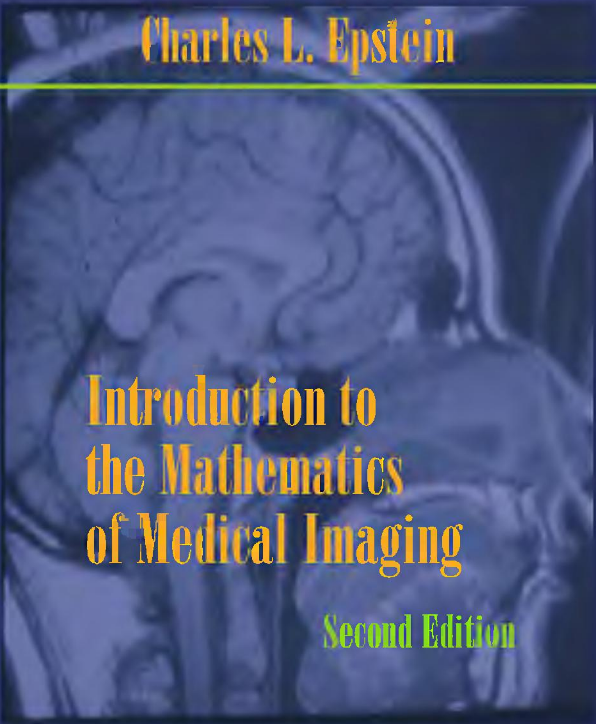
Charles L. Epstein



Introduction to the Mathematics of Medical Imaging

Second Edition

siam.

Introduction to the Mathematics of Medical Imaging

OT102_EpsteinFM1.qxp 8/9/2007 3:27 PM Page 3

Introduction to the Mathematics of Medical Imaging

Second Edition

Charles L. Epstein

University of Pennsylvania
Philadelphia, Pennsylvania

siam.

Society for Industrial and Applied Mathematics • Philadelphia

This SIAM edition is a second edition of the work first published by Pearson Education, Inc., 2003.

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Epstein, Charles L.

Introduction to the mathematics of medical imaging / Charles L. Epstein. -- 2nd ed.

p. cm.

ISBN 978-0-89871-642-9 (alk. paper)

1. Diagnostic imaging--Mathematics. 2. Imaging systems in medicine--Mathematics. 3. Medical physics--Mathematics. I. Title.

RC78.7.D53E676 2008

616.07'54--dc22 2007061801

This book is dedicated to my wife, Jane,
and our children, Leo and Sylvia.

They make it all worthwhile.

Contents

Preface to the second edition

It seems like only yesterday that I was sending the "camera ready" pdf file of this book off to Prentice Hall. Despite a very positive response from the mathematics and engineering communities, Pearson decided, last year, to let the book go out of print. I would like to thank George Lobell, my editor at Prentice Hall, for making it so easy to reacquire the publication rights. Secondly I would like to thank SIAM, and my editors Sarah Granlund and Ann Manning Allen, for making it so easy to prepare this second edition. I would be very remiss if I did not thank Sergei Gelfand, editor at the AMS, who prodded me to get the rights back, so I could prepare a second edition.

The main differences between this edition and the Prentice Hall edition are: 1. A revised section on the relationship between the continuum and discrete Fourier transforms, Section 10.2.2 (reflecting my improved understanding of this problem); 2. A short section on Grangreat's formula, Section 10.2.2, which forms the basis of most of the recent work on cone-beam reconstruction algorithms; 3. A better description of the gridding method, Section 11.8 (many thanks to Leslie Greengard and Jeremy Magland for helping me to understand this properly); 4. A chapter on magnetic resonance imaging, Chapter 14; 5. A short section on noise analysis in MR-imaging, Section 16.3. For the last two items I would like to express my deep gratitude to Felix Wehrli, for allowing me to adapt an article we wrote together for the Elsevier Encyclopedia on Mathematical Physics, and for his enormous hospitality, welcoming me into his research group, the Laboratory for Structural NMR Imaging, at the Hospital of the University of Pennsylvania.

With a bit more experience teaching the course and using the book, I now feel that it is essential for students to have taken at least one semester of undergraduate analysis, beyond calculus, and a semester of linear algebra. Without this level of sophistication, it is difficult to appreciate what all the fuss is about.

I have received a lot of encouragement to prepare this second edition from the many people who used the book, either as a course textbook or for self study. I would like to thank Rafe Mazzeo, Petra Bonfert-Taylor, Ed Taylor, Doug Cochran, John Schotland, Larry Shepp, and Leslie Greengard for their kind words and advice. Finally, I thank my wife, Jane, and our children, Leo and Sylvia, for their forebearance during the endless preparation of the first edition, and their encouragement to produce this second edition.

Charles L. Epstein
May 16, 2007

Preface

Over the past several decades, advanced mathematics has quietly insinuated itself into many facets of our day-to-day life. Mathematics is at the heart of technologies from cellular telephones and satellite positioning systems to online banking and metal detectors. Arguably no technology has had a more positive and profound effect on our lives than medical imaging, and in no technology is the role of mathematics more pronounced or less appreciated. X-ray tomography, ultrasound, positron emission tomography, and magnetic resonance imaging have fundamentally altered the practice of medicine. At the core of each modality is a mathematical model to interpret the measurements and a numerical algorithm to reconstruct an image. While each modality operates on a different physical principle and probes a different aspect of our anatomy or physiology, there is a large overlap in the mathematics used to model the measurements, design reconstruction algorithms, and analyze the effects of noise. In this text we provide a tool kit, with detailed operating instructions, to work on the sorts of mathematical problems that arise in medical imaging. Our treatment steers a course midway between a complete, rigorous mathematical discussion and a cookbook engineering approach.

The target audience for this book is junior or senior math undergraduates with a firm command of multi-variable calculus, linear algebra over the real and complex numbers, and the basic facts of mathematical analysis. Some familiarity with basic physics would also be useful. The book is written in the language of mathematics, which, as I have learned, is quite distinct from the language of physics or the language of engineering. Nonetheless, the discussion of every topic begins at an elementary level and the book should, with a little translation, be usable by advanced science and engineering students with some mathematical sophistication. A large part of the mathematical background material is provided in two appendices.

X-ray tomography is employed as a *pedagogical machine*, similar in spirit to the elaborate devices used to illustrate the principles of Newtonian mechanics. The *physical principles* used in x-ray tomography are simple to describe and require little formal background in physics to understand. This is not the case in any of the other modalities listed nor in less developed modalities like infrared imaging or impedance tomography. The *mathematical* problems that arise in x-ray tomography and the tools used to solve them have a great deal in common with those used in the other imaging modalities. This is why our title is *Introduction to the Mathematics of Medical Imaging* instead of *Introduction to the Mathematics of X-Ray Tomography.* A student with a thorough understanding of the material in this book should be mathematically prepared for further investigations in most subfields of medical imaging.

Very good treatments of the physical principles underlying the other modalities can be found in *Radiological Imaging* by Harrison H. Barrett and William Swindell, [6], *Principles of Computerized Tomographic Imaging* by Avinash C. Kak and Malcolm Slaney, [76], *Foundations of Medical Imaging* by Cho, Jones, Singh, [22], *Image Reconstruction from Projections* by Gabor T. Herman, [52], and *Magnetic Resonance Imaging* by E. Mark Haacke, Robert W. Brown, Michael R. Thompson, Ramesh Venkatesan, [50]. Indeed these books were invaluable sources as I learned the subject myself. My treatment of many topics owes a great deal to these books as well as to the papers of Larry Shepp and Peter Joseph and their collaborators. More advanced treatments of the mathematics and algorithms introduced here can be found in *The Mathematics of Computerized Tomography* by Frank Natterer, [95], and *Mathematical Methods in Image Reconstruction* by Frank Natterer and Frank Wübbelling, [96].

The order and presentation of topics is somewhat nonstandard. The organizing principle of this book is the evolutionary development of an accurate and complete model for x-ray tomography. We start with a highly idealized mathematical model for x-ray tomography and work toward more realistic models of the actual data collected and the algorithms used to reconstruct images. After some preliminary material we describe a continuum, complete data model phrased in terms of the Radon transform. The Fourier transform is introduced as a tool, first to invert the Radon transform and subsequently for image processing. The first refinement of this model is to take account of the fact that real data are always sampled. This entails the introduction of Fourier series, sampling theory, and the finite Fourier transform. After introducing terminology and concepts from filtering theory, we give a detailed synthesis of the foregoing ideas by describing how continuum, shift invariant, linear filters are approximately implemented on finitely sampled data. With these preliminaries in hand, we return to the study of x-ray tomography, per se. Several designs for x-ray computed tomography machines are described, after which we derive the corresponding implementations of the filtered back-projection algorithm. At first we assume that the x-ray beam is one dimensional and monochromatic. Subsequently we analyze the effects of a finite width beam and various sorts of measurement and modeling errors. The last part of the book is concerned with noise analysis. The basic concepts of probability theory are reviewed and applied to problems in imaging. The notion of signal-to-noise ratio (SNR) is introduced and used to analyze the effects of quantum noise on images reconstructed using filtered back-projection. A maximum likelihood algorithm for image reconstruction in positron emission tomography is described. The final chapter introduces the idea of a random process. We describe the random processes commonly encountered in imaging and an elementary example of an optimal filter. We conclude with a brief analysis of noise in the continuum model of filtered back-projection.

The book begins with an introduction to the idea of using a mathematical model as a tool to extract the physical state of system from feasible measurements. In medical imaging, the "state of the system" in question is the anatomy and physiology of a *living* human being. To probe it nondestructively requires considerable ingenuity and sophisticated mathematics. After considering a variety of examples, each a toy problem for some aspect of medical imaging, we turn to a description of x-ray tomography. This leads us to our first mathematical topic, *integral transforms*. The transform of immediate interest is the Radon transform, though we are quickly led to the Abel transform, Hilbert transform, and Fourier transform. Our study of the Fourier transform is dictated by the applications we have in mind, with a strong emphasis on the connection between the smoothness of a function and the decay of its Fourier transform and vice versa. Many of the basic ideas of functional analysis appear as we consider these examples. The concept of a weak derivative, which is ubiquitous in the engineering literature and essential to a precise understanding of the Radon inversion formula, is introduced. This part of the book culminates in a study of the Radon inversion formula. A theme in these chapters is the difference between finite- and infinite-dimensional linear algebra.

The next topics we consider are Fourier series, sampling, and filtering theory. These form the basis for applying the mathematics of the Fourier transform to real-world problems. Chapter 8 is on sampling theory; we discuss the Nyquist theorem, the Shannon–Whittaker interpolation formula, the Poisson summation formula, and the consequences of undersampling. In Chapter 9, on filtering theory, we recast Fourier analysis as a tool for image and signal processing. The chapter concludes with an overview of image processing and a linear systems analysis of some basic imaging hardware. We then discuss the mathematics of approximating continuous time, linear shift invariant filters on finitely sampled data, using the finite Fourier transform.

In Chapters 11 and 12 the mathematical tools are applied to the problem of image reconstruction in x-ray tomography. These chapters are largely devoted to the filtered back-projection algorithm, though other methods are briefly considered. After deriving the reconstruction algorithms, we analyze the point spread function and modulation transfer function of the full measurement and reconstruction process. We use this formalism to analyze a variety of imaging artifacts. Chapter 13 contains a brief description of "algebraic reconstruction techniques," which are essentially methods for solving large, sparse systems of linear equations.

The final topic is noise in the filtered back-projection algorithm. This part of the book begins with an introduction to probability theory. Our presentation uses the language and ideas of measure theory, in a metaphoric rather than a technical way. Chapter 15 concludes with a study of specific probability distributions that are important in imaging. In Chapter 16 we apply probability theory to a variety of problems in medical imaging. This chapter includes the famous resolution-dosage fourth power relation, which shows that to double the resolution in a CT image, keeping the SNR constant, the radiation dosage must be increased by a factor of 16! The chapter ends with an introduction to positron emission tomography and the maximum likelihood algorithm. Chapter 17 introduces the ideas of random processes and their role in signal and image processing. Again the focus is on those processes needed to analyze noise in x-ray imaging. A student with a good grasp of Riemann integration should not have difficulty with the material in these chapters.

Acknowledgments

Perhaps the best reward for writing a book of this type is the opportunity it affords for thanking the many people who contributed to it in one way or another. There are a lot of people to thank, and I address them in roughly chronological order.

First I would like to thank my parents, Jean and Herbert Epstein, for their encouragement to follow my dreams and the very high standards they set for me from earliest childhood. I would also like to thank my father and Robert M. Goodman for imparting the idea that observation and careful thought go a long way toward understanding the world in which we live. Both emphasized the importance of expressing ideas simply but carefully.

My years as an undergraduate at the Massachusetts Institute of Technology not only provided me with a solid background in mathematics, physics, and engineering but also a belief in the unity of scientific enquiry. I am especially grateful for the time and attention Jerry Lettvin lavished on me. My interest in the intricacies of physical measurement surely grew out of our many conversations. I was fortunate to be a graduate student at the Courant Institute, one of the few places where "pure" and "applied" mathematics lived together in harmony. In both word and deed, my thesis advisor, Peter Lax, placed mathematics and its applications on an absolutely equal footing. It was a privilege to be his student. I am very grateful for the enthusiasm that he and his late wife, Anneli, showed for turning my lecture notes into a book.

I would like to acknowledge my friends and earliest collaborators in the enterprise of becoming a scientist—Robert Indik and Carlos Tomei. I would also thank my friends and current collaborators, Gennadi Henkin and Richard Melrose, for the vast wealth of culture and knowledge they have shared with me and their forbearance, while I have been "missing in action."

Coming closer to the present day, I would like to thank Dennis Deturck for his unflagging support, both material and emotional, for the development of my course on medical imaging and this book. Without the financial support he and Jim Gee provided for the spring of 2002, I could not have finished the manuscript. The development of the original course was supported in part by National Science Foundation grant DUE95-52464. I would like to thank Dr. Ben Mann, my program director at the National Science Foundation, for providing both moral and financial support. I am very grateful to Hyunsuk Kang, who transcribed the first version of these notes from my lectures in the spring of 1999. Her hard work provided the solid foundation on which the rest was built. I would also like to thank the students who attended Math 584 in 1999, 2000, 2001, and 2002 for their invaluable input on the content of the course and their numerous corrections to earlier versions of this book. I would also like to thank Paula Airey for flawlessly producing what must have seemed like endless copies of preliminary versions of the book.

John D'Angelo read the entire text and provided me with an extremely useful critique as well as a lot of encouragement. Chris Croke, my colleague at the University of Pennsylvania, also carefully read much of the manuscript while teaching the course and provided many corrections. I would like to thank Phil Nelson for his help with typesetting, publishing, and the writing process and Fred Villars for sharing with me his insights on medicine, imaging, and a host of other topics.

The confidence my editor, George Lobell, expressed in the importance of this project was a strong impetus for me to write this book. Jeanne Audino, my production editor, and Patricia M. Daly, my copy editor, provided the endless lists of corrections that carried my manuscript from its larval state as lecture notes to the polished book that lies before you. I am most appreciative of their efforts

I am grateful to my colleagues in the Radiology Department—Dr. Gabor Herman, Dr. Peter Joseph, Dr. David Hackney, Dr. Felix Wehrli, Dr. Jim Gee, and Brian Avants—for sharing with me their profound, first-hand knowledge of medical imaging. Gabor Herman's computer program, *SNARK93*®, introduced me to the practical side of image reconstruction and was used to make some of the images in the book. I used Dr. Kevin Rosenberg's program *ctsim* to make many other images. I am very grateful for the effort he expended to produce a version of his marvelous program that would run on my computer. David Hackney provided beautiful, state-of-the-art images of brain sections. Felix Wehrli provided the image illustrating aliasing in magnetic resonance imaging and the x-ray CT micrograph of trabecular bone. I am very grateful to Peter Joseph for sharing his encyclopedic first-hand knowledge of x-ray tomography and its development as well as his treasure trove of old, artifact-ridden CT images. Jim Gee and Brian Avants showed me how to use MATLAB® for image processing. Rob Lewitt provided some very useful suggestions and references to the literature.

I would also like the acknowledge my World Wide Web colleagues. I am most appreciative for the x-ray spectrum provided by Dr. Andrew Karellas of the University of Massachusetts, the chest x-ray provided by Drs. David S. Feigen and James Smirniotopoulos of the Uniformed Services University, and the nuclear magnatic resonance spectrum (NMR) provided by Dr. Walter Bauer of the Erlangen-Nürnberg Univerisity. Dr. Bergman of the Virtual Hospital at the University of Iowa (www.vh.org) provided an image of an anatomical section of the brain. The World Wide Web is a resource for imaging science of unparalleled depth, breadth, and openness.

Finally and most of all I would like to thank my wife, Jane, and our children, Leo and Sylvia, for their love, constant support, and daily encouragement through the many, many, moody months.

Charles L. Epstein
Philadelphia, PA
January 8, 2003

How to Use This Book

This chapter is strongly recommended for all readers.

The structure of this book: It is unlikely that you will want to read this book like a novel, from beginning to end. The book is organized somewhat like a hypertext document. Each chapter builds from elementary material to more advanced material, as is also the case with the longer sections. The elementary material in each chapter (or section) depends only on the elementary material in previous chapters. The more advanced material can be regarded as "links" that you may or may not want to follow. Each chapter is divided into sections, labeled $c.s$, where c is the chapter number and s, the section number. Many sections are divided into subsections, labeled $c.s.ss$, with ss the subsection number. Sections, subsections, examples, exercises, and so on that are *essential* for later parts of the book are marked with a star: $\star$. These should not be skipped.

Sections, subsections, exercises, and so on that require more prerequisites than are listed in the preface are marked with an asterisk: $*$. These sections assume a familiarity with the elementary parts of functions of a complex variable or more advanced real analysis. The asterisk is inherited by the subparts of a marked part. The sections without an asterisk do not depend on the sections with an asterisk. All sections with an asterisk can be omitted without any loss in continuity. The book has two appendices. Appendix A contains background material on a variety of topics ranging from very elementary things like numbers and their representation in computers to more advanced topics like spaces of function and approximation theory. This material serves several purposes: It compiles definitions and results used in the book, and provides a larger context for some material and references for further study. Appendix B is a quick review of the definitions and theorems usually encountered in an undergraduate course in mathematical analysis. Many sections begin with a box:

See: A.1, B.3, B.4, B.5.

containing a list of recommended background readings drawn from the appendices and earlier sections.

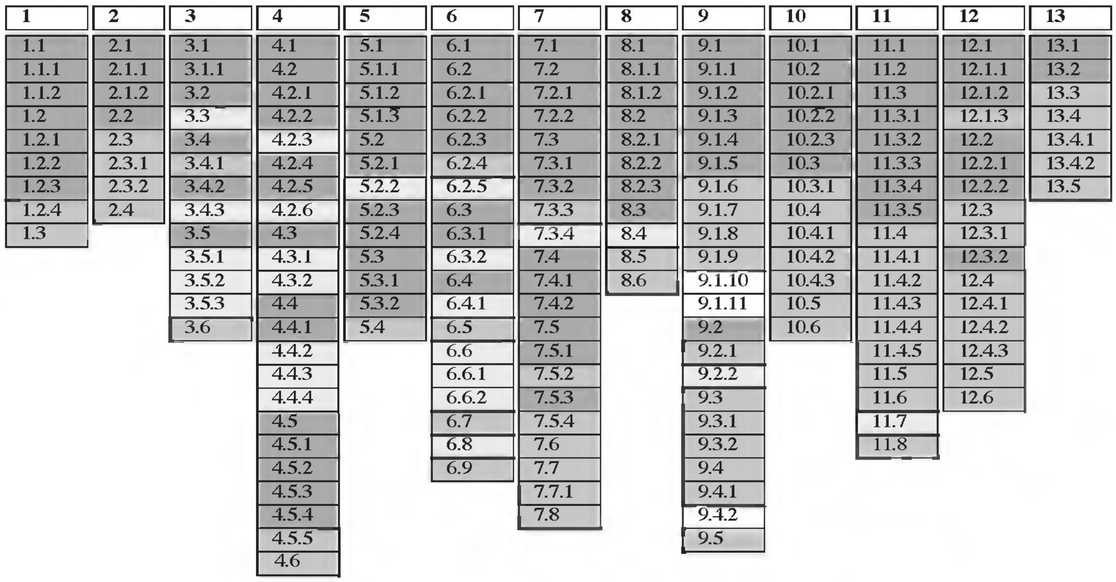
For the student: This book is wide ranging, and it is expected that students with a variety of backgrounds will want to use it. Nonetheless, choices needed to be made regarding notation and terminology. In most cases I use notation and terminology that is standard in the world of mathematics. If you come on a concept or term you do not understand, you should first look in *List of Notations* at the beginning of the book and then in the Index at the end. There are many exercises and examples scattered throughout the text. Doing the exercises and working through the details of the examples will teach you a lot more than simply reading the text. The book does not contain many problems involving machine computation; nonetheless I strongly recommend implementing as many things as you can using MATLAB®, Maple®, and so on.

The World Wide Web is a fantastic resource for medical imaging, with many sites devoted to images and the physics behind imaging. At www.ctsim.org you will find an excellent program, written by Dr. Kevin Rosenberg, that simulates the data collection, processing and postprocessing done by variety of x-ray CT machines. The pictures you will produce with *ctsim* are surely worth thousands of words! A list of additional imaging Web sites can be found at the end of Chapter 11. I hope you enjoy this book and get something out of it. I would be most grateful to hear about which parts of the book you liked and which parts you feel need improvement. Contact me at: cle@math.upenn.edu.

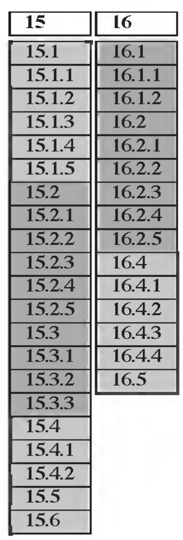
For the instructor: Teaching from this book requires careful navigation. I have taught several courses from the notes that grew into this book. Sometimes the students had a mathematical bent, sometimes a physical bent, and sometimes an engineering bent. I learned that the material covered *and* its presentation must be tailored to the audience. For example, proofs or outlines of proofs are provided for most of the mathematical results. These should be covered in detail for math students and perhaps assigned as reading or skipped for engineering students. How far you go into each topic should depend on how much time you have and the composition of the class. If you are unfamiliar with the practicalities of medical imaging I would strongly recommend that you read the fundamental articles [113], [114], and [55] as well as the parts of [76] devoted to x-ray tomography.

In the following section I have provided suggestions for one- and two-semester courses with either a mathematical or an engineering flavor (or student body). These guidelines should make it easier to fashion coherent courses out of the wide range of material in the text. Exercises are collected at the ends of the sections and subsections. Most develop ideas presented in the text; only a few are of a standard, computational character. I would recommend using computer demonstrations to illustrate as many of the ideas as possible and assigning exercises that ask the students to implement ideas in the book as computer programs. I would be most grateful if you would share your exercises with me, for inclusion in future editions. As remarked previously, Dr. Kevin Rosenberg's program *ctsim*, which is freely available at www.ctsim.org, can be used to provide students with irreplaceable hands-on experience in reconstructing images.

Some suggestions for courses: In the following charts material shaded in *dark gray* forms the basis for classroom presentation; that shaded in *medium gray* can be assigned as reading or used as additional material in class. The *light gray* topics are more advanced enrichment material; if the students have adequate background, some of this material can be selected for presentation in class. Students should be encouraged to read the sections in the appendices listed in the boxes at the beginnings of the sections. I have not recommended spending a lot of class time on the higher-dimensional generalizations of the Fourier transform, Fourier series, and so on. One is tempted to say that "things go just the same." I would instead give a brief outline in class and assign the appropriate sections in the book, *including some exercises*, for home study.



For classes with a good background in probability theory, the material in Chapters 12 and 13 after 12.1.1 could be replaced by the indicated material in Chapters 15 and 16. For this to work, the material in Section 15.2 should be largely review.



Color code:

Material covered in class (dark)

Additional assigned reading (medium)

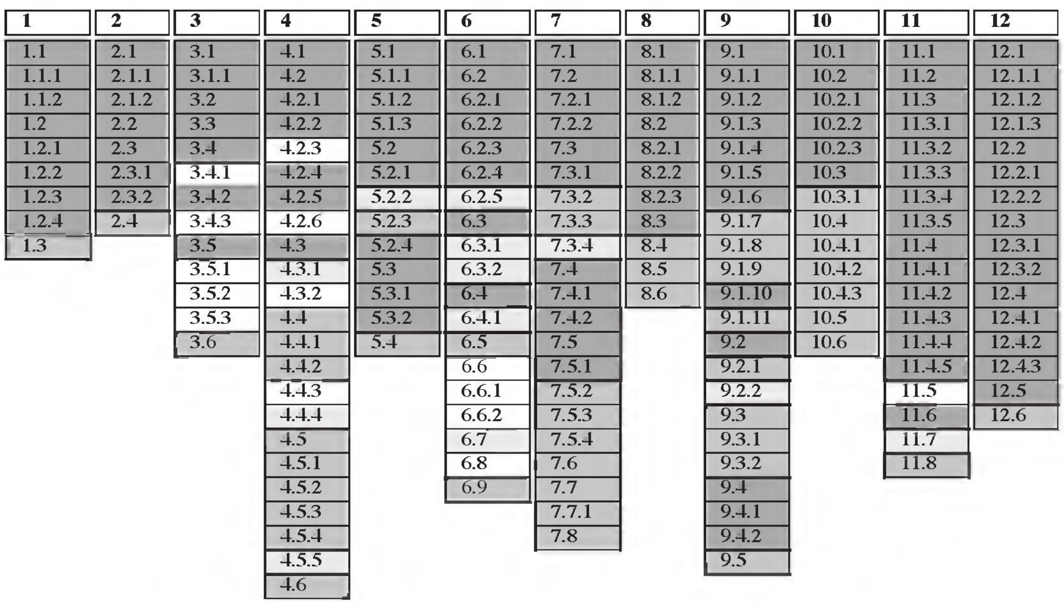
More advanced enrichment material (light)

Figure 1. An outline for a one-semester course with a mathematical emphasis.

The chart in Figure 1 outlines a one semester course with an emphasis on the mathematical aspects of the subject. The dark gray sections assume a background in undergraduate analysis and linear algebra. If students have better preparation, you may want to select some of the light gray topics for classroom presentation.

If at all possible, I would recommend going through the section on generalized functions, A.4.5, when introducing the concept of weak derivatives in Section 4.3. Much of the subsequent red material assumes a familiarity with these ideas. As noted, if students have a good background in measure and integration or probability, you may want to skip the material after 12.1.1 and go instead to the analysis of noise, beginning in Chapter 14.

The chart in Figure 2 outlines a one semester course with an emphasis on the engineering aspects of the subject. For such a class the material in this book could be supplemented with more applied topics taken from, for example, [76] or [6]. The proofs of mathematical results can be assigned as reading or skipped.



Due to time constraints, it may be necessary to choose one of 9.4.1 or 9.4.2 and leave the material not covered in class as a reading assignment.

Figure 2. An outline for a one-semester course with an engineering emphasis.

The chart in Figure 2 gives suggestions for a full-year course with a mathematical emphasis. A full year should afford enough time to introduce generalized functions in Section A.4.5. This should be done along with the material in Section 4.3. This allows inclusion of Section 4.4.4, which covers the Fourier transform on generalized functions.

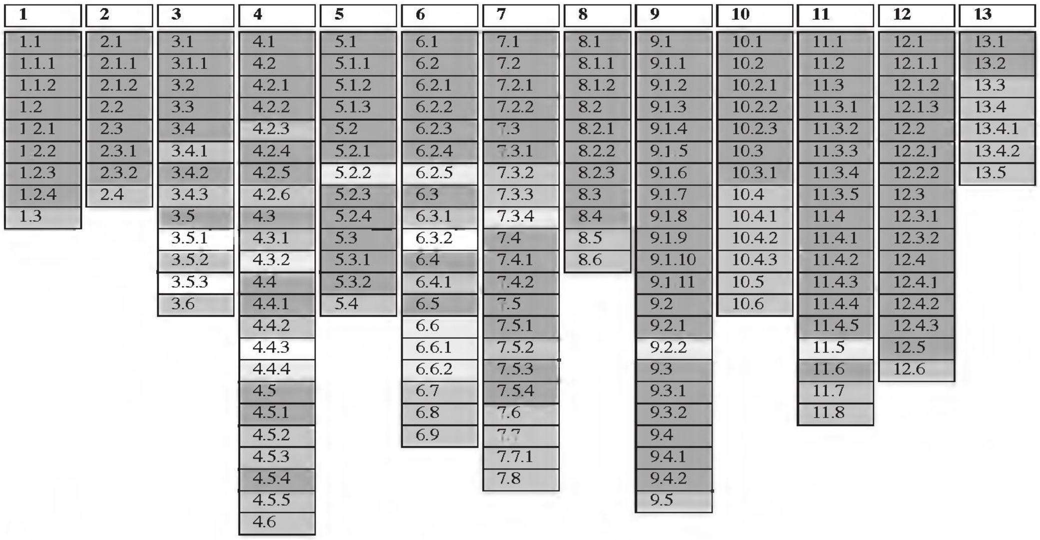
1	2	3	4	5	6	7	8	9	10	11	12	13
1.1	2.1	3.1	4.1	5.1	6.1	7.1	8.1	9.1	10.1	11.1	12.1	13.1
1.1.1	2.1.1	3.1.1	4.2	5.1.1	6.2	7.2	8.1.1	9.1.1	10.2	11.2	12.1.1	13.2
1.1.2	2.1.2	3.2	4.2.1	5.1.2	6.2.1	7.2.1	8.1.2	9.1.2	10.2.1	11.3	12.1.2	13.3
1.2	2.2	3.3	4.2.2	5.1.3	6.2.2	7.2.2	8.2	9.1.3	10.2.2	11.3.1	12.1.3	13.4
1.2.1	2.3	3.4	4.2.3	5.2	6.2.3	7.3	8.2.1	9.1.4	10.2.3	11.3.2	12.2	13.4.1
1.2.2	2.3.1	3.4.1	4.2.4	5.2.1	6.2.4	7.3.1	8.2.2	9.1.5	10.3	11.3.3	12.2.1	13.4.2
1.2.3	2.3.2	3.4.2	4.2.5	5.2.2	6.2.5	7.3.2	8.2.3	9.1.6	10.3.1	11.3.4	12.2.2	13.5
1.2.4	2.4	3.4.3	4.2.6	5.2.3	6.3	7.3.3	8.3	9.1.7	10.4	11.3.5	12.3	
1.3		3.5	4.3	5.2.4	6.3.1	7.3.4	8.4	9.1.8	10.4.1	11.4	12.3.1	
		3.5.1	4.3.1	5.3	6.3.2	7.4	8.5	9.1.9	10.4.2	11.4.1	12.3.2	
		3.5.2	4.3.2	5.3.1	6.4	7.4.1	8.6	9.1.10	10.4.3	11.4.2	12.4	
		3.5.3	4.4	5.3.2	6.4.1	7.4.2		9.1.11	10.5	11.4.3	12.4.1	
		3.6	4.4.1	5.4	6.5	7.5		9.2	10.6	11.4.4	12.4.2	
			4.4.2		6.6	7.5.1		9.2.1		11.4.5	12.4.3	
			4.4.3		6.6.1	7.5.2		9.2.2		11.5	12.5	
			4.4.4		6.6.2	7.5.3		9.3		11.6	12.6	
			4.5		6.7	7.5.4		9.3.1		11.7		
			4.5.1		6.8	7.6		9.3.2		11.8		
			4.5.2		6.9	7.7		9.4				
			4.5.3			7.7.1		9.4.1				
			4.5.4			7.8		9.4.2				
			4.5.5					9.5				
			4.6									

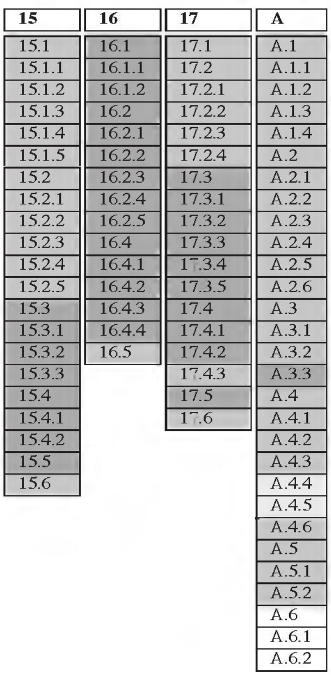
15	16	17	A	B
15.1	16.1	17.1	A.1	B.1
15.1.1	16.1.1	17.2	A.1.1	B.2
15.1.2	16.1.2	17.2.1	A.1.2	B.3
15.1.3	16.2	17.2.2	A.1.3	B.4
15.1.4	16.2.1	17.2.3	A.1.4	B.5
15.1.5	16.2.2	17.2.4	A.2	B.6
15.2	16.2.3	17.3	A.2.1	B.7
15.2.1	16.2.4	17.3.1	A.2.2	B.8
15.2.2	16.2.5	17.3.2	A.2.3	
15.2.3	16.4	17.3.3	A.2.4	
15.2.4	16.4.1	17.3.4	A.2.5	
15.2.5	16.4.2	17.3.5	A.2.6	
15.3	16.4.3	17.4	A.3	
15.3.1	16.4.4	17.4.1	A.3.1	
15.3.2	16.5	17.4.2	A.3.2	
15.3.3		17.4.3	A.3.3	
15.4		17.5	A.4	
15.4.1		17.6	A.4.1	
15.4.2			A.4.2	
15.5			A.4.3	
15.6			A.4.4	
			A.4.5	
			A.4.6	
			A.5	
			A.5.1	
			A.5.2	
			A.6	
			A.6.1	
			A.6.2	

The sections in the appendices, indicated in dark gray, should be presented in class when they are referred to in the boxes at the starts of sections. If at all possible, Section A.4.5, on generalized functions, should be done while doing Section 4.3, thus allowing the inclusion of 4.4.4.

Figure 3. An outline for a one-year course with a mathematical emphasis.

The chart in Figure 3 gives suggestions for a full-year course with an engineering emphasis. As before, this material should be supplemented with more applied material coming from, for example [76] or [6].





Either 9.4.1 or 9.4.2 should be presented in class with the other assigned as reading. If at all possible, the material in A.4.4 and A.4.5 should be presented in class. The medium gray sections of Chapters 15 and 17 should be outlined in class and assigned as reading. Section A.3.3 should be presented when the material is called for in the text.

Figure 4. An outline for a one-year course with an engineering emphasis.

Notational Conventions

DEFINITIONS: $A \stackrel{d}{=} B$ the expression appearing on the right defines the symbol on the left.

SETS: $\{A : P\} \stackrel{d}{=}$ the set of elements A satisfying property P.

CARTESIAN PRODUCT: $A \times B \stackrel{d}{=}$ the set of *ordered* pairs (a, b), where a belongs to the set A and b to the set B.

REPEATED CARTESIAN PRODUCT: $A^n \stackrel{d}{=}$ the set of ordered n-tuples $(a_1, \ldots, a_n)$, where the a_i belong to the set A.

THE NATURAL NUMBERS: $\mathbb{N}$.

THE INTEGERS: $\mathbb{Z}$.

THE RATIONAL NUMBERS: $\mathbb{Q}$.

THE REAL NUMBERS: $\mathbb{R}$.

THE COMPLEX NUMBERS: $\mathbb{C}$.

MULTI-INDEX NOTATION: If $\boldsymbol{j} = (j_1, \ldots, j_n)$ is an n-vector of integers and $\boldsymbol{x} = (x_1, \ldots, x_n)$ is an n-vector, then

$$\boldsymbol{x}^{\boldsymbol{j}} \stackrel{d}{=} x_1^{j_1} \cdots x_n^{j_n}.$$

THE GREATEST INTEGER FUNCTION: For a real number x, $[x] \stackrel{d}{=}$ the largest integer smaller than or equal to x.

INTERVALS: If a and b are real numbers with $a \leq b$, then

$$(a, b) \stackrel{d}{=} \{x \in \mathbb{R} : a < x < b\} \qquad \text{an open interval,}$$

$$[a, b) \stackrel{d}{=} \{x \in \mathbb{R} : a \leq x < b\} \qquad \text{a half-open interval,}$$

$$(a, b] \stackrel{d}{=} \{x \in \mathbb{R} : a < x \leq b\} \qquad \text{a half-open interval,}$$

$$[a, b] \stackrel{d}{=} \{x \in \mathbb{R} : a \leq x \leq b\} \qquad \text{a closed interval.}$$

$$(-\infty, \infty) \stackrel{d}{=} \mathbb{R},$$

$$(a, \infty) \stackrel{d}{=} \{x \in \mathbb{R} : a < x < \infty\} \quad \text{an open positive half-ray,}$$

$$[a, \infty) \stackrel{d}{=} \{x \in \mathbb{R} : a \le x < \infty\} \quad \text{a closed positive half-ray,}$$

$$(-\infty, b) \stackrel{d}{=} \{x \in \mathbb{R} : -\infty < x < b\} \quad \text{an open negative half-ray,}$$

$$(-\infty, b] \stackrel{d}{=} \{x \in \mathbb{R} : -\infty < x \le b\} \quad \text{a closed negative half-ray.}$$

THE EUCLIDEAN NORM: $\|(x_1, \dots, x_n)\|_2 \stackrel{d}{=} \sqrt{\sum_{j=1}^n x_j^2}$.

THE n-DIMENSIONAL UNIT SPHERE: $S^n \stackrel{d}{=} \{\boldsymbol{x} \in \mathbb{R}^{n+1} : \|\boldsymbol{x}\|_2 = 1\}$.

BALLS IN $\mathbb{R}^n$: If $\boldsymbol{a} \in \mathbb{R}^n$ and r is a positive number then

$$B_r(\boldsymbol{a}) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{a}\| < r\}.$$

The ball of radius r centered at $(0, \dots, 0)$ is often denoted B_r.

SEQUENCES: $< \cdot >$ angle brackets enclose the elements of a sequence, for example, $< x_n >$.

INNER PRODUCTS: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is the inner product of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

KRONECKER DELTA: $\delta_{ij} \stackrel{d}{=}$ a square matrix with $\delta_{ii} = 1$ and $\delta_{ij} = 0$ if $i \neq j$.

CONTINUOUS FUNCTIONS: $\mathscr{C}^0(A) \stackrel{d}{=}$ the set of continuous functions defined on the set A.

DIFFERENTIABLE FUNCTIONS: $\mathscr{C}^k(A) \stackrel{d}{=}$ the set of k-times continuously differentiable functions defined on the set A.

SMOOTH FUNCTIONS: $\mathscr{C}^\infty(A) \stackrel{d}{=}$ the set of infinitely differentiable functions defined on the set A.

DERIVATIVES: If f is a function of the variable x then the first derivative is denoted, variously by

$$f', \quad \partial_x f, \text{ or } \quad \frac{df}{dx}.$$

The jth derivative is denoted by

$$f^{[j]}, \quad \partial_x^j f, \text{ or } \quad \frac{d^j f}{dx^j}.$$

INTEGRABLE FUNCTIONS: $L^1(A) \stackrel{d}{=}$ the set of absolutely integrable functions defined on the set A.

SQUARE-INTEGRABLE FUNCTIONS: $L^2(A) \stackrel{d}{=}$ the set of functions, defined on the set A, whose square is absolutely integrable.

"BIG OH" NOTATION: A function f is $O(g(x))$ for x near to x_0 if there is an $\epsilon > 0$ and a constant M so that

$$|f(x)| \leq Mg(x) \quad \text{if } |x - x_0| < \epsilon.$$

"LITTLE OH" NOTATION: A function f is $o(g(x))$ for x near to x_0 if

$$\lim_{x \to x_0} \frac{|f(x)|}{g(x)} = 0.$$

Chapter 1

Measurements and Modeling

quantitative model of a physical system is expressed in the language of mathematics. A qualitative model often precedes a quantitative model. For many years clinicians used medical x-ray images without employing a precise quantitative model. X-rays were thought of as high frequency 'light' with three very useful properties:

1. If x-rays are incident on a human body, some fraction of the incident radiation is absorbed or scattered, though a sizable fraction is transmitted. The fraction absorbed or scattered is proportional to the total density of the material encountered. The overall decrease in the intensity of the x-ray beam is called *attenuation*.

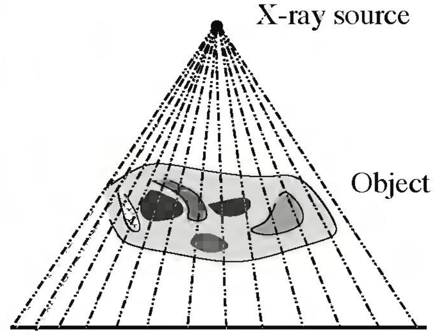
2. A beam of x-ray light travels in a straight line.

3. X-rays darken photographic film. The opacity of the film is a monotone function of the incident energy.

Taken together, these properties mean that using x-rays one can "see through" a human body to obtain a shadow or projection of the internal anatomy on a sheet of film [Figure 1.1(a)].

This model was adequate given the available technology. In their time, x-ray images led to a revolution in the practice of medicine because they opened the door to non-invasive examination of internal anatomy. They are still useful for locating bone fractures, dental caries, and foreign objects, but their ability to visualize soft tissues and more detailed anatomic structure is limited. There are several reasons for this. The principal difficulty is that an x-ray image is a two-dimensional representation of a three-dimensional object. In Figure 1.1(b), the opacity of the film at a point on the film plane is inversely proportional to an average of the density of the object, measured along the line joining the point to the x-ray source. This renders it impossible to deduce the spatial ordering in the missing third dimension.

(a) A old-fashioned chest x-ray image. (Image provided courtesy of Dr. David S. Feigin, ENS Sherri Rudinsky, and Dr. James G. Smirniotopoulos of the Uniformed Services University of the Health Sciences, Dept. of Radiology, Bethesda, MD.)



(b) Depth information is lost in a projection.

Figure 1.1. The world of old fashioned x-rays imaging.

A second problem is connected to the "detector" used in traditional x-ray imaging. Photographic film is used to record the total energy in the x rays that are transmitted through the object. Unfortunately film is rather insensitive to x rays. To get a usable image, a light emitting phosphor is sandwiched with the film. This increases the sensitivity of the overall "detector," but even so, large differences in the intensity of the incident x-ray beam produce small differences in the opacity of film. This means that the contrast between different soft tissues is poor. Beyond this there are other problems caused by the scattering of x rays and noise. Because of these limitations a qualitative theory was adequate for the interpretation of traditional x-ray images.

A desire to improve upon this situation led Alan Cormack, [24], and Godfrey Hounsfield, [64], to independently develop x-ray *tomography* or slice imaging. The first step in their work was to use a quantitative theory for the attenuation of x-rays. Such a theory already existed and is little more than a quantitative restatement of (1) and (2). It is not needed for old fashioned x-ray imaging because traditional x-ray images are read "by eye," and no further processing is done after the film is developed. Both Cormack and Hounsfield realized that mathematics could be used to infer three-dimensional anatomic structure from a large collection of *different* two-dimensional projections. The possibility for making this idea work relied on two technological advances:

1. The availability of scintillation crystals to use as detectors
2. Powerful, digital computers to process the tens of thousands of measurements needed to form a usable image

A detector using a scintillation crystal is about a hundred times more sensitive than photographic film. Increasing the dynamic range in the basic measurements makes possible much finer distinctions. As millions of arithmetic operations are needed for each image, fast computers are a necessity for reconstructing an image from the available measurements. It is an interesting historical note that the mathematics underlying x-ray tomography was done in 1917 by Johan Radon, [105]. It had been largely forgotten, and both Hounsfield and Cormack worked out solutions to the problem of reconstructing an image from its projections. Indeed, this problem had arisen and been solved in contexts as diverse as radio astronomy and statistics.

This book is a detailed exploration of the mathematics that underpins the reconstruction of images in x-ray tomography. While our emphasis is on understanding these mathematical foundations, we constantly return to the practicalities of x-ray tomography. Of particular interest is the relationship between the mathematical treatment of a problem and the realities of numerical computation and physical measurement. There are many different imaging *modalities* in common use today, such as x-ray computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), ultrasound, optical imaging, and electrical impedance imaging. Because each relies on a different physical principle, each provides different information. In every case the mathematics needed to process and interpret the data has a large overlap with that used in x-ray CT. We concentrate on x-ray CT because of the simplicity of the physical principles underlying the measurement process. Detailed descriptions of the other modalities can be found in [91], [76], or [6].

Mathematics is the language in which any quantitative theory or model is eventually expressed. In this introductory chapter we consider a variety of examples of physical systems, measurement processes, and the mathematical models used to describe them. These models illustrate different aspects of more complicated models used in medical imaging. We define the notion of degrees of freedom and relate it to the dimension of a vector space. The chapter concludes by analyzing the problem of reconstructing a region in the plane from measurements of the shadows it casts.

1.1 Mathematical Modeling

The first step in giving a mathematical description of a system is to isolate that system from the universe in which it sits. While it is no doubt true that a butterfly flapping its wings in Siberia in midsummer will affect the amount of rainfall in the Amazon rain forest a decade hence, it is surely a tiny effect, impossible to accurately quantify. A practical model includes the system of interest and the *major* influences on it. Small effects are ignored, though they may come back, as measurement error and noise, to haunt the model. After the system is isolated, we need to find a collection of numerical parameters that describe its state. In this generality these parameters are called *state variables*. In the idealized world of an isolated system the exact measurement of the state variables uniquely determines the state of the system. It may happen that the parameters that give a convenient description of the system are not directly measurable. The mathematical model then describes relations among the state variables. Using these relations, the state of the system can often be determined from feasible measurements. A simple example should help clarify these abstract-sounding concepts.

***Example* 1.1.1.** Suppose the system is a ball on a rod. For simplicity we assume that the ball has radius zero. The state of the system is described by (x, y), the coordinates of the ball. These are the state variables. If the rod is of length r and one end of it is fixed at the point $(0, 0)$, then the state variables satisfy the relation

$$x^2 + y^2 = r^2. \tag{1.1}$$

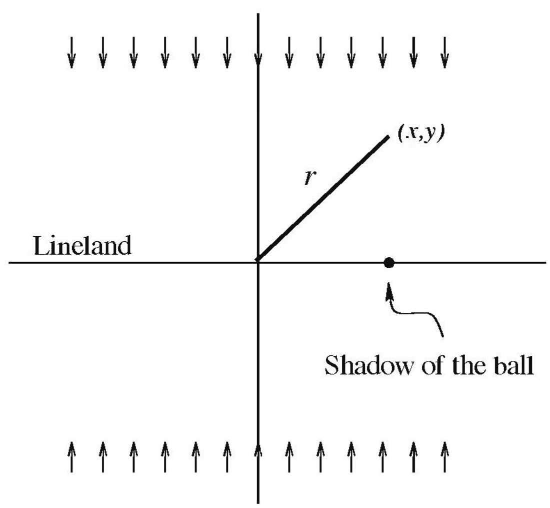


Figure 1.2. A rod of length r casts a shadow on lineland.

Imagine now that one-dimensional creatures, living on the x-axis $\{y = 0\}$, can observe a shadow of the ball, cast by very distant light sources so that the rays of light are perpendicular to the x-axis (Figure 1.2). The line creatures want to predict whether or not the ball is about to collide with their world. Locating the shadow determines the x-coordinate of the ball and using equation (1.1) gives

$$y = \pm\sqrt{r^2 - x^2}.$$

To determine the sign of the y-coordinate requires additional information not available in the model. On the other hand, this information is adequate if one only wants to predict if the ball is about to collide with the x-axis. If the x-axis is illuminated by red light from above and blue light from below, then a ball approaching from below would cast of red shadow while a ball approaching from above would cast a blue shadow. With these additional data, the location of the ball is completely determined.

Ordered pairs of real numbers, $\{(x, y)\}$, are the state variables for the system in Example 1.1.1. Because of the constraint (1.1), not every pair defines a state of this system. Generally we define the *state space* to be values of state variables which correspond to actual states of the system. The state space in Example 1.1.1 is the circle of radius r centered at $(0, 0)$.

Exercises

***Exercise* 1.1.1.** Suppose that in Example 1.1.1 light sources are located at $(0, \pm R)$. What is the relationship between the x-coordinate and the shadow?

***Exercise* 1.1.2.** Suppose that in Example 1.1.1 the ball is tethered to $(0, 0)$ by a string of length r. What relations do the state variables (x, y) satisfy? Is there a measurement the line creatures can make to determine the location of the ball? What is the state space for this system?

***Exercise* 1.1.3.** Suppose that the ball is untethered but is constrained to lie in the region $\{(x, y) : 0 \leq y < R\}$. Assume that the points $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$ do not lie on a line and have $y_j > R$. Show that the shadows cast on the line $y = 0$ by light sources located at these three points determine the location of the ball. Find a formula for (x, y) in terms of the shadow locations. Why are three sources needed?

1.1.1 Finitely Many Degrees of Freedom

See: A.1, B.3, B.4, B.5.

The collection of ordered n-tuples of real numbers

$$\{(x_1, \dots, x_n) : x_j \in \mathbb{R}, j = 1, \dots, n\}$$

is called Euclidean n-space and is denoted by $\mathbb{R}^n$. We often use boldface letters $\boldsymbol{x}, \boldsymbol{y}$ to denote points in $\mathbb{R}^n$, which we sometimes call vectors. Recall that if $\boldsymbol{x} = (x_1, \dots, x_n)$ and $\boldsymbol{y} = (y_1, \dots, y_n)$, then their sum $\boldsymbol{x} + \boldsymbol{y}$ is defined by

$$\boldsymbol{x} + \boldsymbol{y} = (x_1 + y_1, \dots, x_n + y_n), \tag{1.2}$$

and if $a \in \mathbb{R}$, then $a\boldsymbol{x}$ is defined by

$$a\boldsymbol{x} = (ax_1, \dots, ax_n). \tag{1.3}$$

These two operations make $\mathbb{R}^n$ into a real vector space.

***Definition* 1.1.1.** If the state of a system is described by a finite collection of real numbers, then the system has *finitely many degrees of freedom.*

Euclidean n-space is the simplest state space for a system with n degrees of freedom. Most systems encountered in elementary physics and engineering have finitely many degrees of freedom. Suppose that the state of a system is specified by a point $\boldsymbol{x} \in \mathbb{R}^n$. Then the mathematical model is expressed as relations that these variables satisfy. These often take the form of functional relations,

$$\begin{aligned} f_1(x_1, \dots, x_n) &= 0 \\ \vdots \quad & \quad \vdots \\ f_m(x_1, \dots, x_n) &= 0. \end{aligned} \tag{1.4}$$

In addition to conditions like those in (1.4) the parameters describing the state of a system might also satisfy inequalities of the form

$$\begin{aligned} g_1(x_1, \dots, x_n) &\geq 0 \\ \vdots \quad & \quad \vdots \\ g_l(x_1, \dots, x_n) &\geq 0. \end{aligned} \tag{1.5}$$

The state space for the system is then the subset of $\mathbb{R}^n$ consisting of solutions to (1.4) which also satisfy (1.5).

***Definition* 1.1.2.** An equation or inequality that must be satisfied by a point belonging to the state space of a system is called a *constraint*.

Example 1.1.1 considers a system with one degree of freedom. The state space for this system is the subset of $\mathbb{R}^2$ consisting of points satisfying (1.1). If the state variables satisfy constraints, then this generally reduces the number of degrees of freedom.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is linear if it satisfies the conditions

$$\begin{gathered} f(\boldsymbol{x} + \boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y}) \text{ for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n \text{ and} \\ f(a\boldsymbol{x}) = af(\boldsymbol{x}) \text{ for all } a \in \mathbb{R} \text{ and } \boldsymbol{x} \in \mathbb{R}^n. \end{gathered} \tag{1.6}$$

Recall that the dot or inner product is the map from $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{j=1}^{n} x_j y_j. \tag{1.7}$$

Sometimes it is denoted by $\boldsymbol{x} \cdot \boldsymbol{y}$. The Euclidean length of $\boldsymbol{x} \in \mathbb{R}^n$ is defined to be

$$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \left[\sum_{j=1}^{n} x_j^2 \right]^{\frac{1}{2}}. \tag{1.8}$$

From the definition it is easy to establish that

$$\begin{aligned} \langle \boldsymbol{x}, \boldsymbol{y} \rangle &= \langle \boldsymbol{y}, \boldsymbol{x} \rangle \text{ for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \\ \langle a\boldsymbol{x}, \boldsymbol{y} \rangle &= a\langle \boldsymbol{x}, \boldsymbol{y} \rangle \text{ for all } a \in \mathbb{R} \text{ and } \boldsymbol{x} \in \mathbb{R}^n, \\ \langle \boldsymbol{x}_1 + \boldsymbol{x}_2, \boldsymbol{y} \rangle &= \langle \boldsymbol{x}_1, \boldsymbol{y} \rangle + \langle \boldsymbol{x}_2, \boldsymbol{y} \rangle \text{ for all } \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y} \in \mathbb{R}^n. \\ \|c\boldsymbol{x}\| &= |c| \|\boldsymbol{x}\| \text{ for all } c \in \mathbb{R} \text{ and } \boldsymbol{x} \in \mathbb{R}^n. \end{aligned} \tag{1.9}$$

For $\boldsymbol{y}$ a point in $\mathbb{R}^n$, define the function $f_{\boldsymbol{y}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$. The second and third relations in (1.9) show that $f_{\boldsymbol{y}}$ is linear. Indeed, every linear function has a such a representation.

Proposition 1.1.1. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a linear function, then there is a unique vector $\boldsymbol{y}_f$ such that $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{y}_f \rangle$.*

This fact is proved in Exercise 1.1.5. The inner product satisfies a basic inequality called the Cauchy-Schwarz inequality.

Proposition 1.1.2 (Cauchy-Schwarz inequality). *If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, then*

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|. \tag{1.10}$$

A proof of this result is outlined in Exercise 1.1.6. The Cauchy-Schwarz inequality shows that if neither $\boldsymbol{x}$ nor $\boldsymbol{y}$ is zero, then

$$-1 \leq \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} \leq 1;$$

this in turn allows us the define the angle between two vectors.

***Definition* 1.1.3.** If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ are both nonvanishing, then the angle $\theta \in [0, \pi]$, between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined by

$$\cos\theta = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}. \tag{1.11}$$

In particular, two vector are *orthogonal* if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

The Cauchy-Schwarz inequality implies that the Euclidean length satisfies the *triangle inequality*.

Proposition 1.1.3. *For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the following inequality holds:*

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|. \tag{1.12}$$

This is called the triangle inequality.

***Remark* 1.1.1.** The Euclidean length is an example of a *norm* on $\mathbb{R}^n$. A real-valued function N defined on $\mathbb{R}^n$ is a norm provided it satisfies the following conditions:

NON-DEGENERACY:
: $N(\boldsymbol{x}) = 0$ if and only if $\boldsymbol{x} = 0$,

HOMOGENEITY:
: $N(a\boldsymbol{x}) = |a|N(\boldsymbol{x})$ for all $a \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$,

THE TRIANGLE INEQUALITY:
: $N(\boldsymbol{x} + \boldsymbol{y}) \leq N(\boldsymbol{x}) + N(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Any norm provides a way to measure distances. The distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined to be

$$d_N(\boldsymbol{x}, \boldsymbol{y}) \stackrel{d}{=} N(\boldsymbol{x} - \boldsymbol{y}).$$

Suppose that the state of a system is specified by a point in $\mathbb{R}^n$ subject to the constraints in (1.4). If all the functions $\{f_1, \dots, f_m\}$ are linear, then we say that this is a *linear model.* This is the simplest type of model and also the most common in applications. In this case the set of solutions to (1.4) is a *subspace* of $\mathbb{R}^n$. We recall the definition.

***Definition* 1.1.4.** A subset $S \subset \mathbb{R}^n$ is a subspace if

1. the zero vector belongs to S,
2. $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$, then $\boldsymbol{x}_1 + \boldsymbol{x}_2 \in S$,
3. if $c \in \mathbb{R}$ and $\boldsymbol{x} \in S$, then $c\boldsymbol{x} \in S$.

For a linear model it is a simple matter to determine the number of degrees of freedom. Suppose the state space consists of vectors satisfying a single linear equation. In light of Proposition 1.1.1, it can be expressed in the form

$$\langle \boldsymbol{a}_1, \boldsymbol{x} \rangle = 0, \tag{1.13}$$

with $\boldsymbol{a}_1$ a nonzero vector. This is the equation of a *hyperplane* in $\mathbb{R}^n$. The solutions to (1.13) are the vectors in $\mathbb{R}^n$ orthogonal to $\boldsymbol{a}_1$. Recall the following definition:

***Definition* 1.1.5.** The vectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ are *linearly independent* if the only linear combination, $c_1\boldsymbol{v}_1 + \cdots + c_k\boldsymbol{v}_k$, that vanishes has all its coefficients, $\{c_i\}$, equal to zero. Otherwise the vectors are *linearly dependent.*

The dimension of a subspace of $\mathbb{R}^n$ can now be defined.

***Definition* 1.1.6.** Let $S \subset \mathbb{R}^n$ be a subspace. If there is a set of k linearly independent vectors contained in S but any set with $k + 1$ or more vectors is linearly dependent, then the dimension of S equals k. In this case we write $\dim S = k$.

There is a collection of $(n - 1)$ linearly independent n-vectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_{n-1}\}$ so that $\langle \boldsymbol{a}_1, \boldsymbol{x} \rangle = 0$ if and only if

$$\boldsymbol{x} = \sum_{i=1}^{n-1} c_i \boldsymbol{v}_i.$$

The hyperplane has dimension $n - 1$, and therefore a system described by a single linear equation has $n - 1$ degrees of freedom. The general case is not much harder. Suppose that the state space is the solution set of the system of linear equations

$$\begin{aligned} \langle \boldsymbol{a}_1, \boldsymbol{x} \rangle &= 0 \\ \vdots \quad & \quad \vdots \\ \langle \boldsymbol{a}_m, \boldsymbol{x} \rangle &= 0. \end{aligned} \tag{1.14}$$

Suppose that $k \leq m$ is the largest number of linearly independent vectors in the collection $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$. By renumbering, we can assume that $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k\}$ are linearly independent, and for any $l > k$ the vector $\boldsymbol{a}_l$ is a linear combination of these vectors. Hence if $\boldsymbol{x}$ satisfies

$$\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle = 0 \text{ for } 1 \leq i \leq k$$

then it also satisfies $\langle \boldsymbol{a}_l, \boldsymbol{x} \rangle = 0$ for any l greater than k. The argument in the previous paragraph can be applied recursively to conclude that there is a collection of $n - k$ linearly independent vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n-k}\}$ so that $\boldsymbol{x}$ solves (1.14) if and only if

$$\boldsymbol{x} = \sum_{i=1}^{n-k} c_i \boldsymbol{u}_i.$$

Thus the system has $n - k$ degrees of freedom.

A nonlinear model can often be approximated by a linear model. If f is a differentiable function, then the gradient of f at $\boldsymbol{x}$ is defined to be

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \right).$$

From the definition of the derivative it follows that

$$f(\boldsymbol{x}_0 + \boldsymbol{x}_1) = f(\boldsymbol{x}_0) + \langle \boldsymbol{x}_1, \nabla f(\boldsymbol{x}_0) \rangle + e(\boldsymbol{x}_1), \tag{1.15}$$

where the *error* $e(\boldsymbol{x}_1)$ satisfies

$$\lim_{\boldsymbol{x}_1 \to 0} \frac{|e(\boldsymbol{x}_1)|}{\|\boldsymbol{x}_1\|} = 0.$$

In this case we write

$$f(\boldsymbol{x}_0 + \boldsymbol{x}_1) \approx f(\boldsymbol{x}_0) + \langle \boldsymbol{x}_1, \nabla f(\boldsymbol{x}_0) \rangle. \tag{1.16}$$

Suppose that the functions in (1.4) are differentiable and $f_j(\boldsymbol{x}_0) = 0$ for $j = 1, \ldots, m$. Then

$$f_j(\boldsymbol{x}_0 + \boldsymbol{x}_1) \approx \langle \boldsymbol{x}_1, \nabla f_j(\boldsymbol{x}_0) \rangle.$$

For small values of $\boldsymbol{x}_1$ the system of equations (1.4) can be approximated, near to $\boldsymbol{x}_0$, by a system of linear equations,

$$\begin{aligned} \langle \boldsymbol{x}_1, \nabla f_1(\boldsymbol{x}_0) \rangle &= 0 \\ \vdots \quad & \quad \vdots \\ \langle \boldsymbol{x}_1, \nabla f_m(\boldsymbol{x}_0) \rangle &= 0. \end{aligned} \tag{1.17}$$

This provides a linear model that approximates the non-linear model. The accuracy of this approximation depends, in a subtle way, on the collection of vectors $\{\nabla f_j(\boldsymbol{x})\}$, for $\boldsymbol{x}$ near to $\boldsymbol{x}_0$. The simplest situation is when these vectors are linearly independent at $\boldsymbol{x}_0$. In this case the solutions to

$$f_j(\boldsymbol{x}_0 + \boldsymbol{x}_1) = 0, \quad j = 1, \ldots, m,$$

are well approximated, for small $\boldsymbol{x}_1$, by the solutions of (1.17). This is a consequence of the implicit function theorem; see [119].

Often the state variables for a system are divided into two sets, the *input variables,* $(w_1, \dots, w_k)$, and *output variables,* $(z_1, \dots, z_m)$, with constraints rewritten in the form

$$\begin{aligned} F_1(w_1, \dots, w_k) &= z_1 \\ \vdots \qquad & \quad \vdots \\ F_m(w_1, \dots, w_k) &= z_m. \end{aligned} \tag{1.18}$$

The output variables are thought of as being measured; the remaining variables must then be determined by solving this system of equations. For a linear model this amounts to solving a system of linear equations. We now consider some examples of physical systems and their mathematical models.

***Example* 1.1.2.** We would like to find the height of a mountain without climbing it. To that end, the distance x between the point P and the base of the mountain, as well as the angle θ, are measured (Figure 1.3). If x and θ are measured exactly, then the height, h, of the mountain is given by

$$h(x, \theta) = x \tan\theta. \tag{1.19}$$

Measurements are never exact; using the model and elementary calculus, we can relate the error in the measurement θ to the error in the computed value of h. Suppose that x is measured exactly but there is an uncertainty $\Delta\theta$ in the value of θ. Equation (1.16) gives the linear approximation

$$h(x, \theta + \Delta\theta) - h(x, \theta) \approx \frac{\partial h}{\partial\theta}(x, \theta)\Delta\theta.$$

As $\partial_\theta h = x \sec^2\theta$, the height, h_m, predicted from the measurement of the angle is given by

$$h_m = x \tan(\theta + \Delta\theta) \approx x(\tan\theta + \sec^2\theta\,\Delta\theta).$$

The approximate value of the *absolute error* is

$$h_m - h \approx x\frac{\Delta\theta}{\cos^2\theta}.$$

The absolute error is a number with the same units as h; in general, it is not an interesting quantity. If, for example, the true measurement were 10,000 m, then an error of size 1 m would not be too significant. If the true measurement were 2 m, then this error would be significant. To avoid this obvious pitfall, we normally consider the *relative error*. In this problem the relative error is

$$\frac{h_m - h}{h} = \frac{\Delta\theta}{\cos^2\theta\tan\theta} = \frac{\Delta\theta}{\sin\theta\cos\theta}.$$

Generally the relative error is the absolute error divided by the correct value. It is a dimensionless quantity that gives a quantitative assessment of the accuracy of the measurement.

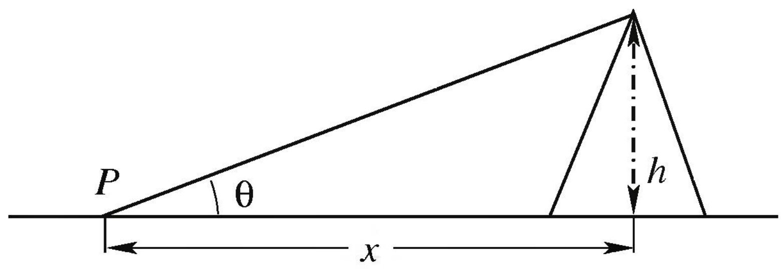


Figure 1.3. Using trigonometry to find the height of a mountain.

If the angle θ is measured from a point too near to or too far from the mountain (i.e., θ is very close to 0 or $\pi/2$), then small measurement errors result in a substantial loss of accuracy. A useful feature of a precise mathematical model is the possibility of estimating how errors in measurement affect the accuracy of the parameters we wish to determine. In Exercise 1.1.13 we consider how to estimate the error entailed in using a linear approximation.

***Example* 1.1.3.** In a real situation we cannot measure the distance to the base of the mountain. Suppose that we measure the angles, θ_1 and θ_2, from two different points, P_1 and P_2, as well as the distance $x_2 - x_1$ between the two points, as shown in Figure 1.4.

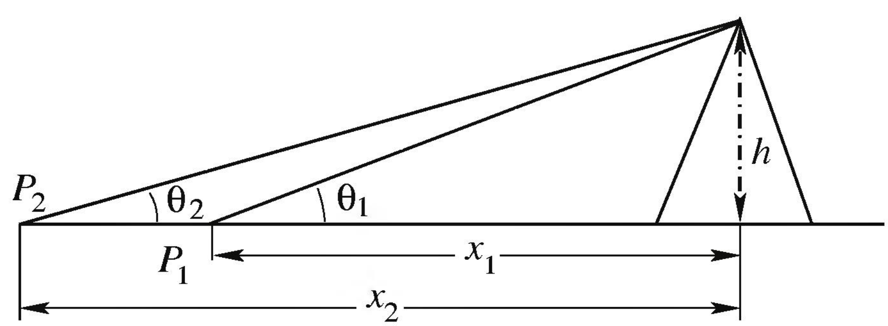


Figure 1.4. A more realistic measurement.

From the previous example we know that

$$\begin{aligned} h &= x_1 \tan\theta_1, \\ h &= x_2 \tan\theta_2. \end{aligned} \tag{1.20}$$

Using these equations and elementary trigonometry, we deduce that

$$x_1 = \frac{x_2 - x_1}{\left[\frac{\tan\theta_1}{\tan\theta_2} - 1\right]}, \tag{1.21}$$

which implies that

$$\begin{aligned} h &= x_1 \tan\theta_1 \\ &= (x_2 - x_1)\frac{\sin\theta_1 \sin\theta_2}{\sin(\theta_1 - \theta_2)}. \end{aligned} \tag{1.22}$$

Thus h can be determined from θ_1, θ_2 and $x_2 - x_1$. With $d = x_2 - x_1$, equation (1.22) expresses h as a function of (d, θ_1, θ_2). At the beginning of this example, $(x_1, \theta_1, x_2, \theta_2, h)$ were the state variables describing our system; by the end we used $(d, \theta_1, \theta_2, h)$. The first three are directly measurable, and the last is an explicit function of the others. The models in this and the previous example, as expressed by equations (1.22) and (1.19), respectively, are nonlinear models.

In this example there are many different ways that the model may fail to capture important features of the physical situation. We now consider a few potential problems.

1. If the shape of a mountain looks like that in Figure 1.5 and we measure the distance and angle at the point P, we are certainly not finding the real height of the mountain. Some a priori information is always incorporated in a mathematical model.

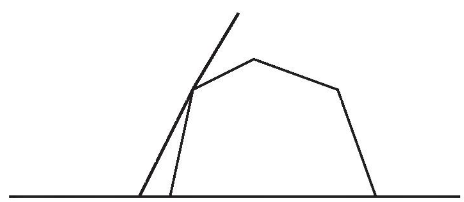


Figure 1.5. Not exactly what we predicted!

2. The curvature of the earth is ignored. A more sophisticated geometric model is needed to correct for such errors. This becomes a significant problem as soon as the distances, x, x_1, x_2, are large compared to the distance to the horizon (about 25 km for a 2-meter-tall person). The approximations used in the model must be adapted to the actual physical conditions of the measurements.

3. The geometry of the underlying measurements could be quite different from the simple Euclidean geometry used in the model. To measure the angles θ_1, θ_2, we would normally use a transit to sight the peak of the mountain. If the mountain is far away, then the light traveling from the mountain to the transit passes through air of varying density. The light is refracted by the air and therefore the ray path is not a straight line, as assumed in the model. To include this effect would vastly complicate the model. This is an important consideration in the similar problem of creating a map of the sky from earth based observations of stars and planets.

Analogous problems arise in medical imaging. If the wavelength of the energy used to probe the human anatomy is very small compared to the size of the structures that are present, then it is reasonable to assume that the waves are not refracted. For example, x-rays can be assumed to travel along straight lines. For energies with wavelengths comparable to the size of structures present in the human anatomy, this assumption is simply wrong. The waves are then bent and diffracted by the medium, and the difficulty of modeling the ray paths is considerable. This is an important issue in ultrasound imaging that remains largely unresolved.

***Example* 1.1.4.** Refraction provides another example of a simple physical system. Suppose that we have two fluids in a tank, as shown in Figure 1.6, and would like to determine the height of the interface between them. Suppose that the refractive indices of the fluids are known. Let n_1 be the refractive index of the upper fluid and n_2 the refractive index of the lower one. Snell's law states that

$$\frac{\sin(\theta_1)}{\sin(\theta_2)} = \frac{n_2}{n_1}.$$

Let h denote the total height of the fluid; then

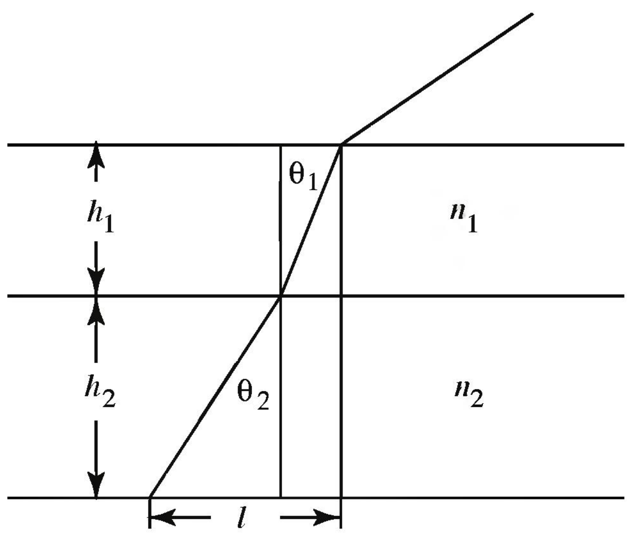


Figure 1.6. Using refraction to determine the height of an interface.

$$h_1 + h_2 = h.$$

The measurement we make is the total displacement, l, of the light ray as it passes through the fluids. It satisfies the relationship

$$h_1 \tan(\theta_1) + h_2 \tan(\theta_2) = l.$$

The heights h_1 and h_2 are easily determined from these three formulæ. The assumption that we know n_1 implies, by Snell's law, that we can determine θ_1 from a measurement of the angle of the light ray above the fluid. If n_2 is also known, then using these observations we can determine θ_2 as well:

$$\sin(\theta_2) = \frac{n_1}{n_2}\sin(\theta_1).$$

The pair (h_1, h_2) satisfies the 2×2 linear system

$$\begin{pmatrix} 1 & 1 \\ \tan(\theta_1) & \tan(\theta_2) \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} h \\ l \end{pmatrix}. \tag{1.23}$$

In Example 2.1.1 we consider a slightly more realistic situation where the refractive index of the lower fluid in not known. By using more measurements, n_2 can also be determined.

Exercises

***Exercise* 1.1.4.** Prove the formulæ in (1.9).

***Exercise* 1.1.5.** Let $\boldsymbol{e}_j \in \mathbb{R}^n$, $j = 1, \ldots, n$ denote the vector with a 1 in the jth place and otherwise zero,

$$\boldsymbol{e}_1 = (1, 0, 0, \ldots, 0), \ \boldsymbol{e}_2 = (0, 1, 0, \ldots, 0), \ldots, \boldsymbol{e}_n = (0, \ldots, 0, 1).$$

1. Show that if $\boldsymbol{x} = (x_1, \ldots, x_n)$, then

$$\boldsymbol{x} = \sum_{j=1}^{n} x_j \boldsymbol{e}_j.$$

2. Use the previous part to prove the existence statement in Proposition 1.1.1; that is, show that there is a vector $\boldsymbol{y}_f$ so that $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{y}_f \rangle$. Give a formula for $\boldsymbol{y}_f$.

3. Show that the uniqueness part of the proposition is equivalent to the statement "If $\boldsymbol{y} \in \mathbb{R}^n$ satisfies

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0 \text{ for all } \boldsymbol{x} \in \mathbb{R}^n,$$

then $\boldsymbol{y} = 0$." Prove this statement.

***Exercise* 1.1.6.** In this exercise we use calculus to prove the Cauchy-Schwarz inequality. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ be nonzero vectors. Define the function

$$F(t) = \langle \boldsymbol{x} + t\boldsymbol{y}, \boldsymbol{x} + t\boldsymbol{y} \rangle.$$

Use calculus to find the value of t, where F assumes its minimum value. By using the fact that $F(t) \geq 0$ for all t, deduce the Cauchy-Schwarz inequality.

***Exercise* 1.1.7.** Show that (1.12) is a consequence of the Cauchy-Schwarz inequality. *Hint*: Consider $\|\boldsymbol{x} + \boldsymbol{y}\|^2$.

***Exercise* 1.1.8.** Define a real-valued function on $\mathbb{R}^n$ by setting

$$N(\boldsymbol{x}) = \max\{|x_1|, \ldots, |x_n|\}.$$

Show that N defines a norm.

***Exercise* 1.1.9.** Let N be a norm on $\mathbb{R}^n$ and define $d(\boldsymbol{x}, \boldsymbol{y}) = N(\boldsymbol{x} - \boldsymbol{y})$. Show that for any triple of points $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3,$ the following estimate holds:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_3) \leq d(\boldsymbol{x}_1, \boldsymbol{x}_2) + d(\boldsymbol{x}_2, \boldsymbol{x}_3).$$

Explain why this is also called the triangle inequality.

***Exercise* 1.1.10.** Let $S \subset \mathbb{R}^n$ be a subspace of dimension k. Show that there exists a collection of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\} \subset S$ such that every vector $\boldsymbol{x} \in S$ has a *unique* representation of the form

$$\boldsymbol{x} = c_1\boldsymbol{v}_1 + \cdots + c_k\boldsymbol{v}_k.$$

***Exercise* 1.1.11.** Let $\boldsymbol{a}$ be a nonzero n-vector. Show that there is a collection of $n-1$ linearly independent n-vectors, $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-1}\}$, so that $\boldsymbol{x}$ solves $\langle \boldsymbol{a}, \boldsymbol{x} \rangle = 0$ if and only if

$$\boldsymbol{x} = \sum_{i=1}^{n-1} c_i \boldsymbol{v}_i$$

for some real constants $\{c_1, \ldots, c_{n-1}\}$.

***Exercise* 1.1.12.** Let $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k\}$ be linearly independent n-vectors. Show that there is a collection of $n-k$ linearly independent n-vectors, $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-k}\}$, so that $\boldsymbol{x}$ solves

$$\langle \boldsymbol{a}_j, \boldsymbol{x} \rangle = 0 \text{ for } j = 1, \ldots, k$$

if and only if

$$\boldsymbol{x} = \sum_{i=1}^{n-k} c_i \boldsymbol{v}_i$$

for some real constants $\{c_1, \ldots, c_{n-k}\}$. *Hint*: Use the previous exercise and an induction argument.

***Exercise* 1.1.13.** If a function f has two derivatives, then Taylor's theorem gives a formula for the error $e(y) = f(x+y) - [f(x) + f'(x)y]$. There exists a z between 0 and y such that

$$e(z) = \frac{f''(z)y^2}{2};$$

see (B.13). Use this formula to bound the error made in replacing $h(x, \theta + \Delta\theta)$ with $h(x, \theta) + \partial_\theta h(x, \theta)\Delta\theta$. *Hint*: Find the value of z between 0 and $\Delta\theta$ that maximizes the error term.

***Exercise* 1.1.14.** In Example 1.1.3 compute the gradient of h to determine how the absolute and relative errors depend on $\theta_1, \theta_2,$ and d.

1.1.2 Infinitely Many Degrees of Freedom

See: A.3, A.5.

In the previous section we examined some simple physical systems with finitely many degrees of freedom. In these examples, the problem of determining the state of the system from feasible measurements reduces to solving systems of finitely many equations in finitely many unknowns. In imaging applications the state of a system is usually described by a function or functions of continuous variables. These systems have infinitely many degrees of freedom. In this section we consider several examples.

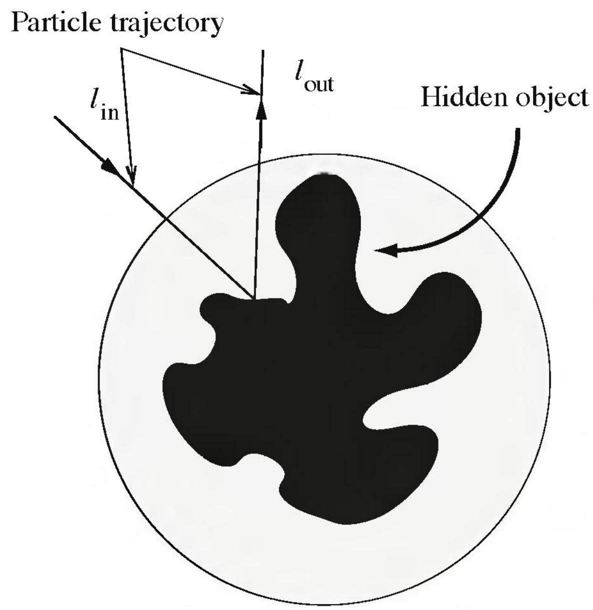
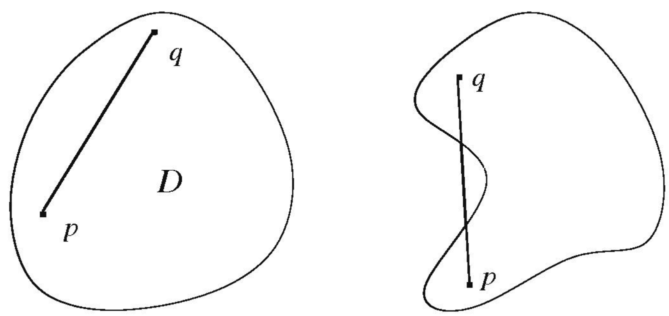


Figure 1.7. Particle scattering can be used to explore the boundary of an unknown region.

***Example* 1.1.5.** Suppose that we would like to determine the shape of a planar region, D, that cannot be seen. The object is lying inside a disk and we can fire particles at the object. Assume that the particles bounce off according to a simple scattering process. Each particle strikes the object once and is then scattered along a straight line off to infinity (Figure 1.7). The outline of the object can be determined by knowing the correspondence between incoming lines, l_{in}, and outgoing lines, l_{out}. Each intersection point $l_{\text{in}} \cap l_{\text{out}}$ lies on the boundary of the object. Measuring $\{l^j_{\text{out}}\}$ for finitely many incoming directions $\{l^j_{\text{in}}\}$ determines finitely many points $\{l^j_{\text{in}} \cap l^j_{\text{out}}\}$ on the boundary of D. In order to use this finite collection of points to make any assertions about the rest of the boundary of D, more information is required. If we know that D consists of a single piece or component, then these points would lie on a single closed curve, though it might be difficult to decide in what order they should appear on the curve. On the other hand, these measurements provide a lot of information about *convex regions.*

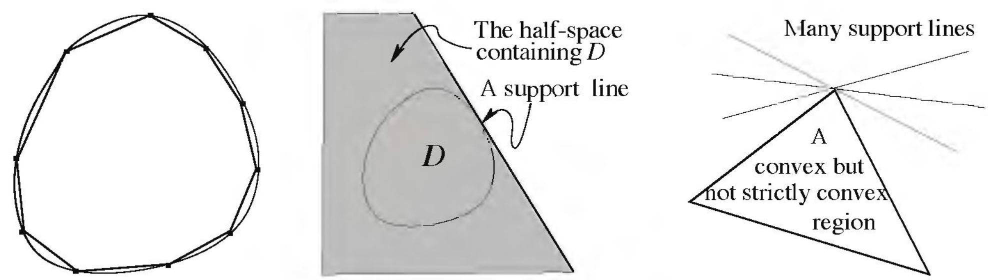
Definition* 1.1.7. A region D in the plane is *convex* if it has the following property: For each pair of points p and q lying in D, the line segment $\overline{pq}$ is also contained in D. See Figure 1.8.



(a) A convex region (b) A non-convex region

Figure 1.8. Convex and non-convex regions.

Convex regions have many special properties. If p and q are on the boundary of D, then the line segment $\overline{pq}$ lies inside of D. From this observation we can show that if $\{p_1, \ldots, p_N\}$ are points on the boundary of a convex region, then the smallest polygon with these points as vertices lies entirely within D [Figure 1.9(a)]. Convexity can also be defined by a property of the boundary of D: For each point p on the boundary of D there is a line l_p that passes through p but is otherwise disjoint from D. This line is called a *support line* through p. If the boundary is smooth at p, then the tangent line to the boundary is the unique support line. A line divides the plane into two half-planes. Let l_p be a support line to D at p. Since the interior of D does not meet l_p it must lie entirely in one of the half-planes determined by this line [see Figure 1.9(b)]. If each support line meets the boundary of D at exactly one point, then the region is *strictly* convex.



(a) Joining the vertices lying on the boundary of a convex region defines an inscribed polygon.

(b) A support line and half-space.

(c) A triangle bounds a convex region that is not strictly convex.

Figure 1.9. Inscribed polygons and support lines.

***Example* 1.1.6.** A triangle is the boundary of a convex region, with each edge of the triangle a support line. As infinitely many points of the boundary belong to each edge, the region bounded by a triangle in not strictly convex. On the other hand, through each vertex of the triangle, there are infinitely many support lines. These observations are illustrated in Figure 1.9(c).

Suppose that the object is convex and more is known about the scattering process: for example, that the angle of incidence is equal to the angle of reflection. From a finite number of incoming and outgoing pairs, $\{(l^i_{\text{in}}, l^i_{\text{out}}) : i = 1, \dots, N\}$, we can now determine an approximation to D with an estimate for the error. The intersection points, $p_i = l^i_{\text{in}} \cap l^i_{\text{out}}$ lie on the boundary of the convex region, D. If we use these points as the vertices of a polygon P^{in}_N, then, as remarked previously, P^{in}_N is completely contained within D. On the other hand, as the angle of incidence equals the angle of reflection, we can also determine the tangent lines $\{l_{p_i}\}$ to the boundary of D at the points $\{p_i\}$. These lines are support lines for D. Hence by intersecting the half-planes that contain D, defined by these tangent lines, we obtain another convex polygon, P^{out}_N, that contains D. Thus with these N-measurements we obtain the both an *inner* and *outer* approximation to D:

$$P^{\text{in}}_N \subset D \subset P^{\text{out}}_N.$$

An example is shown in Figure 1.10.

A convex region is determined by its boundary, and each point on the boundary is, in effect, a state variable. Therefore, the collection of convex regions is a system with *infinitely many degrees of freedom*. A nice description for the state space of smooth convex regions is developed in Section 1.2.2. As we have seen, a convex region can be approximated by polygons. Once the number of sides is fixed, then we are again considering a system with finitely many degrees of freedom. In all practical problems, a system with infinitely many degrees of freedom must eventually be approximated by a system with finitely many degrees of freedom.

***Remark* 1.1.1.** For a non-convex body, the preceding method does not work as the correspondence between incoming and outgoing lines can be complicated: Some incoming lines may undergo multiple reflections before escaping, and in fact some lines might become permanently trapped.

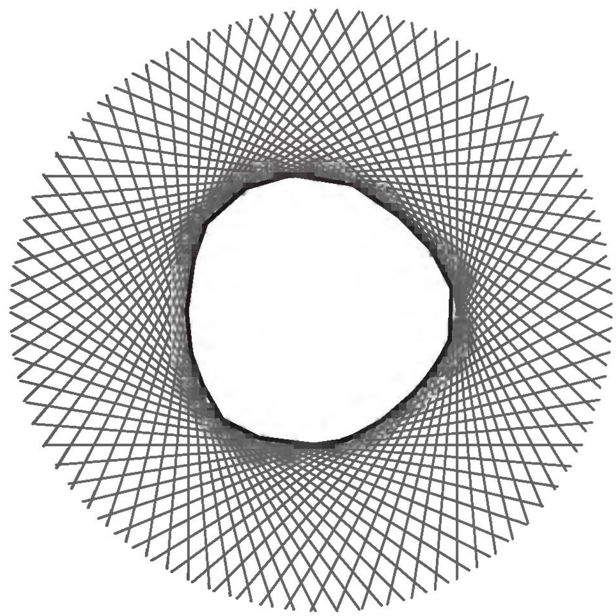


Figure 1.10. An inner and an outer approximation to a convex region.

Exercises

***Exercise* 1.1.15.** Find state variables to describe the set of polygons with n-vertices in the plane. For the case of triangles, find the relations satisfied by your variables. Extra credit: Find a condition, in terms of your parameters, implying that the polygon is convex.

***Exercise* 1.1.16.** Suppose that D_1 and D_2 are convex regions in the plane. Show that their intersection $D_1 \cap D_2$ is also a convex region.

Exercise* 1.1.17. Suppose that D is a possibly non-convex region in the plane. Define a new region D' as the intersection of all the half-planes that contain D. Show that $D = D'$ if and only if D is convex.

***Exercise* 1.1.18.** Find an example of a planar region such that at least one particle trajectory is trapped forever.

1.2 A Simple Model Problem for Image Reconstruction

The problem of image reconstruction in x-ray tomography is sometimes described as reconstructing an object from its "projections." Of course, these are projections under the illumination of x-ray "light." In this section we consider the analogous but simpler problem of determining the outline of a convex object from its shadows. As is also the case in medical applications, we consider a two-dimensional problem. Let D be a convex region in the plane. Imagine that a light source is placed very far from D. Since the light source is very far away, the rays of light are all traveling in essentially the same direction. We can think of them as a collection of parallel lines. We want to measure the shadow that D casts for each position of the light source. To describe the measurements imagine that a screen is placed on the "other side" of D perpendicular to the direction of the light rays (Figure 1.11). In a real apparatus sensors would be placed on the screen, allowing us to determine where the shadow begins and ends.

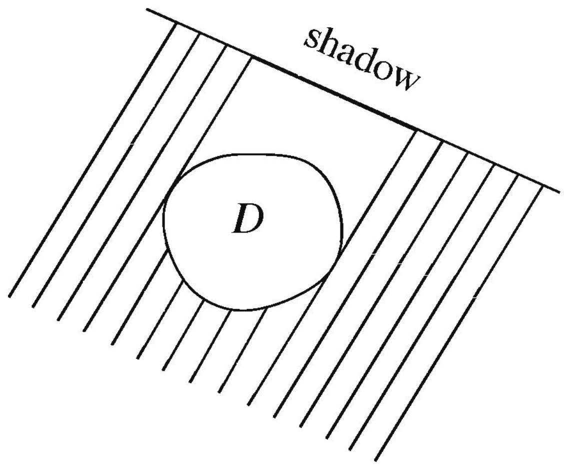


Figure 1.11. The shadow of a convex region.

The region, D, blocks a certain collection of light rays and allows the rest to pass. Locating the shadow amounts to determining the "first" and "last" lines in this family of parallel lines to intersect D. To describe the object completely, we need to rotate the source and detector through π radians, measuring, at each angle, where the shadow begins and ends.

The first and last lines to intersect a region just meet it along its boundary. These lines are therefore tangent to the boundary of D. The problem of reconstructing a region from its shadows is mathematically the same as the problem of reconstructing a region from a knowledge of the tangent lines to its boundary. As a first step in this direction we need a good way to organize our measurements. To that end we give a description for the *space of lines in the plane.*

1.2.1 The Space of Lines in the Plane*

A line in the plane is a set of points that satisfies an equation of the form

$$ax + by = c,$$

where $a^2 + b^2 \neq 0$. We could use (a, b, c) to parameterize the set of lines, but note that we get the same set of points if we replace this equation by

$$\frac{a}{\sqrt{a^2+b^2}}x + \frac{b}{\sqrt{a^2+b^2}}y = \frac{c}{\sqrt{a^2+b^2}}.$$

The coefficients $(\frac{a}{\sqrt{a^2+b^2}}, \frac{b}{\sqrt{a^2+b^2}})$ define a point $\boldsymbol{\omega}$ on the unit circle, $S^1 \subset \mathbb{R}^2$, and the constant $\frac{c}{\sqrt{a^2+b^2}}$ can be any number. The lines in the plane are parameterized by a pair consisting of a unit vector

$$\boldsymbol{\omega} = (\omega_1, \omega_2)$$

and a real number t. The line $l_{t,\boldsymbol{\omega}}$ is the set of points in $\mathbb{R}^2$ satisfying the equation

$$\langle (x, y), \boldsymbol{\omega} \rangle = t. \tag{1.24}$$

The vector $\boldsymbol{\omega}$ is perpendicular to this line (Figure 1.12).

It is often convenient to parameterize the points on the unit circle by a real number; to that end we set

$$\boldsymbol{\omega}(\theta) = (\cos(\theta), \sin(\theta)). \tag{1.25}$$

Since cos and sin are 2π-periodic, it clear that $\boldsymbol{\omega}(\theta)$ and $\boldsymbol{\omega}(\theta + 2\pi)$ are the same point on the unit circle. Using this parameterization for points on the circle, the line $l_{t,\theta} \stackrel{d}{=} l_{t,\boldsymbol{\omega}(\theta)}$ is the set of solutions to the equation

$$\langle (x, y), (\cos(\theta), \sin(\theta)) \rangle = t.$$

Both notations for lines and points on the circle are used in the sequel.

While the parameterization provided by $(t, \boldsymbol{\omega})$ is much more efficient than that provided by (a, b, c), note that the set of points satisfying (1.24) is unchanged if $(t, \boldsymbol{\omega})$ is replaced by $(-t, -\boldsymbol{\omega})$. Thus, as sets,

$$l_{t,\boldsymbol{\omega}} = l_{-t,-\boldsymbol{\omega}}. \tag{1.26}$$

It is not difficult to show that if $l_{t_1,\boldsymbol{\omega}_1} = l_{t_2,\boldsymbol{\omega}_2}$ then either $t_1 = t_2$ and $\boldsymbol{\omega}_1 = \boldsymbol{\omega}_2$ or $t_1 = -t_2$ and $\boldsymbol{\omega}_1 = -\boldsymbol{\omega}_2$.

The pair $(t, \boldsymbol{\omega})$ actually specifies an *oriented line.* That is, we can use these data to define the positive direction along the line. The vector

$$\hat{\boldsymbol{\omega}} = (-\omega_2, \omega_1)$$

is perpendicular to $\boldsymbol{\omega}$ and is therefore parallel to $l_{t,\boldsymbol{\omega}}$. In fact, $\hat{\boldsymbol{\omega}}$ and $-\hat{\boldsymbol{\omega}}$ are both unit vectors that are parallel to $l_{t,\boldsymbol{\omega}}$. The vector $\hat{\boldsymbol{\omega}}$ is selected by using the condition that the 2×2 matrix,

$$\begin{pmatrix} \omega_1 & -\omega_2 \\ \omega_2 & \omega_1 \end{pmatrix},$$

has determinant $+1$. The vector $\hat{\boldsymbol{\omega}}$ *defines* the positive direction or *orientation* of the line $l_{t,\boldsymbol{\omega}}$. This explains how the pair $(t, \boldsymbol{\omega})$ determines an *oriented line.* We summarize these computations in a proposition.

Proposition 1.2.1. *The pairs* $(t, \omega) \in \mathbb{R} \times S^1$ *are in one-to-one correspondence with the set of oriented lines in the plane.*

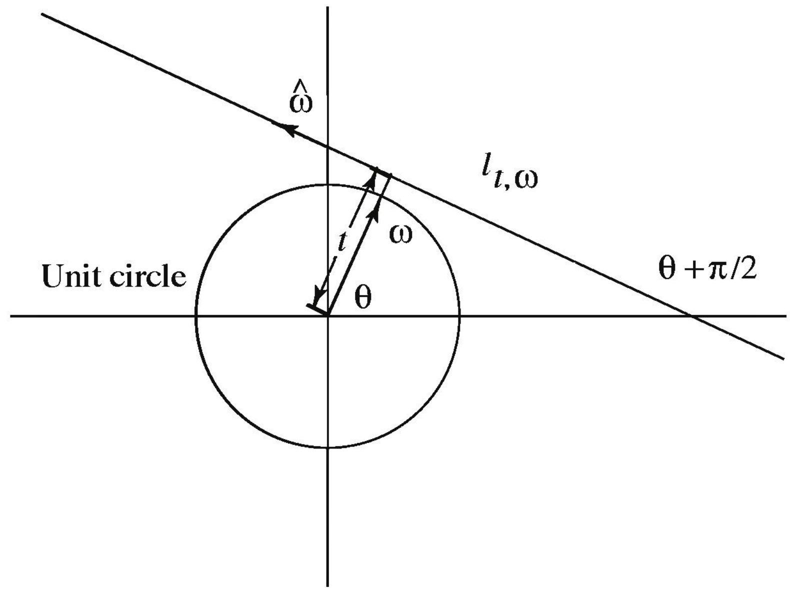


Figure 1.12. Parameterization of oriented lines in the plane.

The vector ω is the direction orthogonal to the line and the number t is called the *affine parameter* of the line; $|t|$ is the distance from the line to the origin of the coordinate system. The pair (t, ω) defines two half-planes

$$H^+_{t,\omega} = \{x \in \mathbb{R}^2 \mid \langle x, \omega\rangle > t\} \text{ and } H^-_{t,\omega} = \{x \in \mathbb{R}^2 \mid \langle x, \omega\rangle < t\}; \tag{1.27}$$

the line $l_{t,\omega}$ is the common boundary of these half-planes. Facing along the line $l_{t,\omega}$ in the direction specified by $\hat{\omega}$, the half-plane $H^-_{t,\omega}$ lies to the left.

Exercises

Exercise* 1.2.1. Show that $l_{t,\omega}$ is given parametrically as the set of points

$$l_{t,\omega} = \{t\omega + s\hat{\omega} \ : \ s \in (-\infty, \infty)\}.$$

Exercise* 1.2.2. Show that if $\omega = (\cos(\theta), \sin(\theta))$, then $\hat{\omega} = (-\sin(\theta), \cos(\theta))$, and as a function of θ :

$$\hat{\omega}(\theta) = \partial_\theta \omega(\theta).$$

***Exercise* 1.2.3.** Suppose that (t, ω) and (t_1, ω_1) are different points in $\mathbb{R} \times S^1$ such that $l_{t_1,\omega_1} = l_{t_2,\omega_2}$. Show that $(t_1, \omega_1) = (-t_2, -\omega_2)$.

***Exercise* 1.2.4.** Show that

$$|t| = \min\{\sqrt{x^2 + y^2} \ : \ (x, y) \in l_{t,\boldsymbol{\omega}}\}.$$

***Exercise* 1.2.5.** Show that if $\boldsymbol{\omega}$ is fixed, then the lines in the family $\{l_{t,\boldsymbol{\omega}} \ : \ t \in \mathbb{R}\}$ are parallel.

***Exercise* 1.2.6.** Show that every line in the family $\{l_{t,\hat{\boldsymbol{\omega}}} \ : \ t \in \mathbb{R}\}$ is orthogonal to every line in the family $\{l_{t,\boldsymbol{\omega}} \ : \ t \in \mathbb{R}\}$.

***Exercise* 1.2.7.** Each choice of direction $\boldsymbol{\omega}$ defines a coordinate system on $\mathbb{R}^2$,

$$(x, y) = t\boldsymbol{\omega} + s\hat{\boldsymbol{\omega}}.$$

Find the inverse, expressing (t, s) as functions of (x, y). Show that the area element in the plane satisfies

$$dx\, dy = dt\, ds.$$

1.2.2 Reconstructing an Object from Its Shadows

Now we can quantitatively describe the shadow. Because there are two lines in each family of parallel lines that are tangent to the boundary of D, we need a way to select one of them. To do this we choose an orientation for the boundary of D; this operation is familiar from Green's theorem in the plane. The positive direction on the boundary is selected so that, when facing in this direction the region lies to the left; the counterclockwise direction is, by convention, the positive direction (Figure 1.13).

Fix a source position $\boldsymbol{\omega}(\theta)$. In the family of parallel lines $\{l_{t,\boldsymbol{\omega}(\theta)} : t \in \mathbb{R}\}$ there are two values of t, $t_0 < t_1$, such that the lines $l_{t_0,\boldsymbol{\omega}(\theta)}$ and $l_{t_1,\boldsymbol{\omega}(\theta)}$ are tangent to the boundary of D (Figure 1.13). Examining the diagram, it is clear that the orientation of the boundary at the point of tangency and that of the oriented line agree for $l_{t_1,\boldsymbol{\omega}}$, and are opposite for $l_{t_0,\boldsymbol{\omega}}$. Define h_D, the *shadow function* of D, by setting

$$h_D(\theta) = t_1 \text{ and } h_D(\theta + \pi) = -t_0. \tag{1.28}$$

The shadow function is completely determined by values of θ belonging to an interval of length π. Because $\boldsymbol{\omega}(\theta) = \boldsymbol{\omega}(\theta + 2\pi)$, the shadow function can be regarded as a 2π-periodic function defined on the whole real line. The mathematical formulation of reconstruction problem is as follows: Can the boundary of the region D be determined from h_D?

As $\boldsymbol{\omega}(\theta) = (\cos(\theta), \sin(\theta))$, the line $l_{h_D(\theta),\boldsymbol{\omega}(\theta)}$ is given parametrically by

$$\{h_D(\theta)(\cos(\theta), \sin(\theta)) + s(-\sin(\theta), \cos(\theta)) \mid s \in (-\infty, \infty)\}.$$

To determine the boundary of D, it would suffice to determine the point of tangency of $l_{h_D(\theta),\boldsymbol{\omega}(\theta)}$ with the boundary of D; in other words, we would like to find the function $s(\theta)$ so that for each θ,

$$(x(\theta), y(\theta)) = h_D(\theta)(\cos(\theta), \sin(\theta)) + s(\theta)(-\sin(\theta), \cos(\theta)) \tag{1.29}$$

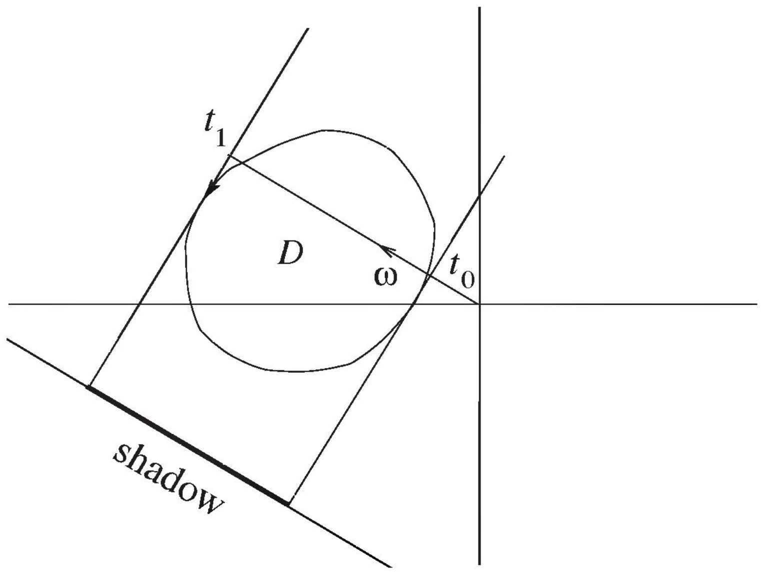


Figure 1.13. The measurement of the shadow.

is a point on the boundary of D. For the remainder of this section we suppose that s is differentiable.

The function s is found by recalling that, at the point of tangency, the direction of the tangent line to D is $\hat{\omega}(\theta)$. For a curve in the plane given parametrically by differentiable functions $(x(\theta), y(\theta))$, the direction of the tangent line is found by differentiating. At a parameter value θ_0 the direction of the tangent line is the same as that of the vector $(x'(\theta_0), y'(\theta_0))$. Differentiating the expression given in (1.29) and using the fact that $\partial_\theta \omega = \hat{\omega}$, we find that

$$(x'(\theta), y'(\theta)) = (h'_D(\theta) - s(\theta))\omega(\theta) + (h_D(\theta) + s'(\theta))\hat{\omega}(\theta). \tag{1.30}$$

Since the tangent line at $(x(\theta), y(\theta))$ is parallel to $\hat{\omega}(\theta)$ it follows from (1.30) that

$$h'_D(\theta) - s(\theta) = 0. \tag{1.31}$$

This gives a parametric representation for the boundary of a convex region in terms of its shadow function: If the shadow function is $h_D(\theta)$, then the boundary of D is given parametrically by

$$(x(\theta), y(\theta)) = h_D(\theta)\omega(\theta) + h'_D(\theta)\hat{\omega}(\theta). \tag{1.32}$$

Note that we have assumed that D is strictly convex and the $h_D(\theta)$ is a differentiable function. This is not always true; for example, if the region D is a polygon, then neither assumption holds.

Let D denote a convex region and h_D its shadow function. We can think of $D \mapsto h_D$ as a mapping from convex regions in the plane to 2π-periodic functions. It is reasonable to enquire if every 2π-periodic function is the shadow function of a convex region. The answer to this question is no. For strictly convex regions with smooth boundaries, we are able to characterize the range of this mapping. If h is twice differentiable, then the tangent vector to the curve defined by

$$(x(\theta), y(\theta)) = h(\theta)\boldsymbol{\omega}(\theta) + h'(\theta)\hat{\boldsymbol{\omega}}(\theta) \tag{1.33}$$

is given by

$$(x'(\theta), y'(\theta)) = (h''(\theta) + h(\theta))\hat{\boldsymbol{\omega}}(\theta).$$

In our construction of the shadow function, we observed that the tangent vector to the curve at $(x(\theta), y(\theta))$ and the vector $\hat{\boldsymbol{\omega}}(\theta)$ point in the same direction. From our formula for the tangent vector, we see that this implies that

$$h''(\theta) + h(\theta) > 0 \text{ for all } \theta \in [0, 2\pi]. \tag{1.34}$$

This gives a necessary condition for a twice differentiable function h to be the shadow function for a strictly convex region with a smooth boundary. Mathematically we are determining the range of the map that takes a convex body $D \subset \mathbb{R}^2$ to its shadow function h_D, under the assumption that h_D is twice differentiable. This is a convenient mathematical assumption, though in an applied context it is likely to be overly restrictive. The state space of the "system" which consists of strictly convex regions with smooth boundaries is parameterized by the set of smooth, 2π-periodic functions satisfying the inequality (1.34). This is an example of a system where the constraint defining the state space is an inequality rather than an equality.

Exercises

***Exercise* 1.2.8.** Justify the definition of $h_D(\theta+\pi)$ in (1.28) by showing that the orientation of the boundary at the point of tangency with $l_{t_0,\boldsymbol{\omega}(\theta)}$ agrees with that of $l_{-t_0,\boldsymbol{\omega}(\theta+\pi)}$.

***Exercise* 1.2.9.** Suppose that D_n is a regular n-gon. Determine the shadow function $h_{D_h}(\theta)$.

***Exercise* 1.2.10.** Suppose that D is a bounded, convex planar region. Show that the shadow function h_D is a continuous function of θ.

***Exercise* 1.2.11.** Suppose that h is a 2π-periodic, twice differentiable function that satisfies (1.34). Show that the curve given by (1.33) is the boundary of a strictly convex region.

***Exercise* 1.2.12.** How is the assumption that D is strictly convex used in the derivation of (1.31)?

***Exercise* 1.2.13.** If h is a differentiable function, then equation (1.33) defines a curve. By plotting examples, determine what happens if the condition (1.34) is not satisfied.

***Exercise* 1.2.14.** Suppose that h is a function satisfying (1.34). Show that the area of D_h is given by the

$$\text{Area}(D_h) = \frac{1}{2}\int_0^{2\pi} [(h(\theta))^2 - (h'(\theta))^2]\, d\theta.$$

Explain why this implies that a function satisfying (1.34) also satisfies the estimate

$$\int_0^{2\pi} (h'(\theta))^2 d\theta < \int_0^{2\pi} (h(\theta))^2 \, d\theta.$$

***Exercise* 1.2.15.** Let h be a smooth 2π-periodic function that satisfies (1.34). Prove that the curvature of the boundary of the region with this shadow function, at the point $h(\theta)\boldsymbol{\omega}(\theta) + h'(\theta)\hat{\boldsymbol{\omega}}(\theta)$, is given by

$$\kappa(\theta) = \frac{1}{h(\theta) + h''(\theta)}. \tag{1.35}$$

***Exercise* 1.2.16.** Suppose that h is a function satisfying (1.34). Show that another parametric representation for the boundary of the region with this shadow function is

$$\theta \mapsto \left(-\int_0^{\theta} (h(s) + h''(s))\sin(s)\, ds, \int_0^{\theta} (h(s) + h''(s))\cos(s)\, ds \right).$$

***Exercise* 1.2.17.** In this exercise we determine which positive functions κ defined on S^1 are the curvatures of closed strictly convex curves. Prove the following result: A positive function κ on S^1 is the curvature of a closed, strictly convex curve (parameterized by its tangent direction) if and only if

$$\int_0^{\infty} \frac{\sin(s)\, ds}{\kappa(s)} = 0 = \int_0^{\infty} \frac{\cos(s)\, ds}{\kappa(s)}.$$

***Exercise* 1.2.18.** Let D be a convex region with shadow function h_D. For a vector $\boldsymbol{v} \in \mathbb{R}^2$, define the translated region

$$D^{\boldsymbol{v}} = \{\boldsymbol{x} + \boldsymbol{v} \; : \; \boldsymbol{x} \in D\}.$$

Find the relation between h_D and $h_{D^{\boldsymbol{v}}}$. Explain why this answer is inevitable in light of the formula (1.35) for the curvature.

***Exercise* 1.2.19.** Let D be a convex region with shadow function h_D. For a rotation $A = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$, define the rotated region

$$D^A = \{A\boldsymbol{x} \; : \; \boldsymbol{x} \in D\}.$$

Find the relation between h_D and h_{D^A}.

Exercise* 1.2.20. If h_1 and h_2 are 2π-periodic functions satisfying (1.34) then they are the shadow functions of convex regions D_1 and D_2. The sum $h_1 + h_2$ also satisfies (1.34) and so is the shadow function of a convex region, D_3. Describe geometrically how D_3 is determined by D_1 and D_2.

Exercise* 1.2.21. Suppose that D is non-convex planar region. The shadow function h_D is defined as before. What information about D is encoded in h_D?

1.2.3 Approximate Reconstructions

See: A.6.2.

In a realistic situation the shadow function is measured at a finite set of angles

$$\{\theta_1, \ldots, \theta_m\}.$$

How can the data, $\{h_D(\theta_1), \ldots, h_D(\theta_m)\}$, be used to construct an approximation to the region D? We consider two different strategies; each relies on the special geometric properties of convex regions. Recall that a convex region always lies in one of the half-planes determined by the support line at any point of its boundary. Since the boundary of D and $l_{h(\theta),\boldsymbol{\omega}(\theta)}$ have the same orientation at the point of contact, it follows that D lies in each of the half-planes

$$H^-_{h(\theta_j),\boldsymbol{\omega}(\theta_j)}, \quad j = 1, \ldots, m;$$

see (1.27). As D lies in each of these half-planes, it also lies in their intersection. This defines a convex polygon

$$P_m = \bigcap_{j=1}^{m} H^-_{h(\theta_j),\boldsymbol{\omega}(\theta_j)}$$

that contains D. This polygon provides one sort of approximation for D from the measurement of a finite set of shadows. It is a stable approximation to D because small errors in the measurements of either the angles θ_j or the corresponding affine parameters $h(\theta_j)$ lead to small changes in the approximating polygon.

The difficulty with using the exact reconstruction formula (1.32) is that h is only known at finitely many values, $\{\theta_j\}$. From this information it is not possible to compute the exact values of the derivatives, $h'(\theta_j)$. We could use a finite difference approximation for the derivative to determine a finite set of points that approximate points on the boundary of D:

$$(x_j, y_j) = h(\theta_j)\boldsymbol{\omega}(\theta_j) + \frac{h(\theta_j) - h(\theta_{j+1})}{\theta_j - \theta_{j+1}}\hat{\boldsymbol{\omega}}(\theta_j).$$

If the measurements were perfect, the boundary of D smooth and the numbers $\{|\theta_j - \theta_{j+1}|\}$ small, then the finite difference approximations to $h'(\theta_j)$ would be accurate and these points would lie close to points on the boundary of D. Joining these points in the given order gives a polygon, P', that approximates D. If $\{h'(\theta_j\}$ could be computed exactly, then P' would be contained in D. With approximate values this cannot be asserted with certainty, though P' should be largely contained within D.

This gives a different way to reconstruct an approximation to D from a finite set of measurements. This method is not as robust as the first technique because it requires the measured data to be differentiated. In order for the finite difference $\frac{h(\theta_j)-h(\theta_{j+1})}{\theta_j-\theta_{j+1}}$ to be a good approximation to $h'(\theta_j)$, it is generally necessary for $|\theta_j - \theta_{j+1}|$ to be small. Moreover, the errors in the measurements of $h(\theta_j)$ and $h(\theta_{j+1})$ must also be small *compared to* $|\theta_j - \theta_{j+1}|$. This difficulty arises in solution of the reconstruction problem in x-ray CT; the exact reconstruction formula calls for the measured data to be differentiated.

In general, measured data are corrupted by noise, and noise is usually non-differentiable. This means that the measurements cannot be used directly to approximate the derivatives of a putative underlying smooth function. This calls for finding a way to improve the accuracy of the measurements. If the errors in individual measurements are random then repeating the same measurement many times and averaging the results should give a good approximation to the true value. This is the approach taken in magnetic resonance imaging. Another possibility is to make a large number of measurements at closely spaced angles $\{(h_j, j\Delta\theta) : j = 1, \dots, N\}$, which are then averaged to give less noisy approximations on a coarser grid. There are many ways to do the averaging. One way is to find a differentiable function, H, belonging to a family of functions of dimension $M < N$ that minimizes the *square error*

$$e(H) = \sum_{j=1}^{N} (h_j - H(j\Delta\theta))^2.$$

For example, H could be taken to be a polynomial of degree $M - 1$, or a continuously differentiable, piecewise cubic function. The reconstruction formula can be applied to H to obtain a different approximation to D. The use of averaging reduces the effects of noise but fine structure in the boundary is also blurred by any such procedure.

Exercises

***Exercise* 1.2.22.** Suppose that the angles $\{\theta_j\}$ can be measured exactly but there is an uncertainty of size ϵ in the measurement of the affine parameters, $h(\theta_j)$. Find a polygon $P_{m,\epsilon}$ that gives the best possible approximation to D and certainly contains D.

***Exercise* 1.2.23.** Suppose that we know that $|h''(\theta)| < M$, and the measurement errors are bounded by $\epsilon > 0$. For what angle spacing $\Delta\theta$ is the error, using a finite difference approximation for h', due to the uncertainty in the measurements equal to that caused by the nonlinearity of h itself?

1.2.4 Can an Object Be Reconstructed from Its Width?

To measure the location of the shadow requires an expensive detector that can accurately locate a transition from light to dark. It would be much cheaper to build a device, similar to the exposure meter in a camera, to measure the length of the shadow region without determining its precise location. It is therefore an interesting question whether or not the boundary of a region can be reconstructed from measurements of the *widths* of its shadows. Let $w_D(\theta)$ denote the width of the shadow in direction θ. A moment's consideration shows that

$$w_D(\theta) = h_D(\theta) + h_D(\theta + \pi). \tag{1.36}$$

Using this formula and Exercise 1.2.11, it is easy to show that w_D does *not* determine D. From Exercise 1.2.11 we know that if h_D has two derivatives such that $h_D'' + h_D > 0$, then h_D is the shadow function of a strictly convex region. Let e be an *odd* smooth function [i.e., $e(\theta) + e(\theta + \pi) \equiv 0$] such that

$$h_D'' + h_D + e'' + e > 0.$$

If $e \not\equiv 0$, then $h_D + e$ is the shadow function for D', a different strictly convex region. Observe that D' has the same *width* of shadow for each direction as D; that is,

$$w_D(\theta) = (h_D(\theta) + e(\theta)) + (h_D(\theta + \pi) + e(\theta + \pi)) = w_{D'}(\theta).$$

To complete this discussion, note that any function expressible as a series of the form

$$e(\theta) = \sum_{j=0}^{\infty} [a_j \sin(2j + 1)\theta + b_j \cos(2j + 1)\theta]$$

is an odd function. This is an infinite-dimensional space of functions. This implies that if w_D is the width of the shadow function for a convex region D, then there is an infinite-dimensional set of regions with the same width of the shadow function. Consequently, the simpler measurement is inadequate to reconstruct the boundary of a convex region. Figure 1.14 shows the unit disk and another region that has constant shadow width equal to 2.

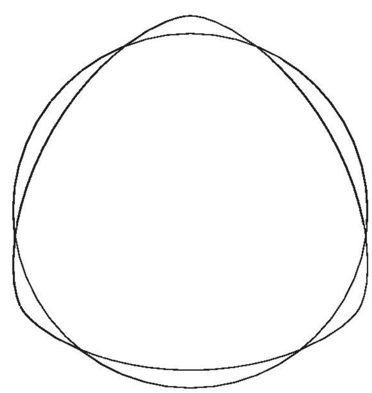


Figure 1.14. Two regions of constant width 2.

Exercises

***Exercise* 1.2.24.** Justify the formula (1.36) for the shadow width.

***Exercise* 1.2.25.** Show that the width function satisfies $w_D'' + w_D > 0$.

***Exercise* 1.2.26.** Is it true that every twice differentiable, π-periodic function, w satisfying $w'' + w > 0$ is the width function of a convex domain?

***Exercise* 1.2.27.** We considered whether or not a convex body is determined by the width of its shadows in order use a less expensive detector. The cheaper detector can only measure the width of the covered region. Can you find a way to use a detector that only measures the length of an illuminated region to locate the edge of the shadow? *Hint*: Cover only half of the detector with photosensitive material.

1.3 Conclusion

By considering examples, we have seen how physical systems can be described using mathematical models. The problem of determining the state of the system from measurements is replaced by that of solving equations or systems of equations. It is important to keep in mind that mathematical models are just models, indeed often toy models. A good model must satisfy two opposing requirements: The model should accurately depict the system under study while at the same time being simple enough to be usable. In addition, it must also have accurate, finite-dimensional approximations.

In mathematics, problems of determining the state of a physical system from feasible measurements are gathered under the rubric of *inverse problems*. The division of problems into inverse problems and *direct problems* is often a matter of history. Usually a physical theory that models how the state of the system determines feasible measurements preceded a description of the inverse process: how to use measurements to determine the state of the system. While many of the problems that arise in medical imaging are considered to be inverse problems, we do not give a systematic development of this subject. The curious reader is referred to the article by Joe Keller, [80], which contains analyses of many classical inverse problems or the book *Introduction to Inverse Problems in Imaging*, [10].

The models used in medical imaging usually involve infinitely many degrees of freedom. The state of the system is described by a function of continuous variables. Ultimately only a finite number of measurements can be made and only a finite amount of time is available to process them. Our analysis of the reconstruction process in x-ray CT passes through several stages. We begin with a description of the complete, perfect data situation. The measurement is described by a function on the space of lines. By finding an explicit inversion formula, we show that the state of the system can be determined from these measurements. The main tool in this analysis is the Fourier transform. We next consider the consequences of having only discrete samples of these measurements. This leads us to sampling theory and Fourier series. In the next chapter we quickly review linear algebra and the theory of linear equations, recasting this material in the language of measurement. The chapter ends with a brief introduction to the issues that arise in the extension of linear algebra to infinite-dimensional spaces.