrix} \end{aligned}$$

So $\Phi'_1 = A\Phi_1$ and Φ_1 is a solution to the system. In a similar fashion it can be shown that Φ_2 is also a solution. Thus e^{At} contains the fundamental solutions to the system.

19.6.1 Homework

Exercise 19.6.1

Exercise 19.6.2

Exercise 19.6.3

Exercise 19.6.4

Exercise 19.6.5

19.7 A Non diagonalizable Test Problem

Now try finding e^{At} using this method for a matrix A which is **not** diagonalizable, and has multiple eigenvalues. Try

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

This is a Jordan block and we know that for this matrix

$$e^{At} = \begin{pmatrix} e^t & te^t \\ 0 & e^t \end{pmatrix}$$

Now let's use the method above to find e^{At} .

1. Find eigenvalues: We know they are $r_1 = 1, r_2 = 1$.
2. Use Cayley–Hamilton Theorem:

$$\begin{aligned} (A - I)(A - I) &= 0 \\ A^2 - 2A + I &= 0 \\ A^2 &= 2A - I \end{aligned}$$

3. $e^{At} = \alpha(t)I + \beta(t)A$

$$Pe^{Jt}P^{-1} = P[\alpha(t)I + \beta(t)J]P^{-1} \quad (\text{Here } A=J)$$

$$\begin{aligned} Pe^{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}t}P^{-1} &= P \left[\alpha(t)I + \beta(t) \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \right] P^{-1} \begin{bmatrix} e^t & te^t \\ 0 & e^t \end{bmatrix} \\ &= \begin{bmatrix} \alpha(t) & 0 \\ 0 & \alpha(t) \end{bmatrix} \begin{bmatrix} \beta(t) & \beta(t) \\ 0 & \beta(t) \end{bmatrix} \end{aligned}$$

So,

$$e^t = \alpha(t) + \beta(t) \quad (1)$$

$$te^t = \beta(t) \quad (2)$$

$$e^t = \alpha(t) + \beta(t) \quad (3)$$

From (2) we get $\beta(t) = te^t$, and using (2), (1) gives us $\alpha(t) = e^t - te^t$. Therefore,

$$\begin{aligned} e^{At} &= (1-t)e^t I + te^t A \\ &= (1-t)e^t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + te^t \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} (1-t)e^t & 0 \\ 0 & (1-t)e^t \end{pmatrix} + \begin{pmatrix} te^t & te^t \\ 0 & te^t \end{pmatrix} \\ &= \begin{pmatrix} e^t & te^t \\ 0 & e^t \end{pmatrix} \end{aligned}$$

and this is what we expected. So this method works for matrices having repeated eigenvalues and for matrices that are not diagonalizable. Again, note that we can find e^{At} and we never have to find P or P^{-1} .

19.7.1 Homework

Exercise 19.7.1

Exercise 19.7.2

Exercise 19.7.3

Exercise 19.7.4

Exercise 19.7.5

19.8 A Non homogeneous Problem

Let's look at a typical nonhomogeneous system to show you the kinds of calculations you have to do to solve a problem like this. It is pretty intense! Consider the non homogeneous problem

$$\mathbf{x}' = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \mathbf{x} + \begin{bmatrix} t \\ \cos(t) \end{bmatrix}$$

and we want $x_1(t_0) = 1$ and $x_2(t_0) = -1$. We have already shown that $P(t) = \int_{t_0}^t e^{A(t-s)} b(s) ds$ is a particular solution to the nonhomogeneous system. Instead of using e^{At} , let's use $\Phi(t)$ to get $Q(t) = \Phi(t) \int_{t_0}^t \Phi^{-1}(s) b(s) ds$ as a particular solution. The check that it is a particular solution we

19.8. A NON HOMOGENEOUS PROBLEM

403

know

$$Q'(t) = \Phi'(t) \int_{t_0}^t \Phi^{-1}(s) b(s) ds + \Phi(t) \Phi^{-1}(t) b(t)$$

where

$$\Phi' = [\Phi'_1 \quad \Phi'_2] = [A\Phi_1 \quad A\Phi_2] = A[\Phi_1 \quad \Phi_2] = A\Phi$$

since Φ_1 and Φ_2 are solutions to the homogeneous system. So then

$$\begin{aligned} Q' &= A\Phi \int_{t_0}^t \Phi^{-1}(s) b(s) ds + b(t) \\ &= AQ + b(t) \end{aligned}$$

and we see that $Q(t)$ satisfies the nonhomogeneous system and is therefore a particular solution. So we see that we don't need to use e^{At} ; all we need is a matrix containing two linearly independent solutions. (Here we used $\Phi(t)$).

Homework

Exercise 19.8.1

Exercise 19.8.2

Exercise 19.8.3

Exercise 19.8.4

Exercise 19.8.5

19.8.1 Louiville's Formula

Now we know that $Q(t) = \Phi(t) \int_{t_0}^t \Phi'(s) b(s) ds$ is also a particular solution where

$$\Phi(t) = \left[\begin{bmatrix} 1 \\ 2 \end{bmatrix} e^{5t} \quad \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-t} \right] = [\mathbf{V}_1 e^{5t} \quad \mathbf{V}_2 e^{-t}]$$

where \mathbf{V}_1 and \mathbf{V}_2 are the corresponding eigenvectors. A quick calculation shows

$$\det \Phi(t) = \det [\mathbf{V}_1 e^{5t} \quad \mathbf{V}_2 e^{-t}] = \det [\mathbf{V}_1 \quad \mathbf{V}_2] e^{4t}$$

where the $\det [\mathbf{V}_1 \quad \mathbf{V}_2]$ is nonzero because the eigenvectors are linearly independent. It is well known the inverse is then

$$\Phi^{-1}(t) = \frac{1}{\det [\mathbf{V}_1 \quad \mathbf{V}_2]} \begin{bmatrix} V_{22}e^{-t} & -V_{12}e^{5t} \\ -V_{12}e^{5t} & V_{11}e^{5t} \end{bmatrix}$$

Now the trace of a matrix is the sum of its diagonal elements and for the matrix B is denoted by $\text{tr}(B)$. Here, $\text{tr}(A) = 4$ and $\det \Phi(0) = \det [\mathbf{V}_1 \quad \mathbf{V}_2]$. Thus, this formula

$$\det \Phi(t) = \det \Phi(0) e^{\int_0^t \text{tr}(A) ds} = \det [\mathbf{V}_1 \quad \mathbf{V}_2] e^{4t}$$

which we have already found. However, if these matrices were not 2×2 we can define the submatrices

$$\Phi_{ij} = (-1)^{i+j} \det \left(\begin{array}{c} \text{submatrix of } \Phi \text{ obtained by} \\ \text{deleting row } i \text{ and column } j \end{array} \right)$$

and form the **adjoint** of Φ $\text{adj}(\Phi)(t) = (\Phi_{ij})^T(t)$. From advanced linear algebra, we find the inverse of $\Phi(t)$ is

$$(\Phi(t))^{-1} = \frac{1}{\det \Phi} \text{adj}(\Phi)$$

with

$$\det \Phi(t) = \det \Phi(t_0) e^{\int_{t_0}^t \text{tr}(A) ds}$$

Now for $t_0 = 0$ we get the determinant we already found which is $\det \Phi(t) = -3e^{4t}$. This formula is called **Louiville’s Formula** and it holds for linear systems.

Homework

Exercise 19.8.6

Exercise 19.8.7

Exercise 19.8.8

Exercise 19.8.9

Exercise 19.8.10

19.8.2 Finding the Particular Solution

How do we find $\text{adj} \Phi$? Using the formulas given above we get that

$$\text{adj } \Phi = \begin{bmatrix} -e^{-t} & -2e^{5t} \\ -e^{-t} & e^{5t} \end{bmatrix}^T$$

which is what we found earlier. So then

$$\Phi^{-1}(t) = \frac{\begin{bmatrix} -e^{-t} & -2e^{5t} \\ -e^{-t} & e^{5t} \end{bmatrix}^T}{-3e^{4t}} = \begin{bmatrix} \frac{1}{3}e^{-5t} & \frac{1}{3}e^{-5t} \\ \frac{2}{3}e^t & -\frac{1}{3}e^t \end{bmatrix}$$

Now, if the inverse was calculated correctly then $\Phi \Phi^{-1} = I$. To verify this we see that

$$\begin{bmatrix} e^{5t} & e^{-t} \\ 2e^{5t} & -e^{-t} \end{bmatrix} \begin{bmatrix} \frac{1}{3}e^{-5t} & \frac{1}{3}e^{-5t} \\ \frac{2}{3}e^t & -\frac{1}{3}e^t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

So now our particular solution looks like

$$\begin{aligned} Q(t) &= \begin{bmatrix} e^{5t} & e^{-t} \\ 2e^{5t} & -e^{-t} \end{bmatrix} \int_0^t \begin{bmatrix} \frac{1}{3}e^{-5s} & \frac{1}{3}e^{-5s} \\ \frac{2}{3}e^s & -\frac{1}{3}e^s \end{bmatrix} \begin{bmatrix} s \\ \cos(s) \end{bmatrix} ds \\ &= \begin{bmatrix} e^{5t} & e^{-t} \\ 2e^{5t} & -e^{-t} \end{bmatrix} \begin{bmatrix} \int_0^t \left[\frac{1}{3}e^{-5s} + \frac{1}{3}\cos(s)e^{-5s} \right] ds \\ \int_0^t \left[\frac{2}{3}e^s - \frac{1}{3}\cos(s)e^s \right] ds \end{bmatrix} \end{aligned}$$

and in evaluating these integrals we will get the particular solution for our example. With this then, the general solution to the nonhomogeneous system is

$$\Phi(t) = c_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} e^{5t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-t} + Q(t)$$

where c_1 and c_2 can be found using the initial conditions given. As you can see this is a very intense process!

Homework

Exercise 19.8.11

Exercise 19.8.12

Exercise 19.8.13

Exercise 19.8.14

Exercise 19.8.15

Chapter 20

Nonlinear Parametric Optimization Theory

Let's look at some optimization problems to see how the material we have been working through applies. First, here is a very relaxed way to think about a classical optimization problem. Consider the task of

$$\begin{aligned} & \min \int_0^t f_o(s, x(s), x'(s), \mu(s)) \text{ subject to} \\ & x'(t) = f(t, x(t), \mu(t)), \quad 0 \leq t \leq T \\ & x(0) = x_0 \end{aligned}$$

As stated, it is not very clear what is going on. We are supposed to find a minimum value of an integral, so there should be some conditions of the integrand f_o to ensure $f_o(s, x(s), \dot{x}(s), \mu(s))$ is an integrable function. This problem probably makes sense even if the arguments of f_o are vectors too. We clearly need to restrict our attention to some class of interesting functions. For example, we could rephrase the problem as

$$\begin{aligned} & \min_{x, x', \mu \in C^1([0, T])} \int_0^t f_o(s, x(s), x'(s), \mu(s)) \text{ subject to} \\ & x'(t) = f(t, x(t), \mu(t)), \quad 0 \leq t \leq T \\ & x(0) = x_0 \end{aligned}$$

This still does not say anything about f_o . A more careful version would be Let's assume $f_o : [0, t] \times D \subset \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous on D with continuous partials on D . The minimization problem is

$$\begin{aligned} & \min_{\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{\mu} \in C^1([0, T])} \int_0^t f_o(s, \boldsymbol{x}(s), \boldsymbol{x}'(s), \boldsymbol{\mu}(s)) \text{ subject to} \\ & \boldsymbol{x}'(t) = f(t, \boldsymbol{x}(t), \boldsymbol{\mu}(t)), \quad 0 \leq t \leq T \\ & \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{aligned}$$

Here is a specific one.

Example 20.0.1

$$\min \int_0^1 \underbrace{[x^2(s) + (x'(s))^2]}_{\text{like energy}} + \underbrace{\mu^2(s)}_{\text{control cost}} ds$$

f₀—not dependent on time

subject to

$$\begin{aligned} x'(t) &= x(t) + \mu(t) + \underbrace{e^{-t}}_{\text{damping term}} & 0 \leq t \leq T \\ x(0) &= 1 \end{aligned}$$

We call x is the state variable and μ the control variable. It's possible that we could have a control constraint which bounds the control. So we could have something like $-1 \leq \mu(t) \leq 1$. We could also have a state constraint which says that $x(t)$ lives in a certain set; i.e. $x(t) \in C$ for some set C . Hence, these problems can be pretty hard.

20.1 The More Precise and Careful Way

Let T be a fixed real number and define some underlying spaces.

- The **control space** is $\mathcal{U} = \{\mu : [0, T] \rightarrow \mathbb{R}^m \mid \mu \text{ has some properties}\}$. These properties might be the control functions are continuous, piecewise continuous (meaning they can have a finite number of discontinuities), piecewise differentiable (meaning they can have a finite number of points where the right and left hand derivatives exists but don't match) and so forth. Each function in the control space is called a **control**.
- The **State Space** is $X = \{x : [0, T] \rightarrow \mathbb{R}^n \mid x' \text{ continuous on } [0, T]\}$. This is the same as $C^1([0, T])$. Each function in the state space is called a **state**.

Let $f_0 : [0, T] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ and assume f_0 is continuous on $[0, T] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$ i.e., $\forall \epsilon > 0, \exists \delta > 0$ so that

$$\|(t, u, v, w) - (t_0, u_0, v_0, w_0)\| < \delta \implies |f_0(t, u, v, w) - f_0(t_0, u_0, v_0, w_0)| < \epsilon$$

The norm $\|\cdot\|$ here is

$$\|(t, u, v, w) - (t_0, u_0, v_0, w_0)\| = \sqrt{|t - t_0|^2 + \|u - u_0\|_{\mathbb{R}^n}^2 + \|v - v_0\|_{\mathbb{R}^n}^2 + \|w - w_0\|_{\mathbb{R}^m}^2}$$

This means that if $x(t)$, $x'(t)$ and $\mu(t)$ are continuous functions on $[0, T]$, then since the composition of continuous functions is continuous, $f(t, x(t), x'(t), \mu(t))$ will be continuous on $[0, T]$. Let's choose $\mathcal{U} = PC[0, T]$, the set of piecewise continuous functions on $[0, T]$. We know that even if $\mu(t)$ is a piecewise continuous function the above will still hold except at the places where the control is not continuous because there is only a finite number of discontinuities. Hence, f_0 will also be piecewise continuous. Now the ODE is

$$\begin{bmatrix} x'_1 \\ \vdots \\ x'_n \end{bmatrix} = \begin{bmatrix} f_1(t, x(t), \mu(t)) \\ \vdots \\ f_n(t, x(t), \mu(t)) \end{bmatrix}, \quad 0 \leq t \leq T, \quad \begin{bmatrix} x_1(0) \\ \vdots \\ x_n(0) \end{bmatrix} = \begin{bmatrix} x_{01} \\ \vdots \\ x_{0n} \end{bmatrix}$$

or

$$\mathbf{x}' = f(t, \mathbf{x}(t), \boldsymbol{\mu}(t)), \quad 0 \leq t \leq T, \quad \mathbf{x}(0) = \mathbf{x}_0$$

where $f : [0, T] \times \Re^n \times \Re^m \rightarrow \Re$ and f is continuous on $[0, T] \times \Re^n \times \Re^m$.

Now, how do we describe the minimization? Let

$$\hat{J}(\mathbf{x}_0, \boldsymbol{\mu}) = \int_0^T f_0(s, \mathbf{x}(s), \mathbf{x}'(s), \boldsymbol{\mu}(s)) ds$$

where μ is a specific control function and \mathbf{x} is the solution to the ODE for that control function choice. Hence, if there is such a solution and the solution is at least piecewise continuous, then for the piecewise continuous control $\boldsymbol{\mu}$, the integrand is piecewise continuous and so the integral exists. But perhaps the ODE does not have a continuous solution for a given control $\boldsymbol{\mu}$ which in other courses we would find would mean the solution \mathbf{x} would have $\|\mathbf{x}\| \rightarrow \infty$. In that case, the integral could go unbounded. Hence, in general, $\hat{J} : \Re^n \times \mathcal{U} \rightarrow \Re \cup \{\pm\infty\}$. We want to find

$$J(x_0) = \inf_{\boldsymbol{\mu} \in \mathcal{U}} \hat{J}(\mathbf{x}_0, \boldsymbol{\mu})$$

This function is called the **Optimal Value Function** and understanding its properties is very important in many application areas. For example, a really good question would be **how does the optimal value function vary when the initial data x_0 changes?** So our full optimization problem is

$$\begin{aligned} & \inf_{\boldsymbol{\mu} \in \mathcal{U}} \hat{J}(\mathbf{x}_0, \boldsymbol{\mu}), \text{ subject to} \\ & \mathbf{x}' = f(t, \mathbf{x}(t), \boldsymbol{\mu}(t)), \quad 0 \leq t \leq T \\ & \mathbf{x}(0) = \mathbf{x}_0 \end{aligned}$$

or

$$\begin{aligned} & \inf_{\boldsymbol{\mu} \in \mathcal{U}} \int_0^T f_0(s, \mathbf{x}(s), \mathbf{x}'(s), \boldsymbol{\mu}(s)) ds, \text{ subject to} \\ & \mathbf{x}' = f(t, \mathbf{x}(t), \boldsymbol{\mu}(t)), \quad 0 \leq t \leq T \\ & \mathbf{x}(0) = \mathbf{x}_0 \end{aligned}$$

The most general way to write this is: Find

$$J(t_0, x_0) = \inf_{\boldsymbol{\mu} \in \mathcal{U}} \hat{J}(t_0, x_0, \boldsymbol{\mu})$$

where

$$\hat{J}(t_0, x_0, \boldsymbol{\mu}) = \ell(t_0, \mathbf{x}_0) + \int_0^T f_0(s, \mathbf{x}(s), \mathbf{x}'(s), \boldsymbol{\mu}(s)) ds$$

subject to

$$\begin{aligned} \mathbf{x}' &= f(t, \mathbf{x}(t), \boldsymbol{\mu}(t)), \quad 0 \leq t \leq T \\ \mathbf{x}(t_0) &= \mathbf{x}_0 \\ \mathbf{x}(t) &\in \Omega(t, \boldsymbol{\mu}(t)) \\ \boldsymbol{\mu}(t) &\in \Lambda(t, \mathbf{x}(t)) \end{aligned}$$

with $\ell(t_0, \mathbf{x}_0)$ a penalty on initial conditions. For example, this could be some random noise added from some probability distribution. The set Ω is a constraint set on the values the state can take at time t given the value of the control at the time. The set Λ is a constraint set on the values the control is permitted to have at time t given the state value at that time. As you may imagine, adding state and control constraints makes this problem much harder to find solutions to. We would like to see that $J(t_0, \mathbf{x}_0)$ is a finite number. So we want to find a control which actually allows us to achieve the **infimum**. This is a very hard problem both theoretically and practically.

Example 20.1.1 A Constrained Optimal Control Problem

Consider this problem which has constraints on the control.

$$\inf \int_0^1 \left[\frac{1}{2} \mathbf{x}^T Q \mathbf{x} + \frac{1}{2} (\mathbf{x}')^T R \mathbf{x}' + \frac{1}{2} \boldsymbol{\mu}^T S \boldsymbol{\mu} \right] ds$$

Subject to:

$$\begin{aligned} \mathbf{x}' &= A\mathbf{x} + B\boldsymbol{\mu} \\ -1 &\leq \boldsymbol{\mu} \leq 1 \\ \mathbf{x}(t_0) &= \mathbf{x}_0 \end{aligned}$$

The idea here is to take a control from any space that we want. Then we solve the differential equation, hopefully obtaining a solution, and then use this solution along with the control to find a value for the integral. Our hope is that there is a control which gives us a smallest value for the integral. For the above problem the solution to the differential equation is given by

$$\begin{aligned} \mathbf{x}(t) &= e^{A(t-t_0)} \mathbf{x}_0 + e^{At} \int_{t_0}^t e^{-As} B \boldsymbol{\mu}(s) ds \\ &= e^{A(t-t_0)} \mathbf{x}_0 + \int_{t_0}^t e^{A(t-s)} B \boldsymbol{\mu}(s) ds. \end{aligned}$$

Now how big is $\mathbf{x}(t)$?

$$\|\mathbf{x}(t)\| \leq \|e^{A(t-t_0)}\| \|\mathbf{x}_0\| + \int_{t_0}^t \|e^{A(t-s)}\| \|B\| \|\boldsymbol{\mu}(s)\| ds$$

We get a nice bound on the first term if all of the eigenvalues have negative real part. We know $e^{At} = Pe^{Jt}P^{-1}$ so the norm of the exponential matrix is determined by the structure of the Jordan Canonical Blocks. Any eigenvalue which is positive or has positive real part will give us a norm which is unbounded over long time scales. We encourage you to work this out. Hence, we value problems where we can guarantee the eigenvalues of A have negative real part. And, provided our control is bounded, we can bound the second term. So we can get a bound on $\mathbf{x}(t)$ in these cases. We get a bound even if the eigenvalues have positive real part but then the norm of the solution can get arbitrary large which is not what we usually want. Now, in general, we can write the solution as

$$\mathbf{x}(t) = \Phi(t)\Phi'(t_0)x_0 + \Phi(t) \int_{t_0}^t \Phi^{-1}(s)B\boldsymbol{\mu}(s)ds$$

where $\Phi(t)$ is the fundamental matrix composed of linearly independent solutions. For nonlinear problems this is very difficult to find and it is here that the problem lies.

Comment 20.1.1 This control problem is called a **LQR** for the **Linear Quadratic regulator** problem. Often there are state variables \mathbf{x} and variables \mathbf{y} which are actually observed, called **observed variables**. This simply adds an extra constraint of the form $\mathbf{y} = C\mathbf{x}$.

20.2 Unconstrained Parametric Optimization

Given $L : \mathbb{R}^n \rightarrow \mathbb{R}$, the problem is to find $\inf_{x \in \mathbb{R}^n} L(x)$. We have already discussed such extremum problems, so we will be brief here. Under sufficient smoothness conditions, if x_0 is a local minimum, then

$$\nabla L(\mathbf{x}_0) = \begin{bmatrix} \frac{\partial L}{\partial x_1}(\mathbf{x}_0) \\ \vdots \\ \frac{\partial L}{\partial x_n}(\mathbf{x}_0) \end{bmatrix} = \mathbf{0}$$

The directional derivative of L in the direction of the unit vector \mathbf{u} from a point \mathbf{x} is

$$D_u L(\mathbf{x}) = \langle \nabla L(\mathbf{x}), \mathbf{u} \rangle$$

$$= (\nabla L(\mathbf{x}))^T \mathbf{u} = \begin{bmatrix} \frac{\partial L}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial L}{\partial x_n}(\mathbf{x}) \end{bmatrix}^T \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

where $u_1^2 + \cdots + u_n^2 = 1$. In two dimensions, \mathbb{R}^2 , assume L has a local min at $\vec{x}_0 = \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}$, i.e. $\exists r > 0$ such that if $x \in B(x_0; r)$, then $L(x) \geq L(x_0)$; a generic illustration of this is shown in Figure 20.1. This looks like a paraboloid locally. In fact, if we look at the cross-section of this

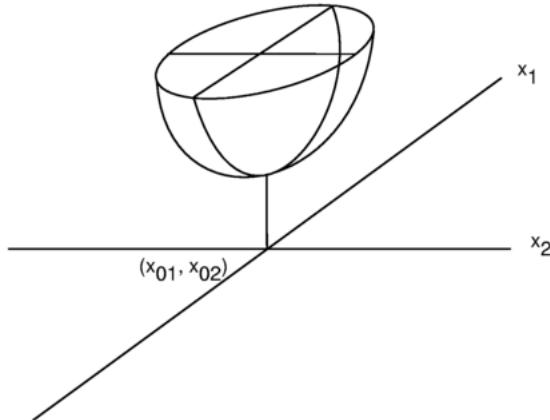


Figure 20.1: Two Dimensional Local Minimum

surface through the line L_θ , $0 \leq \theta \leq 2\pi$, we obtain the two-dimensional slice shown in Figure 20.2. Now restrict your attention to the trace in the L surface above this line lying inside the cylinder of radius $\frac{r}{2}$ about x_0 as shown in Figure 20.3. Then for a given θ , we obtain the parameterized curve given by $x_3(t) = L(x_01 + t \cos(\theta), x_02 + t \sin(\theta))$ for $-\frac{r}{2} < t < \frac{r}{2}$. This is illustrated in Figure 20.4.

Now $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, so

$$\frac{dx_3}{dt} = \left\langle \nabla L(\mathbf{x}), \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \right\rangle = \frac{\partial L}{\partial x_1}(\mathbf{x}) \cos \theta + \frac{\partial L}{\partial x_2}(\mathbf{x}) \sin \theta.$$

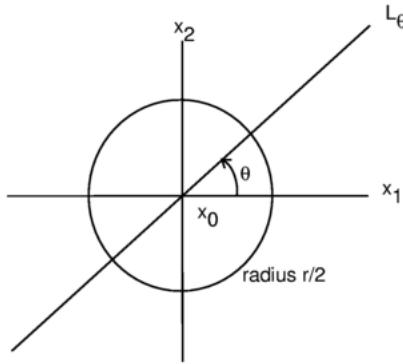


Figure 20.2: L_θ Slice

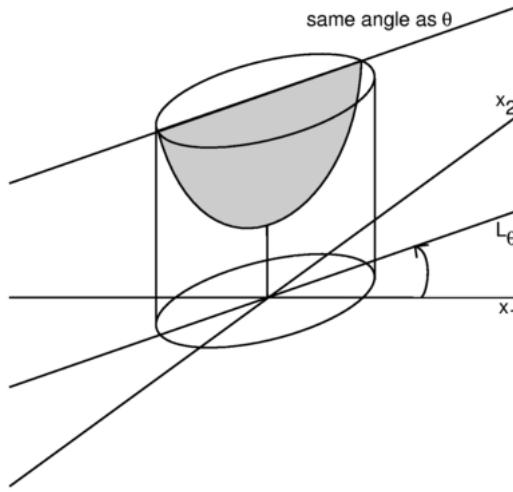


Figure 20.3: Local Cylinder

Then

$$\begin{aligned}\frac{d^2x_3}{dt^2} &= \cos \theta L_{x_1,x_1}(\mathbf{x}) \frac{\partial x_1}{\partial t} + \cos \theta L_{x_1,x_2}(\mathbf{x}) \frac{\partial x_2}{\partial t} + \sin \theta L_{x_2,x_1}(\mathbf{x}) \frac{\partial x_1}{\partial t} + \sin \theta L_{x_2,x_2}(\mathbf{x}) \frac{\partial x_2}{\partial t} \\ &= \cos^2 \theta L_{x_1,x_1}(\mathbf{x}) + 2 \sin \theta \cos \theta L_{x_1,x_2}(\mathbf{x}) + \sin^2 \theta L_{x_2,x_2}(\mathbf{x})\end{aligned}$$

where we assume the mixed order partials match. Of course, if we assume the partials are continuous locally this is true. From this we get the general quadratic form

$$\frac{d^2x_3}{dt^2} = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}^T \mathbf{H}_L(\mathbf{x}) \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}.$$

Hence, at the local minimum \mathbf{x}_0 ,

$$\frac{d^2x_3}{dt^2} = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}^T \mathbf{H}_L(\mathbf{x}_0) \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}.$$

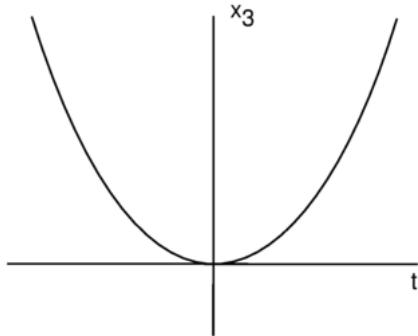


Figure 20.4: Local Cylinder Cross–Section

Since \mathbf{x}_0 is a local minimum

$$\frac{d^2x_3}{dt^2} = \mathbf{E}_\theta^T \mathbf{H}_L(\mathbf{x}_0) \mathbf{E}_\theta > 0 \quad \forall 0 < \theta < 2\pi$$

where

$$\mathbf{E}_\theta = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}$$

Any vector in \mathbb{R}^2 can be written as $r\mathbf{E}_\theta$, so we have shown

$$\mathbf{x}^T \mathbf{H}_L(\mathbf{x}_0) \mathbf{x} > 0 \quad \forall \mathbf{x} \in \mathbb{R}^2$$

i.e. $\mathbf{H}_L(\mathbf{x}_0)$ is a positive definite matrix implying it is diagonalizable and has determinant greater than zero. We note that at $\theta = 0$ we get $L_{x_1, x_1}(\mathbf{x}_0) > 0$. So for the matrix $\mathbf{H}_L(\mathbf{x}_0)$ we know at a minimum:

1. The first principle minor is $L_{x_1, x_1}(\mathbf{x}_0)$ and its determinant is positive.
2. The next principle minor is $\mathbf{H}_L(\mathbf{x}_0)$ itself and since it is positive definite symmetric we know its determinant is positive.

So at the local minimum \mathbf{x}_0 we expect this behavior

- $\nabla L(\mathbf{x}_0) = \mathbf{0}$.
- $\mathbf{H}_L(\mathbf{x}_0)$ is symmetric positive definite
- The first principle minor has positive determinant.
- The second principle minor has positive determinant.

These are all results we have already worked out, but here we used a different method of attack. For this \mathbb{R}^2 case, we have

$$\frac{d^2x_3}{dt^2} = \cos^2 \theta f_{x_1, x_1}(\mathbf{x}_0) + 2 \sin \theta \cos \theta f_{x_1, x_2}(\mathbf{x}_0) + \sin^2 \theta f_{x_2, x_2}(\mathbf{x}_0) \quad \forall 0 < \theta < 2\pi$$

Rewrite this as

$$\frac{\alpha^2}{\alpha^2 + \beta^2} f_{x_1, x_1}(\mathbf{x}_0) + \frac{2\alpha\beta}{\alpha^2 + \beta^2} f_{x_1, x_2}(\mathbf{x}_0) + \frac{\beta^2}{\alpha^2 + \beta^2} f_{x_2, x_2}(\mathbf{x}_0) > 0 \quad \forall \alpha, \beta$$

So

$$\alpha^2 f_{x_1,x_1}(\mathbf{x}_0) + 2\alpha\beta f_{x_1,x_2}(\mathbf{x}_0) + \beta^2 f_{x_2,x_2}(\mathbf{x}_0) > 0 \quad \forall \alpha, \beta$$

Now factoring out $f_{x_1,x_1}(\mathbf{x}_0)$ and completing the square we get the following:

$$f_{x_1,x_1}(\mathbf{x}_0) \left[\alpha^2 + \frac{2\beta f_{x_1,x_2}(\mathbf{x}_0)}{f_{x_1,x_1}(\mathbf{x}_0)} + \frac{\beta^2 (f_{x_1,x_1}(\mathbf{x}_0))^2}{(f_{x_1,x_1}(\mathbf{x}_0))^2} \right] - \frac{\beta^2 (f_{x_1,x_2}(\mathbf{x}_0))^2}{f_{x_1,x_1}(\mathbf{x}_0)} + \beta^2 f_{x_2,x_2}(\mathbf{x}_0) > 0$$

Simplifying,

$$f_{x_1,x_1}(\mathbf{x}_0) \left[\alpha + \frac{\beta f_{x_1,x_2}(\mathbf{x}_0)}{f_{x_1,x_1}(\mathbf{x}_0)} \right]^2 + \beta^2 \left[f_{x_2,x_2}(\mathbf{x}_0) - \frac{(f_{x_1,x_2}(\mathbf{x}_0))^2}{f_{x_1,x_1}(\mathbf{x}_0)} \right] > 0$$

Thus

$$f_{x_1,x_1}(\mathbf{x}_0) \left[\alpha + \frac{\beta f_{x_1,x_2}(\mathbf{x}_0)}{f_{x_1,x_1}(\mathbf{x}_0)} \right]^2 + \beta^2 \left[\frac{f_{x_1,x_1}(\mathbf{x}_0) f_{x_2,x_2}(\mathbf{x}_0) - (f_{x_1,x_2}(\mathbf{x}_0))^2}{f_{x_1,x_1}(\mathbf{x}_0)} \right] > 0$$

Since we know $f_{x_1,x_1}(\mathbf{x}_0) > 0$, this forces

$$f_{x_1,x_1}(\mathbf{x}_0) f_{x_2,x_2}(\mathbf{x}_0) - (f_{x_1,x_2}(\mathbf{x}_0))^2 > 0$$

This is precisely the determinant of the second minor which we expected from our theory to be positive at a minimum. You should go back and look at the argument for the \mathbb{R}^2 case presented in (Peterson (8) 2019). We used a somewhat different argument there based on Hessian approximations and had to work a little more to be able to focus on what happened at the critical point \mathbf{x}_0 . From our more general theory in Chapter 12.2, we know there are similar conditions at the minimum in \mathbb{R}^n .

20.3 Constrained Parametric Optimization

Consider the problem:

$$\inf_{\boldsymbol{\mu} \in \mathbb{R}^m} L(\mathbf{x}, \boldsymbol{\mu})$$

subject to

$$f(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{0}$$

where $f(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{0}$ is the same as saying

$$\begin{aligned} f_1(\mathbf{x}, \boldsymbol{\mu}) &= 0 \\ &\vdots \\ f_n(\mathbf{x}, \boldsymbol{\mu}) &= 0 \end{aligned}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the **state**, $\boldsymbol{\mu} \in \mathbb{R}^m$ is the **control** and $L : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is the **performance index**. The function $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ is the **constraint**. We assume the functions f and L have sufficient smoothness.

Example 20.3.1

$$\inf_{\boldsymbol{\mu} \in \mathbb{R}} x^2 + \boldsymbol{\mu}^2$$

20.3. CONSTRAINED PARAMETRIC OPTIMIZATION

415

subject to

$$x^2 = 2\mu$$

The Figures 20.5 and 20.6 show what is happening.

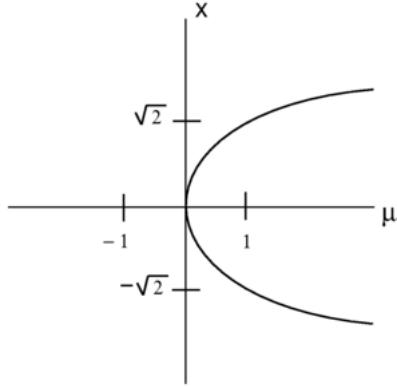


Figure 20.5: Minimize $x^2 + \mu^2$ Subject to $x^2 = 2\mu$

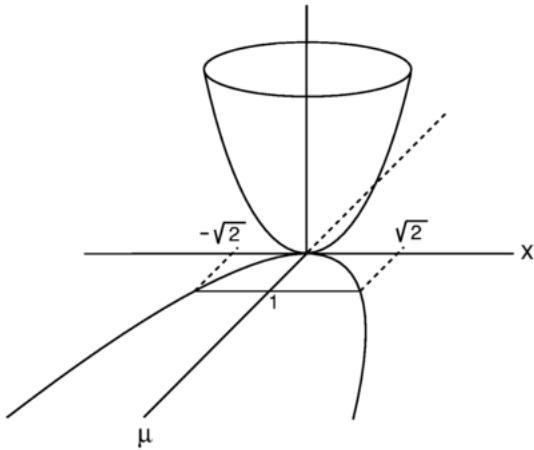


Figure 20.6: Solution to Minimize $x^2 + \mu^2$ Subject to $x^2 = 2\mu$

Here the minimum is clearly at $(0, 0)$. The idea here is to pick μ and then find an x which satisfies the given constraints. This is done until the minimum of $L(x, \mu)$ over all μ is found. Note: Here x and μ are **not** independent of one another. They may be linked in the constraints somehow.

Let's look at the general problem in detail. We start by looking at Hessian approximations in detail. We want to learn how to use them to derive estimates.

20.3.1 Hessian Error Estimates

To save some typing, we will start using the superscript $()^o$ to indicate an expression evaluated at the point (x_0, μ_0) . The superscript $()^\theta$ refers to the expression evaluated at the point on the line segment $[(x_0, \mu_0), (x_0 + \Delta x, \mu_0 + \Delta \mu)]$. In general, given a matrix A , we know that

$$\|Ax\| \leq \|A\|_{Fr} \|x\|$$

and so

$$\|\mathbf{x}^T A \mathbf{x}\| = |<\mathbf{x}, A\mathbf{x}>| \leq \|\mathbf{x}\| \|A\mathbf{x}\| \leq \|A\|_{Fr} \|\mathbf{x}\|^2$$

The Frobenius norm of A is an example of an operator norm and this kind of behavior is studied in greater detail in (Peterson (9) 2019).

Now consider the Hessian approximations of the functions L and f_i .

$$L(\mathbf{x}_0 + \Delta\mathbf{x}, \boldsymbol{\mu}_0 + \Delta\boldsymbol{\mu}) = L^0 + (\nabla_x L^0)^T \Delta\mathbf{x} + (\nabla_{\boldsymbol{\mu}} L^0)^T \Delta\boldsymbol{\mu} + \frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L,x,x}^\theta & \mathbf{H}_{L,x,\mu}^\theta \\ \mathbf{H}_{L,\mu,x}^\theta & \mathbf{H}_{L,\mu,\mu}^\theta \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}$$

where

$$\begin{aligned} \mathbf{H}_{L,x,x} &= \begin{bmatrix} L_{x_1,x_1} & \dots & L_{x_1,x_n} \\ \vdots & & \vdots \\ L_{x_n,x_1} & \dots & L_{x_n,x_n} \end{bmatrix}, \quad \mathbf{H}_{L,x,\mu} = \begin{bmatrix} L_{x_1,\mu_1} & \dots & L_{x_1,\mu_m} \\ \vdots & & \vdots \\ L_{x_n,\mu_1} & \dots & L_{x_n,\mu_m} \end{bmatrix} \\ \mathbf{H}_{L,\mu,x} &= \begin{bmatrix} L_{\mu_1,x_1} & \dots & L_{\mu_1,x_n} \\ \vdots & & \vdots \\ L_{\mu_m,x_1} & \dots & L_{\mu_m,x_n} \end{bmatrix}, \quad \mathbf{H}_{L,\mu,\mu} = \begin{bmatrix} L_{\mu_1,\mu_1} & \dots & L_{\mu_1,\mu_m} \\ \vdots & & \vdots \\ L_{\mu_m,\mu_1} & \dots & L_{\mu_m,\mu_m} \end{bmatrix} \end{aligned}$$

Expand the Hessian term:

$$\begin{aligned} & \frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L,x,x}^\theta & \mathbf{H}_{L,x,\mu}^\theta \\ \mathbf{H}_{L,\mu,x}^\theta & \mathbf{H}_{L,\mu,\mu}^\theta \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L,x,x}^\theta \Delta\mathbf{x} + \mathbf{H}_{L,x,\mu}^\theta \Delta\boldsymbol{\mu} & \mathbf{H}_{L,\mu,x}^\theta \Delta\mathbf{x} \\ \mathbf{H}_{L,\mu,x}^\theta \Delta\mathbf{x} & \mathbf{H}_{L,\mu,\mu}^\theta \Delta\boldsymbol{\mu} \end{bmatrix} \\ &= \frac{1}{2} \left(\mathbf{H}_{L,x,x}^\theta (\Delta\mathbf{x})^T \Delta\mathbf{x} + \mathbf{H}_{L,x,\mu}^\theta (\Delta\mathbf{x})^T \Delta\boldsymbol{\mu} + \mathbf{H}_{L,\mu,x}^\theta (\Delta\boldsymbol{\mu})^T \Delta\mathbf{x} + \mathbf{H}_{L,\mu,\mu}^\theta (\Delta\boldsymbol{\mu})^T \Delta\boldsymbol{\mu} \right) \end{aligned}$$

Assuming sufficient smoothness for L , in any closed and bounded domain $\Omega \subset \mathbb{R}^n \times \mathbb{R}^m$, there exists a bound $B_L(\bar{\Omega})$ such that

$$\max_{\bar{\Omega}} \left(\left| \frac{\partial^2 L}{\partial x_i \partial \mu_j} \right|, \left| \frac{\partial^2 L}{\partial x_i \partial x_j} \right|, \left| \frac{\partial^2 L}{\partial \mu_i \partial \mu_j} \right| \right) < B_L^2(\bar{\Omega})$$

Thus,

$$\begin{aligned} \|\mathbf{H}_{L,x,x}\|_{Fr} &\leq n M_L(\bar{\Omega}), \quad \|\mathbf{H}_{L,x,\mu}\|_{Fr} \leq \sqrt{nm} M_L(\bar{\Omega}) \\ \|\mathbf{H}_{L,\mu,x}\|_{Fr} &\leq \sqrt{nm} M_L(\bar{\Omega}), \quad \|\mathbf{H}_{L,\mu,\mu}\|_{Fr} \leq m M_L(\bar{\Omega}) \end{aligned}$$

Let $M_L(\bar{\Omega}) = \max\{\sqrt{nm}, n, m\} M_L(\bar{\Omega})$. Then we have

$$\begin{aligned} \|\mathbf{H}_{L,x,x}\|_{Fr} &< M_L(\bar{\Omega}), \quad \|\mathbf{H}_{L,x,\mu}\|_{Fr} < M_L(\bar{\Omega}) \\ \|\mathbf{H}_{L,\mu,x}\|_{Fr} &< M_L(\bar{\Omega}), \quad \|\mathbf{H}_{L,\mu,\mu}\|_{Fr} < M_L(\bar{\Omega}) \end{aligned}$$

for any $\bar{\Omega}$ containing $(\mathbf{x}, \boldsymbol{\mu})$. So then

$$\frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L,x,x}^\theta & \mathbf{H}_{L,x,\mu}^\theta \\ \mathbf{H}_{L,\mu,x}^\theta & \mathbf{H}_{L,\mu,\mu}^\theta \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix} \leq \frac{M_L(\bar{\Omega})}{2} [\|\Delta\mathbf{x}\|^2 + 2\|\Delta\mathbf{x}\|\|\Delta\boldsymbol{\mu}\| + \|\Delta\boldsymbol{\mu}\|^2]$$

20.3. CONSTRAINED PARAMETRIC OPTIMIZATION

417

$$= \frac{M_L(\bar{\Omega})}{2} (\|\Delta\mathbf{x}\| + \|\Delta\boldsymbol{\mu}\|)^2$$

We know for two non negative numbers a and b , $(a+b)^2 \leq 2(a^2 + b^2)$. Thus,

$$\frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L_{x,x}}^\theta & \mathbf{H}_{L_{x,\mu}}^\theta \\ \mathbf{H}_{L_{\mu,x}}^\theta & \mathbf{H}_{L_{\mu,\mu}}^\theta \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix} \leq M_L(\bar{\Omega})(\|\Delta\mathbf{x}\|^2 + \|\Delta\boldsymbol{\mu}\|^2)$$

Now

$$L(\mathbf{x}_0 + \Delta\mathbf{x}, \boldsymbol{\mu}_0 + \Delta\boldsymbol{\mu}) - L^0 - \nabla_x L^0 \Delta\mathbf{x} - \nabla_\mu L^0 \Delta\boldsymbol{\mu} = \frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L_{x,x}}^\theta & \mathbf{H}_{L_{x,\mu}}^\theta \\ \mathbf{H}_{L_{\mu,x}}^\theta & \mathbf{H}_{L_{\mu,\mu}}^\theta \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}$$

Let

$$\eta_L(\Delta\mathbf{x}, \Delta\boldsymbol{\mu}) = \frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L_{x,x}}^\theta & \mathbf{H}_{L_{x,\mu}}^\theta \\ \mathbf{H}_{L_{\mu,x}}^\theta & \mathbf{H}_{L_{\mu,\mu}}^\theta \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}$$

Now for any fixed R we can set $\bar{\Omega} = \bar{B}(R, (\mathbf{x}, \boldsymbol{\mu}))$. Then in this ball

$$\Delta L = (\nabla_x L^0)^T \Delta\mathbf{x} + (\nabla_\mu L^0)^T \Delta\boldsymbol{\mu} + \eta_L(\Delta\mathbf{x}, \Delta\boldsymbol{\mu})$$

where $\Delta L = L(\mathbf{x}_0 + \Delta\mathbf{x}, \boldsymbol{\mu}_0 + \Delta\boldsymbol{\mu}) - L^0$. Since all the partials of L are continuous locally, for sufficiently small R , L is differentiable and thus we know and $\eta_L \rightarrow 0$ as $\Delta\mathbf{x}, \Delta\boldsymbol{\mu} \rightarrow 0$ and $\|\eta_L\|/\|(\Delta\mathbf{x}, \Delta\boldsymbol{\mu})\| \rightarrow 0$ as $\|(\Delta\mathbf{x}, \Delta\boldsymbol{\mu})\| \rightarrow 0$. Of course, we have already shown this ourselves as the estimate

$$\eta_L(\Delta\mathbf{x}, \Delta\boldsymbol{\mu}) \leq M_L(\bar{\Omega})(\|\Delta\mathbf{x}\|^2 + \|\Delta\boldsymbol{\mu}\|^2)$$

clearly shows this. Also, in using the same argument as above on each f_i we get

$$\Delta f_i = \nabla_x f_i^0 \Delta\mathbf{x} + \nabla_\mu f_i^0 \Delta\boldsymbol{\mu} + \eta_{f_i}(\Delta\mathbf{x}, \Delta\boldsymbol{\mu})$$

and $\eta_{f_i} \rightarrow 0$ as $\Delta\mathbf{x}, \Delta\boldsymbol{\mu} \rightarrow 0$. Here

$$\eta_{f_i}(\Delta\mathbf{x}, \Delta\boldsymbol{\mu}) = \frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{f_i x,x}^\phi & \mathbf{H}_{f_i x,\mu}^\phi \\ \mathbf{H}_{f_i \mu,x}^\phi & \mathbf{H}_{f_i \mu,\mu}^\phi \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\mu} \end{bmatrix}$$

with

$$|\eta_{f_i}(\Delta\mathbf{x}, \Delta\boldsymbol{\mu})| \leq M_{f_i}(\bar{\Omega})(\|\Delta\mathbf{x}\|^2 + \|\Delta\boldsymbol{\mu}\|^2)$$

for all i , $1 \leq i \leq n$. So in putting all of the Δf_i into one vector we see

$$\underbrace{\begin{bmatrix} \Delta f_1 \\ \Delta f_2 \\ \vdots \\ \Delta f_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} \frac{\partial f_1^0}{\partial x_1} & \cdots & \frac{\partial f_1^0}{\partial x_n} \\ \frac{\partial f_2^0}{\partial x_1} & \cdots & \frac{\partial f_2^0}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_n^0}{\partial x_1} & \cdots & \frac{\partial f_n^0}{\partial x_n} \end{bmatrix}}_{n \times n} \underbrace{\begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \vdots \\ \Delta x_n \end{bmatrix}}_{n \times 1} + \underbrace{\begin{bmatrix} \frac{\partial f_1^0}{\partial \mu_1} & \cdots & \frac{\partial f_1^0}{\partial \mu_m} \\ \frac{\partial f_2^0}{\partial \mu_1} & \cdots & \frac{\partial f_2^0}{\partial \mu_m} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_n^0}{\partial \mu_1} & \cdots & \frac{\partial f_n^0}{\partial \mu_m} \end{bmatrix}}_{n \times m} \underbrace{\begin{bmatrix} \Delta \mu_1 \\ \Delta \mu_2 \\ \vdots \\ \Delta \mu_m \end{bmatrix}}_{m \times 1}$$

$$+ \underbrace{\begin{bmatrix} \eta_{f_1}(\Delta x, \Delta \mu) \\ \eta_{f_2}(\Delta x, \Delta \mu) \\ \vdots \\ \eta_{f_n}(\Delta x, \Delta \mu) \end{bmatrix}}_{n \times 1}$$

We can rewrite this in terms of gradients:

$$\underbrace{\begin{bmatrix} \Delta f_1 \\ \vdots \\ \Delta f_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} (\nabla_{f_1,x}^0)^T \\ \vdots \\ (\nabla_{f_n,x}^0)^T \end{bmatrix}}_{n \times n} \underbrace{\begin{bmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{bmatrix}}_{n \times 1} + \underbrace{\begin{bmatrix} (\nabla_{f_1,\mu}^0)^T \\ \vdots \\ (\nabla_{f_n,\mu}^0)^T \end{bmatrix}}_{n \times m} \underbrace{\begin{bmatrix} \Delta \mu_1 \\ \vdots \\ \Delta \mu_m \end{bmatrix}}_{m \times 1} + \underbrace{\begin{bmatrix} \eta_{f_1}(\Delta x, \Delta \mu) \\ \vdots \\ \eta_{f_n}(\Delta x, \Delta \mu) \end{bmatrix}}_{n \times 1}$$

Now let

$$\mathbf{J}_{\mathbf{f}_x}^0 = \begin{bmatrix} (\nabla_{f_1,x}^0)^T \\ \vdots \\ (\nabla_{f_n,x}^0)^T \end{bmatrix}, \quad \mathbf{J}_{\mathbf{f}_\mu}^0 = \begin{bmatrix} (\nabla_{f_1,\mu}^0)^T \\ \vdots \\ (\nabla_{f_n,\mu}^0)^T \end{bmatrix}, \quad \boldsymbol{\eta}_f = \begin{bmatrix} \eta_{f_1}(\Delta x, \Delta \mu) \\ \vdots \\ \eta_{f_n}(\Delta x, \Delta \mu) \end{bmatrix}$$

Then, we have

$$\Delta f = \mathbf{J}_{\mathbf{f}_x}^0 \Delta x + \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta \mu + \boldsymbol{\eta}_f(\Delta x, \Delta \mu)$$

We also have the estimate

$$\|\boldsymbol{\eta}_f(\Delta x, \Delta \mu)\| \leq M_f(\bar{\Omega})(\|\Delta x\|^2 + \|\Delta \mu\|^2)n$$

where

$$M_f(\bar{\Omega}) = \max(M_{f_1}(\bar{\Omega}), \dots, M_{f_n}(\bar{\Omega})).$$

To summarize, we can write

$$\begin{aligned} \Delta L &= (\nabla_x L^0)^T \Delta x + (\nabla_\mu L^0)^T \Delta \mu + \boldsymbol{\eta}_L(\Delta x, \Delta \mu) \\ \Delta f &= \mathbf{J}_{\mathbf{f}_x}^0 \Delta x + \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta \mu + \boldsymbol{\eta}_f(\Delta x, \Delta \mu) \end{aligned}$$

and we know that the error terms go to zero as Δx and $\Delta \mu$ go to zero.

20.3.2 A First Look at Constraint Satisfaction

Now we also know that constraint satisfaction, i.e. satisfying $f = 0$, implies that $\Delta f = \mathbf{0}$. So then

$$\mathbf{0} = \mathbf{J}_{\mathbf{f}_x}^0 \Delta x + \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta \mu + \boldsymbol{\eta}_f(\Delta x, \Delta \mu)$$

Assuming $(\mathbf{J}_{\mathbf{f}_x}^0)^{-1}$ exists and is continuous, we then get

$$\Delta x = -(\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta \mu - (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \boldsymbol{\eta}_f(\Delta x, \Delta \mu)$$

Note the Δx on both sides of this equation. We can not really solve for Δx of course, but ignoring that we see the above expression yields

$$\|(\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \boldsymbol{\eta}_f(\Delta x, \Delta \mu)\| \leq \|(\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta \mu\| + \|\Delta x\| \leq \|(\mathbf{J}_{\mathbf{f}_x}^0)^{-1}\| \|\mathbf{J}_{\mathbf{f}_\mu}^0 \Delta \mu\| + \|\Delta x\|$$

$$\leq \|(\mathbf{J}_{\mathbf{f}_x}^0)^{-1}\| \|\mathbf{J}_{\mathbf{f}_\mu}^0\| \Delta\mu \| + \|\Delta x\|$$

From our discussions of the bounds due to the smoothness of the functions here, there is an overestimate $B(\bar{\Omega})$ such that

$$\|(\mathbf{J}_{\mathbf{f}_x}^0)^{-1}\| \|\mathbf{J}_{\mathbf{f}_\mu}^0\| \leq B(\bar{\Omega})$$

and so

$$\|(\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \eta_f(\Delta x, \Delta\mu)\| \leq B(\bar{\Omega}) \Delta\mu + \Delta x$$

Using this, we get that the change in performance ΔL is given by

$$\begin{aligned} \Delta L &= (\nabla_x L^0)^T \left(-(\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta\mu - (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \eta_f(\Delta x, \Delta\mu) \right) + (\nabla_\mu L^0)^T \Delta\mu + \eta_L(\Delta x, \Delta\mu) \\ &= -(\nabla_x L^0)^T (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta\mu - (\nabla_x L^0)^T (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \eta_f(\Delta x, \Delta\mu) + (\nabla_\mu L^0)^T \Delta\mu + \eta_L(\Delta x, \Delta\mu) \end{aligned}$$

Letting $\alpha = (\nabla_\mu L^0)^T - (\nabla_x L^0)^T (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{J}_{\mathbf{f}_\mu}^0$ we see for a minimum, $\Delta L > 0$. Thus,

$$0 \leq \alpha \Delta\mu - (\nabla_x L^0)^T (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \eta_f(\Delta x, \Delta\mu) + \eta_L(\Delta x, \Delta\mu)$$

20.3.3 Constraint Satisfaction and the Implicit Function Theorem

We want to show the $\alpha = 0$ at the minimum of L subject to the constraints, but as written Δx and $\Delta\mu$ are dependent of one another. What we need is to find how Δx depends on $\Delta\mu$ so we use to write an inequality equation only in $\Delta\mu$. We will do this by stepping back, rewriting a few things and invoking the Implicit Function Theorem. It is the Implicit Function Theorem that will allow us to write Δx in terms of $\Delta\mu$ only rather than in terms of itself and Δx . At the local minimum $(\mathbf{x}_0, \boldsymbol{\mu}_0)$, we have constraint satisfaction and so

$$\begin{aligned} 0 &\leq \Delta L = (\nabla_x L^0)^T \Delta x + (\nabla_\mu L^0)^T \Delta\mu + \eta_L(\Delta x, \Delta\mu) \\ \mathbf{0} &= \Delta f = \mathbf{J}_{\mathbf{f}_x}^0 \Delta x + \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta\mu + \eta_f(\Delta x, \Delta\mu) \end{aligned}$$

where the Hessian terms in η_L and η_f are evaluated at in between points

$$(\mathbf{x}^\theta, \boldsymbol{\mu}^\theta) = (1 - \theta)(\mathbf{x}_0, \boldsymbol{\mu}_0) + \theta(\mathbf{x}_0 + \Delta x, \boldsymbol{\mu}_0 + \Delta\mu)$$

Here we could assume that $\mathbf{J}_{\mathbf{f}_x}^0)^{-1}$ exists and solve for Δx like we did last time, but we must realize that there is a relationship between Δx and $\Delta\mu$ and that we need to do something else to be precise. We can use the Implicit Function Theorem to write Δx in terms of $\Delta\mu$ to provide this precision. The Implicit Function Theorem, Theorem 13.3.1 has been carefully proven and discussed in Chapter 13. If we translate it to our situation, it would be stated

Implicit Function Theorem

Let $U \subset \mathbb{R}^{n+m}$ be an open set. Let $\mathbf{u} \in U$ be written as $\mathbf{u} = (\mathbf{x}, \boldsymbol{\mu})$ where $\mathbf{x} \in \mathbb{R}^n$ and $\boldsymbol{\mu} \in \mathbb{R}^m$. Assume $f : U \rightarrow \mathbb{R}^m$ has continuous first order partials in U and there is a point $(\mathbf{x}_0, \boldsymbol{\mu}_0) \in U$ satisfying $f(\mathbf{x}_0, \boldsymbol{\mu}_0) = \mathbf{0}$. This is the constraint satisfaction condition. Also assume $\det \mathbf{J}_{\mathbf{f}_x}^0 \neq 0$. Then there is an open set V_0 containing $\boldsymbol{\mu}_0 \in \mathbb{R}^m$ and $g : V_0 \rightarrow \mathbb{R}^n$ with continuous partials so that $g(\boldsymbol{\mu}_0) = \mathbf{x}_0$, $f(g(\boldsymbol{\mu}), \boldsymbol{\mu}) = \mathbf{0}$ on V_0 .

This tells us we have constraint satisfaction: $f(g(\boldsymbol{\mu}_0 + \Delta\mu), \boldsymbol{\mu}_0 + \Delta\mu) = \mathbf{0}$ for all $\boldsymbol{\mu}_0 + \Delta\mu \in V_0$. Also, since $g(\boldsymbol{\mu}_0) = \mathbf{x}_0$, letting $g(\boldsymbol{\mu}_0 + \Delta\mu) = \mathbf{x}$, then $\Delta x = g(\boldsymbol{\mu}_0 + \Delta\mu) - g(\boldsymbol{\mu}_0) = \mathbf{x} - \mathbf{x}_0$ which is our usual notation.

Recall that we had from before

$$0 \leq \Delta L = (\nabla_x L^0)^T \Delta x + (\nabla_\mu L^0)^T \Delta \mu + \frac{1}{2} \begin{bmatrix} \Delta x \\ \Delta \mu \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L_{x,x}}^\theta & \mathbf{H}_{L_{x,\mu}}^\theta \\ \mathbf{H}_{L_{\mu,x}}^\theta & \mathbf{H}_{L_{\mu,\mu}}^\theta \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta \mu \end{bmatrix}$$

Let $\Omega = B(r, \mathbf{x}_0, \boldsymbol{\mu}_0)$ and apply the Implicit Function Theorem. Associated with Ω will be an open set V_0 . Choose a ball $B(\rho, \mathbf{x}_0)$ such that $B\rho, \mathbf{x}_0) \subset V_0$. Then for all $\Delta x \in B(\rho, \boldsymbol{\mu}_0)$ we know

$$\begin{aligned} 0 \leq \Delta L &= (\nabla_x L^0)^T \Delta x + (\nabla_\mu L^0)^T \Delta \mu + \\ &\quad \frac{1}{2} \begin{bmatrix} g(\boldsymbol{\mu}_0 + \Delta \mu) - g(\boldsymbol{\mu}_0) \\ \Delta \mu \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L_{x,x}}^\theta & \mathbf{H}_{L_{x,\mu}}^\theta \\ \mathbf{H}_{L_{\mu,x}}^\theta & \mathbf{H}_{L_{\mu,\mu}}^\theta \end{bmatrix} \begin{bmatrix} g(\boldsymbol{\mu}_0 + \Delta \mu) - g(\boldsymbol{\mu}_0) \\ \Delta \mu \end{bmatrix} \end{aligned}$$

Now consider

$$\mathbf{H}_L^\theta(\Delta \mu) = \frac{1}{2} \begin{bmatrix} g(\boldsymbol{\mu}_0 + \Delta \mu) - g(\boldsymbol{\mu}_0) \\ \Delta \mu \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{L_{x,x}}^\theta & \mathbf{H}_{L_{x,\mu}}^\theta \\ \mathbf{H}_{L_{\mu,x}}^\theta & \mathbf{H}_{L_{\mu,\mu}}^\theta \end{bmatrix} \begin{bmatrix} g(\boldsymbol{\mu}_0 + \Delta \mu) - g(\boldsymbol{\mu}_0) \\ \Delta \mu \end{bmatrix}$$

The matrix in the middle is evaluated at the point

$$\begin{aligned} (\mathbf{x}^\theta, \boldsymbol{\mu}^\theta) &= (1 - \theta)(\mathbf{x}_0, \boldsymbol{\mu}_0) + \theta(\mathbf{x}_0 + \Delta x, \boldsymbol{\mu}_0 + \Delta \mu) = (\mathbf{x}_0, \boldsymbol{\mu}_0) + \theta(\Delta x, \Delta \mu) \\ &= (g(\boldsymbol{\mu}_0), \boldsymbol{\mu}_0) + \theta(g(\boldsymbol{\mu}_0 + \Delta \mu) - g(\boldsymbol{\mu}_0), \Delta \mu) \end{aligned}$$

Note here that

$$(\mathbf{x}^\theta, \boldsymbol{\mu}^\theta) - (\mathbf{x}_0, \boldsymbol{\mu}_0) = \theta(\Delta x, \Delta \mu) = \theta(g(\boldsymbol{\mu}_0 + \Delta \mu) - g(\boldsymbol{\mu}_0), \Delta \mu)$$

is in $B(r, (\mathbf{0}, \mathbf{0}))$ as $(g(\boldsymbol{\mu}_0 + \Delta \mu), \boldsymbol{\mu}_0) + \Delta \mu$ is in $B(r, \mathbf{x}_0, \boldsymbol{\mu}_0)$. Note also that by our assumptions $\mathbf{H}_L^\theta(\Delta \mu)$ is continuous. So now

$$0 \leq \Delta L = (\nabla_x L^0)^T \Delta x + (\nabla_\mu L^0)^T \Delta \mu + H_L^\theta(\Delta \mu)$$

and because constraints are satisfied

$$0 = \mathbf{J}_{\mathbf{f}_x^0} \Delta x + \mathbf{J}_{\mathbf{f}_\mu^0} \Delta \mu + \mathbf{H}_{\mathbf{f}}^\phi(\Delta \mu)$$

where

$$\begin{aligned} \mathbf{H}_{\mathbf{f}_i}^\phi(\Delta \mu) &= \frac{1}{2} \begin{bmatrix} g(\boldsymbol{\mu}_0 + \Delta \mu) - g(\boldsymbol{\mu}_0) \\ \Delta \mu \end{bmatrix}^T \begin{bmatrix} \mathbf{H}_{\mathbf{f}_{i,x}}^\theta & \mathbf{H}_{\mathbf{f}_{i,x,\mu}}^\theta \\ \mathbf{H}_{\mathbf{f}_{i,\mu,x}}^\theta & \mathbf{H}_{\mathbf{f}_{i,\mu,\mu}}^\theta \end{bmatrix} \begin{bmatrix} g(\boldsymbol{\mu}_0 + \Delta \mu) - g(\boldsymbol{\mu}_0) \\ \Delta \mu \end{bmatrix} \\ \mathbf{H}_{\mathbf{f}}^\phi(\Delta \mu) &= \begin{bmatrix} \mathbf{H}_{\mathbf{f}_1}^\phi(\Delta \mu) \\ \vdots \\ \mathbf{H}_{\mathbf{f}_n}^\phi(\Delta \mu) \end{bmatrix} \end{aligned}$$

Hence, $H_f^\phi(\Delta \mu)$ is constructed in the same way as $H_L^\theta(\Delta \mu)$ but the intermediate point is now $(\mathbf{x}^\phi, \boldsymbol{\mu}^\phi)$. Then since we know $\mathbf{J}_{\mathbf{f}_x^0}$ is invertible we have

$$\Delta x = -(\mathbf{J}_{\mathbf{f}_x^0})^{-1} \mathbf{J}_{\mathbf{f}_\mu^0} \Delta \mu - (\mathbf{J}_{\mathbf{f}_x^0})^{-1} \mathbf{H}_{\mathbf{f}}^\phi(\Delta \mu)$$

which is an expression for Δx written in terms of $\Delta\mu$ only. This is what we wanted all along. Now, plugging in for Δx in ΔL we get

$$\begin{aligned} 0 \leq \Delta L &= (\nabla_x L^0)^T \left(-(\mathbf{J}_{f_x}^0)^{-1} \mathbf{J}_{f_\mu}^0 \Delta\mu - (\mathbf{J}_{f_x}^0)^{-1} \mathbf{H}_f^\phi(\Delta\mu) \right) + (\nabla_\mu L^0)^T \Delta\mu + H_L^\theta(\Delta\mu) \\ &= \left((\nabla_\mu L^0)^T - (\nabla_x L^0)^T (\mathbf{J}_{f_x}^0)^{-1} \mathbf{J}_{f_\mu}^0 \right) \Delta\mu - (\nabla_x L^0)^T (\mathbf{J}_{f_x}^0)^{-1} \mathbf{H}_f^\phi(\Delta\mu) + H_L^\theta(\Delta\mu) \end{aligned}$$

Now let $\alpha = (\nabla_\mu L^0)^T - (\nabla_x L^0)^T (\mathbf{J}_{f_x}^0)^{-1} \mathbf{J}_{f_\mu}^0$ and $\beta(\Delta\mu) = -(\nabla_x L^0)^T (\mathbf{J}_{f_x}^0)^{-1} \mathbf{H}_f^\phi(\Delta\mu) + H_L^\theta(\Delta\mu)$. We then have

$$0 \leq \alpha \Delta\mu + \beta(\Delta\mu)$$

for $\Delta\mu \in B(\rho, \mathbf{0})$ and we see β is continuous on $B(\rho, \mathbf{0})$. Our goal is to show that $\alpha = 0$, so that we will have the known necessary conditions for constrained optimization. In trying to show this we will attempt to get a bound on $\beta(\Delta\mu)$ and then try to trap α in between two arbitrarily small quantities. Thus showing that α must be 0. There are two approaches that we will take. Our first attempt will be an incorrect approach and will lead us to a dead end. Our second approach will back off a little, summarizing some of this material once again, and proceed on to the result that we are trying to get to. Namely, that $\alpha = 0$.

20.3.3.1 An Incorrect Approach

Picking up where we left off, since β is continuous on $B(\rho\mu_0)$ we know that given $\epsilon > 0$, there is a $\delta > 0$ so that

$$\Delta\mu \in B(\delta, \mathbf{0}) \implies |\beta(\Delta\mu) - \beta(0)| < \epsilon$$

But $\beta(0) = 0$ so

$$\Delta\mu \in B(\delta, \mathbf{0}) \implies |\beta(\Delta\mu)| < \epsilon$$

Choose $\Delta\mu = cE_i$ for c sufficiently small. Then we have if $|c| < \delta$,

$$\Delta\mu = \begin{bmatrix} 0 \\ \vdots \\ c \\ \vdots \\ 0 \end{bmatrix} \leftarrow j^{th} \text{slot}, \quad \text{and } 0 \leq \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_j \\ \vdots \\ \alpha_m \end{bmatrix}^T \begin{bmatrix} 0 \\ \vdots \\ c \\ \vdots \\ 0 \end{bmatrix} + \epsilon = \alpha_j c + \epsilon$$

So $0 \leq \alpha_j c + \epsilon$ if $|c| < \delta$.

$$\text{Choose } c = \frac{\delta}{2}$$

$$\text{Choose } c = -\frac{\delta}{2}$$

$$0 \leq \alpha_j \frac{\delta}{2} + \epsilon$$

$$0 \leq \alpha_j \left(-\frac{\delta}{2}\right) + \epsilon$$

$$-\epsilon \leq \alpha_j \frac{\delta}{2}$$

$$-\epsilon \leq \alpha_j \left(-\frac{\delta}{2}\right)$$

$$-\frac{2\epsilon}{\delta} \leq \alpha_j$$

$$\frac{-2\epsilon}{\delta} \geq \alpha_j$$

In combining these two inequalities we get

$$-\frac{2\epsilon}{\delta} \leq \alpha_j \leq \frac{2\epsilon}{\delta}$$

We would like to say here that both sides of this inequality go to zero as $\epsilon \rightarrow 0$, but since we do not know anything about how ϵ depends on δ we have no idea what the ratio $\frac{\epsilon}{\delta}$ does as $\epsilon \rightarrow 0$. So we can see that this approach is not going to work! Arghh!

20.3.3.2 The Correct Approach

To back off a little bit, we know that there is always a linkage between the controls and the states. Look at $B(\frac{\rho}{2}, \mu_0)$. Then

$$\begin{aligned} \mathbf{x}_0 &= g(\mu_0) \\ \mathbf{x}_0 + \Delta x &= g(\mu_0 + \Delta \mu) \\ f(g(\mu_0 + \Delta \mu), \mu_0 + \Delta \mu) &\in B(r, \mathbf{x}_0, \mu_0) \\ f(g(\mu_0 + \Delta \mu), \mu_0 + \Delta \mu) &= \mathbf{0}, \Delta \mu \in B(\frac{\rho}{2}, \mu_0) \end{aligned}$$

Constraint satisfaction then gives us

$$\mathbf{0} = \mathbf{J}_{f_x}^0 \Delta x + \mathbf{J}_{f_\mu}^0 \Delta \mu + \mathbf{H}_f^\phi(\Delta \mu)$$

for $\Delta \mu \in B(\frac{\rho}{2}, \mu_0)$. Now examine the last term $\mathbf{H}_f^\phi(\Delta \mu)$. Writing $H_f(\Delta \mu)$ out we see the i^{th} component is

$$\begin{aligned} (H_f(\Delta \mu))_i &= \frac{1}{2}(\Delta x)^T \mathbf{H}_{f_{ixx}}^\phi \Delta x + \frac{1}{2}(\Delta \mu)^T \mathbf{H}_{f_{ix\mu}}^\phi \Delta x \\ &+ \frac{1}{2}(\Delta x)^T \mathbf{H}_{f_{i\mu x}}^\phi \Delta \mu + \frac{1}{2}(\Delta \mu)^T \mathbf{H}_{f_{i\mu\mu}}^\phi \Delta \mu \end{aligned}$$

and so

$$\begin{aligned} 2 \| (H_f(\Delta \mu))_i \| &\leq \|\mathbf{H}_{f_{ixx}}^\phi\| \|\Delta x\|^2 + \|\mathbf{H}_{f_{ix\mu}}^\phi\| \|\Delta x\| \|\Delta \mu\| \\ &+ \|\mathbf{H}_{f_{i\mu x}}^\phi\| \|\Delta x\| \|\Delta \mu\| + \|\mathbf{H}_{f_{i\mu\mu}}^\phi\| \|\Delta \mu\|^2 \end{aligned}$$

But,

$$g(\mu_0 + \Delta \mu) - g(\mu_0) = \mathbf{J}_{g_\mu}^0 \Delta \mu + \frac{1}{2}(\Delta \mu)^T \mathbf{H}_{g_\mu}^\xi \Delta \mu$$

for an intermediate point ξ . Therefore

$$\|\Delta x\| \leq \|\mathbf{J}_{g_\mu}^0\| \|\Delta \mu\| + \frac{1}{2} \|\mathbf{H}_{g_\mu}^\xi\| \|\Delta \mu\|^2$$

Now for any fixed R we can set $\bar{\Omega} = \bar{B}(R, (x, \mu))$ which is a compact set. Since all the partials here are continuous, this means all the gradient and Hessian terms have maximum values on $\bar{\Omega}$. Thus (this is a bit messy as there are a lot of upper bounds!)

$$\begin{aligned} \|\mathbf{H}_{f_{ixx}}^\phi\| &\leq B_{1i}(\bar{\Omega}), \quad \|\mathbf{H}_{f_{ix\mu}}^\phi\| \leq B_{2i}(\bar{\Omega}) \\ \|\mathbf{H}_{f_{i\mu x}}^\phi\| &\leq B_{3i}(\bar{\Omega}), \quad \|\mathbf{H}_{f_{i\mu\mu}}^\phi\| \leq B_{4i}(\bar{\Omega}) \\ \|\mathbf{H}_{g_\mu}^\xi\| &\leq C(\bar{\Omega}) \end{aligned}$$

Thus,

$$2 \|(H_f(\Delta\mu))_i\| \leq B_{1i}(\bar{\Omega}) \|\Delta x\|^2 + B_{2i}(\bar{\Omega}) \|\Delta x\| \|\Delta\mu\| B_{3i}(\bar{\Omega}) \|\Delta x\| \|\Delta\mu\| + B_{4i}(\bar{\Omega}) \|\Delta\mu\|^2$$

Let

$$\begin{aligned} B(\bar{\Omega}) &= \max_{1 \leq i \leq n} \{B_{1i}(\bar{\Omega}), B_{2i}(\bar{\Omega}), B_{3i}(\bar{\Omega}), B_{4i}(\bar{\Omega})\} \\ D(\bar{\Omega}) &= \max\{\|\mathbf{J}_{\mathbf{g}_\mu}^0\|, \frac{1}{2} C(\bar{\Omega})\} \end{aligned}$$

$$\|\Delta x\| \leq D(\bar{\Omega}) \|\Delta\mu\| + D(\bar{\Omega}) \|\Delta\mu\|^2$$

Now assume $\|\Delta\mu\| < 1$, then $\|\Delta x\| \leq 2D(\bar{\Omega})\|\Delta\mu\|$. We can now do our final estimate:

$$\begin{aligned} 2 \max_{1 \leq i \leq n} \|(H_f(\Delta\mu))_i\| &\leq B(\bar{\Omega}) \left(\|\Delta x\|^2 + 2\|\Delta x\| \|\Delta\mu\| + \|\Delta\mu\|^2 \right) \\ &= B(\bar{\Omega})(\|\Delta x\|^2 + \|\Delta\mu\|^2) \leq 2B(\bar{\Omega})(\|\Delta x\|^2 + \|\Delta\mu\|^2) \end{aligned}$$

Thus

$$\max_{1 \leq i \leq n} \|(H_f(\Delta\mu))_i\| \leq B(\bar{\Omega})(\|\Delta x\|^2 + \|\Delta\mu\|^2)$$

Next, using the estimate for $\|\Delta x\|$, we find

$$\max_{1 \leq i \leq n} \|(H_f(\Delta\mu))_i\| \leq B(\bar{\Omega})(4D(\bar{\Omega})^2 + 1) \|\Delta\mu\|^2$$

Let

$$\xi(\bar{\Omega}) = B(\bar{\Omega})(4D(\bar{\Omega})^2 + 1)$$

and we have shown

$$\max_{1 \leq i \leq n} \|(H_f(\Delta\mu))_i\| \leq \xi(\bar{\Omega}) \|\Delta\mu\|^2$$

Thus, to repeat, we have the following constraint equations

$$0 = \mathbf{J}_{\mathbf{f}_x}^0 \Delta x + \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta\mu + \mathbf{H}_f^\phi(\Delta\mu)$$

where

$$\|\mathbf{H}_f^\phi(\Delta\mu)\| = \max_{1 \leq i \leq n} \|(H_f(\Delta\mu))_i\| \leq \xi(\bar{\Omega}) \|\Delta\mu\|^2$$

Using the same argument we used above we will also get that

$$0 \leq \Delta L = (\nabla_x L^0)^T \Delta x + (\nabla_\mu L^0)^T \Delta\mu + H_L^\theta(\Delta\mu)$$

$$\|\mathbf{H}_L^\theta(\Delta\mu)\| \leq \zeta(\bar{\Omega}) \|\Delta\mu\|^2$$

some constant $\zeta(\bar{\Omega})$. And now since we know $\mathbf{J}_{\mathbf{f}_x}^0$ is invertible we know from the constraint equation that

$$\Delta x = -(\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta \mu - (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{H}_{\mathbf{f}}^\phi(\Delta \mu)$$

So putting this into the expression for ΔL we will get an expression that we have seen before. Namely,

$$\begin{aligned} 0 \leq \Delta L &= (\nabla_x^0)^T \left(-(\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{J}_{\mathbf{f}_\mu}^0 \Delta \mu - (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{H}_{\mathbf{f}}^\phi(\Delta \mu) \right) + (\nabla_\mu L^0)^T \Delta \mu + H_L^\theta(\Delta \mu) \\ &= \left(\nabla_\mu L^0 - (\nabla_x L^0)^T (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{J}_{\mathbf{f}_\mu}^0 \right) \Delta \mu - (\nabla_x L^0)^T (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{H}_{\mathbf{f}}^\phi(\Delta \mu) + H_L^\theta(\Delta \mu) \end{aligned}$$

Letting

$$\alpha = (\nabla_\mu L^0)^T - (\nabla_x L^0)^T (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{J}_{\mathbf{f}_\mu}^0$$

and

$$Q(\Delta \mu) = -(\nabla_x L^0)^T (\mathbf{J}_{\mathbf{f}_x}^0)^{-1} \mathbf{H}_{\mathbf{f}}^\phi(\Delta \mu) + H_L^\theta(\Delta \mu)$$

we get

$$0 \leq \Delta L = \alpha \Delta \mu + Q(\Delta \mu).$$

From all of the above work and the fact the $(\mathbf{J}_{\mathbf{f}_x}^0)^{-1}$ exists, we have

$$\begin{aligned} |Q(\Delta \mu)| &\leq \|\nabla_x L^0\| \|(\mathbf{J}_{\mathbf{f}_x}^0)^{-1}\| \|\xi(\bar{\Omega})\| \|\Delta \mu\|^2 + C(\bar{\Omega}) \|\Delta \mu\|^2 \\ &= \left(\|\nabla_x L^0\| \|(\mathbf{J}_{\mathbf{f}_x}^0)^{-1}\| \|\xi(\bar{\Omega})\| + C(\bar{\Omega}) \right) \|\Delta \mu\|^2 \end{aligned}$$

Now let

$$K(\bar{\Omega}) = \|\nabla_x L^0\| \|(\mathbf{J}_{\mathbf{f}_x}^0)^{-1}\| \|\xi(\bar{\Omega})\| + C(\bar{\Omega})$$

and we have shown $|Q(\Delta \mu)| \leq K(\bar{\Omega}) \|\Delta \mu\|^2$. So now we have the inequality

$$0 \leq \alpha \Delta \mu + K \|\Delta \mu\|^2.$$

Notice that this is similar to the inequality that we had at the end of the other approach. This time though we will be able to get the result that we want, i.e. $\alpha = 0$. Now choose all $\Delta \mu_i = 0$ except for $\Delta \mu_j$, and let $\Delta \mu_j = \frac{1}{\sqrt{Kp}}$. For p large enough, we know that $\Delta \mu \in \bar{B}(\frac{p}{2}, \mu_0)$. Looking at the j^{th} component of the inequality we see

$$0 \leq \alpha_j \frac{1}{\sqrt{Kp}} + K \frac{1}{Kp^2} \implies -\frac{1}{p^2} \leq \frac{1}{p\sqrt{K}} \alpha_j \implies -\frac{p\sqrt{K}}{p^2} \leq \alpha_j \implies -\frac{\sqrt{K}}{p} \leq \alpha_j$$

Now choose $\Delta \mu_j = -\frac{1}{\sqrt{Kp}}$. Then the j^{th} component of the inequality becomes

$$0 \leq \alpha_j \left(\frac{-1}{\sqrt{Kp}} \right) + K \frac{1}{Kp^2} \implies \alpha_j \left(\frac{1}{\sqrt{Kp}} \right) \leq \frac{1}{p^2} \implies \alpha_j \leq \frac{\sqrt{K}}{p}$$

Combining these we get that

$$-\frac{\sqrt{K}}{p} \leq \alpha_j \leq \frac{\sqrt{K}}{p}$$

Since p is arbitrarily large, this implies $\alpha_j = 0$. This same argument works for any of the components $\alpha_1, \alpha_2, \dots, \alpha_n$. So we are finally able to conclude that $\alpha = 0$, and we have found the necessary conditions for constrained optimization.

Theorem 20.3.1 Constrained Optimization

Consider the problem

$$\min_{\mu \in \mathbb{R}^m} L(\mathbf{x}, \boldsymbol{\mu})$$

subject to

$$\left. \begin{array}{rcl} f_1(\mathbf{x}, \boldsymbol{\mu}) & = & 0 \\ \vdots & & \\ f_n(\mathbf{x}, \boldsymbol{\mu}) & = & 0 \end{array} \right\} f(\mathbf{x}, \boldsymbol{\mu}) = 0$$

where $\mathbf{x} \in \mathbb{R}^n$ is the **state**, $\boldsymbol{\mu} \in \mathbb{R}^m$ is the **control** and $L : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is the **performance index**. The function $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ is the **constraint**. We assume the functions f and L have continuous partials locally about $(\mathbf{x}_0, \boldsymbol{\mu}_0)$. Then if $(\mathbf{x}_0, \boldsymbol{\mu}_0)$ is a point where L is minimized and $(J_{f_x}(\mathbf{x}_0, \boldsymbol{\mu}_0))^{-1}$ exists, then the following equation must be satisfied:

$$(\nabla_{\boldsymbol{\mu}} L^0)^T - (\nabla_{\mathbf{x}} L^0)^T (J_{f_x}^0)^{-1} J_{f_{\boldsymbol{\mu}}}^0 = \mathbf{0}$$

where $(\cdot)^0$ indicates terms that are evaluated at $(\mathbf{x}_0, \boldsymbol{\mu}_0)$. Hence, finding the points $(\mathbf{x}_0, \boldsymbol{\mu}_0)$ that satisfy this equation are possible candidates for the minima of L subject to $f = \mathbf{0}$.

Proof 20.3.1

We have just finished a very long winded explanation! ■

20.4 Lagrange Multipliers

We could also approach this problem using the Lagrange Multiplier technique. If we let

$$\Phi(\mathbf{x}, \boldsymbol{\mu}, \lambda) = L(\mathbf{x}, \boldsymbol{\mu}) + \lambda^T f(\mathbf{x}, \boldsymbol{\mu})$$

then Φ is minimized when

$$\begin{aligned} \nabla_{\mathbf{x}} \Phi &= \mathbf{0} \implies \nabla_{\mathbf{x}} L^0 + \lambda^T J_{f_x} f^0 = \mathbf{0}, & * \\ \nabla_{\boldsymbol{\mu}} \Phi &= \mathbf{0} \implies \nabla_{\boldsymbol{\mu}} L^0 + \lambda^T J_{f_{\boldsymbol{\mu}}} f^0 = \mathbf{0}, & ** \\ \nabla_{\lambda} \Phi &= \mathbf{0} \implies f(\mathbf{x}, \boldsymbol{\mu}), & \text{constraint equations} \end{aligned}$$

If $J_{f_x} f^0$ is invertible we get from (*) that

$$\lambda^T = -\nabla_{\mathbf{x}} L^0 (J_{f_x} f^0)^{-1}$$

and so using (**) we see

$$\mathbf{0} = \nabla_\mu L^0 - \nabla_x L^0 (\mathbf{J}_{f_x} f^0)^{-1} \mathbf{J}_{f_\mu} f^0$$

which is the transpose of the same equation that we had before. To get some sort of idea what the λ_i 's are we know that to first order

$$\begin{aligned}\Delta L &\approx (\nabla_x L^0)^T \Delta x + (\nabla_\mu L^0)^T \Delta \mu \\ \Delta f &\approx \mathbf{J}_{f_x}^0 \Delta x + \mathbf{J}_{f_\mu}^0 \Delta \mu.\end{aligned}$$

Now suppose we set $\Delta \mu = 0$ and we let Δx vary. Then

$$\begin{aligned}\Delta L &\approx (\nabla_x L^0)^T \Delta x \\ \Delta f &\approx \mathbf{J}_{f_x}^0 \Delta x.\end{aligned}$$

So

$$\Delta L \approx (\nabla_x L^0)^T (\mathbf{J}_{f_x}^0)^{-1} \Delta f \implies \Delta L \approx -\lambda^T \Delta f \implies \Delta L \approx -\sum_i \lambda_i \Delta f_i$$

and thus

$$\lambda_i \approx -\frac{\Delta L}{\Delta f_i}.$$

Here the Lagrange Multipliers λ_i have an interpretation of the cost for violating the constraint f_i , i.e. they are the change in performance with respect to the change in a constraint. Of course, the argument above is very loose! But suggestive arguments can often be proven later with precision. The hardest part of all in some ways is trying to discover conjectures. Think of the pricing interpretation as a challenge! How would you prove it carefully?

Part VII

Summing It All Up



Chapter 21

Summing It All Up

We have now come to the end of these notes. We have not covered all of the things we wanted to but we view that as a plus: there is more to look forward to! In particular, we hope we have encouraged your interest in

- A thorough discussion of the result $\int_{\partial U} \omega = \int_U d\omega$ for 1 - forms ω in \Re^3 .
- The topology of \Re^n and how we generalize this in n dimensional manifolds.
- More ideas in algebraic topology and how it can be used in applied models in various areas of science such as immunology and cognition.
- learning more computational science

So here's to the future! The next book (Peterson (9) 2019) focuses on the study of metric, normed and inner product spaces in more depth. You are welcome to get started on that one next.

Part VIII

References

References

- [1] R. Bate, D. Mueller, and J. White. *Fundamentals of Astrodynamics*. Dover, 1971.
- [2] M. Braun. *Differential Equations and Their Applications*. Springer-Verlag, 1978.
- [3] H. Curtis. *Orbital Mechanics for Engineering Students*. Elsevier, 2014.
- [4] W. Fulton. *Algebraic Topology: A First Course*. Graduate Texts in Mathematics 153, Springer NY, 1995.
- [5] J. Peterson. *Calculus for Cognitive Scientists: Higher Order Models and Their Analysis*. Springer Series on Cognitive Science and Technology, Springer Science+Business Media Singapore Pte Ltd. 152 Beach Road, #22-06/08 Gateway East Singapore 189721, Singapore, 2016. doi: 10.1007/978-981-287-877-9.
- [6] J. Peterson. *Basic Analysis V: Linear Functional Analysis and Topology*. CRC Press, a Division of the Taylor and Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, Florida 33487, 2019. Contract December 15, 2016, 450 pages, 100 illustrations.
- [7] J. Peterson. *Basic Analysis IV: Abstract Measure Theory*. CRC Press, a Division of the Taylor and Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, Florida 33487, 2019. Contract December 15, 2016, 450 pages, 100 illustrations.
- [8] J. Peterson. *Basic Analysis I: Functions of a Real Variable*. CRC Press, a Division of the Taylor and Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, Florida 33487, 2019. Contract December 15, 2016, 450 pages, 100 illustrations.
- [9] J. Peterson. *Basic Analysis III: Mappings on Infinite Dimensional Spaces*. CRC Press, a Division of the Taylor and Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, Florida 33487, 2019. Contract December 15, 2016, 450 pages, 100 illustrations.
- [10] J. Peterson. *Basic Analysis II: Functions of Multiple Variables*. CRC Press, a Division of the Taylor and Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, Florida 33487, 2019. Contract December 15, 2016, 450 pages, 100 illustrations.
- [11] J. Peterson, A. M. Kesson, and N. J. C. King. A Theoretical Model of the West Nile Virus Survival Data. *BMC Immunology*, 18(Suppl 1)(22):24–38, June 21, 2017. URL <http://dx.doi.org/10.1186/s12865-017-0206-z>.
- [12] J. Peterson, A. M. Kesson, and N. J. C. King. A Model of Auto Immune Response. *BMC Immunology*, 18(Suppl 1)(24):48–65, June 21, 2017. URL <http://dx.doi.org/10.1186/s12865-017-0208-x>.
- [13] A. E. Roy. *Orbital Motion*. Adam Hilger, Ltd, Bristol, 1982.
- [14] G. Sutton and O. Biblarz. *Rocket Propulsion Elements*. John Wiley & Sons, Inc, New York, 2001.

- [15] W. Thompson. *Introduction to Space Dynamics*. Dover, 1961.
- [16] R. Wrede and M. Speigel. *Schaum’s Outline on Advanced Calculus*. McGraw - Hill Education, third edition, 2010.

Part IX

Detailed Indices



Index

- Axiom
The Completeness Axiom, 10
- Definition
 \Re^n to \Re^m Continuity, 113
2 Forms, 376
3 Forms, 376
A Refinement of a Partition, 294
Abstract Vector Space, 24
Algebraic Dual of a Vector Space, 364
Balls in \Re^n , 103
Boundary Points of a Subset in \Re^n , 103
Bounded Sets, 9
Cauchy Sequences in \Re^n , 106
Connected Closed Sets, 347
Connected Sets, 371
Conservative Force Field, 344
Critical Points, 224
Darboux Lower and Upper Sums for f on A , 293
Error Form of Differentiability For Scalar Function of n Variables, 192
Exact 1 - forms, 373
Integration of a 1 -form over a smooth path, 368
Least Upper Bound and Greatest Lower Bound, 9
Limit Inferior and Limit Superior of a Sequence, 14
Limit Points, Cluster Points and Accumulation Points, 104
Linear Independence, 26
Linear Independence in a Vector Space, 26
Linear Independence Of N Objects, 24
Linear Independence of Two Objects, 24
Linear Transformations Between Vector Spaces, 64
Norm Convergence in n Dimensions, 104
Norm on a Vector Space, 56
One Forms on an open set U , 363
Open and Closed Sets in \Re^n , 104
Oriented Two Cells, 354
Partial Derivatives, 187
Partitions of the bounded set A , 292
Path Connected Sets, 347
Planes in n Dimensions, 189
- Real Matrices, 64
Second Order Partials, 202
Sequential Compactness, 11
Sequentially Compact Subsets of \Re^n , 109
Sets of Measure Zero, 306
Smooth Functions on \Re^n , 363
Smooth Paths, 366
Smooth Segmented Paths, 373
Tangent Vector Fields, 363
The ℓ^p Sequence Space, 57
The \Re^n to \Re^m Limit, 112
The Characteristic Function of a Set, 309
The Closure of a Set, 105
The Common Refinement of Two Partitions, 295
The Darboux Integral, 298
The Differentiability of a Vector Function of n variables, 209
The Frobenius Norm of a Linear Transformation, 72
The Gradient, 191
The Hessian Matrix, 203
The Inherited Inner Product on a Finite Dimensional Vector Space is Independent of Choice of Orthonormal Basis, 68
The Integration Symbol, 309
The Jacobian, 208
The Line Integral of a Force Field along a Curve, 341
The Lower and Upper Darboux Integral, 298
The Minimum and Maximum of a Set, 10
The Norm of a Partition, 300
The Norm of a Symmetric Matrix, 120
The Oscillation of a Function, 310
The Riemann Integral, 302
The Riemann Sum, 302
The Symmetric Group S_2 , 376
The Symmetric Group S_3 , 376
The Tangent Plane to $z = f(\mathbf{x})$ at \mathbf{x}_0 in \Re^n , 191
The Volume of a Set, 309
Topologically Compact Subsets of \Re^n , 109
Two Dimensional Curves, 339
Vector Spaces
A Basis For A Finite Dimensional Vector Space, 27

- A Basis For An Infinite Dimensional Vector Space, 28
- Linear Independence For Non Finite Sets, 28
- Real Inner Product, 29
- The Span Of A Set Of Vectors, 27

- Finite Difference Techniques for PDE
 - Approximation for Diffusion Equation, 279
 - Central Difference for First Order, 277
 - Central Difference for Second Order, 278
 - Discretization of Underlying Domain, 279
 - Forward Difference for First Order, 277
 - Recursive System for the Diffusion Equation, 281
 - Recursive System for the Diffusion Equation Errors, 282

- Fourier Series
 - Fourier Series for f , 54
 - Normalized Cosine functions, 53

- Functions of Two Variables
 - Continuity, 275
 - Fourth Order Taylor Expansion, 278
 - Partial Derivative Bounds, 276
 - Second Order Taylor Expansion, 276

- General Autoimmune Models
 - The CMN model
 - Assumption 1: the number of infected cells, 274
 - Deviations of C , M and N from nominal values, 274
 - Dynamics, 274
 - Dynamics: the F_1 , F_2 and F_3 nonlinear functions, 274
 - Linearizing F_1 , F_2 and F_3 , 275
 - The $3 \times 3 F_1$, F_2 and F_3 nominal partial values matrix, 275
 - The initial conditions for the dynamics, 274
 - The CMN model: M and N are two distinct populations of cells and C is autoimmune action, 274

- Immune Theory
 - Assumptions
 - Number of infected cells, 274

- Lemma
 - Infimum Tolerance Lemma, 11
 - Supremum Tolerance Lemma, 11

- Linear ODE Systems
 - Complex Eigenvalues, 84
 - First Real Solution, 85
 - General Complex Solution, 84
 - General Real Solution, 85
 - Matrix Representation of the Real Solution, 86

- Rewriting the Real Solution, 86
- Second Real Solution, 85
- Theoretical Analysis, 84
- Worked Out Example, 85

- MultiVariable Calculus
 - Smoothness
 - Continuous first order partials imply mixed second order partials match, 213

- Theorem, 65, 113, 307, 387
 - $(\mathbb{R}^n, \Gamma B30D \cdot \Gamma B30D_p)$ is Complete, 63
 - D in \mathbb{R}^n is closed if and only if it contains all its limit points., 105
 - D is Sequentially Compact if and only if it is closed and bounded, 109
 - D is topologically compact if and only if it is closed and bounded, 111
 - D is topologically compact if and only if sequentially compact if and only if closed and bounded, 111
 - D is topologically compact implies it is closed and bounded, 110
 - S has a maximal element if and only if $\sup(S) \in S$, 10
 - S has a minimal element if and only if $\inf(S) \in S$, 10
 - \mathbb{R}^n is Complete, 106

- Linear Transformation between Two Finite Dimensional Vector Spaces Determines an Equivalence Class of Matrices, 71

- A D Set is Closure if and only if $D = \overline{D}$, 105
- A Hyperrectangle is topologically compact, 110
- A Particular Solution to a linear ODE System, 398

- A Set S is Sequentially Compact if and only if it is Closed and Bounded, 11
- A Simple Recapture Theorem in \mathbb{R}^2 , 372
- All Eigenvalue, 125
- Alternate Definition of the limit inferior and limit superior of a sequence, 17
- An Extension of the Implicit Function Theorem, 248
- Approximation of the Darboux Integral, 300
- Approximation of the Riemann Integral, 305
- At an Interior Extreme Point, The Partials Vanish, 223
- Basic Topology Results in \mathbb{R}^2 , 19
- Best Finite Dimensional Approximation Theorem, 52
- Bolzano - Weierstrass Theorem in \mathbb{R}^n , 107
- Bolzano - Weierstrass Theorem in 2D, 20
- Cauchy-Schwartz Inequality, 29
- Change of Variable for a Linear Map, 319
- Closed Subsets of Hyperrectangles are topologically compact, 111

- Conservative Force Fields Imply $A_y = B_x$, 346
 Constrained Optimization, 425
 Continuous functions on compact domains have a minimum and maximum value, 13
 Definiteness Test for nD Extrema Using Definiteness Of the Hessian, 229
 Derivative of f at an interior point local extremum is zero, 14
 Determinant, 168
 Algorithm, 169
 Properties, 169
 Smoothness, 169
 Differentiable Implies Continuous: Scalar Function of n Variables, 193
 Equivalent Characterizations of Exactness, 374
 Finding where $(x^2 + y^2, 2xy)$ is a Bijection, 234
 Finite Unions of Sets of Measure Zero Also Have Measure Zero, 307
 First Eigenvalue, 122
 Fubini's Theorem for a Rectangle: n dimensional, 334
 Fubini's Theorem on a Rectangle: 2D, 332
 Fubini's Theorem: 2D: a top and bottom closed curve, 335
 Fubini's Theorem: 2D: a top and bottom curve, 334
 Green's Theorem for a SCROC, 354
 Green's Theorem for Oriented 2 - Cells, 357
 Hölder's Inequality, 57
 Hölder's Inequality in \mathbb{R}^n , 59
 Hölder's Theorem for $p = 1$ and $q = \infty$, 58
 If f continuous on a compact domain, it has global extrema, 117
 If f is continuous on a compact domain, its range is compact, 116
 If f is integrable on A and A has measure zero, then the integral of f on A is zero, 316
 If f is integrable over A and f is zero except on a set of measure zero, this the integral is zero, 317
 If the non negative f is integrable on A with value 0, then the measure of the set of points where $f > 0$ is measure zero, 316
 If the Vector Function f is Differentiable at \mathbf{x}_0 , $L(\mathbf{x}_0) = J_f(\mathbf{x}_0)$, 210
 Implicit Function Theorem, 246
 Interior Points of the Range of f , 238
 Inverse Images of open sets under continuous maps are open, 240
 Lebesgue's Theorem, 311
 Linear Transformations Between Finite Dimensional Vector Spaces, 69
 Lower and Upper Darboux Sums and Partition Refinements, 296
 Lower and Upper Darboux Sums are Independent of the choice of Bounding Rectangle, 294
 Lower Darboux Sums are always less than Upper Darboux Sums, 298
 Minkowski's Inequality, 59
 Minkowski's Inequality in \mathbb{R}^n , 61
 Oscillation is Zero if and only Continuous, 311
 Smooth Functions with zero derivatives are constant if and only if the domain is connected, 371
 Subsets of Sets That have Volume, 324
 Sufficient Conditions for the Mixed Order Partials to Match, 205
 Sufficient Conditions On the First Order Partials to Guarantee Differentiability, 200
 The $\alpha - \beta$ Lemma, 57
 The angle function is not exact on $\mathbb{R}^2 \setminus \{0,0\}$, 373
 The Chain Rule for Scalar Functions of n Variables, 197
 The Chain Rule for Vector Functions, 211
 The Change of Basis Mapping, 67
 The Change of Basis Mapping In a Finite Dimensional Inner Product Space, 67
 The Change of Variable Theorem for a Constant Map, 324
 The Change of Variable Theorem for a General Map, 327
 The Completeness of $(\mathbb{R}^n, \Gamma B30D \cdot \Gamma B30D_\infty)$, 62
 The Derivative of the exponential matrix, 393
 The Eigenvalue Representation of the norm of a symmetric matrix, 126
 The Equivalence of the Riemann and Darboux Integral, 303
 The First Inverse Function Theorem, 237
 The Frobenius Norm Fundamental Inequality, 72
 The general solution to a linear ODE system, 395
 The Inverse Function Theorem, 240
 The Mean Value Theorem, 213
 The Potential Function for F , 348
 The range of a continuous function on a sequentially compact domain is also sequentially compact, 12
 The Recapture Theorem for Segmented Paths, 374
 The Riemann Criterion For Darboux Integrability, 298
 The scalar function f is Differentiable implies $Df = (\nabla(f))^T$, 194
 The Second Eigenvalue, 123

- The Solution Space to a linear System of ODE,
397
- The Third Eigenvalue, 124
- The Volume of a Rectangle in \Re^n , 322
- Two dimensional Extrema Test for minima and
maxima, 228
- Two Equivalent Ways To Calculate the Norm
of a Symmetric Matrix, 121
- Volume Zero Implies Measure Zero, 310