Mathematical Analysis

by

Tom Lindstrøm

Department of Mathematics
University of Oslo
2016

Preface

The writing of this book started as an emergency measure when the textbook for the course MAT2400 failed to show up in the spring of 2011. Since then the project has been modified several times according to wishes and demands from students and faculty. In the 2016 version, I have added two new chapters (Chapter 2 on the foundation of calculus and Chapter 6 on differentiation in normed spaces) and removed all the material on measure and integration theory. I have also added two new sections to Chapter 5 on normed spaces and linear operators – some of this material is needed for the new Chapter 6. With these changes, the organization of the material on power series and function spaces had become rather absurd, and I have reorganized it for the current version in what seems a more logical and pedagogical way. This means that the only chapters that are relatively unaltered from last year, are Chapters 1, 3, and 7, although I have made some minor changes (and improvements?) to them as well.

I would like to thank everybody who has pointed out mistakes and weaknesses in previous versions, in particular Geir Ellingsrud, Erik Løw, Nils Henrik Risebro, Nikolai Bjørnestøl Hansen, Bernt Ivar Nødland, Simon Foldvik, Marius Jonsson (who also made the figure of vibrating strings in Chapter 7), Daniel Aubert, Lisa Eriksen, and Imran Ali.

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

Blindern, May 25th, 2016

Tom Lindstrøm

Contents

Chapter 1

Preliminaries: Proofs, Sets, and Functions

Chapters with the word "preliminaries" in the title are never much fun, but they are useful — they provide the reader with the background information necessary 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.

Mathematical 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 way. 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 $A \Longrightarrow B$. As already observed, $A \Longrightarrow B$ and $B \Longrightarrow A$ 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 $A \Longleftrightarrow B$. A typical example is:

"A triangle is equilateral if and only if all three angels are 60°"

When we want to prove that $A \Longleftrightarrow B$, it is often convenient to prove $A \Longrightarrow B$ and $B \Longrightarrow A$ separately.

If you think a little, you will realize that "$A \Longrightarrow B$" and "not-$B \Longrightarrow$ not-A" mean exactly the same thing — they both say that whenever A happens, so does B. This means that instead of proving "$A \Longrightarrow B$", we might just a well prove "not-$B \Longrightarrow$ not-A". This is called a *contrapositive proof*, and is convenient when the hypothesis "not-B" gives us more to work on 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 cancelling 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. □

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 $n = 1, 2, 3, \ldots$. A typical statement one may want to prove in this way, is

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

The basic observation behind the technique is:

1.1.3 (Induction Principle) *Assume that $P(n)$ is a statement about natural numbers $n = 1, 2, 3, \ldots$. Assume 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

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

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.

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

We then have

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

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 (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

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

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

$\mathbb{Q}$, 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

$$A = \{a \,|\, P(a)\}$$

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

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

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

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

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 | by a colon : and write $A = \{a : P(a)\}$ and $A = \{a \in B : P(a)\}$ instead of $A = \{a \,|\, P(a)\}$ and $A = \{a \in B \,|\, P(a)\}$, e.g. in expressions like $\{\|\alpha \mathbf{x}\| : |\alpha| < 1\}$ where there are lots of vertical bars already.

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

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

and

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

respectively. 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:

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

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

Proposition 1.2.1 *For all sets $B, A_1, A_2, \ldots, A_n$*

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

and

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

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. $x \in B \cap (A_1 \cup A_2 \cup \ldots \cup A_n)$. 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 $x \in (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)$. This proves that

$$B \cap (A_1 \cup A_2 \cup \ldots \cup A_n) \subseteq (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)$$

To prove the opposite inclusion, assume that $x \in (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)$. 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 $x \in A_1 \cup A_2 \cup \ldots \cup A_n$, and hence $x \in B \cap (A_1 \cup A_2 \cup \ldots \cup A_n)$. This proves that

$$B \cap (A_1 \cup A_2 \cup \ldots \cup A_n) \supseteq (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)$$

As we now have inclusion in both directions, (1.2.1) follows. □

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

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

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

$$A^c = U \setminus A = \{a \in U \,|\, a \notin A\}$$

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

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

$$(A_1 \cup A_2 \cup \ldots \cup A_n)^c = A_1^c \cap A_2^c \cap \ldots \cap A_n^c \tag{1.2.3}$$

and

$$(A_1 \cap A_2 \cap \ldots \cap A_n)^c = A_1^c \cup A_2^c \cup \ldots \cup A_n^c \tag{1.2.4}$$

(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 of De Morgan's laws: 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 $x \in (A_1 \cup A_2 \cup \ldots \cup A_n)^c$. Then $x \notin A_1 \cup A_2 \cup \ldots \cup A_n$, and hence for all i, $x \notin A_i$. This means that for all i, $x \in A_i^c$, and hence $x \in A_1^c \cap A_2^c \cap \ldots \cap A_n^c$.

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

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 pairs (a, b) where $a \in A$ and $b \in B$. If we have more sets $A_1, A_2, \ldots, A_n$, the cartesian product $A_1 \times A_2 \times \cdots \times A_n$ consists of all n-tuples $(a_1, a_2, \ldots, a_n)$ where $a_1 \in A_1, a_2 \in A_2, \ldots, a_n \in A_n$. If all the sets are the same (i.e. $A_i = A$ for all i), we usually write A^n instead of $A \times A \times \cdots \times A$. 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 $A_1 \cup A_2 \cup \ldots \cup A_n = U$ if and only if $A_1^c \cap A_2^c \cap \ldots \cap A_n^c = \emptyset$.
8. 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.,

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

and so do the laws of De Morgan:

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

Families are often given as *indexed sets*. This means we 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

$$\mathcal{A} = \{A_i \mid i \in I\},$$

and use notation such as

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

or alternatively

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

for unions and intersections

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

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

6. Prove De Morgan's laws for families:

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

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 $A, B \in \mathcal{A}$, the $A \cup B \in \mathcal{A}$.

   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 $A_1, A_2, \ldots, A_n \in \mathcal{A}$ for an $n \in \mathbb{N}$, then

$$A_1 \cup A_2 \cup \ldots \cup A_n \in \mathcal{A}$$

   (*Hint:* Use induction.)

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

$$A_1 \cap A_2 \cap \ldots \cap A_n \in \mathcal{A}$$

   (*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 $f : X \to Y$ from X to Y as a *rule* 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 three sets X, Y, Z and functions $f : X \to Y$ and $g : Y \to Z$, we can define a *composite function* $h : X \to Z$ by $h(x) = g(f(x))$. This composite function is often denoted by $g \circ f$, and hence $g \circ f(x) = g(f(x))$.

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*. 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*. 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 $f : X \to Y$ we have*

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

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 $x \in f^{-1}(\bigcup_{B\in\mathcal{B}} B)$. □

For forward images the situation is more complicated:

[1] Set theoretically, a function from X to Y is a subset f of $X \times Y$ such that for each $x \in A$, 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 in $y \in Y$ such that $(x, y) \in f$, and we are back to our usual notation.

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

$$f(\bigcup_{A\in\mathcal{A}} A) = \bigcup_{A\in\mathcal{A}} f(A) \quad \textit{and} \quad f(\bigcap_{A\in\mathcal{A}} A) \subseteq \bigcap_{A\in\mathcal{A}} f(A)$$

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 $\bigcup_{A\in\mathcal{A}} f(A) \subseteq f(\bigcup_{A\in\mathcal{A}})$. 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. □

Example 1: Let $X = \{x_1, x_2\}$ and $Y = \{y\}$. Define $f : X \to Y$ 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 $f(A_1\cap A_2) \neq f(A_1)\cap f(A_2)$. ♣

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 $x_1 \in A_1$ og $x_2 \in A_2$; 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 $f : X \to Y$ 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 $f : X \to Y$ we have*

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

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 $y \in f(\bigcap_{A\in\mathcal{A}} A)$. □

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

Proposition 1.4.4 *Assume that $f : X \to Y$ 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 complemet 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 $f(x) \in B^c$. □

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 does not hold in general.

We end this section by taking a look at three important properties a function can have. We have already defined a function $f : X \to Y$ 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 someimes called *injections, surjections, and bijections*, respectively.

If $f : X \to Y$ 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 $g : Y \to X$ by

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

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.

If $f : X \to Y$ and $g : Y \to Z$ are bijective, so is their composition $g \circ f$, and $(g \circ f)^{-1} = (f^{-1}) \circ (g^{-1})$ (see Exercise 7 below).

Exercises for Section 1.4

1. Let $f : \mathbb{R} \to \mathbb{R}$ be the function $f(x) = x^2$. Find $f([-1, 2])$ and $f^{-1}([-1, 2])$.
2. Let $g : \mathbb{R}^2 \to \mathbb{R}$ 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 $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ is neither injective nor surjective. What if we change the definition fo $f(x) = x^3$?
4. Show that a strictly increasing function $f : \mathbb{R} \to \mathbb{R}$ is injective. Does it have to be surjective?

5. Prove the second part of Proposition 1.4.1.

6. Find a function $f : X \to Y$ 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. In this problem f, g are functions $f : X \to Y$ and $g : Y \to Z$.

   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 $(g \circ f)^{-1} = (f^{-1}) \circ (g^{-1})$.

8. Given a set Z, we let $\mathrm{id}_Z : Z \to Z$ be the *identity map* $\mathrm{id}(z) = z$ for all $z \in Z$.

   a) Show that if $f : X \to Y$ is bijective with inverse function $g : Y \to X$, then $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$.

   b) Assume that $f : X \to Y$ and $g : Y \to X$ are two functions such that $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$. Show that f and g are bijective, and that $g = f^{-1}$.

9. 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 $f : X \to Y$, but it may also be used to denote inverse images $f^{-1}(B)$ of sets under arbitrary functions $f : X \to Y$. The only instances where this might have caused real confusion, is when $f : X \to Y$ 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 interpretation of C always 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 parallell, 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

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

for equality and

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

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 non-overlapping pieces. More precisely, if X is a set, a *partition* $\mathcal{P}$ of X is a family of 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

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

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:

$$[x] = \{y \in X \,|\, x \sim y\}$$

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 $[x] \subseteq [y]$. □

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

$$x \sim y \Longleftrightarrow x - y \in U$$

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 $x - y \in U$, and concequently $y - x = (-1)(x - y) \in U$ as subspaces are closed under multiplication by scalars. Hence $y \sim x$.

(iii) *Transitive:* If $x \sim y$ and $y \sim z$, then $x - y \in U$ and $y - z \in U$. Since subspaces are closed under addition, this means that $x - z = (x - y) + (y - z) \in U$, 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. ♣

.

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 an equivalence relation on $\mathcal{L}$.

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

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

   Show that $\sim$ is an equivalence relation. What does 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 on $\sim$ on $\mathcal{M}$ by

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

   Show that $\sim$ is an equivalence relation.

1.6 Countability

A set A is called *countable* if it possible to make a list $a_1, a_2, \ldots, a_n, \ldots$ 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 $A = \{a_1, a_2, \ldots, a_m\}$ are obviously countable[2] as they can be listed

$$a_1, a_2, \ldots, a_m, a_m, a_m, \ldots$$

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

(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

$$1, 2, 3, \ldots$$

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

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

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 $A \times B$.*

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

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

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 indicies sum to 2, then we list the elements $(a_2, b_1), (a_1, b_2)$ where the indicies sum to 3, then the elements $(a_3, b_1), (a_2, b_2), (a_1, b_3)$ where the indicies sum to 4 etc.) □

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

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

Proposition 1.6.2 *If the sets $A_1, A_2, \ldots, A_n, \ldots$ 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 $A_i = \{a_{i1}, a_{i2}, \ldots, a_{in}, \ldots\}$ be a listing of the i-th set. Then

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

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

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 $(a_1, b_1), (a_2, b_2), (a_3, b_3), \ldots$. But then $\frac{a_1}{b_1}, \frac{a_2}{b_2}, \frac{a_3}{b_3}, \ldots$ 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). □

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 $r_1, r_2, r_3, \ldots$. Let us write down the decimal expansions of the numbers on the list:

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

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

Exercises to Section 1.6

1. Show that a subset of a countable set is countable.
2. Show that if $A_1, A_2, \ldots A_n$ are countable, then $A_1 \times A_2 \times \cdots A_n$ is countable.
3. Show that the set of all finite sequences $(q_1, q_2, \ldots, q_k)$, $k \in \mathbb{N}$, of rational numbers is countable.
4. Show that if A is an *infinite*, countable set, then there is a list $a_1, a_2, a_3, \ldots$ 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) $f : A \to B$.
5. Show that the set of all subsets of $\mathbb{N}$ is uncountable (*Hint:* Try to modify the proof of Theorem 1.6.4.)

Chapter 2

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 in the rest of the book. How much new you will find here, depends on your calculus courses. Have you followed a fairly theoretical calculus sequence, 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 here 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 mathematical analysis, it becomes an indispensable tool that you have to master – the subject matter becomes so complicated that there is no other way to get a good grasp of it. 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 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 ϵ-δ 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 $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ in $\mathbb{R}^m$ is

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

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 $|x - y| = |y - x|$.

That the concept of distance between points is based on the notions of absolute values and norms may seem bad news to you if you are uncomfortable with these notions, but don't despair: there isn't really that much about absolute values and norms that you need to know 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 point; 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 our 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 $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $|x_n - a| < \epsilon$ for all $n \geq N$. We write $\lim_{n\to\infty} x_n = a$.*

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 this 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 $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{a}$*

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 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* $\{\mathbf{x}_n\}$ *og* $\{\mathbf{y}_n\}$ *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 alle $n \geq N_2$. 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. □

Remark: Many get confused when $\frac{\epsilon}{2}$ shows up in the proof above and takes over the rôle 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 $\mathbf{x}_n \to \mathbf{a}$, we can tackle any "epsilon-challenge", including half of the original epsilon.

Continuity

Let us now see how we can use the notion of distance to define continuity. Intuitively, one often says that a function $f : \mathbb{R} \to \mathbb{R}$ 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 $\epsilon > 0$ that we have chosen, 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 $f : \mathbb{R} \to \mathbb{R}$ is* continuous *at a point $a \in \mathbb{R}$ if for every $\epsilon > 0$ 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 judiciuos choice of δ. If the function is discontinuous at a, player II has a winning strategy – he can choose an $\epsilon > 0$ that no choice of $\delta > 0$ can parry.

Let us for a change take a look at 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.

Let us now take a look at a more complex example of the ϵ-δ-technique where we combine convergence and continuity.

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

Proof: Assume first that f is continuous at a, and that $\lim_{n\to\infty} x_n = a$. 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 $|x - a| < \delta$. But we know that x_n converges to a, and hence 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 $f(a)$). □

The proof above shows how we can combine different forms of dependence. Note in particular how old quantities reappear in new rôles – 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 rôles, 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 $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ 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 $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $\mathbf{a}$ if and only if $\lim_{k\to\infty} \mathbf{F}(\mathbf{x}_k) = \mathbf{F}(\mathbf{a})$ 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 $\mathbf{F} : A \to \mathbb{R}^m$ 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 many 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 student 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 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's take a look at the example:

Proposition 2.1.9 *Assume that $g : \mathbb{R} \to \mathbb{R}$ 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 $g(x) \to g(a)$, 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, $|g(a)|$, we can control quite easily as it is a constant. What about the other factor $|g(x)|$? Since $g(x) \to g(a) \neq 0$, 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 impression 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. □

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. Use the definition of continuity to show that if $f : \mathbb{R} \to \mathbb{R}$ is continuous at a point a, then the function $g(x) = Mf(x)$, where M is a constant, is also continuous at a.
3. Use the definition of continuity to show that if $f, g : \mathbb{R} \to \mathbb{R}$ are continuous at a point a, then so is $f + g$.
4. Use the definition of continuity to show that if $f, g : \mathbb{R} \to \mathbb{R}$ is 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.) Then combine this result 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.
5. Use the definition of continuity to show that if $f(x) = \frac{1}{\sqrt{x}}$ is continuous at all points $a > 0$.
6. 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 *upper bounded* 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 *lower bounded* 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 *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, upper bounded subset A of $\mathbb{R}$ 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 in not upper bounded, 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, lower bounded subset A of $\mathbb{R}$ 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 in not lower bounded, and we then put

$$\inf A = -\infty$$

Here is a simply 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's take a look at an example.

Example 2: The set

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

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

$$B = \{x \in \mathbb{Q} \,|\, x^2 < 2\}$$

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, but that the principle guarantees that we don't find similar gaps in $\mathbb{R}$. ♣

Now that we have realized that 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 trivilality, but is actually a very powerful tool.

Theorem 2.2.2 *All monotone, bounded sequences in* $\mathbb{R}$ *converge to a number in* $\mathbb{R}$.

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

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

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, $a - \epsilon$ can not be an upper bound, and hence there must be an a_N such that at $a_N > a - \epsilon$. 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. □

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

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

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. They will, however, always have upper and lower limits captured by the notions of *limit superior*, lim sup, and *limit inferior*, lim inf. You may not have seen these notions in your calculus courses, but we shall need them now and then later in the book.

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

$$M_n = \sup\{a_k \,|\, k \geq n\}$$

and

$$m_n = \inf\{a_k \,|\, k \geq n\}$$

We allow that $M_n = \infty$ and that $m_n = -\infty$ as may well occur. Note 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

$$\lim_{n\to\infty} M_n \qquad \text{and} \qquad \lim_{n\to\infty} m_n$$

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

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

and the *limit inferior* to be

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} m_n$$

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.

The following relationship should be no surprise:

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

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

if and only if

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

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

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

$$\lim_{n\to\infty} m_n = \liminf_{n\to\infty} a_n = b\,,$$
$$\lim_{n\to\infty} M_n = \limsup_{n\to\infty} a_n = b\,,$$

we clearly have $\lim_{n\to\infty} a_n = b$ by "squeezing".

We now assume that $\lim_{n\to\infty} a_n = b$ 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 $n \geq N$. Since this holds for all $\epsilon > 0$, we have $\limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = b$ □

Cauchy sequences

As there is no natural way to order the points in $\mathbb{R}^m$ when $m > 1$, it is not natural to use upper and lower bounds to describe the completeness of $\mathbb{R}^m$. Instead we shall use the notion of Cauchy sequences that also generalizes 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 idea level – in a space "without holes", the terms in a Cauchy sequence ought to be squeezed towards 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

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

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 $n \leq N$. 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. □

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

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

is an upper bound for the sequence and that

$$k = \min\{a_1, a_2, \ldots a_{N-1}, x_N - 1\}$$

is a lower bound. □

We can now complete the first part of our program. The proof relies on the Completeness Principle through Theorem 2.2.2 and Proposition 2.2.4.

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

$$M = \limsup_{n\to\infty} a_n \qquad \text{og} \qquad m = \liminf_{n\to\infty} a_n$$

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 $n \geq N$. This means that $m_k \geq a_N - \epsilon$ and $M_k \leq a_N + \epsilon$ for $k \geq N$. Consequently $M_k - m_k \leq 2\epsilon$ for all $k \geq N$. Hence $M - m \leq 2\epsilon$ for every $\epsilon > 0$, and this is only possible if $M = m$. □

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

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

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 $\{\mathbf{a}_n\}$. □

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 converge and deduce the Completeness Principle. The Complete 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 generalixe 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 $\limsup_{n\to\infty} a_n$ and $\liminf_{n\to\infty} a_n$.
5. Let $a_n = \cos\frac{n\pi}{2}$. Find $\limsup_{n\to\infty} a_n$ and $\liminf_{n\to\infty} a_n$.
6. Complete the proof of Proposition 2.2.3 for the cases $b = \infty$ and $b = -\infty$.
7. Show that
$$\limsup_{n\to\infty}(a_n + b_n) \le \limsup_{n\to\infty} a_n + \limsup_{n\to\infty} b_n$$
and
$$\liminf_{n\to\infty}(a_n + b_n) \ge \liminf_{n\to\infty} a_n + \liminf_{n\to\infty} b_n$$
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 $\limsup_{n\to\infty} a_n b_n = ab$ (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 lim sup and lim inf for functions $f : \mathbb{R} \to \mathbb{R}$. Let $a \in \mathbb{R}$, and define (note that we exclude $x = a$ in these definitions)
$$M_\epsilon = \sup\{f(x) \mid x \in (a-\epsilon, a+\epsilon),\ x \ne a\}$$
$$m_\epsilon = \inf\{f(x) \mid x \in (a-\epsilon, a+\epsilon),\ x \ne a\}$$
for $\epsilon > 0$ (we allow $M_\epsilon = \infty$ and $m_\epsilon = -\infty$).
   a) Show that M_ϵ decreases and m_ϵ increases as $\epsilon \to 0$.
   b) Show that $\limsup_{x\to a} f(x) = \lim_{\epsilon\to 0^+} M_\epsilon$ and $\liminf_{x\to a} f(x) = \lim_{\epsilon\to 0^+} m_\epsilon$ exist (we allow $\pm\infty$ as values).
   c) Show that $\lim_{x\to a} f(x) = b$ if and only if $\limsup_{x\to a} f(x) = \liminf_{x\to a} f(x) = b$

d) Find $\liminf_{x\to 0} \sin\frac{1}{x}$ and $\limsup_{x\to 0} \sin\frac{1}{x}$

10. Assume that $\{\mathbf{a}_n\}$ is a sequence in $\mathbb{R}^m$, and write the terms on component form
$$\mathbf{a}_n = (a_n^{(1)}, a_n^{(2)}, \ldots, a_n^{(m)})$$
Show that $\{\mathbf{a}_n\}$ converges if and only if all of the component sequences $\{a_n^{(k)}\}$, $k = 1, 2, \ldots, m$ 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 cannot change sign without intersecting the x-axis.

Theorem 2.3.1 (The Intermediate Value Theorem) *Assume that $f : [a, b] \to \mathbb{R}$ 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, $f(c) = \lim_{n\to\infty} f(x_n)$, 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 $z_n \to c$. Using proposition 2.1.5 again, we get $f(c) = \lim_{n\to\infty} f(z_n)$, 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 $f(c) = 0$. □

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

Example 1: Define a function $f : \mathbb{Q} \to \mathbb{Q}$ 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 doesn't hold if we replace $\mathbb{R}$ by $\mathbb{Q}$.

What is happening here? The function graph sneaks trough 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

$$n_1 < n_2 < \ldots < n_k < \ldots$$

are the indicies 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 nestet in shorter and shorter intervals, $\{y_k\}$ is a Cauchy sequence and hence converges. □

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

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

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. □

We shall see a typical example of how the Bolzano-Weierstrass Theorem is used in the proof of the next result.

The Extreme Value Theorem

Finding maximal and minimal values of functions are 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 $f : [a, b] \to \mathbb{R}$ 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 $f(x_n) \to M$

(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 $c = \lim_{k\to\infty} y_k$ belongs to $[a, b]$. By construction $f(y_k) \to M$, but on the other hand, $f(y_k) \to f(c)$ 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 $[a, b]$. □.

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

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

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

Proof: Assume for contradiction that $f'(c) > 0$ (the case where $f'(c) < 0$ can be treated similarly). Since

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

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. □

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

Lemma 2.3.6 (Rolle's Theorem) *Assume that $f : [a, b] \to \mathbb{R}$ 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: If f is constant, $f'(x) = 0$ at *all* inner points x, and there is nothing more to prove. According to the Extreme Value Theorem, the function has minimum and maximum points, and if it is not constant, at least one of these must be at an inner point c (here we are using that the value at the end points are equal). According to the previous lemma, $f'(c) = 0$. □

We are now ready to prove the theorem. It says that for differentiable functions there is in each interval a point where the instantaneous growth of the function equals it average growth over the interval.

Theorem 2.3.7 (The Mean Value Theorem) *Assume that $f : [a, b] \to \mathbb{R}$ 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

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

□

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 in higher (including infinite!) dimensions.

Exercises for section 2.3

In exercises 1-4 you are asked to show that the results above would not have hold if 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.

1. Show that the function $f : \mathbb{Q} \to \mathbb{Q}$ defined by $f(x) = \frac{1}{x^2-2}$ is continuous at all $x \in \mathbb{Q}$, but that it is unbounded on $[0, 2]$. Compare to the Extremal Value Theorem.

2. Show that the function $f : \mathbb{Q} \to \mathbb{Q}$ defined by $f(x) = x^3 - 6x$ is continuous at all $x \in \mathbb{Q}$, but that it does not have a maximum in $[0, 2]_{\mathbb{Q}}$, where $[0, 2]_{\mathbb{Q}} = [0, 2] \cap \mathbb{Q}$. Compare to the Extremal Value Theorem.

3. Show that the function $f : \mathbb{Q} \to \mathbb{Q}$ defined by $f(x) = x^3 - 9x$ satisfies $f(0) = f(3) = 0$, but that there are no rational points in the interval $[0, 3]$ 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.
8. Carry out the proof of Lemma 2.3.5 when $f'(c) < 0$.
9. Assume that f og 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 $x, y \in [a, b]$.

Chapter 3

Metric Spaces

Many of the arguments you have seen in several variable calculus are almost identical to the corresponding arguments in one 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 that the notion of distance between numbers that you need in one 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 suprisingly similar to those we have already seen in one and several variable 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 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 $d : X \times X \to \mathbb{R}$ which measures the distance $d(x, y)$ beween 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 pont 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 non-empty set X and a function $d : X \times X \to [0, \infty)$ such that:*

(i) (Positivity) For all $x, y \in X$, $d(x, y) \geq 0$ with equality if and only if $x = y$.

(ii) (Symmetry) For all $x, y \in X$, $d(x, y) = d(y, x)$.

(iii) (Triangle inequality) For all $x, y, z \in X$

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

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 $d(x, y) = |x-y|$, $(\mathbb{R}, d)$ 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

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

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 $\mathbf{x} = (x_1, x_2, \ldots, x_K)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_K)$ 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 coding 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 $f : [a, b] \to \mathbb{R}$. Then

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

is a metric on X. This metric determines the distance beween two functions by measuring the distance at the x-value where the graphs are most apart. This means that the distance between two functions may be large even if the functions 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 $f(x)$ og $g(x)$ 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 are not very useful in applications, they are important to as they are totally different from the metrics we are used to from $\mathbb{R}^n$ and may help sharpen our intuition of how a metric can be. Let X be any non-empty 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 non-empty 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 a new class of 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 *from (X, d_X) to (Y, d_Y) is a bijection $i : X \to Y$ 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 $i : X \to Y$ 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 (real) vector space, a function $\|\cdot\| : V \to \mathbb{R}$ 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 $x_1, x_2, \ldots, x_n$ 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. In this problem you can use the Inverse Triangle Inequality.

   a) Assume that $\{x_n\}$ is a sequence in a metric space X converging to x. Show that $d(x_n, y) \to d(x, y)$ 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 $d(x_n, y_n) \to d(x, y)$.

9. Assume that d_1 og 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.

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

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

    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 $X \times Y$.

11. Let X be a non-empty set, and let $\rho : X \times X \to \mathbb{R}$ 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 $d : X \times X \to \mathbb{R}$ by

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

    Show that d is a metric on X.

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

13. 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 $i : \mathbb{R}^n \to \mathbb{R}^n$ given by $i(\mathbf{x}) = U\mathbf{x} + \mathbf{b}$ is an isometry.

3.2 Convergence and continuity

We begin our study of metric spaces by defining convergence. A sequence $\{x_n\}$ in a metric space X is just an ordered collection $\{x_1, x_2, x_3, \ldots, x_n, \ldots\}$ of elements in X enumerated by the natural numbers.

Definition 3.2.1 *Let (X, d) be a metric space. A sequencee $\{x_n\}$ in X* converges *to a point $a \in X$ if there for every $\epsilon > 0$ exists an $N \in \mathbb{N}$ such that $d(x_n, a) < \epsilon$ for all $n \geq N$. We write $\lim_{n\to\infty} x_n = a$ or $x_n \to a$.*

Note that this definition exactly mimics the definition of convergence in $\mathbb{R}$ and $\mathbb{R}^n$. Here is an alternative formulation.

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

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. □

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 can not be certain without a proof.

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

Proof: Assume that $\lim_{n\to\infty} x_n = a$ and $\lim_{n\to\infty} x_n = b$. 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

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

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

Note how we use 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

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

As the terminology suggests, we think of $\mathrm{B}(a; r)$ as a ball around a with radius r. Note that $x \in \mathrm{B}(a; r)$ 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 $\mathrm{B}(a; \epsilon)$ around a.

Let us now see how we can define continuity in metric spaces.

Definition 3.2.4 *Assume that (X, d_X), (Y, d_Y) are two metric spaces. A function $f : X \to Y$ 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 $d_X(x, a) < \delta$.*

This definition says exactly the same as 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 og d_Y to measure the distances.

A more geometric formulation of the definition is to say that for any open ball $\mathrm{B}(f(a);\epsilon)$ around $f(a)$, there is an open ball $\mathrm{B}(a,\delta)$ around a such that $f(\mathrm{B}(a;\delta)) \subseteq \mathrm{B}(f(a);\epsilon)$ (make a drawing!).

There is a close connection between continuity and convergence which 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 $f : X \to Y$ 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: $(i) \Longrightarrow (ii)$: 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 $d_Y(f(x_n), f(a)) < \epsilon$ 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 $n \geq N$.

$(ii) \Longrightarrow (i)$ We argue contrapositively: Assume 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 $f(a)$. □

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 $f : X \to Y$ and $g : Y \to Z$ are two functions, and let $h : X \to Z$ 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. □

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

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

Occasionally, we need to study functions which 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 $f : A \to Y$ 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 often 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 $f : A \to Y$:*

(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 $f, g : X \to \mathbb{R}$ 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 $f : X \to \mathbb{R}$ 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 $f : X \to Y$ 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 $id : \mathbb{R} \to \mathbb{R}$ be the identity function $id(x) = x$. Show that

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

is continuous, but that

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

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

8. 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 $f : X \to Y$ 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 i part b), but this time we have a function $g : Y \to X$. 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

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

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

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

   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*

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

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

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

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 i Section 3.1) consists either of only one point or the entire space X.

If A is a subset of X and x is a point in X, there are three possibilities:

(i) There is a ball $\mathrm{B}(x; r)$ around x which is contained in A. In this case x is called an *interior* point of A.

(ii) There is a ball $\mathrm{B}(x; r)$ around x which is contained in the complement A^c. In this case x is called an *exterior point* of A.

(iii) All balls $\mathrm{B}(x; r)$ 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.

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 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. 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.

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 $\mathrm{B}(a; r)$ 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 can not have any. This means that A is open. □

The following observation may seem obvious, but needs to be proved:

Lemma 3.3.5 *All open balls* $\mathrm{B}(a;r)$ *are open sets, while all closed balls* $\overline{\mathrm{B}}(a;r)$ *are closed sets.*

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

The next result shows that closed sets are indeed closed as far as sequences are concerned:

Proposition 3.3.6 *Assume that* F *is a subset of a metric space* X*. The following are equivalent:*

(i) F *is closed.*

(ii) If $\{x_n\}$ *is a convergent sequence of elements in* F*, then the limit* $a = \lim_{n\to\infty} x_n$ *always 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 $\mathrm{B}(a;\epsilon)$ around a which only contains points from the complement of F. But then a sequence from F can never get inside $\mathrm{B}(a,\epsilon)$, and hence cannot converge to a.

Assume now that 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 $\mathrm{B}(a;\frac{1}{n})$. Then $\{x_n\}$ is a sequence from F that converges to a point a which is not in F. □

An open set containing x is called a *neighborhood* of x[1]. The next result is rather silly, but also quite useful.

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

Lemma 3.3.7 *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 $\mathrm{B}(x_0, r) \subseteq U_{x_0}$. But then $\mathrm{B}(x_0, r) \subseteq U$, which shows that x_0 is an interior point of U. □

In Proposition 3.2.5 we gave a characterization of continuity in terms of sequences. We shall now prove three characterizations in terms of open and closed sets. The first one characterizes continuity at a point.

Proposition 3.3.8 *Let $f : X \to Y$ 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: $(i) \Longrightarrow (ii)$: 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 $U = \mathrm{B}(x_0, \delta)$.

$(ii) \Longrightarrow (i)$ 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 neighbohood 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. □

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

Proposition 3.3.9 *The following are equivalent for a function $f : X \to Y$ 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: $(i) \Longrightarrow (ii)$: 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 3.3.7, $f^{-1}(V)$ is open.

$(ii) \Longrightarrow (i)$ Assume that the inverse images of open sets are open. To prove that f is continuous at an arbitrary point x_0, Proposition 3.3.8 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 $U = f^{-1}(V)$. □

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, we can translate it into a statement about closed sets:

Proposition 3.3.10 *The following are equivalent for a function $f : X \to Y$ 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: $(i) \Longrightarrow (ii)$: 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.

$(ii) \Longrightarrow (i)$ 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. □

Mathematicians usually sum up the last two theorems by saying that openness and closedness are preserved under inverse, continuous images. Be aware that these properties 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 $f, g : \mathbb{R} \to \mathbb{R}$ be the continuous functions defined by

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

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. ♣

We end this section with two simple but useful observations on open and closed sets.

Proposition 3.3.11 *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 $G_1, G_2, \ldots, G_n$ is a finite *collection of open sets, then the intersection $G_1 \cap G_2 \cap \ldots \cap G_n$ is open.*

Proof: Left to the reader (see Exercise 12, where you are also asked to show that the intersection of infinitely many open sets is not necessarily open). □

Proposition 3.3.12 *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 $F_1, F_2, \ldots, F_n$ is a finite *collection of closed sets, then the union $F_1 \cup F_2 \cup \ldots \cup F_n$ is closed.*

Proof: Left to the reader (see Exercise 13, where you are also asked to show that the union of infinitely many closed sets is not necessarily closed). □

Propositions 3.3.11 and 3.3.12 are the starting points for *topology*, an even more abstract theory of nearness.

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 $f : X \to Y$ are continuous.
2. Give a geometric description of the ball $\mathrm{B}(a; r)$ 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 non-empty, 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. Prove the second part of Lemma 3.3.5, i.e. prove that a closed ball $\overline{\mathrm{B}}(a;r)$ is always a closed set.
5. Assume that $f : X \to Y$ and $g : Y \to Z$ are continuous functions. Use Proposition 3.3.9 to show that the composition $g \circ f : X \to Z$ is continuous.
6. 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.
7. Assume that A is a subset of a metric space X. The *interior* A° of A is the set consisting of all interior points of A. Show that A° is open.
8. Assume that A is a subset of a metric space X. The *closure* $\overline{A}$ of A is the set consisting of all interior points plus all boundary points of A.
   a) Show that $\overline{A}$ is closed.
   b) Let $\{a_n\}$ be a sequence from A converging to a point a. Show that $a \in \overline{A}$.
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 $G \subseteq A$ 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)
10. 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)
11. Let (X, d) be a metric space and give $\mathbb{R}$ the usual metric. Assume that $f : X \to \mathbb{R}$ 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 $a \in \mathbb{R}$.
12. Prove Proposition 3.3.11. Find an example of an infinite collection of open sets $G_1, G_2, \ldots$ whose intersection is *not* open.
13. Prove Proposition 3.3.12. Find an example of an infinite collection of closed sets $F_1, F_2, \ldots$ whose union is *not* closed.

3.4 Complete spaces

One of the reasons why calculus in $\mathbb{R}^n$ is so successful, is that $\mathbb{R}^n$ is a complete space. We shall now generalize this notion to metric spaces. The key concept is that of a Cauchy sequence:

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. □

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

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

We know from Section 2.2 that $\mathbb{R}^n$ is complete, but that $\mathbb{Q}$ is not when we use the usual metric $d(x, y) = |x - y|$. The complete spaces are in many ways the "nice" metric spaces, and we shall spend much time studying their properties. We shall also spend some time showing how we can make non-complete spaces complete. Example 5 in Section 3.1 (where X is the space of all continuous $f : [a, b] \to \mathbb{R}$) shows some interesting cases; X with the metric d_1 is complete, but not X with the metrics d_2 and d_3. 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.

The following proposition is quite 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 3.3.6 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 $\mathrm{B}(a, \frac{1}{n})$ 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. □

The nice thing about complete spaces is that we can prove that sequences converge to a limit without actually constructing or specifying the limit — all we need is to prove that the sequence is a Cauchy sequence. To prove that a sequence has the Cauchy property, we only need to work with the given terms of the sequence and not the unknown limit, and this often makes the arguments easier. As an example of this technique, we shall now prove an important theorem that will be useful later in the book, but first we need some definitions.

A function $f : X \to X$ 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. 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 $f^{\circ n}(x) = f(f(f(\ldots f(x)\ldots)))$ is the result of iterating f exactly n times. If $f(a) = a$, we say that a is a *fixed point* for f.

Theorem 3.4.5 (Banach's Fixed Point Theorem) *Assume that (X, d) is a complete metric space and that $f : X \to X$ 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*

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

converges to a.

Proof: Let us first show that f can not 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

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

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 $n \to \infty$, 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

$$d(x_n, x_{n+k}) \leq d(x_n, x_{n+1}) + d(x_{n+1}, x_{n+2}) + \ldots + d(x_{n+k-1}, x_{n+k}) =$$

$$= d(f^{\circ n}(x_0), f^{\circ n}(x_1)) + d(f^{\circ (n+1)}(x_0), f^{\circ (n+1)}(x_1)) + \ldots$$

$$\ldots + d(f^{\circ (n+k-1)}(x_0), f^{\circ (n+k-1)}(x_1)) \leq$$

$$\leq s^n d(x_0, x_1) + s^{n+1} d(x_0, x_1) + \ldots + s^{n+k-1} d(x_0, x_1) =$$

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

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. □

In Section 4.7 we shall use Banach's Fixed Point Theorem to prove the existence of solutions to quite general differential equations.

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* $\mathrm{diam}(A)$ of A is defined by
$$\mathrm{diam}(A) = \sup\{d(x, y) \mid x, y \in A\}$$
Let $\{A_n\}$ be a collection of subsets of X such that $A_{n+1} \subseteq A_n$ and $\mathrm{diam}(A_n) \to 0$, 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 $f : [0, 1] \to [0, 1]$ 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 $f : X \to X$ 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 sequence $\{y_n\}$ 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
$$n_1 < n_2 < n_3 < \ldots < n_k < \ldots$$
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 does 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* compact *if every sequence in K has a subsequence converging to a point in K. The space (X, d) is* compact *if X a compact set, i.e. if all sequences in X have a convergent subsequence.*

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 can not be compact, and then we show that if K is not bounded, it can not 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 does 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 $\lim_{k\to\infty} d(y_k, b) = \infty$. It is easy to see that $\{y_k\}$ can not 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 $d(y_k, b) \to \infty$, we must have $d(y_k, y) \to \infty$. Hence $\{x_n\}$ has no convergent subsequences, and K can not be compact. □

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 2.3.2: 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, $\mathbf{a} \in A$. □

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 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. □

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. □

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. □

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 $f : X \to Y$ 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 $y = f(x) \in f(K)$. □

So far we have only proved technical results about the nature of compact sets. The next result gives the first indication why these sets are useful.

Theorem 3.5.10 (The Extreme Value Theorem) *Assume that K is a non-empty, compact subset of a metric space (X, d) and that $f : K \to \mathbb{R}$ 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

$$M = \sup\{f(x) \mid x \in K\}$$

(if F is unbounded, we put $M = \infty$) and choose a sequence $\{x_n\}$ in K such that $\lim_{n\to\infty} f(x_n) = M$. Since K is compact, $\{x_n\}$ has a subsequence $\{x_{n_k}\}$ converging to a point $c \in K$. Then on the one hand $\lim_{k\to\infty} f(x_{n_k}) = M$, and on the other $\lim_{k\to\infty} f(x_{n_k}) = f(c)$ 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. □.

Remark: As already mentioned, it is possible to give a shorter proof of the Extreme Value Theorem by using Proposition 3.5.8. According to it, 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 $\mathrm{B}(a_1, \epsilon), \mathrm{B}(a_2, \epsilon), \ldots, \mathrm{B}(a_n, \epsilon)$ of balls with centers in A and radius ϵ that cover A (i.e. $A \subseteq \mathrm{B}(a_1, \epsilon) \cup \mathrm{B}(a_2, \epsilon) \cup \ldots \cup \mathrm{B}(a_n, \epsilon)$).*

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 $\mathrm{B}(x_1, \epsilon)$ does not cover A, we can choose $x_2 \in A \setminus \mathrm{B}(x_1, \epsilon)$. Since $\mathrm{B}(x_1, \epsilon)$ and $\mathrm{B}(x_2, \epsilon)$ do not cover A, we can choose $x_3 \in A \setminus \big(\mathrm{B}(x_1, \epsilon) \cup \mathrm{B}(x_2, \epsilon)\big)$. Continuing in this way, we get a sequence $\{x_n\}$ such that

$$x_n \in A \setminus \big(\mathrm{B}(x_1, \epsilon) \cup \mathrm{B}(x_2, \epsilon) \cup \ldots \cup (\mathrm{B}(x_{n-1}, \epsilon)\big)$$

This means that $d(x_n, x_m) \geq \epsilon$ for all $n, m \in \mathbb{N}$, $n > m$, and hence $\{x_n\}$ has no convergent subsequence. □

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.

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 $B_1^1, B_2^1, \ldots, B_{k_1}^1$ 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 $B_1^2, B_2^2, \ldots, B_{k_2}^2$ of radius $\frac{1}{2}$ that cover A. At least one of these ball must contain infinitely many of the terms from the sequence that lies 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

$$S_1 \cap \ S_2 \cap \ldots \cap S_k$$

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 that 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

$$x_{n_k} \in S_1 \cap S_2 \cap \ldots \cap S_k$$

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. □

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. Assume that (X, d) is a metric space and that $f : X \to [0, \infty)$ 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.
6. Let (X, d) be a compact metric space, and assume that $f : X \to \mathbb{R}$ 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 $x \in X$.

7. Assume that $f : X \to Y$ 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.

8. 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.

9. The Bolzano-Weierstrass Theorem 2.3.3 says that any bounded sequence in $\mathbb{R}^n$ has a convergent subsequence. Use it to prove that a subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.

10. Let (X, d) be a metric space.

    a) Assume that $K_1, K_2, \ldots, K_n$ is a finite collection of compact subsets of X. Show that the union $K_1 \cup K_2 \cup \ldots \cup K_n$ 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.

11. Let (X, d) be a metric space. Assume that $\{K_n\}$ is a sequence of non-empty, compact subsets of X such that $K_1 \supseteq K_2 \supseteq \ldots \supseteq K_n \supseteq \ldots$. Prove that $\bigcap_{n \in \mathbb{N}} K_n$ is non-empty.

12. Let (X, d_X) and (Y, d_Y) be two metric spaces. Assume that (X, d_X) is compact, and that $f : X \to Y$ is bijective and continuous. Show that the inverse function $f^{-1} : Y \to X$ is continuous.

13. 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.

14. Assume that (X, d) is compact and that $f : X \to X$ 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, so-called *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 each open covering $\mathcal{O}$ of K, there is a* finite *number of elements $O_1, O_2, \ldots, O_n$ in $\mathcal{O}$ such that*

$$K \subseteq O_1 \cup O_2 \cup \ldots \cup O_n$$

(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 a sequence $\{x_n\}$ converges to x. Then there must be an open ball $\mathrm{B}(x; r)$ 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 $\{x_n\}$ which 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 $\mathrm{B}(x; r_x)$ around x which only contains finitely many terms of the sequence. The family $\{\mathrm{B}(x, r_x) \,:\, x \in K\}$ is an open covering of K, but it cannot have a finite subcovering since any such subcovering $\mathrm{B}(x_1, r_{x_1}), \mathrm{B}(x_2, r_{x_2}), \ldots, \mathrm{B}(x_m, r_{x_m})$ can only contain finitely many of the infinitely many terms in the sequence. □

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 spece X. Define a function $f : A \to \mathbb{R}$ by*

$$f(x) = \sup\{r \in \mathbb{R} \,|\, r < 1 \text{ and } \mathrm{B}(x;r) \subseteq O \text{ for some } O \in \mathcal{O}\}$$

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 $\mathrm{B}(x;r) \subseteq O$. 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 out 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 $\mathrm{B}(x,r) \subseteq O$. For any such r, $\mathrm{B}(y, r - d(x,y)) \subseteq O$ since $\mathrm{B}(y, r - d(x,y)) \subset \mathrm{B}(x,r)$. 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. □

We are now ready for the main theorem:

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 Extremal Value Theorem, 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 $\mathcal{B}(x, \frac{r}{2})$ is contained in a set $O \in \mathcal{B}$. Since K is compact, it is totally bounded, and hence there is a finite collection of balls $\mathcal{B}(x_1, \frac{r}{2})$, $\mathcal{B}(x_2, \frac{r}{2})$, $\ldots$, $\mathcal{B}(x_n, \frac{r}{2})$ that covers K. Each ball $\mathcal{B}(x_i, \frac{r}{2})$ is contained in a set $O_i \in \mathcal{O}$, and hence $O_1, O_2, \ldots, O_n$ is a finite subcovering of $\mathcal{O}$. □

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

$$K \cap F_1 \cap F_2 \cap \ldots \cap F_n \neq \emptyset$$

for all finite collections $F_1, F_2, \ldots, F_n$ 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 $I_1, I_2, \ldots, I_n$ of sets from $\mathcal{I}$ such that
$$[0,1] \subseteq I_1 \cup I_1 \cup \ldots \cup I_n$$

2. Let $\{K_n\}$ be a decrasing 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 exactly the same problem as 3.5.11, but this time you should do it with the methods in this section).

3. Assume that $f : X \to Y$ 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 $K_1, K_2, \ldots, K_n$ are compact subsets of a metric space X. Use the open covering property to show that $K_1 \cup K_2 \cup \ldots \cup K_n$ is compact.

5. Use the open covering property to show that a closed subset of a compact set is compact.

6. Assume that $f : X \to Y$ 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 $\{\mathrm{B}(x, \frac{\delta(x)}{2})\}_{x\in K}$ is an open covering of K, and that it has a finite subcovering $\mathrm{B}(x_1, \frac{\delta(x_1)}{2})$, $\mathrm{B}(x_2, \frac{\delta(x_2)}{2})$, $\ldots$, $\mathrm{B}(x_n, \frac{\delta(x_n)}{2})$.

   c) Put $\delta = \min\{\frac{\delta(x_1)}{2}, \frac{\delta(x_2)}{2}, \ldots, \frac{\delta(x_n)}{2}\}$, 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 $K \cap \left(\bigcap_{F\in\mathcal{F}} F\right) \neq \emptyset$ 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 important 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, but before we turn to this, we need to take a look at a concept that will be important in many different contexts throughout the book.

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

$$y_1 = 1.4 = \frac{14}{10},\ \ y_2 = 1.41 = \frac{141}{100},\ \ y_3 = 1.414 = \frac{1414}{1000},\ \ y_4 = 1.4142 = \frac{14142}{10000}, \ldots$$

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 isometric[2], they are the same for all practical purposes, and we usually talk about *the* completion of a metric space.

[2]Recall from Section 3.1 that an *isometry* from (X, d_X) to (Y, d_Y) is a bijection $i : X \to Y$ 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 isomorphic; i.e. when there is an isomorphism between them.

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 $i : Y \to Z$. 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 $d_Z(x_n, \hat{x}_n) = d_X(x_n, \hat{x}_n) = d_Y(x_n, \hat{x}_n) \to 0$ as $n \to \infty$.

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

$$d_Z(i(y), i(\hat{y})) = \lim_{n\to\infty} d_Z(x_n, \hat{x}_n) = \lim_{n\to\infty} d_X(x_n, \hat{x}_n) =$$

$$= \lim_{n\to\infty} d_Y(x_n, \hat{x}_n) = d_Y(y, \hat{y})$$

(we are using repeatedly that if $\{u_n\}$ and $\{v_n\}$ are sequences in a metric space converging to u and v, respectively, then $d(u_n, v_n) \to d(u, v)$, see Exercise 3.1.8 b). It remains to prove that i is a bijection. Injectivity follows immediately from 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. □

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 $\sim$ on $\mathcal{X}$ by defining

$$\{x_n\} \sim \{y_n\} \quad \Longleftrightarrow \quad \lim_{n\to\infty} d(x_n, y_n) = 0$$

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

$$\bar{d}([x_n], [y_n]) = \lim_{n\to\infty} d(x_n, y_n)$$

We now have our completion $(\bar{X}, \bar{d})$. 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 $\bar{x} =$

$[x, x, x \ldots]$ be the equivalence class of the constant sequence $\{x, x, x, \ldots\}$, and put

$$D = \{\bar{x} \,|\, x \in X\}$$

We then prove that D is dense in X and that 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's 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 $\lim_{n\to\infty} d(x_n, y_n)$ exists.*

Proof: As $\mathbb{R}$ is complete, it suffices to show that $\{d(x_n, y_n)\}$ is a Cauchy sequence. We have

$$|d(x_n, y_n) - d(x_m, y_m)| = |d(x_n, y_n) - d(x_m, y_n) + d(x_m, y_n) - d(x_m, y_m)| \leq$$

$$\leq |d(x_n, y_n) - d(x_m, y_n)| + |d(x_m, y_n) - d(x_m, y_m)| \leq d(x_n, x_m) + d(y_n, y_m)$$

where we have used the inverse triangle inequality (Proposition 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. □

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

$$\{x_n\} \sim \{y_n\} \iff \lim_{n\to\infty} d(x_n, y_n) = 0$$

Lemma 3.7.6 *$\sim$ is an equivalence relation.*

Proof: We have to check the three properties in Definition 1.5.2:

Reflexivity: Since $\lim_{n\to\infty} d(x_n, x_n) = 0$, the relation is reflexive.

Symmetry: Since $\lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(y_n, x_n)$, the relation is symmetric.

Transitivity: Assume that $\{x_n\} \sim \{y_n\}$ og $\{y_n\} \sim \{z_n\}$. Then $\lim_{n\to\infty} d(x_n, y_n)$ $= \lim_{n\to\infty} d(y_n, z_n) = 0$, and consequently

$$0 \leq \lim_{n\to\infty} d(x_n, z_n) \leq \lim_{n\to\infty} \big(d(x_n, y_n) + d(y_n, z_n)\big) =$$

$$= \lim_{n\to\infty} d(x_n, y_n) + \lim_{n\to\infty} d(y_n, z_n) = 0$$

which shows that $\{x_n\} = \{y_n\}$. □

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* $\lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(\hat{x}_n, \hat{y}_n)$.

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 $\lim_{n\to\infty} d(x_n, \hat{x}_n) = \lim_{n\to\infty} d(\hat{y}_n, y_n) = 0$, we get

$$\lim_{n\to\infty} d(x_n, y_n) \leq \lim_{n\to\infty} d(\hat{x}_n, \hat{y}_n)$$

By reversing the roles of elements with and without hats, we get the opposite inequality. □

We may now define a function $\bar{d} : \bar{X} \times \bar{X} \to [0, \infty)$ by

$$\bar{d}([x_n], [y_n]) = \lim_{n\to\infty} d(x_n, y_n)$$

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.

Positivity: Clearly $\bar{d}([x_n], [y_n]) = \lim_{n\to\infty} d(x_n, y_n) \geq 0$, and by definition of the equivalence relation, we have equality if and only if $[x_n] = [y_n]$.

Symmetry: Since the underlying metric d is symmetric, we have

$$\bar{d}([x_n], [y_n]) = \lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(y_n, x_n) = \bar{d}([y_n], [x_n])$$

Triangle inequality: For all equivalence classes $[x_n], [y_n], [z_n]$, we have

$$\bar{d}([x_n], [z_n]) = \lim_{n\to\infty} d(x_n, z_n) \leq \lim_{n\to\infty} d(x_n, y_n) + \lim_{n\to\infty} d(y_n, z_n) =$$

$$= \bar{d}([x_n], [y_n]) + \bar{d}([y_n], [z_n])$$

□

For each $x \in X$, let $\bar{x}$ be the equivalence class of the constant sequence $\{x, x, x, \ldots\}$. Since $\bar{d}(\bar{x}, \bar{y}) = \lim_{n\to\infty} d(x, y) = d(x, y)$, the mapping $x \to \bar{x}$ is an embedding[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 $\bar{X}$.

[3]Recall Definition 3.1.3

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 $\bar{d}([x_n], \bar{x}) = \lim_{n\to\infty} d(x_n, x_N) \leq \frac{\epsilon}{2} < \epsilon$. □

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 *Any Cauchy sequences in D converges to an element in $\bar{X}$.*

Proof: Let $\{\bar{u}_k\}$ be a Cauchy sequence in D. Since $d(u_n, u_m) = \bar{d}(\bar{u}_n, \bar{u}_m)$, $\{u_n\}$ is a Cauchy sequence in X, and gives rise to an element $[u_n]$ in $\bar{X}$. To see that $\{\bar{u}_k\}$ converges to $[u_n]$, note that $\bar{d}(\bar{u}_k, [u_n]) = \lim_{n\to\infty} d(u_k, u_n)$. Since $\{u_n\}$ is a Cauchy sequence, this limit decreases to 0 as k goes to infinity. □

The lemma above isn't enough to prove that $\bar{X}$ is complete as it may have "new" Cauchy sequences that doesn't come from Cauchy sequences in X. However, since D is dense, this is not a big problem:

Lemma 3.7.11 *$(\bar{X}, \bar{d})$ is complete.*

Proof: Let $\{x_n\}$ be a Cauchy sequence in $\bar{X}$. Since D is dense in $\bar{X}$, there is for each n a $y_n \in D$ such that $\bar{d}(x_n, y_n) < \frac{1}{n}$. It is easy to check that since $\{x_n\}$ is a Cauchy sequence, so is $\{y_n\}$. By the previous lemma, $\{y_n\}$ converges to an element in $\bar{X}$, and by construction $\{x_n\}$ must converge to the same element. Hence $(\bar{X}, \bar{d})$ 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, $x \to \bar{x}$ is an isometry from X to D). All we have to do, is to replace the elements $\bar{x}$ in D by the original elements x in X, and we have found a completion of X. □

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 $i : X \to Y$ 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/compact if and only if $i(A)$ is open/closed/compact.

Chapter 4

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 $f : X \to Y$ 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 $f : X \to Y$ 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 $d_Y(f(x), f(y)) < \epsilon$ for all points $x, y \in X$ such that $d_X(x, y) < \delta$.*

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 $f : \mathbb{R} \to \mathbb{R}$ 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 $f : X \to Y$ 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 $n \in \mathbb{N}$. □

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 $f : X \to Y$. 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 the 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 $f : X \to X$. Then $\mathcal{F}$ is equicontinuous, since we can 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 $f \in \mathcal{F}$. ♣

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 $f : (0,1) \to \mathbb{R}$ given by $f(x) = \frac{1}{x}$ is not uniformly continuous.
3. A function $f : X \to Y$ 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 $x, y \in X$. Asume that $\mathcal{F}$ is a collection of functions $f : X \to Y$ with Lipschitz constant K. Show that $\mathcal{F}$ is equicontinuous.
4. Let $f : \mathbb{R} \to \mathbb{R}$ 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 continuous functions can converge to a limit function f: *pointwise convergence* and *uniform convergence*. The distinction is rather simililar 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 $f_n : X \to Y$. We say that $\{f_n\}$* converges pointwise *to a function $f : X \to Y$ if $f_n(x) \to f(x)$ 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 $f_n : X \to Y$. We say that $\{f_n\}$* converges uniformly *to a function $f : X \to Y$ 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 $f_n : X \to Y$. For any function $f : X \to Y$ the following are equivalent:*

(i) $\{f_n\}$ *converges uniformly to* f.

(ii) $\sup\{d_Y(f_n(x), f(x)) \,|\, x \in X\} \to 0$ *as* $n \to \infty$.

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 $\sup\{d_Y(f_n(x), f(x)) \,|\, x \in X\} \leq \epsilon$ 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 $\sup\{d_Y(f_n(x), f(x)) \,|\, x \in X\} \to 0$.

$(ii) \implies (i)$ Assume that $\sup\{d_Y(f_n(x), f(x)) \,|\, x \in X\} \to 0$ as $n \to \infty$. Given an $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $\sup\{d_Y(f_n(x), f(x)) \,|\, x \in X\} < \epsilon$ 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. □

Here is an example which shows clearly the distinction between pointwise and uniform convergence:

Example 1 Let $f_n : [0, 1] \to \mathbb{R}$ be the function in Figure 1. It is constant zero except on the interval $[0, \frac{1}{n}]$ where it looks like a tent of height 1.

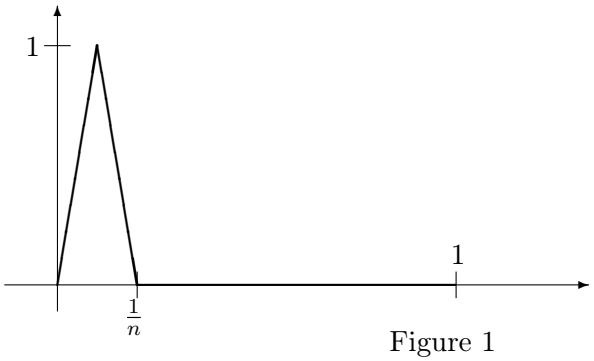


Figure 1

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, $\sup\{d_Y(f_n(x), f(x)) \mid x \in [0,1]\} = 1$ 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 $f_n : \mathbb{R} \to \mathbb{R}$ be the function in Figure 2.

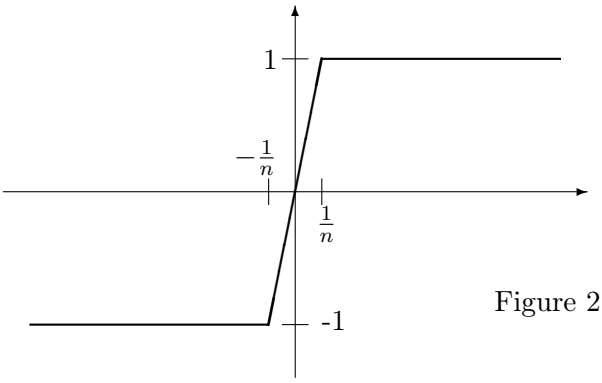


Figure 2

It is defined by

$$f_n(x) = \begin{cases} -1 & \text{if } x \leq -\frac{1}{n} \\ nx & \text{if } -\frac{1}{n} < x < \frac{1}{n} \\ 1 & \text{if } \frac{1}{n} \leq x \end{cases}$$

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 $f_n : X \to Y$ 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

$$d_Y(f(x), f(a)) \leq d_Y(f(x), f_N(x)) + d_Y(f_N(x), f_N(a)) + d_Y(f_N(a), f(a)) <$$
$$< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$$

and hence f is continuous at a. □

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.

Exercises for Section 4.2

1. Let $f_n : \mathbb{R} \to \mathbb{R}$ be defined by $f_n(x) = \frac{x}{n}$. Show that $\{f_n\}$ converges pointwise, but not uniformly to 0.
2. Let $f_n : (0, 1) \to \mathbb{R}$ be defined by $f_n(x) = x^n$. Show that $\{f_n\}$ converges pointwise, but not uniformly to 0.
3. The function $f_n : [0, \infty) \to \mathbb{R}$ 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 $f_n : (0, \infty) \to \mathbb{R}$ 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 $f_n : \mathbb{R} \to \mathbb{R}$ and assume that the sequence $\{f_n\}$ of continuous functions converges uniformly to $f : \mathbb{R} \to \mathbb{R}$ 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 $f_n : [a, b] \to \mathbb{R}$ are continuous functions converging uniformly to f. Show that
   $$\int_a^b f_n(x)\, dx \to \int_a^b f(x)\, dx$$
   Find an example which shows that this is not necessarily the case if $\{f_n\}$ only converges pointwise to f.

8. Let $f_n : \mathbb{R} \to \mathbb{R}$ 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 derivates 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 $f_n(x_n) \to f(x)$. 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 $f_n : X \to Y$ converges uniformly to f, and that $g : Y \to Z$ 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 $f_n : X \to \mathbb{R}$ 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 $n \in N$, then the partial sums $s_N(x) = \sum_{n=0}^{N} f_n(x)$ converge uniformly to a continuous function $s : X \to \mathbb{R}$ as $N \to \infty$. (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 $f_n : X \to \mathbb{R}$ 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 $x \in \mathrm{B}(a; r)$.

    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 has 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.

Example 1: Let $f_n : [0,1] \to \mathbb{R}$ be the function in the figure.

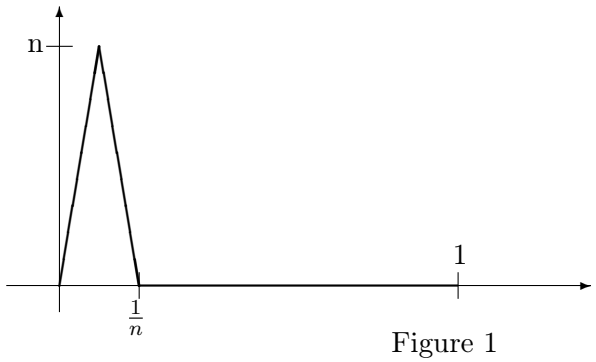


Figure 1

It is given by the formula

$$f_n(x) = \begin{cases} 2n^2x & \text{if } 0 \le x < \frac{1}{2n} \\ -2n^2x + 2n & \text{if } \frac{1}{2n} \le x < \frac{1}{n} \\ 0 & \text{if } \frac{1}{n} \le x \le 1 \end{cases}$$

but it is much easier just to work from the picture. The sequence $\{f_n\}$ converges pointwise to 0, but the integrals do not 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 do not converge to the integral of f. If the convergence is *uniform*, this can not 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 $x \in [a, b]$:

$$|F(x) - F_n(x)| = |\int_a^x (f(t) - f_n(t))\,dt\,| \leq \int_a^x |f(t) - f_n(t)|\,dt \leq$$

$$\leq \int_a^x \frac{\epsilon}{b-a}\,dt \leq \int_a^b \frac{\epsilon}{b-a}\,dt = \epsilon$$

This shows that $\{F_n\}$ converges uniformly to F on $[a, b]$. □

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 $\int_a^x f(t)\,dt$ according to 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. □

Let us reformulate this result in terms of series. Recall that a series of functions $\sum_{n=0}^{\infty} v_n(x)$ converges pointwise/unifomly to a function f on an interval I if an only if the sequence $\{s_N\}$ of partial sum $s_N(x) = \sum_{n=0}^{N} v_n(x)$ converges pointwise/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*

$$\sum_{n=0}^{\infty} \int_{x_0}^{x} v_n(t)\,dt = \int_{x_0}^{x} \sum_{n=0}^{\infty} v_n(t)\,dt$$

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,

$$\int_{x_0}^{x} f(t)\,dt = \lim_{N\to\infty} \int_{x_0}^{x} s_N(t)\,dt = \lim_{N\to\infty} \int_{x_0}^{x} \sum_{n=0}^{N} v_n(t)\,dt$$

Since

$$\lim_{N\to\infty} \int_{x_0}^{x} \sum_{n=0}^{N} v_n(t)\,dt = \lim_{N\to\infty} \sum_{n=0}^{N} \int_{x_0}^{x} v_n(t)\,dt = \sum_{n=0}^{\infty} \int_{x_0}^{x} v_n(t)\,dt\,,$$

the corollary follows. □

The corollary tell 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 later, 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 $v_n : A \to \mathbb{R}$ defined on a set A, and assume that there is a convergent series $\sum_{n=0}^{\infty} M_n$ of positive numbers such that $|v_n(x)| \le M_n$ 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 $m > n$,

$$|s_m(x) - s_n(x)| = |\sum_{k=n+1}^{m} v_k(x)\,| \le \sum_{k=n+1}^{m} |v_k(x)| \le \sum_{k=n+1}^{m} M_k = |S_m - S_n|\,,$$

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 $s : A \to \mathbb{R}$.

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 $n \geq N$,

$$|s(x) - s_n(x)| = |\sum_{k=n+1}^{\infty} v_k(x)| \leq \sum_{k=n+1}^{\infty} |v_k(x)| \leq \sum_{k=n+1}^{\infty} M_k < \epsilon$$

for all $x \in A$, and hence $\{s_n\}$ converges uniformly to s on A. □

Example 1: 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 intergrate termwise to get

$$\int_0^x f(t)\, dt = \sum_{n=1}^{\infty} \int_x \frac{\cos nt}{n^2}\, dt = \sum_{n=1}^{\infty} \frac{\sin nx}{n^3}$$

♣

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 2: 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. ♣

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 all wriggle equally much, and the derivatives do not converge.

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 $\{f(x_0)\}$ 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 se 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

$$f_n(x) - f_n(x_0) \to \int_{x_0}^{x} g(t)\,dt \qquad \text{uniformly on } [a,b]$$

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 $f'(x) = g(x)$. □

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*

$$\left(\sum_{n=0}^{\infty} u_n(x)\right)' = \sum_{n=0}^{\infty} u_n'(x)$$

The corollary tells us that under rather strong conditions, we can differentiate the series $\sum_{n=0}^{\infty} u_n(x)$ term by term.

Example 3: Summing a geometric series, we see that

$$\frac{1}{1-e^{-x}} = \sum_{n=0}^{\infty} e^{-nx} \qquad \text{for } x > 0 \tag{4.3.1}$$

If we can differentiate term by term on the right hand side, we shall get

$$\frac{e^{-x}}{(1-e^{-x})^2} = \sum_{n=1}^{\infty} n e^{-nx} \qquad \text{for } x > 0 \tag{4.3.2}$$

To check that this is correct, we must check the convergence of the differentiated series (4.2.2). Choose an interval $[a,b]$ where $a > 0$, then

$ne^{-nx} \leq ne^{-na}$ for all $x \in [a,b]$. Using, e.g., the ratio test, it is easy to see that the series $\sum_{n=0}^{\infty} ne^{-na}$ converges, and hence $\sum_{n=0}^{\infty} ne^{-nx}$ converges uniformly on $[a,b]$ by Weierstrass' M-test. The corollary now tells us that the sum of the sequence (4.2.2) is the derivative of the sum of the sequence (4.2.1), i.e.

$$\frac{e^{-x}}{(1-e^{-x})^2} = \sum_{n=1}^{\infty} ne^{-nx} \qquad \text{for } x \in [a,b]$$

Since $[a,b]$ is an arbitrary subinterval of $(0,\infty)$, we have

$$\frac{e^{-x}}{(1-e^{-x})^2} = \sum_{n=1}^{\infty} ne^{-nx} \qquad \text{for all } x > 0$$

♣

Exercises for Section 4.3

1. Show that $\sum_{n=0}^{\infty} \frac{\cos(nx)}{n^2+1}$ converges uniformly on $\mathbb{R}$.
2. Does the series $\sum_{n=0}^{\infty} ne^{-nx}$ in Example 3 converge uniformly on $(0,\infty)$?
3. Let $f_n : [0,1] \to \mathbb{R}$ be defined by $f_n(x) = nx(1-x^2)^n$. Show that $f_n(x) \to 0$ for all $x \in [0,1]$, but that $\int_0^1 f_n(x)\,dx \to \frac{1}{2}$.
4. Explain in detail how Corollary 4.3.6 follows from Proposition 4.3.5.
5. a) Show that series $\sum_{n=1}^{\infty} \frac{\cos\frac{x}{n}}{n^2}$ converges uniformly on $\mathbb{R}$.
   b) Show that $\sum_{n=1}^{\infty} \frac{\sin\frac{x}{n}}{n}$ converges to a continuous function f, and that
   $$f'(x) = \sum_{n=1}^{\infty} \frac{\cos\frac{x}{n}}{n^2}$$
6. One can show that
   $$x = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx) \qquad \text{for } x \in (-\pi,\pi)$$
   If we differentiate term by term, we get
   $$1 = \sum_{n=1}^{\infty} 2(-1)^{n+1} \cos(nx) \qquad \text{for } x \in (-\pi,\pi)$$
   Is this a correct formula?
7. 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 $f'(x) = -\sum_{n=1}^{\infty} \frac{\ln x}{n^x}$.

4.4 Applications to power series

In this section, we shall illustrate the theory in previous section by applying it to the power series you know from calculus. If you are not familiar with lim sup and lim inf, you should read the discussion in Section 2.2 before you continue.

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

$$\frac{1}{R} = \limsup_{n\to\infty} \sqrt[n]{|c_n|}$$

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 $\sum_{n=0}^{\infty} c_n(x-a)^n$, 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 $\limsup_{n\to\infty} \sqrt[n]{|c_n|} = \frac{1}{R}$, 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 $\limsup_{n\to\infty} \sqrt[n]{|c_n|} = \frac{1}{R}$, 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 $|x-a| < R$. □

Remark: When we want to find the radius of convergence, it is occasionally convenient to compute a slightly different limit such as $\lim_{n\to\infty} \sqrt[n+1]{c_n}$ or $\lim_{n\to\infty} \sqrt[n-1]{c_n}$ instead of $\lim_{n\to\infty} \sqrt[n]{c_n}$. 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 endpoint. Although the convergence is uniform on all subintervals $[a-r, a+r]$, it is not in general uniform on $(a-R, a+R)$.

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*

$$f'(x) = \sum_{n=1}^{\infty} nc_n(x-a)^{n-1} = \sum_{n=0}^{\infty}(n+1)c_{n+1}(x-a)^n \qquad \text{for } x \in (a-R, a+R)$$

and

$$\int_a^x f(t)\,dt = \sum_{n=0}^{\infty} \frac{c_n}{n+1}(x-a)^{n+1} = \sum_{n=1}^{\infty} \frac{c_{n-1}}{n}(x-a)^n \qquad \text{for } x \in (a-R, a+R)$$

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

$$\limsup_{n\to\infty} \sqrt[n+1]{|(n+1)c_{n+1}|} = \limsup_{n\to\infty} \sqrt[n]{|c_n|} = \frac{1}{R}$$

(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 $\lim_{n\to\infty} \sqrt[n+1]{n+1} = 1$, this is not hard to show (see Exercise 6). Applying Corollary 4.2.6 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 $x \in (a-R, a+R)$. □

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 *Let $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ 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

$$f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)\cdot \ldots \cdot (n-k+1)c_n(x-a)^{n-k}$$

Putting $x = a$, we get $f^{(k)}(a) = k!c_k$, and the corollary follows. □

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 $a \pm R$? It turns out that the sum is also continuous at the endpoint, but that this is surprisingly intricate to prove.

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 $b_n \to 0$ as $n \to \infty$, 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 $N \to \infty$. □

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]

Proof: We already know that f is continuous in the open interval $(a-R, a+R)$, and that 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 assume that $|x| < R$, we may apply the second version of Abel's summation formula with $a_n = c_n R^n$ and $b_n = \left(\frac{x}{n}\right)^n$ to get

$$f(x) = \sum_{n=0}^{\infty} f_n(R) \left(\left(\frac{x}{R}\right)^n - \left(\frac{x}{R}\right)^{n+1} \right) = \left(1 - \frac{x}{R}\right) \sum_{n=0}^{\infty} f_n(R) \left(\frac{x}{R}\right)^n$$

where $f_n(R) = \sum_{k=0}^{n} c_k R^k$. Summing a geometric series, we see that we also have

$$f(R) = \left(1 - \frac{x}{R}\right) \sum_{n=0}^{\infty} f(R) \left(\frac{x}{R}\right)^n$$

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 series may go to infinity when $x \to R$. Hence we need to control the tail of the sequence before we exploit the factor

[1] I use $\lim_{x \uparrow b}$ and $\lim_{x \downarrow b}$ for one-sided limits, also denoted by $\lim_{x \to b^-}$ and $\lim_{x \to b^+}$.

$(1 - x/R)$. Fortunately, this is not difficult: Since $f_n(R) \to f(R)$, we first pick an $N \in \mathbb{N}$ such that $|f_n(R) - f(R)| < \frac{\epsilon}{2}$ for $n \geq N$. Then

$$|f(x) - f(R)| \leq \left(1 - \frac{x}{R}\right) \sum_{n=0}^{N-1} |f_n(R) - f(R)| \left(\frac{x}{R}\right)^n +$$

$$+ \left(1 - \frac{x}{R}\right) \sum_{n=N}^{\infty} |f_n(R) - f(R)| \left(\frac{x}{R}\right)^n \leq$$

$$\leq \left(1 - \frac{x}{R}\right) \sum_{n=0}^{N-1} |f_n(R) - f(R)| \left(\frac{x}{R}\right)^n + \left(1 - \frac{x}{R}\right) \sum_{n=0}^{\infty} \frac{\epsilon}{2} \left(\frac{x}{R}\right)^n =$$

$$= \left(1 - \frac{x}{R}\right) \sum_{n=0}^{N-1} |f_n(R) - f(R)| \left(\frac{x}{R}\right)^n + \frac{\epsilon}{2}$$

where we have summed a geometric series. Now the sum is finite, and 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. □

Let us take a look at a famous example.

Example 1: Summing a geometric series, we clearly have

$$\frac{1}{1 + x^2} = \sum_{n=0}^{\infty} (-1)^n x^{2n} \qquad \text{for } |x| < 1$$

Integrating, we get

$$\arctan x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n + 1} \qquad \text{for } |x| < 1$$

Using the Alternating Series Test, we see that the series converges even for $x = 1$. By Abel's Theorem

$$\frac{\pi}{4} = \arctan 1 = \lim_{x \uparrow 1} \arctan x = \lim_{x \uparrow 1} \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n + 1} = \sum_{n=0}^{\infty} (-1)^n \frac{1}{2n + 1}$$

Hence we have proved

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$$

This is often called Leibniz' or Gregory's formula for π, but it was actually first discovered by the Indian mathematician Madhava (ca. 1350 – 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

$$\frac{1}{1+x} = \sum_{n=0}^{\infty} (-x)^n \qquad \text{for } |x| < 1$$

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

1. Find power series with radius of convergence 0, 1, 2, and ∞.
2. Find power series with radius of convergence 1 that converge at both, one and neither of the endpoints.
3. Show that for any polynomial P, $\lim_{n\to\infty} \sqrt[n]{|P(n)|} = 1$.
4. Use the result in Exercise 3 to find the radius of convergence:
   a) $\sum_{n=0}^{\infty} \frac{2^n x^n}{n^3+1}$
   b) $\sum_{n=0}^{\infty} \frac{2n^2+n-1}{3n+4} x^n$
   c) $\sum_{n=0}^{\infty} n x^{2n}$
5. a) Explain that $\frac{1}{1-x^2} = \sum_{n=0}^{\infty} x^{2n}$ for $|x| < 1$,
   b) Show that $\frac{2x}{(1-x^2)^2} = \sum_{n=0}^{\infty} 2n x^{2n-1}$ 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. Let $\sum_{n=0}^{\infty} c_n (x-a)^n$ be a power series.
   a) Show that the radius of convergence is given by

$$\frac{1}{R} = \limsup_{n\to\infty} \sqrt[n+k]{|c_n|}$$

   for any integer k.

b) Show that $\lim_{n\to\infty} \sqrt[n+1]{n+1} = 1$ (write $\sqrt[n+1]{n+1} = (n+1)^{\frac{1}{n+1}}$).

c) Prove the formula

$$\limsup_{n\to\infty} \sqrt[n+1]{|(n+1)c_{n+1}|} = \limsup_{n\to\infty} \sqrt[n]{|c_n|} = \frac{1}{R}$$

in the proof of Corollary 4.4.2.

7. 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 $\ln 2 = \sum_{n=0}^{\infty}(-1)^n \frac{1}{n+1}$.

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 $\lim_{n\to\infty} nc_n = 0$, 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 $\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} n|c_n| = 0$.

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 $1 - x^n \le n(1-x)$ for $|x| < 1$.

e) Let N_x be the integer such that $N_x \le \frac{1}{1-x} < N_x + 1$ Show that

$$\sum_{n=0}^{N_x} c_n(1-x^n) \le (1-x)\sum_{n=0}^{N_x} n|c_n| \le \frac{1}{N_x}\sum_{n=0}^{N_x} n|c_n| \to 0$$

as $x \uparrow 1$.

f) Show that

$$\left|\sum_{n=N_x+1}^{\infty} c_n x^n\right| \le \sum_{n=N_x+1}^{\infty} n|c_n|\frac{x^n}{n} \le \frac{d_x}{N_x}\sum_{n=0}^{\infty} x^n$$

where $d_x \to 0$ as $x \uparrow 1$. Show that $\sum_{n=N_x+1}^{\infty} c_n x^n \to 0$ as $x \uparrow 1$.

g) Prove Tauber's theorem.

4.5 The spaces $B(X,Y)$ 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 this 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 to prove the existence of solutions to differential equations. In this section, we shall consider spaces of bounded functions while 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 space, a function $f : X \to Y$ 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 $f, g : X \to Y$ are two bounded functions, then there is a number K such that $d_Y(f(x), g(x)) \leq K$ for all $u, v \in X$. 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

$$d_Y(f(x), g(x)) \leq d_Y(f(x), f(a)) + d_Y(f(a), g(a)) + d_Y(g(a), g(x))$$

$$\leq M_a + d_Y(f(a), g(a)) + N_a$$

we can take $K = M_a + d_Y(f(a), g(a)) + N_a$.

We now let

$$B(X,Y) = \{f : X \to Y \mid f \text{ is bounded}\}$$

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 the 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 $C(X,Y)$; 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. □

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

$$\sup\{d_Y(f_n(x),f(x)) \,|\, x \in X\} \to 0$$

This just means that $\rho(f_n,f) \to 0$, which is to say that $\{f_n\}$ converges to f in $(B(X,Y),\rho)$. □

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 $f \in B(X,Y)$.

Fix 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 $f : X \to Y$. 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 all $x \in X$ and

all $n, m \geq N$, $d_Y(f_n(x), f_m(x)) < \frac{\epsilon}{2}$. If we let $m \to \infty$, we see that for all $x \in X$ and all $n \geq N$

$$d_Y(f_n(x), f(x)) = \lim_{m\to\infty} d_Y(f_n(x), f_m(x)) \leq \frac{\epsilon}{2}$$

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 $B(X,Y)$. □

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 $f, g : [0,1] \to \mathbb{R}$ be given by $f(x) = x$, $g(x) = x^2$. Find $\rho(f,g)$.
2. Let $f, g : [0, 2\pi] \to \mathbb{R}$ 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 $\lim_{r\to 0} \rho(f_r, f) = 0$.
   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 $\lim_{r\to 0} \rho(f_r, f) = 0$ for all $f \in B(\mathbb{R}, \mathbb{R})$?

4.6 The spaces $C_b(X,Y)$ and $C(X,Y)$ of 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 hence we now turn to spaces of continuous functions

As before, we assume that (X, d_X) and (Y, d_Y) are metric spaces. We define

$$C_b(X, Y) = \{f : X \to Y \mid f \text{ is continuous and bounded}\}$$

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 3.3.6, 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 $C_b(X, Y)$. □

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 it 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. □

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 $f : X \to Y$ 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 (Theorem 3.5.10) if we can show that it is continiuous. 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). □

If we define

$$C(X,Y) = \{f : X \to Y \mid f \text{ is continuous}\},$$

the proposition above tell 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 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 $f : X \to \mathbb{R}$.

3. Assume that $f : \mathbb{R} \to \mathbb{R}$ is a bounded continuous function. If $u \in C([0,1], \mathbb{R})$, we define $L(u) : [0,1] \to \mathbb{R}$ 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,

$$\rho(f,g) = \sup\{d_Y(f(x),g(x)) \,|\, x \in X\}$$

may be ∞, and a metric can not 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 $f : X \to Y$ is continuous with respect to $\bar{d}$ if and only if it is continuous with respect to d. Show the same for functions $g : Y \to X$.

g) For functions $f, g \in C(X,Y)$, define

$$\bar{\rho}(f,g) = \sup\{\bar{d}(f(x),g(x) \,|\, x \in X\}$$

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

Consider a system of differential equations

$$\begin{aligned} y_1'(t) &= f_1(t, y_1(t), y_2(t), \ldots, y_n(t)) \\ y_2'(t) &= f_2(t, y_1(t), y_2(t), \ldots, y_n(t)) \\ \vdots \quad &\vdots \qquad\qquad \vdots \qquad\quad \vdots \\ y_n'(t) &= f_n(t, y_1(t), y_2(t), \ldots, y_n(t)) \end{aligned}$$

with initial conditions $y_1(0) = Y_1$, $y_2(0) = Y_2$, $\ldots$, $y_n(0) = Y_n$. In this section we shall use Banach's Fixed Point Theorem 3.4.5 and the completeness of $C([0,a], \mathbb{R}^n)$ to prove that under reasonable conditions such systems have a unique solution.

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

$$\mathbf{f}(t, \mathbf{y}(t)) = \begin{pmatrix} f_1(t, y_1(t), y_2(t), \ldots, y_n(t)) \\ f_2(t, y_1(t), y_2(t), \ldots, y_n(t)) \\ \vdots \\ f_n(t, y_1(t), y_2(t), \ldots, y_n(t)) \end{pmatrix}$$

In this notation, the system becomes

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \qquad \mathbf{y}(0) = \mathbf{y}_0 \tag{4.7.1}$$

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.

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}(s))\, ds \tag{4.7.2}$$

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

$$\mathbf{u}(t) = y_0 + \int_0^t \mathbf{f}(s, \mathbf{z}(s))\, ds$$

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 to prove that F has a unique fixed point, but first we have to introduce a crucial condition. We say that the function $\mathbf{f} : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n$ 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 $\mathbf{f} : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and uniformly Lipschitz with Lipschitz constant K on $[0, \infty)$. If $a < \frac{1}{K}$, the map*

$$F : C([0, a], \mathbb{R}^n) \to C([0, a], \mathbb{R}^n)$$

defined by

$$F(\mathbf{z})(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(t, \mathbf{z}(t))\, dt$$

is a contraction.

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: Let $\mathbf{v}, \mathbf{w}$ be two elements in $C([0, a], \mathbb{R}^n)$, and note that for any $t \in [0, a]$

$$\|F(\mathbf{v})(t) - F(\mathbf{w})(t)\| = \| \int_0^t \big(\mathbf{f}(s, \mathbf{v}(s)) - \mathbf{f}(s, \mathbf{w}(s))\big)\, ds\| \leq$$

$$\leq \int_0^t \|\mathbf{f}(s,\mathbf{v}(s)) - \mathbf{f}(s,\mathbf{w}(s))\|\, ds \leq \int_0^t K\|\mathbf{v}(s) - \mathbf{w}(s)\|\, ds \leq$$

$$\leq K\int_0^t \rho(\mathbf{v},\mathbf{w})\, ds \leq K\int_0^a \rho(\mathbf{v},\mathbf{w})\, ds = Ka\,\rho(\mathbf{v},\mathbf{w})$$

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 $\mathbf{f} : [0,\infty) \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and uniformly Lipschitz on $[0,\infty)$. Then the initial value problem*

$$\mathbf{y}'(t) = \mathbf{f}(t,\mathbf{y}(t)), \qquad \mathbf{y}(0) = \mathbf{y}_0 \tag{4.7.3}$$

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

$$F : C([0,a],\mathbb{R}^n) \to C([0,a],\mathbb{R}^n)$$

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 tells us that F has a unique fixed point $\mathbf{y}$. This means that the integral equation

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s,\mathbf{y}(s))\, ds \tag{4.7.4}$$

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

$$\mathbf{y}(t) = \mathbf{y}(a) + \int_a^t \mathbf{f}(s,\mathbf{y}(s))\, ds =$$

$$= \mathbf{y}_0 + \int_0^a \mathbf{f}(s,\mathbf{y}(s))\, ds + \int_a^t \mathbf{f}(s,\mathbf{y}(s))\, ds = \mathbf{y}_0 + \int_0^t \mathbf{f}(s,\mathbf{y}(s))\, ds$$

Continuing this procedure to new intervals $[2a,3a]$, $[3a,4a]$, we see that the integral equation (4.7.3) 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. □

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.

Exercises to Section 4.7

1. Solve the initial value problem

$$y' = 1 + y^2, \qquad y(0) = 0$$

and show that the solution is only defined on the interval $[0, \pi/2)$.

2. Show that the functions

$$y(t) = \begin{cases} 0 & \text{if } 0 \le t \le a \\ (t-a)^{\frac{3}{2}} & \text{if } t > a \end{cases}$$

where $a \ge 0$ are all solutions of the initial value problem

$$y' = \frac{3}{2} y^{\frac{1}{3}}, \qquad y(0) = 0$$

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

$$u''(t) = g(t, u(t), u'(t)) \qquad u(0) = a, u'(0) = b \qquad (*)$$

where $g : [0, \infty) \times \mathbb{R}^2 \to \mathbb{R}$ is a given function. Define a function $\mathbf{f} : [0, \infty) \times \mathbb{R}^2 \to \mathbb{R}^2$ 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

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \qquad \mathbf{y}(0) = \begin{pmatrix} a \\ b \end{pmatrix},$$

then u is a solution of the original problem $(*)$.

4.8 Compact subsets of $C(X, \mathbb{R}^m)$

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 $d(x, y) \leq \delta$.

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

$$a_1 = 1.4 = \frac{14}{10}, \;\; a_2 = 1.41 = \frac{141}{100}, \;\; a_3 = 1.414 = \frac{1414}{1000}, \;\; a_4 = 1.4142 = \frac{14142}{10000}, \ldots$$

Recall 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 (X, d) spaces 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 Theorem 3.5.13). This mean 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 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 belong to, etc. Since $d(x, a_n) < \frac{1}{n}$, $\{a_n\}$ 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 $\frac{1}{m}$. □

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 $f : X \to Y$. 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 dense sets of the kind described above.

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 $n, m \geq N$,

$$d_Y(g_n(x), g_m(x)) \leq$$
$$\leq d_Y(g_n(x), g_n(a)) + d_Y(g_n(a), g_m(a)) + d_Y(g_m(a), g_m(x)) <$$
$$< \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{3\epsilon}{4}$$

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 $C(X, Y)$. □

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

$$A = \{a_1, a_2 \ldots, a_n, \ldots\}$$

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 $a = a_1, a_2, \ldots, a_j$. 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 $a = a_1, a_2, \ldots, a_j$. 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. □

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. □

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 x, 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. □

We are finally ready to prove the main theorem:

Theorem 4.8.9 (Arzelà-Ascoli's 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, if 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 can not 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

$$f_{n_{k_j}}(x_{n_{k_j}}) \to f(a) \quad \text{and} \quad f_{n_{k_j}}(y_{n_{k_j}}) \to f(a)$$

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. □

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 $\mathrm{B}(a, r)$, $a \in X$, $r > 0$, 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 $f : A \to \mathbb{R}$ 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 $\bar{f} : X \to \mathbb{R}$ by putting $\bar{f}(x) = \lim_{n\to\infty} f(a_n)$ where $\{a_n\}$ is a sequence from a converging to x. We call f the *continuous extension of f to X*. Show that $\bar{f}$ is uniformly continuous.

   c) Let $f : \mathbb{Q} \to \mathbb{R}$ 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 uniformly convergent subsequence.

5. A function $f : [-1, 1] \to \mathbb{R}$ is called *Lipschitz continuous with Lipschitz constant* $K \in \mathbb{R}$ if

   $$|f(x) - f(y)| \leq K|x \breve{} y|$$

   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 $\sigma : [0, \infty) \to [0, \infty)$ be a nondecreasing, continuous function such that $\sigma(0) = 0$. We say that σ is a *modulus of continuity* for a function $f : X \to Y$ 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

$$\mathcal{K} = \{f : X \to \mathbb{R}^n \ : \ f(x_0) = \mathbf{0} \text{ and } \sigma \text{ is modulus of continuity for } f\}$$

   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 $\overline{\mathrm{B}}(a; r)$ centered at a that is compact. (Recall that $\overline{\mathrm{B}}(a; r) = \{x \in X \ : \ d(a, x) \leq r\}$). Show that $\mathbb{R}^m$ is locally compact, but that $C([0, 1], \mathbb{R})$ is not.

4.9 Differential equations revisited

In Section 4.7, we used Banach's Fixed Point Theorem to study initial value problems of the form

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0 \tag{4.9.1}$$

or equivalently

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}(s))\, ds \tag{4.9.2}$$

In this section we shall see how Arzelà-Ascoli's Theorem can be used to prove 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

$$T = \{t_0, t_1, t_2, t_3 \ldots\}$$

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 $\mathbf{f}(t_0, \mathbf{y}_0)$, 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 $\mathbf{f}(t_1, \hat{\mathbf{y}}(t_1))$ so that we have

$$\hat{\mathbf{y}}(t) = \hat{\mathbf{y}}(t_1) + \mathbf{f}(t_0, \hat{\mathbf{y}}(t_1))(t - t_1)$$

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

$$\mathbf{f}(t_i, \hat{\mathbf{y}}(t_i))(t_{i+1} - t_i) = \int_{t_i}^{t_{i+1}} \mathbf{f}(t_i, \hat{\mathbf{y}}(t_i)\, ds\,,$$

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

$$\hat{\mathbf{y}}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \hat{\mathbf{y}}(s))\, ds + \int_0^t \big(\mathbf{f}(\underline{s}, \hat{\mathbf{y}}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}(s))\big)\, ds$$

(observe that there is one s and one $\underline{s}$ term in the last integral) where the last term measures how much $\hat{\mathbf{y}}$ "deviates" from being a solution of equation (4.9.2).

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 expext the squence $\{\hat{\mathbf{y}}_n\}$ to converge to a solution of (2). It turns out that in the most general case we can not quite prove this, but we can instead use the Arzelà-Ascoli Theorem to find a *subsequence* converging to a solution.

Before we turn to the proof, it will useful to see how intergals 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 $\mathbf{f} : [0,\infty) \times \mathbb{R}^m \to \mathbb{R}^m$ be a continuous function, and assume that $\{\hat{\mathbf{y}}_k\}$ is a sequence of continuous functions $\hat{\mathbf{y}}_k : [0,a] \to \mathbb{R}^m$ 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

$$\|I_n(t) - I(t)\| = \| \int_0^t \big(\mathbf{f}(s, \hat{\mathbf{y}}_n(s)) - \mathbf{f}(s, \mathbf{y}(s))\big)\, ds\| \leq$$

$$\leq \int_0^t \|\mathbf{f}(s, \hat{\mathbf{y}}_n(s)) - \mathbf{f}(s, \mathbf{y}(s))\|\, ds < \int_0^a \frac{\epsilon}{a}\, ds = \epsilon$$

for all $t \in [0,a]$, and hence $\{I_k\}$ converges uniformly to I. □

We are now ready for the main result.

Theorem 4.9.2 *Assume that $\mathbf{f} : [0,\infty) \times \mathbb{R}^m \to \mathbb{R}^m$ is a continuous function and that $\mathbf{y}_0 \in \mathbb{R}^m$. Then there exists a positive real number a and a function $\mathbf{y} : [0,a] \to \mathbb{R}^m$ such that $\mathbf{y}(0) = \mathbf{y}_0$ and*

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) \quad \textit{for all } t \in [0,a]$$

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 intervall (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 $i = 1, 2, \ldots, m$. 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, put the point is that our approximate solutions of the differential equation can never leave the area

$$[Y_1 - Ma, Y_1 + Ma] \times [Y_2 - Ma, Y_2 + Ma] \times \cdots \times [Y_m - Ma, Y + ma]$$

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

$$\hat{\mathbf{y}}_{n_k}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))\, ds + \int_0^t \big(\mathbf{f}(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))\big)\, ds \quad (4.9.3)$$

and according to the lemma

$$\int_0^t \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))\, ds \to \int_0^t \mathbf{f}(s, \mathbf{y}(s))\, ds \quad \text{uniformly for } t \in |0, a]$$

If we can only prove that

$$\int_0^t \big(\mathbf{f}(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))\big)\, ds \to 0 \quad (4.9.4)$$

we will get

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}(s))\, ds$$

as $k \to \infty$ 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

$$\|(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - (s, \hat{\mathbf{y}}_{n_k}(s))\| \leq \|(\Delta t, M\Delta t, \ldots, M\Delta t)\| = \sqrt{1 + nM^2}\, \Delta t\,,$$

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

$$\| \int_0^t \big(\mathbf{f}(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))\| < \int_0^a \frac{\epsilon}{a}\, ds = \epsilon$$

and hence

$$\int_0^t \big(\mathbf{f}(\underline{s}, \hat{\mathbf{y}}_{n_k}(\underline{s})) - \mathbf{f}(s, \hat{\mathbf{y}}_{n_k}(s))\big)\, ds \to 0$$

as $k \to \infty$. As already observed, this completes the proof. □

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 one. 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 $t \to \frac{\pi}{2}^-$.

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 $\mathbf{f}_n : [a, b] \to \mathbb{R}^m$ 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 a given point, but passes through it:

   Teorem *Assume that* $\mathbf{f} : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ *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* $\mathbf{y} : [t_0 - a, t_0 + a] \to \mathbb{R}^m$ *such that* $\mathbf{y}(t_0) = \mathbf{y}_0$ *and*

   $$y'(t) = \mathbf{f}(t, \mathbf{y}(t)) \quad \text{for all } t \in [t_0 - a, t_0 + a]$$

   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 $C([a, b], \mathbb{R})$

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* $C([a, b], \mathbb{R})$ *for all* $a, b \in \mathbb{R}$*,* $a < b$*. In other words, for each continuous function* $f : [a, b] \to \mathbb{R}$*, there is a sequence of polynomials* $\{p_n\}$ *converging uniformly to* f*.*

The proof I shall give (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 straight forward scaling argument to extend it to all closed and bounded intervals.

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) 1000 first tosses, you would expect your award to be close to $f(x)$. 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 $\mathrm{E}(Y_i) = x$ and variance $\mathrm{Var}(Y_i) = x(1-x)$. We thus have

$$\mathrm{E}(X_N) = \frac{1}{N}(\mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \cdots \mathrm{E}(Y_N)) = x$$

and (using that the Y_i's are independent)

$$\mathrm{Var}(X_N) = \frac{1}{N^2}(\mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \cdots + \mathrm{Var}(Y_N)) = \frac{1}{N}x(1-x)$$

(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 $\mathrm{E}(f(X_N))$ to converge to $f(x)$.

To see what this has to do with polynomials, let us compute the average award $\mathrm{E}(f(X_N))$. Since the probability of getting exactly k heads in N tosses is $\binom{N}{k}x^k(1-x)^{n-k}$, we get

$$\mathrm{E}(f(X_N)) = \sum_{k=0}^{N} f(\frac{k}{N})\binom{N}{k}x^k(1-x)^{N-k}$$

Our expectation that $\mathrm{E}(f(X_N)) \to f(x)$ as $N \to \infty$, can therefore be rephrased as

$$\sum_{k=0}^{N} f(\frac{k}{N})\binom{N}{k}x^k(1-x)^{N-k} \to f(x) \quad N \to \infty$$

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

$$p_N(x) = \sum_{k=0}^{N} f(\frac{k}{N})\binom{N}{k}x^k(1-x)^{N-k}$$

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

$$\mathbf{1}_{\{|x-X|<\delta\}} = \begin{cases} 1 & \text{if } |x-X| < \delta \\ \\ 0 & \text{otherwise} \end{cases}$$

and oppositely for $\mathbf{1}_{\{|x-X_|\geq\delta\}}$.

Lemma 4.10.2 (Chebyshev's Inequality) *For a bounded random variable X with mean x*

$$\mathrm{E}(\mathbf{1}_{\{|x-X|\geq\delta\}}) \leq \frac{1}{\delta^2}\mathrm{Var}(X)$$

Proof: Since $\delta^2\mathbf{1}_{\{|x-X|\geq\delta\}}) \leq (x-X)^2$, we have

$$\delta^2\mathrm{E}(\mathbf{1}_{\{|x-X|\geq\delta\}}) \leq \mathrm{E}((x-X)^2) = \mathrm{Var}(X)$$

Dividing by δ^2, we get the lemma. □

We are now ready to prove that the Bernstein polynomials converge.

Proposition 4.10.3 *If $f : [0,1] \to \mathbb{R}$ is a continuous function, the* Bernstein polynomials

$$p_N(x) = \sum_{k=0}^{N} f(\frac{k}{N})\binom{N}{k}x^k(1-x)^{N-k}$$

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 has to be uniformly continuous, and hence we can find a $\delta > 0$ such that $|f(u) - f(v)| < \frac{\epsilon}{2}$ whenever $|u-v| < \delta$. Since $p_n(x) = \mathrm{E}(f(X_n))$, we have

$$|f(x) - p_n(x)| = |f(x) - \mathrm{E}(f(X_n))| = |\mathrm{E}(f(x) - f(X_n))| \leq \mathrm{E}(|f(x) - f(X_n)|)$$

We split the last expectation into two parts: the cases where $|x - X_n| < \delta$ and the rest:

$$\mathrm{E}(|f(x)-f(X_n)|) = \mathrm{E}(\mathbf{1}_{\{|x-X_n|<\delta\}}|f(x)-f(X_n)|)+\mathrm{E}(\mathbf{1}_{\{|x-X_n|\geq\delta\}}|f(x)-f(X_n)|)$$

The idea is that the first term is always small since f is continuous and that the second part will be small when N is large because X_N then is unlikely to deviate much from x. Here are the details:

By choice of δ, we have for the first term

$$\mathrm{E}(\mathbf{1}_{\{|x-X_n|<\delta\}}|f(x)-f(X_n)|) \leq \mathrm{E}\left(\mathbf{1}_{\{|x-X_n|<\delta\}}\frac{\epsilon}{2}\right) \leq \frac{\epsilon}{2}$$

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

$$\mathrm{E}(\mathbf{1}_{\{|x-X_n|\geq\delta\}}|f(x)-f(X_n)|) \leq 2M\mathrm{E}(\mathbf{1}_{\{|x-X_n|\geq\delta\}}) \leq$$

$$\leq \frac{2M}{\delta^2}\mathrm{Var}(X_n) = \frac{2Mx(1-x)}{\delta^2 n} \leq \frac{M}{2\delta^2 n}$$

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

$$\mathrm{E}(\mathbf{1}_{\{|x-X_n|\geq\delta\}}|f(x)-f(X_n)|) < \frac{\epsilon}{2}$$

for all $n \geq N$. Combining all the inequalities above, we see that if $n \geq N$, we have for all $x \in [0,1]$

$$|f(x)-p_n(x)| \leq \mathrm{E}(|f(x)-f(X_n)|) =$$

$$= \mathrm{E}(\mathbf{1}_{\{|x-X_n|<\delta\}}|f(x)-f(X_n)|) + \mathrm{E}(\mathbf{1}_{\{|x-X_n|\geq\delta\}}|f(x)-f(X_n)|) <$$

$$< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and hence the Bernstein polynomials p_n converge uniformly to f. □

To get Weierstrass' result, we just have to move functions from arbitrary intervals $[a,b]$ 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 pynomials 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

$$p_n(x) = q_n(\frac{x-a}{b-a})$$

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 if the limit function 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. There is an extension of the result called the Stone-Weierstrass Theorem which extends the result to much more general 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 poynomials 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 in the point 0.
4. Assume that $f : [a, b] \to \mathbb{R}$ is a continuous function such that $\int_a^b f(x)x^n \, dx = 0$ for all $n = 0, 1, 2, 3, \ldots$.
   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 $\binom{N}{k} = \frac{N(N-1)(N-2)\cdot\ldots\cdot(N-k+1)}{1\cdot 2\cdot 3\cdot\ldots\cdot k}$ 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 $\mathrm{E}(X_N) = \frac{1}{N}(\mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \cdots \mathrm{E}(Y_N)) = x$)

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 $\mathrm{Var}(X_N) = \frac{1}{N} x(1-x)$). *Hint:* Write $(\frac{k}{N} - x)^2 = \frac{1}{N^2}\left(k(k-1) + (1-2xN)k + N^2x^2\right)$ 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

$$|f(x) - p_n(x)| \leq \sum_{\{k:|\frac{k}{n}-x|<\delta\}} |f(x) - f(k/n)| \binom{n}{k} x^n (1-x)^{n-k} +$$

$$+ \sum_{\{k:|\frac{k}{n}-x|\geq\delta\}} |f(x) - f(k/n)| \binom{n}{k} x^n (1-x)^{n-k} \leq$$

$$< \frac{\epsilon}{2} + \sum_{\{k:|\frac{k}{n}-x|\geq\delta\}} |f(x) - f(k/n)| \binom{n}{k} x^n (1-x)^{n-k}$$

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:

$$\sum_{\{k:|\frac{k}{n}-x|\geq\delta\}} |f(x) - f(k/n)| \binom{n}{k} x^n (1-x)^{n-k} \leq$$

$$\leq 2M \sum_{\{k:|\frac{k}{n}-x|\geq\delta\}} \binom{n}{k} x^n (1-x)^{n-k} \leq$$

$$\leq 2M \sum_{k=0}^{n} \left(\frac{\frac{k}{n} - x}{\delta}\right)^2 \binom{n}{k} x^n (1-x)^{n-k} \leq \frac{2M}{n\delta^2} x(1-x) \leq \frac{M}{2n\delta^2}$$

g) Explain why we can get $|f(x) - p_n(x)| < \epsilon$ by chosing n large enough, and explain why this proves Proposition 4.10.2.

Chapter 5

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 the linear maps between them. This is in many ways an extension of theory you are all 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 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* $\|\cdot\|:V\to\mathbb{R}$ *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) $\|\mathbf{u}+\mathbf{v}\|\leq\|\mathbf{u}\|+\|\mathbf{v}\|$ *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 $\mathbf{x}=(x_1,x_2\ldots,x_n)$. 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 $\mathbf{z}=(z_1,z_2\ldots,z_n)$. ♣

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 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:
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 $\mathbf{u} = \mathbf{v}$.
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 $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$.
Triangle inequality: By part (iii) of the definition above, we see that for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$:

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\| = \|(\mathbf{u} - \mathbf{w}) + (\mathbf{w} - \mathbf{v})\| \leq$$

$$\leq \|\mathbf{u} - \mathbf{w}\| + \|\mathbf{w} - \mathbf{v}\| = d(\mathbf{u}, \mathbf{w}) + d(\mathbf{w}, \mathbf{v})$$

□

Whenever we refer to notions such as convergence, continuity, openness, closedness, completeness, compactness etc. in a normed vector space, we shall be 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

$$\|\mathbf{x} - \mathbf{x}_n\| = d(\mathbf{x}, \mathbf{x}_n) \to 0 \quad \text{as} \quad n \to \infty$$

Remark: The Inverse Triangle Inequality (recall Proposition 3.1.4)

$$|d(x, y) - d(x, z)| \leq d(y, z) \tag{5.1.1}$$

is a useful tool in metric spaces. In normed spaces, it is most conveniently expressed as

$$|\,\|\mathbf{u}\| - \|\mathbf{v}\|\,| \leq \|\mathbf{u} - \mathbf{v}\| \tag{5.1.2}$$

(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 $\mathbf{x} + \mathbf{y}$.

(ii) 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 $\lim_{n\to\infty} \|\mathbf{x} - \mathbf{x}_n\| = 0$. As $|\|\mathbf{x}_n\| - \|\mathbf{x}\|| \leq \|\mathbf{x} - \mathbf{x}_n\|$ by the inverse triangle inequality, it follows that $\lim_{n\to\infty} |\|\mathbf{x}_n\| - \|\mathbf{x}\|| = 0$, 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

$$\|\alpha\mathbf{x} - \alpha_n\mathbf{x}_n\| = \|(\alpha\mathbf{x} - \alpha\mathbf{x}_n) + (\alpha\mathbf{x}_n - \alpha_n\mathbf{x}_n)\|$$

$$\leq \|\alpha\mathbf{x} - \alpha\mathbf{x}_n\| + \|\alpha\mathbf{x}_n - \alpha_n\mathbf{x}_n\| = |\alpha|\|\mathbf{x} - \mathbf{x}_n\| + |\alpha - \alpha_n|\|\mathbf{x}_n\|$$

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. □

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 $f_n : [0,1] \to \mathbb{R}$ be the function in Figure 1. It is constant zero except on the interval $[0, \frac{1}{n}]$ where it looks like a tent of height 1.

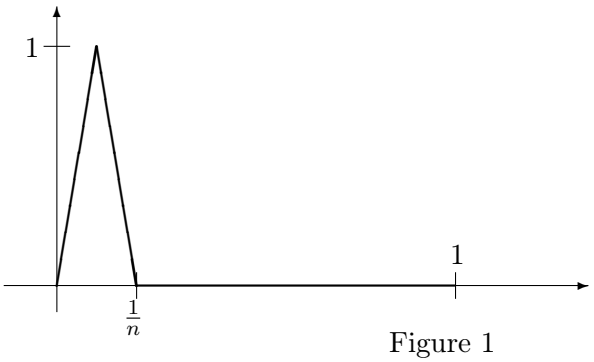


Figure 1

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.

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 on 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

$$\|f_n - f\| = \sup\{|f_n(x) - f(x)| \,:\, x \in [0,1]\} = \sup\{|f(x)| \,:\, x \in [0,1]\} = 1$$

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 $\mathbf{x} \in V$,*

$$\|\mathbf{x}\|_1 \leq K_1\|\mathbf{x}\|_2 \quad \textit{and} \quad \|\mathbf{x}\|_2 \leq K_2\|\mathbf{x}\|_1$$

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 (Y, d) is a metric space, and a map $f : Y \to X$ is continuous with respect to one of the norms, it is also continuous with respect to the other. Likewise, if a map $g : X \to Y$ 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 with 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 $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ be the usual basis in $\mathbb{R}^n$, and put

$$B = \max\{|\mathbf{e}_1|, |\mathbf{e}_2|, \ldots, |\mathbf{e}_n|\}$$

For $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \ldots + x_n\mathbf{e}_n$, we have

$$|\mathbf{x}| = |x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \ldots + x_n\mathbf{e}_n| \leq |x_1||\mathbf{e}_1| + |x_2||\mathbf{e}_2| + \ldots + |x_n||\mathbf{e}_n|$$

$$\leq B(|x_1| + |x_2| + \ldots + |x_n|) \leq nB \max_{1\leq i\leq n} |x_i|$$

Since

$$\max_{1\leq i\leq n} |x_i| = \sqrt{\max_{1\leq i\leq n} |x_i|^2} \leq \sqrt{x_1^2 + x_2^2 + \ldots x_n^2} = \|\mathbf{x}\|$$

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 $f : \mathbb{R}^n \to [0, \infty)$ 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

$$|\frac{\mathbf{x}}{\|x\|}| \geq a$$

which implies V

$$\frac{1}{a}|\mathbf{x}| \geq \|x\|$$

Hence we can choose $K_2 = \frac{1}{a}$, and the theorem is proved. □

The theorem above can be extended to all finite dimensional vector spaces by a simple trick (see Exercise 11).

We shall end this section with a brief look at product spaces. Assume that $(V_1, \|\cdot\|_1)$, $(V_2, \|\cdot\|_2)$, ..., $(V_n, \|\cdot\|_n)$ are vector spaces over $\mathbb{K}$. As usual,

$$V = V_1 \times V_2 \times \ldots \times V_n$$

is the set of all n-tuples $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, where $\mathbf{x}_i \in V_i$ for $i = 1, 2, \ldots, n$. If we define addition and scalar multiplication by

$$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) + (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) = (\mathbf{x}_1 + \mathbf{y}_1, \mathbf{x}_2 + \mathbf{y}_2, \ldots, \mathbf{x}_n + \mathbf{y}_n)$$

and

$$\alpha(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = (\alpha\mathbf{x}_1, \alpha\mathbf{x}_2, \ldots, \alpha\mathbf{x}_n) \ ,$$

V becomes a vector space over $\mathbb{K}$. It is easy to check that

$$\|\mathbf{x}\| = \|\mathbf{x}_1\|_1 + \|\mathbf{x}_2\|_2 + \ldots + \|\mathbf{x}_n\|_n$$

is a norm on V, and hence $(V, \|\cdot\|)$ is a normed space, called the *product* of $(V_1, \|\cdot\|_1)$, $(V_2, \|\cdot\|_2)$, $\ldots$, $(V_n, \|\cdot\|_n)$.

Proposition 5.1.8 *If the spaces $(V_1, \|\cdot\|_1)$, $(V_2, \|\cdot\|_2)$, $\ldots$, $(V_n, \|\cdot\|_n)$ 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 2 really is a norm.
4. Prove Proposition 5.1.4b).
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 the claim in the opening sentence of the proof of Theorem 5.1.7: that it suffices to prove that all norms are equivalent with the euclidean norm.
9. Check that the product $(V, \|\cdot\|)$ of normed spaces $(V_1, \|\cdot\|_1)$, $(V_2, \|\cdot\|_2)$, $\ldots$, $(V_n, \|\cdot\|_n)$ really is a normed space (you should check that V is a linear space as well as that $\|\cdot\|$ is a norm).
10. Prove Proposition 5.1.8.
11. Assume that V is a finite dimensional vector space with a basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$.
    a) Show that the function $T : \mathbb{R}^n \to V$ defined by
    $$T(x_1, x_2, \ldots, x_n) = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \ldots x_n\mathbf{e}_n$$
    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 $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ 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

$$\mathbf{x} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \ldots + \alpha_n \mathbf{v}_n$$

in a *unique* way. If such a (finite) set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ exists, we say that V is *finite dimensional* with dimension n (all bases have the same number of all 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 $\{\mathbf{u}_n\}_{n=1}^{\infty}$ is a sequence of elements in a normed vector space, we define the infinite sum $\sum_{n=1}^{\infty} \mathbf{u}_n$ as the limit of the partial sums $\mathbf{s}_n = \sum_{k=1}^{n} \mathbf{u}_k$ provided this limit exists; i.e.*

$$\sum_{n=1}^{\infty} \mathbf{u}_n = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbf{u}_k$$

When the limit exists, we say that the series converges*; otherwise it* diverges.

Remark: The notation $\mathbf{u} = \sum_{n=1}^{\infty} \mathbf{u}_n$ 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 $\mathbf{u} = \sum_{n=1}^{\infty} \mathbf{u}_n$ with respect to $\|\cdot\|_1$, but not with respect to $\|\cdot\|_2$, as $\|\mathbf{u} - \mathbf{s}_n\|_1 \to 0$ does not necesarily imply $\|\mathbf{u} - \mathbf{s}_n\|_2 \to 0$ (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 $\{\mathbf{e}_n\}_{n=1}^{\infty}$ be a sequence of elements in a normed vector space V. We say that $\{\mathbf{e}_n\}$ is a* basis[1] *for V if for each $\mathbf{x} \in V$ there is a unique sequence $\{\alpha_n\}_{n=1}^{\infty}$ from $\mathbb{K}$ such that*

$$\mathbf{x} = \sum_{n=1}^{\infty} \alpha_n \mathbf{e}_n$$

[1]Strictly 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 ia another kind of basis called a *Hamel basis* which does not require the space to be normed, but which is less practical for applications.

Not all normed spaces have a basis; there are, e.g., spaces so big 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 3: Let c_0 be the set of all sequences $\mathbf{x} = \{x_n\}_{n\in\mathbb{N}}$ of real numbers such that $\lim_{n\to\infty} x_n = 0$. It is not hard to check that $\{c_0\}$ is a vector space and that

$$\|\mathbf{x}\| = \sup\{|x_n| \,:\, n \in \mathbb{N}\}$$

is a norm on c_0. Let $\mathbf{e}_n = (0, 0, \ldots, 0, 1, 0, \ldots)$ be the sequence that is 1 on element number n and 0 elsewhere. Then $\{\mathbf{e}_n\}_{n\in\mathbb{N}}$ is a basis for c_0 with $\mathbf{x} = \sum_{n=1}^{\infty} x_n \mathbf{e}_n$. ♣

If a normed vector space is complete, we shall call it a *Banach space*. The next theorem provides an efficient method for checking that a normed space is complete. We say that a series $\sum_{n=1}^{\infty} \mathbf{u}_n$ in V *converges absolutely* if $\sum_{n=1}^{\infty} \|\mathbf{u}_n\|$ converges (note that $\sum_{n=1}^{\infty} \|\mathbf{u}_n\|$ 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 $\sum_{n=0}^{\infty} \mathbf{u}_n$ 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 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

$$\|\mathbf{s}_m - \mathbf{s}_n\| = \|\sum_{k=n+1}^{m} \mathbf{u}_k\| \leq \sum_{k=n+1}^{m} \|\mathbf{u}_k\| = |S_m - S_n| < \epsilon$$

when $n, m \geq N$, and hence $\{\mathbf{s}_n\}$ is a Cauchy sequence. Since V is complete, the series $\sum_{n=0}^{\infty} \mathbf{u}_n$ 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 2.5.5. □

Exercises for Section 5.2

1. Prove that the set $\{\mathbf{e}_n\}_{n\in\mathbb{N}}$ 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, 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 $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{K}$ 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 are the dot products in $\mathbb{R}^n$ and $\mathbb{C}^n$. If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are two real vectors, we define

$$\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x} \cdot \mathbf{y} = x_1y_1 + x_2y_2 + \ldots + x_ny_n$$

If $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ and $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ are two complex vectors, we define

$$\langle \mathbf{z}, \mathbf{w}\rangle = \mathbf{z} \cdot \mathbf{w} = z_1\overline{w_1} + z_2\overline{w_2} + \ldots + z_n\overline{w_n}$$

♣

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 $f, g : [a, b] \to \mathbb{R}$ are continuous functions, we get a complex valued function $h : [a, b] \to \mathbb{C}$ 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 separately.

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 $f : [a, b] \to \mathbb{R}$, and define the inner product by

$$\langle f, g\rangle = \int_a^b f(t)g(t)\, dt$$

[2]Strictly speaking, we are defining *positive definite* inner products, but they are the only inner products we have use for.

For the complex case, let V be the set of all continuous, complex valued functions $h : [a, b] \to \mathbb{C}$ 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 products. ♣

Given an inner product $\langle \cdot, \cdot \rangle$, we define $\| \cdot \| : V \to [0, \infty)$ 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, we 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* $\mathbf{u}_1$, $\mathbf{u}_2$, $\ldots$, $\mathbf{u}_n$ *in* V,

$$\|\mathbf{u}_1 + \mathbf{u}_2 + \ldots + \mathbf{u}_n\|^2 = \|\mathbf{u}_1\|^2 + \|\mathbf{u}_2\|^2 + \ldots + \|\mathbf{u}_n\|^2$$

Proof: We have

$$\|\mathbf{u}_1 + \mathbf{u}_2 + \ldots + \mathbf{u}_n\|^2 = \langle \mathbf{u}_1 + \mathbf{u}_2 + \ldots + \mathbf{u}_n, \mathbf{u}_1 + \mathbf{u}_2 + \ldots + \mathbf{u}_n \rangle =$$

$$= \sum_{1 \le i,j \le n} \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \|\mathbf{u}_1\|^2 + \|\mathbf{u}_2\|^2 + \ldots + \|\mathbf{u}_n\|^2$$

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 1 shows the idea.

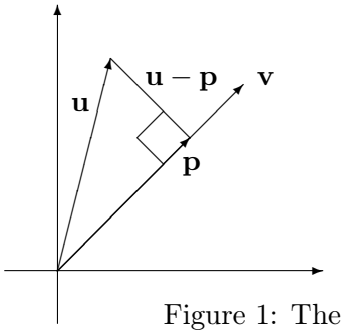


Figure 1: The projection $\mathbf{p}$ of $\mathbf{u}$ on $\mathbf{v}$

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 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

$$\|\mathbf{u}\|^2 = \|\mathbf{u} - \mathbf{p}\|^2 + \|\mathbf{p}\|^2 \geq \|\mathbf{p}\|^2$$

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 4.6.3, we have

$$\|\mathbf{u}\|^2 \geq \frac{|\langle \mathbf{u}, \mathbf{v} \rangle|^2}{\|\mathbf{v}\|^2}$$

and the proposition follows. □

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 $\mathrm{Re}(z)$ refers to the real part a of a complex number $z = a + ib$):

$$\|\mathbf{u} + \mathbf{v}\|^2 = \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle =$$

$$= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \overline{\langle \mathbf{u}, \mathbf{v} \rangle} + \langle \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + 2\mathrm{Re}(\langle \mathbf{u}, \mathbf{v} \rangle) + \langle \mathbf{v}, \mathbf{v} \rangle \leq$$

$$\leq \|\mathbf{u}\|^2 + 2\|\mathbf{u}\|\|\mathbf{v}\| + \|\mathbf{v}\|^2 = (\|\mathbf{u}\| + \|\mathbf{v}\|)^2$$

where we have used that according to Cauchy-Schwarz' inequality, we have $\mathrm{Re}(\langle \mathbf{u}, \mathbf{v} \rangle) \leq |\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\|\|\mathbf{v}\|$. □

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* $\alpha \in \mathbb{C}$ *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. □

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 $\sum_{n=0}^{\infty} \mathbf{w}_n$ converges in V, then

$$\|\sum_{n=0}^{\infty} \mathbf{w}_n\| = \lim_{N\to\infty} \|\sum_{n=0}^{N} \mathbf{w}_n\|$$

(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, $\lim_{n\to\infty}\langle \mathbf{u}_n, \mathbf{v}\rangle = \langle \lim_{n\to\infty} \mathbf{u}_n, \mathbf{v}\rangle$ for all $\mathbf{v} \in V$.

(iv) If the series $\sum_{n=0}^{\infty} \mathbf{w}_n$ converges in V, then

$$\langle \sum_{n=1}^{\infty} \mathbf{w}_n, \mathbf{v}\rangle = \sum_{n=1}^{\infty} \langle \mathbf{w}_n, \mathbf{v}\rangle$$

.

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 $\mathbf{u}_n \to \mathbf{u}$. To show that $\langle \mathbf{u}_n, \mathbf{v}\rangle \to \langle \mathbf{u}, \mathbf{v}\rangle$, is suffices to prove that $\langle \mathbf{u}_n, \mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle = \langle \mathbf{u}_n - \mathbf{u}, \mathbf{v}\rangle \to 0$. But by Cauchy-Schwarz' inequality

$$|\langle \mathbf{u}_n - \mathbf{u}, \mathbf{v}\rangle| \leq \|\mathbf{u}_n - \mathbf{u}\|\|\mathbf{v}\| \to 0$$

since $\|\mathbf{u}_n - \mathbf{u}\| \to 0$ by assumption.

(iv) We use (iii) with $\mathbf{u} = \sum_{n=1}^{\infty} \mathbf{w}_n$ and $\mathbf{u}_n = \sum_{k=1}^{n} \mathbf{w}_k$. Then

$$\langle \sum_{n=1}^{\infty} \mathbf{w}_n, \mathbf{v}\rangle = \langle \mathbf{u}, \mathbf{v}\rangle = \lim_{n\to\infty} \langle \mathbf{u}_n, \mathbf{v}\rangle = \lim_{n\to\infty} \langle \sum_{k=1}^{n} \mathbf{w}_k, \mathbf{v}\rangle =$$

$$= \lim_{n\to\infty} \sum_{k=1}^{n} \langle \mathbf{w}_k, \mathbf{v}\rangle = \sum_{n=1}^{\infty} \langle \mathbf{w}_n, \mathbf{v}\rangle$$

□

We shall now generalize some notions from linear algebra to our new setting. If $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is a finite set of elements in V, we define the *span*

$$\mathrm{Sp}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$$

of $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ to be the set of all linear combinations

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \ldots + \alpha_n\mathbf{u}_n, \qquad \text{where } \alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{K}$$

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 $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is an orthonormal set and $\mathbf{u} \in V$, we define the *projection of* $\mathbf{u}$ on $\mathrm{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ by

$$\mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u}) = \langle \mathbf{u}, \mathbf{e}_1\rangle\mathbf{e}_1 + \langle \mathbf{u}, \mathbf{e}_2\rangle\mathbf{e}_2 + \cdots + \langle \mathbf{u}, \mathbf{e}_n\rangle\mathbf{e}_n$$

This terminology is justified by the following result.

Proposition 5.3.8 *Let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ be an orthonormal set in V. For every $\mathbf{u} \in V$, the projection $\mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})$ is the element in $\mathrm{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ closest to $\mathbf{u}$. Moreover, $\mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})$ is orthogonal to all elements in $\mathrm{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$.*

Proof: We first prove the orthogonality. It suffices to prove that

$$\langle \mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u}), \mathbf{e}_i \rangle = 0 \tag{5.3.1}$$

for each $i = 1, 2, \ldots, n$, as we then have

$$\langle \mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u}), \alpha_1\mathbf{e}_1 + \cdots + \alpha_n\mathbf{e}_n \rangle =$$

$$= \overline{\alpha}_1 \langle \mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u}), \mathbf{e}_1 \rangle + \ldots + \overline{\alpha}_n \langle \mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u}), \mathbf{e}_n \rangle = 0$$

for all $\alpha_1\mathbf{e}_1 + \cdots + \alpha_n\mathbf{e}_n \in \mathrm{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$. To prove formula (5.3.1), just observe that for each $\mathbf{e}_i$

$$\langle \mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u}), \mathbf{e}_i \rangle = \langle \mathbf{u}, \mathbf{e}_i \rangle - \langle \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u}), \mathbf{e}_i \rangle$$

$$= \langle \mathbf{u}, \mathbf{e}_i \rangle - \big(\langle \mathbf{u}, \mathbf{e}_i \rangle \langle \mathbf{e}_1, \mathbf{e}_i \rangle + \langle \mathbf{u}, \mathbf{e}_2 \rangle \langle \mathbf{e}_2, \mathbf{e}_i \rangle + \cdots + \langle \mathbf{u}, \mathbf{e}_n \rangle \langle \mathbf{e}_n, \mathbf{e}_i \rangle \big) =$$

$$= \langle \mathbf{u}, \mathbf{e}_i \rangle - \langle \mathbf{u}, \mathbf{e}_i \rangle = 0$$

To prove that the projection is the element in $\mathrm{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ closest to $\mathbf{u}$, let $\mathbf{w} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \cdots + \alpha_n\mathbf{e}_n$ be another element in $\mathrm{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$. Then $\mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u}) - \mathbf{w}$ is in $\mathrm{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, and hence orthogonal to $\mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})$ by what we have just proved. By the Pythagorean theorem

$$\|\mathbf{u} - \mathbf{w}\|^2 = \|\mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})\|^2 + \|\mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u}) - \mathbf{w}\|^2 > \|\mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})\|^2$$

□

As an immediate consequence of the proposition above, we get:

Corollary 5.3.9 (Bessel's inequality) *Let* $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n, \ldots\}$ *be an orthonormal sequence in* V*. For any* $\mathbf{u} \in V$*,*

$$\sum_{i=1}^{\infty} |\langle \mathbf{u}, \mathbf{e}_i \rangle|^2 \leq \|\mathbf{u}\|^2$$

Proof: Since $\mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})$ is orthogonal to $\mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})$, we get by the Pythagorean theorem that for any n

$$\|\mathbf{u}\|^2 = \|\mathbf{u} - \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})\|^2 + \|\mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})\|^2 \geq \|\mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})\|^2$$

Using the Pythagorean Theorem again, we see that

$$\|\mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}(\mathbf{u})\|^2 = \|\langle \mathbf{u}, \mathbf{e}_1 \rangle \mathbf{e}_1 + \langle \mathbf{u}, \mathbf{e}_2 \rangle \mathbf{e}_2 + \cdots + \langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n\|^2 =$$

$$= \|\langle \mathbf{u}, \mathbf{e}_1 \rangle \mathbf{e}_1\|^2 + \|\langle \mathbf{u}, \mathbf{e}_2 \rangle \mathbf{e}_2\|^2 + \cdots + \|\langle \mathbf{u}, \mathbf{e}_n \rangle \mathbf{e}_n\|^2 =$$

$$= |\langle \mathbf{u}, \mathbf{e}_1 \rangle|^2 + |\langle \mathbf{u}, \mathbf{e}_2 \rangle|^2 + \cdots + |\langle \mathbf{u}, \mathbf{e}_n \rangle|^2$$

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 $n \to \infty$, the corollary follows. □

We have now reached the main result of this section. Recall from Definition 5.2.2 that $\{\mathbf{e}_i\}$ is a *basis* for V if any element $\mathbf{u}$ in V can be written as a linear combination $\mathbf{u} = \sum_{i=1}^{\infty} \alpha_i \mathbf{e}_i$ in a unique way. The theorem tells us that if the basis is orthonormal, the coeffisients α_i are easy to find; they are simply given by $\alpha_i = \langle \mathbf{u}, \mathbf{e}_i\rangle$.

Theorem 5.3.10 (Parseval's Theorem) *If $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n, \ldots\}$ is an orthonormal basis for V, then for all $\mathbf{u} \in V$, we have $\mathbf{u} = \sum_{i=1}^{\infty} \langle \mathbf{u}, \mathbf{e}_i\rangle \mathbf{e}_i$ and $\|\mathbf{u}\|^2 = \sum_{i=1}^{\infty} |\langle \mathbf{u}, \mathbf{e}_i\rangle|^2$.*

Proof: Since $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n, \ldots\}$ is a basis, we know that there is a unique sequence $\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots$ from $\mathbb{K}$ such that $\mathbf{u} = \sum_{n=1}^{\infty} \alpha_n \mathbf{e}_n$. This means that $\|\mathbf{u} - \sum_{n=1}^{N} \alpha_n \mathbf{e}_n\| \to 0$ as $N \to \infty$. Since the projection $\mathrm{P}_{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N}(\mathbf{u}) = \sum_{n=1}^{N} \langle \mathbf{u}, \mathbf{e}_n\rangle \mathbf{e}_n$ is the element in $\mathrm{Sp}\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N\}$ closest to $\mathbf{u}$, we have

$$\|\mathbf{u} - \sum_{n=1}^{N} \langle \mathbf{u}, \mathbf{e}_n\rangle \mathbf{e}_n\| \leq \|\mathbf{u} - \sum_{n=1}^{N} \alpha_n \mathbf{e}_n\| \to 0 \qquad \text{as } N \to \infty$$

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 $\mathbf{u} = \sum_{n=1}^{\infty} \langle \mathbf{u}, \mathbf{e}_n\rangle \mathbf{e}_n = \lim_{N\to\infty} \sum_{n=1}^{N} \langle \mathbf{u}, \mathbf{e}_n\rangle \mathbf{e}_n$, we have (recall Proposition 5.3.7(ii))

$$\|\mathbf{u}\|^2 = \lim_{N\to\infty} \|\sum_{n=1}^{N} \langle \mathbf{u}, \mathbf{e}_n\rangle \mathbf{e}_n\|^2 = \lim_{N\to\infty} \sum_{n=1}^{N} |\langle \mathbf{u}, \mathbf{e}_n\rangle|^2 = \sum_{n=1}^{\infty} |\langle \mathbf{u}, \mathbf{e}_n\rangle|^2$$

□

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 $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n, \ldots\}$. 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

$$\|\mathbf{s}_m - \mathbf{s}_n\|^2 = \| \sum_{k=n+1}^{m} \alpha_n \mathbf{e}_n \|^2 = \sum_{k=n+1}^{m} |\alpha_n|^2$$

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,

$$\langle \mathbf{u}, \mathbf{e}_i \rangle = \langle \sum_{n=1}^{\infty} \alpha_n \mathbf{e}_n, \mathbf{e}_i \rangle = \sum_{n=1}^{\infty} \langle \alpha_n \mathbf{e}_n, \mathbf{e}_i \rangle = \alpha_i$$

□

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
$$\int_a^b u(t)v'(t)\, dt = \Big[\, u(t)v(t) \Big]_a^b - \int_a^b u'(t)v(t)\, dt$$
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 $\langle \mathbf{u}_n, \mathbf{v}_n \rangle \to \langle \mathbf{u}, \mathbf{v} \rangle$.

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 $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ be an orthonormal set in an inner product space V. Show that the projection $P = \mathrm{P}_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n}$ 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

$$\langle \mathbf{u}, \mathbf{v}\rangle = \frac{1}{4}\left(\|\mathbf{u}+\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2\right)$$

    b) Show that if V is an inner product space over $\mathbb{C}$, then

$$\langle \mathbf{u}, \mathbf{v}\rangle = \frac{1}{4}\left(\|\mathbf{u}+\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2 + i\|\mathbf{u}+i\mathbf{v}\|^2 - i\|\mathbf{u}-i\mathbf{v}\|^2\right)$$

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,

$$\sum_{n=1}^{N} x_n y_n \le \left(\sum_{n=1}^{N} x_n^2\right)^{\frac{1}{2}} \left(\sum_{n=1}^{N} y_n^2\right)^{\frac{1}{2}}$$

    by 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 $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n, \ldots\}$. 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 important functions are the linear maps. The same holds for infinitely dimensional spaces, but here the linear maps are most often referred to as linear operators:

Definition 5.4.1 *Assume that V and W are two vector spaces over $\mathbb{K}$. A function $A : V \to W$ 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* $\mathbf{u}, \mathbf{v} \in V$.

Combining (i) and (ii), we see that

$$A(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha A(\mathbf{u}) + \beta A(\mathbf{v})$$

Using induction, this can be generalized to

$$A(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u_2} + \cdots + \alpha_n\mathbf{u}_n) = \alpha_1 A(\mathbf{u}_1) + \alpha_2 A(\mathbf{u_2}) + \cdots + \alpha_n A(\mathbf{u}_n) \quad (5.4.1)$$

It is also useful to observe 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 $A : V \to \mathbb{R}$ defined by

$$A(u) = \int_a^b u(x)\, dx$$

is a linear functional, while the function $B : V \to V$ 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. A function $f : (a, b) \to \mathbb{R}$ is said to be *infinitely differentiable* if it has derivatives of all orders at all points in (a, b), i.e. if $f^{(n)}(x)$ exists for all $n \in \mathbb{N}$ and all $x \in (a, b)$. Let U be the space of all infinitely differentiable functions, and define $D : U \to U$ by $Du(x) = u'(x)$. Then D is a linear operator. ♣

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* $A : V \to W$ *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. 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

$$\|A\| = \sup\{\|A(\mathbf{u})\|_W \ : \ \|\mathbf{u}\|_V = 1\} \tag{5.4.2}$$

We call $\|A\|$ the *operator norm* of A. The name is justified in Proposition 5.4.7 below.

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

$$|B(u)(x)| = |\int_a^x u(t)\,dt| \leq \int_a^x |u(t)|\,dt \leq \int_a^x \|u\|\,du = \|u\|(x-a) \leq \|u\|(b-a)$$

which implies that $\|B(u)\| \leq (b-a)\|u\|$ for all $u \in V$. ♣

Example 4: If we let U 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 $\|D(u_n)\| = \|n\cos nx\| \to \infty$ as $n \to \infty$. 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 $\frac{\|A\mathbf{u}_n\|_W}{\|\mathbf{u}_n\|_V} = M_n \geq n$. If we put $\mathbf{v}_n = \frac{\mathbf{u}_n}{M_n\|\mathbf{u}_n\|_V}$, we see that $\mathbf{v}_n \to \mathbf{0}$, but $A(\mathbf{v}_n)$ does not converge to $A(\mathbf{0}) = \mathbf{0}$ since $\|A(\mathbf{v}_n)\|_W = \|A(\frac{\mathbf{u}_n}{M_n\|\mathbf{u}_n\|_V})\| = \frac{\|A(\mathbf{u}_n)\|_W}{M_n\|\mathbf{u}_n\|_V} = \frac{M_n\|\mathbf{u}_n\|_V}{M_n\|\mathbf{u}_n\|_V} = 1$. By Proposition 3.2.5, this means that A is not contiuous at $\mathbf{0}$. □

Let us sum up the two lemmas in a theorem:

Theorem 5.4.5 *For linear operators $A : V \to W$ 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)$\Longrightarrow$(ii)$\Longrightarrow$(iii)$\Longrightarrow$(i). As (ii)$\Longrightarrow$(iii) is obvious, we just have to observe that (i)$\Longrightarrow$(ii) by Lemma 5.4.3 and (iii)$\Longrightarrow$(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 $A : V \to W$.*

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

$$\|\alpha A\| = \sup\left\{ \frac{\|\alpha A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \ : \ \mathbf{u} \neq \mathbf{0} \right\} = \sup\left\{ \frac{|\alpha|\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \ : \ \mathbf{u} \neq \mathbf{0} \right\}$$

$$= |\alpha| \sup\left\{ \frac{\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \ : \ \mathbf{u} \neq \mathbf{0} \right\} = |\alpha|\|A\|$$

(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!);

$$\|A + B\| = \sup\left\{ \frac{\|(A+B)(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \ : \ \mathbf{u} \neq \mathbf{0} \right\}$$

$$\leq \sup\left\{ \frac{\|(A(\mathbf{u})\| + \|B(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \ : \ \mathbf{u} \neq \mathbf{0} \right\}$$

$$\leq \sup\left\{ \frac{\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \ : \ \mathbf{u} \neq \mathbf{0} \right\} + \sup\left\{ \frac{\|B(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \ : \ \mathbf{u} \neq \mathbf{0} \right\}$$

$$= \|A\| + \|B\|$$

□

The spaces $\mathcal{L}(V, W)$ will play a central rôle 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.

$$A(\mathbf{u}) = \lim_{n\to\infty} A_n(\mathbf{u}) \qquad \text{for all } \mathbf{u} \in V$$

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

$$A(\alpha\mathbf{u}) = \lim_{n\to\infty} A_n(\alpha\mathbf{u}) = \alpha \lim_{n\to\infty} A_n(\mathbf{u}) = \alpha A(\mathbf{u})$$

and

$$A(\mathbf{u}+\mathbf{v}) = \lim_{n\to\infty} A_n(\mathbf{u}+\mathbf{v}) = \lim_{n\to\infty} A_n(\mathbf{u}) + \lim_{n\to\infty} A_n(\mathbf{v}) = A(\mathbf{u}) + A(\mathbf{v})$$

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, $|\,\|A_n\| - \|A_m\|\,| \le \|A_n - A_m\|$, 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 $M \ge \|A_n\|$ for all n. Thus for all $\mathbf{u} \neq \mathbf{0}$, we have

$$\frac{\|A_n(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \le M$$

and hence, by definition of A,

$$\frac{\|A(\mathbf{u})\|_W}{\|\mathbf{u}\|_V} \le M$$

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 $n, m \ge N$. This means that

$$\|A_n(\mathbf{u}) - A_m(\mathbf{u})\| \le \epsilon\|\mathbf{u}\|$$

for all $\mathbf{u} \in V$. If we let m go to infinity, we get (recall Proposition 5.1.4(i))

$$\|A_n(\mathbf{u}) - A(\mathbf{u})\| \le \epsilon\|\mathbf{u}\|$$

for all $\mathbf{u}$, which means that $\|A_n - A\| \le \epsilon$. This shows that $\{A_n\}$ converges to A, and the proof is complete. □

Exercises for Section 5.4

1. Prove Formula (5.4.1).
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. Prove formula (5.4.2).
5. Define $F : C([0,1],\mathbb{R}) \to \mathbb{R}$ 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 $A : U \to V$ and $B : V \to W$ 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 $A : \mathbb{R}^d \to W$ 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 $A : V \to \mathbb{K}$ be a linear functional. The *kernel* of A is defined by

$$\ker(A) = \{\mathbf{v} \in V \ : \ A(\mathbf{v}) = \mathbf{0}\} = A^{-1}(\{\mathbf{0}\})$$

   a) Show that if A is bounded, $\ker(A)$ is closed. (*Hint:* Recall Proposition 3.3.10)

   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 $A(a)$ is a number, this makes sense).

   b) Show that $A(\mathbf{b}) = 1$ and that there is a ball $\mathrm{B}(\mathbf{b}; r)$ around $\mathbf{b}$ contained in $\ker A^c$.

   c) Show that if $\mathbf{u} \in \mathrm{B}(\mathbf{0}; r)$ (where r is as in b) above), then $\|A(\mathbf{u})\|_W \leq 1$. (*Hint:* Assume for contradiction that $\mathbf{u} \in \mathrm{B}(\mathbf{0}, r)$, but $\|A(\mathbf{u})\|_W > 1$, and show that $A(\mathbf{b} - \frac{\mathbf{u}}{A(\mathbf{u})}) = 0$ although $\mathbf{b} - \frac{\mathbf{u}}{A(\mathbf{u})} \in \mathrm{B}(\mathbf{b}; r)$.)

   d) Use a) and c) to prove:

   Teorem: *Assume that $(V, \|\cdot\|_V)$ is a normed spaces over $\mathbb{K}$. A linear functional $A : V \to \mathbb{K}$ 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 $A : V \to \mathbb{R}$ 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 $\left(\sum_{i=1}^\infty \beta_i^2\right)^{\frac{1}{2}} \leq \|A\|$.

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 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 it 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 $\mathrm{B}(x; r) \subseteq G$ contains elements from A. The following definition catches our intiution of a set that is not dense anywhere.

Definition 5.5.1 *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, for all nonempty, open sets $G \subseteq X$, there is a ball $\mathrm{B}(x; r) \subseteq G$ 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 5.5.2 *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]

[3]Most books refer to meager sets as "sets of first category" while comeager sets are

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 $\mathbb{Q}$.

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 5.5.3 (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 $\mathrm{B}(a; r)$, our task is to find an element $x \in \mathrm{B}(a; r)$ which does not belong to M.

We first observe that since N_1 is nowhere dense, there is a ball $\mathrm{B}(a_1; r_1)$ inside $\mathrm{B}(a; r)$ which does not intersect N_1. By shrinking the radius r_1 slightly if necessary, we may assume that the *closed* ball $\overline{\mathrm{B}}(a_1; r_1)$ is contained in $\mathrm{B}(a; r)$, does not intersect N_1, and has radius less than 1. Since N_2 is nowhere dense, there is a ball $\mathrm{B}(a_2; r_2)$ inside $\mathrm{B}(a_1; r_1)$ which does not intersect N_2. By shrinking the radius r_2 if necessary, we may assume that the closed ball $\overline{\mathrm{B}}(a_2; r_2)$ does not intersect N_2 and has radius less than $\frac{1}{2}$. Continuing in this way, we get a sequence $\{\overline{\mathrm{B}}(a_k; r_k)\}$ of closed balls, each contained in the previous, such that $\overline{\mathrm{B}}(a_k; r_k)$ 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 $\overline{\mathrm{B}}(a_k; r_k)$ is closed, and the "tail" $\{a_n\}_{n=k}^{\infty}$ of the sequence belongs to $\overline{\mathrm{B}}(a_k; r_k)$, the limit x also belongs to $\overline{\mathrm{B}}(a_k; r_k)$. This means that for all k, $x \notin N_k$, and hence $x \notin M$. Since $\overline{\mathrm{B}}(a_1; r_1) \subseteq \mathrm{B}(a; r)$, we see that $x \in \mathrm{B}(a; r)$, and the theorem is proved. □

As an immediate consequence we have:

Corollary 5.5.4 *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.

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 nondescriptive that it should be abandoned in favor of more evocative terms.

Lemma 5.5.5 *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 $\mathrm{B}(x; r_1)$ around x that does not intersect F. Since G is open, there is a ball $\mathrm{B}(x; r_2)$ around x that is contained in G. If we choose $r = \min\{r_1, r_2\}$, the ball $\mathrm{B}(x; r)$ is contained in G and does not intersect F, and hence F is nowhere dense. □

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.

Definition 5.5.6 *Let (X, d) be a metric space. A family $\mathcal{F}$ of functions $f : X \to \mathbb{R}$ is called* pointwise bounded *if for each $x \in X$, there is a constant $M_x \in \mathbb{R}$ such that $|f(x)| \leq M_x$ for all $f \in \mathcal{F}$.*

Note that the constant M_x may vary from point to point, and that there need not be a constant M such that $|f(x)| \leq M$ for all f and all x (a simple example is $\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R} \,|\, f(x) = kx \text{ for } k \in [-1, 1}\}$, where $M_x = |x|$). The next result shows that although we cannot guarantee boundedness on all of X, we can under reasonable assumptions guarantee boundedness on a part of X.

Proposition 5.5.7 *Let (X, d) be a complete metric space, and assume that $\mathcal{F}$ is a pointwise bounded family of continuous functions $f : X \to \mathbb{R}$. Then there exists an open, nonempty set G and a constant $M \in \mathbb{R}$ such that $|f(x)| \leq M$ for all $f \in \mathcal{F}$ and all $x \in G$.*

Proof: For each $n \in \mathbb{N}$ and $f \in \mathcal{F}$, the set $f^{-1}([-n, n])$ is closed as it is the inverse image of a closed set under a continuous function (recall Proposition 3.3.10). As intersections of closed sets are closed (Proposition 3.3.12)

$$A_n = \bigcap_{f \in \mathcal{F}} f^{-1}([-n, n])$$

is also closed. Since $\mathcal{F}$ is pointwise bounded, $X = \bigcup_{n \in \mathbb{N}} A_n$, and Corollary 5.5.4 tells us that not all A_n can be nowhere dense. If A_{n_0} is not nowhere dense, it contains an open set G by the lemma above. By definition of A_{n_0},

we see that $|f(x)| \leq n_0$ for all $f \in \mathcal{F}$ and all $x \in A_{n_0}$ (and hence all $x \in G$). □

You may doubt the usefulness of this theorem as we only know that the result holds for *some* open set G, but the point is that if we have extra information on the the family $\mathcal{F}$, the sole existence of such a set may be exactly what we need to pull through a more complex argument. This is what happens in the next result where we return to the setting of normed spaces.

Theorem 5.5.8 ***(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 $\lim_{n\to\infty} A_n(\mathbf{u})$ exists for all $\mathbf{u} \in V$ (we say that the sequence $\{A_n\}$* converges pointwise*). Then the function $A : V \to W$ defined by*

$$A(\mathbf{u}) = \lim_{n\to\infty} A_n(\mathbf{u})$$

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), and we concentrate on the boundedness. Define $f_n : V \to \mathbb{R}$ by $f_n(u) = \|A_n(\mathbf{u})\|$. Since the sequence $\{A_n(\mathbf{u})\}$ converges for any $\mathbf{u}$, the sequence $\{f_n(\mathbf{u})\}$ is bounded. Hence $\{f_n\}$ is a pointwise bounded family in the terminology of the proposition above, and there exist an open set G and a constant M such that $f_n(\mathbf{u}) \leq M$ for all $\mathbf{u} \in G$ and all $n \in \mathbb{N}$. In other words, $\|A_n(\mathbf{u})\| \leq M$ for all $\mathbf{u} \in G$ and all $n \in \mathbb{N}$. As $A(\mathbf{u}) = \lim_{n\to\infty} A_n(\mathbf{u})$, this means that $\|A(\mathbf{u})\| \leq M$ for all $\mathbf{u} \in G$.

To show that A is bounded, pick a point $\mathbf{a} \in G$ and a radius $r > 0$ such that the closed ball $\overline{\mathrm{B}}(\mathbf{a}, r)$ is contained in G. Since for any $\mathbf{u} \in V$, we must have $\mathbf{a} + \frac{\mathbf{u}}{\|\mathbf{u}\|} r \in \overline{\mathrm{B}}(\mathbf{a}, r) \subseteq G$, we see that

$$\|A(\mathbf{a} + \frac{\mathbf{u}}{\|\mathbf{u}\|} r)\| \leq M$$

and hence by linearity

$$\|A(\mathbf{a}) + \frac{r}{\|\mathbf{u}\|} A(\mathbf{u})\| \leq M$$

Playing with the triangle inequality, we now get

$$\|\frac{r}{\|\mathbf{u}\|} A(\mathbf{u})\| = \|A(\mathbf{a}) + \frac{r}{\|\mathbf{u}\|} A(\mathbf{u}) - A(\mathbf{a})\|$$

$$\leq \|A(\mathbf{a}) + \frac{r}{\|\mathbf{u}\|} A(\mathbf{u})\| + \|A(\mathbf{a})\| \leq 2M$$

and hence

$$\|A(\mathbf{u})\| \leq \frac{2M}{r}\|\mathbf{u}\|$$

which shows that A is bounded. □

The Banach-Steinhaus Theorem is one of several important results about linear operators that rely on Baire's Category Theorem. We shall meet more examples in the next section.

For our next 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 can not be written as a countable union of compact sets. We need a lemma.

Lemma 5.5.9 *A compact subset K of $C([0,1],\mathbb{R})$ is nowhere dense.*

Proof: Since compact sets are closed, it suffices (by Lemma 5.5.5) to show that each ball $\mathrm{B}(f;\epsilon)$ contains elements that are not in K. By Arzelà-Ascoli's Theorem, we know that compact sets are equicontinuous, and hence we need only prove that $\mathrm{B}(f;\epsilon)$ 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 $\{f + g_n\}$ is not equicontinuous (see Exercise 9 for help to prove this), all these functions can not be in K, and hence $\mathrm{B}(f;\epsilon)$ contains elements that are not in K. □

Proposition 5.5.10 *$C([0,1],\mathbb{R})$ 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 5.5.4. Since the lemma tells us that all compact sets are nowhere dense, the theorem follows. □

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 5.5

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 $\mathrm{B}(a_0;r_0) \subseteq X$, there is a ball $\mathrm{B}(x;r) \subseteq \mathrm{B}(a_0;r_0)$ 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.
7. Show that if $N_1, N_2, \ldots, N_k$ are nowhere dense, so is $N_1 \cup N_2 \cup \ldots N_k$.
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 5.5.8 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 equicontinous family of functions $h_n : [0,1] \to \mathbb{R}$ and $k : [0,1] \to \mathbb{R}$ 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?
11. Show that in a complete space, a closed set is meager if and only if it is nowhere dense.
12. 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
13. Assume that a sequence $\{f_n\}$ of continuous functions $f_n : [0,1] \to \mathbb{R}$ 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 $[0,1]$.

14. 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 $f_N(x) > 1$.

15. Let (X, d) be a metric space. A point $x \in X$ is called *isolated* if there is an $\epsilon > 0$ such that $\mathrm{B}(x; \epsilon) = \{x\}$.

    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 $[0, 1]$ can *not* 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.

5.6 A group of famous theorems

In this section, we shall use Baire's Category Theorem 5.5.3 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 $f : X \to Y$ between metric spaces is continuous if the inverse image $f^{-1}(O)$ of every open set O is open (recall Proposition 2.3.9). There is a dual notion for forward images.

Definition 5.6.1 *A function $f : X \to Y$ 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 5.6.2 (Open Mapping Theorem) *Assume that X, Y are two complete, normed spaces, and that $A : X \to Y$ 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 5.6.3 *Assume that $A : X \to Y$ is a linear operator from one normed space to another. If there is a ball $\mathrm{B}(\mathbf{0}, t)$ around the origin in X whose image $A(\mathrm{B}(\mathbf{0}, t))$ contains a ball $\mathrm{B}(\mathbf{0}, s)$ around the origin in Y, then A is open.*

Proof: Assume that $O \subseteq X$ 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 $\mathrm{B}(\mathbf{a}, \frac{t}{N}) \subseteq O$. The idea is that since A is linear, we should have $A(B(\mathbf{a}, \frac{t}{N})) \supseteq \mathrm{B}(A(\mathbf{a}), \frac{s}{N})$, and since $A(O) \supseteq A(B(\mathbf{a}, \frac{t}{N}))$, the lemma will follow.

It remains to check that we really have $A(B(\mathbf{a}, \frac{t}{N})) \supseteq \mathrm{B}(A(\mathbf{a}), \frac{s}{N})$. Let $\mathbf{y}$ be an arbitrary element of $\mathrm{B}(A(\mathbf{a}), \frac{s}{N})$; then $\mathbf{y} = A(\mathbf{a}) + \frac{1}{N}\mathbf{v}$ where $\mathbf{v} \in \mathrm{B}(\mathbf{0}, s)$. We know there is a $\mathbf{u} \in \mathrm{B}(\mathbf{0}, t)$ 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 $y \in A(B(\mathbf{a}, \frac{t}{N}))$. □

The next step is the crucial one.

Lemma 5.6.4 *Assume that X, Y are two complete, normed spaces, and that $A : X \to Y$ is a surjective, linear operator. Then there is a ball $\mathrm{B}(\mathbf{0}, r)$ such that the closure $\overline{A(\mathrm{B}(\mathbf{0}, r))}$ of the image $A(\mathrm{B}(\mathbf{0}, r))$ contains an open ball $\mathrm{B}(\mathbf{0}, s)$.*

Proof: Since A is surjective, $Y = \bigcup_{n\in\mathbb{N}} A(\mathrm{B}(\mathbf{0}, n))$. By Corollary 5.5.4, the sets $A(\mathrm{B}(\mathbf{0}, n))$ cannot all be nowhere dense. If $A(\mathrm{B}(\mathbf{0}, n))$ fails to be nowhere dense, so does it's closure $\overline{A(\mathrm{B}(\mathbf{0}, n))}$, and by Lemma 5.5.5, $\overline{A(\mathrm{B}(\mathbf{0}, n))}$ contains an open ball $\mathrm{B}(\mathbf{b}, s)$.

We have to "move" the ball $\mathrm{B}(\mathbf{b}, s)$ to the origin. Note that if $\mathbf{y} \in \mathrm{B}(\mathbf{0}, s)$, then both $\mathbf{b}$ and $\mathbf{b} + \mathbf{y}$ belong to $\mathrm{B}(\mathbf{b}, s)$ and hence to $\overline{A(\mathrm{B}(\mathbf{0}, n))}$. Consequently there are sequences $\{\mathbf{u}_k\}$, $\{\mathbf{v}_k\}$ from $\mathrm{B}(\mathbf{0}, n)$ 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 $\|\mathbf{v}_k - \mathbf{u}_k\| \leq \|\mathbf{u}_k\| + \|\mathbf{v}_k\| < 2n$, and $\mathbf{y}$ is an arbitrary element in $\mathrm{B}(\mathbf{0}, s)$, we get that $\mathrm{B}(\mathbf{0}, s) \subseteq \overline{A(\mathrm{B}(\mathbf{0}, 2n))}$. Hence the lemma is proved with $r = 2n$. □

To prove the theorem, we need to get rid of the closure in $\overline{A(\mathrm{B}(\mathbf{0}, r))}$. It is important to understand what this means. That the ball $\mathrm{B}(\mathbf{0}, s)$ is contained in $A(\mathrm{B}(\mathbf{0}, r))$, means that every $\mathbf{y} \in \mathrm{B}(\mathbf{0}, s)$ is the image $\mathbf{y} = A(\mathbf{x})$ of an

element $\mathbf{x} \in \mathrm{B}(\mathbf{0}, r)$; that $\mathrm{B}(\mathbf{0}, s)$ is contained in *the closure* $\overline{A(\mathrm{B}(\mathbf{0}, r))}$, means that every $\mathbf{y} \in \mathrm{B}(\mathbf{0}, s)$ can be approximated arbitrarily well by images $\mathbf{y} = A(\mathbf{x})$ of elements $\mathbf{x} \in \mathrm{B}(\mathbf{0}, r)$; i.e., for every $\epsilon > 0$, there is an $\mathbf{x} \in \mathrm{B}(\mathbf{0}, r)$ 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 $q > 0$, $\mathrm{B}(\mathbf{0}, qs)$ is contained in $\overline{A(\mathrm{B}(\mathbf{0}, qr))}$. In particular, $\mathrm{B}(\mathbf{0}, \frac{s}{2^k}) \subseteq \overline{A(\mathrm{B}(\mathbf{0}, \frac{r}{2^k}))}$ for all $k \in \mathbb{N}$. We shall use this repeatedly in the proof below.

Proof of Open Mapping Theorem: Let r and s be as in the lemma above. According to Lemma 5.5.3 it suffices to prove that $A(\mathrm{B}(\mathbf{0}, 2r)) \supseteq \mathrm{B}(\mathbf{0}, s)$. This means that given a $\mathbf{y} \in \mathrm{B}(\mathbf{0}, s)$, we must show that there is an $\mathbf{x} \in \mathrm{B}(\mathbf{0}, 2r)$ 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 $\mathbf{x}_1 \in \mathrm{B}(\mathbf{0}, r)$ 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 $\mathbf{y} - A(\mathbf{x}_1) \in \mathrm{B}(\mathbf{0}, \frac{s}{2})$, and hence there is an $\mathbf{x}_2 \in \mathrm{B}(\mathbf{0}, \frac{r}{2})$ 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 $\mathbf{y} - (A(\mathbf{x}_1)) + A(\mathbf{x}_2)) \in \mathrm{B}(\mathbf{0}, \frac{s}{4})$, and hence there is an $\mathbf{x}_3 \in \mathrm{B}(\mathbf{0}, \frac{r}{4})$ such that $\|\big(\mathbf{y} - (A(\mathbf{x}_1) + A(\mathbf{x}_2))\big) - A(\mathbf{x}_3)\| < \frac{s}{8}$, 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 $\|\mathbf{y} - A(\mathbf{x}_1 + \mathbf{x}_2 + \ldots \mathbf{x}_n)\| < \frac{s}{2^n}$. The sequence $\{\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_n\}$ is a Cauchy sequence, and since X is complete, it converges to an element $\mathbf{x} = \sum_{n=1}^{\infty} \mathbf{x}_n$. Since A is continuous, $A(\mathbf{x}) = \lim_{n\to\infty} A(\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_n)$, and since $\|\mathbf{y} - A(\mathbf{x}_1 + \mathbf{x}_2 + \ldots \mathbf{x}_n)\| < \frac{s}{2^n}$, 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 $\mathbf{x} \in \mathrm{B}(\mathbf{0}, 2r)$ such that $\mathbf{y} = A(\mathbf{x})$, and the proof is complete. □

The Open Mapping Theorem has an immediate consequence that will be important to in the next chapter.

Theorem 5.6.5 (Bounded Inverse Theorem) *Assume that X, Y are two complete, normed spaces, and that $A : X \to Y$ is a bijective, bounded, linear operator. Then the inverse A^{-1} is also bounded.*

Proof: According to the Open Mapping Theorem, A is open. Hence for any open set $O \subseteq X$, we see that $(A^{-1})^{-1}(O) = A(O)$ is open. This shows that A^{-1} is continuous, which is the same as bounded. □

The next theorem needs a little introduction. Assume that $A : X \to Y$ is a linear operator between two normed spaces. The *graph* of A is the set

$$G(A) = \{(\mathbf{x}, A(\mathbf{x})) \mid \mathbf{x} \in X\}$$

$G(A)$ is clearly a subset of the product space $X \times Y$, and since A is linear, it is easy to check that it is actually a subspace of $X \times Y$ (see Exercise 3 if you need help).

Theorem 5.6.6 (Closed Graph Theorem) *Assume that X, Y are two complete, normed spaces, and that $A : X \to Y$ is a linear operator. Then A is bounded if and only if $G(A)$ is a* closed *subspace of $X \times Y$.*

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 $X \times Y$, then $(\mathbf{x}, \mathbf{y})$ belong to $G(A)$, i.e. $\mathbf{y} = G(\mathbf{x})$. But if $\{(\mathbf{x}_n, A(\mathbf{x}_n))\}$ converges to $(\mathbf{x}, \mathbf{y})$, then $\{\mathbf{x}_n\}$ converges to $\mathbf{x}$ in X and $\{A(\mathbf{x}_n)\}$ converges to $\mathbf{y}$ in Y. Since A is continuous, this means that $\mathbf{y} = A(\mathbf{x})$ (recall Proposition 3.2.9). Hence the limit belongs to $G(A)$, and $G(A)$ is closed.

The other direction is a very clever trick. If $G(A)$ is closed, it is complete as a closed subspace of the complete space $X \times Y$ (remember Proposition 5.1.8). Define $\pi : G(A) \to X$ 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?). □

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 Banach-Steinhaus Theorem 5.5.8 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 5.6

1. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. Show that f is *not* open.

2. Assume that $A : X \to Y$ is a linear operator. Show that if $\mathrm{B}(\mathbf{0}, s)$ is contained in $\overline{A(\mathrm{B}(\mathbf{0}, r))}$, then $\mathrm{B}(\mathbf{0}, qs)$ is contained in $\overline{A(\mathrm{B}(\mathbf{0}, qr))}$ 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 $X \times Y$. 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 $id : X \to X$, the map that sends all elements to themselves.)

6. Assume that X, Y, and Z are complete, normed spaces and that $A : X \to Z$ and $B : Y \to Z$ are two bounded, linear maps. Assume that for every $x \in X$, the equation $A(x) = B(y)$ has a unique solution $y = C(x)$. Show that $C : X \to Y$ is a bounded, linear operator. (*Hint:* Use the Closed Graph Theorem).

7. Assume that $(X; \|\cdot\|_X)$ and $(Y; \|\cdot\|_Y)$ are two complete, normed spaces, and that $A : X \to Y$ is an injective, bounded, linear operator. Show that the following are equivalent:

   (i) The image $A(X)$ is a closed subspace of Y.

   (ii) A is *bounded below*, i.e., there is a real number $a > 0$ such that $\|A(\mathbf{x})\|_Y \geq a\|\mathbf{x}\|_X$ for all $\mathbf{x} \in X$.

8. We shall look at an example which illustrates 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 $A : l_2 \to l_2$ by

$$A(\{x_1, x_2, x_3, \ldots, x_n, \ldots\}) = \{x_1, \frac{x_2}{2}, \frac{x_3}{3}, \ldots, \frac{x_n}{n}, \ldots\}$$

   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 $\mathbf{x} \in X$. Show that A is injective, but *not* bounded below.

   c) Let Y be the image of A, i.e., $Y = A(l_2)$. Explain that Y is a subspace of l_2, but that Y 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 $A : l_2 \to Y$. Show that the inverse $A^{-1} : Y \to l_2$ 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 $l_2 \times Y$ (*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?

Chapter 6

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, as new functions derived from old ones according to certain rules etc. If we think of functions of several variables, there is even more variety – we have directional derivatives, partial derivatives, gradients, Jacobi matrices, total derivatives etc. In this chapter we shall extend the notion even further, to normed spaces, and we need a unifying idea to hold on to.

Perhaps somewhat surprisingly, this idea will be *linear approximation*: Our derivatives will always be linear approximations to functional differences of the kind $f(a+r)-f(a)$ for small r. Recall that if $f : \mathbb{R} \to \mathbb{R}$ is a function of one variable, $f(a+r) - f(a) \approx f'(a)r$ for small r; if $f : \mathbb{R}^n \to \mathbb{R}$ is 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 $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is a vector valued function, $\mathbf{F}(\mathbf{a}+\mathbf{r})-\mathbf{F}(\mathbf{a}) \approx \mathbf{F}'(\mathbf{a})\mathbf{r}$ for small $\mathbf{r}$, where $\mathbf{F}'(\mathbf{a})$ is the Jacobi 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 weak, 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 smaller and smaller *even compared to the size of* r, i.e., the fraction

$$\frac{f(a+r) - f(a) - f'(a)r}{r}$$

goes to 0 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 separate as you zoom in.

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 $\mathbb{C}$. I shall use the symbol $\|\cdot\|$ to denote the norms in both spaces – it should always be clear from the context which one is meant. Our first task will be to define derivatives of functions $\mathbf{F} : X \to Y$. The following definition should not be surprising after the discussion above.

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 $\mathbf{F} : O \to Y$. If $\mathbf{a}$ is a point in O, a bounded, linear map $A : X \to Y$ is called a* derivative of $\mathbf{F}$ *at $\mathbf{a}$ if*

$$\sigma(\mathbf{r}) = \mathbf{F}(\mathbf{a}+\mathbf{r}) - \mathbf{F}(\mathbf{a}) - A(\mathbf{r})$$

goes to $\mathbf{0}$ faster than $\mathbf{r}$, i.e., if

$$\lim_{\mathbf{r}\to\mathbf{0}} \frac{\|\sigma(\mathbf{r})\|}{\|\mathbf{r}\|} = 0$$

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 $\mathbf{F}$ can not 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 any $\mathbf{x}$ 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 inequality, 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 $t \to 0$, the expression on the right goes to 0, and hence $\|A(\mathbf{x}) - B(\mathbf{x})\|$ must be 0, which means that $A(\mathbf{x}) = B(\mathbf{x})$. □

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 $\mathbf{F} : O \to Y$. 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 is 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 straight forward generalization of what you are used to. If $\mathbf{F}$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, the Jacobi matrix is just the matrix of $\mathbf{F}'(\mathbf{a})$ with respect to the standard bases in $\mathbb{R}^n$ and $\mathbb{R}^m$.

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

$$\sigma(\mathbf{r}) = \mathbf{F}(\mathbf{a}+\mathbf{r}) - \mathbf{F}(\mathbf{a}) - \mathbf{F}'(\mathbf{a})(\mathbf{r}) \tag{6.1.1}$$

and check that $\sigma(\mathbf{r})$ goes to $\mathbf{0}$ faster than $\mathbf{r}$, i.e., that

$$\lim_{\mathbf{r}\to\mathbf{0}} \frac{\|\sigma(\mathbf{r})\|}{\|\mathbf{r}\|} = 0 \tag{6.1.2}$$

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 $\mathbf{F} : O \to Y$ 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

$$\lim_{\mathbf{r}\to\mathbf{0}} \mathbf{F}(\mathbf{a}+\mathbf{r}) = \mathbf{F}(\mathbf{a})$$

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 $\mathbf{F} : X \to Y$ 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 $\mathbf{a}$. Also recall that if $\mathbf{F}$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, then the Jacobi matrix of $\mathbf{F}$ is just the matrix of $\mathbf{F}$ (with respect to the standard bases in $\mathbb{R}^n$ and $\mathbb{R}^m$).

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 $\mathbf{F} : X \to Y$ is constant. The $\mathbf{F}$ 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 $\mathbf{F}, \mathbf{G} : O \to Y$ 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

$$\sigma(\mathbf{r}) = \big(\mathbf{F}(\mathbf{a}+\mathbf{r}) + \mathbf{G}(\mathbf{a}+\mathbf{r})\big) - \big(\mathbf{F}(\mathbf{a}) + \mathbf{G}(\mathbf{a})\big) - \big(\mathbf{F}'(\mathbf{a})(\mathbf{r}) + \mathbf{G}'(\mathbf{a})(\mathbf{r})\big)$$

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. □

As we need not have a notion of multiplication in our target space Y, there is no canonical generalization of the product rule[1], 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 $\alpha : O \to \mathbb{K}$ and $\mathbf{F} : O \to Y$ 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

$$(\alpha\mathbf{F})'(\mathbf{a}) = \alpha\mathbf{F}'(\mathbf{a})$$

[1]Strictly speaking, this is not quite true. There is a notion of *bilinear maps* that can be used to formulate an extremely general version of the product rule, but we postpone this discussion till Proposition 6.8.5.

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 linear map!), we see that

$$\sigma(\mathbf{r}) = \mathbf{G}(\mathbf{a}+\mathbf{r}) - \mathbf{G}(\mathbf{a}) - \mathbf{G}'(\mathbf{a})(\mathbf{r})$$

$$= \alpha(\mathbf{a}+\mathbf{r})\mathbf{F}(\mathbf{a}+\mathbf{r}) - \alpha(\mathbf{a})\mathbf{F}(\mathbf{a}) - \alpha'(\mathbf{a})(\mathbf{r})\mathbf{F}(\mathbf{a}) - \alpha(\mathbf{a})\mathbf{F}'(\mathbf{a})(\mathbf{r})$$

$$= \big(\alpha(\mathbf{a}) + \alpha'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r})\big)\big(\mathbf{F}(\mathbf{a}) + \mathbf{F}'(\mathbf{a})(\mathbf{r}) + \sigma_2(\mathbf{r})\big)$$

$$-\alpha(\mathbf{a})\mathbf{F}(\mathbf{a}) - \alpha'(\mathbf{a})(\mathbf{r})\mathbf{F}(\mathbf{a}) - \alpha(\mathbf{a})\mathbf{F}'(\mathbf{a})(\mathbf{r})$$

$$= \alpha(\mathbf{a})\sigma_2(\mathbf{r}) + \alpha'(\mathbf{a})(\mathbf{r})\mathbf{F}'(\mathbf{a})(\mathbf{r}) + \alpha'(\mathbf{a})(\mathbf{r})\sigma_2(\mathbf{r}) + \sigma_1(\mathbf{r})\mathbf{F}(\mathbf{a})$$

$$+\sigma_1(\mathbf{r})\mathbf{F}'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r})\sigma_2(\mathbf{r})$$

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 five 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 $\mathbf{r}$. □

Before we prove the Chain Rule, it's useful to agree on notation. If A, B, C are three sets, and $g : A \to B$ and $f : B \to C$ are two functions, the *composite* function $f \circ g : A \to C$ is defined in the usual way by

$$(f \circ g)(a) = f(g(a)) \quad \text{for all } a \in A$$

If g and f are linear maps, it is easy to check that $f \circ g$ is also a linear map.

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 $\mathbf{G} : O_1 \to O_2$ and $\mathbf{F} : O_2 \to Z$ 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

$$\sigma_1(\mathbf{r}) = \mathbf{G}(\mathbf{a}+\mathbf{r}) - \mathbf{G}(\mathbf{a}) - \mathbf{G}'(\mathbf{a})(\mathbf{r}) \tag{6.1.3}$$

and

$$\sigma_2(\mathbf{s}) = \mathbf{F}(\mathbf{b}+\mathbf{s}) - \mathbf{F}(\mathbf{b}) - \mathbf{F}'(\mathbf{b})(\mathbf{s}) \tag{6.1.4}$$

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

$$\sigma(\mathbf{r}) = \mathbf{H}(\mathbf{a}+\mathbf{r}) - \mathbf{H}(\mathbf{a}) - \mathbf{H}'(\mathbf{a})(\mathbf{r})$$

$$= \mathbf{F}(\mathbf{G}(\mathbf{a}+\mathbf{r})) - \mathbf{F}(\mathbf{G}(\mathbf{a})) - \mathbf{F}'(\mathbf{G}(\mathbf{a}))(\mathbf{G}'(\mathbf{a})(\mathbf{r}))$$

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

$$\sigma(\mathbf{r}) = \mathbf{F}(\mathbf{b}+\mathbf{s}) - \mathbf{F}(\mathbf{b}) - \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r}))$$

$$= \mathbf{F}'(\mathbf{b})(\mathbf{s}) + \sigma_2(\mathbf{s}) - \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r}))$$

$$= \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r})) + \sigma_2(\mathbf{s}) - \mathbf{F}'(\mathbf{b})(\mathbf{G}'(\mathbf{a})(\mathbf{r}))$$

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

$$\frac{\|\mathbf{s}\|}{\|\mathbf{r}\|} = \frac{\|\mathbf{G}'(\mathbf{a})(\mathbf{r}) + \sigma_1(\mathbf{r})\|}{\|\mathbf{r}\|} \leq \frac{\|\mathbf{G}'(\mathbf{a})(\mathbf{r})\|}{\|\mathbf{r}\|} + \frac{\|\sigma_1(\mathbf{r})\|}{\|\mathbf{r}\|}$$

Since 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. □

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 $\mathbf{F} : O \to Y$. 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*

$$\mathbf{F}'(\mathbf{a}; \mathbf{r}) = \lim_{t \to 0} \frac{\mathbf{F}(\mathbf{a} + t\mathbf{r}) - \mathbf{F}(\mathbf{a})}{t}$$

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 $\mathbf{F} : O \to Y$ 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. □

Remark: In the literature, the terms *Fréchet differentiability* and *Gâteaux differentiability* are often used to distinguish between two different notions of differentiability, especially when the spaces are infinite dimensional. "Fréchet differentiable" is the same as we have called "differentiable", while "Gâteaux differentiable" means that all directional derivatives exist. We have just proved that Fréchet differentiability implies Gâteaux differentiability, but the opposite implication does not hold as you may know from calculus (see Exercise 11).

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

$$\mathbf{F}'(a) = \lim_{t\to 0} \frac{\mathbf{F}(a+t) - \mathbf{F}(a)}{t} \tag{6.1.5}$$

As

$$\mathbf{F}'(a)(1) = \lim_{t\to 0} \frac{\mathbf{F}(a + t\cdot 1) - \mathbf{F}(a)}{t} = \lim_{t\to 0} \frac{\mathbf{F}(a+t) - \mathbf{F}(a)}{t}$$

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 ruke becomes

$$\mathbf{H}'(a) = \mathbf{F}'(\mathbf{G}(a))(\mathbf{G}'(a))$$

It may be useful to end this section with an example:

Example 1: Let $X = Y = C([0,1], \mathbb{R})$ with the usual supremum norm, $\|y\| = \sup\{|y(s)| : s \in [0,1]\}$. We first consider the map $\mathbf{F} : X \to Y$ given by

$$\mathbf{F}(y)(x) = \int_0^x y(s)\, ds$$

It is easy to check that $\mathbf{F}$ is a bounded, linear map, and by Proposition 6.1.5, $\mathbf{F}'(y)(r) = \mathbf{F}(r)$, i.e.

$$\mathbf{F}'(y)(r)(x) = \mathbf{F}(r)(x) = \int_0^x r(s)\, ds$$

To get a nonlinear example, we may instead consider

$$\mathbf{G}(y)(x) = \int_0^x y(s)^2\, ds$$

In this case, it is not quite obvious what $\mathbf{G}'$ is, and it is then often a good idea to find the directional derivatives first as they are given by simple limits. We get

$$\begin{aligned}\mathbf{G}'(y;r)(x) &= \lim_{t\to 0} \frac{\mathbf{G}(y+tr)(x) - \mathbf{G}(y)(x)}{t} \\ &= \lim_{t\to 0} \frac{\int_0^x (y(s)+tr(s))^2\, ds - \int_0^x y(s)^2\, ds}{t} \\ = \lim_{t\to 0} \int_0^x \left[2y(s)r(s) + tr(s)^2\right]\, ds &= \int_0^x 2y(s)r(s)\, ds\end{aligned}$$

This isn't quite enough, though, as the existence of directional derivatives doesn't guarantee differentiability. We need to check that

$$\sigma(r) = \mathbf{G}(y+r) - \mathbf{G}(y) - \mathbf{G}'(y;r)$$

goes to zero faster than r. A straightforward computations shows that

$$\sigma(r)(x) = \int_0^x r(s)^2\, ds \le \int_0^1 \|r\|\, ds = \|r\|^2$$

which means that $\|\sigma\| \le \|r\|^2$, and hence σ goes to zero faster than r. Thus

$$\mathbf{G}'(y)(r)(x) = \int_0^x 2y(s)r(s)\, ds$$

Exercises for Section 6.1

1. Prove Proposition 6.1.6.

2. Assume that X and Y are two normed spaces. A function $\mathbf{F} : X \to Y$ is called *affine* if there is a linear map $A : X \to Y$ 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 $\mathbf{F}, \mathbf{G} : X \to Y$ 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 linear spaces and that $B : X \to Y$ and $A : Y \to Z$ are linear maps. Show that $C = A \circ B$ is a linear map from $X \to Z$.

5. Let X, Y, Z, V be normed spaces and assume that $\mathbf{H} : X \to Y$, $\mathbf{G} : Y \to Z$, $\mathbf{F} : Z \to V$ are functions such that $\mathbf{H}$ is differentiable at $\mathbf{a}$, $\mathbf{G}$ is differentiable at $\mathbf{b} = \mathbf{H}(\mathbf{a})$ og $\mathbf{F}$ 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. Towards 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 $A : \mathbb{R} \to Y$ 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 $\|A\| = \|y\|$. 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 $J(\mathbf{a})$ is the Jacobi matrix of $\mathbf{F}$ at $\mathbf{a}$. Show that

$$\mathbf{F}'(\mathbf{a})(\mathbf{r}) = J(\mathbf{a})\mathbf{r}$$

   where the expression on the right is the product the matrix $J(\mathbf{a})$ and the column vector $\mathbf{r}$.

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 $\mathbf{F}, \mathbf{G} : X \to Y$ are differentiable at $\mathbf{a} \in X$. Show that the function $h : X \to \mathbb{R}$ 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 $f : X \to \mathbb{R}$ is differentiable at all points $\mathbf{x} \in X$.

a) Assume that $\mathbf{r} : \mathbb{R} \to X$ is differentiable at a point $a \in \mathbb{R}$. Show that the function $h(t) = f(\mathbf{r}(t))$ is differentiable at a and that (using the notation of formula (6.1.5))

$$h'(a) = f'(\mathbf{r}(a))(\mathbf{r}'(a))$$

b) If $\mathbf{a}, \mathbf{b}$ are two points in X, and $\mathbf{r}$ is the parametrized line

$$\mathbf{r}(s) = \mathbf{a} + s(\mathbf{b} - \mathbf{a}), \qquad s \in \mathbb{R}$$

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 the next section.

10. Let X be a normed space and assume that the function $F : X \to \mathbb{R}$ 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, $f : \mathbb{R}^2 \to \mathbb{R}$ 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 that continuity fails, consider what happens along the curve $y = x^2$.)

6.2 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 rôle 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 6.2.1 (Mean Value Theorem) *Let a, b be two real numbers, $a < b$. Assume that Y is a normed space and that $\mathbf{F} : [a, b] \to Y$ and $g : [a, b] \to \mathbb{R}$ are two continuous functions which are differentiable at all points $t \in (a, b)$ with $\|\mathbf{F}'(t)\| \leq g'(t)$. Then*

$$\|\mathbf{F}(b) - \mathbf{F}(a)\| \leq g(b) - g(a)$$

Proof: We shall prove that if $\epsilon > 0$, then

$$\|\mathbf{F}(t) - \mathbf{F}(a)\| \leq g(t) - g(a) + \epsilon + \epsilon(t-a) \tag{6.2.1}$$

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 (6.2.1) 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 $b \neq c$. 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 (6.2.1), and hence does *not* belong to C. This is the contradiction we have been looking for. □

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 work, 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

$$g'_+(t) = \lim_{s \to t^+} \frac{g(s) - g(t)}{s - t}$$

and

$$\mathbf{F}'_+(t) = \lim_{s \to t^+} \frac{\mathbf{F}(s) - \mathbf{F}(t)}{s - t}$$

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 easier to see.

Corollary 6.2.2 *Assume that Y is a normed space and that $\mathbf{F} : [a, b] \to Y$ is a continuous map which is differentiable at all points $t \in (a, b)$ with $\|\mathbf{F}'(t)\| \leq k$. Then*

$$\|\mathbf{F}(b) - \mathbf{F}(a)\| \leq k(b - a)$$

Proof: Use the Mean Value Theorem with $g(t) = kt$. □

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

$$\mathbf{r}(t) = \mathbf{a} + t(\mathbf{b} - \mathbf{a}), \qquad t \in [0, 1]$$

connecting $\mathbf{a}$ and $\mathbf{b}$ also belongs to C, i.e., $\mathbf{r}(t) \in C$ for all $t \in [0, 1]$.

Corollary 6.2.3 *Assume that X, Y are normed spaces and that $\mathbf{F} : O \to Y$ is a function defined on a subset O of X. Assume that C is a convex subset of O and that $\mathbf{F}$ is differentiable at all points in $\mathbf{x} \in C$ with $\|\mathbf{F}'(\mathbf{x})\| \leq K$. 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

$$\|\mathbf{F}(\mathbf{b}) - \mathbf{F}(\mathbf{a})\| = \|\mathbf{H}(1) - \mathbf{H}(0)\| \leq K\|\mathbf{b} - \mathbf{a}\|(1 - 0) = K\|\mathbf{b} - \mathbf{a}\|$$

□

Exercises for Section 6.2

1. In this problem X and Y are two normed spaces and O is an open, convex subset of X.

   a) Assume that $\mathbf{F} : O \to Y$ is differentiable with $\mathbf{F}'(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x} \in O$. Show that $\mathbf{F}$ is constant.

   b) Assume that $\mathbf{G}, \mathbf{H} : O \to Y$ 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 $\mathbf{x} \in O$.

   c) Assume that $\mathbf{F} : O \to Y$ is differentiable and that $\mathbf{F}'$ is constant on O. Show that there exist a bounded, linear map $G : X \to Y$ and a constant $\mathbf{C} \in Y$ such that $\mathbf{F} = G + \mathbf{C}$ on O

2. 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 $\mathbf{F} : [a, b] \to Y$ and $g : [a, b] \to \mathbb{R}$ are two continuous functions. Assume further that except for finitely many points $t_1 < t_2 < \ldots < t_n$, $\mathbf{F}$ 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 $[t_i, t_{i+1}]$.)

3. We shall prove the following theorem (which you might want to compare to Proposition 4.3.5):

   Theorem: *Assume that X is a normed spaces, 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 $\mathbf{F}_n : O \to Y$ such that:*

   (i) 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 G is a function from O to the set $\mathcal{L}(X, Y)$ of bounded, linear maps from X to Y).

   (ii) 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 $\mathbf{F}$ and $\mathbf{F}' = \mathbf{G}$ on O.

   a) 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.

   b) Show that $\{\mathbf{F}_n\}$ converges uniformly to a function $\mathbf{F}$ on O.

   c) 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 $n \in \mathbb{N}$

$$\begin{aligned}\|\mathbf{F}(\mathbf{x}+\mathbf{r}) - \mathbf{F}(\mathbf{x}) - \mathbf{G}(\mathbf{x})(\mathbf{r})\| \leq \|\mathbf{F}(\mathbf{x}+\mathbf{r}) - \mathbf{F}(\mathbf{x}) - (\mathbf{F}_n(\mathbf{x}+\mathbf{r}) - \mathbf{F}_n(\mathbf{x}))\| \\ + \|\mathbf{F}_n(\mathbf{x}+\mathbf{r}) - \mathbf{F}_n(\mathbf{x}) - \mathbf{F}'_n(\mathbf{x})(\mathbf{r})\| \\ + \|\mathbf{F}'_n(\mathbf{x})(\mathbf{r}) - \mathbf{G}(\mathbf{x})(\mathbf{r})\|\end{aligned}$$

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 $m \to \infty$.)

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 $\mathbf{F}' = \mathbf{G}$.

6.3 Partial derivatives

From calculus you remember the notion of a partial derivative: If f is a function of n variables $x_1, x_2, \ldots, x_n$, 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

$$\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$$

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

$$X = X_1 \times X_2 \times \ldots \times X_n$$

of normed spaces $X_1, X_2, \ldots, X_n$, and that the norm on X is the product norm $\|(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)\| = \|\mathbf{x}_1\| + \|\mathbf{x}_2\| + \cdots + \|\mathbf{x}_n\|$ (see Section 5.1). A function $\mathbf{F} : X \to Y$ from X into a normed space Y, will be expressed as

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

If $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ is a point in X, we can define functions $\mathbf{F}^i_{\mathbf{a}} : X_i \to Y$ by

$$\mathbf{F}^i_{\mathbf{a}}(x_i) = \mathbf{F}(\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}, \mathbf{x}_i, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_n)$$

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 $\mathbf{F}^i_{\mathbf{a}}$ is a function from X_i to Y, it's derivative at $\mathbf{a}_i$ (if it exists) is a 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 6.3.1 *If $\mathbf{F}^i_{\mathbf{a}}$ 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*

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}_i}(\mathbf{a}) \quad \textit{or} \quad \mathbf{F}'_{\mathbf{x}_i}(\mathbf{a})$$

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 6.3.2 *Assume that U is an open subset of $X_1 \times X_2 \times \ldots \times X_n$ and that $\mathbf{F} : U \to Y$ is differentiable at $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) \in U$. Then the maps $\mathbf{F}^i_{\mathbf{a}}$ 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 $\hat{\mathbf{r}}_i = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{r}_i, \mathbf{0}, \ldots, \mathbf{0})$. Moreover, for all $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) \in X$,

$$\mathbf{F}'(\mathbf{a})(\mathbf{r}) = \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \ldots + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n)$$

Proof: To show that $\mathbf{F}^i_{\mathbf{a}}$ is differentiable at $\mathbf{a}$ 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

$$\sigma_i(\mathbf{r}_i) = \mathbf{F}^i_{\mathbf{a}}(\mathbf{a}_i + \mathbf{r}_i) - \mathbf{F}^i_{\mathbf{a}}(\mathbf{a}_i) - \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_i)$$

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 $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ in X, we have $\mathbf{r} = \hat{\mathbf{r}}_1 + \hat{\mathbf{r}}_2 + \ldots + \hat{\mathbf{r}}_n$, and since $\mathbf{F}'(\mathbf{a})(\cdot)$ is linear

$$\begin{aligned}\mathbf{F}'(\mathbf{a})(\mathbf{r}) &= \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_1) + \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_2) + \ldots + \mathbf{F}'(\mathbf{a})(\hat{\mathbf{r}}_n) \\ &= \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \ldots + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n)\end{aligned}$$

by what we have already shown. □

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 6.3.3 *Assume that U is an open subset of $X_1 \times X_2 \times \ldots \times X_n$ and that $\mathbf{F} : U \to Y$ is continuous at $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$. 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*

$$\mathbf{F}'(\mathbf{a})(\mathbf{r}) = \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \ldots + \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n)$$

for all $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) \in X$.

Proof: We have to prove that

$$\sigma(\mathbf{r}) = \mathbf{F}(\mathbf{a}+\mathbf{r}) - \mathbf{F}(\mathbf{a}) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) - \ldots - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n)$$

goes to zero faster than $\mathbf{r}$. To simplify notation, let us write $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ for $\mathbf{a}+\mathbf{r}$. Observe that we can write $\mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) - \mathbf{F}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ as a telescoping sum:

$$\begin{aligned}&\mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) - \mathbf{F}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) \\ &\qquad = \mathbf{F}(\mathbf{y}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) - \mathbf{F}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) \\ &\qquad + \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{a}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) \\ &\qquad \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ &\qquad + \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{a}_n)\end{aligned}$$

Hence

$$\begin{aligned}\sigma(\mathbf{r}) &= \mathbf{F}(\mathbf{y}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) - \mathbf{F}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{y}_1 - \mathbf{a}_1) \\ &+ \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{a}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{y}_2 - \mathbf{a}_2) \\ &\quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ &+ \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{y}_n - \mathbf{a}_n)\end{aligned}$$

It suffices to prove that the i-th line of this expression goes to zero faster that $\mathbf{r} = \mathbf{y} - \mathbf{a}$. 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 to replace the difference $\mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{a}_n)$ by $\frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{c}_n)(\mathbf{y}_n - \mathbf{a}_n)$ for some $\mathbf{c}_n$ between $\mathbf{a}_n$ and $\mathbf{y}_n$, and then used the continuity of the partial derivative. In the present, more complicated setting, we have to use the Mean Value Theorem of the previous section instead (or, more precisely, its corollary 6.2.3). To do so, we first introduce a function $\mathbf{G}$ defined by

$$\mathbf{G}(\mathbf{z}_n) = \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{z}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{z}_n - \mathbf{a}_n)$$

for all $\mathbf{z}_n \in X_n$ that are close enough to $\mathbf{a}_n$ for the expression to be defined. Note that

$$\mathbf{G}(\mathbf{y}_n) - \mathbf{G}(\mathbf{a}_n) = \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{y}_n - \mathbf{a}_n)$$

which is the quantity we need to prove goes to zero faster than $\mathbf{y} - \mathbf{a}$.

The derivative of $\mathbf{G}$ is

$$\mathbf{G}'(\mathbf{z}_n) = \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{z}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})$$

and hence by Corollary 6.2.3,

$$\|\mathbf{G}(\mathbf{y}_n) - \mathbf{G}(\mathbf{a}_n)\| \leq K\|\mathbf{y}_n - \mathbf{a}_n\|$$

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

$$\mathbf{G}(\mathbf{y}_n) - \mathbf{G}(\mathbf{a}_n) = \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) - \mathbf{F}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{a}_n) - \frac{\partial \mathbf{F}}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{y}_n - \mathbf{a}_n)$$

goes to zero faster than $\mathbf{y} - \mathbf{a}$, and the theorem follows. □

We shall also take a brief look at the dual situation where $\mathbf{F} : X \to Y_1 \times Y_2 \times \ldots \times Y_m$ is a function *into* a product space. Clearly, $\mathbf{F}$ has components $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_m$ such that

$$\mathbf{F}(\mathbf{x}) = (\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}), \ldots, \mathbf{F}_m(\mathbf{x}))$$

Proposition 6.3.4 *Assume that $X, Y_1, Y_2, \ldots, Y_m$ are normed spaces and that U is an open subset of X. A function $\mathbf{F} : U \to Y_1 \times Y_2 \times \ldots \times Y_m$ 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*

$$\mathbf{F}'(\mathbf{a}) = (\mathbf{F}'_1(\mathbf{a}), \mathbf{F}'_2(\mathbf{a}), \ldots, \mathbf{F}'_m(\mathbf{a}))$$

(where this equation means that $\mathbf{F}'(\mathbf{a})(\mathbf{r}) = (\mathbf{F}'_1(\mathbf{a})(\mathbf{r}), \mathbf{F}'_2(\mathbf{a})(\mathbf{r}), \ldots, \mathbf{F}'_m(\mathbf{a})(\mathbf{r}))$.)

Proof: Clearly,

$$\sigma(\mathbf{r}) = (\sigma_1(\mathbf{r}), \ldots, \sigma_m(\mathbf{r}))$$

$$= (\mathbf{F}_1(\mathbf{a}+\mathbf{r}) - \mathbf{F}_1(\mathbf{a}) - \mathbf{F}'_1(\mathbf{a})(\mathbf{r}), \ldots, \mathbf{F}_m(\mathbf{a}+\mathbf{r}) - \mathbf{F}_m(\mathbf{a}) - \mathbf{F}'_m(\mathbf{a})(\mathbf{r}))$$

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 $\mathbf{r}$. □

If we combine the proposition above with Theorem 6.3.3, we get

Proposition 6.3.5 *Assume that U is an open subset of $X_1 \times X_2 \times \ldots \times X_n$ and that $\mathbf{F} : U \to Y_1 \times Y_2 \times \ldots \times Y_m$ is continuous at $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$. 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*

$$\begin{aligned}\mathbf{F}'(\mathbf{a})(\mathbf{r}) = \Big(& \frac{\partial \mathbf{F}_1}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}_1}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \ldots + \frac{\partial \mathbf{F}_1}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n), \\ & \frac{\partial \mathbf{F}_2}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}_2}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \ldots + \frac{\partial \mathbf{F}_2}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n), \ldots, \\ & \frac{\partial \mathbf{F}_m}{\partial \mathbf{x}_1}(\mathbf{a})(\mathbf{r}_1) + \frac{\partial \mathbf{F}_m}{\partial \mathbf{x}_2}(\mathbf{a})(\mathbf{r}_2) + \ldots + \frac{\partial \mathbf{F}_m}{\partial \mathbf{x}_n}(\mathbf{a})(\mathbf{r}_n) \Big)\end{aligned}$$

for all $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) \in X$.

Exercises for Section 6.3

1. Assume that $f : \mathbb{R}^2 \to \mathbb{R}$ 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. Let X be a normed space and consider to differentiable functions $F, G : X \to \mathbb{R}$. Define the *Lagrange function* $H : X \times \mathbb{R}$ by

$$H(\mathbf{x}, \lambda) = F(\mathbf{x}) + \lambda G(\mathbf{x})$$

   a) Show that

$$\frac{\partial H}{\partial \mathbf{x}}(\mathbf{x}, \lambda) = F'(\mathbf{x}) + \lambda G'(x)$$

$$\frac{\partial H}{\partial \lambda}(\mathbf{x}, \lambda) = G(x)$$

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 $F'(\mathbf{a}) + \lambda_0 G'(\mathbf{a}) = 0$.

3. Let X be a real inner product space and define $F : X \times X \to \mathbb{R}$ 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 $\frac{\partial F}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y})(\mathbf{s})$?

4. Let $G : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be differentiable at the point $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^n \times \mathbb{R}^m$. Show that

$$\frac{\partial G}{\partial \mathbf{x}}(\mathbf{a}, \mathbf{b})(\mathbf{r}) = \left(\frac{\partial G}{\partial x_1}(\mathbf{a}, \mathbf{b}), \frac{\partial G}{\partial x_2}(\mathbf{a}, \mathbf{b}), \ldots, \frac{\partial G}{\partial x_n}(\mathbf{a}, \mathbf{b}) \right) \cdot \mathbf{r}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. What is $\frac{\partial G}{\partial \mathbf{y}}(\mathbf{a}, \mathbf{b})(\mathbf{s})$?

5. 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 $F : A \to \mathbb{R}$ by $F(t, f) = \int_0^t f(s)\, ds$. Show that the partial derivates $\frac{\partial F}{\partial t}(t, f)$ 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 $i_t : C([0, 1], \mathbb{R}) \to \mathbb{R}$ is the map defined by $i_t(g) = \int_0^t g(s)\, ds$.

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 $F : A \to \mathbb{R}$ by $F(t, f) = f(t)$. Show that if f is differentiable at t, then the partial derivates $\frac{\partial F}{\partial t}(t, f)$ and $\frac{\partial F}{\partial f}(t, f)$ exist and that $\frac{\partial F}{\partial t}(t, f) = f'(t)$, $\frac{\partial F}{\partial f}(t, f) = e_t$, where $e_t : C([0, 1], \mathbb{R}) \to \mathbb{R}$ is the evaluation function $e_t(g) = g(t)$.

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 $\mathbf{F} : [a, b] \to X$, 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 $\Pi = \{x_0, x_1, \ldots, x_n\}$ from $[a, b]$ such that

$$x = x_0 < x_1 < x_2 < \ldots < x_n = b$$

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 set of points $S = \{c_1, c_2, \ldots, c_n\}$ such that $x_{i-1} \leq c_i \leq x_i$, i.e., a collection 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 $\mathbf{F}:[a,b]\to X$ be a continuous function from a real interval to a normed space. Assume that $\Pi=\{x_0,x_1,\ldots,x_n\}$ 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 $\|\mathbf{F}(c)-\mathbf{F}(d)\|\le M$. 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*

$$|R(\mathbf{F},\Pi,S)-R(\mathbf{F},\hat{\Pi},\hat{S})|\le M(b-a)$$

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

$$x_{i-1}=y_j<y_{j+1}<\cdots<y_m=x_i$$

The selection S picks a point c_i in the interval $[x_{i-1},x_i]$ and the selection $\hat{S}$ picks point $d_{j+1}\in[y_j,y_{j+1}],d_{j+2}\in[y_{j+1},y_{j+2}],\ldots,d_m\in[y_{m-1},y_m]$. The contributions to the two Riemann sums are

$$\mathbf{F}(c_i)(x_i-x_{i-1})=\mathbf{F}(c_i)(y_{j+1}-y_j)+\mathbf{F}(c_i)(y_{j+2}-y_{j+1})+\cdots+\mathbf{F}(c_i)(y_m-y_{m-1})$$

and

$$\mathbf{F}(d_j)(y_{j+1}-y_j)+\mathbf{F}(d_{j+1})(y_{j+2}-y_{j+1})+\cdots+\mathbf{F}(d_m)(y_m-y_{m-1})$$

By the triangle inequality, the difference between these two expressions are less than

$$\begin{gathered}\|\mathbf{F}(c_i)-\mathbf{F}(d_j)\|(y_{j+1}-y_j)+\|\mathbf{F}(c_i)-\mathbf{F}(d_{j+1})\|(y_{j+2}-y_{j+1})+\\ \cdots+\|\mathbf{F}(c_i)-\mathbf{F}(d_m)\|(y_m-y_{m-1})\\ \le M(y_{j+1}-y_j)+M(y_{j+2}-y_{j+1})+\cdots+M(y_m-y_{m-1})\end{gathered}$$

$$= M(x_i - x_{i-1})$$

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. □

The next lemma brings us closer to the point.

Lemma 6.4.2 *Let $\mathbf{F} : [a,b] \to X$ 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 $M = \frac{\epsilon}{2(b-a)}$. □

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. If X is complete, the sequence must converge 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 $\mathbf{F} : [a,b] \to X$ 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 $|\Pi_n| \to 0$.*

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 $\mathbf{F}, \mathbf{G} : [a,b] \to X$ be continuous functions from a real interval to a complete, normed space. Then*

$$\int_a^b \big(\alpha\mathbf{F}(x) + \beta\mathbf{G}(x)\big)\,dx = \alpha\int_a^b \mathbf{F}(x)\,dx + \beta\int_a^b \mathbf{G}(x)\,dx$$

for all $\alpha, \beta \in \mathbb{R}$.

Proof: Pick sequences $\{\Pi_n\}$, $\{S_n\}$ of partitions and selections such that $|\Pi_n| \to 0$. Then

$$\begin{aligned}\int_a^b \big(\alpha\mathbf{F}(x) + \beta\mathbf{G}(x)\big)\,dx &= \lim_{n\to\infty} R(\alpha\mathbf{F} + \beta\mathbf{G}, \Pi_n, S_n)\\ &= \lim_{n\to\infty} \big(\alpha R(\mathbf{F}, \Pi_n, S_n) + \beta R(\mathbf{G}, \Pi_n, S_n)\big)\\ &= \alpha\lim_{n\to\infty} R(\mathbf{F}, \Pi_n, S_n) + \beta\lim_{n\to\infty} R(\mathbf{G}, \Pi_n, S_n)\\ &= \alpha\int_a^b \mathbf{F}(x)\,dx + \beta\int_a^b \mathbf{G}(x)\,dx\end{aligned}$$

□

Proposition 6.4.5 *Let $\mathbf{F} : [a,b] \to X$ 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. □

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 $\mathbf{F} : [a,b] \to X$ be a continuous function from a real interval to a complete, normed space. Define a function $I : [a,b] \to X$ by*

$$\mathbf{I}(x) = \int_a^x \mathbf{F}(t)\,dt$$

Then $\mathbf{F}$ 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

$$\sigma(r) = \int_x^{x+r} \mathbf{F}(t)\,dt - \mathbf{F}(x)r = \int_x^{x+r} \big(\mathbf{F}(t) - \mathbf{F}(x)\big)\,dt$$

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. □

We shall also need a version of the fundamental theorem that works in the opposite direction.

Corollary 6.4.7 *Let $\mathbf{F} : (a,b) \to X$ be a continuous function from a real interval to a complete, normed space. Assume that $\mathbf{F}$ is differentiable in (a,b) and that $\mathbf{F}'$ is continuous 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 $\mathbf{G} : [c,d] \to X$ by $\mathbf{G}(x) = \mathbf{F}(x) - \int_c^x \mathbf{F}'(t)\,dt$. According to the Fundamental Theorem, $\mathbf{G}'(x) = \mathbf{F}'(x) - \mathbf{F}'(x) = \mathbf{0}$ for all $x \in (c,d)$. If we apply the Mean Value Theorem 6.2.1 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. □

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

$$\int_a^b \mathbf{F}(t)\,dt = \int_b^a \mathbf{F}(t)\,dt \tag{6.4.1}$$

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 exercise above).
3. Let X be a complete, normed space. Assume that $\mathbf{F} : \mathbb{R} \to X$ and $g : \mathbb{R} \to \mathbb{R}$ are two functions with continuous derivatives such that $\|\mathbf{F}'(t)\| \leq g'(t)$ for all $t \in [a, b]$, Show that $\|\mathbf{F}(b) - \mathbf{F}(a)\| \leq g(b) - g(a)$.
4. Let X be a complete, normed space. Assume that $\mathbf{F} : \mathbb{R} \to X$ is continuous function. Show that there is a unique, continuous function $\mathbf{G} : [a, b] \to X$ such that $\mathbf{G}(a) = \mathbf{0}$ and $\mathbf{G}'(t) = \mathbf{F}(t)$ for all $t \in (a, b)$.
5. Let X be a complete, normed space. Assume that $\mathbf{F} : \mathbb{R} \to X$ and $g : \mathbb{R} \to \mathbb{R}$ 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 $a > b$).

6.5 Inverse Function Theorem

From single variable calculus, you know that if a function $f : \mathbb{R} \to \mathbb{R}$ 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 $\mathbf{F} : U \to Y$ 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 6.5.1 (Inverse Function Theorem) *Assume that X and Y are complete normed spaces, that U is an open subset of X, and that $\mathbf{F} : U \to Y$ is a differentiable function. If $\mathbf{F}'$ is continuous at a point $\mathbf{a} \in U$ where $\mathbf{F}'(\mathbf{a})$ is invertible, then $\mathbf{F}$ has a local inverse at $\mathbf{a}$. This inverse $\mathbf{G}$ 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 (5.6.5), the inverse of a bijective linear map 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

$$\mathbf{F}(\mathbf{x}) = \mathbf{y} \tag{6.5.1}$$

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 (6.5.1):

$$\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{a}) = \mathbf{y} - \mathbf{b}$$

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

$$\mathbf{x}' + \mathbf{H}(\mathbf{x}') = \mathbf{z}\,, \tag{6.5.2}$$

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 6.5.2 (Perturbation Lemma) *Assume that X is a Banach space (a complete normed space). Let $\overline{B}(\mathbf{0}, r)$ be a closed ball around the origin in X, and assume that the function $\mathbf{H} : \overline{B}(\mathbf{0}, r) \to X$ is such that $\mathbf{H}(\mathbf{0}) = \mathbf{0}$ and*

$$\|\mathbf{H}(\mathbf{u}) - \mathbf{H}(\mathbf{v})\| \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in \overline{B}(\mathbf{0}, r)$$

Then the function $\mathbf{L} : \overline{B}(\mathbf{0}, r) \to X$ defined by $\mathbf{L}(\mathbf{x}) = \mathbf{x} + \mathbf{H}(\mathbf{x})$ is injective, and the ball $\overline{B}(\mathbf{0}, \frac{r}{2})$ is contained in the image $\mathbf{L}(\overline{B}(\mathbf{0}, r))$.

Proof: To show that $\mathbf{L}$ is injective, we assume that $\mathbf{L}(\mathbf{x}) = \mathbf{L}(\mathbf{y})$ and need to prove that $\mathbf{x} = \mathbf{y}$. By definition of $\mathbf{L}$,

$$\mathbf{x} + \mathbf{H}(\mathbf{x}) = \mathbf{y} + \mathbf{H}(\mathbf{y}),$$

that is

$$\mathbf{x} - \mathbf{y} = \mathbf{H}(\mathbf{y}) - \mathbf{H}(\mathbf{x}),$$

which gives us

$$\|\mathbf{x} - \mathbf{y}\| = \|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y})\|$$

According to the assumptions, $\|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y})\| \leq \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|$, and thus the equality above is only possible if $\|\mathbf{x} - \mathbf{y}\| = 0$, i.e. if $\mathbf{x} = \mathbf{y}$.

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 $\overline{B}(\mathbf{0}, r)$. □

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 $I : X \to X$ is the identity map $I(\mathbf{x}) = \mathbf{x}$. However, the arguments that brought us from formula (6.5.1) to (6.2.2) will later help us convert this very special case to the general.

Lemma 6.5.3 *Let X be a Banach space. Assume that U is an open set in X containing $\mathbf{0}$, and that $\mathbf{L} : U \to X$ 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 6.2.3, this means that

$$\|\mathbf{H}(\mathbf{u}) - \mathbf{H}(\mathbf{v})\| \leq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\| \qquad \text{for all } \mathbf{u}, \mathbf{v} \in \overline{B}(\mathbf{0}, r)$$

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 $\mathbf{y} \to 0$, we only have to consider $\mathbf{y} \in \overline{B}(\mathbf{0}, \frac{r}{2})$. For each such $\mathbf{y}$, we know there is a unique $\mathbf{x}$ in $\overline{B}(\mathbf{0}, r)$ 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

$$\sigma_M(\mathbf{y}) = \mathbf{M}(\mathbf{y}) - \mathbf{y} = \mathbf{x} - \mathbf{L}\big(\mathbf{x}\big) = -\big(\mathbf{L}(\mathbf{x}) - \mathbf{L}(\mathbf{0}) - I(\mathbf{x})\big) = -\sigma_L(\mathbf{x})$$

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

$$\lim_{\mathbf{y}\to\mathbf{0}} \frac{\|\sigma_M(\mathbf{y})\|}{\|\mathbf{y}\|} = \lim_{\mathbf{x}\to\mathbf{0}} \frac{\|\sigma_L(\mathbf{x})\|}{\|\mathbf{x}\|} \cdot \frac{\|\mathbf{x}\|}{\|\mathbf{y}\|} = 0$$

since $\lim_{\mathbf{x}\to\mathbf{0}} \frac{\|\sigma_L(\mathbf{x})\|}{\|\mathbf{x}\|} = 0$ 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 (6.5.1) into (6.5.2).

We begin by defining $\mathbf{L}$ by

$$\mathbf{L}(\mathbf{z}) = A\left(\mathbf{F}(\mathbf{z} + \mathbf{a}) - \mathbf{b}\right)$$

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

$$\mathbf{L}(\mathbf{0}) = A\left(\mathbf{F}(\mathbf{a}) - \mathbf{b}\right) = \mathbf{0}$$

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 $\overline{B}(\mathbf{0}, r)$ 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 $\mathbf{L}(\mathbf{z}) = A\left(\mathbf{F}(\mathbf{z} + \mathbf{a}) - \mathbf{b}\right)$, we get

$$\mathbf{F}(\mathbf{x}) = A^{-1}\mathbf{L}(\mathbf{x} - \mathbf{a}) + \mathbf{b}$$

for alle $\mathbf{x} \in \overline{B}(\mathbf{a}, r)$. 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 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 $\mathbf{F}'(\bar{\mathbf{a}})^{-1}$ by definition. □

Many applications of the Inverse Function Theorem are to functions $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$. Since 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 $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ 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 6.5

1. Show that the function $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ 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 Jacobi matrix of the function $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ 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. Recall from linear algebra (or prove!) that a linear map $A : \mathbb{R}^n \to \mathbb{R}^m$ can only be invertible if $n = m$. Show that a differentiable function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ can only have a differentiable, local inverse if $n = m$.

4. Let X, Y be two complete normed spaces and assume that $O \subseteq X$ is open. Show that if $\mathbf{F} : O \to Y$ 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.

5. 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 $\mathbf{P}_n : \mathcal{M}_n \to \mathcal{M}_n$ by $\mathbf{P}_n(A) = A^n$. Show that 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 $\exp : \mathcal{M}_n \to \mathcal{M}_n$ 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?

6. Let X, Y be two complete normed spaces, and let $\mathcal{L}(X,Y)$ be the space of all continuous, linear maps $A : X \to Y$. 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 $\mathbf{F} : \mathcal{L}(X,Y) \to \mathcal{L}(X,Y)$ 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).

7. Define $f : \mathbb{R} \to \mathbb{R}$ 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?

6.6 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

$$\mathbf{F}(\mathbf{x}, \mathbf{G}(\mathbf{x})) = \mathbf{0} \tag{6.6.1}$$

Even when we cannot solve the equation explicitly, it would be helpful to know that there exists a function $\mathbf{G}$ satisfying equation (6. 6.1) – especially if we also got to know a few of its properties. The Inverse Function Theorem 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 6.6.1 (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 $\mathbf{F} : U \to Z$ 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 $\mathbf{G} : V \to Y$ 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*

$$\mathbf{G}'(\mathbf{x}) = -\left(\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{G}(x))\right)^{-1} \circ \frac{\partial \mathbf{F}}{\partial \mathbf{x}}(\mathbf{x}, \mathbf{G}(x)) \tag{6.6.2}$$

for all $\mathbf{x} \in V$.

Proof: Define a function $\mathbf{H} : U \to X \times Z$ 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, we first have to check that $\mathbf{H}'(\mathbf{a})$ is a bijection. According to Proposition 6.3.5, 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 5). Hence $\mathbf{H}$ satisfies the conditions of the Inverse Function Theorem, 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 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 $(\mathbf{x}, \mathbf{G}(\mathbf{x}))$. □

The point $\mathbf{G}(\mathbf{x})$ in the implicit function theorem is "locally unique" in the following sense.

Corollary 6.6.2 *Let the setting be as in the Implicit Function Theorem. 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

$$(\mathbf{x}, \mathbf{y}) = \mathbf{K}(\mathbf{H}(\mathbf{x}, \mathbf{y})) = \mathbf{K}(\mathbf{x}, \mathbf{F}(\mathbf{x}, \mathbf{y})) = (\mathbf{x}, \mathbf{L}(\mathbf{F}(\mathbf{x}, \mathbf{y})))$$

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

$$(\mathbf{x}, \mathbf{y}_1) = (\mathbf{x}, \mathbf{L}(\mathbf{F}(\mathbf{x}, \mathbf{y}_1))) = (\mathbf{x}, \mathbf{L}(\mathbf{0})) = (\mathbf{x}, \mathbf{L}(\mathbf{F}(\mathbf{x}, \mathbf{y}_2))) = (\mathbf{x}, \mathbf{y}_2)$$

and thus $\mathbf{y}_1 = \mathbf{y}_2$. □

Remark: We cannot expect more than local existence and local uniqueness for 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 sufficintly near $\sin b$. ♣

Exercises for Section 6.6

1. Work through the example in the remark above.
2. Let $f : \mathbb{R}^3 \to \mathbb{R}$ 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 $g(-1, 2) = 0$ og $f(x, y, g(x, y)) = -4$. 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. 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 $\frac{\partial \phi}{\partial y}(x, y(x)) \neq 0$.

5. Show that $\mathbf{H}'(\mathbf{a},\mathbf{b})$ in the proof of Theorem 6.6.1 is a bijection from $X \times Y$ to $X \times Z$.

6. 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 an equation $\phi(x(t), y(t), z(t)) = 0$. Find an expression for $x'(t)$ in this case.

7. 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

   $$\phi(X(y,z),y,z) = 0 \qquad \phi(x,Y(x,z),z) = 0 \qquad \text{and} \qquad \phi(x,y,Z(x,y)) = 0$$

   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.

8. Deduce the Inverse Function Theorem from the Implicit Function Theorem by applying the latter to the function $\mathbf{H}(\mathbf{x},\mathbf{y}) = \mathbf{x} - \mathbf{F}(\mathbf{y})$.

9. (Lagrange multipliers) Let X, Y, Z be complete normed spaces and assume that $f : X \times Y \to \mathbb{R}$ and $\mathbf{F} : X \times Y \to Z$ 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

   $$A = \{(\mathbf{x},\mathbf{y}) \,|\, \mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{0}\}$$

   We assume that $f(\mathbf{x},\mathbf{y})$ has a local maximun (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 $\frac{\partial f}{\partial\mathbf{x}}(\mathbf{x}_0,\mathbf{y}_0) + \frac{\partial f}{\partial y}(\mathbf{x}_0,\mathbf{y}_0)(\mathbf{G}'(\mathbf{x}_0)) = 0$.

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.

6.7 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 compactnesss 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.

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 $\mathbf{F} : [a, b] \to X$ with the norm

$$\|\mathbf{F}\|_0 = \sup\{\|\mathbf{F}(t)\| \ : \ t \in [a, b]\}$$

(the reason why we 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.

Let $\mathbf{F} : [a, b] \to X$ where X is a normed space. If $t \in (a, b)$ is an interior point of $[a, b]$, we have already introduced the notation

$$\mathbf{F}'(t) = \lim_{r \to 0} \frac{\mathbf{F}(t + r) - \mathbf{F}(t)}{r}$$

and we now extend it to the end points by using one-sided derivatives:

$$\mathbf{F}'(a) = \lim_{r \to 0^+} \frac{\mathbf{F}(a+r) - \mathbf{F}(a)}{r}$$

$$\mathbf{F}'(b) = \lim_{r \to 0^-} \frac{\mathbf{F}(b+r) - \mathbf{F}(b)}{r}$$

We are now ready to define the new spaces we shall be working with in this section.

Definition 6.7.1 *A function $\mathbf{F} : [a,b] \to X$ from an interval to a normed space is* continuously differentiable *if the function $\mathbf{F}'$ is defined and continuous on all of $[a,b]$. The set of all continuously differentiable functions is denoted by $C^1([a,b],X)$, and the norm on this space is defined by*

$$\|\mathbf{F}\|_1 = \|\mathbf{F}\|_0 + \|\mathbf{F}'\|_0$$

$$= \sup\{\|\mathbf{F}(x)\| \,:\, x \in [a,b]\} + \sup\{\|\mathbf{F}'(x)\| \,:\, x \in [a,b]\}$$

Remark: A word on notation may 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.[2] 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 $f_n : [0,2\pi] \to \mathbb{R}$ 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

[2] The system become even clearer if one writes $C^0([a,b],X)$ for $C([a,b],X)$, as is often done

$C^1([0,2\pi],\mathbb{R})$. 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 f_n gets closer and closer to the constant function 0, the derivative f_n' does not approach the derivative of 0. The point is that 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 6.7.2 *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 $\mathbf{G}:[a,b]\to X$. 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

$$\mathbf{F}(x) = \mathbf{y} + \int_a^x \mathbf{G}(t)\,dt \tag{6.7.1}$$

Note that by the Fundamental Theorem of Calculus in Section 6.4, $\mathbf{F}' = \mathbf{G}$, 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 (6.7.1) above, we get

$$\begin{aligned}
\|\mathbf{F}(x)-\mathbf{F}_n(x)\| &= \|\mathbf{y}-\mathbf{F}_n(a) + \int_a^x \big(\mathbf{G}(t)-\mathbf{F}_n'(t)\big)\,dt\| \\
&\leq \|\mathbf{y}-\mathbf{F}_n(a)\| + \|\int_a^x \big(\mathbf{G}(t)-\mathbf{F}_n'(t)\big)\,dt\| \\
&\leq \|\mathbf{y}-\mathbf{F}_n(a)\| + \int_a^x \|\mathbf{G}-\mathbf{F}_n'\|_0\,dt \\
&\leq \|\mathbf{y}-\mathbf{F}_n(a)\| + \|\mathbf{G}-\mathbf{F}_n'\|_0(b-a)
\end{aligned}$$

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 that 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 to normed spaces U and V is a bounded, bijective, linear map $T : U \to V$ whose inverse is also bounded. In this terminology, the conditions of the Implicit Function Theorem requires that $\frac{\partial \mathbf{F}}{\partial \mathbf{y}}(\mathbf{x},\mathbf{y})$ is an isomorphism.

If $c \in [a,b]$, the space

$$C^1_c([a,b],X) = \{\mathbf{F} \in C^1([a,b],X) \ : \ \mathbf{F}(c) = \mathbf{0}\}$$

consists of those functions in $C^1([a,b],X)$ that have value zero at c. As $C^1_c([a,b],X)$ is a closed subset of the complete space $C^1([a,b],X)$, it is itself a complete space.

Proposition 6.7.3 *Let X be a complete, normed space, and define $D : C^1_c([a,b],X) \to C([a,b],X)$ 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 $\mathbf{F} \in C^1_c([a,b],X)$ and – by the Fundamental Theorem of Calculus – $D\mathbf{F} = \mathbf{F}' = \mathbf{G}$.

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 6.5.7, 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 $C^1_c([a,b],X)$ and $C([a,b],X)$ are complete, it now follows from the Bounded Inverse Theorem 5.6.5 that D^{-1} is bounded, and hence D is an isometry. □

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 $\lim_{\mathbf{r}\to\mathbf{0}} \frac{\sigma(\mathbf{r},\mathbf{x})}{\|\mathbf{r}\|} = \mathbf{0}$ *uniformly* in $\mathbf{x}$? This is not necessarily the case, but the next lemma gives us the positive information we shall need.

Lemma 6.7.4 *Let X, Y be two normed spaces and let $\mathbf{F}: X \to Y$ be a continuously differentiable function. Assume that $\mathbf{G}: [a,b] \to X$ 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

$$\lim_{n\to\infty} \frac{\|\sigma_{\mathbf{F}}(\mathbf{r}_n,t_n)\|}{\|\mathbf{r}_n\|} = 0$$

Proof: We shall apply the Mean Value Theorem (or, more precisely, its Corollary 6.2.2) to the function

$$\mathbf{H}(s) = \mathbf{F}(\mathbf{G}(t_n)+s\mathbf{r}_n) - \mathbf{F}(\mathbf{G}(t_n)) - s\mathbf{F}'(\mathbf{G}(t_n))(\mathbf{r}_n)$$

where $s \in [0,1]$ (note that $\sigma_{\mathbf{F}}(\mathbf{r}_n,t_n) = \mathbf{H}(1) = \mathbf{H}(1) - \mathbf{H}(0)$). Differentiating, we get

$$\mathbf{H}'(s) = \mathbf{F}'(\mathbf{G}(t_n)+s\mathbf{r}_n)(\mathbf{r}_n) - \mathbf{F}'(\mathbf{G}(t_n))(\mathbf{r}_n)$$

and hence

$$\|\mathbf{H}'(s)\| \le \|\mathbf{F}'(\mathbf{G}(t_n)+s\mathbf{r}_n) - \mathbf{F}'(\mathbf{G}(t_n))\|\|\mathbf{r}_n\|$$

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

$$\mathbf{H}'(s) \le \epsilon\|\mathbf{r}_n\|$$

for all such n. Applying Corollary 6.2.2, we now get

$$\|\sigma_{\mathbf{F}}(\mathbf{r}_n, t_n)\| = \|\mathbf{H}(1) - \mathbf{H}(0)\| \leq \epsilon \|\mathbf{r}_n\|$$

and the lemma is proved. □

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 $\mathbf{F} : X \to Y$ between the underlying spaces. If we now have a continuous function $\mathbf{G} : [a,b] \to X$, we can change it to a continuous function $K : [a,b] \to Y$ 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

$$\Omega_{\mathbf{F}} : C([a,b],X) \to C([a,b],Y)$$

(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 6.7.5 (Omega Rule) *Let X, Y be two normed spaces and let U be an open subset of X. Assume that $\mathbf{F} : U \to Y$ is a continuously differentiable function (i.e. $\mathbf{F}'$ is defined and continuous in all of U). Define a function $\Omega_{\mathbf{F}} : C([a,b],U) \to C([a,b],Y)$ 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 $C([a,b],Y)$, we can evaluate it at a point $\mathbf{G} \in C([a,b],U)$. Now $\Omega'_{\mathbf{F}}(\mathbf{G})$ is a linear map from $C([a,b],U)$ to $C([a,b],Y)$, and

we can evaluate it at a point $\mathbf{H} \in C([a,b],U)$ to get $\Omega'_{\mathbf{F}}(\mathbf{G})(\mathbf{H}) \in C([a,b],Y)$. 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

$$\lim_{\mathbf{H}\to 0} \frac{\|\sigma_\Omega(\mathbf{H})\|_0}{\|\mathbf{H}\|_0} = \lim_{\mathbf{H}\to 0} \frac{\sup_{t\in[a,b]} \|\sigma_\Omega(\mathbf{H}(t))\|}{\|\mathbf{H}\|_0} = 0$$

Since $\mathbf{F}$ is differentiable, we know that for each $t \in [a,b]$,

$$\sigma_{\mathbf{F}}(\mathbf{r},t) = \mathbf{F}(\mathbf{G}(t) + \mathbf{r}) - \mathbf{F}(\mathbf{G}(t)) - \mathbf{F}'(\mathbf{G}(t))(\mathbf{r}) \tag{6.7.2}$$

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

$$\lim_{\mathbf{H}\to 0} \frac{\sup_{t\in[a,b]} \|\sigma_{\mathbf{F}}(\mathbf{H}(t),t)\|}{\|\mathbf{H}\|_0} = 0 \tag{6.7.3}$$

Assume not, then there must be an $\epsilon > 0$ and sequences $\{\mathbf{H}_n\}$, $\{t_n\}$ such that $\mathbf{H}_n \to \mathbf{0}$ and

$$\frac{\|\sigma_{\mathbf{F}}(\mathbf{H}_n(t_n),t_n)\|}{\|\mathbf{H}_n\|_0} > \epsilon$$

for all n. As $\|\mathbf{H}_n(t)\| \le \|\mathbf{H}_n\|_0$ 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

$$\lim_{k\to\infty} \frac{\|\sigma_{\mathbf{F}}(\mathbf{H}_{n_k}(t_{n_k}),t_{n_k})\|}{\|\mathbf{H}_{n_k}(t_{n_k})\|} = 0$$

This contradicts the assumption above, and the theorem is proved. □

The Omega Rule still holds when we replace $C([a,b],U)$ by $C^1([a,b],U)$:

Corollary 6.7.6 *Let X, Y be two normed spaces and let U be an open subset of X. Assume that $\mathbf{F} : U \to Y$ is a continuously differentiable function. Define a function $\Omega_{\mathbf{F}} : C^1([a,b], U) \to C([a,b], Y)$ 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.

$$\lim_{\|\mathbf{H}\|_0 \to 0} \frac{\sigma_\Omega(\mathbf{H})}{\|\mathbf{H}\|_0} = \mathbf{0}$$

We need to prove the corresponding statement for $C^1([a,b], U)$; i.e.,

$$\lim_{\|\mathbf{H}\|_1 \to 0} \frac{\sigma_\Omega(\mathbf{H})}{\|\mathbf{H}\|_1} = \mathbf{0}$$

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

$$\lim_{\|\mathbf{H}\|_1 \to 0} \frac{\sigma_\Omega(\mathbf{H})}{\|\mathbf{H}\|_0} = \mathbf{0} \quad \text{since} \quad \lim_{\|\mathbf{H}\|_0 \to 0} \frac{\sigma_\Omega(\mathbf{H})}{\|\mathbf{H}\|_0} = \mathbf{0}$$

As $\|\mathbf{H}\|_1 \geq \|\mathbf{H}\|_0$, this implies that

$$\lim_{\|\mathbf{H}\|_1 \to 0} \frac{\sigma_\Omega(\mathbf{H})}{\|\mathbf{H}\|_1} = \mathbf{0}$$

and the corollary is proved. □

We are finally ready to take a look at differential equations. If X is a Banach space, O is an open subset of X, and $\mathbf{H} : \mathbb{R} \times O \to X$ is a continuously differentiable function, we shall consider equations of the form

$$\mathbf{y}'(t) = \mathbf{H}(t, \mathbf{y}(t)) \quad \text{where } \mathbf{y}(0) = \mathbf{x} \in O \tag{6.7.4}$$

Our primary goal is to prove the existence of local solutions defined on a small intervall $[-a, a]$, 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 $\mathbf{F} : O \times C^1_0([-1,1], O) \to C([-1,1], X)$ 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 $\mathbf{z} \in C^1_0([-1,1], O)$ satisfies the equation

$$\mathbf{F}(\mathbf{x}, \mathbf{z}) = \mathbf{0} \tag{6.7.5}$$

then $\mathbf{y}(t) = \mathbf{x} + \mathbf{z}(t)$ is a solution to equation (6.7.4) (note that since $\mathbf{z} \in C^1_0([-1,1], O)$, $\mathbf{z}(0) = \mathbf{0}$). The idea is to use the Implicit Function Theorem to prove that for all $\mathbf{x} \in O$ and all sufficiently small t, equation (6.7.5) 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 $\mathbf{z}_0 \in C^1_0([-1,1], O)$ 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

$$\mathbf{u}'(t) = a\mathbf{H}(at, \mathbf{u}(t)), \qquad \mathbf{u}(0) = \mathbf{x} \in O \tag{6.7.6}$$

where $a \in \mathbb{R}$, and assume for the time being that $a \neq 0$. Note that if $\mathbf{y}$ is a solution of (6.7.4), then $\mathbf{u}(t) = \mathbf{y}(at)$ is a solution of (6.7.6), and if $\mathbf{u}$ is a solution of (6.7.6), then $\mathbf{y}(t) = \mathbf{u}(\frac{t}{a})$ is a solution of (6.7.4). Hence to solve (6.7.4) locally, it suffices to solve (6.7.6) for some $a \neq 0$. The point is that the "uninteresting" point $a = 0$ 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

$$\mathbf{F} : \mathbb{R} \times U \times C^1_0([-1,1], O) \to C([-1,1], X)$$

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 Proposistion 6.7.3, 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 6.7.6), we get

$$\frac{\partial}{\partial \mathbf{z}}(a\mathbf{H}(at, \mathbf{x} + \mathbf{z}(t))) = a\frac{\partial \mathbf{H}}{\partial \mathbf{y}}(a, \mathbf{x} + \mathbf{z}(t))$$

(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

$$\frac{\partial}{\partial \mathbf{z}}\mathbf{F}(a, \mathbf{x}, \mathbf{z}) = D - a\frac{\partial \mathbf{H}}{\partial \mathbf{y}}(a, \mathbf{x} + \mathbf{z}(t))$$

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 6.7.3, the conditions of the Implicit Function Theorem 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 $\mathbf{G} : U \to C_0^1([-1, 1], O)$ such that

$$\mathbf{F}(a, \mathbf{x}, \mathbf{G}(a, \mathbf{x})) = \mathbf{0} \quad \text{for all } (a, \mathbf{x}) \in U \tag{6.7.7}$$

i.e.

$$\mathbf{G}'(a, \mathbf{x})(t) = a\mathbf{H}(at, \mathbf{x} + \mathbf{G}(a, \mathbf{x})(t)) \tag{6.7.8}$$

Choose a and r so close to 0 that U contains all points $(a, \mathbf{x})$ where $\mathbf{x} \in \mathrm{B}(\mathbf{x}_0, r)$. For each $\mathbf{x} \in \mathrm{B}(\mathbf{x}_0, r)$, we define a function $\mathbf{y}_\mathbf{x} : [-a, a] \to O$ by

$$\mathbf{y}_\mathbf{x}(t) = \mathbf{x} + \mathbf{G}(a, \mathbf{x})(t/a)$$

Differentiating and using (6.7.8), we get

$$\mathbf{y}'_\mathbf{x}(t) = \mathbf{G}'(a, \mathbf{x})(t/a) \cdot \frac{1}{a} = a\,\mathbf{H}(a(t/a), \mathbf{x} + \mathbf{G}(a, \mathbf{x})(t/a)) \cdot \frac{1}{a} = \mathbf{H}(t, \mathbf{y}_x(t))$$

Hence $\mathbf{y}_x$ is a solution of (6.7.4) on the interval $[-a, a]$.

It's time to stop and sum up the situation:

Theorem 6.7.7 *Let X be a complete normed space and O an open subset of X. Assume that $\mathbf{H} : \mathbb{R} \times O \to X$ 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 $\mathrm{B}(\mathbf{x}_0, r) \subseteq O$ and an interval $[-a, a]$ such that for each $\mathbf{x} \in \mathrm{B}(\mathbf{x}_0, r)$, 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 $\mathrm{B}(\mathbf{x}_0, r)$ to $C^1([-a, a], X)$.

Proof: If we choose an initial value $\mathbf{x}_0$, the argument above not only gives us a solution for this initial value, but for all initial values $\mathbf{x}$ in a ball $\mathrm{B}(\mathbf{x}_0, r)$ 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, $\mathbf{y}_x$ 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 inital 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 $\mathbf{z}_1, \mathbf{z}_2 : [-1, 1] \to U$ 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 $\mathbf{z}_1, \mathbf{z}_2 \in C^1_0([-1, 1], U)$ 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 6.6.2.

This proves uniqueness for a short interval $[-a, a]$, but could the two solutions split later? Assume that they do and put $t_0 = \inf\{t > a \ : \ \mathbf{y}_1(y) \neq \mathbf{y}_2(t)\}$. By continuity, $\mathbf{y}_1(t_0) = \mathbf{y}_2(t)$, 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 $\mathbf{y}_0 = \mathbf{y}_1(t_0) = \mathbf{y}_2(t)$ 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. □

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 6.7

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 $C^1_c([a, b], X)$ is a closed subspace of $C^1([a, b], X)$ and explain why this means that $C^1_c([a, b], X)$ is complete.
3. Check the claim in the text that if y is a solution of (6.7.4), then $u(t) = y(at)$ is a solution of (6.7.6), and that if u is a solution of (6.7.6) for $a \neq 0$, then $y(t) = u(t/a)$ is a solution of (6.7.4).

4. Define $\mathbf{H} : C([0,1],\mathbb{R}) \to C([0,1],\mathbb{R})$ 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 derivative.

5. Show that
$$I(f)(t) = \int_0^t f(t)\,dt$$
defines a bounded linear map $I : C([0,1],\mathbb{R}) \to C_1([0,1],\mathbb{R})$. What is $\|I\|$?

6. In the setting of Theorem 6.7.7, 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
$$y'(t) = y(t), \qquad \mathbf{y}(0) = x$$
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 $f, g : \mathbb{R} \to \mathbb{R}$ are continuous functions.

   a) Show that the unique solution $y(t,x)$ to the problem
   $$y'(t) + f(t)y(t) = g(t), \qquad y(0) = x$$
   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
$$y'(t) = |y(t)| \qquad y(0) = x$$
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 differentiably on x.

6.8 Multilinear maps

So far we have only 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. Before we turn to higher derivatives, we shall therefore 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 6.8.1 *Assume that $X_1, X_2, \ldots, X_n, Y$ are linear spaces. A function $A : X_1 \times X_2 \times \ldots X_n \to Y$ is* multilinear *if it is linear in each variable in the following sense: For all indices $i \in \{1, 2, \ldots, n\}$ and all elements $\mathbf{r}_1 \in X_1, \ldots, \mathbf{r_i} \in X_i, \ldots, \mathbf{r}_n \in X_n$, we have*

(i) $A(\mathbf{r}_1, \ldots, \alpha\mathbf{r}_i, \ldots, \mathbf{r}_n) = \alpha A(\mathbf{r}_1, \ldots, \mathbf{r}_i, \ldots, \mathbf{r}_n)$ *for all* $\alpha \in \mathbb{K}$.

(ii) $A(\mathbf{r}_1, \ldots, \mathbf{r}_i + \mathbf{s}_i, \ldots, \mathbf{r}_n) = A(\mathbf{r}_1, \ldots, \mathbf{r}_i, \ldots, \mathbf{r}_n) + A(\mathbf{r}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{r}_n)$ *for all* $\mathbf{s}_i \in \mathbf{X}_i$.

A multilinear map $A : X_1 \times X_2 \to Y$ 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 $\mathbf{a}_1 = (a_{11}, a_{12}, \ldots, a_{1n})$, $\mathbf{a}_2 = (a_{21}, a_{22}, \ldots, a_{2n})$, $\ldots$, $\mathbf{a}_n = (a_{n1}, a_{n2}, \ldots, a_{nn})$ be n vectors in $\mathbb{R}^n$, and let A be the matrix having $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ as rows. The function from $\mathbb{R}^n$ to $\mathbb{R}$ defined by $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) \mapsto \det(A)$ 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. $A(\mathbf{r}_1, \ldots, \mathbf{0}, \ldots, \mathbf{r}_n) = \mathbf{0}$. This is because by rule (i) of Definition 6.8.1,

$$A(\mathbf{r}_1, \ldots, \mathbf{0}, \ldots, \mathbf{r}_n) = A(\mathbf{r}_1, \ldots, 0\mathbf{0}, \ldots, \mathbf{r}_n)$$

$$= 0A(\mathbf{r}_1, \ldots, \mathbf{0}, \ldots, \mathbf{r}_n) = \mathbf{0}$$

Our next observation is that

$$A(\alpha_1\mathbf{x}_1, \alpha_2\mathbf{x}_2, \ldots, \alpha_n\mathbf{x}_n) = \alpha_1\alpha_2 \ldots \alpha_n A(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

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, $\ldots X_n$ are normed spaces. If we have nonzero vectors $\mathbf{x}_1 \in X_1$, $\mathbf{x}_2 \in X_2$, $\ldots$, $\mathbf{x}_n \in X_n$, we may rescale them to unit vectors $\mathbf{u}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}$, $\mathbf{u}_2 = \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|}$, $\ldots$, $\mathbf{u}_n = \frac{\mathbf{x}_n}{\|\mathbf{x}_n\|}$, and hence

$$A(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = A(\|\mathbf{x}_1\|\mathbf{u}_1, \|\mathbf{x}_2\|\mathbf{u}_2, \ldots, \|\mathbf{x}_n\|\mathbf{u}_n)$$

$$= \|\mathbf{x}_1\|\|\mathbf{x}_2\| \dots \|\mathbf{x}_n\| A(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)$$

which shows that the size of $A(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ grows like the product of the norms $\|\mathbf{x}_1\|$, $\|\mathbf{x}_2\|$, ..., $\|\mathbf{x}_n\|$. This suggests the following definition:

Definition 6.8.2 *Assume that $X_1, X_2, \dots, X_n, Y$ are normed spaces. A multilinear map $A : X_1 \times X_2 \times \dots X_n \to Y$ 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 $\mathbf{x}_1 \in X_1$, $\mathbf{x}_2 \in X_2$,..., $\mathbf{x}_n \in X_n$.

Just as for linear maps (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 $X_1 \times X_2 \times \dots X_n$).

Proposition 6.8.3 *For a multilinear map $A : X_1 \times X_2 \times \dots X_n \to Y$ 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) $\Longrightarrow$(ii) $\Longrightarrow$(iii) $\Longrightarrow$(i). As (ii) obviously implies (iii), it suffices to prove that (i) $\Longrightarrow$ (ii) and (iii) $\Longrightarrow$ (i).

(i) $\Longrightarrow$(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 $\mathbf{x}_1 \in X_1$, $\mathbf{x}_2 \in X_2$,..., $\mathbf{x}_n \in X_n$, and let $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ be an element in $X = X_1 \times X_2 \times \dots X_n$. 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

$$\begin{aligned}
A(\mathbf{x}) - A(\mathbf{a}) =& A(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) - A(\mathbf{a}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \\
&+ A(\mathbf{a}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) - A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{x}_n) \\
&\quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
&+ A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{x}_n) - A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \\
=& A(\mathbf{x}_1 - \mathbf{a}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \\
&+ A(\mathbf{a}_1, \mathbf{x}_2 - \mathbf{a}_2, \dots, \mathbf{x}_n) \\
&\quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
&+ A(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{x}_n - \mathbf{a}_n)
\end{aligned}$$

by multilinearity, and hence

$$\begin{aligned}\|A(\mathbf{x})-A(\mathbf{a})\| \leq & \|A(\mathbf{x}_1-\mathbf{a}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)\| \\ & +\|A(\mathbf{a}_1,\mathbf{x}_2-\mathbf{a}_2,\ldots,\mathbf{x}_n)\| \\ & \quad\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\ & +\|A(\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{x}_n-\mathbf{a}_n)\| \\ \leq & K\|\mathbf{x}_1-\mathbf{a_1}\|\|\mathbf{x}_2\|\ldots\|\mathbf{x}_n\| \\ & +K\|\mathbf{a}_1\|\|\mathbf{x}_2-\mathbf{a_2}\|\ldots\|\mathbf{x}_n\| \\ & \quad\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\ & +K\|\mathbf{a}_1\|\|\mathbf{a}_2\|\ldots\|\mathbf{x}_n-\mathbf{a_2}\|\end{aligned}$$

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

$$\begin{aligned}\|A(\mathbf{x})-A(\mathbf{a})\| \leq & K(\|\mathbf{a}\|+1)^{n-1}(\|\mathbf{x}_1-\mathbf{a}_1\|+\|\mathbf{x}_2-\mathbf{a}_2\|+\cdots\|\mathbf{x}_n-\mathbf{a}_n\|) \\ \leq & K(\|\mathbf{a}\|+1)^{n-1}\|\mathbf{x}-\mathbf{a}\|\end{aligned}$$

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) $\Longrightarrow$ (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 $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is an arbitrary element in X with nonzero components, define

$$\mathbf{u} = \left(\frac{\delta\mathbf{x}_1}{2n\|\mathbf{x}_1\|}, \frac{\delta\mathbf{x}_2}{2n\|\mathbf{x}_2\|}, \ldots, \frac{\delta\mathbf{x}_n}{2n\|\mathbf{x}_n\|}\right)$$

and note that since

$$\|\mathbf{u}\| = \|\mathbf{u}_1\| + \|\mathbf{u}_2\| + \cdots\|\mathbf{u}_n\| \leq n \cdot \frac{\delta}{2n} = \frac{\delta}{2} < \delta$$

we have $\|A(\mathbf{u})\| < 1$. Hence

$$\begin{aligned}\|A(\mathbf{x})\| &= \|A\left(\frac{2n\|\mathbf{x}_1\|}{\delta}\mathbf{u}_1, \frac{2n\|\mathbf{x}_2\|}{\delta}\mathbf{u}_2, \ldots, \frac{2n\|\mathbf{x}_n\|}{\delta}\mathbf{u}_n\right)\| \\ &= \left(\frac{2n}{\delta}\right)^n \|\mathbf{x}_1\|\|\mathbf{x}_2\|\ldots\|\mathbf{x}_n\|A(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_n) \\ &\leq \left(\frac{2n}{\delta}\right)^n \|\mathbf{x}_1\|\|\mathbf{x}_2\|\ldots\|\mathbf{x}_n\|\end{aligned}$$

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 $A : X_1 \times \ldots \times X_n \to Z$ is

a multilinear map, we are looking for derivatives $A'(\mathbf{a}_1, \ldots, \mathbf{a}_n)(\mathbf{r}_1, \ldots \mathbf{r}_n)$ at a point $(\mathbf{a}_1, \ldots \mathbf{a}_n) \in X_1 \times \ldots \times X_n$ and in the direction of a vector $(\mathbf{r}_1, \ldots, \mathbf{r}_n) \in X_1 \times \ldots \times X_n$.

Proposition 6.8.4 *Assume that $X_1, \ldots, X_n, Z$ are normed vector spaces, and that $A : X_1 \times \ldots \times X_n \to Z$ is a continuous multilinear map. Then A is differentiable and*

$$A'(\mathbf{a}_1, \ldots, \mathbf{a}_n)(\mathbf{r}_1, \ldots, \mathbf{r}_n) \\ = A(\mathbf{a}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{r_n}) + A(\mathbf{a}_1, \ldots, \mathbf{r}_{n-1}, \mathbf{a_n}) + \ldots + A(\mathbf{r}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$$

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

$$A(\mathbf{a}_1 + \mathbf{r}_1, \mathbf{a}_2 + \mathbf{r}_2) - A(\mathbf{a}_1, \mathbf{a}_2) = A(\mathbf{a}_1, \mathbf{a}_2 + \mathbf{r}_2) + A(\mathbf{r}_1, \mathbf{a}_2 + \mathbf{r}_2) - A(\mathbf{a}_1, \mathbf{a}_2)$$

$$= A(\mathbf{a}_1, \mathbf{a}_2) + A(\mathbf{a}_1, \mathbf{r}_2) + A(\mathbf{r}_1, \mathbf{a}_2) + A(\mathbf{r}_1, \mathbf{r}_2) - A(\mathbf{a}_1, \mathbf{a}_2)$$

$$= A(\mathbf{a}_1, \mathbf{r}_2) + A(\mathbf{r}_1, \mathbf{a}_2) + A(\mathbf{r}_1, \mathbf{r}_2),$$

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 $\|\mathbf{r}_1\| + \|\mathbf{r}_2\|$. □

Multilinear maps may be thought of as generalized products, and they give rise to a generalized product rule for derivatives.

Proposition 6.8.5 *Assume that $X, Y_1, \ldots, Y_n, U$ are normed spaces and that O is an open subset of X. Assume further that $\mathbf{F}_1 : O \to Y_2$, $\mathbf{F}_2 : O \to Y_2$, $\ldots$, $\mathbf{F}_n : O \to Y_n$ are differentiable at a point $\mathbf{a} \in O$. If $A : Y_1 \times Y_2 \times \ldots \times Y_n \to U$ is a multilinear map, then the composed function $\mathbf{H}(\mathbf{x}) = A(\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}), \ldots, \mathbf{F}_n(\mathbf{x}))$ is differentiable at $\mathbf{a}$ with*

$$\mathbf{H}'(\mathbf{a})(\mathbf{r}) = A(\mathbf{F}_1(\mathbf{a}), \ldots \mathbf{F}_{n-1}(\mathbf{a}), \mathbf{F}'_n(\mathbf{a})(\mathbf{r}))$$

$$+A(\mathbf{F}_1(\mathbf{a}), \ldots, \mathbf{F}'_{n-1}(\mathbf{a})(\mathbf{r}), \mathbf{F}_{n-1}(\mathbf{a})) + \ldots + A(\mathbf{F'}_1(\mathbf{a})(\mathbf{r}), \mathbf{F}_2(\mathbf{a}), \ldots, \mathbf{F}_n(\mathbf{a}))$$

Proof: Let $\mathbf{K} : X \to Y_1 \times Y_2 \times \ldots \times Y_n$ be defined by

$$\mathbf{K}(\mathbf{x}) = (\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x})), \ldots, \mathbf{F}_n(\mathbf{x}))$$

Then $\mathbf{H}(\mathbf{x}) = A(\mathbf{K}(\mathbf{x}))$, and by the Chain Rule and the proposition above

$$\mathbf{H}'(\mathbf{a})(\mathbf{r}) = A'(\mathbf{K}(\mathbf{a}))(\mathbf{K}'(\mathbf{a})(\mathbf{r}))$$

$$= A(\mathbf{F}_1(\mathbf{a}), \ldots \mathbf{F}_{n-1}(\mathbf{a}), \mathbf{F}'_n(\mathbf{a})(\mathbf{r})) + A(\mathbf{F}_1(\mathbf{a}), \ldots, \mathbf{F}'_{n-1}(\mathbf{a})(\mathbf{r}), \mathbf{F}_{n-1}(\mathbf{a}))$$
$$+ \ldots + A(\mathbf{F}'_1(\mathbf{a})(\mathbf{r}), \mathbf{F}_2(\mathbf{a}), \ldots, \mathbf{F}_n(\mathbf{a}))$$

□

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 and sum the results.

Exercises for Section 6.8

1. Show that the maps in Example 1 really are multilinear.
2. Prove the general case of Proposition 6.8.4.
3. Let X be a normed space and Y an inner product space. Assume that $\mathbf{F}, \mathbf{G} : X \to Y$ 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 $A : X^n \to Y$ is called *alternating* if $A(\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_j, \ldots) = -A(\ldots, \mathbf{a}_j, \ldots, \mathbf{a}_i, \ldots)$ 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, $A : X^n \to Y$ 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. $A(\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_i, \ldots) = \mathbf{0}$.

   c) Show the converse of b): If $B : X^n \to Y$ 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 $i \neq j$,

   $$A(\ldots, \mathbf{a}_i + s\mathbf{a}_j, \ldots, \mathbf{a}_j, \ldots) = A(\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_j, \ldots)$$

   for all s.

e) Show that if $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ are linearly dependent, then

$$A(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \mathbf{0}$$

f) Assume now that X is an n-dimensional vector space and that $\mathbf{v}_1$,$\mathbf{v}_2$,..., $\mathbf{v}_n$ is a basis for X. Let B be another alternating, multilinear map such that

$$A(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) = B(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$$

Show that $B = A$. (*Hint:* Show first that if $i_1, i_2, \ldots, i_n \in \{1, 2, \ldots, n\}$, then $A(\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \ldots, \mathbf{v}_{i_n}) = B(\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \ldots, \mathbf{v}_{i_n})$.)

g) Show that the determinant is the only alternating, multilinear map $\det : \mathbb{R}^n \to \mathbb{R}$ such that $\det(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n) = 1$ (here $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ is the standard basis in $\mathbb{R}^n$.)

6.9 Higher order derivatives

We are now ready to look at higher order derivatives. Just as in one-variable calculus, we obtain these by differentiating over and over again, but the difference is that in our present setting, the higher order derivatives become increasingly complicated objects, and it is important to look at them from the right perspective. But let us begin from the beginning.

If X, Y are two normed spaces, O is an open subset of X, and $F : O \to Y$ is a differentiable function, we know that the derivative $\mathbf{F}'(\mathbf{a})$ at a point $\mathbf{a} \in O$ is a 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 mean that we can think of the derivative as a function $\mathbf{F}' : O \to \mathcal{L}(X, Y)$ 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 6.9.1 *Assume that X, Y are two normed spaces, O is an open subset of X, and $F : O \to Y$ is a differentiable function. If the derivative $\mathbf{F}' : O \to \mathcal{L}(X, Y)$ 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 a 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 derivates and how should we think of them? Since the first derivate 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 $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ in X^n to elements $\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)\ldots(\mathbf{r}_n)$ 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 $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) \mapsto \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)\ldots(\mathbf{r}_n)$ is linear in one variable at the time, and hence $\mathbf{F}^{(n)}$ acts like a multilinear map.

Let us formalize this argument.

Proposition 6.9.2 *Assume that X, Y are two normed spaces, that O is an open subset of X, and that $F : O \to Y$ is an n times differentiable function. Then for each $\mathbf{a} \in O$, the function defined by*

$$(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) \mapsto \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)\ldots(\mathbf{r}_n)$$

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 $X \to \mathcal{L}(X, Y)$ 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. □

Remark: We now have two ways to think of higher order derivates. One is to think of them as linear maps

$$\mathbf{F}^{(n)}(\mathbf{a}) : X \to \mathcal{L}(X \to \mathcal{L}(X, \ldots, \mathcal{L}(X, Y)\ldots)]$$

the other is to think of them as multilinear maps

$$\mathbf{F}^{(n)}(\mathbf{a}) : X^n \to Y$$

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 $\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ instead of $\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)\ldots(\mathbf{r}_n)$ 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 $f : \mathbb{R}^n \to \mathbb{R}$, 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., $\mathbf{r} = (r_1, r_2, \ldots, r_n)$.

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 $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ and $\mathbf{s} = (s_1, s_2, \ldots, s_n)$, 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 $\mathbf{r} = (r_1, r_2, \ldots, r_n)$, $\mathbf{s} = (s_1, s_2, \ldots, s_n)$, and $\mathbf{t} = (t_1, t_2, \ldots, t_n)$. The pattern should now be clear. ♣

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 expressions are equal, but if $\mathbf{F}$ is twice differentiable, they are.

Theorem 6.9.3 *Let X and Y be two normed spaces, and let O be an open subset of X. Assume that $\mathbf{F} : O \to Y$ 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

$$\Lambda(h) = \big[\mathbf{F}(\mathbf{a}+h\mathbf{r}+h\mathbf{s}) - \mathbf{F}(\mathbf{a}+h\mathbf{r})\big] - \big[\mathbf{F}(\mathbf{a}+h\mathbf{s}) - \mathbf{F}(\mathbf{a})\big]$$

$$\approx \mathbf{F}'(\mathbf{a}+h\mathbf{r})(h\mathbf{s}) - \mathbf{F}'(\mathbf{a})(h\mathbf{s}) \approx \mathbf{F}''(\mathbf{a})(h\mathbf{r})(h\mathbf{s}) = h^2\mathbf{F}(\mathbf{a})(\mathbf{r})(\mathbf{s})$$

However, if we arrange the terms differently, we get

$$\Lambda(h) = \big[\mathbf{F}(\mathbf{a}+h\mathbf{r}+h\mathbf{s}) - \mathbf{F}(\mathbf{a}+h\mathbf{s})\big] - \big[\mathbf{F}(\mathbf{a}+h\mathbf{r}) - \mathbf{F}(\mathbf{a})\big]$$

$$\approx \mathbf{F}'(\mathbf{a}+h\mathbf{s})(h\mathbf{r}) - \mathbf{F}'(\mathbf{a})(h\mathbf{r}) \approx \mathbf{F}''(\mathbf{a})(h\mathbf{s})(h\mathbf{r}) = h^2\mathbf{F}(\mathbf{a})(\mathbf{s})(\mathbf{r})$$

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

$$\lim_{h\to 0} \frac{\Lambda(h)}{h^2} = \mathbf{F}''(\mathbf{a})(\mathbf{r})(\mathbf{s})$$

By symmetry, we will then also have $\lim_{h\to 0} \frac{\Lambda(h)}{h^2} = \mathbf{F}''(\mathbf{a})(\mathbf{s})(\mathbf{r})$, and the theorem will be proved.

We begin by observing that since $\mathbf{F}$ is twice differentiable at $\mathbf{a}$,

$$\sigma(\mathbf{u}) = \mathbf{F}'(\mathbf{a}+\mathbf{u}) - \mathbf{F}'(\mathbf{a}) - \mathbf{F}''(\mathbf{a})(\mathbf{u}) \tag{6.9.1}$$

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 h is so small that $|h|(\|\mathbf{r}\| + \|\mathbf{s}\|) < \delta$.

We shall first use formula (6.9.1) with $\mathbf{u} = h\mathbf{s}$. Since all the terms in formula (6.9.1) 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

$$\Lambda(h) - \mathbf{F}''(\mathbf{a})(h\mathbf{s})(h\mathbf{r}) =$$

$$= \big[\mathbf{F}(\mathbf{a}+h\mathbf{r}+h\mathbf{s}) - \mathbf{F}(\mathbf{a}+h\mathbf{r}) - \mathbf{F}'(\mathbf{a}+h\mathbf{s})(h\mathbf{r}) + \mathbf{F}'(\mathbf{a})(h\mathbf{r})\big]$$
$$-\big[\mathbf{F}(\mathbf{a}+h\mathbf{s}) - \mathbf{F}(\mathbf{a})\big] + \sigma(h\mathbf{s})(h\mathbf{r})$$
$$= \mathbf{G}(h) - \mathbf{G}(0) + \sigma(h\mathbf{s})(h\mathbf{r})$$

where

$$\mathbf{G}(t) = \mathbf{F}(\mathbf{a}+t\mathbf{r}+h\mathbf{s}) - \mathbf{F}(\mathbf{a}+t\mathbf{r}) - \mathbf{F}'(\mathbf{a}+h\mathbf{s})(t\mathbf{r}) + \mathbf{F}'(\mathbf{a})(t\mathbf{r})$$

$$= \mathbf{F}(\mathbf{a}+t\mathbf{r}+h\mathbf{s}) - \mathbf{F}(\mathbf{a}+t\mathbf{r}) - t\mathbf{F}'(\mathbf{a}+h\mathbf{s})(\mathbf{r}) + t\mathbf{F}'(\mathbf{a})(\mathbf{r})$$

Hence

$$\|\Lambda(h) - \mathbf{F}''(\mathbf{a})(h\mathbf{s})(h\mathbf{r})\| \leq \|\mathbf{G}(h) - \mathbf{G}(0)\| + \|\sigma(h\mathbf{s})\|\|h\mathbf{r}\| \tag{6.9.2}$$

$$\leq \|\mathbf{G}(h) - \mathbf{G}(0)\| + h^2\epsilon\|\mathbf{r}\|\|\mathbf{s}\|$$

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 6.2.2), we have

$$\|\mathbf{G}(h) - \mathbf{G}(0)\| \leq |h| \sup\{\|\mathbf{G}'(t)\| \,:\, t \text{ lies between 0 and } h\} \tag{6.9.3}$$

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 (6.9.1):

$$\mathbf{F}'(\mathbf{a}+t\mathbf{r}+h\mathbf{s}) = \mathbf{F}'(\mathbf{a}) + \mathbf{F}''(\mathbf{a})(t\mathbf{r}+h\mathbf{s}) + \sigma(t\mathbf{r}+h\mathbf{s})$$
$$\mathbf{F}'(\mathbf{a}+t\mathbf{r}) = \mathbf{F}'(\mathbf{a}) + \mathbf{F}''(\mathbf{a})(t\mathbf{r}) + \sigma(t\mathbf{r})$$
$$\mathbf{F}'(\mathbf{a}+h\mathbf{s}) = \mathbf{F}'(\mathbf{a}) + \mathbf{F}''(\mathbf{a})(h\mathbf{s}) + \sigma(h\mathbf{s})$$

If we substitute these expressions into the formula for $\mathbf{G}'(t)$ and use the linearity of $\mathbf{F}''(\mathbf{a})$, we get

$$\mathbf{G}'(t) = \sigma(t\mathbf{r}+h\mathbf{s})(\mathbf{r}) - \sigma(t\mathbf{r})(\mathbf{r}) - \sigma(\mathbf{s})(\mathbf{r})$$

and hence

$$\begin{aligned}\|\mathbf{G}'(t)\| &\leq \|\mathbf{r}\| \left(\|\sigma(t\mathbf{r}+h\mathbf{s})\| + \|\sigma(t\mathbf{r})\| + \|\sigma(h\mathbf{s})\|\right) \\ &\leq \epsilon\|\mathbf{r}\| \left(\|t\mathbf{r}+h\mathbf{s}\| + \|t\mathbf{r}\| + \|h\mathbf{s}\|\right) \\ &\leq 2|h|\epsilon\|\mathbf{r}\|(\|\mathbf{r}\| + \|\mathbf{s}\|)\end{aligned}$$

since $\|t\mathbf{r}+h\mathbf{s}\|$, $\|t\mathbf{r}\|$ and $\|h\mathbf{s}\|$ are less than δ, and $|h|$ is less than $|t|$. By (6.9.3), this means that

$$\|\mathbf{G}(h) - \mathbf{G}(0)\| \leq 2h^2\epsilon\|\mathbf{r}\|(\|\mathbf{r}\| + \|\mathbf{s}\|)$$

and hence by (6.9.2)

$$\begin{aligned}\|\Lambda(h) - \mathbf{F}''(\mathbf{a})(h\mathbf{s})(h\mathbf{r})\| &\leq 2h^2\epsilon\|\mathbf{r}\|(\|\mathbf{r}\| + \|\mathbf{s}\|) + h^2\epsilon\|\mathbf{r}\|\|\mathbf{s}\| \\ &= h^2\epsilon\left(2\|\mathbf{r}\|^2 + 3\|\mathbf{r}\|\|\mathbf{s}\|\right)\end{aligned}$$

Dividing by h^2, we get

$$\|\frac{\Lambda(h)}{h^2} - \mathbf{F}''(\mathbf{a})(\mathbf{s})(\mathbf{r})\| \leq \epsilon\left(2\|\mathbf{r}\|^2 + 3\|\mathbf{r}\|\|\mathbf{s}\|\right)$$

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 $\lim_{h\to 0}\frac{\Lambda(h)}{h^2} = \mathbf{F}''(\mathbf{a})(\mathbf{r})(\mathbf{s})$. As we have already observed, this is sufficient to prove the theorem. □

The theorem generalizes to higher order derivatives.

Theorem 6.9.4 *Let X and Y be two normed spaces, and let O be an open subset of X. Assume that $\mathbf{F} : O \to Y$ 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 $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$ and $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ are the same elements of X but in different order, then*

$$\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) = \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$$

Proof: According to the previous result, we can always interchange two neighbor elements:

$$\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1, \ldots, \mathbf{r}_i, \mathbf{r}_{i+1}, \ldots, \mathbf{r}_n) = \mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}_1, \ldots, \mathbf{r}_{i+1}, \mathbf{r}_i, \ldots, \mathbf{r}_n)$$

and the result follows by observing that we can obtain any permutation of $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$ 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 concentrating on one variable at the time, always works to give us what we want, although it may not always be the fastest method. □

Remark: For functions $\mathbf{F} : \mathbb{R} \to Y$ (or $\mathbf{F} : \mathbb{C} \to Y$) we have been using the simplified notation $\mathbf{F}'(a)$ for what is really $\mathbf{F}'(a)(1)$. We extend this to higher order derivatives by writing $\mathbf{F}^{(n)}(a)$ for what is formally $\mathbf{F}^{(n)}(a)(1)(1)\dots(1)$. Note that this is in agreement with the intuitive idea that

$$\mathbf{F}^{(n)}(a) = \lim_{t\to 0} \frac{\mathbf{F}^{(n-1)}(a+t) - \mathbf{F}^{(n-1)}(a)}{t}$$

The derivatives $\mathbf{F}^{(n)}(a)$ will figure prominently in the next section.

Exercises for Section 6.9

1. Assume that $f : \mathbb{R}^2 \to \mathbb{R}$ 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

   $$f''(\mathbf{a})(\mathbf{e}_i, \mathbf{e}_j) = \frac{\partial f^2}{\partial x_j \partial x_i}(\mathbf{a})$$

   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

$$\mathbf{F}^{(n)}(a) = \lim_{t\to 0} \frac{\mathbf{F}^{(n-1)}(a+t) - \mathbf{F}^{(n-1)}(a)}{t}$$

in the Remark above.

5. Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable and let $Hf(\mathbf{a})$ be the Hesse matrix at $\mathbf{a}$:

$$Hf(\mathbf{a}) = \begin{pmatrix} \frac{\partial f^2}{\partial x_1^2}(\mathbf{a}) & \frac{\partial f^2}{\partial x_1 \partial x_2}(\mathbf{a}) & \dots & \frac{\partial f^2}{\partial x_1 \partial x_n}(\mathbf{a}) \\ \frac{\partial f^2}{\partial x_2 \partial x_1}(\mathbf{a}) & \frac{\partial f^2}{\partial x_2^2}(\mathbf{a}) & \dots & \frac{\partial f^2}{\partial x_2 \partial x_n}(\mathbf{a}) \\ \vdots & \vdots & \dots & \vdots \\ \frac{\partial f^2}{\partial x_n \partial x_1}(\mathbf{a}) & \frac{\partial f^2}{\partial x_n \partial x_2}(\mathbf{a}) & \dots & \frac{\partial f^2}{\partial x_n^2}(\mathbf{a}) \end{pmatrix}$$

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 $f : \mathbb{R}^2 \to \mathbb{R}$ 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

$$\frac{\partial^2 f}{\partial y \partial x}(0,0) = \lim_{h\to 0} \frac{\frac{\partial f}{\partial x}(0,h) - \frac{\partial f}{\partial x}(0,0)}{h}$$

Show in a similar way that $\frac{\partial^2 f}{\partial x \partial y}(0,0) = 1$.

6.10 Taylor's Formula

We shall end this chapter by taking a look at Taylor's formula. In single variable calculus, this formula says that

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} + R_n f(x;a)$$

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 *Tayloyr polynomials* $\sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}$ become better and better approximations to f.

We shall now generalize Taylor's formula to the setting we have been working with in this chapter. First we shall look at functions $\mathbf{F} : \mathbb{R} \to Y$ defined on the real line, but taking values in a normed space Y, and then we shall generalize one step further to functions $\mathbf{F} : X \to Y$ between two normed spaces.

We start by a simple observation (note that we are writing $\mathbf{F}^{(n)}(a)$ for $\mathbf{F}^{(n)}(a)(1)(1)\dots(1)$ as explained at the end of the previous section):

Lemma 6.10.1 *Let Y be a normed space, and assume that $\mathbf{F} : [0,1] \to Y$ is $n+1$ times continuously differentiable in $[0,1]$. Then*

$$\frac{d}{dt}\left(\sum_{k=0}^{n} \frac{1}{k!}(1-t)^k \mathbf{F}^{(k)}(t)\right) = \frac{1}{n!}(1-t)^n \mathbf{F}^{(n+1)}(t)$$

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)

$$\begin{aligned}&\frac{d}{dt}\left(\sum_{k=0}^{n} \frac{1}{k!}(1-t)^k \mathbf{F}^{(k)}(t)\right)\\ &\quad = \mathbf{F}'(t) + \sum_{k=1}^{n}\left(-\frac{1}{(k-1)!}(1-t)^{k-1}\mathbf{F}^{(k)}(t) + \frac{1}{k!}(1-t)^k \mathbf{F}^{(k+1)}(t)\right)\end{aligned}$$

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 one from the previous line, and that the second term cancels with one from the next line (telescoping sum). All you are left with, is the very last term

$$\frac{1}{n!}(1-t)^n \mathbf{F}^{(n+1)}(t)$$

□

We have now have our first version of Taylor's formula:

Proposition 6.10.2 *Let Y be a normed space, and assume that $\mathbf{F} : [0,1] \to Y$ 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 $\mathbf{G}(t) = \sum_{k=0}^{n} \frac{1}{k!}(1-t)^k \mathbf{F}^{(k)}(t)$. Then

$$\mathbf{G}'(t) = \frac{d}{dt}\left(\sum_{k=0}^{n} \frac{1}{k!}(1-t)^k \mathbf{F}^{(k)}(t)\right) = \frac{1}{n!}(1-t)^n \mathbf{F}^{(n+1)}(t)$$

by the lemma. If we use the Fundamental Theorem of Calculus (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. □

In practice, the following corollary is usually more handy than the proposition above.

Lemma 6.10.3 *Let Y be a normed space, and assume that $\mathbf{F} : [0,1] \to Y$ 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

$$\|\int_0^1 \frac{1}{n!}(1-t)^n \mathbf{F}^{(n+1)}(t)\, dt\| \leq \frac{M}{(n+1)!}$$

Let

$$\mathbf{H}(t) = \int_0^t \frac{1}{n!}(1-t)^n \mathbf{F}^{(n+1)}(t)\, dt$$

and note that

$$\|\mathbf{H}'(t)\| = \|\frac{1}{n!}(1-t)^n \mathbf{F}^{(n+1)}(t)\| \leq \frac{M}{n!}(1-t)^n$$

By the Mean Value Theorem (6.2.1), we get

$$\|\mathbf{H}(1)\| = \|\mathbf{H}(1) - \mathbf{H}(0)\| \leq \int_0^1 \frac{M}{n!}(1-t)^n\, dt = \frac{M}{(n+1)!}$$

□

We are now ready to extend Taylor's formula to functions defined on a normed space X, and to keep the expressions short, we need the following notation. If $\mathbf{h} \in X$, we write $\mathbf{h}^n$ for the element $(\mathbf{h}, \mathbf{h}, \ldots, \mathbf{h}) \in X^n$ which has all components equal to $\mathbf{h}$.

Theorem 6.10.4 (Taylor's Formula) *Let X, Y be normed spaces, and assume that $\mathbf{F} : O \to Y$ is 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*

$$\mathbf{F}(\mathbf{a}+\mathbf{h}) = \sum_{k=0}^{n} \frac{1}{k!}\mathbf{F}^{(k)}(\mathbf{a})(\mathbf{h}^k) + \int_0^1 \frac{(1-t)^n}{n!}\mathbf{F}^{(n+1)}(\mathbf{a}+t\mathbf{h})(\mathbf{h}^{n+1})\,dt$$

Proof: Define a function $\mathbf{G} : [0,1] \to Y$ by

$$\mathbf{G}(t) = \mathbf{F}(\mathbf{a}+t\mathbf{h})$$

and note that by the chain rule, $\mathbf{G}^{(k)}(t) = \mathbf{F}^{(k)}(\mathbf{a}+t\mathbf{h})(\mathbf{h}^k)$ for $k = 1, 2, \ldots, n+1$. Applying Proposition 6.10.2 to $\mathbf{G}$, we get

$$\mathbf{F}(\mathbf{a}+\mathbf{h}) = \mathbf{G}(1) = \sum_{k=0}^{n} \frac{1}{k!}\mathbf{G}^{(k)}(0) + \int_0^1 \frac{1}{n!}(1-t)^n\mathbf{G}^{(n+1)}(t)\,dt$$

$$= \sum_{k=0}^{n} \frac{1}{k!}\mathbf{F}^{(k)}(\mathbf{a})(\mathbf{h}^k) + \int_0^1 \frac{(1-t)^n}{n!}\mathbf{F}^{n+1}(t)(\mathbf{h}^{n+1})\,dt$$

□

Remark: As in the one-dimensional case, we refer to

$$\sum_{k=0}^{n} \frac{1}{k!}\mathbf{F}^{(k)}(\mathbf{a})(\mathbf{h}^k)$$

as *the Taylor polynomial of f of degree n at $\mathbf{a}$*.

Again we have a corollary that is often easier to apply in practice.

Corollary 6.10.5 *Let X, Y be normed spaces, and assume that $\mathbf{F} : O \to Y$ is an $n+1$ times continuously differentiable function defined on an open, convex subset O of X. Assume that $\mathbf{a}, \mathbf{a}+\mathbf{h} \in O$, and that $\|\mathbf{F}^{(n+1)}(\mathbf{a}+t\mathbf{h})\| \leq M$ for all $t \in [0,1]$. Then*

$$\|\mathbf{F}(\mathbf{a}+\mathbf{h}) - \sum_{k=0}^{n} \frac{1}{k!}\mathbf{F}^{(k)}(\mathbf{a})(\mathbf{h}^k)\| \leq \frac{M\|h\|^{n+1}}{(n+1)!}$$

Proof: This result follows from Corollary 6.10.3 the same way the previous result followed from Proposition 6.10.2, using that $\|\mathbf{F}^{(n+1)}(\mathbf{a}+t\mathbf{h})(\mathbf{h}^n)\| \leq \|\mathbf{F}^{(n+1)}(\mathbf{a}+t\mathbf{h})\|\|\mathbf{h}\|^{n+1}$. The details are left to the reader. □

In some ways the version of Taylor's formula we have presented above is deceptively simple as the higher order derivatives $\mathbf{F}^{(k)}$ are actually quite

complicated objects. If we look at a multivariable function $f : \mathbb{R}^n \to \mathbb{R}$, we know from Example 1 in section 6.9 that

$$f'(\mathbf{a})(\mathbf{r}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a}) r_i$$

$$f''(\mathbf{a})(\mathbf{r})(\mathbf{s}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}) r_i s_j$$

$$f'''(\mathbf{a})(\mathbf{r})(\mathbf{s})(\mathbf{t}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial^3 f}{\partial x_k \partial x_j \partial x_i}(\mathbf{a}) r_i s_j t_k$$

In general

$$f^{(k)}(\mathbf{a})(\mathbf{r}_1)(\mathbf{r}_2)\dots(\mathbf{r}_k) = \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\dots\sum_{i_k=1}^{n} \frac{\partial^k f}{\partial x_{i_k}\dots\partial x_{i_2}\partial x_{i_1}}(\mathbf{a}) r_k^{(i_k)} \cdots r_2^{(i_2)} r_1^{(i_1)}$$

where $\mathbf{r}_i = (r_i^{(1)}, r_i^{(2)}, \dots, r_i^{(k)})$. The Taylor polynomials can now be written

$$\sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(\mathbf{a})(\mathbf{h}^k) = \sum_{k=0}^{n} \frac{1}{k!} \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\dots\sum_{i_k=1}^{n} \frac{\partial^k f}{\partial x_{i_k}\dots\partial x_{i_2}\partial x_{i_1}} h_{i_k} \cdot \dots \cdot h_{i_2} h_{i_1}$$

where $\mathbf{h} = (h_1, h_2, \dots, h_k)$. This is the version we normally use for functions of several real variables (but see Exercise 5 below for a more efficient way of organizing the terms).

In the results above, we have assumed that $\mathbf{F}$ is $n+1$ times differentiable although we are only interested in the Taylor polynomial of order n. This has the advantage of giving us good estimates for the error in terms of the $(n+1)$-st derivative, but for theoretical purposes it is interesting to see what can be obtained if we only have n derivatives.

Theorem 6.10.6 *Let X, Y be normed spaces and let O be an open subset of X. Assume that $\mathbf{F} : O \to Y$ is n times differentiable at a point $\mathbf{a} \in O$. Then*

$$\|\mathbf{F}(\mathbf{a}+\mathbf{h}) - \sum_{k=0}^{n} \frac{1}{k!} \mathbf{F}^{(k)}(\mathbf{a})(\mathbf{h}^k)\|$$

goes to zero faster than $\|\mathbf{h}\|^n$ as $\mathbf{h}$ goes to zero, i.e.

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{F}(\mathbf{a}+\mathbf{h}) - \sum_{k=0}^{n} \frac{1}{k!} \mathbf{F}^{(k)}(\mathbf{a})(\mathbf{h}^k)\|}{\|\mathbf{h}\|^n} = 0$$

I'll leave the proof to the reader (see Exercises 7 and 8 for help). For $n = 1$, the statement is just the definition of differentiability, and the proof proceeds by (a somewhat intricate) induction on n.

Exercises for Section 6.10

1. Write out the Taylor polynomials of order 1, 2, and 3 of a function $f : \mathbb{R}^2 \to \mathbb{R}$ in terms of its partial derivatives.
2. Find the Taylor polynomial of degree 2 at $\mathbf{a} = \mathbf{0}$ of the function $f(x, y) = \sin(xy)$. Use Corollary 6.10.5 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 6.10.5 to estimate the error term.
4. Consider functions $f : \mathbb{R}^2 \to \mathbb{R}$.
   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 $\Delta f(x, y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}(x, y)$ of f at (x, y).
5. The formula
   $$\sum_{k=0}^{n} \frac{1}{k!} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_k=1}^{n} \frac{\partial^k f}{\partial x_{i_k} \dots \partial x i_2 \partial x_{i_1}} h_{i_k} \cdot \ldots \cdot h_{i_2} h_{i_1} \tag{6.10.1}$$
   for the Taylor polynomials of a function $f : \mathbb{R}^n \to \mathbb{R}$ is rather inefficient as the same derivative shows up many times, only with the differentiations performed in different order. *Multiindices* give us a better way of keeping track of partial derivatives. A multiindex α of order n is just an n-tuple $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ where all the entries $\alpha_1, \alpha_2, \dots, \alpha_n$ are nonnegative integers. We let $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$ and introduce the notation
   $$D^\alpha f(\mathbf{a}) = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_n^{\alpha_n}}(\mathbf{a})$$
   (note that since α_i may be 0, we don't necessarily differentiate with respect to all variables).
   a) If $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ is a multiindex, we define
   $$\alpha! = \alpha_1! \alpha_2! \cdot \ldots \cdot \alpha_n!$$
   (recall tha1 $0! = 1$). Show that if you have α_1 indistinguishable objects of type 1, α_2 indistinguishable objects of type 2 etc., then you can order the objects in
   $$\frac{|\alpha|}{\alpha_1! \alpha_2! \cdot \ldots \cdot \alpha_n!}$$
   distinguishable ways.

b) Show that the Taylor polynomial in formula (6.10.1) above can now be written

$$\sum_{|\alpha|\leq N} \frac{1}{\alpha!} D^\alpha f(\mathbf{a})\mathbf{h}^\alpha$$

where $\mathbf{h}^\alpha = h_1^{\alpha_1} h_2^{\alpha_2} \cdot \ldots \cdot h_n^{\alpha_n}$.

c) Use the formula in b) to write out the Taylor polynomial of order 3 of a function $f : \mathbb{R}^3 \to \mathbb{R}$.

6. Let X be a normed space and assume that $f : X \to \mathbb{R}$ is three times continuously differentiable at $\mathbf{a} \in X$. Assume that $f'(\mathbf{a}) = 0$ and that $\mathbf{f}''(\mathbf{a})$ is strictly positive definite in the following sense: There exists an $\epsilon > 0$ such that

$$f''(\mathbf{a})(\mathbf{r}, \mathbf{r}) \geq \epsilon \|\mathbf{r}\|^2$$

for all $\mathbf{r} \in X$. Show that f has a local minimum at $\mathbf{a}$.

7. In this problem we shall prove Theorem 6.10.6 for functions $f : \mathbb{R} \to \mathbb{R}$. You will be asked to prove the full theorem in the next exercise, but it an advantage to look at one-dimensional case first as the main idea is much easier to spot there. To be precise, we shall prove:

Theorem: *Let O be an open subset of $\mathbb{R}$ and assume that $f : O \to Y$ 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 $\mathbf{h}$ goes to zero, i.e.

$$\lim_{h\to 0} \frac{f(a+h) - \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(a)h^k}{h^n} = 0$$

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

$$\sigma(h) = f(a+h) - f(a) - f'(a)(h) - \ldots - \frac{1}{n!} f^{(n)}(a)h^n$$

Differentiate this expression to get

$$\sigma'(h) = f'(a+h) - f'(a) - \ldots - \frac{1}{(n-1)!} f^{(n)}(a)(h^{n-1}$$

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 Theorem 6.10.6 holds for f and complete the induction argument.

8. In this problem we shall prove Theorem 6.10.6. If you haven't done so already, it may be a good idea to do Exercise 7 first as it will show you the basic idea in a less cluttered context.

   a) Check that for $n = 1$ the statement follows immediately from the definition of differentiability.

   The rest of the proof is by induction on n, but we need some preliminary information on differentiation of functions of the form $\mathbf{h} \mapsto \mathbf{F}^k(\mathbf{a})(\mathbf{h}^{(k)})$.

   b) Assume that $A : X^k \to Y$ is a bounded multilinear map, and define $\mathbf{G}(\mathbf{h}) = A(\mathbf{h}, \mathbf{h}, \ldots, \mathbf{h})$. Show that

   $$\mathbf{G}'(\mathbf{h})(\mathbf{r}) = A(\mathbf{r}, \mathbf{h}, \ldots, \mathbf{h}) + A(\mathbf{h}, \mathbf{r}, \ldots, \mathbf{h}) + \ldots + A(\mathbf{h}, \mathbf{h}, \ldots, \mathbf{r})$$

   (recall Proposition 6.8.4).

   c) Show that if $\mathbf{F} : X \to Y$ is as in the theorem and $k \leq n$, then the derivative of the function

   $$\mathbf{G}_k(\mathbf{h}) = \mathbf{F}^k(\mathbf{a})(\mathbf{h}^{(k)})$$

   is

   $$\mathbf{G}_k'(\mathbf{h})(\mathbf{r}) = k\mathbf{F}^k(\mathbf{a})(\mathbf{r}, \mathbf{h}, \ldots, \mathbf{h})$$

   d) Define a function σ by

   $$\sigma(\mathbf{h}) = \mathbf{F}(\mathbf{a} + \mathbf{h}) - \mathbf{F}(\mathbf{a}) - \mathbf{F}'(\mathbf{a})(\mathbf{h}) - \ldots - \frac{1}{n!}\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{h}^n)$$

   and show that

   $$\sigma'(\mathbf{h})(\mathbf{r}) = \mathbf{F}'(\mathbf{a} + \mathbf{h})(\mathbf{r}) - \mathbf{F}'(\mathbf{a})(\mathbf{r}) - \frac{1}{2}\mathbf{F}''(\mathbf{a})(\mathbf{r}, \mathbf{h}) - \ldots \\ \ldots - \frac{1}{(n-1)!}\mathbf{F}^{(n)}(\mathbf{a})(\mathbf{r}, \mathbf{h}, \ldots, \mathbf{h})$$

   e) Assume that the theorem holds for all $n - 1$ times differentiable functions. Apply it to $\mathbf{F}'$ (as a function from X to $\mathcal{L}(X, Y)$), and explain that $\|\sigma'(\mathbf{h})\|$ goes to zero faster than $\|\mathbf{h}\|^{n-1}$; i.e. that for every $\epsilon > 0$, there is a $\delta > 0$ such that

   $$\|\sigma'(\mathbf{h})\| \leq \epsilon\|\mathbf{h}\|^{n-1}$$

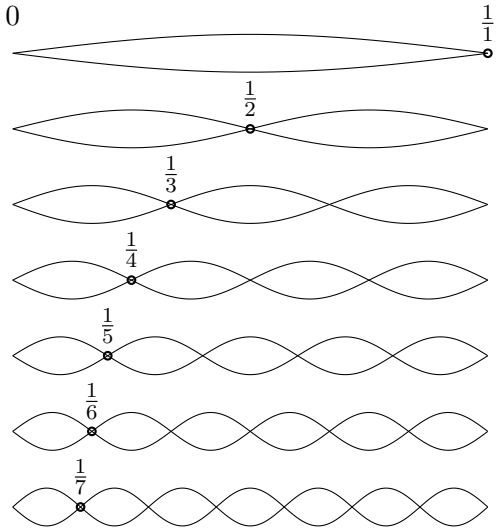
   when $\|\mathbf{h}\| \leq \delta$.

   f) Show that $\|\sigma(\mathbf{h})\| \leq \epsilon\|\mathbf{h}\|^n$ when $\|\mathbf{h}\| \leq \delta$. Conclude that Theorem 6.10.6 holds for $\mathbf{F}$ and complete the induction argument.

Chapter 7

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 above shows some possibilities. If we start with a simple sine curve $f_1(x) = C_1 \sin(\pi x)$, the string will oscillate up an 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 ponts (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

$$f(x) = \sum_{n=1}^{\infty} C_n \sin(n\pi x) \tag{7.0.1}$$

of the starting positions above. The motion of the string will now 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 (7.0.1). Or – perhaps it is? Perhaps any reasonable starting position for the string can be written in the form (7.0.1)? 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 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 allow constant terms and cosine functions in his series. Hence he looked for representations of the form

$$f(x) = A + \sum_{n=1}^{\infty} \big(C_n \sin(n\pi x) + D_n \cos(n\pi x)\big) \tag{7.0.2}$$

with $A, C_n, D_n \in \mathbb{R}$. 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 $[0, 1]$ by $[-\pi, \pi]$ (it is easy to go from one interval to another by scaling the functions, and $[-\pi, \pi]$ has certain notational advantages), and we shall replace $\sin nx$ and $\cos nx$ 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. We begin by taking a closer look at the relationship between complex exponentials and trigonometric functions.

7.1 Complex exponential functions

You may remember the name Fourier from the section 5.3 on inner product spaces, and we shall now see how the abstract Fourier analysis presented there, can be turned into concrete Fourier analysis of functions on the real line. Before we do so, it will be convenient to take a brief look at the functions that will serve as elements of our orthonormal basis. 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:

$$e^{iy} = \cos y + i \sin y \tag{7.1.1}$$

Since we also have

$$e^{-iy} = \cos(-y) + i \sin(-y) = \cos y - i \sin y$$

we may add and subtract to get

$$\cos y = \frac{e^{iy} + e^{-iy}}{2} \tag{7.1.2}$$

$$\sin y = \frac{e^{iy} - e^{-iy}}{2i} \tag{7.1.3}$$

Formulas (7.1.1)-(7.1.3) give us important connections between complex exponetials and trigonometric functions that we shall exploit in the next sections.

We need some information about functions $f : \mathbb{R} \to \mathbb{C}$ of the form

$$f(x) = e^{(a+ib)x} = e^{ax} \cos bx + ie^{ax} \sin bx, \qquad \text{where } a \in \mathbb{R}$$

If we differentiate f by differentiating the real and complex parts separately, we get

$$f'(x) = ae^{ax} \cos bx - be^{ax} \sin bx + iae^{ax} \sin bx + ibe^{ax} \cos bx =$$

$$= ae^{ax} (\cos bx + i \sin bx) + ibe^{ax} (\cos bx + i \sin bx) = (a + ib)e^{(a+ib)x}$$

and hence we have the formula

$$\left(e^{(a+ib)x}\right)' = (a + ib)e^{(a+ib)x} \tag{7.1.4}$$

that we would expect from the real case. Antidifferentiating, we see that

$$\int e^{(a+ib)x}\, dx = \frac{e^{(a+ib)x}}{a + ib} + C \tag{7.1.5}$$

where $C = C_1 + iC_2$ is an arbitrary, complex constant.

Note that if we multiply by the conjugate $a - ib$ in the numerator and the denominator, we get

$$\frac{e^{(a+ib)x}}{a+ib} = \frac{e^{(a+ib)x}(a-ib)}{(a+ib)(a-ib)} = \frac{e^{ax}}{a^2+b^2}(\cos bx + i\sin bx)(a-ib) =$$

$$= \frac{e^{ax}}{a^2+b^2}\big(a\cos bx + b\sin bx + i(a\sin bx - b\cos bx)\big)$$

Hence (7.1.5) may also be written

$$\int \big(e^{ax}\cos bx \quad + \quad ie^{ax}\sin bx\big)\,dx =$$

$$= \quad \frac{e^{ax}}{a^2+b^2}\big(a\cos bx + b\sin bx + i(a\sin bx - b\cos bx)\big)$$

Separating the real and the imaginary parts, we get

$$\int e^{ax}\cos bx\,dx = \frac{e^{ax}}{a^2+b^2}\big(a\cos bx + b\sin bx\big) \tag{7.1.6}$$

and

$$\int e^{ax}\sin bx\,dx = \frac{e^{ax}}{a^2+b^2}\big(a\sin bx - b\cos bx\big) \tag{7.1.7}$$

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 (7.1.5).

We shall be particularly interested in 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 7.1.1 *For all $n, m \in \mathbb{Z}$ we have*

$$\int_{-\pi}^{\pi} e_n(x)\overline{e_m(x)}\,dx = \begin{cases} 0 & \text{if } n \neq m \\ 2\pi & \text{if } n = m \end{cases}$$

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. □

The proposition shows that the family $\{e_n\}_{n\in\mathbb{Z}}$ is almost orthonormal with respect to the inner product

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx.$$

The only problem is that $\langle e_n, e_n\rangle$ is 2π and not 1. We could fix this by replacing e_n by $\frac{e_n}{\sqrt{2\pi}}$, but instead we shall choose to change the inner product to

$$\langle f, g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx.$$

Abusing terminology slightly, we shall refer to this at the *L_2-inner product on $[-\pi, \pi]$*. The norm it induces will be called the *L_2-norm* $\|\cdot\|_2$. It is defined by

$$\|f\|_2 = \langle f, f\rangle^{\frac{1}{2}} = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |f(x)|^2\,dx\right)^{\frac{1}{2}}$$

The Fourier coefficients of a function f with respect to $\{e_n\}_{n\in\mathbb{Z}}$ are defined by

$$\langle f, e_n\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{e_n(x)}\,dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx$$

From Section 5.3 we know that $f = \sum_{n=-\infty}^{\infty}\langle f, e_n\rangle e_n$ (where the series converges in L_2-norm) provided f belongs to a space where $\{e_n\}_{n\in\mathbb{Z}}$ is a basis. We shall study this question in detail in the next sections. For the time being, we look at an example of how to compute Fourier coefficients.

Example 1: We shall compute the Fourier coefficients α_n of the function $f(x) = x$. By definition

$$\alpha_n = \langle f, e_n\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} xe^{-inx}\,dx$$

It is easy to check that $\alpha_0 = \int_{-\pi}^{\pi} x\,dx = 0$. For $n \neq 0$, we use integration by parts (see Exercise 8) with $u = x$ and $v' = e^{-inx}$. We get $u' = 1$ and

$v = \frac{e^{-inx}}{-in}$, and:

$$\alpha_n = -\frac{1}{2\pi}\left[x\frac{e^{-inx}}{in}\right]_{-\pi}^{\pi} + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{-inx}}{in}\,dx =$$

$$= \frac{(-1)^{n+1}}{in} - \frac{1}{2\pi}\left[\frac{e^{-inx}}{n^2}\right]_{-\pi}^{\pi} = \frac{(-1)^{n+1}}{in}$$

The Fourier series becomes

$$\sum_{n=-\infty}^{\infty}\alpha_n e_n = \sum_{n=-\infty}^{-1}\frac{(-1)^{n+1}}{in}e^{inx} + \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{in}e^{inx} =$$

$$= \sum_{n=1}^{\infty}\frac{2(-1)^{n+1}}{n}\sin(nx)$$

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.

Exercises for Section 7.1

1. Show that $\langle f, g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx$ is an inner product on $C([-\pi, \pi], \mathbb{C})$.
2. Deduce the formulas for $\sin(x + y)$ and $\cos(x + y)$ from the rule $e^{i(x+y)} = e^{ix}e^{iy}$.
3. 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 $\int \sin x \cos 4x\,dx$.
4. Find the Fourier series of $f(x) = e^x$.
5. Find the Fourier series of $f(x) = x^2$.
6. Find the Fourier sries of $f(x) = \sin\frac{x}{2}$.
7. 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 multiplum 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 multiplum of 2π.

d) Use the result in c) to find formulas for $\sum_{k=0}^{n} \cos(kx)$ and $\sum_{k=0}^{n} \sin(kx)$.

8. 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.

7.2 Fourier series

Recall from the previous section that the functions

$$e_n(x) = e^{inx}, \quad n \in \mathbb{Z}$$

form an orthonormal set with respect to the L_2-inner product

$$\langle f, g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx$$

The Fourier coefficients of a continuous function $f : [-\pi, \pi] \to \mathbb{C}$ with respect to this set are given by

$$\alpha_n = \langle f, e_n\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{e_n(x)}\,dx$$

From Parseval's theorem 5.3.10, we know that if $\{e_n\}$ is a basis (for whatever space we are working with), then

$$f(x) = \sum_{n=-\infty}^{\infty} \alpha_n e_n(x)$$

where the series converges in the L_2-norm, i.e.

$$\lim_{N\to\infty} \|f - \sum_{n=-N}^{N} \alpha_n e_n\|_2 = 0$$

At this stage, life becomes complicated in two ways. First, we don't know yet that $\{e_n\}_{n\in\mathbb{Z}}$ is a basis for $C([-\pi, \pi], \mathbb{C})$, and second, we don't really know what L_2-convergence means. It turns out that L_2-convergence is quite weak, and that a sequence may converge in L_2-norm without actually converging at any point! This means that we would also like to investigate other forms for convergence (pointwise, uniform etc.).

Let us begin by observing that since $e_n(-\pi) = e_n(\pi)$ for all $n \in \mathbb{Z}$, any function that is the pointwise limit of a series $\sum_{n=-\infty}^{\infty} \alpha_n e_n$ must also satisfy this periodicity assumption. Hence it is natural to introduce the following class of functions:

Definition 7.2.1 *Let C_P be the set of all continuous functions $f : [-\pi, \pi] \to \mathbb{C}$ such that $f(-\pi) = f(\pi)$. A function in C_P is called a* trigonometric polynomial *if it is of the form $\sum_{n=-N}^{N} \alpha_n e_n$ where $N \in \mathbb{N}$ and each $\alpha_n \in \mathbb{C}$.*

To distinguish it from the L_2-norm, we shall denote the supremum norm on $C([-\pi, \pi], \mathbb{C})$ by $\| \cdot \|_\infty$, i.e.

$$\|f\|_\infty = \sup\{|f(x)| \, : \, x \in [-\pi.\pi]\}$$

Note that the metric generated by $\| \cdot \|_\infty$ is the metric ρ that we studied in Chapter 4. Hence convergence with respect to $\| \cdot \|_\infty$ is the same as uniform convergence.

Theorem 7.2.2 *The trigonometric polynomials are dense in C_P in the $\| \cdot \|_\infty$-norm. Hence for any $f \in C_P$ there is a sequence $\{p_n\}$ of trigonometric polynomials which converges uniformly to f.*

It is possible to prove this result from Weierstrass' Approximation Theorem 3.10.1, but the proof is technical and not very informative. In the next section, we shall get a more informative proof from ideas we have to develop anyhow, and we postpone the proof till then. In the meantime we look at some consequences.

Corollary 7.2.3 *For all $f \in C_P$, the Fourier series $\sum_{n=-\infty}^{\infty} \langle f, e_n \rangle e_n$ converges to f in L_2-norm, i.e. $\lim_{N\to\infty} \|f - \sum_{n=-N}^{N} \langle f, e_n \rangle e_n\|_2 = 0$.*

Proof: Given $\epsilon > 0$, we must show that there is an $N \in \mathbb{N}$ such that $\|f - \sum_{n=-M}^{M} \langle f, e_n \rangle e_n\|_2 < \epsilon$ when $M \geq N$. According to the theorem above, there is a trigonometric polynomial $p(x) = \sum_{n=-N}^{N} \alpha_n e_n$ such that $\|f - p\|_\infty < \epsilon$. Hence

$$\|f - p\|_2 = \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x) - p(x)|^2 \, dx \right)^{\frac{1}{2}} < \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \epsilon^2 \, dx \right)^{\frac{1}{2}} = \epsilon$$

According to Proposition 5.3.8, $\|f - \sum_{n=-M}^{M} \langle f, e_n \rangle e_n\|_2 \leq \|f - p\|_2$ for all $M \geq N$, and the corollary follows. □

The corollary above is rather unsatisfactory. It is particularly inconvenient that it only applies to periodic functions such that $f(-\pi) = f(\pi)$ (although we can not have *pointwise convergence* to functions violating this condition, we may well have L_2-convergence as we soon shall see). To get a better result, we introduce a bigger space D of piecewise continuous functions.

Definition 7.2.4 *A function $f : [-\pi, \pi] \to \mathbb{C}$ is said to be* piecewise continuous with one sided limits *if there exists a finite set of points*

$$-\pi = a_0 < a_1 < a_2 < \ldots < a_{n-1} < a_n = \pi$$

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 $f(a_0) = f(a_n) = \frac{1}{2}(f(a_n^-) + f(a_0^+))$

The collection of all such functions will be denoted by D.

Remark: Part (iii) is only included for technical reasons (we must specify the values at the jump points to make D an inner product space), but it reflects how Fourier series behave — at jump points they always choose the average value. The treatment of the end points may seem particularly strange; why should we enforce the average rule even here? The reason is that since the trigonometric polynomials are 2π-periodic, they regard 0 and 2π as the "same" point, and hence it is natural to compare the right limit at 0 to the left limit at 2π.

Note that the functions in D are bounded and integrable, that the sum and product of two functions in D are also in D, and that D is an inner product space over $\mathbb{C}$ with the L_2-inner product. The next lemma will allow us to extend the corollary above to D.

Lemma 7.2.5 *C_P is dense in D in the L_2-norm, i.e. for each $f \in D$ and each $\epsilon > 0$, there is a $g \in C_P$ such that $\|f - g\|_2 < \epsilon$.*

Proof: I only sketch the main idea of the proof, leaving the details to the reader. Assume that $f \in D$ and $\epsilon > 0$ are given. To construct g, choose a very small $\delta > 0$ (it is your task to figure out how small) and construct g as follows: Outside the (nonoverlapping) intervals $(a_i - \delta, a_i + \delta)$, we let g agree with f, but in each of these intervals, g follows the straight line connecting the points $(a_i - \delta, f(a_i - \delta))$ and $(a_i + \delta, f(a_i + \delta))$ on f's graph. Check that if we choose δ small enough, $\|f - g\|_2 < \epsilon$ (In making your choice, you have to take $M = \sup\{|f(x)| : x \in [-\pi, \pi]\}$ into account, and you also have to figure ut what to do at the endpoints $-\pi, \pi$ of the interval). □

We can now extend the corollary above from C_P to D.

Theorem 7.2.6 *For all $f \in D$, the Fourier series $\sum_{n=-\infty}^{\infty}\langle f, e_n\rangle e_n$ converges to f in L_2-norm, i.e. $\lim_{N\to\infty}\|f - \sum_{n=-N}^{N}\langle f, e_n\rangle e_n\|_2 = 0$.*

Proof: Assume that $f \in D$ and $\epsilon > 0$ are given. By the lemma, we know that there is a $g \in C_P$ such that $\|f - g\|_2 < \frac{\epsilon}{2}$, and by Corollary 7.2.3 there is a trigonometric polynomial $p = \sum_{n=-N}^{N}\alpha_n e_n$ such that $\|g - p\|_2 < \frac{\epsilon}{2}$. The triangle inequality now tells us that

$$\|f - p\|_2 \leq \|f - g\|_2 + \|g - p\|_2 < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Invoking Proposition 5.3.8 again, we see that for $M \geq N$, we have

$$\|f - \sum_{n=-M}^{M}\langle f, \mathbf{e}_n\rangle e_n\|_2 \leq \|f - p\|_2 < \epsilon$$

and the theorem is proved. □

The theorem above is satisfactory in the sense that we know that the Fourier series of f converges to f for a reasonably wide class of functions. However, we still have things to attend to: We haven't proved Theorem 7.2.2 yet, and we would really like to prove that Fourier series converge pointwise (or even uniformly) for a reasonable class of functions. We shall take a closer look at these questions in the next sections.

Exercises for Section 7.2

1. Show that C_P is a closed subset of $C([-\pi, \pi], \mathbb{C})$
2. 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.
3. In this problem we shall show that if $f : [-\pi, \pi] \to \mathbb{R}$ is a *realvalued* function, then the Fourier series $\sum_{n=-\infty}^{\infty}\alpha_n e_n$ can be turned into a sine/cosine-series of the form (7.2.2).
   a) Show that if $\alpha_n = a_n + ib_n$ are Fourier coefficients of f, then $\alpha_{-n} = \overline{\alpha}_n = a_n - ib_n$.
   b) Show that $a_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx$ and $b_n = -\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\,dx$.

c) Show that the Fourier series can be written

$$\alpha_0 + \sum_{n=1}^{\infty} \big(2a_n \cos(nx) - 2b_n \sin(nx)\big)$$

4. Complete the proof of Lemma 7.2.5.

7.3 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 are orthonormal. 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 $f \in D$ periodically outside the basic interval such that $f(x + 2\pi) = f(x)$ for all $x \in \mathbb{R}$. The figure shows the extension graphically; in part a) we have the original function, and in part b) (a part of) the periodic extension. As there is no danger of confusion, we shall denote the original function and the extension by the same symbol f.

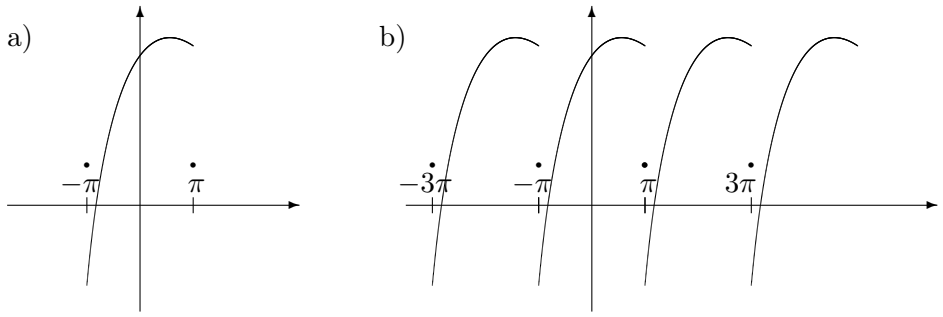


Figure 1

To see the point of this extension more clearly, assume that we have a function $f : [-\pi, \pi] \to \mathbb{R}$. 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 now longer 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

$$\int_{-\pi}^{\pi} f(x)\,dx = \int_{-\pi+a}^{\pi+a} f(u-a)\,du = \int_{-\pi+a}^{\pi} f(u-a)\,du + \int_{\pi}^{\pi+a} f(u-a)\,du$$

$$= \int_{-\pi+a}^{\pi} f(u-a)\,du + \int_{-\pi}^{-\pi+a} f(u-a)\,du = \int_{-\pi}^{\pi} f(u-a)\,du$$

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.

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 $f \in D$ the way we did above, 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".

Let us begin 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. Since

$$\alpha_n = \langle f, e_n \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-int}\,dt$$

we have

$$s_N(x) = \frac{1}{2\pi} \sum_{n=-N}^{N} \left(\int_{-\pi}^{\pi} f(t) e^{-int}\,dt \right) e^{inx} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \sum_{n=-N}^{N} e^{in(x-t)}\,dt =$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-u) \sum_{n=-N}^{N} e^{inu}\,du$$

where we in the last step has 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 geomtric 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:

$$D_N(u) = \frac{e^{-iNu} - e^{i(N+1)u}}{1 - e^{iu}} = \frac{e^{-i(N+\frac{1}{2})u} - e^{i(N+\frac{1}{2})u}}{e^{-i\frac{u}{2}} - e^{i\frac{u}{2}}} = \frac{\sin((N+\frac{1}{2})u)}{\sin\frac{u}{2}}$$

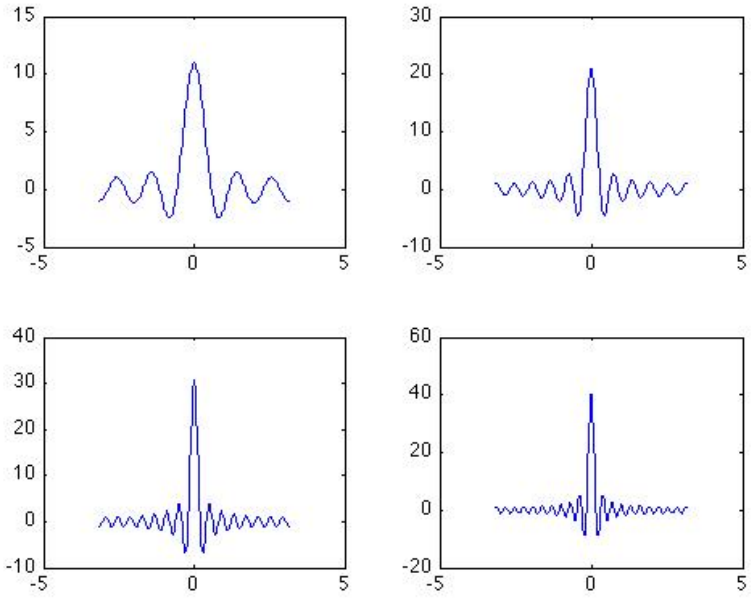
where we have used the formula $\sin x = \frac{e^{ix} - e^{-ix}}{2i}$ twice in the last step. This formula gives us a nice, compact expression for $D_N(u)$. If we substitute it into the formula above, we get

$$s_N(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-u)\, \frac{\sin((N+\frac{1}{2})u)}{\sin\frac{u}{2}}\, du$$

If we want to prove that 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, Dirichlet used this approach to prove:

Theorem 7.3.1 (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. In 1873, the German mathematician Paul Du Bois-Reymond explained why they failed by constructing a periodic, continuous function whose Fourier series diverges at a dense set of points.



It turns out that the theory for pointwise convergence of Fourier series is quite complicated, and 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 which is easier to handle, and which will also give us some tools that are useful in the proof of Dini's test. This alternative notion of convergence is called *Cesaro convergence* or *convergence in Cesaro 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. The figure above 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 $2n + 1$. 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 Cesaro convergence in the next section, we get another set of kernels — the *Fejér kernels* — and they turn out not to have this problem. This is the main reason why Cesaro convergence works much better than ordinary convergence for Fourier series.

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 7.3.2 *The Dirichlet kernel $D_n(t)$ is an even, realvalued function such that $|D_n(t)| \leq D_n(0) = 2n+1$ for all t. For all n,*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(t)\,dt = 1$$

but

$$\lim_{n\to\infty}\int_{-\pi}^{\pi} |D_n(t)|\,dt \to \infty$$

Proof: That D_n is realvalued 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

$$D_n(t) = |\sum_{k=-n}^{n} e^{ikt}| \leq \sum_{k=-n}^{n} |e^{ikt}| = 2n+1 = D_n(0)$$

Similarly for the integral

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(t)\,dt = \sum_{k=-n}^{n} \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{e}^{ikt}\,dt = 1$$

as all integrals except the one for $k=0$ is 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

$$\int_{-\pi}^{\pi} |D_n(t)|\,dt = \int_{0}^{\pi} 2|D_n(t)|\,dt \geq \int_{0}^{\pi} \frac{4|\sin((n+\frac{1}{2})t)|}{|t|}\,dt =$$

$$= \int_{0}^{(n+\frac{1}{2})\pi} \frac{4|\sin z|}{z}\,dz \geq \sum_{k=1}^{n}\int_{(k-1)\pi}^{k\pi} \frac{4|\sin z|}{k\pi}\,dz = \frac{8}{\pi}\sum_{k=1}^{n}\frac{1}{k}$$

The expression on the right goes to infinity since the series diverges. □

Exercises for Section 7.3

1. Let $f : [-\pi, \pi] \to \mathbb{C}$ 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 $\lim_{t\to 0}\frac{\sin((n+\frac{1}{2})t)}{\sin\frac{t}{2}}$.
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 7.3.2 (the part that proves that $\lim_{n\to\infty}\int_{-\pi}^{\pi}|D_n(t)|\,dt = \infty$).

7.4 The Fejér kernel

Before studying the Fejér kernel, we shall take a look at a generalized notion of convergence for sequences. Certain sequences such at

$$0, 1, 0, 1, 0, 1, 0, 1, \ldots$$

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 7.4.1 *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* Cesaro mean *if*

$$a = \lim_{n\to\infty} S_n = \lim_{n\to\infty} \frac{a_0 + a_1 + \cdots + a_{n-1}}{n}$$

We shall write $a = C\text{-}\lim_{n\to\infty} a_n$.

The sequence at the beginning of the section converges to $\frac{1}{2}$ in Cesaro mean, but diverges in the ordinary sense. Let us prove that the opposite can not happen:

Lemma 7.4.2 *If $\lim_{n\to\infty} a_n = a$, then $C\text{-}\lim_{n\to\infty} a_n = a$.*

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 $M = \max\{|a_k - a| : k = 0, 1, 2, \ldots\}$, we have for any $n \geq K$:

$$|S_n - a| = \left|\frac{(a_0 - a) + (a_1 - a) + \cdots + (a_{K-1} - a) + (a_K - a) + \cdots (a_{n-1} - a)}{n}\right| \leq$$

$$\leq \left|\frac{(a_0 - a) + (a_1 - a) + \cdots + (a_{K-1} - a)}{n}\right| + \left|\frac{(a_K - a) + \cdots (a_{n-1} - a)}{n}\right| \leq \frac{MK}{n} + \frac{\epsilon}{2}$$

Choosing n large enough, we get $\frac{MK}{n} < \frac{\epsilon}{2}$, and the lemma follows. □

The idea behind the Fejér kernel is to show that the partial sums $s_n(x)$ converge to $f(x)$ in Cesaro 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

$$s_k(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-u)D_k(u)\,du$$

where D_k is the Dirichlet kernel, we get

$$S_n(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-u)\left(\frac{1}{n}\sum_{k=0}^{n-1} D_k(u)\right)\,du = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-u)F_n(u)\,du$$

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 7.4.3 *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

$$\sum_{k=0}^{n-1}\sin((k+\frac{1}{2})u) = \frac{1}{2i}\left(\frac{1-e^{inu}+1-e^{-inu}}{e^{-i\frac{u}{2}} - e^{i\frac{u}{2}}}\right) = \frac{1}{2i}\left(\frac{e^{inu}-2+e^{-inu}}{e^{i\frac{u}{2}} - e^{-i\frac{u}{2}}}\right) =$$

$$= \frac{1}{2i} \cdot \frac{(e^{i\frac{nu}{2}} - e^{-\frac{nu}{2}})^2}{e^{i\frac{u}{2}} - e^{-i\frac{u}{2}}} = \frac{\left(\frac{e^{i\frac{nu}{2}} - e^{-\frac{nu}{2}})}{2i}\right)^2}{\frac{e^{i\frac{u}{2}} - e^{-i\frac{u}{2}}}{2i}} = \frac{\sin^2(\frac{nu}{2})}{\sin\frac{u}{2}}$$

and thus

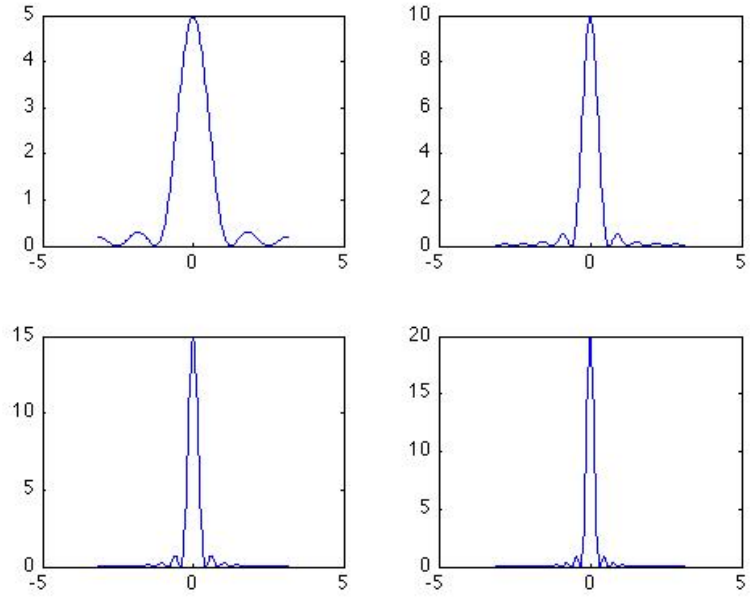
$$F_n(u) = \frac{1}{n\sin(\frac{u}{2})} \sum_{k=0}^{n-1} \sin((k+\frac{1}{2})u) = \frac{\sin^2(\frac{nu}{2})}{n\sin^2\frac{u}{2}}$$

To prove that $F_n(0) = n$, we just have to sum an arithmetic series

$$F_n(0) = \frac{1}{n}\sum_{k=0}^{n-1} D_k(0) = \frac{1}{n}\sum_{k=0}^{n-1}(2k+1) = n$$

□

The figure below shows the Fejer 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 values, 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 7.4.4 *For all n, the Fejér kernel F_n is an even, positive function such that*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} F_n(x)\,dx = 1$$

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 Proposition 7.3.2, we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} F_n(x)\,dx = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{n}\sum_{k=0}^{n-1} D_k\,dx = \frac{1}{n}\sum_{k=0}^{n-1}\frac{1}{2\pi}\int_{-\pi}^{\pi} D_k\,dx = \frac{1}{n}\sum_{k=0}^{n-1} 1 = 1$$

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

$$F_n(x) = \frac{\sin^2(\frac{nx}{2})}{n\sin^2\frac{x}{2}} \leq \frac{1}{n(\frac{2}{\pi}\frac{x}{2})^2} \leq \frac{\pi^2}{nx^2}$$

□

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 Cesaro mean. We have already observed that

$$S_n(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-u)F_n(u)\,du$$

If we introduce a new variable $t = -u$ and use that F_n is even, we get

$$S_n(x) = \frac{1}{2\pi}\int_{\pi}^{-\pi} f(x+t)F_n(-t)\,(-dt) =$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x+t)F_n(t)\,dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x+u)F_n(u)\,du$$

If we combine the two expressions we now have for $S_n(x)$, we get

$$S_n(x) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \left(f(x+u) + f(x-u)\right)F_n(u)\,du$$

Since $\frac{1}{2\pi}\int_{-\pi}^{\pi} F_n(u)\,du = 1$, we also have

$$f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)F_n(u)\,du$$

Hence

$$S_n(x) - f(x) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \big(f(x+u) + f(x-u) - 2f(x)\big) F_n(u)\, du$$

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. Here are the technical details.

Theorem 7.4.5 *If $f \in D$, then S_n converges to f on $[-\pi, \pi]$, i.e. the Fourier series converges in Cesaro mean. The convergence is uniform on each subinterval $[a, b] \subseteq [-\pi, \pi]$ 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

$$|S_n(x) - f(x)| \leq \frac{1}{4\pi} \int_{-\pi}^{\pi} |f(x+u) + f(x-u) - 2f(x)|\, F_n(u)\, du =$$

$$= \frac{1}{4\pi} \int_{-\delta}^{\delta} |f(x+u) + f(x-u) - 2f(x)|\, F_n(u)\, du +$$

$$+ \frac{1}{4\pi} \int_{-\pi}^{-\delta} |f(x+u) + f(x-u) - 2f(x)|\, F_n(u)\, du +$$

$$+ \frac{1}{4\pi} \int_{\delta}^{\pi} |f(x+u) + f(x-u) - 2f(x)|\, F_n(u)\, du$$

For the first integral we have

$$\frac{1}{4\pi} \int_{-\delta}^{\delta} |f(x+u) + f(x-u) - 2f(x)|\, F_n(u)\, du \leq$$

$$\leq \frac{1}{4\pi} \int_{-\delta}^{\delta} \epsilon F_n(u)\, du \leq \frac{1}{4\pi} \int_{-\pi}^{\pi} \epsilon F_n(u)\, du = \frac{\epsilon}{2}$$

For the second integral we get

$$\frac{1}{4\pi} \int_{-\pi}^{-\delta} |f(x+u) + f(x-u) - 2f(x)|\, F_n(u)\, du \leq$$

$$\leq \frac{1}{4\pi} \int_{-\pi}^{-\delta} 4\|f\|_\infty \frac{\pi^2}{n\delta^2}\, du = \frac{\pi^2 \|f\|_\infty}{n\delta^2}$$

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 $x \in [a, b]$. One might think that this follows immediately from the fact that a continuous function on a compact interval $[a, b]$ is uniformly continuous, but we need to be a little careful as $x + u$ or $x - u$ may be outside the interval $[a, b]$ even if x is inside. The quickest way to fix this, is to observe that since f is in D, it must be continuous — and hence uniformly continuous — on a slightly larger interval $[a - \eta, b + \eta]$. This means that we can use the same $\delta < \eta$ for all x and $x \pm u$ in $[a-\eta, b+\eta]$, and this clinches the argument.□

We have now finally proved Theorem 7.2.2 which we restate here:

Corollary 7.4.6 *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: According to the theorem, the sums $S_N(x) = \frac{1}{N}\sum_{n=0}^{N-1} s_n(x)$ converge uniformly to f. Since each s_n is a trigonometric polynomial, so are the S_N's. □

Exercises to Section 7.4

1. Let $\{a_n\}$ be the sequence $1, 0, 1, 0, 1, 0, 1, 0, \ldots$. Prove that C-$\lim_{n\to\infty} a_n = \frac{1}{2}$.
2. Assume that $\{a_n\}$ and $\{b_n\}$ converge in Cesaro mean. Show that
$$\text{C}-\lim_{n\to\infty}(a_n + b_n) = \text{C}-\lim_{n\to\infty} a_n + \text{C}-\lim_{n\to\infty} b_n$$
3. Check that $F_n(0) = n$ by computing $\lim_{u\to 0} \frac{\sin^2(\frac{nu}{2})}{n\sin^2\frac{u}{2}}$.
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. Assume that $f \in C_P$. Work through the details of the proof of Theorem 7.4.5 and check that S_n converges uniformly to f.

7.5 The Riemann-Lebesgue lemma

The Riemann-Lebesgue lemma is a seemingly simple observation about the size of the Fourier coefficients, but it turns out to be a very efficient tool in the study of pointwise convergence.

Theorem 7.5.1 (Riemann-Lebesgue Lemma) *If $f \in D$ and*

$$\alpha_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx, \quad n \in \mathbb{Z},$$

are the Fourier coefficients of f, then $\lim_{|n|\to\infty} \alpha_n \to 0$.

Proof: According to Bessel's inequality 5.3.9, $\sum_{n=-\infty}^{\infty} |\alpha_n|^2 \leq \|f\|_2^2 < \infty$, and hence $\alpha_n \to 0$ as $|n| \to \infty$. □

Remark: We are cheating a little here as we only prove the Riemann-Lebesgue lemma for function which are in D and hence square integrable. The lemma holds for integrable functions in general, but even in that case the proof is quite easy.

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 7.5.2 *If $f \in D$ and $[a,b] \subseteq [-\pi,\pi]$, then*

$$\lim_{|n|\to\infty} \int_a^b f(x)e^{-inx}\,dx = 0$$

Also

$$\lim_{|n|\to\infty} \int_a^b f(x)\cos(nx)\,dx = \lim_{|n|\to\infty} \int_a^b f(x)\sin(nx)\,dx = 0$$

Proof: Let g be the function (this looks more horrible than it is!)

$$g(x) = \begin{cases} 0 & \text{if } x \notin [a,b] \\ f(x) & \text{if } x \in (a,b) \\ \frac{1}{2}\lim_{x\downarrow a} f(x) & \text{if } x = a \\ \frac{1}{2}\lim_{x\uparrow b} f(x) & \text{if } x = b \end{cases}$$

then g is in D, and

$$\int_a^b f(x)e^{-inx}\,dx = \int_{-\pi}^{\pi} g(x)e^{-inx}\,dx = 2\pi\alpha_n$$

where α_n is the Fourier coefficient of g. By the Riemann-Lebesgue lemma, $\alpha_n \to 0$. The last two parts follows from what we have just proved and the

identities $\sin(nx) = \frac{e^{inx}-e^{-inx}}{2i}$ and $\cos(nx) = \frac{e^{inx}+e^{-inx}}{2}$ □

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. If the function f is relatively smooth, the positive and negative contributions cancel more and more as n increases, and in the limit there is nothing left. This argument also indicates 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 7.5.3 *If $f \in D$ and $[a,b] \subseteq [-\pi,\pi]$, then*

$$\lim_{|n|\to\infty} \int_a^b f(x) \sin\left((n+\frac{1}{2})x\right) dx = 0$$

Proof: Follows from the corollary above and the identity

$$\sin\left((n+\frac{1}{2})x\right) = \sin(nx)\cos\frac{x}{2} + \cos(nx)\sin\frac{x}{2}$$

□

Exercises to Section 7.5

1. Work out the details of the $\sin(nx)$- and $\cos(nx)$-part of Corollary 7.5.2.
2. Work out the details of the proof of Corollary 7.5.3.
3. a) 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.
   b) Let f be an integrable function, and assume that for each $\epsilon > 0$ there is a trigonometric polynomial such that $\frac{1}{2\pi}\int_{-\pi}^{\pi} |f(t) - p(t)|\, dt < \epsilon$. Show that if $\alpha_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-int}\, dt$ are the Fourier coefficients of f, then $\lim_{|n|\to\infty} \alpha_n = 0$.

7.6 Dini's test

We shall finally take a serious look at pointwise convergence of Fourier series. As aready 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.

Recall from Section 7.3 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

$$s_N(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-u) D_N(u)\, du$$

If we change variable in the intergral and use the symmetry of D_N, we see that we also get

$$s_N(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x+u) D_N(u)\, du$$

Combining these two expressions, we get

$$s_N(x) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \big(f(x+u) + f(x-u)\big) D_N(u)\, du$$

Since $\frac{1}{2\pi} \int_{-\pi}^{\pi} D_N(u)\, du = 1$, we also have

$$f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) D_N(u)\, du$$

and hence

$$s_N(x) - f(x) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \big(f(x+u) + f(x-u) - 2f(x)\big) D_N(u)\, du$$

(note that the we are now doing exactly the same to the Dirichlet kernel as we did to the Fejér kernel in Section 7.4). 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:

Lemma 7.6.1 *Let $f \in D$ and assume that there is a $\eta > 0$ such that*

$$\lim_{N\to\infty} \frac{1}{4\pi} \int_{-\eta}^{\eta} \big(f(x+u) + f(x-u) - 2f(x)\big) D_N(u)\, du = 0$$

Then the Fourier series $\{s_N(x)\}$ converges to $f(x)$.

Proof: Note that since $\frac{1}{\sin \frac{x}{2}}$ is a bounded function on $[\eta, \pi]$, Corollary 7.5.3 tells us that

$$\lim_{N\to\infty} \frac{1}{4\pi} \int_{\eta}^{\pi} \big(f(x+u) + f(x-u) - 2f(x)\big) D_N(u)\, du =$$

$$= \lim_{N\to\infty} \frac{1}{4\pi} \int_\eta^\pi \big[\big(f(x+u)+f(x-u)-2f(x)\big)\, \frac{1}{\sin\frac{u}{2}}\big] \sin\big((N+\frac{1}{2})u\big)\, du = 0$$

The same obviously holds for the integral from $-\pi$ to $-\eta$, and hence

$$s_N(x) - f(x) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \big(f(x+u)+f(x-u)-2f(x)\big) D_N(u)\, du =$$

$$= \frac{1}{4\pi} \int_{-\pi}^{\eta} \big(f(x+u)+f(x-u)-2f(x)\big) D_N(u)\, du +$$

$$+ \frac{1}{4\pi} \int_{-\eta}^{\eta} \big(f(x+u)+f(x-u)-2f(x)\big) D_N(u)\, du +$$

$$+ \frac{1}{4\pi} \int_{\eta}^{\pi} \big(f(x+u)+f(x-u)-2f(x)\big) D_N(u)\, du$$

$$\to 0+0+0 = 0$$

□

Theorem 7.6.2 (Dini's test) *Let $x \in [-\pi,\pi]$, and assume that there is a $\delta > 0$ such that*

$$\int_{-\delta}^{\delta} \left| \frac{f(x+u)+f(x-u)-2f(x)}{u} \right| du < \infty$$

Then the Fourier series converges to the function f at the point x, i.e. $s_N(x) \to f(x)$.

Proof: According to the lemma, it suffices to prove that

$$\lim_{N\to\infty} \frac{1}{4\pi} \int_{-\delta}^{\delta} \left(f(x+u)+f(x-u)-2f(x)\right) D_N(u)\, du = 0$$

Given an $\epsilon > 0$, we have to show that if $N \in \mathbb{N}$ is large enough, then

$$\frac{1}{4\pi} \int_{-\delta}^{\delta} \big(f(x+u)+f(x-u)-2f(x)\big) D_N(u)\, du < \epsilon$$

Since the integral in the theorem converges, there is an $\eta > 0$ such that

$$\int_{-\eta}^{\eta} \left| \frac{f(x+u)+f(x-u)-2f(x)}{u} \right| du < \epsilon$$

Since $|\sin v| \geq \frac{2|v|}{\pi}$ for $v \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ (make a drawing), we have $|D_N(u)| = |\frac{\sin((N+\frac{1}{2})u)}{\sin\frac{u}{2}}| \leq \frac{\pi}{|u|}$ for $u \in [-\pi,\pi]$. Hence

$$|\frac{1}{4\pi} \int_{-\eta}^{\eta} \big(f(x+u)+f(x-u)-2f(x)\big) D_N(u)\, du| \leq$$

$$\leq \frac{1}{4\pi}\int_{-\eta}^{\eta} |f(x+u)+f(x-u)-2f(x)|\,\frac{\pi}{|u|}\,du < \frac{\epsilon}{4}$$

By Corollary 7.5.3 we can get

$$\frac{1}{4\pi}\int_{\eta}^{\delta} \big(f(x+u)+f(x-u)-2f(x)\big)D_N(u)\,du$$

as small as we want by choosing N large enough and similarly for the integral from $-\delta$ to $-\eta$. In particular, we can get

$$\frac{1}{4\pi}\int_{-\delta}^{\delta} \big(f(x+u)+f(x-u)-2f(x)\big)D_N(u)\,du =$$

$$= \frac{1}{4\pi}\int_{-\delta}^{-\eta} \big(f(x+u)+f(x-u)-2f(x)\big)D_N(u)\,du+$$

$$+\frac{1}{4\pi}\int_{-\eta}^{\eta} \big(f(x+u)+f(x-u)-2f(x)\big)D_N(u)\,du+$$

$$+\frac{1}{4\pi}\int_{\eta}^{\delta} \big(f(x+u)+f(x-u)-2f(x)\big)D_N(u)\,du$$

less than ϵ, and hence the theorem is proved. □

Dini's test has some immediate consequences that we leave to the reader to prove.

Corollary 7.6.3 *If $f \in D$ is differentiable at a point x, then the Fourier series converges to $f(x)$ at this point.*

We may even extend this result to one-sided derivatives:

Corollary 7.6.4 *Assume $f \in D$ 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. Then the Fourier series $s_N(x)$ converges to $f(x)$ at this point.

Exercises to Section 7.6

1. Show that the Fourier series $\sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx)$ in Example 7.1.1 converges to $f(x) = x$ for $x \in (-\pi, \pi)$. What happens in the endpoints?

2. Prove Corollary 7.6.3

3. Prove Corollary 7.6.4

4. 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

$$\int_{-\infty}^{\infty} \frac{\sin x}{x}\, dx$$

5. 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

$$P_r(x) = \frac{1 - r^2}{1 - 2r\cos x + r^2}$$

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

$$\lim_{r\uparrow 1} \sum_{n=-\infty}^{\infty} c_n r^{|n|} e^{inx} = f(x)$$

7.7 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 7.7.1 *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*

$$g(x) = \int_0^x f(t)\,dt = \alpha_0 x + \sum_{n\in\mathbb{Z}, n\neq 0} -\frac{i\alpha_n}{n}\left(e^{inx}-1\right)$$

where the convergence of the series is uniform.

Proof: By Cauchy-Schwarz's inequality we have

$$|g(x)-t_n(x)| = |\int_0^{\pi} (f(t)-s_n(t))\,dt| \leq \int_{-\pi}^{\pi} |f(t)-s_n(t)|\,dt \leq$$

$$\leq 2\pi\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |f(s)-s_n(s)|\cdot 1\,ds\right) = 2\pi\langle |f-s_n|, 1\rangle \leq$$

$$\leq 2\pi \|f-s_n\|_2 \|1\|_2 = 2\pi\|f-s_n\|_2$$

By Theorem 7.2.6, we see that $\|f - s_n\|_2 \to 0$, and hence t_n converges uniformly to $g(x)$. □

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 7.1, 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 7.7.2 *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 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

$$\beta_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f'(t) e^{-int}\, dt = \frac{1}{2\pi} \left[f(t) e^{-int} \right]_{-\pi}^{\pi} - \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)(-ine^{-int})\, dt =$$

$$= 0 + in \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-int}\, dt = in\alpha_n$$

which shows that $\sum_{n=-\infty}^{\infty} in\alpha_n e^{inx}$ is the Fourier series of f'. The convergence follows from Corollary 7.6.3. □

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.

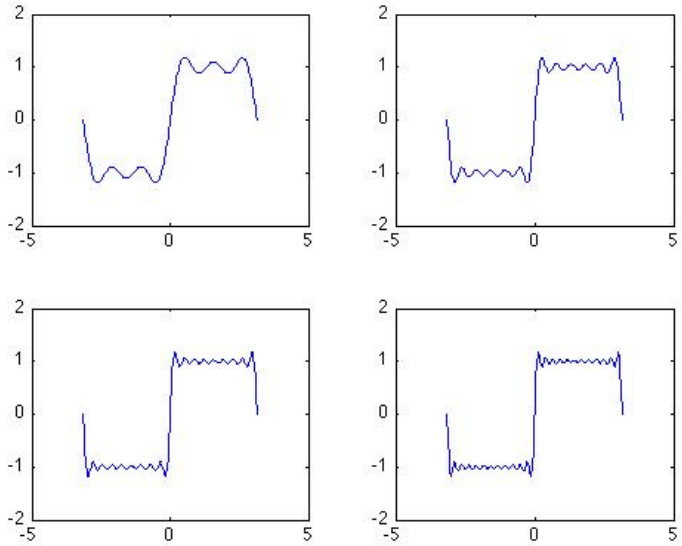
Exercises for Section 7.7

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 7.7.1).

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 *Gibbs' phenomenon*. Let $f : [-\pi, \pi] \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} -1 & \text{for } x < 0 \\ 0 & \text{for } x = 0 \\ 1 & \text{for } x > 1 \end{cases}$$



The figure above 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 solution of problem 3.

a) Show that the partial sums can be expressed as

$$s_{2n-1}(x) = \frac{4}{\pi} \sum_{k=1}^{n} \frac{\sin((2k-1)x)}{2k-1}$$

b) Use problem 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 $x_+ = \frac{\pi}{2n}$.

d) Show that

$$s_{2n-1}(\pm\frac{\pi}{2n}) = \pm\frac{4}{\pi} \sum_{k=1}^{n} \frac{\sin \frac{(2k-1)\pi}{2n}}{2k-1}$$

e) Show that $s_{2n-1}(\pm\frac{\pi}{2n}) \to \pm\frac{2}{\pi} \int_0^\pi \frac{\sin x}{x}\, dx$ 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. Gibbs showed that this number holds in general for functions in D.