Introduction to Real Analysis

Supplementary notes for MATH/MTHE 281

Andrew D. Lewis

This version: 2018/01/09

This version: 2018/01/09

Table of Contents

Chapter 1

Set theory and terminology

The principle purpose of this chapter is to introduce the mathematical notation and language that will be used in the remainder of these volumes. Much of this notation is standard, or at least the notation we use is generally among a collection of standard possibilities. In this respect, the chapter is a simple one. However, we also wish to introduce the reader to some elementary, although somewhat abstract, mathematics. The secondary objective behind this has three components.

1. We aim to provide a somewhat rigorous foundation for what follows. This means being fairly clear about defining the (usually) somewhat simple concepts that arise in the chapter. Thus "intuitively clear" concepts like sets, subsets, maps, etc., are given a fairly systematic and detailed discussion. It is at least interesting to know that this can be done. And, if it is not of interest, it can be sidestepped at a first reading.
2. This chapter contains some results, and many of these require very simple proofs. We hope that these simple proofs might be useful to readers who are new to the world where everything is proved. Proofs in other chapters in these volumes may not be so useful for achieving this objective.
3. The material is standard mathematical material, and should be known by anyone purporting to love mathematics.

Do I need to read this chapter? Readers who are familiar with standard mathematical notation (e.g., who understand the symbols $\in$, $\subseteq$, $\cup$, $\cap$, $\times$, $f\colon S \to T$, $\mathbb{Z}_{>0}$, and $\mathbb{Z}$) can simply skip this chapter in its entirety. Some ideas (e.g., relations, orders, Zorn's Lemma) may need to be referred to during the course of later chapters, but this is easily done.

Readers not familiar with the above standard mathematical notation will have some work to do. They should certainly read Sections 1.1, 1.2, and 1.3 closely enough that they understand the language, notation, and main ideas. And they should read enough of Section 1.4 that they know what objects, familiar to them from their being human, the symbols $\mathbb{Z}_{>0}$ and $\mathbb{Z}$ refer to. The remainder of the material can be overlooked until it is needed later. •

Contents

Section 1.1

Sets

The basic ingredient in modern mathematics is the set. The idea of a set is familiar to everyone at least in the form of "a collection of objects." In this section, we shall not really give a definition of a set that goes beyond that intuitive one. Rather we shall accept this intuitive idea of a set, and move forward from there. This way of dealing with sets is called ***naïve set theory***. There are some problems with naïve set theory, as described in Section 1.8.1, and these lead to a more formal notion of a set as an object that satisfies certain axioms, those given in Section 1.8.2. However, these matters will not concern us much at the moment.

Do I need to read this section? Readers familiar with basic set theoretic notation can skip this section. Other readers should read it, since it contains language, notation, and ideas that are absolutely commonplace in these volumes. •

1.1.1 Definitions and examples

First let us give our working definition of a set. A ***set*** is, for us, a well-defined collection of objects. Thus one can speak of everyday things like "the set of red-haired ladies who own yellow cars." Or one can speak of mathematical things like "the set of even prime numbers." Sets are therefore defined by describing their ***members*** or ***elements***, i.e., those objects that are in the set. When we are feeling less formal, we may refer to an element of a set as a ***point*** in that set. The set with no members is the ***empty set***, and is denoted by $\emptyset$. If S is a set with member x, then we write $x \in S$. If an object x is *not* in a set S, then we write $x \notin S$.

1.1.1 Examples (Sets)

1. If S is the set of even prime numbers, then $2 \in S$.
2. If S is the set of even prime numbers greater than 3, then S is the empty set.
3. If S is the set of red-haired ladies who own yellow cars and if x = Ghandi, then $x \notin S$. •

If it is possible to write the members of a set, then they are usually written between braces { }. For example, the set of prime numbers less that 10 is written as $\{2, 3, 5, 7\}$ and the set of physicists to have won a Fields Prize as of 2005 is {Edward Witten}.

A set S is a ***subset*** of a set T if $x \in S$ implies that $x \in T$. We shall write $S \subseteq T$, or equivalently $T \supseteq S$, in this case. If $x \in S$, then the set $\{x\} \subseteq S$ with one element, namely x, is a ***singleton***. Note that x and $\{x\}$ are different things. For example, $x \in S$ and $\{x\} \subseteq S$. If $S \subseteq T$ and if $T \subseteq S$, then the sets S and T are ***equal***, and we write $S = T$. If two sets are not equal, then we write $S \neq T$. If $S \subseteq T$ and if $S \neq T$, then S is a ***proper*** or ***strict*** subset of T, and we write $S \subset T$ if we wish to emphasise this fact.

1.1.2 Notation (Subsets and proper subsets) We adopt a particular convention for denoting subsets and proper subsets. That is, we write $S \subseteq T$ when S is a subset of T, allowing for the possibility that $S = T$. When $S \subseteq T$ and $S \neq T$ we write $S \subset T$. In this latter case, many authors will write $S \subsetneq T$. We elect not to do this. The convention we use is consistent with the convention one normally uses with inequalities. That is, one normally writes $x \leq y$ and $x < y$. It is not usual to write $x \leqslant y$ in the latter case. •

Some of the following examples may not be perfectly obvious, so may require sorting through the definitions.

1.1.3 Examples (Subsets)

1. For any set S, $\emptyset \subseteq S$ (see Exercise 1.1.1).
2. $\{1, 2\} \subseteq \{1, 2, 3\}$.
3. $\{1, 2\} \subset \{1, 2, 3\}$.
4. $\{1, 2\} = \{2, 1\}$.
5. $\{1, 2\} = \{2, 1, 2, 1, 1, 2\}$. •

A common means of defining a set is to define it as the subset of an existing set that satisfies conditions. Let us be slightly precise about this. A ***one-variable predicate*** is a statement which, in order that its truth be evaluated, needs a single argument to be specified. For example, $P(x)$ = "x is blue" needs the single argument x in order that it be decided whether it is true or not. We then use the notation

$$\{x \in S \mid P(x)\}$$

to denote the members x of S for which the predicate P is true when evaluated at x. This is read as something like, "the set of x's in S such that $P(x)$ holds."

For sets S and T, the ***relative complement*** of T in S is the set

$$S - T = \{x \in S \mid x \notin T\}.$$

Note that for this to make sense, we do not require that T be a subset of S. It is a common occurrence when dealing with complements that one set be a subset of another. We use different language and notation to deal with this. If S is a set and if $T \subseteq S$, then $S \setminus T$ denotes the ***absolute complement*** of T in S, and is defined by

$$S \setminus T = \{x \in S \mid x \notin T\}.$$

Note that, if we forget that T is a subset of S, then we have $S \setminus T = S - T$. Thus $S - T$ is the more general notation. Of course, if $A \subseteq T \subseteq S$, one needs to be careful when using the words "absolute complement of A," since one must say whether one is taking the complement in T or the larger complement in S. For this reason, we prefer the notation we use rather the commonly encountered notation A^{C} or A' to refer to the absolute complement. Note that one should not talk about the absolute complement to a set, without saying within which subset the complement is being

taken. To do so would imply the existence of "a set containing all sets," an object that leads one to certain paradoxes (see Section 1.8).

A useful set associated with every set S is its ***power set***, by which we mean the set

$$\mathbf{2}^S = \{A \mid A \subseteq S\}.$$

The reader can investigate the origins of the peculiar notation in Exercise 1.1.3.

1.1.2 Unions and intersections

In this section we indicate how to construct new sets from existing ones.

Given two sets S and T, the ***union*** of S and T is the set $S \cup T$ whose members are members of S *or* T. The ***intersection*** of S and T is the set $S \cap T$ whose members are members of S *and* T. If two sets S and T have the property that $S \cap T = \emptyset$, then S and T are said to be ***disjoint***. For sets S and T their ***symmetric complement*** is the set

$$S \triangle T = (S - T) \cup (T - S).$$

Thus $S \triangle T$ is the set of objects in union $S \cup T$ that do not lie in the intersection $S \cap T$. The symmetric complement is so named because $S \triangle T = T \triangle S$. In Figure 1.1 we

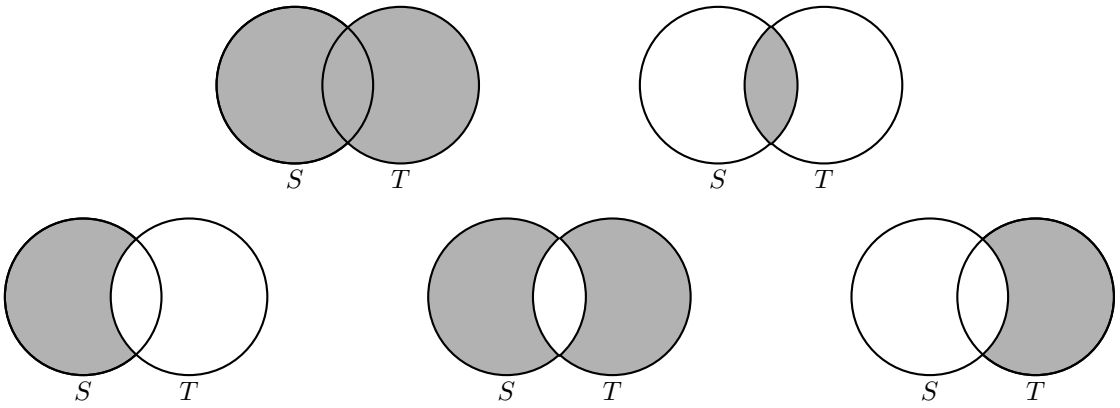


Figure 1.1 $S \cup T$ (top left), $S \cap T$ (top right), $S - T$ (bottom left), $S \triangle T$ (bottom middle), and $T - S$ (bottom right)

give Venn diagrams describing union, intersection, and symmetric complement.

The following result gives some simple properties of pairwise unions and intersections of sets. We leave the straightforward verification of some or all of these to the reader as Exercise 1.1.5.

1.1.4 Proposition (Properties of unions and intersections) *For sets* S *and* T, *the following statements hold:*

(i) $S \cup \emptyset = S$;

(ii) $S \cap \emptyset = \emptyset$;

(iii) $S \cup S = S$;

(iv) $S \cap S = S$;
(v) $S \cup T = T \cup S$ (***commutativity***);
(vi) $S \cap T = T \cap S$ (***commutativity***);
(vii) $S \subseteq S \cup T$;
(viii) $S \cap T \subseteq S$;
(ix) $S \cup (T \cup U) = (S \cup T) \cup U$ (***associativity***);
(x) $S \cap (T \cap U) = (S \cap T) \cap U$ (***associativity***);
(xi) $S \cap (T \cup U) = (S \cap T) \cup (S \cap U)$ (***distributivity***);
(xii) $S \cup (T \cap U) = (S \cup T) \cap (S \cup U)$ (***distributivity***).

We may more generally consider not just two sets, but an arbitrary collection $\mathcal{S}$ of sets. In this case we *posit* the existence of a set, called the ***union*** of the sets $\mathcal{S}$, with the property that it contains each element of each set $S \in \mathcal{S}$. Moreover, one can specify the subset of this big set to *only* contain members of sets from $\mathcal{S}$. This set we will denote by $\cup_{S \in \mathcal{S}} S$. We can also perform a similar construction with intersections of an arbitrary collection $\mathcal{S}$ of sets. Thus we denote by $\cap_{S \in \mathcal{S}} S$ the set, called the ***intersection*** of the sets $\mathcal{S}$, having the property that $x \in \cap_{S \in \mathcal{S}} S$ if $x \in S$ for every $S \in \mathcal{S}$. Note that we do not need to posit the existence of the intersection.

Let us give some properties of general unions and intersections as they relate to complements.

1.1.5 Proposition (De Morgan's[1] Laws) *Let* T *be a set and let* $\mathcal{S}$ *be a collection of subsets of* T. *Then the following statements hold:*

(i) $T \setminus (\cup_{S \in \mathcal{S}} S) = \cap_{S \in \mathcal{S}} (T \setminus S)$;
(ii) $T \setminus (\cap_{S \in \mathcal{S}} S) = \cup_{S \in \mathcal{S}} (T \setminus S)$.

Proof (i) Let $x \in T \setminus (\cup_{S \in \mathcal{S}})$. Then, for each $S \in \mathcal{S}$, $x \notin S$, or $x \in T \setminus S$. Thus $x \in \cap_{S \in \mathcal{S}} (T \setminus S)$. Therefore, $T \setminus (\cup_{S \in \mathcal{S}}) \supseteq \cap_{S \in \mathcal{S}} (T \setminus S)$. Conversely, if $x \in \cap_{S \in \mathcal{S}} (T \setminus S)$, then, for each $S \in \mathcal{S}$, $x \notin S$. Therefore, $x \notin \cup_{S \in \mathcal{S}}$. Therefore, $x \in T \setminus (\cup_{S \in \mathcal{S}})$, thus showing that $\cap_{S \in \mathcal{S}} (T \setminus S) \subseteq T \setminus (\cup_{S \in \mathcal{S}})$. It follows that $T \setminus (\cup_{S \in \mathcal{S}}) = \cap_{S \in \mathcal{S}} (T \setminus S)$.

(ii) This follows in much the same manner as part (i), and we leave the details to the reader. ■

1.1.6 Remark (Showing two sets are equal) Note that in proving part (i) of the preceding result, we proved two things. First we showed that $T \setminus (\cup_{S \in \mathcal{S}}) \subseteq \cap_{S \in \mathcal{S}} (T \setminus S)$ and then we showed that $\cap_{S \in \mathcal{S}} (T \setminus S) \subseteq T \setminus (\cup_{S \in \mathcal{S}})$. This is the standard means of showing that two sets are equal; first show that one is a subset of the other, and then show that the other is a subset of the one. •

For general unions and intersections, we also have the following generalisation of the distributive laws for unions and intersections. We leave the straightforward proof to the reader (Exercise 1.1.6)

[1] Augustus De Morgan (1806–1871) was a British mathematician whose principal mathematical contributions were to analysis and algebra.

1.1.7 Proposition (Distributivity laws for general unions and intersections) *Let* T *be a set and let* $\mathcal{S}$ *be a collection of sets. Then the following statements hold:*

(i) $\mathrm{T} \cap (\cup_{\mathrm{S} \in \mathcal{S}} \mathrm{S}) = \cup_{\mathrm{S} \in \mathcal{S}} (\mathrm{T} \cap \mathrm{S})$*;*

(ii) $\mathrm{T} \cup (\cap_{\mathrm{S} \in \mathcal{S}} \mathrm{S}) = \cap_{\mathrm{S} \in \mathcal{S}} (\mathrm{T} \cup \mathrm{S})$*.*

There is an alternative notion of the union of sets, one that retains the notion of membership in the original set. The issue that arises is this. If $S = \{1, 2\}$ and $T = \{2, 3\}$, then $S \cup T = \{1, 2, 3\}$. Note that we lose with the usual union the fact that 1 is an element of S only, but that 2 is an element of both S and T. Sometimes it is useful to retain these sorts of distinctions, and for this we have the following definition.

1.1.8 Definition (Disjoint union) ***missing stuff*** For sets S and T, their ***disjoint union*** is the set

$$S \mathbin{\mathring{\cup}} T = \{(S, x) \mid x \in S\} \cup \{(T, y) \mid y \in T\}. \qquad \bullet$$

Let us see how the disjoint union differs from the usual union.

1.1.9 Example (Disjoint union) Let us again take the simple example $S = \{1, 2\}$ and $T = \{2, 3\}$. Then $S \cup T = \{1, 2, 3\}$ and

$$S \mathbin{\mathring{\cup}} T = \{(S, 1), (S, 2), (T, 2), (T, 3)\}.$$

We see that the idea behind writing an element in the disjoint union as an ordered pair is that the first entry in the ordered pair simply keeps track of the set from which the element in the disjoint union was taken. In this way, if $S \cap T \neq \emptyset$, we are guaranteed that there will be no "collapsing" when the disjoint union is formed. •

1.1.3 Finite Cartesian products

As we have seen, if S is a set and if $x_1, x_2 \in S$, then $\{x_1, x_2\} = \{x_2, x_1\}$. There are times, however, when we wish to keep track of the order of elements in a set. To accomplish this and other objectives, we introduce the notion of an ordered pair. First, however, in order to make sure that we understand the distinction between ordered and unordered pairs, we make the following definition.

1.1.10 Definition (Unordered pair) If S is a set, an ***unordered pair*** from S is any subset of S with two elements. The collection of unordered pairs from S is denoted by $S^{(2)}$. •

Obviously one can talk about unordered collections of more than two elements of a set, and the collection of subsets of a set S comprised of k elements is denoted by $S^{(k)}$ and called the set of ***unordered* k*-tuples***.

With the simple idea of an unordered pair, the notion of an ordered pair is more distinct.

1.1.11 Definition (Ordered pair and Cartesian product) Let S and T be sets, and let $x \in S$ and $y \in T$. The ***ordered pair*** of x and y is the set $(x, y) = \{\{x\}, \{x, y\}\}$. The ***Cartesian product*** of S and T is the set

$$S \times T = \{(x, y) \mid x \in S,\ y \in T\}. \qquad \bullet$$

The definition of the ordered pair seems odd at first. However, it is as it is to secure the objective that if two ordered pairs (x_1, y_1) and (x_2, y_2) are equal, then $x_1 = x_2$ and $y_1 = y_2$. The reader can check in Exercise 1.1.8 that this objective is in fact achieved by the definition. It is also worth noting that the form of the ordered pair as given in the definition is seldom used after its initial introduction.

Clearly one can define the Cartesian product of any finite number of sets $S_1, \ldots, S_k$ inductively. Thus, for example, $S_1 \times S_2 \times S_3 = (S_1 \times S_2) \times S_3$. Note that, according to the notation in the definition, an element of $S_1 \times S_2 \times S_3$ should be written as $((x_1, x_2), x_3)$. However, it is immaterial that we define $S_1 \times S_2 \times S_3$ as we did, or as $S_1 \times S_2 \times S_3 = S_1 \times (S_2 \times S_3)$. Thus we simply write elements in $S_1 \times S_2 \times S_3$ as (x_1, x_2, x_3), and similarly for a Cartesian product $S_1 \times \cdots \times S_k$. The Cartesian product of a set with itself k-times is denoted by S^k. That is,

$$S^k = \underbrace{S \times \cdots \times S}_{k\text{-times}}.$$

In Section 1.6.2 we shall indicate how to define Cartesian products of more than finite collections of sets.

Let us give some simple examples.

1.1.12 Examples (Cartesian products)

1. If S is a set then note that $S \times \emptyset = \emptyset$. This is because there are no ordered pairs from S and $\emptyset$. It is just as clear that $\emptyset \times S = \emptyset$. It is also clear that, if $S \times T = \emptyset$, then either $S = \emptyset$ or $T = \emptyset$.
2. If $S = \{1, 2\}$ and $T = \{2, 3\}$, then

$$S \times T = \{(1, 2), (1, 3), (2, 2), (2, 3)\}. \qquad \bullet$$

Cartesian products have the following properties.

1.1.13 Proposition (Properties of Cartesian product) *For sets* S, T, U, *and* V, *the following statements hold:*

(i) $(\mathrm{S} \cup \mathrm{T}) \times \mathrm{U} = (\mathrm{S} \times \mathrm{U}) \cup (\mathrm{T} \times \mathrm{U})$;

(ii) $(\mathrm{S} \cap \mathrm{U}) \times (\mathrm{T} \cap \mathrm{V}) = (\mathrm{S} \times \mathrm{T}) \cap (\mathrm{U} \times \mathrm{V})$;

(iii) $(\mathrm{S} - \mathrm{T}) \times \mathrm{U} = (\mathrm{S} \times \mathrm{U}) - (\mathrm{T} \times \mathrm{U})$.

Proof Let us prove only the first identity, leaving the remaining two to the reader. Let $(x, u) \in (S \cup T) \times U$. Then $x \in S \cup T$ and $u \in U$. Therefore, x is an element of at least one of S and T. Without loss of generality, suppose that $x \in S$. Then $(x, u) \in S \times U$ and so $(x, u) \in (S \times U) \cup (T \times U)$. Therefore, $(S \cup T) \times U = (S \times U) \cup (T \times U)$. Conversely, suppose that $(x, u) \in (S \times U) \cup (T \times U)$. Without loss of generality, suppose that $(x, u) \in S \times U$. Then $x \in S \subseteq S \cup T$ and $u \in U$. Therefore, $(x, u) \in (S \cup T) \times U$. Thus $(S \times U) \cup (T \times U) \subseteq (S \cup T) \times U$, giving the result. ■

1.1.14 Remark ("Without loss of generality") In the preceding proof, we twice employed the expression "without loss of generality." This is a commonly encountered expression, and is frequently used in one of the following two contexts. The first, as above, indicates that one is making an arbitrary selection, but that were another arbitrary selection to have been made, the same argument holds. This is a more or less straightforward use of "without loss of generality." A more sophisticated use of the expression might indicate that one is making a simplifying assumption, and that this is okay, because it can be shown that the general case follows easily from the simpler one. The trick is to then understand *how* the general case follows from the simpler one, and this can sometimes be nontrivial, depending on the willingness of the writer to describe this process. •

Exercises

1.1.1 Prove that the empty set is a subset of every set.
Hint: *Assume the converse and arrive at an absurdity.*

1.1.2 Let S be a set, let $A, B, C \subseteq S$, and let $\mathscr{A}, \mathscr{B} \subseteq \mathbf{2}^S$.

(a) Show that $A \triangle \emptyset = A$.

(b) Show that $(S \setminus A) \triangle (S \setminus B) = A \triangle B$.

(c) Show that $A \triangle C \subseteq (A \triangle B) \cup (B \triangle C)$.

(d) Show that

$$\Big(\cup_{A \in \mathscr{A}} A\Big) \triangle \Big(\cup_{B \in \mathscr{B}} B\Big) \subseteq \cup_{(A,B) \in \mathscr{A} \times \mathscr{B}} (A \triangle B),$$
$$\Big(\cap_{A \in \mathscr{A}} A\Big) \triangle \Big(\cap_{B \in \mathscr{B}} B\Big) \subseteq \cap_{(A,B) \in \mathscr{A} \times \mathscr{B}} (A \triangle B),$$
$$\cap_{(A,B) \in \mathscr{A} \times \mathscr{B}} (A \triangle B) \subseteq \Big(\cap_{A \in \mathscr{A}} A\Big) \triangle \Big(\cup_{B \in \mathscr{B}} B\Big).$$

1.1.3 If S is a set with n members, show that $\mathbf{2}^S$ is a set with 2^n members.

1.1.4 Let S be a set with m elements. Show that the number of subsets of S having k distinct elements is $\binom{m}{k} = \frac{m!}{k!(m-k)!}$.

1.1.5 Prove as many parts of Proposition 1.1.4 as you wish.

1.1.6 Prove Proposition 1.1.7.

1.1.7 Let S be a set with n members and let T be a set with m members. Show that $S \overset{\circ}{\cup} T$ is a set with nm members.

1.1.8 Let S and T be sets, let $x_1, x_2 \in S$, and let $y_1, y_2 \in T$. Show that $(x_1, y_1) = (x_2, y_2)$ if and only if $x_1 = x_2$ and $y_1 = y_2$.

Section 1.2

Relations

Relations are a fundamental ingredient in the description of many mathematical ideas. One of the most valuable features of relations is that they allow many useful constructions to be explicitly made only using elementary ideas from set theory.

Do I need to read this section? The ideas in this section will appear in many places in the series, so this material should be regarded as basic. However, readers looking to proceed with minimal background can skip the section, referring back to it when needed. •

1.2.1 Definitions

We shall describe in this section "binary relations," or relations between elements of two sets. It is possible to define more general sorts of relations where more sets are involved. However, these will not come up for us.

1.2.1 Definition (Relation) A ***binary relation from* S *to* T** (or simply a ***relation from* S *to* T**) is a subset of $S \times T$. If $R \subseteq S \times T$ and if $(x, y) \in R$, then we shall write $x\,R\,y$, meaning that x and y are related by R. A relation from S to S is a ***relation in* S**. •

The definition is simple. Let us give some examples to give it a little texture.

1.2.2 Examples (Relations)

1. Let S be the set of husbands and let T be the set of wives. Define a relation R from S to T by asking that $(x, y) \in R$ if x is married to y. Thus, to say that x and y are related in this case means to say that x is married to y.
2. Let S be a set and consider the relation R in the power set $\mathbf{2}^S$ of S given by

$$R = \{(A, B) \mid A \subseteq B\}.$$

   Thus A is related to B if A is a subset of B.
3. Let S be a set and define a relation R in S by

$$R = \{(x, x) \mid x \in S\}.$$

   Thus, under this relation, two members in S are related if and only if they are equal.
4. Let S be the set of integers, let k be a positive integer, and define a relation R_k in S by

$$R_k = \{(n_1, n_2) \mid n_1 - n_2 = k\}.$$

   Thus, if $n \in S$, then all integers of the form $n + mk$ for an integer m are related to n. •

1.2.3 Remark ("If" versus "if and only if") In part 3 of the preceding example we used the expression "if and only if" for the first time. It is, therefore, worth saying a few words about this commonly used terminology. One says that statement A holds "if and only if" statement B holds to mean that statements A and B are exactly equivalent. Typically, this language arises in theorem statements. In proving such theorems, it is important to note that one must prove *both* that statement A implies statement B *and* that statement B implies statement A.

To confuse matters, when stating a definition, the convention is to use "if" rather than "if and only if". It is not uncommon to see "if and only if" used in definitions, the thinking being that a definition makes the thing being defined as equivalent to what it is defined to be. However, there is a logical flaw here. Indeed, suppose one is defining "X" to mean that "Proposition A applies". If one writes "X if and only if Proposition A applies" then this makes no sense. Indeed the "only if" part of this statement says that the statement "Proposition A applies" if "X" holds. But "X" is undefined except by saying that it holds when "Proposition A applies". •

In the next section we will encounter the notion of the inverse of a function; this idea is perhaps known to the reader. However, the notion of inverse also applies to the more general setting of relations.

1.2.4 Definition (Inverse of a relation) If $R \subseteq S \times T$ is a relation from S to T, then the ***inverse*** of R is the relation R^{-1} from T to S defined by

$$R^{-1} = \{(y, x) \in T \times S \mid (x, y) \in R\}.$$ •

There are a variety of properties that can be bestowed upon relations to ensure they have certain useful attributes. The following is a partial list of such properties.

1.2.5 Definition (Properties of relations) Let S be a set and let R be a relation in S. The relation R is:

(i) ***reflexive*** if $(x, x) \in R$ for each $x \in S$;
(ii) ***irreflexive*** if $(x, x) \notin R$ for each $x \in S$;
(iii) ***symmetric*** if $(x_1, x_2) \in R$ implies that $(x_2, x_1) \in R$;
(iv) ***antisymmetric*** if $(x_1, x_2) \in R$ and $(x_2, x_1) \in R$ implies that $x_1 = x_2$;
(v) ***transitive*** if $(x_1, x_2) \in R$ and $(x_2, x_3) \in R$ implies that $(x_1, x_3) \in R$. •

1.2.6 Examples (Example 1.2.2 cont'd)

1. The relation of inclusion in the power set $\mathbf{2}^S$ of a set S is reflexive, antisymmetric, and transitive.
2. The relation of equality in a set S is reflexive, symmetric, antisymmetric, and transitive.
3. The relation R_k in the set S of integers is reflexive, symmetric, and transitive. •

1.2.2 Equivalence relations

In this section we turn our attention to an important class of relations, and we indicate why these are important by giving them a characterisation in terms of a decomposition of a set.

1.2.7 Definition (Equivalence relation, equivalence class) An ***equivalence relation*** in a set S is a relation R that is reflexive, symmetric, and transitive. For $x \in S$, the set of elements of S related to x is denoted by $[x]$, and is the ***equivalence class*** of x with respect to R. An element x' in an equivalence class $[x]$ is a ***representative*** of that equivalence class. The set of equivalence classes is denoted by S/R (typically pronounced as **S** ***modulo*** **R**). •

It is common to denote that two elements $x_1, x_2 \in S$ are related by an equivalence relation by writing $x_1 \sim x_2$. Of the relations defined in Example 1.2.2, we see that those in parts 3 and 4 are equivalence relations, but that in part 2 is not.

Let us now characterise equivalence relations in a more descriptive manner. We begin by defining a (perhaps seemingly unrelated) notion concerning subsets of a set.

1.2.8 Definition (Partition of a set) A ***partition*** of a set S is a collection $\mathscr{A}$ of subsets of S having the properties that

(i) two distinct subsets in $\mathscr{A}$ are disjoint and

(ii) $S = \cup_{A \in \mathscr{A}} A$. •

We now prove that there is an exact correspondence between equivalence classes associated to an equivalence relation.

1.2.9 Proposition (Equivalence relations and partitions) *Let* S *be a set and let* R *be an equivalence relation in* S*. Then the set of equivalence classes with respect to* R *is a partition of* S*.*

Conversely, if $\mathscr{A}$ is a partition of S*, then the relation*

$$\{(x_1, x_2) \mid x_1, x_2 \in A \text{ for some } A \in \mathscr{A}\}$$

is an equivalence relation in S*.*

Proof We first claim that two distinct equivalence classes are disjoint. Thus we let $x_1, x_2 \in S$ and suppose that $[x_1] \neq [x_2]$. Suppose that $x \in [x_1] \cap [x_2]$. Then $x \sim x_1$ and $x \sim x_2$, or, by transitivity of R, $x_1 \sim x$ and $x \sim x_2$. By transitivity of R, $x_1 \sim x_2$, contradicting the fact that $[x_1] \neq [x_2]$. To show that S is the union of its equivalence classes, merely note that, for each $x \in S$, $x \in [x]$ by reflexivity of R.

Now let $\mathscr{A}$ be a partition and defined R as in the statement of the proposition. Let $x \in S$ and let A be the element of $\mathscr{A}$ that contains x. Then clearly we see that $(x, x) \in R$ since $x \in A$. Thus R is reflexive. Next let $(x_1, x_2) \in R$ and let A be the element of $\mathscr{A}$ such that $x_1, x_2 \in A$. Clearly then, $(x_2, x_1) \in R$, so R is symmetric. Finally, let $(x_1, x_2), (x_2, x_3 \in R$. Then there are elements $A_{12}, A_{23} \in \mathscr{A}$ such that $x_1, x_2 \in A_{12}$ and such that $x_2, x_3 \in A_{23}$. Since A_{12} and A_{23} have the point x_2 in common, we must have $A_{12} = A_{23}$. Thus $(x_1, x_3 \in A_{12} = A_{23}$, giving transitivity of R. ■

Exercises

1.2.1 In a set S define a relation $R = \{(x, y) \in S \times S \mid x = y\}$.

(a) Show that R is an equivalence relation.

(b) Show that $S/R = S$.

Section 1.3

Maps

Another basic concept in all of mathematics is that of a map between sets. Indeed, many of the interesting objects in mathematics are maps of some sort. In this section we review the notation associated with maps, and give some simple properties of maps.

Do I need to read this section? The material in this section is basic, and will be used constantly throughout the series. Unless you are familiar already with maps and the notation associated to them, this section is essential reading. •

1.3.1 Definitions and notation

We begin with the definition.

1.3.1 Definition (Map) For sets S and T, a ***map*** from S to T is a relation R from S to T having the property that, for each $x \in S$, there exists a unique $y \in T$ such that $(x, y) \in R$. The set S is the ***domain*** of the map and the set T is the ***codomain*** of the map. The set of maps from S to T is denoted by T^S.[2] •

By definition, a map is a relation. This is not how one most commonly thinks about a map, although the definition serves to render the concept of a map in terms of concepts we already know. Suppose one has a map from S to T defined by a relation R. Then, given $x \in S$, there is a single $y \in T$ such that x and y are related. Denote this element of T by $f(x)$, since it is defined by x. When one refers to a map, one more typically refers to the assignment of the element $f(x) \in T$ to $x \in S$. Thus one refers to the map as f, leaving aside the baggage of the relation as in the definition. Indeed, this is how we from now on will think of maps. The definition above does, however, have some use, although we alter our language, since we are now thinking of a map as an "assignment." We call the set

$$\text{graph}(f) = \{(x, f(x)) \mid x \in S\} \subseteq S \times T$$

(which we originally called the map in Definition 1.3.1) the ***graph*** of the map $f\colon S \to T$.

If one wishes to indicate a map f with domain S and codomain T, one typically writes $f\colon S \to T$ to compactly express this. If one wishes to *define* a map by saying what it does, the notation

$$\begin{aligned} f\colon S &\to T \\ x &\mapsto \text{what } x \text{ gets mapped to} \end{aligned}$$

[2] The idea behind this notation is the following. A map from S to T assigns to each point in S a point in T. If S and T are finite sets with k and l elements, respectively, then there are l possible values that can be assigned to each of the k elements of S. Thus the set of maps has l^k elements.

is sometimes helpful. Sometimes we shall write this in the text as $f\colon x \mapsto$ "what x gets mapped to". Note the distinct uses of the symbols "$\to$" and "$\mapsto$".

1.3.2 Notation (f versus f(x)) Note that a map is denoted by "f". It is quite common to see the expression "consider the map $f(x)$". Taken literally, these words are difficult to comprehend. First of all, x is unspecified. Second of all, even if x were specified, $f(x)$ is an element of T, not a map. Thus it is considered bad form mathematically to use an expression like "consider the map $f(x)$". However, there are times when it is quite convenient to use this poor notation, with an understanding that some compromises are being made. For instance, in this volume, we will be frequently dealing simultaneously with functions of both time (typically denoted by t) and frequency (typically denoted by ν). Thus it would be convenient to write "consider the map $f(t)$" when we wish to write a map that we are considering as a function of time, and similarly for frequency. Nonetheless, we shall refrain from doing this, and shall consistently use the mathematically precise language "consider the map f". •

The following is a collection of examples of maps. Some of these examples are not just illustrative, but also define concepts and notation that we will use throughout the series.

1.3.3 Examples (Maps)

1. There are no maps having $\emptyset$ as a domain or codomain since there are no elements in the empty set.
2. If S is a set and if $T \subseteq S$, then the map $i_T\colon T \to S$ defined by $i_T(x) = x$ is called the ***inclusion*** of T in S.
3. The inclusion map $i_S\colon S \to S$ of a set S into itself (since $S \subseteq S$) is the ***identity map***, and we denote it by id_S.
4. If $f\colon S \to T$ is a map and if $A \subseteq S$, then the map from A to T which assigns to $x \in A$ the value $f(x) \in T$ is called the ***restriction*** of f to A, and is denoted by $f|A\colon A \to T$.
5. If S is a set with $A \subseteq S$, then the map χ_A from S to the integers defined by
$$\chi_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A, \end{cases}$$
is the ***characteristic function*** of A.
6. If $S_1, \dots, S_k$ are sets, if $S_1 \times \cdots \times S_k$ is the Cartesian product, and if $j \in \{1, \dots, k\}$, then the map
$$\begin{aligned} \mathrm{pr}_j\colon S_1 \times \cdots \times S_j \times \cdots \times S_k &\to S_j \\ (x_1, \dots, x_j, \dots, x_k) &\mapsto x_j \end{aligned}$$
is the ***projection onto the* jth *factor***.
7. If R is an equivalence relation in a set S, then the map $\pi_R\colon S \to S/R$ defined by $\pi_R(x) = [x]$ is called the ***canonical projection*** associated to R.

8. If S, T, and U are sets and if $f\colon S \to T$ and $g\colon T \to U$ are maps, then we define a map $g \circ f\colon S \to U$ by $g \circ f(x) = g(f(x))$. This is the ***composition*** of f and g.
9. If S and $T_1, \dots, T_k$ are sets then a map $f\colon S \to T_1 \times \cdots \times T_k$ can be written as

$$f(x) = (f_1(x), \dots, f_k(x))$$

for maps $f_j\colon S \to T_j$, $j \in \{1, \dots, k\}$. In this case we will write $f = f_1 \times \cdots \times f_k$. •

Next we introduce the notions of images and preimages of points and sets.

1.3.4 Definition (Image and preimage) Let S and T be sets and let $f\colon S \to T$ be a map.

(i) If $A \subseteq S$, then $f(A) = \{f(x) \mid x \in A\}$.

(ii) The ***image*** of f is the set $\text{image}(f) = f(S) \subseteq T$.

(iii) If $B \subseteq T$, then $f^{-1}(B) = \{x \in S \mid f(x) \in B\}$ is the ***preimage*** of B under f. If $B = \{y\}$ for some $y \in T$, then we shall often write $f^{-1}(y)$ rather that $f^{-1}(\{y\})$. •

Note that one can think of f as being a map from $\mathbf{2}^S$ to $\mathbf{2}^T$ and of f^{-1} as being a map from $\mathbf{2}^T$ to $\mathbf{2}^S$. Here are some elementary properties of f and f^{-1} thought of in this way.

1.3.5 Proposition (Properties of images and preimages) *Let* S *and* T *be sets, let* $\mathrm{f}\colon \mathrm{S} \to \mathrm{T}$ *be a map, let* $\mathrm{A} \subseteq \mathrm{S}$ *and* $\mathrm{B} \subseteq \mathrm{T}$*, and let* $\mathscr{A}$ *and* $\mathscr{B}$ *be collections of subsets of* S *and* T*, respectively. Then the following statements hold:*

(i) $\mathrm{A} \subseteq \mathrm{f}^{-1}(\mathrm{f}(\mathrm{A}))$;

(ii) $\mathrm{f}(\mathrm{f}^{-1}(\mathrm{B})) \subseteq \mathrm{B}$;

(iii) $\cup_{\mathrm{A}\in\mathscr{A}}\mathrm{f}(\mathrm{A}) = \mathrm{f}(\cup_{\mathrm{A}\in\mathscr{A}}\mathrm{A})$;

(iv) $\cup_{\mathrm{B}\in\mathscr{B}}\mathrm{f}^{-1}(\mathrm{B}) = \mathrm{f}^{-1}(\cup_{\mathrm{B}\in\mathscr{B}}\mathrm{B})$;

(v) $\cap_{\mathrm{A}\in\mathscr{A}}\mathrm{f}(\mathrm{A}) = \mathrm{f}(\cap_{\mathrm{A}\in\mathscr{A}}\mathrm{A})$;

(vi) $\cap_{\mathrm{B}\in\mathscr{B}}\mathrm{f}^{-1}(\mathrm{B}) = \mathrm{f}^{-1}(\cap_{\mathrm{B}\in\mathscr{B}}\mathrm{B})$.

Proof We shall prove only some of these, leaving the remainder for the reader to complete.

(i) Let $x \in A$. Then $x \in f^{-1}(f(x))$ since $f(x) = f(x)$.

(iii) Let $y \in \cup_{A\in\mathscr{A}} f(A)$. Then $y = f(x)$ for some $x \in \cup_{A\in\mathscr{A}} A$. Thus $y \in f(\cup_{A\in\mathscr{A}} A)$. Conversely, let $y \in f(\cup_{A\in\mathscr{A}} A)$. Then, again, $y = f(x)$ for some $x \in \cup_{A\in\mathscr{A}} A$, and so $y \in \cup_{A\in\mathscr{A}} f(A)$.

(vi) Let $x \in \cap_{B\in\mathscr{B}} f^{-1}(B)$. Then, for each $B \in \mathscr{B}$, $x \in f^{-1}(B)$. Thus $f(x) \in B$ for all $B \in \mathscr{B}$ and so $f(x) \in \cap_{B\in\mathscr{B}} B$. Thus $x \in f^{-1}(\cap_{B\in\mathscr{B}} B)$. Conversely, if $x \in f^{-1}(\cap_{B\in\mathscr{B}} B)$, then $f(x) \in B$ for each $B \in \mathscr{B}$. Thus $x \in f^{-1}(B)$ for each $B \in \mathscr{B}$, or $x \in \cap_{B\in\mathscr{B}} f^{-1}(B)$. ■

1.3.2 Properties of maps

Certain basic features of maps will be of great interest.

1.3.6 Definition (Injection, surjection, bijection) Let S and T be sets. A map $f\colon S \to T$ is:

(i) ***injective***, or an ***injection***, if $f(x) = f(y)$ implies that $x = y$;

(ii) ***surjective***, or a ***surjection***, if $f(S) = T$;

(iii) ***bijective***, or a ***bijection***, if it is both injective and surjective. •

1.3.7 Remarks (One-to-one, onto, 1–1 correspondence)

1. It is not uncommon for an injective map to be said to be ***1–1*** or ***one-to-one***, and that a surjective map be said to be ***onto***. In this series, we shall exclusively use the terms injective and surjective, however. These words appear to have been given prominence by their adoption by Bourbaki (see footnote on page **??**).
2. If there exists a bijection $f\colon S \to T$ between sets S and T, it is common to say that there is a ***1–1 correspondence*** between S and T. This can be confusing if one is familiar with the expression "1–1" as referring to an injective map. The words "1–1 correspondence" mean that there is a bijection, not an injection. In case S and T are in 1–1 correspondence, we shall also say that S and T are ***equivalent***. •

Closely related to the above concepts, although not immediately obviously so, are the following notions of inverse.

1.3.8 Definition (Left-inverse, right-inverse, inverse) Let S and T be sets, and let $f\colon S \to T$ be a map. A map $g\colon T \to S$ is:

(i) a ***left-inverse*** of f if $g \circ f = \mathrm{id}_S$;

(ii) a ***right-inverse*** of f if $f \circ g = \mathrm{id}_T$;

(iii) an ***inverse*** of f if it is both a left- and a right-inverse. •

In Definition 1.2.4 we gave the notion of the inverse of a relation. Functions, being relations, also possess inverses in the sense of relations. We ask the reader to explore the relationships between the two concepts of inverse in Exercise 1.3.7.

The following result relates these various notions of inverse to the properties of injective, surjective, and bijective.

1.3.9 Proposition (Characterisation of various inverses) *Let* S *and* T *be sets and let* f: S → T *be a map. Then the following statements hold:*

(i) f *is injective if and only if it possesses a left-inverse;*

(ii) f *is surjective if and only if it possess a right-inverse;*

(iii) f *is bijective if and only if it possesses an inverse;*

(iv) there is at most one inverse for f*;*

(v) if f *possesses a left-inverse and a right-inverse, then these necessarily agree.*

Proof (i) Suppose that f is injective. For $y \in \mathrm{image}(f)$, define $g(y) = x$ where $f^{-1}(y) = \{x\}$, this being well-defined since f is injective. For $y \notin \mathrm{image}(f)$, define $g(y) = x_0$ for some $x_0 \in S$. The map g so defined is readily verified to satisfy $g \circ f = \mathrm{id}_S$, and so is a left-inverse. Conversely, suppose that f possesses a left-inverse g, and let $x_1, x_2 \in S$ satisfy $f(x_1) = f(x_2)$. Then $g \circ f(x_1) = g \circ f(x_2)$, or $x_1 = x_2$. Thus f is injective.

(ii) Suppose that f is surjective. For $y \in T$ let $x \in f^{-1}(y)$ and define $g(y) = x$.[3] With g so defined it is easy to see that $f \circ g = \mathrm{id}_T$, so that g is a right-inverse. Conversely, suppose that f possesses a right-inverse g. Now let $y \in T$ and take $x = g(y)$. Then $f(x) = f \circ g(y) = y$, so that f is surjective.

(iii) Since f is bijective, it possesses a left-inverse g_L and a right-inverse g_R. We claim that these are equal, and each is actually an inverse of f. We have

$$g_L = g_L \circ \mathrm{id}_T = g_L \circ f \circ g_R = \mathrm{id}_S \circ g_R = g_R,$$

showing equality of g_L and g_R. Thus each is a left- and a right-inverse, and therefore an inverse for f.

(iv) Let g_1 and g_2 be inverses for f. Then, just as in part (iii),

$$g_1 = g_1 \circ \mathrm{id}_T = g_1 \circ f \circ g_2 = \mathrm{id}_S \circ g_2 = g_2.$$

(v) This follows from the proof of part (iv), noting that there we only used the facts that g_1 is a left-inverse and that g_2 is a right-inverse. ■

In Figure 1.2 we depict maps that have various of the properties of injectivity,

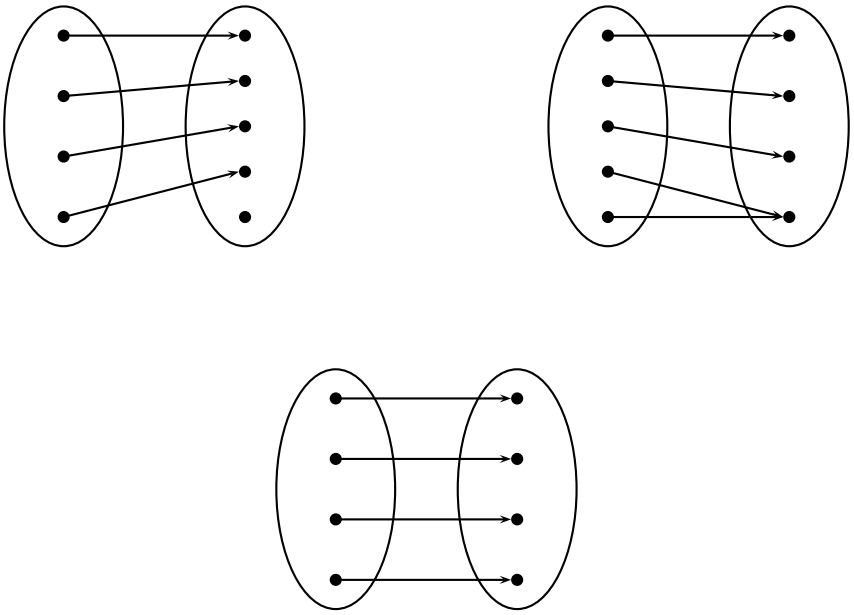


Figure 1.2 A depiction of maps that are injective but not surjective (top left), surjective but not injective (top right), and bijective (bottom)

surjectivity, or bijectivity. From these cartoons, the reader may develop some intuition for Proposition 1.3.9. In the case that $f\colon S \to T$ is a bijection, we denote its unique inverse by $f^{-1}\colon T \to S$. The confluence of the notation f^{-1} introduced when discussing preimages is not a problem, in practice.

[3] Note that the ability to choose an x from each set $f^{-1}(y)$ requires the Axiom of Choice (see Section 1.8.3).

It is worth mentioning at this point that the characterisation of left- and right-inverses in Proposition 1.3.9 is not usually very helpful. Normally, in a given setting, one will want these inverses to have certain properties. For vector spaces, for example, one may want left- or right-inverses to be linear (see ***missing stuff***), and for topological spaces, for another example, one may want a left- or right-inverse to be continuous (see Chapter **??**).

1.3.3 Graphs and commutative diagrams

Often it is useful to be able to understand the relationship between a number of maps by representing them together in a diagram. We shall be somewhat precise about what we mean by a diagram by making it a special instance of a graph. We shall encounter graphs in ***missing stuff***, although for the present purposes we merely use them as a means of making precise the notion of a commutative diagram.

First the definitions for graphs.

1.3.10 Definition (Graph) A ***graph*** is a pair (V, E) where V is a set, an element of which is called a ***vertex***, and E is a subset of the set $V^{(2)}$ of unordered pairs from V, an element of which is called an ***edge***. If $\{v_1, v_2\} \in E$ is an edge, then the vertices v_1 and v_2 are the ***endvertices*** of this edge. •

In a graph, it is the way that vertices and edges are related that is of interest. To capture this structure, the following language is useful.

1.3.11 Definition (Adjacent and incident) Let (V, E) be a graph. Two vertices $v_1, v_2 \in V$ are ***adjacent*** if $\{v_1, v_2\} \in E$ and a vertex $v \in V$ and an edge $e \in E$ are ***incident*** if there exists $v' \in V$ such that $e = \{v, v'\}$. •

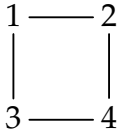
One typically represents a graph by placing the vertices in some sort of array on the page, and then drawing a line connecting two vertices if there is a corresponding edge associated with the two vertices. Some examples make this process clear.

1.3.12 Examples (Graphs)

1. Consider the graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\}.$$

There are many ways one can lay out the vertices on the page, but for this diagram, it is most convenient to arrange them in a square. Doing so gives rise to the following representation of the graph:

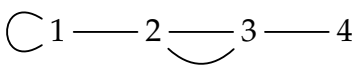


The vertices 1 and 2 are adjacent, but the vertices 1 and 4 are not. The vertex 1 and the edge $\{1, 2\}$ are incident, but the vertex 1 and the edge $\{3, 4\}$ are not.

2. For the graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{\{1, 2\}, \{2, 3\}, \{2, 3\}, \{3, 4\}\}$$

we have the representation



Note that we allow the same edge to appear twice, and we allow for an edge to connect a vertex to itself. We observe that the vertices 2 and 3 are adjacent, but the vertices 1 and 3 are not. Also, the vertex 3 and the edge $\{2, 3\}$ are incident, but the vertex 4 and the edge $\{1, 2\}$ are not. •

Often one wishes to attach "direction" to vertices. This is done with the following notion.

1.3.13 Definition (Directed graph) A ***directed graph***, or ***digraph***, is a pair (V, E) where V is a set an element of which is called a ***vertex*** and E is a subset of the set $V \times V$ of ordered pairs from V an element of which is called an ***edge***. If $e = (v_1, v_2) \in E$ is an edge, then v_1 is the ***source*** for e and v_2 is the ***target*** for e. •

Note that every directed graph is certainly also a graph, since one can assign an unordered pair to every ordered pair of vertices.

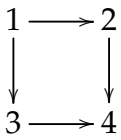
The examples above of graphs are easily turned into directed graphs, and we see that to represent a directed graph one needs only to put a "direction" on an edge, typically via an arrow.

1.3.14 Examples (Directed graphs)

1. Consider the directed graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{(1, 2), (1, 3), (2, 4), (3, 4)\}.$$

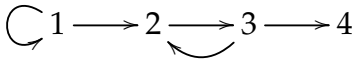
A convenient representation of this directed graph is as follows:



2. For the directed graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{(1, 1), (1, 2), (2, 3), (2, 3), (3, 4)\}$$

we have the representation



•

Of interest in graph theory is the notion of connecting two, perhaps nonadjacent, vertices with a sequence of edges (the notion of a sequence is familiar, but will be made precise in Section 1.6.3). This is made precise as follows.

1.3.15 Definition (Path)

(i) If (V, E) is a graph, a ***path*** in the graph is a sequence $(a_j)_{j\in\{1,\dots,k\}}$ in $V \cup E$ with the following properties:
   (a) $a_1, a_k \in V$;
   (b) for $j \in \{1, \dots, k-1\}$, if $a_j \in V$ (resp. $a_j \in E$), then $a_{j+1} \in E$ (resp. $a_{j+1} \in V$).

(ii) If (V, E) is a directed graph, a ***path*** in the graph is a sequence $(a_j)_{j\in\{1,\dots,k\}}$ in $V \cup E$ with the following properties:
   (a) $(a_j)_{j\in\{1,\dots,k\}}$ is a path in the graph associated to (V, E);
   (b) for $j \in \{2, \dots, k-1\}$, if $a_j \in E$, then $a_j = (a_{j-1}, a_{j+1})$.

(iii) If $(a_j)_{j\in\{1,\dots,k\}}$ is a path, the ***length*** of the path is the number of edges in the path.

(iv) For a path $(a_j)_{j\in\{1,\dots,k\}}$, the ***source*** is the vertex a_1 and the ***target*** is the vertex a_k. •

Let us give some examples of paths for graphs and for directed graphs.

1.3.16 Examples (Paths)

1. For the graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\},$$

   there are an infinite number of paths. Let us list a few:
   (a) (1), (2), (3), and (4);
   (b) $(4, \{3, 4\}, 3, \{1, 3\}, 1)$;
   (c) $(1, \{1, 2\}, 2, \{2, 4\}, 4, \{3, 4\}, 3, \{1, 3\}, 1)$;
   (d) $(1, \{1, 2\}, 2, \{1, 2\}, 1, \{1, 2\}, 2, \{1, 2\}, 1)$.
   Note that for this graph there are infinitely many paths.
2. For the directed graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{(1, 2), (1, 3), (2, 4), (3, 4)\},$$

   there are a finite number of paths:
   (a) (1), (2), (3), and (4);
   (b) $(1, (1, 2), 2)$;
   (c) $(1, (1, 2), 2, (2, 4), 4)$;
   (d) $(1, (1, 3), 3)$;
   (e) $(1, (1, 3), 3, (2, 4), 4)$;
   (f) $(2, (2, 4))$;
   (g) $(3, (3, 4), 4)$.
3. For the graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{\{1, 2\}, \{2, 3\}, \{2, 3\}, \{3, 4\}\}$$

   some examples of paths are:

(a) (1), (2), (3), and (4);
(b) $(1, \{1,2\}, 2, \{2,3\}, 3, \{2,3\}, 2, \{1,2\}, 1)$;
(c) $(4, \{3,4\}, 3)$.

There are an infinite number of paths for this graph.

4. For the directed graph (V, E) with

$$V = \{1,2,3,4\}, \quad E = \{(1,1), (1,2), (2,3), (2,3), (3,4)\}$$

some paths include:

(a) (1), (2), (3), and (4);
(b) $(1, (1,2), 2, (2,3), 3, (3,2), 2, (2,3), 3, (3,4), 4)$;
(c) $(3, (3,4), 4)$.

This directed graph has an infinite number of paths by virtue of the fact that the path $(2, (2,3), 3, (3,2), 2)$ can be repeated an infinite number of times. •

1.3.17 Notation (Notation for paths of nonzero length) For paths which contain at least one edge, i.e., which have length at least 1, the vertices in the path are actually redundant. For this reason we will often simply write a path as the sequence of edges contained in the path, since the vertices can be obviously deduced. •

There is a great deal one can say about graphs, a little of which we will say in ***missing stuff***. However, for our present purposes of defining diagrams, the notions at hand are sufficient. In the definition we employ Notation 1.3.17.

1.3.18 Definition (Diagram, commutative diagram) Let (V, E) be a directed graph.

(i) A ***diagram*** on (V, E) is a family $(S_v)_{v \in V}$ of sets associated with each vertex and a family $(f_e)_{e \in E}$ of maps associated with each edge such that, if $e = (v_1, v_2)$, then f_e has domain S_{v_1} and codomain S_{v_2}.
(ii) If $P = (e_j)_{j \in \{1,\dots,k\}}$ is a path of nonzero length in a diagram on (V, E), the ***composition*** along P is the map $f_{e_k} \circ \cdots \circ f_{e_1}$.
(iii) A diagram is ***commutative*** if, for every two vertices $v_1, v_2 \in V$ and any two paths P_1 and P_2 with source v_1 and target v_2, the composition along P_1 is equal to the composition along P_2. •

The notion of a diagram, and in particular a commutative diagram is straightforward.

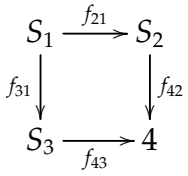
1.3.19 Examples (Diagrams and commutative diagrams)

1. Let S_1, S_2, S_3, and S_4 be sets and consider maps $f_{21} \colon S_1 \to S_2$, $f_{31} \colon S_1 \to S_3$, $f_{42} \colon S_2 \to S_4$, and $f_{43} \colon S_3 \to S_4$.[4] ***missing stuff*** Note that if we assign set S_j to j for each $j \in \{1,2,3,4\}$, then this gives a diagram on (V, E) where

$$V = \{1,2,3,4\}, \quad E = \{(1,2), (1,3), (2,4), (3,4)\}.$$

[4] It might seem more natural to write, for example, $f_{12} \colon S_1 \to S_2$ to properly represent the normal order of the domain and codomain. However, we instead write $f_{21} \colon S_1 \to S_2$ for reasons having to do with conventions that will become convenient in .

This diagram can be represented by

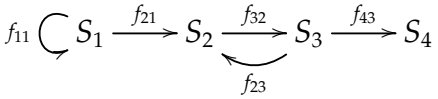


The diagram is commutative if and only if $f_{42} \circ f_{21} = f_{43} \circ f_{31}$.

2. Let S_1, S_2, S_3, and S_4 be sets and let $f_{11} \colon S_1 \to S_1$, $f_{21} \colon S_1 \to S_2$, $f_{32} \colon S_2 \to S_3$, $f_{23} \colon S_3 \to S_2$, and $f_{43} \colon S_3 \to S_4$ be maps. This data then represents a commutative diagram on the directed graph (V, E) where

$$V = \{1, 2, 3, 4\}, \quad E = \{(1, 1), (1, 2), (2, 3), (2, 3), (3, 4)\}.$$

The diagram is represented as



While it is possible to write down conditions for this diagram to be commutative, there will be infinitely many such conditions. In practice, one encounters commutative diagrams with only finitely many paths with a given source and target. This example, therefore, is not so interesting as a commutative diagram, but is more interesting as a signal flow graph, as we shall see ***missing stuff***. •

Exercises

1.3.1 Let S, T, U, and V be sets, and let $f \colon S \to T$, $g \colon T \to U$, and $h \colon U \to V$ be maps. Show that $h \circ (g \circ f) = (h \circ g) \circ f$.

1.3.2 Let S, T, and U be sets and let $f \colon S \to T$ and $g \colon T \to U$ be maps. Show that $(g \circ f)^{-1}(C) = f^{-1}(g^{-1}(C))$ for every subset $C \subseteq U$.

1.3.3 Let S and T be sets, let $f \colon S \to T$, and let $B \subseteq T$. Show that $f^{-1}(T \setminus B) = S \setminus f^{-1}(B)$.

1.3.4 If S, T, and U are sets and if $f \colon S \to T$ and $g \colon T \to U$ are bijections, then show that $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

1.3.5 Let S, T and U be sets and let $f \colon S \to T$ and $g \colon T \to U$ be maps.

(a) Show that if f and g are injective, then so too is $g \circ f$.

(b) Show that if f and g are surjective, then so too is $g \circ f$.

1.3.6 Let S and T be sets, let $f \colon S \to T$ be a map, and let $A \subseteq S$ and $B \subseteq T$. Do the following:

(a) show that if f is injective then $A = f^{-1}(f(A))$;

(b) show that if f is surjective then $f(f^{-1}(B)) = B$.

1.3.7 Let S and T be sets and let $f \colon S \to T$ be a map.

(a) Show that if f is invertible as a map, then "the relation of its inverse is the inverse of its relation." (Part of the question is to precisely understand the statement in quotes.)

(b) Show that the inverse of the relation defined by f is itself the relation associated to a function if and only if f is invertible.

1.3.8 Show that equivalence of sets, as in Remark 1.3.7–2, is an "equivalence relation"[5] on collection of all sets.

[5]The quotes are present because the notion of equivalence relation, as we have defined it, applies to sets. However, there is no set containing all sets; see Section 1.8.1.

Section 1.4

Construction of the integers

It can be supposed that the reader has some idea of what the set of integers is. In this section we actually give the set of integers a *definition*. As will be seen, this is not overly difficult to do. Moreover, the construction has little bearing on what we do. We merely present it so that the reader can be comfortable with the fact that the integers, and so subsequently the rational numbers and the real numbers (see Section 2.1), have a formal definition.

Do I need to read this section? Much of this section is not of importance in the remainder of this series. The reader should certainly know what the sets $\mathbb{Z}_{>0}$ and $\mathbb{Z}$ are. However, the details of their construction should be read only when the inclination strikes. •

1.4.1 Construction of the natural numbers

The natural numbers are the numbers 1, 2, 3, and so on, i.e., the "counting numbers." As such, we are all quite familiar with them in that we can recognise, in the absence of trickery, when we are presented with 4 of something. However, what is 4? This is what we endeavour to define in this section.

The important concept in defining the natural numbers is the following.

1.4.1 Definition (Successor) Let S be a set. The ***successor*** of S is the set $S^+ = S \cup \{S\}$. •

Thus the successor is a set whose elements are the elements of S, plus an additional element which is the set S itself. This seems, and indeed is, a simple enough idea. However, it does make possible the following definition.

1.4.2 Definition (0, 1, 2, etc.)

(i) The number ***zero***, denoted by 0, is the set $\emptyset$.

(ii) The number ***one***, denoted by 1, is the set 0^+.

(iii) The number ***two***, denoted by 2, is the set 1^+.

(iv) The number ***three***, denoted by 3, is the set 2^+.

(v) The number ***four***, denoted by 4, is the set 3^+.

This procedure can be inductively continued to define any finite nonnegative integer. •

The procedure above is well-defined, and so gives meaning to the symbol "k" where k is any nonnegative finite number. Let us give the various explicit ways of

writing the first few numbers:

$$\begin{aligned}
0 &= \emptyset, \\
1 = 0^+ &= \{0\} &&= \{\emptyset\}, \\
2 = 1^+ &= \{0, 1\} &&= \{\emptyset, \{\emptyset\}\}, \\
3 = 2^+ &= \{0, 1, 2\} &&= \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \\
4 = 3^+ &= \{0, 1, 2, 3\} &&= \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\}.
\end{aligned}$$

This settles the matter of defining any desired number. We now need to indicate how to talk about the *set* of numbers. This necessitates an assumption. As we shall see in Section 1.8.2, this assumption is framed as an axiom in axiomatic set theory.

1.4.3 Assumption There exists a set containing $\emptyset$ and all subsequent successors. •

We are now almost done. The remaining problem is that the set guaranteed by the assumption may contain more than what we want. However, this is easily remedied as follows. Let S be the set whose existence is guaranteed by Assumption 1.4.3. Define a collection $\mathscr{A}$ of subsets of S by

$$\mathscr{A} = \{A \subseteq S \mid \emptyset \in A \text{ and } n^+ \in A \text{ if } n \in A\}.$$

Note that $S \in \mathscr{A}$ so that $\mathscr{A}$ is nonempty. The following simple result is now useful.

1.4.4 Lemma *With $\mathscr{A}$ as above, if $\mathscr{B} \subseteq \mathscr{A}$, then $(\cap_{B \in \mathscr{B}} B) \in \mathscr{A}$.*

Proof For each $B \in \mathscr{B}$, $\emptyset \in B$. Thus $\emptyset \in \cap_{B \in \mathscr{B}} B$. Also let $n \in \cap_{B \in \mathscr{B}} B$. Since $n^+ \in B$ for each $B \in \mathscr{B}$, $n^+ \in \cap_{B \in \mathscr{B}} B$. Thus $(\cap_{B \in \mathscr{B}} B) \in \mathscr{A}$, as desired. ■

The lemma shows that $\cap_{A \in \mathscr{A}} A \in \mathscr{A}$. Now we have the following definition of the *set* of numbers.

1.4.5 Definition (Natural numbers) Let S and $\mathscr{A}$ be as defined above.

(i) The set $\cap_{A \in \mathscr{A}} A$ is denoted by $\mathbb{Z}_{\geq 0}$, and is the set of ***nonnegative integers***.

(ii) The set $\mathbb{Z}_{\geq 0} \setminus \{0\}$ is denoted by $\mathbb{Z}_{>0}$, and is the set of ***natural numbers***. •

1.4.6 Remark (Convention concerning $\mathbb{Z}_{>0}$ and $\mathbb{Z}_{\geq 0}$) There are two standard conventions concerning notation for nonnegative and positive integers. Neither agree with our notation. The two more or less standard bits of notation are:

1. $\mathbb{N}$ is the set of natural numbers and something else, maybe $\mathbb{Z}_{\geq 0}$, denotes the set of nonnegative integers;
2. $\mathbb{N}$ is the set of nonnegative integers (these are called the natural numbers in this scheme) and something else, maybe $\mathbb{N}^*$, denotes the set of natural numbers (called the positive natural numbers in this scheme).

Neither of these schemes is optimal on its own, and since there is no standard here, we opt for notation that is more logical. This will not cause the reader problems we hope, and may lead some to adopt our entirely sensible notation. •

Next we turn to the definition of the usual operations of arithmetic with the set $\mathbb{Z}_{\geq 0}$. That is to say, we indicate how to "add" and "multiply." First we consider addition.

1.4.7 Definition (Addition in $\mathbb{Z}_{\geq 0}$) For $k \in \mathbb{Z}_{\geq 0}$, inductively define a map $a_k \colon \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}$, called ***addition by*** $\mathbf{k}$, by

(i) $a_k(0) = k$;

(ii) $a_k(j^+) = (a_k(j))^+$, $j \in \mathbb{Z}_{>0}$.

We denote $a_k(j) = k + j$. •

Upon a moments reflection, it is easy to convince yourself that this formal definition of addition agrees with our established intuition. Roughly speaking, one defines $k + (j + 1) = (k + j) + 1$, where, by definition, the operation of adding 1 means taking the successor. With these definitions it is straightforward to verify such commonplace assertions as "$1 + 1 = 2$."

Now we define multiplication.

1.4.8 Definition (Multiplication in $\mathbb{Z}_{\geq 0}$) For $k \in \mathbb{Z}_{\geq 0}$, inductively define a map $m_k \colon \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}$, called ***multiplication by*** $\mathbf{k}$, by

(i) $m_k(0) = 0$;

(ii) $m_k(j^+) = m_k(j) + k$.

We denote $m_k(j) = k \cdot j$, or simply kj where no confusion can arise. •

Again, this definition of multiplication is in concert with our intuition. The definition says that $k \cdot (j + 1) = k \cdot j + k$. For $k, m \in \mathbb{Z}_{\geq 0}$, define k^m recursively by $k^0 = 1$, and $k^{m^+} = k^m \cdot k$. The element $k^m \in \mathbb{Z}_{\geq 0}$ is the mth ***power*** of k.

Let us verify that addition and multiplication in $\mathbb{Z}_{\geq 0}$ have the expected properties. In stating the properties, we use the usual order of operation rules one learns in high school; in this case, operations are done with the following precedence: (1) operations enclosed in parentheses, (2) multiplication, then (3) addition.

1.4.9 Proposition (Properties of arithmetic in $\mathbb{Z}_{\geq 0}$) *Addition and multiplication in $\mathbb{Z}_{\geq 0}$ satisfy the following rules:*

(i) $\mathrm{k}_1 + \mathrm{k}_2 = \mathrm{k}_2 + \mathrm{k}_1$, $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{\geq 0}$ (***commutativity*** *of addition);*

(ii) $(\mathrm{k}_1 + \mathrm{k}_2) + \mathrm{k}_3 = \mathrm{k}_1 + (\mathrm{k}_2 + \mathrm{k}_3)$, $\mathrm{k}_1, \mathrm{k}_2, \mathrm{k}_3 \in \mathbb{Z}_{\geq 0}$ (***associativity*** *of addition);*

(iii) $\mathrm{k} + 0 = \mathrm{k}$, $\mathrm{k} \in \mathbb{Z}_{\geq 0}$ (***additive identity****);*

(iv) $\mathrm{k}_1 \cdot \mathrm{k}_2 = \mathrm{k}_2 \cdot \mathrm{k}_1$, $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{\geq 0}$ (***commutativity*** *of multiplication);*

(v) $(\mathrm{k}_1 \cdot \mathrm{k}_2) \cdot \mathrm{k}_3 = \mathrm{k}_1 \cdot (\mathrm{k}_2 \cdot \mathrm{k}_3)$, $\mathrm{k}_1, \mathrm{k}_2, \mathrm{k}_3 \in \mathbb{Z}_{\geq 0}$ (***associativity*** *of multiplication);*

(vi) $\mathrm{k} \cdot 1 = \mathrm{k}$, $\mathrm{k} \in \mathbb{Z}_{\geq 0}$ (***multiplicative identity****);*

(vii) $\mathrm{j} \cdot (\mathrm{k}_1 + \mathrm{k}_2) = \mathrm{j} \cdot \mathrm{k}_1 + \mathrm{j} \cdot \mathrm{k}_2$, $\mathrm{j}, \mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{\geq 0}$ (***distributivity****);*

(viii) $\mathrm{j}^{\mathrm{k}_1} \cdot \mathrm{j}^{\mathrm{k}_2} = \mathrm{j}^{\mathrm{k}_1 + \mathrm{k}_2}$, $\mathrm{j}, \mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{\geq 0}$*;*

(ix) *if* $\mathrm{j}_1 + \mathrm{k} = \mathrm{j}_2 + \mathrm{k}$ *then* $\mathrm{j}_1 = \mathrm{j}_2$, $\mathrm{j}_1, \mathrm{j}_2, \mathrm{k} \in \mathbb{Z}_{\geq 0}$ (***cancellation law for addition****);*

(x) *if* $\mathrm{j}_1 \cdot \mathrm{k} = \mathrm{j}_2 \cdot \mathrm{k}$ *then* $\mathrm{j}_1 = \mathrm{j}_2$, $\mathrm{j}_1, \mathrm{j}_2, \mathrm{k} \in \mathbb{Z}_{>0}$ (***cancellation law for multiplication****).*

Proof We shall prove these in logical order, rather than the order in which they are stated.

(ii) We prove this by induction on k_3. For $k_3 = 0$ we have $(k_1 + k_2) + 0 = k_1 + k_2$ and $k_1 + (k_2 + 0) = k_1 + k_2$, giving the result in this case. Now suppose that $(k_1 + k_2) + j = k_1 + (k_2 + j)$ for $j \in \{0, 1, \ldots, k_3\}$. Then

$$(k_1 + k_2) + k_3^+ = ((k_1 + k_2) + k_3)^+ = (k_1 + (k_2 + k_3))^+ = k_1 + (k_2 + k_3)^+ = k_1 + (k_2 + k_3^+),$$

where we have used the definition of addition, the induction hypothesis, and then twice used the definition of addition.

(i) We first claim that $0 + k = k$ for all $k \in \mathbb{Z}_{\geq 0}$. It is certainly true, by definition, that $0 + 0 = 0$. Now suppose that $0 + j = j$ for $j \in \{0, 1, \ldots, k\}$. Then

$$0 + k^+ = 0 + (k + 1) = (0 + k) + 1 = k + 1 = k^+.$$

We next claim that $k_1^+ + k_2 = (k_1 + k_2)^+$ for $k_1, k_2 \in \mathbb{Z}_{\geq 0}$. We prove this by induction on k_2. For $k_2 = 0$ we have $k_1^+ + 0 = k_1^+$ and $(k_1 + 0)^+ = k_1^+$, using the definition of addition. This gives the claim for $k_2 = 0$. Now suppose that $k_1^+ + j = (k_1 + j)^+$ for $j \in \{0, 1, \ldots, k_2\}$. Then

$$k_1^+ + k_2^+ = k_1^+ + (k_2 + 1) = (k_1^+ + k_2) + 1 = (k_1^+ + k_2)^+,$$

as desired.

We now complete the proof of this part of the result by induction on k_1. For $k_1 = 0$ we have $0 + k_2 = k_2 = k_2 + 0$, using the first of our claims above and the definition of addition. Now suppose that $j + k_2 = k_2 + j$ for $j \in \{0, 1, \ldots, k_1\}$. Then

$$k_1^+ + k_2 = (k_1 + k_2)^+ = (k_2 + k_1)^+ = k_2 + k_1^+,$$

using the second or our claims above and the definition of addition.

(iii) This is part of the definition of addition.

(vii) We prove the this by induction on k_2. First note that for $k_2 = 0$ we have $j \cdot (k_1 + 0) = j \cdot k_1$ and $j \cdot k_1 + j \cdot 0 = j \cdot k_1 + 0 = j \cdot k_1$, so the result holds when $k_2 = 0$. Now suppose that $j \cdot (k_1 + k) = j \cdot k_1 + j \cdot k$ for $k \in \{0, 1, \ldots, k_2\}$. Then we have

$$\begin{aligned} j \cdot (k_1 + k_2^+) &= j \cdot (k_1 + k_2)^+ = j \cdot (k_1 + k_2) + j \\ &= (j \cdot k_1 + j \cdot k_2) + j = j \cdot k_1 + (j \cdot k_2 + j) \\ &= j \cdot k_1 + j \cdot k_2^+, \end{aligned}$$

as desired, where we have used, in sequence, the definition of addition, the definition of multiplication, the induction hypothesis, the associativity of addition, and the definition of multiplication.

(iv) We first prove by induction on k that $0 \cdot k = 0$ for $k \in \mathbb{Z}_{\geq 0}$. For $k = 0$ the claim holds by definition of multiplication. So suppose that $0 \cdot j = 0$ for $j \in \{0, 1, \ldots, k\}$ and then compute $0 \cdot k^+ = 0 \cdot k + 0 = 0$, as desired.

We now prove the result by induction on k_2. For $k_2 = 0$ we have $k_1 \cdot 0 = 0$ by definition of multiplication. We also have $k_2 \cdot 0 = 0$ by the first part of the proof. So now suppose that $k_1 \cdot j = j \cdot k$ for $j \in \{0, 1, \ldots, k_2\}$. We then have

$$k_1 \cdot k_2^+ = k_1 \cdot k_2 + k_1 = k_2 \cdot k_1 + k_1 = k_1 + k_2 \cdot k_1 = (1 + k_2) \cdot k_1 = k_2^+ \cdot k_1,$$

where we have used, in sequence, the definition of multiplication, the induction hypothesis, commutativity of addition, distributivity, commutativity of addition, and the definition of addition.

(v) We prove this part of the result by induction on k_3. For $k_3 = 0$ we have $(k_1 \cdot k_2) \cdot 0 = 0$ and $k_1 \cdot (k_2 \cdot 0) = k_1 \cdot 0 = 0$. Thus the result is true when $k_3 = 0$. Now suppose that $(k_1 \cdot k_2) \cdot j = k_1 \cdot (k_2 \cdot j)$ for $j \in \{0, 1, \dots, k_3\}$. Then

$$(k_1 \cdot k_2) \cdot k_3^+ = (k_1 \cdot k_2) \cdot k_3 + k_1 \cdot k_2 = k_1 \cdot (k_2 \cdot k_3) + k_1 \cdot k_2 = k_1 \cdot (k_2 \cdot k_3 + k_2) = k_1 \cdot (k_2 \cdot k_3^+),$$

where we have used, in sequence, the definition of multiplication, the induction hypothesis, distributivity, and the definition of multiplication.

(vi) This follows from the definition of multiplication.

(viii) We prove the result by induction on k_1. The result is obviously true for $k_2 = 0$, so suppose that $j^{k_1+l} = j^{k_1} \cdot j^l$ for $l \in \{1, \dots, k_2\}$. Then

$$j^{k_1+k_2^+} = j^{(k_1+k_2)^+} = j^{k_1+k_2} \cdot j = j^{k_1} \cdot j^{k_2} \cdot j = j^{k_1} \cdot j^{k_2^+},$$

as desired.

(ix) We prove the result by induction on k. Since

$$j_1 + 0 = j_1, \quad j_2 + 0 = j_2,$$

the assertion holds for all $j_1, j_2 \in \mathbb{Z}_{\geq 0}$ and for $k = 0$. Now suppose the result holds for all $j_1, j_2 \in \mathbb{Z}_{\geq 0}$ and for $k \in \{0, 1, \dots, m\}$. Then

$$j_1 + (m + 1) = (j_1 + m) + 1, \quad j_2 + (m + 1) = (j_2 + m) + 1$$

and so

$$(j_1 + m) + 1 = (j_2 + m) + 1 \implies j_1 + m = j_2 + m \implies j_1 = j_2,$$

using the induction hypotheses. Thus the result holds for $k = m + 1$, completing our proof by induction.

(x) We prove this result by induction on j_1. First take $j_1 = 1$ and assume that $1 \cdot k = j_2 \cdot k$ for all $j_2, k \in \mathbb{Z}_{>0}$. If $j_2 = 1$ then we conclude that the assertion holds. If $j_2 \neq 1$, then $j_2 = j_2' + 1$ for some $j_2' \in \mathbb{Z}_{>0}$ and so we have

$$1 \cdot k = (j_2' + 1) \cdot k = j_2' \cdot k + 1 \cdot k,$$

giving $j_2' \cdot k = 0$ using the cancellation rule for addition. But the definition of multiplication by j_2' implies that we must have $k = 0$, which is not the case since we are assuming that $k \in \mathbb{Z}_{>0}$. Thus the assertion holds for $j_1 = 1$ and for all $j_2, k \in \mathbb{Z}_{>0}$. Now assume that the assertion holds for $j_2 \in \{1, \dots, m\}$ and assume that $(m + 1) \cdot k = j_2 \cdot k$ for all $j_2, k \in \mathbb{Z}_{>0}$. We first assert that $j_2 \neq 1$. Indeed, if $j_2 = 1$ we have $m \cdot k = 0$ using the cancellation law for addition, and, as above, this cannot be since $k \in \mathbb{Z}_{>0}$. Therefore, $j_2 = j_2' + 1$ for some $j_2' \in \mathbb{Z}_{>0}$ and so

$$(m + 1) \cdot k = (j_2' + 1) \cdot k \implies m \cdot k = j_2' \cdot k$$

by the cancellation law for addition. Thus, by the induction hypothesis, $m = j_2'$ and so $j_2 = m + 1$, which gives this part of the lemma. ■

1.4.2 Two relations on $\mathbb{Z}_{\geq 0}$

Another property of the naturals that we would all agree they ought to have is an "order." Thus we should have a means of saying when one natural number is less than another. To get started at this, we have the following result.

1.4.10 Lemma *For* $\mathrm{j}, \mathrm{k} \in \mathbb{Z}_{\geq 0}$*, exactly one of the following possibilities holds:*

(i) $\mathrm{j} \subset \mathrm{k}$*;*

(ii) $\mathrm{k} \subset \mathrm{j}$*;*

(iii) $\mathrm{j} = \mathrm{k}$*.*

Proof For $k \in \mathbb{Z}_{\geq 0}$ define

$$S(k) = \{j \in \mathbb{Z}_{>0} \mid \ j \subset k, k \subset j, \text{ or } j = k\}.$$

We shall prove by induction that $S(k) = \mathbb{Z}_{\geq 0}$ for each $k \in \mathbb{Z}_{\geq 0}$.

First take the case of $k = 0$. Since $\emptyset$ is a subset of every set, $0 \in S(0)$. Now suppose that $j \in S(0)$ for $j \in \mathbb{Z}_{\geq 0}$. We have the following cases.

1. $j \in 0$: This is impossible since 0 is the empty set.
2. $0 \in j$: In this case $0 \in j^+$.
3. $0 = j$: In this case $0 \in j^+$.

Thus $j \in S(0)$ implies that $j^+ \in S(0)$, and so $S(0) = \mathbb{Z}_{\geq 0}$.

Now suppose that $S(m) = \mathbb{Z}_{\geq 0}$ for $m \in \{0, 1, \ldots, k\}$. We will show that $S(k^+) = \mathbb{Z}_{\geq 0}$. Clearly $0 \in S(k^+)$. So suppose that $j \in S(k^+)$. We again have three cases.

1. $j \in k^+$: We have the following two subcases.
   (a) $j = k$: Here we have $j^+ = k^+$.
   (b) $j \in k$: Since $j^+ \in S(k)$ by the induction hypothesis, we have the following three cases.
      i. $k \in j^+$: This is impossible since $j \in k$.
      ii. $j^+ \in k$: Here $j^+ \in k^+$.
      iii. $j^+ = k$: Here again, $j^+ \in n^+$.
2. $k^+ \in j$: In this case $k^+ \in j^+$.
3. $k^+ = j$: In this case $k^+ \in j^+$.

In all cases we conclude that $j^+ \in S(k^+)$, and this completes the proof. ■

It is easy to show that $j \in k$ if and only if $j \subseteq k$, and that, if $j \in k$ but $j \neq k$, then $j \subset k$ (see Exercise 1.4.2). With this result, it is now comparatively easy to prove the following.

1.4.11 Proposition (Order[6] on $\mathbb{Z}_{\geq 0}$) *On* $\mathbb{Z}_{\geq 0}$ *define two relations* $<$ *and* $\leq$ *by*

$$\begin{aligned} \mathrm{j} < \mathrm{k} &\iff \mathrm{j} \subset \mathrm{k}, \\ \mathrm{j} \leq \mathrm{k} &\iff \mathrm{j} \subseteq \mathrm{k}. \end{aligned}$$

Then

(i) $<$ *and* $\leq$ *are transitive,*

(ii) $<$ *is irreflexive;*

(iii) $\leq$ *is reflexive and antisymmetric.*

Furthermore, for any $\mathrm{j}, \mathrm{k} \in \mathbb{Z}_{\geq 0}$*, either* $\mathrm{j} \leq \mathrm{k}$ *or* $\mathrm{k} \leq \mathrm{j}$*.*

The following rewording of the final part of the result is distinguished.

[6] We have not introduced the notion of order yet, but refer the reader to Section 1.5.

1.4.12 Corollary (Trichotomy Law for $\mathbb{Z}_{\geq 0}$) *For* $j, k \in \mathbb{Z}_{\geq 0}$*, exactly one of the following possibilities holds:*

(i) $j < k$*;*

(ii) $k < j$*;*

(iii) $j = k$*.*

Of course, the symbols "$<$" and "$\leq$" have their usual meaning, which is "less than" and "less than or equal to," respectively. We shall explore such matters in more depth and generality in Section 1.5.

We shall also sometimes write "$j > k$" (resp. "$j \geq k$") for "$k < j$" (resp. "$k \leq j$"). The symbols "$>$" and "$\geq$" then have their usual meaning as "greater than" and "greater than or equal to," respectively.

The relations $<$ and $\leq$ satisfy some natural properties with respect to addition and multiplication in $\mathbb{Z}_{\geq 0}$. Let us record these, leaving their proof as Exercise 1.4.3.

1.4.13 Proposition (Relation between addition and multiplication and <) *For* $j, k, m \in \mathbb{Z}_{\geq 0}$*, the following statements hold:*

(i) if $j < k$ *then* $j + m < k + m$*;*

(ii) if $j < k$ *and if* $m \neq 0$ *then* $m \cdot j < m \cdot k$*.*

1.4.3 Construction of the integers from the natural numbers

Next we construct negative numbers to arrive at a definition of the integers. The construction renders the integers as the set of equivalence classes under a prescribed equivalence relation in $\mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$. The equivalence relation is defined formally as follows:

$$(j_1, k_1) \sim (j_2, k_2) \iff j_1 + k_2 = k_1 + j_2. \tag{1.1}$$

It is a simple exercise to check that this is indeed an equivalence relation.

We now define the integers.

1.4.14 Definition (Integers) The set of ***integers*** is the set $\mathbb{Z} = (\mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0})/\sim$, where $\sim$ is the equivalence relation in (1.1). •

Now let us try to understand this definition by understanding the equivalence classes under the relation of (1.1). Key to this is the following result.

1.4.15 Lemma *Let* Z *be the subset of* $\mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$ *defined by*

$$Z = \{(k, 0) \mid k \in \mathbb{Z}_{>0}\} \cup \{(0, k) \mid k \in \mathbb{Z}_{>0}\} \cup \{(0, 0)\},$$

and define a map $f_Z \colon Z \to \mathbb{Z}$ *by* $f_Z(j, k) = [(j, k)]$*. Then* f_Z *is a bijection.*

Proof First we show that f_Z is injective. Suppose that $f_Z(j_1, k_1) = f_Z(j_2, k_2)$. This means that $(j_1, k_1) \sim (j_2, k_2)$, or that $j_1 + k_2 = k_1 + j_2$. If $(j_1, k_1) = (0, 0)$, then this means that $k_2 = j_2$, which means that $(j_2, k_2) = (0, 0)$ since this is the only element of Z whose entries agree. If $j_1 = 0$ and $k_1 > 0$, then we have $k_2 = k_1 + j_2$. Since at least one of j_2

and k_2 must be zero, we then deduce that it must be that j_2 is zero (or else the equality $k_2 = k_1 + j_2$) cannot hold. This then also gives $k_2 = k_1$. A similar argument holds if $j_1 > 0$ and $k_1 = 0$. This shows injectivity of f_Z.

Next we show that f_Z is surjective. Let $[(j,k)] \in \mathbb{Z}$. By the Trichotomy Law, we have three cases.

1. $j = k$: We claim that $[(j,j)] = f_Z(0,0)$. Indeed, we need only note that $(0,0) \sim (j,j)$ since $0 + j = 0 + j$.
2. $j < k$: Let $m \in \mathbb{Z}_{>0}$ be defined such that $j + m = k$. (Why can this be done?) We then claim that $f_Z(0,m) = [(j,k)]$. Indeed, since $0 + k = m + j$, this is so.
3. $k < j$: Here we let $m \in \mathbb{Z}_{>0}$ satisfy $k + m = j$, and, as in the previous case, we can easily check that $f_Z(m,0) = [(j,k)]$. ■

With this in mind, we introduce the following notation to denote an integer.

1.4.16 Notation (Notation for integers) Let $[(j,k)] \in \mathbb{Z}$.

(i) If $f_Z^{-1}[(j,k)] = [(0,0)]$ then we write $[(j,k)] = 0$.

(ii) If $[(j,k)] = [(m,0)]$, $m > 0$, then we write $[(j,k)] = m$. Such integers are ***positive***.

(iii) If $[(j,k)] = [(0,m)]$, $m > 0$, then we write $[(j,k)] = -m$. Such integers are ***negative***.

An integer is ***nonnegative*** if it is either positive or zero, and an integer is ***nonpositive*** if it is either negative or zero. •

This then relates the equivalence class definition of integers to the notion we are more familiar with: positive and negative numbers. We can also define the familiar operations of addition and multiplication of integers.

1.4.17 Definition (Addition and multiplication in $\mathbb{Z}$) Define the operations of ***addition*** and ***multiplication*** in $\mathbb{Z}$ by

(i) $[(j_1,k_1)] + [(j_2,k_2)] = [(j_1 + j_2, k_1 + k_2)]$ and

(ii) $[(j_1,k_1)] \cdot [(j_2,k_2)] = [(j_1 \cdot j_2 + k_1 \cdot k_2, j_1 \cdot k_2 + k_1 \cdot j_2)]$,

respectively, for $[(j_1,k_1)], [(j_2,k_2)] \in \mathbb{Z}$. As with multiplication in $\mathbb{Z}_{\geq 0}$, we shall sometimes omit the "$\cdot$". •

These definitions do not *a priori* make sense; this needs to be verified.

1.4.18 Lemma *The definitions for addition and multiplication in $\mathbb{Z}$ a well-defined in that they do not depend on the choice of representative.*

Proof Let $(j_1,k_1) \sim (\tilde{j}_1,\tilde{k}_1)$ and $(j_2,k_2) \sim (\tilde{j}_2,\tilde{k}_2)$. Thus

$$j_1 + \tilde{k}_1 = k_1 + \tilde{j}_1, \quad j_2 + \tilde{k}_2 = k_2 + \tilde{j}_2.$$

It therefore follows that

$$(\tilde{j}_1 + \tilde{j}_2) + (k_1 + k_2) = (\tilde{k}_1 + \tilde{k}_2) + (j_1 + j_2),$$

which gives the independence of addition on representative.

To verify the well-definedness of multiplication, we first see that

$$j_2 \cdot (j_1 + \tilde{k}_1) + k_2 \cdot (\tilde{j}_1 + k_1) + \tilde{j}_1 \cdot (j_2 + \tilde{k}_2) + \tilde{k}_1 \cdot (\tilde{j}_2 + k_2) \\ = j_2 \cdot (k_1 + \tilde{j}_1) + k_2 \cdot (j_1 + \tilde{k}_1) + \tilde{j}_1 \cdot (k_2 + \tilde{j}_2) + \tilde{k}_1 \cdot (j_2 + \tilde{k}_2),$$

and expanding this and rearranging gives

$$(j_1 \cdot j_2 + k_1 \cdot k_2 + \tilde{k}_1 \cdot \tilde{j}_2 + \tilde{j}_1 \cdot \tilde{k}_2) + (\tilde{k}_1 \cdot j_2 + \tilde{j}_1 \cdot k_2 + \tilde{j}_1 \cdot j_2 + \tilde{k}_1 \cdot k_2) \\ = (k_1 \cdot j_2 + j_1 \cdot k_2 + \tilde{j}_1 \cdot \tilde{j}_2 + \tilde{k}_1 \cdot \tilde{k}_2) + (\tilde{k}_1 \cdot j_2 + \tilde{j}_1 \cdot k_2 + \tilde{j}_1 \cdot j_2 + \tilde{k}_1 \cdot k_2).$$

Using the cancellation law for addition we then have

$$(\tilde{j}_1 \cdot \tilde{j}_2 + \tilde{k}_1 \cdot \tilde{k}_2) + (j_1 \cdot k_2 + k_1 \cdot j_2) = (\tilde{j}_1 \cdot \tilde{k}_2 + \tilde{k}_1 \cdot \tilde{j}_2) + (j_1 \cdot j_2 + k_1 \cdot k_2),$$

which gives the independence of multiplication on representative. ■

As with elements of $\mathbb{Z}_{\geq 0}$, we can define powers for integers. Let $k \in \mathbb{Z}$ and $m \in \mathbb{Z}_{\geq 0}$. We define k^m recursively as follows. We take $k^0 = 1$ and define $k^{m^+} = k^m \cdot k$. We call k^m the mth ***power*** of k. Note that, at this point, k^m only makes sense for $m \in \mathbb{Z}_{\geq 0}$.

Finally, we give the properties of addition and multiplication in $\mathbb{Z}$. Some of these properties are as for $\mathbb{Z}_{\geq 0}$. However, there is a useful new feature that arises in $\mathbb{Z}$ that mirrors our experience with negative numbers. In the statement of the result, it is convenient to denote an integer as in Notation 1.4.16, rather than as in the definition.

1.4.19 Proposition (Properties of addition and multiplication in $\mathbb{Z}$) *Addition and multiplication in $\mathbb{Z}$ satisfy the following rules:*

(i) $k_1 + k_2 = k_2 + k_1$, $k_1, k_2 \in \mathbb{Z}$ (***commutativity*** *of addition);*

(ii) $(k_1 + k_2) + k_3 = k_1 + (k_2 + k_3)$, $k_1, k_2, k_3 \in \mathbb{Z}$ (***associativity*** *of addition);*

(iii) $k + 0 = k$, $k \in \mathbb{Z}$ (***additive identity***);

(iv) $k + (-1 \cdot k) = 0$, $k \in \mathbb{Z}$ (***additive inverse***);

(v) $k_1 \cdot k_2 = k_2 \cdot k_1$, $k_1, k_2 \in \mathbb{Z}$ (***commutativity*** *of multiplication);*

(vi) $(k_1 \cdot k_2) \cdot k_3 = k_1 \cdot (k_2 \cdot k_3)$, $k_1, k_2, k_3 \in \mathbb{Z}$ (***associativity*** *of multiplication);*

(vii) $k \cdot 1 = k$, $k \in \mathbb{Z}$ (***multiplicative identity***);

(viii) $j \cdot (k_1 + k_2) = j \cdot k_1 + j \cdot k_2$, $j, k_1, k_2 \in \mathbb{Z}$ (***distributivity***);

(ix) $j^{k_1} \cdot j^{k_2} = j^{k_1 + k_2}$, $j \in \mathbb{Z}$, $k_1, k_2 \in \mathbb{Z}_{\geq 0}$.

Moreover, if we define $i_{\mathbb{Z}_{\geq 0}} : \mathbb{Z}_{\geq 0} \to \mathbb{Z}$ by $i_{\mathbb{Z}_{\geq 0}}(k) = [(k, 0)]$, then addition and multiplication in $\mathbb{Z}$ agrees with that in $\mathbb{Z}_{\geq 0}$:

$$i_{\mathbb{Z}_{\geq 0}}(k_1) + i_{\mathbb{Z}_{\geq 0}}(k_2) = i_{\mathbb{Z}_{\geq 0}}(k_1 + k_2), \quad i_{\mathbb{Z}_{\geq 0}}(k_1) \cdot i_{\mathbb{Z}_{\geq 0}}(k_2) = i_{\mathbb{Z}_{\geq 0}}(k_1 \cdot k_2).$$

Proof These follow easily from the definitions of addition and multiplication, using the fact that the corresponding properties hold for $\mathbb{Z}_{\geq 0}$. We leave the details to the reader as Exercise 1.4.4. We therefore only prove the new property (iv). For this, we

suppose without loss of generality that $k \in \mathbb{Z}_{\geq 0}$, i.e., $k = [(k, 0)]$. Then $-k = [(0, k)]$ so that

$$k + (-k) = [(k + 0, 0 + k)] = [(k, k)] = [(0, 0)] = 0,$$

as claimed. ■

We shall make the convention that $-1 \cdot k$ be written as $-k$, whether k be positive or negative. We shall also, particularly as we move along to things of more substance, think of $\mathbb{Z}_{\geq 0}$ as a subset of $\mathbb{Z}$, without making explicit reference to the map $i_{\mathbb{Z}_{\geq 0}}$.

1.4.4 Two relations in $\mathbb{Z}$

Finally we introduce in $\mathbb{Z}$ two relations that extend the relations $<$ and $\leq$ for $\mathbb{Z}_{\geq 0}$. The following result is the analogue of Proposition 1.4.11.

1.4.20 Proposition (Order on $\mathbb{Z}$) *On $\mathbb{Z}$ define two relations $<$ and $\leq$ by*

$$\begin{aligned} [(\mathrm{j}_1, \mathrm{k}_1)] < [(\mathrm{j}_2, \mathrm{k}_2)] &\iff \mathrm{j}_1 + \mathrm{k}_2 < \mathrm{k}_1 + \mathrm{j}_2, \\ [(\mathrm{j}_1, \mathrm{k}_1)] \leq [(\mathrm{j}_2, \mathrm{k}_2)] &\iff \mathrm{j}_1 + \mathrm{k}_2 \leq \mathrm{k}_1 + \mathrm{j}_2. \end{aligned}$$

*Then**missing stuff***

(i) $<$ *and* $\leq$ *are transitive,*

(ii) $<$ *is irreflexive, and*

(iii) $\leq$ *is reflexive.*

Furthermore, for any $\mathrm{j}, \mathrm{k} \in \mathbb{Z}$, *either* $\mathrm{j} \leq \mathrm{k}$ *or* $\mathrm{k} \leq \mathrm{j}$.

Proof First one must show that the relations are well-defined in that they do not depend on the choice of representative. Thus let $[(j_1, k_1)] \sim [(\tilde{j}_1, \tilde{k}_1)]$ and $[(j_2, k_2)] \sim [(\tilde{j}_2, \tilde{k}_2)]$, so that

$$j_1 + \tilde{k}_1 = k_1 + \tilde{j}_1, \quad j_2 + \tilde{k}_2 = k_2 + \tilde{j}_2.$$

Now suppose that the relation $j_1 + k_2 < k_1 + j_2$ holds. Now perform the following steps:

1. add $\tilde{j}_1 + k_1 + j_2 + \tilde{k}_2 + j_1 + \tilde{k}_1 + k_2 + \tilde{j}_2$ to both sides of the relation;
2. observe that $j_1 + k_2 + k_1 + j_2$ appears on both sides of the relation;
3. observe that $j_1 + \tilde{k}_1$ appears on one side of the relation and that $\tilde{j}_1 + k_1$ appears on the other;
4. observe that $k_2 + \tilde{j}_2$ appears on one side of the relation and that $j_2 + \tilde{k}_2$ appears on the other.

After simplification using the above observations, and using Proposition 1.4.13, we note that the relation $\tilde{j}_1 + \tilde{k}_2 < \tilde{k}_1 + \tilde{j}_2$ holds, which gives independence of the definition of $<$ on the choice of representative. The same argument works for the relation $\leq$.

The remainder of the proof follows in a fairly straightforward manner from the corresponding assertions for $\mathbb{Z}_{\geq 0}$, and we leave the details to the reader as Exercise 1.4.6. ■

As with the natural numbers, the last assertion of the previous result has a standard restatement.

1.4.21 Corollary (Trichotomy Law for $\mathbb{Z}$) *For* $j, k \in \mathbb{Z}$, *exactly one of the following possibilities holds:*

(i) $j < k$;

(ii) $k < j$;

(iii) $j = k$.

Similarly with $\mathbb{Z}_{\geq 0}$, we shall also write "$j > k$" for "$k < j$" and "$j \geq k$" for "$k \leq j$." It is also easy to directly verify that the relations $<$ and $\leq$ have the expected properties with respect to positive and negative integers. These are given in Exercise 1.4.7, for the interested reader.

We also have the following extension of Proposition 1.4.13 that relates addition and multiplication to the relations $<$ and $\leq$. We again leave these to the reader to verify in Exercise 1.4.8.

1.4.22 Proposition (Relation between addition and multiplication and <) *For* $j, k, m \in \mathbb{Z}$, *the following statements hold:*

(i) *if* $j < k$ *then* $j + m < k + m$;

(ii) *if* $j < k$ *and if* $m > 0$ *then* $m \cdot j < m \cdot k$;

(iii) *if* $j < k$ *and if* $m < 0$ *then* $m \cdot k < m \cdot j$;

(iv) *if* $0 < j, k$ *then* $0 < j \cdot k$.

1.4.5 The absolute value function

On the set of integers there is an important map that assigns a nonnegative integer to each integer.

1.4.23 Definition (Integer absolute value function) The ***absolute value function*** on $\mathbb{Z}$ is the map from $\mathbb{Z}$ to $\mathbb{Z}_{\geq 0}$, denoted by $k \mapsto |k|$, defined by

$$|k| = \begin{cases} k, & 0 < k, \\ 0, & k = 0, \\ -k, & k < 0. \end{cases}$$

•

The absolute value has the following properties.

1.4.24 Proposition (Properties of absolute value on $\mathbb{Z}$) *The following statements hold:*

(i) $|k| \geq 0$ *for all* $k \in \mathbb{Z}$;

(ii) $|k| = 0$ *if and only if* $k = 0$;

(iii) $|j \cdot k| = |j| \cdot |k|$ *for all* $j, k \in \mathbb{Z}$;

(iv) $|j + k| \leq |j| + |k|$ *for all* $j, k \in \mathbb{Z}$ (***triangle inequality***).

Proof Parts (i) and (ii) follow directly from the definition of $|\cdot|$.

(iii) We first note that $|-k| = |k|$ for all $k \in \mathbb{Z}$. Now, if $0 \leq j, k$, then the result is clear. If $j < 0$ and $k \geq 0$, then

$$|j \cdot k| = |-1 \cdot (-j) \cdot k| = |(-j) \cdot k| = |-j| \cdot |k| = |j| \cdot |k|.$$

A similar argument holds when $k < 0$ and $j \geq 0$.

(iv) We consider various cases.

1. $|j| \leq |k|$:
   (a) $j, k \geq 0$: Here $|j + k| = j + k$, and $|j| = j$ and $|k| = k$. So the result is obvious.
   (b) $j < 0, k \geq 0$: Here one can easily argue, using the definition of addition, that $0 < j + k$. From Proposition 1.4.22 we have $j + k < 0 + k = k$. Therefore, $|j + k| < |k| < |j| + |k|$, again by Proposition 1.4.22.
   (c) $k < 0, j \geq 0$: This follows as in the preceding case, swapping j and k.
   (d) $j, k < 0$: Here $|j+k| = |-j+(-k)| = |-(j+k)| = -(j+k)$, and $|j| = -j$ and $|k| = -k$, so the result follows immediately.
2. $|k| \leq |j|$: The argument here is the same as the preceding one, but swapping j and k. ■

Exercises

1.4.1 Let $k \in \mathbb{Z}_{>0}$. Show that $k \subseteq \mathbb{Z}_{>0}$; thus k is both an element of $\mathbb{Z}_{>0}$ and a subset of $\mathbb{Z}_{>0}$.

1.4.2 Let $j, k \in \mathbb{Z}_{\geq 0}$. Do the following:
   (a) show that $j \in k$ if and only if $j \subseteq k$;
   (b) show that if $j \subset k$, then $k \notin j$ (and so $j \in k$ by the Trichotomy Law).

1.4.3 Prove Proposition 1.4.13.

1.4.4 Complete the proof of Proposition 1.4.19.

1.4.5 For $j_1, j_2, k \in \mathbb{Z}$, prove the distributive rule $(j_1 + j_2) \cdot k = j_1 \cdot k + j_2 \cdot k$.

1.4.6 Complete the proof of Proposition 1.4.20.

1.4.7 Show that the relations $<$ and $\leq$ on $\mathbb{Z}$ have the following properties:
   1. $[(0, j)] < [(0, 0)]$ for all $j \in \mathbb{Z}_{>0}$;
   2. $[(0, j)] < [(k, 0)]$ for all $j, k \in \mathbb{Z}_{>0}$;
   3. $[(0, j)] < [(0, k)]$, $j, k, \in \mathbb{Z}_{\geq 0}$, if and only if $k < j$;
   4. $[(0, 0)] < [(j, 0)]$ for all $j \in \mathbb{Z}_{>0}$;
   5. $[(j, 0)] < [(k, 0)]$, $j, k \in \mathbb{Z}_{\geq 0}$, if and only if $j < k$;
   6. $[(0, j)] \leq [(0, 0)]$ for all $j \in \mathbb{Z}_{\geq 0}$;
   7. $[(0, j)] \leq [(k, 0)]$ for all $j, k \in \mathbb{Z}_{\geq 0}$;
   8. $[(0, j)] \leq [(0, k)]$, $j, k, \in \mathbb{Z}_{\geq 0}$, if and only if $k \leq j$;
   9. $[(0, 0)] \leq [(j, 0)]$ for all $j \in \mathbb{Z}_{\geq 0}$;
   10. $[(j, 0)] \leq [(k, 0)]$, $j, k \in \mathbb{Z}_{\geq 0}$, if and only if $j \leq k$.

1.4.8 Prove Proposition 1.4.22.

Section 1.5

Orders of various sorts

In Section 1.4 we defined two relations, denoted by $<$ and $\leq$, on both $\mathbb{Z}_{\geq 0}$ and $\mathbb{Z}$. Here we see that these relations have additional properties that fall into a general class of relations called orders. There are various classes or orders, having varying degrees of "strictness," as we shall see.

Do I need to read this section? Much of the material in this section is not used widely in the series, so perhaps can be overlooked until it is needed. •

1.5.1 Definitions

Let us begin by defining the various types of orders we consider.

1.5.1 Definition (Partial order, total order, well order) Let S be a set and let R be a relation in S.

(i) R is a ***partial order*** in S if it is reflexive, transitive, and antisymmetric.

(ii) A ***partially ordered set*** is a pair (S, R) where R is a partial order in S.

(iii) R is a ***strict partial order*** in S if it is irreflexive and transitive.

(iv) A ***strictly partially ordered set*** is a pair (S, R) where R is a strict partial order in S.

(v) R is a ***total order*** in S if it is a partial order and if, for each $x_1, x_2 \in S$, either $(x_1, x_2) \in R$ or $(x_2, x_1) \in R$.

(vi) A ***totally ordered set*** is a pair (S, R) where R is a total order in S.

(vii) R is a ***well order*** in S if it is a partial order and if, for every nonempty subset $A \subseteq S$, there exists an element $x \in A$ such that $(x, x') \in R$ for every $x' \in A$.

(viii) A ***well ordered set*** is a pair (S, R) where R is a well order in S. •

1.5.2 Remark (Mathematical structures as ordered pairs) In the preceding definitions we see four instances of an "X set," where X is some property, e.g., a partial order. In such cases, it is common practice to do as we have done and write the object as an ordered pair, in the cases above, as (S, R). The practice dictates that the first element in the ordered pair be the name of the set, and that the second specifies the structure.

In many cases one simply wishes to refer to the set, with the structure being understood. For example, one might say, "Consider the partially ordered set S..." and not make explicit reference to the partial order. Both pieces of language are in common use by mathematicians, and in mathematical texts. •

Let us consider some simple examples of partial and strict partial orders.

1.5.3 Examples (Partial orders)

1. Consider the relation $R = \{(k_1, k_2) \mid k_1 \leq k_2\}$ in either $\mathbb{Z}_{\geq 0}$ or $\mathbb{Z}$. Then one verifies that R is a partial order. In fact, it is both a total order and a well order.
2. Consider the relation $R = \{(k_1, k_2) \mid k_1 \leq k_2\}$ in either $\mathbb{Z}_{\geq 0}$ or $\mathbb{Z}$. Here one can verify that R is a strict partial order.
3. Let S be a set and consider the relation R in $\mathbf{2}^S$ defined by $R = \{(A, B) \mid A \subseteq B\}$. Here one can see that R is a partial order, but it is generally neither a total order nor a well order (cf. Exercise 1.5.2).
4. Let S be a set and consider the relation R in $\mathbf{2}^S$ defined by $R = \{(A, B) \mid A \subset B\}$. In this case R can be verified to be a strict partial order.
5. A well order R is a total order. Indeed, for $(x_1, x_2) \in R$, there exists an element $x \in \{x_1, x_2\}$ such that $(x, x') \in R$ for every $x' \in \{x_1, x_2\}$. But this implies that either $(x_1, x_2) \in R$ or $(x_2, x_1) \in R$, meaning that R is a total order. •

Motivated by the first and second of these examples, we utilise the following more or less commonplace notation for partial orders.

1.5.4 Notation ($\preceq$ and $\prec$) If R is a partial order in S, we shall normally write $x_1 \preceq x_2$ for $(x_1, x_2) \in R$, and shall refer to $\preceq$ as the partial order. In like manner, if R is a strict partial order in S, we shall write $x_1 \prec x_2$ for $(x_1, x_2) \in R$. We shall also use $x_1 \succeq x_2$ and $x_1 \succ x_2$ to stand for $x_2 \preceq x_1$ and $x_2 \prec x_1$, respectively. •

There is a natural way of associating to every partial order a strict partial order, and vice versa.

1.5.5 Proposition (Relationship between partial and strict partial orders) *Let* S *be a set.*

(i) If $\preceq$ is a partial order in S*, then the relation $\prec$ defined by*

$$x_1 \prec x_2 \iff x_1 \preceq x_2 \text{ and } x_1 \neq x_2$$

is a strict partial order in S*.*

(ii) If $\prec$ is a strict partial order in S*, then the relation $\preceq$ defined by*

$$x_1 \preceq x_2 \iff x_1 \prec x_2 \text{ or } x_1 = x_2$$

is a partial order in S*.*

Proof This is a straightforward matter of verifying that the definitions are satisfied. ■

When talking about a partial order $\preceq$, the symbol $\prec$ will always refer to the strict partial order as in part (i) of the preceding result. Similarly, given a strict partial order $\prec$, the symbol $\preceq$ will always refer to the partial order as in part (ii) of the preceding result.

1.5.6 Examples (Example 1.5.3 cont'd)

1. One can readily verify that $<$ is the strict partial order associated with the partial order $\leq$ in either $\mathbb{Z}_{\geq 0}$ or $\mathbb{Z}$, and that $\leq$ is the partial order associated to $<$.
2. It is also easy to verify that, for a set S, $\subset$ is the strict partial order in $\mathbf{2}^S$ associated to the partial order $\subseteq$, and that $\subseteq$ is the partial order associated to $\subset$. •

1.5.2 Subsets of partially ordered sets

Surrounding subsets of a partially ordered set $(S, \preceq)$ there is some useful language. For the following definition, it is helpful to think of an order, be it partial, strictly partial, or whatever, as a relation, and to use the notation of a relation. Thus we refer to an order as R, and not as $\preceq$.

1.5.7 Definition (Restriction of an order) Let S be a set and let R be a partial order, (resp. strict partial order, total order, well order) in S. For a subset $T \subseteq S$, the ***restriction*** of R to T is the partial order (resp. strict partial order, total order, well order) in T defined by

$$R|T = R \cap \{(x_1, x_2) \in S \times S \mid x_1, x_2 \in T\}.$$ •

It is a trivial matter to see that if R is an order, then its restriction to T is an order having the same properties as R, as is tacitly assumed in the definition. The notion of the restriction of an order allows us to talk unambiguously about the order on a subset of a given set, and we shall do this freely in this section.

Since most of this section is language, let us begin with some simple language associated with points.

1.5.8 Definition (Comparing elements in a partially ordered set) Let $(S, \preceq)$ be a partially ordered set.

(i) A point $x_1 \in S$ is ***less*** than or ***smaller*** than x_2, or equivalently is a ***predecessor*** of x_2, if $x_1 \preceq x_2$.

(ii) A point $x_1 \in S$ is ***greater*** than or ***larger*** than x_2, or equivalently is a ***successor*** of x_2, if $x_1 \succeq x_2$.

(iii) A point x' is ***between*** x_1 and x_2 if $x_1 \preceq x'$ and if $x' \preceq x_2$.

Similarly, let $(S, \prec)$ be a strictly partially ordered set.

(iv) A point $x_1 \in S$ is ***strictly less*** than or ***strictly smaller*** than x_2, or equivalently is a ***strict predecessor*** of x_2, if $x_1 \prec x_2$.

(v) A point $x_1 \in S$ is ***strictly greater*** than or ***strictly larger*** than x_2, or equivalently is a ***strict successor*** of x_2, if $x_1 \succ x_2$.

(vi) A point x' is ***strictly between*** x_1 and x_2 if $x_1 \prec x'$ and if $x' \prec x_2$.

(vii) If $x_1 < x_2$ and there exists no $x' \in S$ that is strictly between x_1 and x_2, then x_1 is the ***immediate predecessor*** of x_2. •

Next we talk about some language attached to subsets of a partially ordered set.

1.5.9 Definition (Segment, least, greatest, minimal, maximal) Let $(S, \preceq)$ be a partially ordered set.

(i) The ***initial segment*** determined by $x \in S$ is the set $\underline{\text{seg}}(x) = \{x' \in S \mid x' \preceq S\}$.

(ii) A ***least***, ***smallest***, or ***first*** element in S is an element $x \in S$ with the property that $x \preceq x'$ for every $x' \in S$.

(iii) A ***greatest***, ***largest***, or ***last*** element in S is an element $x \in S$ with the property that $x' \preceq x$ for every $x' \in S$.

(iv) A ***minimal*** element of S is an element $x \in S$ with the property that $x \preceq x'$ implies that $x' = x$.

(v) A ***maximal*** element of S is an element $x \in S$ with the property that $x \prec x'$ implies that $x' = x$.

Now let $(S, \preceq)$ be a partially ordered set.

(vi) The ***strict initial segment*** determined by $x \in S$ is the set $\text{seg}(x) = \{x' \in S \mid x' \prec S\}$. •

The least and greatest elements of a set, if they exist, are unique. This is easy to prove (Exercise 1.5.4).

Let us give an example that distinguishes between least and minimal.

1.5.10 Example (Least and minimal are different) Let S be a set and consider the partially ordered set $(\mathbf{2}^S \setminus \emptyset, \subseteq)$. Then any singleton is a minimal element of $\mathbf{2}^S \setminus \emptyset$. However, unless S is itself a set with only one member, then $\mathbf{2}^S$ has no least element, i.e., there is no subset which is contained in every other subset. •

Next we turn to two important concepts related to partial orders.

1.5.11 Definition (Greatest lower bound and least upper bound) Let $(S, \preceq)$ be a partially ordered set and let $A \subseteq S$.

(i) An element $x \in S$ is a ***lower bound*** for A if $x \preceq x'$ for every $x' \in A$.

(ii) An element $x \in S$ is an ***upper bound*** for A if $x' \preceq x$ for every $x' \in A$.

(iii) If, in the set of lower bounds for A, there is a greatest element, this is the ***greatest lower bound***, or the ***infimum***, of E. This is denoted by $\inf(A)$.

(iv) If, in the set of upper bounds for A, there is a least element, this is the ***least upper bound***, or the ***supremum***, of E. This is denoted by $\sup(A)$.

Now let $(S, \prec)$ be a strictly partially ordered set and let $A \subseteq S$.

(v) An element $x \in S$ is a ***strict lower bound*** for A if $x \prec x'$ for every $x' \in A$.

(vi) An element $x \in S$ is a ***strict upper bound*** for A if $x' \prec x$ for every $x' \in A$. •

Let us give some examples that illustrate the various possibilities arising from the preceding definitions. The examples will be given for lower bounds, but similar examples can be conjured to give similar conclusions for upper bounds.

1.5.12 Examples (Greatest lower bounds)

1. A subset $A \subseteq S$ may have no lower bounds. For example, the set of negative integers has no lower bound if we use the standard partial order in $\mathbb{Z}$.
2. A subset $A \subseteq S$ may have a greatest lower bound in A. For example, the set of nonnegative integers has as lower bounds all nonpositive integers. The greatest of these lower bounds is 0, which is itself a nonnegative integer.
3. A subset $A \subseteq S$ may have a greatest lower bound that is not an element of A. To see this, let S be the set of nonpositive integers, let A be the set of negative integers, and define a partial order $\preceq$ in S by

$$k_1 \preceq k_2 \quad \Longleftrightarrow \quad \begin{cases} k_1 \leq k_2,\ k_1, k_2 \in A, & \text{or} \\ k_1 = k_2 = 0, & \text{or} \\ k_1 = 0,\ k_2 \in A. \end{cases}$$

Thus this is the usual partial order in $A \subseteq S$, and one declares 0 to be less than all elements of A. In this case, 0 is the only lower bound for A, and so is, therefore, the greatest lower bound. But $0 \notin A$. •

1.5.3 Zorn's Lemma

Zorn's[7] Lemma comes up frequently in mathematics during the course of nonconstructive existence proofs. Since some of these proofs appear in this series and are important, we state Zorn's Lemma.

1.5.13 Theorem (Zorn's Lemma) *Every partially ordered set* $(S, \preceq)$ *in which every totally ordered subset has an upper bound contains at least one maximal member.*

Proof Suppose that every totally ordered subset has an upper bound, but that S has no maximal member. By assumption, if $A \subseteq S$ is a totally ordered subset, then there exists an upper bound x for A. Since S has no maximal element, there exists $x' \in S$ such that $x < x'$. Therefore, x' is a strict upper bound for A. Thus we have shown that every totally ordered subset possesses a strict upper bound. Let b be a function from the collection of totally ordered subsets into S having the property that $b(A)$ is a strict upper bound for A.[8]

A **b-*set*** is a subset B of S that is well ordered and has the property that, for every $x \in B$, we have $x = b(\mathrm{seg}_B(x))$, where $\mathrm{seg}_B(x)$ denotes the strict initial segment of x in B.

1 Lemma *If* B_1 *and* B_2 *are unequal* b*-sets, then one of the following statements holds:*

(i) there exists $\mathrm{x}_1 \in \mathrm{B}_1$ *such that* $\mathrm{B}_2 = \mathrm{seg}_{\mathrm{B}_1}(\mathrm{x}_1)$*;*

(ii) there exists $\mathrm{x}_2 \in \mathrm{B}_2$ *such that* $\mathrm{B}_1 = \mathrm{seg}_{\mathrm{B}_2}(\mathrm{x}_2)$*.*

Proof If $B_2 \subset B_1$, then we claim that (i) holds. Take x_1 to be the least member of $B_1 - B_2$. We claim that $B_2 = \mathrm{seg}_{B_1}(x_1)$. First of all, if $x \in B_2$, then $x < x_1$ since x_1 is the least

[7] Max August Zorn (1906–1993) was a German mathematician who did work in the areas of set theory, algebra, and topology.

[8] The existence of the function b relies on the Axiom of Choice (see Section 1.8.3).

member of $B_1 - B_2$. Therefore, $B_2 \subseteq \mathrm{seg}_{B_1}(x_1)$. Now suppose that $\mathrm{seg}_{B_1}(x_1) - B_2 \neq \emptyset$, and let x be the least member of this set. Note that for any $x' \in B_2$ we therefore have $x' < x$, contradicting the fact that x_1 is the least member of $B_1 - B_2$. Thus we must have $\mathrm{seg}_{B_1}(x_1) - B_2 = \emptyset$, and so $B_2 = \mathrm{seg}_{B_1}(x_1)$.

We now suppose that $B_2 - B_1 \neq \emptyset$. Let x_2 be the least member of $B_2 - B_1$. If $x \in \mathrm{seg}_{B_2}(x_2)$ then $x < x_2$ and x must therefore be an element of B_1, or else this contradicts the definition of x_2. Now suppose that $B_1 \setminus \mathrm{seg}_{B_2}(x_2) \neq \emptyset$ and let y_1 be the least member of this set. If $y \in \mathrm{seg}_{B_1}(y_1)$ and $y' \in B_2$ satisfies $y' < y$, then $y' \in \mathrm{seg}_{B_1}(y_1)$. If z is the least member of $B_2 \setminus \mathrm{seg}_{B_1}(y_1)$, we then have $\mathrm{seg}_{B_2}(z) = \mathrm{seg}_{B_1}(y_1)$. Therefore

$$z = b(\mathrm{seg}_{B_2}(z)) = b(\mathrm{seg}_{B_1}(y_1)) = y_1.$$

Since $y_1 \in B_1$, $z = y_1 \neq x_2$. Since $z \leq x_2$, it follows that $z < x_2$. Thus $y_1 = z \in \overline{\mathrm{seg}_{B_2}(x_2)}$. This, however, contradicts the choice of y_1, so we conclude that $B_1 \setminus \mathrm{seg}_{B_2}(x_2) = \emptyset$, and so that $B_1 = \mathrm{seg}_{B_2}(x_2)$. Thus (ii) holds.

A swapping of the rôles of B_1 and B_2 will complete the proof. ▼

2 Lemma *The union of all* b*-sets is a* b*-set.*

Proof Let U denote the union of all b-sets. First we must show that U is well ordered. Let $A \subseteq U$ and let $x \in A$. Then there is a b-set B such that $x \in B$. We claim that $\mathrm{seg}_A(x) \subseteq B$. Indeed, if $x' < x$ then, by Lemma 1, either $x' \in B$ or x' does not lie in any b-set. Since A lies in the union of all b-sets, it must be the case that $x' \in B$. Thus $\mathrm{seg}_A(x)$ is a subset of the well ordered set B, and as such has a least element x_0. This is clearly also a least element for A, so U is well ordered.

Next, let $x \in U$ and let B be a b-set such that $x \in B$. Our above argument shows that $\mathrm{seg}_U(x) \subseteq B$ so that $\mathrm{seg}_U(x) = \mathrm{seg}_B(x)$. Therefore, $x = b(\mathrm{seg}_B(x)) = b(\mathrm{seg}_U(x))$. This completes the proof. ▼

To complete the proof, let U be the union of all b-sets and let $x = b(U)$. Then we claim that $U \cup \{x\}$ is a b-set. That $U \cup \{x\}$ is well ordered follows since U is well ordered and since x is an upper bound for U. Since U is the union of all b-sets, it must hold that $x \in U$. However, this contradicts the fact that x is a strict upper bound for U. ■

1.5.4 Induction and recursion

In some of the proofs we have given in this section, and in our definition of $\mathbb{Z}_{\geq 0}$, we have used the idea of induction. This idea is an eminently reasonable one. One starts with a fact or a definition that applies to the element $0 \in \mathbb{Z}_{\geq 0}$, and a rule for extending this from the jth number to the $(j+1)$st number, and then asserts that the fact or definition applies to all elements of $\mathbb{Z}_{\geq 0}$. In this section we formulate this principle in a more general setting that the set $\mathbb{Z}_{\geq 0}$, namely for a well ordered set.

Since the result will have to do with a property being true for the elements of a well ordered set, let us formally say that a ***property*** defined in a set S is a map $P \colon S \to \{\text{true}, \text{false}\}$. A property is ***true***, or ***holds***, at x if $P(x) = \text{true}$.

1.5.14 Theorem (Principle of Transfinite Induction) *Let* $(\mathrm{W}, \preceq)$ *be a well ordered set and let* P *be a property defined in* W*. Suppose that, for every* $\mathrm{w} \in \mathrm{W}$*, the fact that* $\mathrm{P}(\mathrm{w}')$ *is true for every* $\mathrm{w}' \prec \mathrm{w}$ *implies that* $\mathrm{P}(\mathrm{w})$ *is true. Then* $\mathrm{P}(\mathrm{w})$ *is true for every* $\mathrm{w} \in \mathrm{W}$*.*

Proof Suppose that the hypothesis is true, but the conclusion is false. Then

$$F = \{w \in W \mid P(w) = \text{false}\} \neq \emptyset.$$

Let w be the least element of F Therefore, for $w' < w$ it must hold that $P(w') = \text{true}$. But then the hypotheses imply that $P(w) = \text{true}$, so that $w \in W \setminus F$. This is a contradiction. ■

Next we turn to the process of defining something using recursion. As we did for induction, let us first consider doing this for $\mathbb{Z}_{\geq 0}$. What we wish to define is a map $f\colon \mathbb{Z}_{\geq 0} \to S$. The idea for doing this is that, if, for each $k \in \mathbb{Z}_{\geq 0}$, one knows the value of f on the first k elements of $\mathbb{Z}_{\geq 0}$, and if one knows a rule for then giving the value of f at $k+1$, then the f extends uniquely to a function on all of $\mathbb{Z}_{\geq 0}$. To give a concrete example, if $S = \mathbb{Z}$ and if we define $f(k+1) = 2 \cdot f(k)$, then the resulting function $f\colon \mathbb{Z}_{\geq 0} \to \mathbb{Z}$ is determined by its value at 0: $f(k) = 2^k \cdot f(0)$.

To state the general theorem requires some notation. We let W be a well ordered set and let S be a set. For $w \in W$, we let $\text{seq}_S(w)$ be the set of maps from $\text{seg}(w)$ into S. We then let $\text{Seq}_S(W)$ be the set of all maps of the form $g\colon \text{seq}_S(w) \to S$. The idea is that an element of $\text{S}_S(W)$ tells us how to extend a map from $\text{seg}(w)$ to give its value at w.

The desired result is now the following.

1.5.15 Theorem (Transfinite recursion) *Let* $(\text{W}, \preceq)$ *be a well ordered set and let* S *be a set. Given a member* $\text{g} \in \text{Seq}_\text{S}(\text{W})$*, there exists a unique map* $\text{f}_\text{g}\colon \text{W} \to \text{S}$ *such that* $\text{f}_\text{g}(\text{w}) = \text{g}(\text{f}|\,\text{seg}(\text{w}))$.

Proof That there can be only one map f_g as in the theorem statement follows from the Principle of Transfinite Induction (take $P(w) = \text{true}$ if and only if $f_g(w) = g(f_g|\,\text{seg}(w))$).

So we shall prove the existence of f_g. Define

$$\begin{aligned}\mathscr{C}_g = \{&A \subseteq W \times S| \\ &w \in W, h \in \text{seq}_S(w), (w', h(w')) \in A \text{ for all } w' \in \text{seg}(w) \implies (w, g(h)) \in A\}.\end{aligned}$$

Note that $W \times S \in \mathscr{C}_g$, so that C_g is not empty. It is easy to check that the intersection of members of $\mathscr{C}_g$ is also a member of $\mathscr{C}_g$. Therefore we let $F_g = \cap_{A \in \mathscr{C}_g} A$, and note that $F_g \in \mathscr{C}_g$. We shall show that F_g is the graph of a function f_g that satisfies the conditions in the theorem statement.

First we need to show that, for each $w \in W$, there exists exactly one $x \in S$ such that $(w, x) \in F_g$. Define

$$A_g = \{w \in W \mid \text{ there exists exactly one } x \in S \text{ such that } (w, x) \in F_g\}.$$

For $w \in W$, we claim that if $\text{seg}(w) \subseteq A_g$, then $w \in A_g$. Indeed, if $\text{seg}(w) \subseteq A_g$, define $h \in \text{seq}_S(w)$ by $h(w') = x'$ where $x' \in S$ is the unique element such that $(w', x') \in A_g$. Since $F_g \in \mathscr{C}_g$, there exists some $x \in S$ such that $(w, x) \in F_g$. Suppose that $x \neq g(h)$. We claim that $F_g - \{(w, x)\} \in \mathscr{C}_g$. Let $w' \in W$ and let $h' \in \text{seg}_S(w')$ satisfy $(w'', h'(w'')) \in F_g - \{(w, x)\}$ for all $w'' \in \text{seg}(w')$. If $w' = w$ then $h' = h$ by the uniqueness assertion of the theorem, and therefore $(w', g(h')) \in F_g - \{(w, x)\}$ since $x \neq g(h) = g(h')$. On the other hand, if $w' \neq w$ then $(w', g(h')) \in F_g - \{(w, x)\}$ since $F_g \in \mathscr{C}_g$. Thus, indeed, $F_g - \{(w, x)\} \in \mathscr{C}_g$, contradicting the fact that F_g is the intersection of all sets in $\mathscr{C}_g$. Thus we can conclude

that $x = g(h)$, and therefore that there is exactly one $x \in S$ such that $(w, x) \in F_g$. By the Principle of Transfinite Induction, we can then conclude that for *every* $w \in W$, there is exactly one $x \in S$ such that $(w, x) \in F_g$. Thus F_g is the graph of a map $f_g \colon W \to S$.

It remains to verify that $f_g(w) = g(f_g|\operatorname{seg}(w))$. This, however, follows easily from the definition of F_g. ■

One of the features of transfinite induction and transfinite recursion that requires some getting used to is that, unlike the usual induction with natural numbers as the well ordered set, one does not begin the induction or recursion by starting at 0 (or, in the case of a well ordered set, the least element), and proceeding element by element. Rather, one deals with initial segments. The reason for this is that in a well ordered set one may not have an immediate predecessor for every element, so that cannot be part of the induction/recursion; so the initial segment serves this purpose instead.

1.5.5 Zermelo's Well Ordering Theorem

The final topic in this section is a somewhat counterintuitive one. It says that every set possesses as well order.

1.5.16 Theorem (Zermelo's[9] Well Ordering Theorem) *For every set* S*, there is a well order in* S*.*

Proof Define

$$\mathscr{W} = \{(W, \preceq_W) \mid W \subseteq S \text{ and } \preceq_W \text{ is a well order on } W\}.$$

Since $\emptyset \in \mathscr{W}$, $\mathscr{W}$ is nonempty. Define a partial order $\preceq$ on $\mathscr{W}$ by

$$W_1 \preceq W_2 \iff W_2 \text{ is similar to a segment of } W_1.$$

Suppose that $\mathscr{T}$ is a totally ordered subset of $\mathscr{W}$.

1 Lemma *The set* $\cup_{\mathrm{A}\in\mathscr{T}}\mathrm{A}$ *has a unique well ordering, denoted by* $\lesssim$*, such that* $\mathrm{A}' \lesssim \cup_{\mathrm{A}\in\mathscr{T}}$ *for all* $\mathrm{A}' \in \mathscr{T}$*.*

Proof Let $x_1, x_2 \in \cup_{A\in\mathscr{T}} A$, and let $W_1, W_2 \in \mathscr{T}$ have the property that $x_1 \in W_1$ and $x_2 \in W_2$. Note that since either $W_1 = W_2$, $W_1 \preceq W_2$, or $W_2 \preceq W_1$, it must be the case that x_1 and x_2 lie in the same set from $\mathscr{C}$, let us call this W. The order in $\cup_{A\in\mathscr{T}} A$ is then defined by giving to the points x_1 and x_2 their order in W. This is unambiguous since $\mathscr{T}$ is totally ordered. It is then a simple exercise, left to the reader, that this is a well order. ▼

The lemma ensures that the hypotheses of Zorn's Lemma apply to the totally ordered subsets of $\mathscr{W}$, and therefore the conclusions of Zorn's Lemma ensure that there is a maximal element W in $\mathscr{W}$. We claim that this maximal element is S. Suppose this is not the case, and that $x \in S - W$. We claim that $W \cup \{x\} \in \mathscr{W}$. To see this, simply define a well order on $W \cup \{x\}$ by asking that points in W have their usual order, and that x be greater that all points in W. The result is easily verified to be a well order on $W \cup \{x\}$, so contradiction the maximality of W. This completes the proof. ■

[9] Ernst Friedrich Ferdinand Zermelo (1871–1953) was a German mathematician whose mathematical contributions were mainly in the area of set theory.

It might be surprising that it should be possible to well order any set. A well order can be thought of as allowing an arranging of the elements in a set, starting from the least element, and moving upwards in order:

$$x_0 < x_1 < x_2 < \cdots .$$

The complicated thing to understand here are the "$\cdots$," since they only mean "and so on" with an appropriate interpretation of these words (this is entirely related to the idea of ordinal numbers discussed in Section 1.7.1). As an example, the reader might want to imagine trying to order the real numbers (which we define in Section 2.1). It might seem absurd that it is possible to well order the real numbers. However, this is one of the many counterintuitive consequences arising from set theory, in this case directly related to the Axiom of Choice (Section 1.8.3).

1.5.6 Similarity

Between partially ordered sets, there are classes of maps that are distinguished by their preserving of the order relation. In this section we look into these and some of their properties, particularly with respect to well orders.

1.5.17 Definition (Similarity) If $(S, \preceq_S)$ and $(T, \preceq_T)$ are partially ordered sets, a bijection $f : S \to T$ is a ***similarity***, and $(S, \preceq_S)$ and $(T, \preceq_T)$ are said to be ***similar***, if $f(x_1) \preceq_T f(x_2)$ if and only if $x_1 \preceq_S x_2$. •

Now we prove a few results relating to similarities between well ordered sets. These shall be useful in our discussion or ordinal numbers in Section 1.7.1.

1.5.18 Proposition (Similarities of a well ordered set with itself) *If* $(S, \preceq)$ *is a well ordered set and if* $f : S \to S$ *is a similarity, then* $x \preceq f(x)$ *for each* $x \in S$.

Proof Define $A = \{x \in S \mid f(x) \prec x\}$ and let x be the least element of A. Then, for any $x' < x$, we have $x' \preceq f(x')$. In particular, $f(x) \preceq f \circ f(x)$. But $f(x) < x$ implies that $f \circ f(x) < f(x)$, giving a contradiction. Thus $A = \emptyset$. ■

1.5.19 Proposition (Well ordered sets are similar in at most one way) *If* $f, g : S \to T$ *are similarities between well ordered sets* $(S, \preceq_S)$ *and* $(T, \preceq_T)$, *then* $f = g$.

Proof Let $h = f^{-1} \circ g$, and note that h is a similarity from S to itself. By Proposition 1.5.18 this implies that $x \preceq_S h(x)$ for each $x \in S$. Thus

$$\begin{aligned} & x \preceq_S f^{-1} \circ g(x), \qquad x \in S \\ \implies \quad & f(x) \preceq_T g(x), \qquad x \in S. \end{aligned}$$

Reversing the argument gives $g(x) \preceq_T f(x)$ for every $x \in S$. This gives the result. ■

1.5.20 Proposition (Well ordered sets are not similar to their segments) *If* $(S, \prec)$ *is a well ordered set and if* $x \in S$, *then* S *is not similar to* $\operatorname{seg}(x)$.

Proof If $f(x) \in \operatorname{seg}(x)$ then $f(x) < x$, contradiction Proposition 1.5.18. ■

The final result is the deepest of the results we give here, because it gives a rather simple structure to the collection of all well ordered sets.

1.5.21 Proposition (Comparing well ordered sets) *If* $(\mathrm{S}, \preceq_{\mathrm{S}})$ *and* $(\mathrm{T}, \preceq_{\mathrm{T}})$ *are well ordered sets, then one of the following statements holds:*

(i) S *and* T *are similar;*

(ii) there exists $\mathrm{x} \in \mathrm{S}$ *such that* seg(x) *and* T *are similar;*

(iii) there exists $\mathrm{y} \in \mathrm{T}$ *such that* seg(y) *and* S *are similar.*

Proof Define

$$S_0 = \{x \in S \mid \text{ there exists } y \in T \text{ such that } \mathrm{seg}(x) \text{ is similar to } \mathrm{seg}(y)\},$$

noting that S_0 is nonempty, since the segment of the least element in S is similar to the segment of the least element in T. Define $f\colon S_0 \to T$ by $f(x) = y$ where $\mathrm{seg}(x)$ is similar to $\mathrm{seg}(y)$. Note that this uniquely defines f by Propositions 1.5.19 and 1.5.20. We then take $T_0 = \mathrm{image}(f)$. If $S_0 = S$, then the result immediately follows. If $S_0 \subset S$, then we claim that $S_0 = \mathrm{seg}(x_0)$ for some $x_0 \in S$. Indeed, we simply take x_0 to be the least strict upper bound for S_0, and then apply the definition of S_0 to see that $S_0 = \mathrm{seg}(x_0)$. We next claim that $T_0 = T$. Indeed, suppose that $T_0 \subset T$, let y_0 be the least strict upper bound for T_0, and let x_0 be the least strict upper bound for S_0. We claim that $\mathrm{seg}(x_0)$ is similar to $\mathrm{seg}(y_0)$. Indeed, if this is not the case, then there exists $y < y_0$ such that $\mathrm{seg}(y)$ is not similar to a segment in S. However, this contradicts the definition of T_0. ■

1.5.7 Notes

The proof of Zorn's Lemma we give is from the paper of [Lewin 1991].

Exercises

1.5.1 Show that any set S possesses a partial order.

1.5.2 Give conditions on S under which the partial order $\subseteq$ on $\mathbf{2}^S$ is

(a) a total order or

(b) a well-order.

1.5.3 Given two partially ordered sets $(S, \preceq_S)$ and $(T, \preceq_T)$, we define a relation $\preceq_{S\times T}$ in $S \times T$ by

$$(x_1, y_1) \preceq_{S\times T} (x_2, y_2) \quad \Longleftrightarrow \quad (x_1 \prec_S x_2) \text{ or } (x_1 = x_2 \text{ and } y_1 \preceq_T y_2).$$

This is called the ***lexicographic order*** on $S \times T$. Show the following:

(a) the lexicographic order is a partial order;

(b) if $\preceq_S$ and $\preceq_T$ are total orders, then the lexicographic order is a total order.

1.5.4 Show that a partially ordered set $(S, \preceq)$ possesses at most one least element and/or at most one greatest element.

Section 1.6

Indexed families of sets and general Cartesian products

In this section we discuss general collections of sets, and general collections of members of sets. In Section 1.1.3 we considered Cartesian products of a finite collection of sets. In this section, we wish to extend this to allow for an arbitrary collection of sets. The often used idea of an index set is introduced here, and will come up on many occasions in the text.

Do I need to read this section? The idea of a general family of sets, and notions related to it, do not arise in a lot of places in these volumes. But they do arise. The ideas here are simple (although the notational nuances can be confusing), and so perhaps can be read through. But the reader in a rush can skip the material, knowing they can look back on it if necessary. •

1.6.1 Indexed families and multisets

Recall that when talking about sets, a set is determined only by the concept of membership. Therefore, for example, the sets $\{1, 2, 2, 1, 2\}$ and $\{1, 2\}$ are the same since they have the same members. However, what if one wants to consider a set with two 1's and three 2's? The way in which one does this is by the use of an index to label the members of the set.

1.6.1 Definition (Indexed family of elements) Let A and S be sets. An ***indexed family of elements*** of S with ***index set*** A is a map $f\colon A \to S$. The element $f(a) \in S$ is sometimes denoted as x_a and the indexed family is denoted as $(x_a)_{a\in A}$. •

missing stuff

With the notion of an indexed family we can make sense of "repeated entries" in a set, as is shown in the first of these examples.

1.6.2 Examples (Indexed family)

1. Consider the two index sets $A_1 = \{1, 2, 3, 4, 5\}$ and $A_2 = \{1, 2\}$ and let S be the set of natural numbers. Then the functions $f_1\colon A_1 \to S$ and $f_2\colon A_2 \to S$ defined by

$$f_1(1) = 1,\ f_1(2) = 2,\ f_1(3) = 2,\ f_1(4) = 1,\ f_1(5) = 2,$$
$$f_2(1) = 1,\ f_2(2) = 2,$$

give the indexed families $(x_1 = 1, x_2 = 2, x_3 = 2, x_4 = 1, x_5 = 2)$ and $(x_1 = 1, x_2 = 2)$, respectively. In this way we can arrive at a set with two 1's and three 2's, as desired. Moreover, each of the 1's and 2's is assigned a specific place in the list $(x_1, \dots, x_5)$.

2. Any set S gives rise in a natural way to an indexed family of elements of S indexed by S itself: $(x)_{x\in S}$. •

We can then generalise this notion to an indexed family of sets as follows.

1.6.3 Definition (Indexed family of sets) Let A and S be sets. An ***indexed family of subsets*** of S with ***index set*** A is an indexed family of elements of $\mathbf{2}^S$ with index set A. Thus an indexed family of subsets of S is denoted by $(S_a)_{a\in A}$ where $S_a \subseteq S$ for $a \in A$. •

We use the notation $\cup_{a\in A} S_a$ and $\cap_{a\in A} S_a$ to denote the union and intersection of an indexed family of subsets indexed by A. Similarly, when considering the disjoint union of an indexed family of subsets indexed by A, we define this to be

$$\mathring{\cup}_{a\in A} S_a = \cup_{a\in A}(\{a\} \times S_a).$$

Thus an element in the disjoint union has the form (a, x) where $x \in S_a$. Just as with the disjoint union of a pair of sets, the disjoint union of a family of sets keeps track of the set that element belongs to, now labelled by the index set A, along with the element. A family of sets $(S_a)_{a\in A}$ is ***pairwise disjoint*** if, for every distinct $a_1, a_2 \in A$, $S_{a_1} \cap S_{a_2} = \emptyset$.

Often when one writes $(S_a)_{a\in A}$, one omits saying that the family is "indexed by A," this being understood from the notation. Moreover, many authors will say things like, "Consider the family of sets $\{S_a\}$," so omitting any reference to the index set. In such cases, the index set is usually understood (often it is $\mathbb{Z}_{>0}$). However, we shall not use this notation, and will always give a symbol for the index set.

Sometimes we will simply say something like, "Consider a family of sets $(S_a)_{a\in A}$." When we say this, we tacitly suppose there to be a set S which contains each of the sets S_a as a subset; the union of the sets S_a will serve to give such a set.

There is an alternative way of achieving the objective of allowing sets where the same member appears multiple times.

1.6.4 Definition (Multiset, submultiset) A ***multiset*** is an ordered pair (S, ϕ) where S is a set and $\phi\colon S \to \mathbb{Z}_{\geq 0}$ is a map. A multiset (T, ψ) is a ***submultiset*** of (S, ϕ) if $T \subseteq S$ and if $\psi(x) \leq \phi(x)$ for every $x \in T$. •

This is best illustrated by examples.

1.6.5 Examples (Multisets)

1. The multiset alluded to at the beginning of this section is (S, ϕ) with $S = \{1, 2\}$, and $\phi(1) = 2$ and $\phi(2) = 3$. Note that some information is lost when considering the multiset (S, ϕ) as compared to the indexed family $(1, 2, 2, 1, 2)$; the order of the elements is now immaterial and only the number of occurrences is accounted for.
2. Any set S can be thought of as a multiset (S, ϕ) where $\phi(x) = 1$ for each $x \in S$.
3. Let us give an example of how one might use the notion of a multiset. Let $P \subseteq \mathbb{Z}_{>0}$ be the set of prime numbers and let S be the set $\{2, 3, 4, \dots\}$ of integers greater than 1. As we shall prove in Corollary **??**, every element $n \in S$ can be written in a unique way as $n = p_1^{k_1} \cdots p_m^{k_m}$ for distinct primes $p_1, \dots, p_m$ and for $k_1, \dots, k_m \in \mathbb{Z}_{>0}$. Therefore, for every $n \in S$ there exists a unique multiset (P, ϕ_n)

defined by

$$\phi_n(p) = \begin{cases} k_j, & p = p_j, \\ 0, & \text{otherwise}, \end{cases}$$

understanding that $k_1, \dots, k_m$ and $p_1, \dots, p_m$ satisfy $n = p_1^{k_1} \cdots p_m^{k_m}$. •

1.6.6 Notation (Sets and multisets from indexed families of elements) Let A and S be sets and let $(x_a)_{a \in A}$ be an indexed family of elements of S. If for each $x \in S$ the set $\{a \in A \mid x_a = x\}$ is finite, then one can associate to $(x_a)_{a \in A}$ a multiset (S, ϕ) by

$$\phi(x) = \operatorname{card}\{a \in A \mid x_a = x\}.$$

This multiset is denoted by $\{x_a\}_{a \in A}$. One also has a subset of S associated with the family $(x_a)_{a \in A}$. This is simply the set

$$\{x \in S \mid x = x_a \text{ for some } a \in A\}.$$

This set is denoted by $\{x_a \mid a \in A\}$. Thus we have three potentially quite different objects:

$$(x_a)_{a \in A}, \quad \{x_a\}_{a \in A}, \quad \{x_a \mid a \in A\},$$

arranged in decreasing order of information prescribed (be sure to note that the multiset in the middle is only defined when the sets $\{a \in A \mid x_a = x\}$ are finite). This is possibly confusing, although there is not much in it, really.

For example, the indexed family $(1, 2, 2, 1, 2)$ gives the multiset denoted $\{1, 1, 2, 2, 2\}$ and the set $\{1, 2\}$. Now, this is truly confusing since there is no notational discrimination between the *set* $\{1, 1, 2, 2, 2\}$ (which is simply the set $\{1, 2\}$) and the *multiset* $\{1, 1, 2, 2, 2\}$ (which is not the set $\{1, 2\}$). However, the notation is standard, and the hopefully the intention will be clear from context.

If the map $a \mapsto x_a$ is injective, i.e., the elements in the family $(x_a)_{a \in A}$ are distinct, then the three objects are in natural correspondence with one another. For this reason we can sometimes be a bit lax in using one piece of notation over another. •

1.6.2 General Cartesian products

Before giving general definitions, it pays to revisit the idea of the Cartesian product $S_1 \times S_2$ of sets S_1 and S_2 as defined in Section 1.1.3 (the reason for our change from S and T to S_1 and S_2 will become clear shortly). Let $A = \{1, 2\}$, and let $f : A \to S_1 \cup S_2$ be a map satisfying $f(1) \in S_1$ and $f(2) \in S_2$. Then $(f(1), f(2)) \in S_1 \times S_2$. Conversely, given a point $(x_1, x_2) \in S_1 \times S_2$, we define a map $f : A \to S_1 \cup S_2$ by $f(1) = x_1$ and $f(2) = x_2$, noting that $f(1) \in S_1$ and $f(2) \in S_2$.

The punchline is that, for a pair of sets S_1 and S_2, their Cartesian product is in 1–1 correspondence with maps f from $A = \{1, 2\}$ to $S_1 \cup S_1$ having the property that $f(x_1) \in S_1$ and $f(x_2) \in S_2$. There are two things to note here: (1) the use of the set A to label the sets S_1 and S_2 and (2) the alternative characterisation of the Cartesian product.

Now we generalise the Cartesian product to families of sets.

1.6.7 Definition (Cartesian product) The ***Cartesian product*** of a family of sets $(S_a)_{a\in A}$ is the set

$$\prod_{a\in A} S_a = \{f : A \to \cup_{a\in A} S_a \mid f(a) \in S_a\}.$$ •

Note that the analogue to the ordered pair in a general Cartesian product is simply the set $f(A)$ for some $f \in \prod_{a\in A} S_a$. The reader should convince themselves that this is indeed the appropriate generalisation.

1.6.3 Sequences

The notion of a sequence is very important for us, and we give here a general definition for sequences in arbitrary sets.

1.6.8 Definition (Sequence, subsequence) Let S be a set.

(i) A ***sequence*** in S is an indexed family $(x_j)_{j\in\mathbb{Z}_{>0}}$ of elements of S with index set $\mathbb{Z}_{>0}$.

(ii) A ***subsequence*** of a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in S is a map $f : A \to S$ where

(a) $A \subseteq \mathbb{Z}_{>0}$ is a nonempty set with no upper bound and

(b) $f(k) = x_k$ for all $k \in A$.

If the elements in the set A are ordered as $j_1 < j_2 < j_3 < \cdots$, then the subsequence may be written as $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$. •

Note that in a sequence the location of the elements is important, and so the notation $(x_j)_{j\in\mathbb{Z}_{>0}}$ is the correct choice. It is, however, not uncommon to see sequences denoted $\{x_j\}_{j\in\mathbb{Z}_{>0}}$. According to Notation 1.6.6 this would imply that the same element in S could only appear in the list $(x_j)_{j\in\mathbb{Z}_{>0}}$ a finite number of times. However, this is often not what is intended. However, there is seldom any real confusion induced by this, but the reader should simply be aware that our (not uncommon) notational pedantry is not universally followed.

1.6.4 Directed sets and nets

What we discuss in this section is a generalisation of the notion of a sequence. A sequence is a collection of objects where there is a natural order to the objects inherited from the total order of $\mathbb{Z}_{>0}$.

First we define the index sets for this more general type of sequence.

1.6.9 Definition (Directed set) A ***directed set*** is a partially ordered set $(D, \preceq)$ with the property that, for $x, y \in D$, there exists $z \in D$ such that $x \preceq z$ and $y \preceq z$. •

Thus for any two elements in a directed set D it is possible to find an element greater than either, relative to the specified partial order. Let us give some examples to clarify this.

1.6.10 Examples (Directed sets)

1. The set $(\mathbb{Z}_{>0}, \leq)$ is a directed set since clearly one can find a natural number exceeding any two specified natural numbers.
2. The partially ordered set $([0, \infty), \leq)$ is similarly a directed set.
3. The partially ordered set $((0, 1], \geq)$ is also a directed set since, given $x, y \in (0, 1]$, one can find an element of $(0, 1]$ which is smaller than either x or y.
4. Next take $D = \mathbb{R} \setminus \{x_0\}$ and consider the partial order $\preceq$ on D defined by $x \preceq y$ if $|x - x_0| \leq |y - y_0|$. This may be shown to be a directed set since, given two elements $x, y \in \mathbb{R} \setminus \{x_0\}$, one can find another element of $\mathbb{R} \setminus \{x_0\}$ which is closer to x_0 than either x or y.
5. Let S be a set with more than one element and consider the partially ordered set $(\mathbf{2}^S \setminus \{\emptyset\}, \preceq)$ specified by $A \preceq B$ if $A \supseteq B$. This is readily verified to be a partial order. However, this order does not make $(S, \supseteq)$ a directed set. Indeed, suppose that $A, B \in \mathbf{2}^S \setminus \{\emptyset\}$ are disjoint. Since the only set contained in both A and B is the empty set, it follows that there is no element $T \in \mathbf{2}^S \setminus \{\emptyset\}$ for which $A \supseteq T$ and $B \supseteq T$. •

The next definition is of the generalisation of sequences built on the more general notion of index set given by a directed set.

1.6.11 Definition (Net) Let $(D, \preceq)$ be a directed set. A ***net*** in a set S defined on D is a map $\phi \colon D \to S$ from D into S. •

As with a sequence, it is convenient to instead write $\{x_\alpha\}_{\alpha \in D}$ where $x_\alpha = \phi(\alpha)$ for a net. The idea here is that a net generalises the notion of a sequence to the case where the index set may not be countable and where the order is more general than the total order of $\mathbb{Z}$.

Exercises

1.6.1

Section 1.7

Ordinal numbers, cardinal numbers, cardinality

The notion of cardinality has to do with the "size" of a set. For sets with finite numbers of elements, there is no problem with "size." For example, it is clear what it means for one set with a finite number of elements to be "larger" or "smaller" than another set with a finite number of elements. However, for sets with infinite numbers of elements, can one be larger than another? If so, how can this be decided? In this section we see that there is a set, called the ***cardinal numbers***, which exactly characterises the "size" of all sets, just as natural numbers characterise the "size" if finite sets.

Do I need to read this section? The material in this section is used only slightly, so it can be thought of as "cultural," and hopefully interesting. Certainly the details of constructing the ordinal numbers, and then the cardinal numbers, plays no essential rôle in these volumes. The idea of cardinality comes up, but only in the simple sense of Theorem 1.7.12. •

1.7.1 Ordinal numbers

Ordinal numbers generalise the natural numbers. Recall from Section 1.4.1 that a natural number is a set, and moreover, from Section 1.4.2, a well ordered set. Indeed, the number $k \in \mathbb{Z}_{\geq 0}$ is, by definition,

$$k = \{0, 1, \ldots, k-1\}.$$

Moreover, note that, for every $j \in k$, $j = \text{seg}(j)$. This motivates our definition of the ordinal numbers.

1.7.1 Definition (Ordinal number) An ***ordinal number*** is a well ordered set $(o, \leq)$ with the property that, for each $x \in o$, $x = \text{seg}(x)$. •

Let us give some examples of ordinal numbers. The examples we give are all of "small" ordinals. We begin our constructions in a fairly detailed way, and then we omit the details as we move on, since the idea becomes clear after the initial constructions.

1.7.2 Examples (Ordinal numbers)

1. As we saw before we stated Definition 1.7.1, each nonnegative integer is an ordinal number.
2. The set $\mathbb{Z}_{\geq 0}$ is an ordinal number. This is easily verified, but discomforting. We are saying that the set of numbers is itself a new kind of number, an ordinal number. Let us call this ordinal number ω. Pressing on. . .
3. The successor $\mathbb{Z}^+_{\geq 0} = \mathbb{Z}_{\geq 0} \cup \{\mathbb{Z}_{\geq 0}\}$ is also an ordinal number, in just the same manner as a natural number is an ordinal number. This ordinal number is denoted by $\omega + 1$.

4. One carries on in this way defining ordinal numbers $\omega + (k+1) = (\omega + k)^+$.
5. Next we assume that there is a set containing ω and all of its successors. In axiomatic set theory, this follows from a construction like that justifying Assumption 1.4.3, along with another axiom (the Axiom of Substitution; see Section 1.8.2) saying, essentially, that we can repeat the process. Just as we did with the definition of $\mathbb{Z}_{\geq 0}$, we take the smallest of these sets of successors to arrive at a net set that is to ω as ω is to 0. As was $\omega = \mathbb{Z}_{\geq 0}$, we well order this set by the partial order $\subseteq$. This set is then clearly an ordinal number, and is denoted by $\omega 2$.
6. One now proceeds to construct the successors $\omega 2 + 1 = \omega 2^+$, $\omega 2 + 2 = (\omega 2 + 1)^+$, and so on. These new sets are also ordinal numbers.
7. The preceding process yields ordinal numbers $\omega, \omega 2, \omega 3$, and so on.
8. We now again apply the same procedure to define an ordinal number that is contains ω, $\omega 2$, etc. This set we denote by ω^2.
9. One then defines $\omega^2 + 1 = (\omega^2)^+$, $\omega^2 + 2 = (\omega^2 + 1)^+$, etc., noting that these two are all ordinal numbers.
10. Next comes $\omega^2 + \omega$, which is the set containing all ordinal numbers $\omega^2 + 1$, $\omega^2 + 2$, etc.
11. Then comes $\omega^2 + \omega + 1$, $\omega^2 + \omega + 2$, etc.
12. Following these is $\omega^2 + \omega 2$, $\omega^2 + \omega 2 + 1$, and so on.
13. Then comes $\omega^2 + \omega 3$, $\omega^2 + \omega 3 + 1$, and so on.
14. After ω^2, $\omega^2 + \omega$, $\omega^2 + \omega 2$, and so on, we arrive at $\omega^2 2$.
15. One then arrives at $\omega^2 2 + 1, \ldots, \omega^2 2 + \omega, \ldots, \omega^2 2 + \omega 2$, etc.
16. After $\omega^2 2$, $\omega^2 3$, and so on comes ω^3.
17. After ω, ω^2, ω^3, etc., comes ω^ω.
18. After ω, ω^ω, ω^{ω^ω}, etc., comes ϵ_0. The entire construction starts again from ϵ_0. Thus we get to $\epsilon_0 + 1$, $\epsilon_0 + 2$, and so on reproducing all of the above steps with an ϵ_0 in front of everything.
19. Then we get $\epsilon_0 2$, $\epsilon_0 3$, and so on up to $\epsilon_0 \omega$.
20. These are followed by $\epsilon_0 \omega^2$, $\epsilon_0 \omega^3$ and so on up to $\epsilon_0 \omega^\omega$.
21. Then comes $\epsilon_0 \omega^{\omega^\omega}$, etc.
22. These are followed by ϵ_0^2.
23. We hope the reader is getting the point of these constructions, and can produce more such ordinals derived from the natural numbers. •

The above constructions of examples of ordinal numbers suggests that there are a lot of them. However, the concrete constructions do not really do justice to the number of ordinals. The ordinals that are elements of $\mathbb{Z}_{\geq 0}$ are called ***finite*** ordinals, and all other ordinals are ***transfinite***. All of the ordinals we have named above are called "countable" (see Definition 1.7.13). There are many other ordinals not included in the above list, but before we can appreciate this, we first have to describe some properties of ordinals.

First we note that ordinals are exactly defined by similarity. More precisely, we have the following result.

1.7.3 Proposition (Similar ordinals are equal) *If* o_1 *and* o_2 *are similar ordinal numbers then* $\mathrm{o}_1 = \mathrm{o}_2$.

Proof Let $f : o_1 \to o_2$ be a similarity and define

$$S = \{x \in o_1 \mid f(x) = x\}.$$

We wish to show that $S = o_1$. Suppose that $\mathrm{seg}(x) \subseteq S$ for $x \in o_1$. Then x is the least element of $\mathrm{seg}(x)$ and, since f is a similarity, $f(x)$ is the least element of $f(\mathrm{seg}(x))$. Therefore, x and $f(x)$ both have $\mathrm{seg}(x)$ as their strict initial segment, by definition of S. Thus, by the definition of ordinal numbers, $x = f(x)$. The result now follows by the Principle of Transfinite Induction. ■

The next result gives a rather rigid structure to any set of ordinal numbers.

1.7.4 Proposition (Sets of ordinals are always well ordered) *If* O *is a set of ordinal numbers, then this set is well ordered by* $\subseteq$.

Proof First we claim that O is totally ordered. Let $o_1, o_2 \in O$ and note that these are both well ordered sets. Therefore, by Proposition 1.5.21, either $o_1 = o_2$, o_1 is similar to a strict initial segment in o_2, or o_2 is similar to a strict initial segment in o_1. In either of the last two cases, it follows from Proposition 1.7.3 that either o_1 is *equal* to a strict initial segment in o_2, or vice versa. Thus, either $o_1 \leq o_2$ or $o_2 \leq o_1$. Thus O is totally ordered, a fact we shall assume in the remainder of the proof.

Let $o \in O$. If $o \leq o'$ for every $o' \in O$, then o is the least member of O, and so O has a least member, namely o. If o is not the least member of O, then there exists $o' \in O$ such that $o' < o$. Thus $o' \in o$ and so the set $o \cap E$ is nonempty. Let o_0 be the least element of o. We claim that o_0 is also the least element of O. Indeed, let $o' \in O$. If $o' < o$ then $o' \in o \cap E$ and so $o_0 \leq o'$. If $o \leq o'$ then $o_0 < o'$, so showing that o_0 is indeed the least element of O. ■

Our constructions in Example 1.7.2, and indeed the definition of an ordinal number, suggest the true fact that every ordinal number has a successor that is an ordinal number. However, it may not be the case that an ordinal number has an immediate predecessor. For example, each of the ordinals that are natural numbers has an immediate predecessor, but the ordinal ω does not have an immediate predecessor. That is to say, there is no largest ordinal number strictly less ω.

Recall that the set $\mathbb{Z}_{\geq 0}$ was defined by being the smallest set, having a certain property, that contains all nonnegative integers. One can then ask, "Is there a set containing all ordinal numbers?" It turns out the definition of the ordinal numbers prohibits this.

1.7.5 Proposition (Burali-Forti[10] Paradox) *There is no set* $\mathbb{O}$ *having the property that, if* o *is an ordinal number, then* $\mathrm{o} \in \mathbb{O}$.

[10]Cesare Burali-Forti (1861–1931) was an Italian mathematician who made contributions to mathematical logic.

Proof Suppose that such a set $\mathbb{O}$ exists. We claim that $\operatorname{supp}\mathbb{O}$ exists and is an ordinal number. Indeed, we claim that $\operatorname{supp}\mathbb{O} = \cup_{o\in\mathbb{O}} o$. Note that the set $\cup_{o\in\mathbb{O}} o$ is well ordered by inclusion by Proposition 1.7.4. Clearly, $\cup_{o\in\mathbb{O}}$ is the smallest such set containing each $o \in \mathbb{O}$. Moreover, it is also clear from Proposition 1.7.4 that if $o' \in \cup_{o\in\mathbb{O}}$, then $o' = \operatorname{seg}(o')$. Thus $\operatorname{supp}\mathbb{O}$ exists, and is an ordinal number. Moreover, this order number is greater than all those in $\mathbb{O}$, thus showing that $\mathbb{O}$ cannot exist. ■

For our purposes, the most useful feature of the ordinal numbers is the following.

1.7.6 Theorem (Ordinal numbers can count the size of a set) *If $(S, \preceq)$ is a well ordered set, then there exists a unique ordinal number o_S with the property that S and o_S are similar.*

Proof The uniqueness follows from Proposition 1.7.3. Let $x_0 \in S$ have the property that if $x < x_0$ then $\operatorname{seg}(x)$ is similar to some (necessarily unique) ordinal. (Why does x_0 exist?) Now let $P(x, o)$ be the proposition "o is an ordinal number similar to $\operatorname{seg}(x)$". Then define the set of ordinal numbers

$$o_0 = \{o \mid \text{ for each } x \in \operatorname{seg}(x_0), \text{ there exists } o \text{ such that } P(x, o) \text{ holds}\}.$$

One can easily verify that o_0 is itself an ordinal number that is similar to $\operatorname{seg}(x_0)$. Therefore, the Principle of Transfinite Induction can be applied to show that S is similar to an ordinal number. ■

This theorem is important, because it tells us that the ordinal numbers are the same, essentially, as the well ordered sets. Thus one can use the two concepts interchangeably; this is not obvious from the definition of an ordinal number.

It is also possible to define addition and multiplication of ordinal numbers. Since we will not make use of this, let us merely sketch how this goes. For ordinal numbers o_1 and o_2, let $(S_1, \preceq_1)$ and $(S_2, \preceq_2)$ be well ordered sets similar to o_1 and o_2, respectively. Define a partial order in $S_1 \mathbin{\mathring{\cup}} S_2$ by

$$(i_1, x_1) \preceq_+ (i_2, x_2) \iff \begin{cases} i_1 = i_2,\ x_1 \preceq_{i_1}, & \text{or} \\ i_1 < i_2. \end{cases}$$

One may verify that this is a well order. Then define $o_1 + o_2$ as the unique ordinal number equivalent to the well ordered set $(S_1 \mathbin{\mathring{\cup}} S_2, \preceq_+)$. To define product of o_1 and o_2, on the Cartesian product $S_1 \times S_2$ consider the partial order

$$(x_1, x_2) \preceq_\times (y_1, y_2) \iff \begin{cases} x_2 \prec_2 y_2, & \text{or} \\ x_2 = y_2,\ x_1 \prec_1 y_1. \end{cases}$$

Again, this is verifiable as being a well order. One then defines $o_1 \cdot o_2$ to be the unique ordinal number similar to the well ordered set $(S_1 \times S_2, \preceq_\times)$. One must exercise care when dealing with addition and multiplication of ordinals, since, for example, neither addition nor multiplication are commutative. For example, $1 + \omega \neq \omega + 1$ (why?). However, since we do not make use of this arithmetic, we shall not explore this further. It is worth noting that the notation in Example 1.7.2 is derived from ordinal arithmetic. Thus, for example, $\omega 2 = \omega \cdot 2$, etc.

1.7.2 Cardinal numbers

The cardinal numbers, as mentioned at the beginning of this section, are intended to be measures of the size of a set. If one combines the Zermelo's Well Ordering Theorem (Theorem 1.5.16) and Theorem 1.7.6, one might be inclined to say that the ordinal numbers are suited to this task. Indeed, simply place a well order on the set of interest by Theorem 1.5.16, and then use the associated ordinal number, given by Theorem 1.7.6, to define "size." The problem with this construction is that this notion of the "size" of a set would depend on the choice of well ordering. As an example, let us take the set $\mathbb{Z}_{\geq 0}$. We place two well orderings on $\mathbb{Z}_{\geq 0}$, one being the natural well ordering $\leq$ and the other being defined by

$$k_1 \preceq k_2 \iff \begin{cases} k_1 \leq k_2,\ k_1, k_2 \in \mathbb{Z}_{>0}, & \text{or} \\ k_1 = k_2 = 0, & \text{or} \\ k_1 = 0,\ k_2 \in \mathbb{Z}_{>0}. \end{cases}$$

Thus, for the partial order $\preceq$, one places 0 after all other natural numbers. One then verifies that $(\mathbb{Z}_{\geq 0}, \preceq)$ is similar to the ordinal number ω and that $(\mathbb{Z}_{\geq 0}, \preceq)$ is similar to the ordinal number $\omega + 1$. Thus, even in a fairly simple example of a non-finite set, we see that the well order can change the size, if we go with size being determined by ordinals.

Therefore, we introduce a special subset of ordinals.

1.7.7 Definition (Cardinal number) A *cardinal number* is an ordinal number c with the property that, for all ordinal numbers o for which there exists a bijection from c to o, we have $c \leq o$. •

In other words, a cardinal number is the least ordinal number in a collection of ordinal numbers that are equivalent. Note that finite ordinals are only equivalent with a single ordinal, namely themselves. However, transfinite ordinals may be equivalent to different transfinite ordinals. The following example illustrates this.

1.7.8 Example (Equivalent transfinite ordinals) We claim that there is a 1–1 correspondence between ω and $\omega + 1$. We can establish this correspondence explicitly by defining a map $f\colon \omega \to \omega + 1$ by

$$f(x) = \begin{cases} \omega, & x = 0, \\ x - 1, & x \in \mathbb{Z}_{>0}, \end{cases}$$

where $x - 1$ denotes the immediate predecessor of $x \in \mathbb{Z}_{>0}$.

One can actually check that *all* of the ordinal numbers presented in Example 1.7.2 are equivalent to ω! This is a consequence of Proposition 1.7.16 below. Accepting this as fact for the moment, we see that the only ordinals from Example 1.7.2 that are cardinal numbers are the elements of $\mathbb{Z}_{\geq 0}$ along with ω. •

Certain of the facts about ordinal numbers translate directly to equivalent facts about cardinal numbers. Let us record these

1.7.9 Proposition (Properties of cardinal numbers) *The following statements hold:*

(i) *if* c_1 *and* c_2 *are similar cardinal numbers then* $\mathrm{c}_1 = \mathrm{c}_2$*;*

(ii) *if* $\mathbb{C}$ *is a set of cardinal numbers, then this set is well ordered by* $\subseteq$*;*

(iii) *there is no set* $\mathbb{C}$ *having the property that, if* c *is an cardinal number, then* $\mathrm{c} \in \mathbb{C}$ *(**Cantor's paradox**).*[11]

Proof The only thing that does not follow immediately from the corresponding results for ordinal numbers is Cantor's Paradox. The proof of this part of the result goes exactly as does that of Proposition 1.7.5. One only needs to verify that, if $\mathbb{C}$ is any set of cardinal numbers, then there exists a cardinal number greater or equal to $\operatorname{supp}\mathbb{C}$. This, however, is clear since $\operatorname{supp}\mathbb{C}$ is an ordinal number strictly greater than any element of $\mathbb{C}$, meaning that there is a corresponding cardinal number c equivalent to $\operatorname{supp}\mathbb{C}$. Thus $c \geq \operatorname{supp}\mathbb{C}$. ■

1.7.3 Cardinality

Cardinality is the measure of the "size" of a set that we have been after. The following result sets the stage for the definition.

1.7.10 Lemma *For a set* S *there exists a unique cardinal number* card(S) *such that* S *and* card(S) *are equivalent.*

Proof By Theorem 1.7.6 there exists an ordinal number o_S that is similar to S, and therefore equivalent to S. Any ordinal equivalent to o_S is therefore also equivalent to S, since equivalence of sets is an "equivalence relation" (Exercise 1.3.8). Therefore, the result follows by choosing the unique least element in the set of ordinals equivalent to o_S. ■

With this fact at hand, the following definition makes sense.

1.7.11 Definition (Cardinality) The ***cardinality*** of a set S is the unique cardinal number $\mathrm{card}(S)$ that is equivalent to S. •

The next result indicates how one often deals with cardinality in practice. The important thing to note is that, provided one is interested only in *comparing* cardinalities of sets, then one need not deal with the complication of cardinal numbers.***missing stuff***

1.7.12 Theorem (Cantor–Schröder–Bernstein[12] Theorem) *For sets* S *and* T*, the following statements are equivalent:*

(i) card(S) = card(T)*;*

[11]Georg Ferdinand Ludwig Philipp Cantor (1845–1918) was born in Denmark, grew up in St. Petersburg, and lived much of his mathematical life in Germany. He made many important contributions to set theory and logic. He is regarded as the founder of set theory as we now know it.

[12]Friedrich Wilhelm Karl Ernst Schröder (1814–1902) was a German mathematician whose work was in the area of mathematical logic. Felix Bernstein (1878–1956) was born in Germany. Despite his name being attached to a basic result in set theory, Bernstein's main contributions were in the areas of statistics, mathematical biology, and actuarial mathematics.

(ii) *there exists a bijection* $\mathrm{f}\colon \mathrm{S} \to \mathrm{T}$;

(iii) *there exists injections* $\mathrm{f}\colon \mathrm{S} \to \mathrm{T}$ *and* $\mathrm{g}\colon \mathrm{T} \to \mathrm{S}$;

(iv) *there exists surjections* $\mathrm{f}\colon \mathrm{S} \to \mathrm{T}$ *and* $\mathrm{g}\colon \mathrm{T} \to \mathrm{S}$.

Proof It is clear from Lemma 1.7.10 that (i) and (ii) are equivalent. It is also clear that (ii) implies both (iii) and (iv).

(iii) $\implies$ (ii) We start with a lemma.

1 Lemma *If* $\mathrm{A} \subseteq \mathrm{S}$ *and if there exists an injection* $\mathrm{f}\colon \mathrm{S} \to \mathrm{A}$, *then there exists a bijection* $\mathrm{g}\colon \mathrm{S} \to \mathrm{A}$.

Proof Define $B_0 = S \setminus A$ and then inductively define B_j, $j \in \mathbb{Z}_{>0}$, by $B_{j+1} = f(B_j)$. We claim that the sets $(B_j)_{j\in\mathbb{Z}_{\geq 0}}$ (this notation for a family of sets will be made clear in Section 1.6.1) are pairwise disjoint. Suppose not and let $(j,k) \in \mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$ be the least pair, with respect to the lexicographic ordering (see Exercise 1.5.3), for which $B_j \cap B_k \neq \emptyset$. Since clearly $B_0 \cap B_j = \emptyset$ for $j \in \mathbb{Z}_{>0}$, we can assume that $j = \tilde{j} + 1$ and $k = \tilde{k} + 1$ for $\tilde{j}, \tilde{k} \in \mathbb{Z}_{\geq 0}$, and so therefore that $B_j = f(B_{\tilde{j}})$ and $B_k = f(B_{\tilde{k}})$. Thus $f(B_{\tilde{j}} \cap B_{\tilde{k}}) \neq \emptyset$ by Proposition 1.3.5, and so $B_{\tilde{j}} \cap B_{\tilde{k}} \neq \emptyset$. Since $(\tilde{j}, \tilde{k})$ is less that (j,k) with respect to the lexicographic order, we have a contradiction.

Now let $B = \cup_{j\in\mathbb{Z}_{\geq 0}} B_j$ and define $g\colon S \to A$ by

$$g(x) = \begin{cases} f(x), & x \in B, \\ x, & x \notin B. \end{cases}$$

For $x \in B$, $g(x) = f(x) \in A$. For $x \notin B$, we have $x \in A$ by definition of B_0, so that g indeed takes values in A. By definition g is injective. Also, let $x \in A$. If $x \notin B$ then $g(x) = x$. If $x \in B$ then $x \in B_{j+1}$ for some $j \in \mathbb{Z}_{\geq 0}$. Since $B_{j+1} = f(B_j)$, $x \in \text{image}(g)$, so showing that g is surjective. ▼

We now continue with the proof of this part of the theorem. Note that $g \circ f\colon S \to g(T)$ is injective (cf. Exercise 1.3.5). Therefore, by the preceding lemma, there exists a bijection $h\colon S \to g(T)$. Since g is injective, $g\colon T \to g(T)$ is bijective, and let us denote the inverse by, abusing notation, $g^{-1}\colon g(T) \to T$. We then define $b\colon S \to T$ by $b = g^{-1} \circ h$, and leave it to the reader to perform the easy verification that b is a bijection.

(iv) $\implies$ (iii) Since f is surjective, by Proposition 1.3.9 there exists a right inverse $f_R\colon T \to S$. Thus $f \circ f_R = \mathrm{id}_T$. Thus f is a *left*-inverse for f_R, implying that f_R is injective, again by Proposition 1.3.9. In like manner, g being surjective implies that there is an injective map from S to T, namely a right-inverse for g. ■

Distinguished names are given to certain kinds of sets, based on their cardinality. Recall that ω is the cardinal number corresponding to the set of natural numbers.

1.7.13 Definition (Finite, countable, uncountable) A set S is:

(i) ***finite*** if $\text{card}(S) \in \mathbb{Z}_{\geq 0}$;

(ii) ***infinite*** if $\text{card}(S) \geq \omega$;

(iii) ***countable*** if $\text{card}(S) \in \mathbb{Z}_{\geq 0}$ or if $\text{card}(S) = \omega$;

(iv) ***countably infinite*** if $\text{card}(S) = \omega$;

(v) ***uncountable***, or ***uncountably infinite***, if $\text{card}(S) > \omega$. •

Let us give some examples illustrating the distinctions between the various notions of set size.

1.7.14 Examples (Cardinality)

1. All elements of $\mathbb{Z}_{\geq 0}$ are, of course, finite sets.
2. The set $\mathbb{Z}_{\geq 0}$ is countably infinite. Indeed, $\text{card}(\mathbb{Z}_{\geq 0}) = \omega$.
3. We claim that $\mathbf{2}^{\mathbb{Z}_{\geq 0}}$ is uncountable. More generally, we claim that, for any set S, $\text{card}(S) < \text{card}(\mathbf{2}^S)$. To see this, we shall show that any map $f\colon S \to \mathbf{2}^S$ is not surjective. For such a map, let

$$A_f = \{x \in S \mid x \notin f(x)\}.$$

We claim that $A_f \notin \text{image}(f)$. Indeed, suppose that $A_f = f(x)$. If $x \in A_f$ then $x \notin f(x) = A_f$ by definition of A_f; a contradiction. On the other hand, if $x \notin A_f$, then $x \in f(x) = A_f$; again a contradiction. We thus conclude that $A_f \notin \text{image}(f)$. Thus there is no surjective map from S to $\mathbf{2}^S$. There is, however, a surjective map from $\mathbf{2}^S$ to S; for example, for any $x_0 \in S$, the map

$$g(A) = \begin{cases} x, & A = \{x\}, \\ x_0, & \text{otherwise} \end{cases}$$

is surjective. Thus S is "smaller than" $\mathbf{2}^S$, or $\text{card}(S) < \text{card}(\mathbf{2}^S)$. •

1.7.15 Remark (Uncountable sets exist, Continuum Hypothesis) A consequence of the last of the preceding examples is that fact that uncountable sets exist since $\mathbf{2}^{\mathbb{Z}_{\geq 0}}$ has a cardinality strictly greater than that of $\mathbb{Z}_{\geq 0}$.

It is usual to denote the countable ordinal by $\aleph_0$ (pronounced "aleph zero" or "aleph naught"). The smallest uncountable ordinal is then denoted by $\aleph_1$. An easy way to characterise $\aleph_1$ is as follows. Note that the cardinal $\aleph_0$ has the property that each of its initial segments is finite. In like manner, $\aleph_1$ has the property that each of its segments is countable. This does not *define* $\aleph_1$, but perhaps gives the reader some idea what it is.

It is conjectured that there are no cardinal numbers between $\aleph_0$ and $\aleph_1$; this conjecture is called the ***Continuum Hypothesis***. For readers prepared to accept the existence of the real numbers (or to look ahead to Section 2.1), we comment that $\text{card}(\mathbb{R}) = \text{card}(\mathbf{2}^{\mathbb{Z}_{\geq 0}})$ (see Exercise 1.7.5). From this follows a slightly more concrete statement of the Continuum Hypothesis, namely the conjecture that $\text{card}(\mathbb{R}) = \aleph_1$. Said yet otherwise, the Continuum Hypothesis is the conjecture that, among the subsets of $\mathbb{R}$, the only possibilities are (1) countable sets and (2) sets having the same cardinality as $\mathbb{R}$. •

It is clear the finite union of finite sets is finite. The following result, however, is less clearly true.

1.7.16 Proposition (Countable unions of countable sets are countable) *Let $(S_j)_{j\in\mathbb{Z}_{\geq 0}}$ be a family of sets, each of which is countable. Then $\cup_{j\in\mathbb{Z}_{\geq 0}} S_j$ is countable.*

Proof Let us explicitly enumerate the elements in the sets S_j, $j \in \mathbb{Z}_{\geq 0}$. Thus we write $S_j = (x_{jk})_{k \in \mathbb{Z}_{\geq 0}}$. We now indicate how one constructs a surjective map f from $\mathbb{Z}_{\geq 0}$ to $\cup_{j \in \mathbb{Z}_{\geq 0}} S_j$:

$$f(0) = x_{00},\ f(1) = x_{01},\ f(2) = x_{10},\ f(3) = x_{02},\ f(4) = x_{11},\ f(5) = x_{20},\ f(6) = x_{03},\ f(7) = x_{12},\ f(8) = x_{21},\ f(9) = x_{30},\ f(10) = x_{04}, \ldots.$$

We leave it to the reader to examine this definition and convince themselves that, if it were continued indefinitely, it would include every element of the set $\cup_{j \in \mathbb{Z}_{>0}} S_j$ in the domain of f. ■

For cardinal numbers one can define arithmetic in a manner similar to, but not the same as, that for ordinal numbers. Given cardinal numbers c_1 and c_2 we let S_1 and S_2 be sets equivalent to (not necessarily similar to, note) c_1 and c_2, respectively. We then define $c_1 + c_2 = \text{card}(S_1 \mathbin{\mathring{\cup}} S_2)$ and $c_1 \cdot c_2 = \text{card}(S_1 \times S_2)$. Note that cardinal number arithmetic is not just ordinal number arithmetic restricted to the cardinal numbers. That is to say, for example, the sum of two cardinal numbers is *not* the ordinal sum of the cardinal numbers thought of as ordinal numbers. It is easy to see this with an example. If S and T are two countably infinite sets, then so too is $S \mathbin{\mathring{\cup}} T$ a countably infinite set (this is Proposition 1.7.16). Therefore, $\text{card}(S) + \text{card}(T) = \text{card}(S \mathbin{\mathring{\cup}} T) = \omega = \text{card}(S) = \text{card}(T)$. We can also define exponentiation of cardinal numbers. For cardinal numbers c_1 and c_2 we, as above, let S_1 and S_2 be sets equivalent to c_1 and c_2, respectively. We then define $c_1^{c_2} = \text{card}(S_1^{S_2})$, where we recall that $S_1^{S_2}$ denotes the set of maps from S_2 to S_1.

The only result that we shall care about concerning cardinal arithmetic is the following.

1.7.17 Theorem (Sums and products of infinite cardinal numbers) *If* c *is an infinite cardinal number then*

(i) $\mathrm{c} + \mathrm{k} = \mathrm{c}$ *for every finite cardinal number* k,

(ii) $\mathrm{c} = \mathrm{c} + \mathrm{c}$, *and*

(iii) $\mathrm{c} = \mathrm{c} \cdot \mathrm{c}$.

Proof (i) Let S and T be disjoint sets such that $\text{card}(S) = c$ and $\text{card}(T) = k$. Let $g \colon T \to \{1, \ldots, k\}$ be a bijection. Since S is infinite, we may suppose that S contains $\mathbb{Z}_{>0}$ as a subset. Define $f \colon S \cup T \to S$ by

$$f(x) = \begin{cases} g(x), & x \in T, \\ x + k, & x \in \mathbb{Z}_{>0} \subseteq S, \\ x, & x \in S \setminus \mathbb{Z}_{>0}. \end{cases}$$

This is readily seen to be a bijection, and so gives the result by definition of cardinal addition.

(ii) Let S be a set such that $\text{card}(S) = c$ and define

$$G(S) = \{(f, A) \mid A \subseteq S,\ f \colon A \times \{0, 1\} \to A \text{ is a bijection}\}.$$

If $A \subseteq S$ is countably infinite, then $\text{card}(A \times \{0,1\}) = \text{card}(A)$, and so $G(S)$ is not empty. Place a partial order $\preceq$ on $G(S)$ by $(f_1, A_1) \preceq (f_2, A_2)$ if $A_1 \subseteq A_2$ and if $f_2|A_1 = f_1$. This is readily verified to be a partial order. Moreover, if $\{(f_j, A_j) \mid j \in J\}$ is a totally ordered subset, then we define an upper bound (f, A) as follows. We take $A = \cup_{j\in J} A_j$ and $f(x,k) = f_j(x,k)$ where $j \in J$ is defined such that $x \in A_j$. One can now use Zorn's Lemma to assert the existence of a maximal element of $G(S)$ which we denote by (f, A). We claim that $S \setminus A$ is finite. Indeed, if $S \setminus A$ is infinite, then there exists a countably infinite subset B of $S \setminus A$. Let g be a bijection from $B \times \{0,1\}$ to B and note that the map $f \times g \colon (A \cup B) \times \{0,1\} \to A \cup B$ defined by

$$f \times g(x,k) = \begin{cases} f(x,k), & x \in A, \\ g(x,k), & x \in B \end{cases}$$

if then a bijection, thus contradicting the maximality of (f, A). Thus $S \setminus A$ is indeed finite. Finally, since $(f, A) \in G(S)$, we have $\text{card}(A) + \text{card}(A) = \text{card}(A)$. Also, $\text{card}(S) = \text{card}(A) + \text{card}(A \setminus S)$. Since $\text{card}(S \setminus A)$ is finite, by part (i) this part of the theorem follows.

(iii) Let S be a set such that $\text{card}(S) = c$ and define

$$F(S) = \{(f, A) \mid A \subseteq S,\ f \colon A \times A \to A f \text{ is a bijection}\}.$$

If $A \subseteq S$ is countably infinite, then $\text{card}(A \times A) = \text{card}(A)$ and so there exists a bijection from $A \times A$ to A. Thus $F(S)$ is not empty. Place a partial order $\preceq$ on $F(S)$ by asking that $(f_1, A_1) \preceq (f_2, A_2)$ if $A_1 \subseteq A_2$ and $f_2|A_1 \times A_1 = f_1$; we leave to the reader the straightforward verification that this is a partial order. Moreover, if $\{(f_j, A_j) \mid j \in J\}$ is a totally ordered subset, it is easy to define an upper bound (f, A) for this set as follows. Take $A = \cup_{j\in J} A_j$ and define $f(x,y) = f_j(x,y)$ where $j \in J$ is defined such that $(x,y) \in A_j \times A_j$. Thus, by Zorn's Lemma, there exists a maximal element (f, A) of $F(S)$. By definition of $F(S)$ we have $\text{card}(A)\,\text{card}(A) = \text{card}(A)$. We now show that $\text{card}(A) = \text{card}(S)$.

Clearly $\text{card}(A) \leq \text{card}(S)$ since $A \subseteq S$. Thus suppose that $\text{card}(A) < \text{card}(S)$. We now use a lemma.

1 Lemma *If* $\mathfrak{c}_1$ *and* $\mathfrak{c}_2$ *are cardinal numbers at least one of which is infinite, and if* $\mathfrak{c}_3$ *is the larger of* $\mathfrak{c}_1$ *and* $\mathfrak{c}_2$*, then* $\mathfrak{c}_1 + \mathfrak{c}_2 = \mathfrak{c}_3$.

Proof Let S_1 and S_2 be disjoint sets such that $\text{card}(S_1) = c_1$ and $\text{card}(S_2) = c_2$. Since $c_1 \leq c_3$ and $c_2 \leq c_3$ it follows that $c_1 + c_2 = c_3 + c_3$. Also, $\text{card}(c_3) \leq \text{card}(c_1) + \text{card}(c_2)$. The lemma now follows from part (ii). ▼

From the lemma we know that $\text{card}(S)$ is the larger of $\text{card}(A)$ and $\text{card}(S \setminus A)$, i.e., that $\text{card}(S) = \text{card}(S \setminus A)$. Therefore $\text{card}(A) < \text{card}(S \setminus A)$. Thus there exists a subset $B \subseteq (S \setminus A)$ such that $\text{card}(B) = \text{card}(A)$. Therefore,

$$\text{card}(A \times B) = \text{card}(B \times A) = \text{card}(B \times B) = \text{card}(A) = \text{card}(B).$$

Therefore,

$$\text{card}((A \times B) \cup (B \times A) \cup (B \times B)) = \text{card}(B)$$

by part (ii). Therefore, there exists a bijection g from $(A \times B) \cup (B \times A) \cup (B \times B)$ to B. Thus we can define a bijection $f \times g$ from

$$(A \cup B) \times (A \cup B) = (A \times A) \cup (A \times B) \cup (B \times A) \cup (B \times B)$$

to $A \cup B$ by

$$f \times g(x,y) = \begin{cases} f(x,y), & (x,y) \in A \times A, \\ g(x,y), & \text{otherwise.} \end{cases}$$

Since $A \subseteq (A \cup B)$ and since $f \times g|(A \times A) = f$, this contradicts the maximality of (f, A). Thus our assumption that $\text{card}(A) < \text{card}(S)$ is invalid. ■

The following corollary will be particularly useful.

1.7.18 Corollary (Sum and product of a countable cardinal and an infinite cardinal) *If* $\mathfrak{c}$ *is an infinite cardinal number then*

(i) $\mathfrak{c} \leq \mathfrak{c} + \text{card}(\mathbb{Z}_{>0})$ *and*

(ii) $\mathfrak{c} \leq \mathfrak{c} \cdot \text{card}(\mathbb{Z}_{>0})$.

Proof This follows from Theorem 1.7.17 since $\text{card}(\mathbb{Z}_{>0})$ is the smallest infinite cardinal number, and so $\text{card}(\mathbb{Z}_{>0}) \leq c$. ■

Exercises

1.7.1 Show that every element of an ordinal number is an ordinal number.

1.7.2 Show that any finite union of finite sets is finite.

1.7.3 Show that the Cartesian product of a finite number of countable sets is countable.

1.7.4 For a set S, as per Definition 1.3.1, let 2^S denote the collection of maps from the set S to the set 2. Show that $\text{card}(2^S) = \text{card}(\mathbf{2}^S)$, so justifying the notation $\mathbf{2}^S$ as the collection of subsets of S.
Hint: *Given a subset* A ⊆ S, *think of a natural way of assigning a map from* S *to* 2.

In the next exercise you will show that $\text{card}(\mathbb{R}) = \text{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$. We refer to Section 2.1 for the definition of the real numbers. There the reader can also find the definition of the rational numbers, as these are also used in the next exercise.

1.7.5 Show that $\text{card}(\mathbb{R}) = \text{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$ by answering the following questions.
Define $f_1 \colon \mathbb{R} \to \mathbf{2}^{\mathbb{Q}}$ by

$$f_1(x) = \{q \in \mathbb{Q} \mid \ q \leq x\}.$$

(a) Show that f_1 is injective to conclude that $\text{card}(\mathbb{R}) \leq \text{card}(\mathbf{2}^{\mathbb{Q}})$.

(b) Show that $\text{card}(\mathbf{2}^{\mathbb{Q}}) = \text{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$, and conclude that $\text{card}(\mathbb{R}) \leq \text{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$.

Let $\{0,2\}^{\mathbb{Z}_{>0}}$ be the set of maps from $\mathbb{Z}_{>0}$ to $\{0,2\}$, and regard $\{0,2\}^{\mathbb{Z}_{>0}}$ as a subset of $[0,1]$ by thinking of $\{0,2\}^{\mathbb{Z}_{>0}}$ as being a sequence representing a decimal expansion in base 3. That is, to $f \colon \mathbb{Z}_{>0} \to \{0,2\}$ assign the real number

$$f_2(f) = \sum_{j=1}^{\infty} \frac{f(j)}{3^j}.$$

Thus f_2 is a map from $\{0,2\}^{\mathbb{Z}_{>0}}$ to $[0,1]$.

(c) Show that f_2 is injective so that $\mathrm{card}(\{0,2\}^{\mathbb{Z}_{>0}}) \leq \mathrm{card}([0,1])$.
(d) Show that $\mathrm{card}([0,1]) \leq \mathrm{card}(\mathbb{R})$.
(e) Show that $\mathrm{card}(\{0,2\}^{\mathbb{Z}_{>0}}) = \mathrm{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$, and conclude that $\mathrm{card}(\mathbf{2}^{\mathbb{Z}_{>0}}) \leq \mathrm{card}(\mathbb{R})$.
Hint: *Use Exercise 1.7.4.*

This shows that $\mathrm{card}(\mathbb{R}) = \mathrm{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$, as desired.

Section 1.8

Some words on axiomatic set theory

The account of set theory in this chapter is, as we said at the beginning of Section 1.1, called "naïve set theory." It turns out that the lack of care in saying what a set *is* in naïve set theory causes some problems. We indicate the nature of these problems in Section 1.8.1. To get around these problems, the presently accepted technique is the define a set as an element of a collection of objects satisfying certain axioms. This is called ***axiomatic set theory***, and we refer the reader to the notes at the end of the chapter for references. The most commonly used such axioms are those of Zermelo–Fränkel set theory, and we give these in Section 1.8.2. There are alternative collections of axioms, some equivalent to the Zermelo–Fränkel axioms, and some not. We shall not discuss this here. An axiom commonly, although not incontroversially, accepted is the Axiom of Choice, which we discuss in Section 1.8.3. We also discuss the Peano Axioms in Section 1.8.4, as these are the axioms of arithmetic. We close with a discussion of some of the issues in set theory, since these are of at least cultural interest.

Do I need to read this section? The material in this section is used exactly nowhere else in the texts. However, we hope the reader will find the informal presentation, and historical slant, interesting. •

1.8.1 Russell's Paradox

Russell's Paradox[13] is the following. Let S be the set of all sets that are not members of themselves. For example, the set P of prime numbers is in S since the set of prime numbers is not a prime number. However, the set N of all things that are not prime numbers is in S since the set of all things that are not prime numbers is not a prime number. Now argue as follows. Suppose that $S \in S$. Then S is a set that does not contain itself as a member; that is, $S \notin S$. Now suppose that $S \notin S$. Then S is a set that does not contain itself as a member; that is, $S \in S$. This is clearly absurd, so the set S cannot exist, although there seems to be nothing wrong with its definition. That a contradiction can be derived from the naïve version of set theory means that it is ***inconsistent***.

A consequence of Russell's Paradox is that there is no set containing all sets. Indeed, let S be any set. Then define

$$T = \{x \in S \mid x \notin x\}.$$

We claim that $T \notin S$. Indeed, suppose that $T \in S$. Then either $T \in T$ or $T \notin T$. In the first instance, since $T \in S$, $T \notin T$. In the second instance, again since $T \in S$, we have

[13] So named for Bertrand Arthur William Russell (1872–1970), who was a British philosopher and mathematician. Russell received a Nobel prize for literature in recognition of his popular writings on philosophy.

$T \notin T$. This is clearly a contradiction, and so we have concluded that, for every set S, there exists something that is not in A. Thus there can be no "set of sets."

Another consequence of Russell's Paradox is the ridiculous conclusion that everything is true. This is a simply logical consequence of the fact that, if a contradiction holds, then all statements hold. Here a contradiction means that a proposition P and its negation $\neg P$ both hold. The argument is as follows. Consider a proposition P'. Then P or P' holds, since P holds. However, since $\neg P$ holds and either P or P' holds, it must be the case that P' holds, no matter what P' is!

Thus the contradiction arising from Russell's Paradox is unsettling since it now calls into question any conclusions that might arise from our discussion of set theory. Various attempts were made to eliminate the eliminate the inconsistency in the naïve version of set theory. The presently most widely accepted of these attempts is the collection of axioms forming Zermelo–Fränkel set theory.

1.8.2 The axioms of Zermelo–Fränkel set theory

The axioms we give here are the culmination of the work of Ernst Friedrich Ferdinand Zermelo (1871–1953) and Adolf Abraham Halevi Fränkel (1891–1965).[14] The axioms were constructed in an attempt to arrive at a basis for set theory that was free of inconsistencies. At present, it is unknown whether the axioms of Zermelo–Fränkel set theory, abbreviated ***ZF***, are consistent.

Here we shall state the axioms, give a slight discussion of them, and indicate some of the places in the chapter where the axioms were employed.

The first axiom merely says that two sets are equal if they have the same elements. This is not controversial, and we have used this axiom out of hand throughout the chapter.

Axiom of Extension For sets S and T, if $x \in S$ if and only if $x \in T$, then $S = T$. •

The next axiom indicates that one can form the set of elements for which a certain property holds. Again, this is not controversial, and is an axiom we have used throughout the chapter.

Axiom of Separation For a set S and a property P defined in S, there exists a set A such that $x \in A$ if and only if $x \in S$ and $P(x) = \text{true}$. •

We also have an axiom which says that one can extract two members from two sets, and think of these as members of another set. This is another uncontroversial axiom that we have used without much fuss.

Axiom of the Unordered Pair For sets S_1 and S_2 and for $x_1 \in S_1$ and $x_2 \in S_2$, there exists a set T such that $x \in T$ if and only if $x = x_1$ or $x = x_2$. •

To form the union of two sets, one needs an axiom asserting that the union exists. This is natural, and we have used it whenever we use the notion of union, i.e., frequently.

[14] Fränkel was a German mathematician who worked primarily in the areas of set theory and mathematical logic.

Axiom of Union For sets S_1 and S_2 there exists a set T such that $x \in T$ if and only if $x \in S_1$ or $x \in S_2$. •

The existence of the power set is also included in the axioms. It is natural and we have used it frequently.

Axiom of the Power Set For a set S there exists a set T such that $A \in T$ if and only if $A \subseteq S$. •

When we constructed the set of natural numbers, we needed an axiom to ensure that this set existed (cf. Assumption 1.4.3). This axiom is the following.

Axiom of Infinity There exists a set S such that

(i) $\emptyset \in S$ and

(ii) for each $x \in S$, $x^+ \in S$. •

When we constructed a large number of ordinal numbers in Example 1.7.2, we repeatedly used an axiom, the essence of which was, "The same principle used to assert the existence of $\mathbb{Z}_{\geq 0}$ can be applied to this more general setting." Let us now state this idea more formally.

Axiom of Substitution For a set S, if for all $x \in S$ there exists a unique y such that $P(x, y)$ holds, then there exists a set T and a map $f \colon S \to T$ such that $f(x) = y$ where $P(x, y) = \text{true}$. •

The idea is that, for each $x \in S$, the collection of objects y for which $P(x, y)$ holds forms a set. Let us illustrate how the Axiom of Substitution can be used to define the ordinal number $\omega 2$, as in Example 1.7.2. For $k \in \mathbb{Z}_{\geq 0}$ we define

$$P(k, y) = \begin{cases} \text{true}, & y = \omega + k, \\ \text{false}, & \text{otherwise.} \end{cases}$$

The Axiom of Substitution then says that there is a set T and a map $f \colon \mathbb{Z}_{\geq 0} \to T$ such that $f(k) = \omega + k$. The ordinal number $\omega 2$ is then simply the image of the map f.

The final axiom in ZF is the one whose primary purpose is to eliminate inconsistencies such as those arising from Russell's Paradox.

Axiom of Regularity For each nonempty set S there exists $x \in S$ such that $x \cap S = \emptyset$. •

The Axiom of Regularity rules out sets like $S = \{S\}$ whose only members are themselves. It is no great loss having to live without such sets.

1.8.3 The Axiom of Choice

The Axiom of Choice has its origins in Zermelo's proof of his theorem that every set can be well ordered. In order to prove the theorem, he had to introduce a new axiom in addition to those accepted at the time to characterise sets. The new axiom is the following.

Axiom of Choice For each family $(S_a)_{a\in A}$ of nonempty sets, there exists a function, $f : A \to \cup_{a\in A} S_a$, called a ***choice function***, having the property that $f(a) \in S_a$. •

The combination of the axioms of ZF with the Axiom of Choice is sometimes called ***ZF with Choice***, or ***ZFC***. Work of Cohen[15] shows that the Axiom of Choice is independent of the axioms of ZF. Thus, when one adopts ZFC, the Axiom of Choice is really something additional that one is adding to one's list of assumptions of set theory.

At first glance, the Axiom of Choice, at least in the form we give it, does not seem startling. It merely says that, from any collection of sets, it is possible to select an element from each set. A trivial rephrasing of the Axiom of Choice is that, for any family $(S_a)_{a\in A}$ of nonempty sets, the Cartesian product $\prod_{a\in A} S_a$ is nonempty.

What is less settling about the Axiom of Choice is that it can lead to some non-intuitive conclusions. For example, as mentioned above, Zermelo's Well Ordering Theorem follows from the Axiom of Choice. Indeed, the two are equivalent. Let us, in fact, list the equivalence of the Axiom of Choice with two other important results from the chapter, one of which is Zermelo's Well Ordering Theorem.

1.8.1 Theorem (Equivalents of the Axiom of Choice) *If the axioms of ZF hold, then the following statements are equivalent:*

(i) the Axiom of Choice holds;

(ii) Zorn's Lemma holds;

(iii) Zermelo's Well Ordering Theorem holds.

Proof Let us suppose that the proofs we give of Theorems 1.5.13 and 1.5.16 are valid using the axioms of ZF. This is true, and can be verified, if tediously. One only needs to check that no constructions, other than those allowed by the axioms of ZF were used in the proofs. Assuming this, the implications (i) $\implies$ (ii) and (ii) $\implies$ (iii) hold, since these are what is used in the proofs of Theorems 1.5.13 and 1.5.16. It only remains to prove the implication (iii) $\implies$ (i). However, this is straightforward. Let $(S_a)_{a\in A}$ be a family of sets. By Zermelo's Well Ordering Theorem, well order each of these sets, and then define a choice function by assigning to $a \in A$ the least member of S_a. ■

There are, in fact, many statements that are equivalent to the Axiom of Choice. For example, the fact that a surjective map possesses a right-inverse is equivalent to the Axiom of Choice. In Exercise 1.8.1 we give a few of the more easily proved equivalents of the Axiom of Choice. At the time of its introduction, the equivalence of the Axiom of Choice with Zermelo's Well Ordering Theorem led many mathematicians to reject the validity of the Axiom of Choice. Zermelo, however, countered that many mathematicians implicitly used the Axiom of Choice without saying so. This then led to much activity in mathematics along the lines of deciding which results *required* the Axiom of Choice for their proof. Results can then be divided into three groups, in ascending order of "goodness," where the Axiom of Choice is deemed "bad":

[15] Paul Joseph Cohen was born in the United States in 1934, and has made outstanding contributions to the foundations of mathematics and set theory.

1. results that are equivalent to the Axiom of Choice;
2. results that are not equivalent to the Axiom of Choice, but can be shown to require it for their proof;
3. results that are true, whether or not the Axiom of Choice holds.

Somewhat more startling is that, if one accepts the Axiom of Choice, then it is possible to derive results which seem absurd. Perhaps the most famous of these is the ***Banach–Tarski Paradox***,[16] which says, very roughly, that it is possible to divide a sphere into a finite number of pieces and then reassemble them, while maintaining their shape, into two spheres of equal volume. Said in this way, the result seems impossible. However, if one looks at the result carefully, the nature of the pieces into which the sphere is divided is, obviously, extremely complicated. In the language of Chapter **??**, they are nonmeasurable sets. Such sets correspond poorly with our intuition, and indeed require the Axiom of Choice to assert their existence. We shall give a proof of the Banach–Tarski Paradox in Section **??**.

On the flip side of this is the fact that there are statements that seem like they *must* be true, and that are equivalent to the Axiom of Choice. One such statement is the Trichotomy Law for the real numbers, which says that, given two real numbers x and y, either $x < y$, $y < x$, or $x = y$. If rejecting the Axiom of Choice means rejecting the Trichotomy Law for real numbers, then many mathematicians would have to rethink the way they do mathematics!***missing stuff***

Indeed, there is a branch of mathematics that is dedicated to just this sort of rethinking, and this is called ***constructivism***; see the notes at the end of the chapter for references. The genesis of this branch of mathematics is the dissatisfaction, often arising from applications of the Axiom of Choice, with nonconstructive proofs in mathematics (for example, our proof that a surjective map possesses a right-inverse).

In this book, we will unabashedly assume the validity of the Axiom of Choice. In doing so, we follow in the mainstream of contemporary mathematics.

1.8.4 Peano's axioms

Peano's axioms[17] were derived in order to establish a basis for arithmetic. They essentially give those properties of the set of "numbers" that allow the establishment of the usual laws for addition and multiplication of natural numbers. ***Peano's axioms*** are these:

1. $0 = \emptyset$ is a number;
2. if k is a number, the successor of k is a number;
3. there is no number for which 0 is a successor;
4. if $j^+ = k^+$ then $j = k$ for all numbers j and k;

[16] Stefan Banach (1892–1945) was a well-known Polish mathematician who made significant and foundational contributions to functional analysis. Alfred Tarski (1902–1983) was also Polish, and his main contributions were to set theory and mathematical logic.

[17] Named after Giuseppe Peano (1858–1932), an Italian mathematician who did work with differential equations and set theory.

5. if S is a set of numbers containing 0 and having the property that the successor of every element of S is in S, then S contains the set of numbers.

Peano's axioms, since they led to the integers, and so there to the rational and real numbers (as in Section 2.1), were once considered as the basic ingredient from which all the rest of mathematics stemmed. This idea, however, received a blow with the publication of a paper by Kurt Gödel[18]. Gödel showed that in any logical system sufficiently general to include the Peano axioms, there exist statements whose truth cannot be validated within the axioms of the system. Thus, this showed that any system built on arithmetic could not possibly be self-contained.

1.8.5 Discussion of the status of set theory

In this section, we have painted a picture of set theory that suggests it is something of a morass of questionable assumptions and possibly unverifiable statements. There is some validity in this, in the sense that there are many fundamental questions unanswered. However, we shall not worry much about these matters as we proceed onto more concrete topics.

1.8.6 Notes

There are many general references for axiomatic set theory. We cite [Suppes 1960]***missing stuff***

The independence of the Axiom of Choice from the ZF axioms was proved in [Cohen 1963]. An interesting book on the Axiom of Choice is that of Moore [1982]. Constructivism is discussed by [Bridges and Richman 1987], for example. It is the paper of Gödel [1931] where the incompleteness of axiomatic systems which contain the Peano axioms is proved.

Exercises

1.8.1 Prove the following result.

Theorem *If the axioms of ZF hold, then the following statements are equivalent:*

(i) the Axiom of Choice holds;

(ii) for any family $(S_a)_{a\in A}$ *of sets, the Cartesian product* $\prod_{a\in A} S_a$ *is nonempty;*

(iii) every surjective map possesses a right inverse.

[18]Kurt Gödel (1906–1978) was born in a part of the Austro-Hungarian Empire that is now Czechoslovakia. He made outstanding contributions to the subject of mathematical logic.

Section 1.9

Some words about proving things

Rigour is an important part of the presentation in this series, and if you are so unfortunate as to be using these books as a text, then hopefully you will be asked to prove some things, for example, from the exercises. In this section we say a few (almost uselessly) general things about techniques for proving things. We also say some things about poor proof technique, much (but not all) of which is delivered with tongue in cheek. The fact of the matter is that the best way to become proficient at proving things is to (1) read a lot of (needless to say, good) proofs, and (2) most importantly, get lots of practice. What is certainly true is that it much easier to begin your theorem-proving career by proving simple things. In this respect, the proofs and exercises in this chapter are good ones. Similarly, many of the proofs and exercises in Chapters **??** and **??** provide a good basis for honing one's theorem-proving skills. By contrast, some of the results in Chapter 2 are a little more sophisticated, while still not difficult. As we progress through the preparatory material, we shall increasingly encounter material that is quite challenging, and so proofs that are quite elaborate. The neophyte should not be so ambitious as to tackle these early on in their mathematical development.

Do I need to read this section? Go ahead, read it. It will be fun. •

1.9.1 Legitimate proof techniques

The techniques here are the principle ones use in proving simple results. For very complicated results, many of which appear in this series, one is unlikely to get much help from this list.

1. *Proof by definition:* Show that the desired proposition follows directly from the given definitions and assumptions. Theorems that have already been proven to follow from the definitions and assumptions may also be used. Proofs of this sort are often abbreviated by "This is obvious." While this may well be true, it is better to replace this hopelessly vague assertion with something more meaningful like "This follows directly from the definition."
2. *Proof by contradiction:* Assume that the hypotheses of the desired proposition hold, but that the conclusions are false, and make no other assumption. Show that this leads to an impossible conclusion. This implies that the assumption must be false, meaning the desired proposition is true.
3. *Proof by induction:* In this method one wishes to prove a proposition for an enumerable number of cases, say $1, 2, \ldots, n, \ldots$. One first proves the proposition for case 1. Then one proves that, if the proposition is true for the nth case, it is true for the $(n + 1)$st case.
4. *Proof by exhaustion:* One proves the desired proposition to be true for all cases. This method only applies when there is a *finite* number of cases.

5. *Proof by contrapositive:* To show that proposition A implies proposition B, one shows that proposition B *not* being true implies that proposition A is *not* true. It is common to see newcomers get proof by contrapositive and proof by contradiction confused.
6. *Proof by counterexample:* This sort of proof is typically useful in showing that some general assertion *does not* hold. That is to say, one wishes to show that a certain conclusion does not follow from certain hypotheses. To show this, it suffices to come up with a single example for which the hypotheses hold, but the conclusion does not. Such an example is called a ***counterexample***.

1.9.2 Improper proof techniques

Many of these seem so simple that a first reaction is, "Who would be dumb enough to do something so obviously incorrect." However, it is easy, and sometimes tempting, to hide one of these incorrect arguments inside something complicated.

1. *Proof by reverse implication:* To prove that A implies B, shows that B implies A.
2. *Proof by half proof:* One is required to show that A and B are equivalent, but one only shows that A implies B. Note that the appearance of "if and only if" means that you have two implications to prove!
3. *Proof by example:* Show only a single case among many. Assume that only a single case is sufficient (when it is not) or suggest that the proof of this case contains most of the ideas of the general proof.
4. *Proof by picture:* A more convincing form of proof by example. Pictures can provide nice illustrations, but suffice in no part of a rigorous argument.
5. *Proof by special methods:* You are allowed to divide by zero, take wrong square roots, manipulate divergent series, etc.
6. *Proof by convergent irrelevancies:* Prove a lot of things related to the desired result.
7. *Proof by semantic shift:* Some standard but inconvenient definitions are changed for the statement of the result.
8. *Proof by limited definition:* Define (or implicitly assume) a set S, for which all of whose elements the desired result is true, then announce that in the future only members of the set S will be considered.
9. *Proof by circular cross-reference:* Delay the proof of a lemma until many theorems have been derived from it. Use one or more of these theorems in the proof of the lemma.
10. *Proof by appeal to intuition:* Cloud-shaped drawings frequently help here.
11. *Proof by elimination of counterexample:* Assume the hypothesis is true. Then show that a counterexample cannot exist. (This is really just a well-disguised proof by reverse implication.) A common variation, known as "begging the question" involves getting deep into the proof and then using a step that assumes the hypothesis.

12. *Proof by obfuscation:* A long plotless sequence of true and/or meaningless syntactically related statements.
13. *Proof by cumbersome notation:* Best done with access to at least four alphabets and special symbols. Can help make proofs by special methods look more convincing.
14. *Proof by cosmology:* The negation of a proposition is unimaginable or meaningless.
15. *Proof by reduction to the wrong problem:* To show that the result is true, compare (reduce/translate) the problem (in)to another problem. This is valid if the other problem is then solvable. The error lies in comparing to an unsolvable problem.

Exercises

1.9.1 Find the flaw in the following inductive "proof" of the fact that, in any class, if one selects a subset of students, they will have received the same grade.

> Suppose that we have a class with students $S = \{S_1, \ldots, S_m\}$. We shall prove by induction on the size of the subset that any subset of students receive the same grade. For a subset $\{S_{j_1}\}$, the assertion is clearly true. Now suppose that the assertion holds for all subsets of S with k students with $k \in \{1, \ldots, l\}$, and suppose we have a subset $\{S_{j_1}, \ldots, S_{j_l}, S_{j_{l+1}}\}$ of $l + 1$ students. By the induction hypothesis, the students from the set $\{S_{j_1}, \ldots, S_{j_l}\}$ all receive the same grade. Also by the induction hypothesis, the students from the set $\{S_2, \ldots, S_{j_l}, S_{j_{l+1}}\}$ all receive the same grade. In particular, the grade received by student $S_{j_{l+1}}$ is the same as the grade received by student S_{j_l}. But this is the same as the grade received by students $S_{j_1}, \ldots, S_{j_{l-1}}$, and so, by induction, we have proved that all students receive the same grade.

In the next exercise you will consider one of Zeno's paradoxes. Zeno[19] is best known for having developed a collection of paradoxes, some of which touch surprisingly deeply on mathematical ideas that were not perhaps fully appreciated until the 19th century. Many of his paradoxes have a flavour similar to the one we give here, which may be the most commonly encountered during dinnertime conversations.

1.9.2 Consider the classical problem of the Achilles chasing the tortoise. A tortoise starts off a race T seconds before Achilles. Achilles, of course, is faster than the tortoise, but we shall argue that, despite this, Achilles will actually never overtake the tortoise.

> At time T when Achilles starts after the tortoise, the tortoise will be some distance d_1 ahead of Achilles. Achilles will reach this point after some time t_1. But, during the time it took Achilles to travel distance d_1, the tortoise will have moved along to some point d_2 ahead of d_1. Achilles will then take a time t_2 to travel the distance

[19] Zeno of Elea (~490BC–~425BC) was an Italian born philosopher of the Greek school.

d_2. But by then the tortoise will have travelled another distance d_3. This clearly will continue, and when Achilles reaches the point where the tortoise was at some moment before, the tortoise will have moved inexorably ahead. Thus Achilles will never actually catch up to the tortoise.

What is the flaw in the argument?

Chapter 2

Real numbers and their properties

Real numbers and functions of real numbers form an integral part of mathematics. Certainly all students in the sciences receive basic training in these ideas, normally in the form of courses on calculus and differential equations. In this chapter we establish the basic properties of the set of real numbers and of functions defined on this set. In particular, using the construction of the integers in Section 1.4 as a starting point, we *define* the set of real numbers, thus providing a fairly firm basis on which to develop the main ideas in these volumes. We follow this by discussing various structural properties of the set of real numbers. These cover both algebraic properties (Section 2.2.1) and topological properties (Section 2.5). After this, we discuss important ideas like continuity and differentiability of real-valued functions of a real variable.

Do I need to read this chapter? Yes you do, unless you already know its contents. While the construction of the real numbers in Section 2.1 is perhaps a little bit of an extravagance, it does set the stage for the remainder of the material. Moreover, the material in the remainder of the chapter is, in some ways, the backbone of the mathematical presentation. We say this for two reasons.

1. The technical material concerning the structure of the real numbers is, very simply, assumed knowledge for reading everything else in the series.
2. The *ideas* introduced in this chapter will similarly reappear constantly throughout the volumes in the series. But here, many of these ideas are given their most concrete presentation and, as such, afford the inexperienced reader the opportunity to gain familiarity with useful techniques (e.g., the $\epsilon - \delta$ formalism) in a setting where they presumably possess some degree of comfort. This will be crucial when we discuss more abstract ideas in Chapters **??**, **??**, and **??**, to name a few. •

Contents

Section 2.1

Construction of the real numbers

In this section we undertake to define the set of real numbers, using as our starting point the set $\mathbb{Z}$ of integers constructed in Section 1.4. The construction begins by building the rational numbers, which are defined, loosely speaking, as fractions of integers. We know from our school days that every real number can be arbitrarily well approximated by a rational number, e.g., using a decimal expansion. We use this intuitive idea as our basis for defining the set of real numbers from the set of rational numbers.

Do I need to read this section? If you feel comfortable with your understanding of what a real number is, then this section is optional reading. However, it is worth noting that in Section 2.1.2 we first use the $\epsilon - \delta$ formalism that is so important in the analysis featured in this series. Readers unfamiliar/uncomfortable with this idea may find this section a good place to get comfortable with this idea. It is also worth mentioning at this point that the $\epsilon - \delta$ formalism is one with which it is difficult to become fully comfortable. Indeed, PhD theses have been written on the topic of how difficult it is for students to fully assimilate this idea. We shall not adopt any unusual pedagogical strategies to address this matter. However, students are well-advised to spend some time understanding $\epsilon - \delta$ language, and instructors are well-advised to appreciate the difficulty students have in coming to grips with it. •

2.1.1 Construction of the rational numbers

The set of rational numbers is, roughly, the set of fractions of integers. However, we do not know what a fraction is. To define the set of rational numbers, we introduce an equivalence relation $\sim$ in $\mathbb{Z} \times \mathbb{Z}_{>0}$ by

$$(j_1, k_1) \sim (j_2, k_2) \iff j_1 \cdot k_2 = j_2 \cdot k_1.$$

We leave to the reader the straightforward verification that this is an equivalence relation. Using this relation we define the rational numbers as follows.

2.1.1 Definition (Rational numbers) A ***rational number*** is an element of $(\mathbb{Z} \times \mathbb{Z}_{>0})/\sim$. The set of rational numbers is denoted by $\mathbb{Q}$. •

2.1.2 Notation (Notation for rationals) For the rational number $[(j, k)]$ we shall typically write $\frac{j}{k}$, reflecting the usual fraction notation. We shall also often write a typical rational number as "q" when we do not care which equivalence class it comes from. We shall denote by 0 and 1 the rational numbers $[(0, 1)]$ and $[(1, 1)]$, respectively •

The set of rational numbers has many of the properties of integers. For example, one can define addition and multiplication for rational numbers, as well as a total

order in the set of rationals. However, there is an important construction that can be made for rational numbers that cannot generally be made for integers, namely that of division. Let us see how this is done.

2.1.3 Definition (Addition, multiplication, and division in $\mathbb{Q}$) Define the operations of ***addition***, ***multiplication***, and ***division*** in $\mathbb{Q}$ by

(i) $[(j_1,k_1)] + [(j_2,k_2)] = [(j_1 \cdot k_2 + j_2 \cdot k_1, k_1 \cdot k_2)]$,

(ii) $[(j_1,k_1)] \cdot [(j_2,k_2)] = [(j_1 \cdot j_2, k_1 \cdot k_2)]$, and

(iii) $[(j_1,k_1)]/[(j_2,k_2)] = [(j_1 \cdot k_2, k_1 \cdot j_2)]$ (we will also write $\frac{[(j_1,k_1)]}{[(j_2,k_2)]}$ for $[(j_1,k_1)]/[(j_2,k_2)]$),

respectively, where $[(j_1,k_1)], [(j_2,k_2)] \in \mathbb{Q}$ and where, in the definition of division, we require that $j_2 \neq 0$. We will sometimes omit the "$\cdot$" when in multiplication. •

We leave to the reader as Exercise 2.1.1 the straightforward task of showing that these definitions are independent of choice of representatives in $\mathbb{Z} \times \mathbb{Z}_{>0}$. We also leave to the reader the assertion that, with respect to Notation 2.1.2, the operations of addition, multiplication, and division of rational numbers assume the familiar form:

$$\frac{j_1}{k_1} + \frac{j_2}{k_2} = \frac{j_1 \cdot k_2 + j_2 \cdot k_1}{k_1 \cdot k_2}, \quad \frac{j_1}{k_1} \cdot \frac{j_2}{k_2} = \frac{j_1 \cdot j_2}{k_2 \cdot k_2}, \quad \frac{\frac{j_1}{k_1}}{\frac{j_2}{k_2}} = \frac{j_1 \cdot k_2}{k_1 \cdot j_2}.$$

For the operation of division, it is convenient to introduce a new concept. Given $[(j,k)] \in \mathbb{Q}$ with $j \neq 0$, we define $[(j,k)]^{-1} \in \mathbb{Q}$ by $[(k,j)]$. With this notation, division then can be written as $[(j_1,k_1)]/[(j_2,k_2)] = [(j_1,k_1)] \cdot [(j_2,k_2)]^{-1}$. Thus division is really just multiplication, as we already knew. Also, if $q \in \mathbb{Q}$ and if $k \in \mathbb{Z}_{\geq 0}$, then we define $q^k \in \mathbb{Q}$ inductively by $q^0 = 1$ and $q^{k^+} = q^k \cdot q$. The rational number q^k is the kth ***power*** of q.

Let us verify that the operations above satisfy the expected properties. Note that there are now some new properties, since we have the operation of division, or multiplicative inversion, to account for. As we did for integers, we shall write $-q$ for $-1 \cdot q$.

2.1.4 Proposition (Properties of addition and multiplication in $\mathbb{Q}$) *Addition and multiplication in $\mathbb{Q}$ satisfy the following rules:*

(i) $q_1 + q_2 = q_2 + q_1$, $q_1, q_2 \in \mathbb{Q}$ (***commutativity*** *of addition);*

(ii) $(q_1 + q_2) + q_3 = q_1 + (q_2 + q_3)$, $q_1, q_2, q_3 \in \mathbb{Q}$ (***associativity*** *of addition);*

(iii) $q + 0 = q$, $q \in \mathbb{Q}$ (***additive identity***);

(iv) $q + (-q) = 0$, $q \in \mathbb{Q}$ (***additive inverse***);

(v) $q_1 \cdot q_2 = q_2 \cdot q_1$, $q_1, q_2 \in \mathbb{Q}$ (***commutativity*** *of multiplication);*

(vi) $(q_1 \cdot q_2) \cdot q_3 = q_1 \cdot (q_2 \cdot q_3)$, $q_1, q_2, q_3 \in \mathbb{Q}$ (***associativity*** *of multiplication);*

(vii) $q \cdot 1 = q$, $q \in \mathbb{Q}$ (***multiplicative identity***);

(viii) $q \cdot q^{-1} = 1$, $q \in \mathbb{Q} \setminus \{0\}$ (***multiplicative inverse***);

(ix) $r \cdot (q_1 + q_2) = r \cdot q_1 + r \cdot q_2$, $r, q_1, q_2 \in \mathbb{Q}$ (***distributivity***);

(x) $q^{k_1} \cdot q^{k_2} = q^{k_1+k_2}$, $q \in \mathbb{Q}$, $k_1, k_2 \in \mathbb{Z}_{\geq 0}$.

Moreover, if we define $\mathrm{i}_{\mathbb{Z}} \colon \mathbb{Z} \to \mathbb{Q}$ *by* $\mathrm{i}_{\mathbb{Z}}(\mathrm{k}) = [(\mathrm{k}, 1)]$, *then addition and multiplication in* $\mathbb{Q}$ *agrees with that in* $\mathbb{Z}$*:*

$$\mathrm{i}_{\mathbb{Z}}(\mathrm{k}_1) + \mathrm{i}_{\mathbb{Z}}(\mathrm{k}_2) = \mathrm{i}_{\mathbb{Z}}(\mathrm{k}_1 + \mathrm{k}_2), \quad \mathrm{i}_{\mathbb{Z}}(\mathrm{k}_1) \cdot \mathrm{i}_{\mathbb{Z}}(\mathrm{k}_2) = \mathrm{i}_{\mathbb{Z}}(\mathrm{k}_1 \cdot \mathrm{k}_2).$$

Proof All of these properties follow directly from the definitions of addition and multiplication, using Proposition 1.4.19. ■

Just as we can naturally think of $\mathbb{Z}_{\geq 0}$ as being a subset of $\mathbb{Z}$, so too can we think of $\mathbb{Z}$ as a subset of $\mathbb{Q}$. Moreover, we shall very often do so without making explicit reference to the map $i_{\mathbb{Z}}$.

Next we consider on $\mathbb{Q}$ the extension of the partial order $\leq$ and the strict partial order $<$.

2.1.5 Proposition (Order on $\mathbb{Q}$) *On* $\mathbb{Q}$ *define two relations* $<$ *and* $\leq$ *by*

$$\begin{aligned} [(\mathrm{j}_1, \mathrm{k}_1)] < [(\mathrm{j}_2, \mathrm{k}_2)] &\iff \mathrm{j}_1 \cdot \mathrm{k}_2 < \mathrm{k}_1 \cdot \mathrm{j}_2, \\ [(\mathrm{j}_1, \mathrm{k}_1)] \leq [(\mathrm{j}_2, \mathrm{k}_2)] &\iff \mathrm{j}_1 \cdot \mathrm{k}_2 \leq \mathrm{k}_1 \cdot \mathrm{j}_2. \end{aligned}$$

Then $\leq$ *is a total order and* $<$ *is the corresponding strict partial order.*

Proof First let us show that the relations defined make sense, in that they are independent of choice of representative. Thus we suppose that $[(j_1, k_1)] = [(\tilde{j}_1, \tilde{k}_1)]$ and that $[(j_2, k_2)] = [(\tilde{j}_2, \tilde{k}_2)]$. Then

$$\begin{aligned} & [(j_1, k_1)] \leq [(j_2, k_2)] \\ \iff \quad & j_1 \cdot k_2 \leq k_1 \cdot j_2 \\ \iff \quad & j_1 \cdot k_2 \cdot j_2 \cdot \tilde{k}_2 \cdot \tilde{j}_1 \cdot k_1 \leq k_1 \cdot j_2 \cdot \tilde{j}_2 \cdot k_1 \cdot j_1 \cdot \tilde{k}_1 \\ \iff \quad & (\tilde{j}_1 \cdot \tilde{k}_2) \cdot (j_1 \cdot j_2 \cdot k_1 \cdot k_2) \leq (\tilde{j}_2 \cdot \tilde{k}_1) \cdot (j_1 \cdot j_2 \cdot k_1 \cdot k_2) \\ \iff \quad & \tilde{j}_1 \cdot \tilde{k}_2 \leq \tilde{j}_2 \cdot \tilde{k}_1. \end{aligned}$$

This shows that the definition of $\leq$ is independent of representative. Of course, a similar argument holds for $<$.

That $\leq$ is a partial order, and that $<$ is its corresponding strict partial order, follow from a straightforward checking of the definitions, so we leave this to the reader.

Thus we only need to check that $\leq$ is a total order. Let $[(j_1, k_1)], [(j_2, k_2)] \in \mathbb{Q}$. Then, by the Trichotomy Law for $\mathbb{Z}$, either $j_1 \cdot k_2 < k_1 \cdot j_2$, $k_1 \cdot j_2 < j_1 \cdot k_2$, or $j_1 \cdot k_2 = k_1 \cdot j_2$. But this directly implies that either $[(j_1, k_1)] < [(j_2, k_2)]$, $[(j_2, k_2)] < [(j_1, k_1)]$, or $[(j_1, k_1)] = [(j_2, k_2)]$, respectively. ■

The total order on $\mathbb{Q}$ allows a classification of rational numbers as follows.

2.1.6 Definition (Positive and negative rational numbers) A rational number $q \in \mathbb{Q}$ is:

(i) ***positive*** if $0 < q$;
(ii) ***negative*** if $q < 0$;
(iii) ***nonnegative*** if $0 \leq q$;
(iv) ***nonpositive*** if $q \leq 0$.

The set of positive rational numbers is denoted by $\mathbb{Q}_{>0}$ and the set of nonnegative rational numbers is denoted by $\mathbb{Q}_{\geq 0}$. •

As we did with natural numbers and integers, we isolate the Trichotomy Law.

2.1.7 Corollary (Trichotomy Law for $\mathbb{Q}$) *For $\mathrm{q}, \mathrm{r} \in \mathbb{Q}$, exactly one of the following possibilities holds:*

(i) $\mathrm{q} < \mathrm{r}$;

(ii) $\mathrm{r} < \mathrm{q}$;

(iii) $\mathrm{q} = \mathrm{r}$.

The following result records the relationship between the order on $\mathbb{Q}$ and the arithmetic operations.

2.1.8 Proposition (Relation between addition and multiplication and <) *For $\mathrm{q}, \mathrm{r}, \mathrm{s} \in \mathbb{Q}$, the following statements hold:*

(i) if $\mathrm{q} < \mathrm{r}$ *then* $\mathrm{q} + \mathrm{s} < \mathrm{r} + \mathrm{s}$;

(ii) if $\mathrm{q} < \mathrm{r}$ *and if* $\mathrm{s} > 0$ *then* $\mathrm{s} \cdot \mathrm{q} < \mathrm{s} \cdot \mathrm{r}$;

(iii) if $\mathrm{q} < \mathrm{r}$ *and if* $\mathrm{s} < 0$ *then* $\mathrm{s} \cdot \mathrm{r} < \mathrm{s} \cdot \mathrm{q}$;

(iv) if $0 < \mathrm{q}, \mathrm{r}$ *then* $0 < \mathrm{q} \cdot \mathrm{r}$;

(v) if $\mathrm{q} < \mathrm{r}$ *and if either*

(a) $0 < \mathrm{q}, \mathrm{r}$ *or*

(b) $\mathrm{q}, \mathrm{r} < 0$,

then $\mathrm{r}^{-1} < \mathrm{q}^{-1}$.

Proof (i) Write $q = [(j_q, k_q)]$, $r = [(j_r, k_r)]$, and $s = [(j_s, k_s)]$. Since $q < r$, $j_q \cdot k_r \leq j_r \cdot k_q$. Therefore,

$$\begin{aligned} & j_q \cdot k_r \cdot k_s^2 < j_r \cdot k_q \cdot k_s^2 \\ \implies \quad & j_q \cdot k_r \cdot k_s^2 + j_s \cdot k_q \cdot k_r \cdot k_s < j_r \cdot k_q \cdot k_s^2 + j_2 \cdot k_q \cdot k_r \cdot k_s, \end{aligned}$$

using Proposition 1.4.22. This last inequality is easily seen to be equivalent to $q + s < r + s$.

(ii) Write $q = [(j_q, k_q)]$, $r = [(j_r, k_r)]$, and $s = [(j_s, k_s)]$. Since $s > 0$ it follows that $j_s > 0$. Since $q \leq r$ it follows that $j_q \cdot k_r \leq j_r \cdot k_q$. From Proposition 1.4.22 we then have

$$j_q \cdot j_s \cdot j_s \cdot k_s \leq j_r \cdot k_q \cdot j_s \cdot k_s,$$

which is equivalent to $s \cdot q \leq s \cdot r$ by definition of multiplication.

(iii) The result here follows, as does (ii), from Proposition 1.4.22, but now using the fact that $j_s < 0$.

(iv) This is a straightforward application of the definition of multiplication and <.

(v) This follows directly from the definition of <. ■

The final piece of structure we discuss for rational numbers is the extension of the absolute value function defined for integers.

2.1.9 Definition (Rational absolute value function) The ***absolute value function*** on $\mathbb{Q}$ is the map from $\mathbb{Q}$ to $\mathbb{Q}_{\geq 0}$, denoted by $q \mapsto |q|$, defined by

$$|q| = \begin{cases} q, & 0 < q, \\ 0, & q = 0, \\ -q, & q < 0. \end{cases}$$

•

The absolute value function on $\mathbb{Q}$ has properties like that on $\mathbb{Z}$.

2.1.10 Proposition (Properties of absolute value on $\mathbb{Q}$) *The following statements hold:*

(i) $|q| \geq 0$ *for all* $q \in \mathbb{Q}$;

(ii) $|q| = 0$ *if and only if* $q = 0$;

(iii) $|r \cdot q| = |r| \cdot |q|$ *for all* $r, q \in \mathbb{Q}$;

(iv) $|r + q| \leq |r| + |q|$ *for all* $r, q \in \mathbb{Q}$ (***triangle inequality***);

(v) $|q^{-1}| = |q|^{-1}$ *for all* $q \in \mathbb{Q} \setminus \{0\}$.

Proof Parts (i), (ii), and (v), follow directly from the definition, and part (iii) follows in the same manner as the analogous statement in Proposition 1.4.24. Thus we have only to prove part (iv). We consider various cases.

1. $|r| \leq |q|$:
   (a) $0 \geq r, q$: Since $|r + q| = r + q$, and $|r| = r$ and $|q| = q$, this follows directly.
   (b) $r < 0$, $0 \leq q$: Let $r = [(j_r, k_r)]$ and $q = [(j_q, k_q)]$. Then $r < 0$ gives $j_r < 0$ and $0 \leq q$ gives $j_q \geq 0$. We now have

$$|r + q| = \left|\frac{j_r \cdot k_q + j_q \cdot k_r}{k_r \cdot k_q}\right| = \frac{|j_r \cdot k_q + j_q \cdot k_r|}{k_r \cdot k_q}$$

   and

$$|r| + |q| = \frac{|j_r| \cdot k_q + |j_q| \cdot k_r}{k_r \cdot k_q}.$$

   Therefore,

$$\begin{aligned} |r + q| &= \frac{|j_r \cdot k_q + j_q \cdot k_r|}{k_r \cdot k_q} \\ &\leq \frac{|j_r| \cdot k_q + |j_q| \cdot k_r}{k_r \cdot k_q} \\ &= |r| + |q|, \end{aligned}$$

   where we have used Proposition 2.1.8.
   (c) $r, q < 0$: Here $|r + q| = |-r + (-q)| = |-(r + q)| = -(r + q)$, and $|r| = -r$ and $|q| = -q$, so the result follows immediately.
2. $|q| \leq |r|$: This argument is the same as above, swapping r and q. ■

2.1.11 Remark Having been quite fussy about how we arrived at the set of integers and the set of rational numbers, and about characterising their important properties, we shall now use standard facts about these, some of which we may not have proved, but which can easily be proved using the definitions of $\mathbb{Z}$ and $\mathbb{Q}$. Some of the arithmetic properties of $\mathbb{Z}$ and $\mathbb{Q}$ that we use without comment are in fact proved in Section **??** in the more general setting of rings. However, we anticipate that most readers will not balk at the instances where we use unproved properties of integers and rational numbers. •

2.1.2 Construction of the real numbers from the rational numbers

Now we use the rational numbers as the building block for the real numbers. The idea of this construction, which was originally due to Cauchy[1], is the intuitive idea that the rational numbers may be used to approximate well a real number. For example, we learn in school that any real number is expressible as a decimal expansion (see Exercise 2.4.8 for the precise construction of a decimal expansion). However, any finite length decimal expansion (and even some infinite length decimal expansions) is a rational number. So one could *define* real numbers as a limit of decimal expansions in some way. The problem is that there may be multiple decimal expansions giving rise to the same real number. For example, the decimal expansions 1.0000 and 0.9999... represent the same real number. The way one gets around this potential problem is to use equivalence classes, of course. But equivalence classes of what? This is where we begin the presentation, proper.

2.1.12 Definition (Cauchy sequence, convergent sequence) Let $(q_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{Q}$. The sequence:

(i) is a ***Cauchy sequence*** if, for each $\epsilon \in \mathbb{Q}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < \epsilon$ for $j, k \geq N$;

(ii) ***converges to*** $\mathbf{q_0}$ if, for each $\epsilon \in \mathbb{Q}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_0| < \epsilon$ for $j \geq N$.

(iii) is ***bounded*** if there exists $M \in \mathbb{Q}_{>0}$ such that $|q_j| < M$ for each $j \in \mathbb{Z}_{>0}$. •

The set of Cauchy sequences in $\mathbb{Q}$ is denoted by CS($\mathbb{Q}$). A sequence converging to q_0 has q_0 as its ***limit***. •

The idea of a Cauchy sequence is that the terms in the sequence can be made arbitrarily close as we get to the tail of the sequence. A convergent sequence, however, gets closer and closer to its limit as we get to the tail of the sequence. Our instinct is probably that there is a relationship between these two ideas. One thing that is true is the following.

2.1.13 Proposition (Convergent sequences are Cauchy) *If a sequence* $(q_j)_{j\in\mathbb{Z}_{>0}}$ *converges to* q_0*, then it is a Cauchy sequence.*

Proof Let $\epsilon \in \mathbb{Q}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_0| < \frac{\epsilon}{2}$ for $j \geq N$. Then, for $j, k \geq N$ we have

$$|q_j - q_k| = |q_j - q_0 - q_k + q_0| = |q_j - q_0| + |q_k - q_0| < \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon,$$

using the triangle inequality of Proposition 2.1.10. ■

Cauchy sequences have the property of being bounded.

[1]The French mathematician Augustin Louis Cauchy (1789–1857) worked in the areas of complex function theory, partial differential equations, and analysis. His collected works span twenty-seven volumes.

2.1.14 Proposition (Cauchy sequences are bounded) *If* $(q_j)_{j\in\mathbb{Z}_{>0}}$ *is a Cauchy sequence, then it is bounded.*

Proof Choose $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < 1$ for $j, k \in \mathbb{Z}_{>0}$. Then take M_N to be the largest of the nonnegative rational numbers $|q_1|, \dots, |q_N|$. Then, for $j \geq N$ we have, using the triangle inequality,

$$|q_j| = |q_j - q_N + q_N| \leq |q_j - q_N| + |q_N| < 1 + M_N,$$

giving the result by taking $M = M_N + 1$. ■

The question as to whether there are nonconvergent Cauchy sequences is now the obvious one.

2.1.15 Example (Nonconvergent Cauchy sequences in $\mathbb{Q}$ exist) If one already knows the real numbers exist, it is somewhat easy to come up with Cauchy sequences in $\mathbb{Q}$. However, to fabricate one "out of thin air" is not so easy.

For $k \in \mathbb{Z}_{>0}$, since $2k + 5 > k + 4$, it follows that $2^{2k+5} - 2^{k+4} > 0$. Let m_k be the smallest nonnegative integer for which

$$m_k^2 \geq 2^{2k+5} - 2^{k+4}. \tag{2.1}$$

The following contains a useful property of m_k.

1 Lemma $m_k^2 \leq 2^{2k+5}$.

Proof First we show that $m_k \leq 2^{k+3}$. Suppose that $m_k > 2^{k+3}$. Then

$$(m_k - 1)^2 > (2^{k+3} - 1)^2 = 2^{2k+6} - 2^{k+4} + 1 = 2(2^{2k+5} - 2^{k+4}) + 1) > 2^{2k+5} - 2^{k+4},$$

which contradicts the definition of m_k.

Now suppose that $m_k^2 > 2^{2k+5}$. Then

$$(m_k - 1)^2 = m_k^2 - 2m_k + 1 > 2^{2k+5} - 2^{k+4} + 1 > 2^{2k+5} - 2^{k+4},$$

again contradicting the definition of m_k. ▼

Now define $q_k = \frac{m_k}{2^{k+2}}$.

2 Lemma $(q_k)_{k\in\mathbb{Z}_{>0}}$ *is a Cauchy sequence.*

Proof By Lemma 1 we have

$$q_k^2 = \frac{m_k^2}{2^{2k+4}} \leq \frac{2^{2k+5}}{2^{2k+4}} = 2, \qquad k \in \mathbb{Z}_{>0},$$

and by (2.1) we have

$$q_k^2 = \frac{m_k^2}{2^{2k+4}} \geq \frac{2^{2k+5}}{2^{2k+4}} - \frac{2^{k+4}}{2^{2k+4}} = 2 - \frac{1}{2^k}, \qquad k \in \mathbb{Z}_{>0}.$$

Summarising, we have

$$2 - \frac{1}{2^k} \leq q_k^2 \leq 2, \qquad k \in \mathbb{Z}_{>0}. \tag{2.2}$$

Then, for $j, k \in \mathbb{Z}_{>0}$ we have

$$2 - \frac{1}{2^k} \leq q_k^2 \leq 2,\ 2 - \frac{1}{2^j} \leq q_k^2 \leq 2 \quad \Longrightarrow \quad -\frac{1}{2^j} \leq q_j^2 - q_k^2 \leq \frac{1}{2^k}.$$

Next we have, from (2.1),

$$q_k^2 = \frac{m_k^2}{2^{2k+4}} \geq \frac{2^{2k+5}}{2^{2k+4}} - \frac{2^{k+4}}{2^{2k+4}} = 2 - \frac{1}{2^k}, \qquad k \in \mathbb{Z}_{>0},$$

from which we deduce that $q_k^2 \geq 1$, which itself implies that $q_k \geq 1$. Next, using this fact and $(q_j - q_k)^2 = (q_j + q_k)(q_j - q_k)$ we have

$$-\frac{1}{2^j}\frac{1}{q_j + q_k} \leq q_j - q_k \leq \frac{1}{2^j}\frac{1}{q_j + q_k} \quad \Longrightarrow \quad -\frac{1}{2^{j+1}} \leq q_j - q_k \leq \frac{1}{2^{k+1}}, \qquad j, k \in \mathbb{Z}_{>0}. \tag{2.3}$$

Now let $\epsilon \in \mathbb{Q}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $\frac{1}{2^{N+1}} < \epsilon$. Then we immediately have $|q_j - q_k| < \epsilon$, $j, k \geq N$, using (2.3). ▼

The following result gives the character of the limit of the sequence $(q_k)_{k \in \mathbb{Z}_{>0}}$, were it to be convergent.

3 Lemma *If* q_0 *is the limit for the sequence* $(q_k)_{k \in \mathbb{Z}_{>0}}$, *then* $q_0^2 = 2$.

Proof We claim that if $(q_k)_{k \in \mathbb{Z}_{>0}}$ converges to q_0, then $(q_k^2)_{k \in \mathbb{Z}_{>0}}$ converges to q_0^2. Let $M \in \mathbb{Q}_{>0}$ satisfy $|q_k| < M$ for all $k \in \mathbb{Z}_{>0}$, this being possible by Proposition 2.1.14. Now let $\epsilon \in \mathbb{Q}_{>0}$ and take $N \in \mathbb{Z}_{>0}$ such that

$$|q_k - q_0| < \frac{\epsilon}{M + |q_0|}.$$

Then

$$|q_k^2 - q_0^2| = |q_k - q_0||q_k + q_0| < \epsilon,$$

giving our claim.

Finally, we prove the lemma by proving that $(q_k^2)_{k \in \mathbb{Z}_{>0}}$ converges to 2. Indeed, let $\epsilon \in \mathbb{Q}_{>0}$ and note that, if $N \in \mathbb{Z}_{>0}$ is chosen to satisfy $\frac{1}{2^N} < \epsilon$. Then, using (2.2), we have

$$|q_k^2 - 2| \leq \frac{1}{2^k} < \epsilon, \qquad k \geq N,$$

as desired. ▼

Finally, we have the following result, which is contained in the mathematical works of Euclid.

4 Lemma *There exists no* $\mathrm{q}_0 \in \mathbb{Q}$ *such that* $\mathrm{q}_0^2 = 2$.

Proof Suppose that $q_0^2 = [(j_0, k_0)]$ and further suppose that there is no integer m such that $q_0 = [(mj_0, mk_0)]$. We then have

$$q_0^2 = \frac{j_0^2}{k_0^2} = 2 \qquad \Longrightarrow \qquad j_0^2 = 2k_0^2.$$

Thus j_0^2 is even, and then so too is j_0 (why?). Therefore, $j_0 = 2\tilde{j}_0$ and so

$$q_0^2 = \frac{4\tilde{j}_0^2}{k_0^2} = 2 \qquad \Longrightarrow \qquad k_0^2 = 2\tilde{j}_0^2$$

which implies that k_0^2, and hence k_0 is also even. This contradicts our assumption that there is no integer m such that $q_0 = [(mj_0, mk_0)]$. ▼

With these steps, we have constructed a Cauchy sequence that does not converge. •

Having shown that there are Cauchy sequences that do not converge, the idea is now to define a real number to be, essentially, that to which a nonconvergent Cauchy sequence would converge if only it could. First we need to allow for the possibility, realised in practice, that different Cauchy sequences may converge to the same limit.

2.1.16 Definition (Equivalent Cauchy sequences) Two sequences $(q_j)_{j\in\mathbb{Z}_{>0}}, (r_j)_{j\in\mathbb{Q}} \in \mathrm{CS}(\mathbb{Q})$ are ***equivalent*** if the sequence $(q_j - r_j)_{j\in\mathbb{Z}_{>0}}$ converges to zero. We write $(q_j)_{j\in\mathbb{Z}_{>0}} \sim (r_j)_{j\in\mathbb{Z}_{>0}}$ if the two sequences are equivalent. •

We should verify that this notion of equivalence of Cauchy sequences is indeed an equivalence relation.

2.1.17 Lemma *The relation* $\sim$ *defined in* $\mathrm{CS}(\mathbb{Q})$ *is an equivalence relation.*

Proof It is clear that the relation $\sim$ is reflexive and symmetric. To prove transitivity, suppose that $(q_j)_{j\in\mathbb{Z}_{>0}} \sim (r_j)_{j\in\mathbb{Z}_{>0}}$ and that $(r_j)_{j\in\mathbb{Z}_{>0}} \sim (s_j)_{j\in\mathbb{Z}_{>0}}$. For $\epsilon \in \mathbb{Q}_{>0}$ let $N \in \mathbb{Z}_{>0}$ satisfy

$$|q_j - r_j| < \tfrac{\epsilon}{2},\ |r_j - s_j| < \tfrac{\epsilon}{2}, \qquad j \geq N.$$

Then, using the triangle inequality,

$$|q_j - s_j| = |q_j - r_j + r_j - s_j| \leq |q_j - r_j| + |r_j - s_j| < \epsilon, \qquad j \geq \mathbb{Z}_{>0},$$

showing that $(q_j)_{j\in\mathbb{Z}_{>0}} \sim (s_j)_{j\in\mathbb{Z}_{>0}}$. ■

We are now prepared to define the set of real numbers.

2.1.18 Definition (Real numbers) A ***real number*** is an element of $\mathrm{CS}(\mathbb{Q})/\sim$. The set of real numbers is denoted by $\mathbb{R}$. •

The definition encodes, in a precise way, our intuition about what a real number is. In the next section we shall examine some of the properties of the set $\mathbb{R}$.

Let us give the notation we will use for real numbers, since clearly we do not wish to write these explicitly as equivalence classes of Cauchy sequences.

2.1.19 Notation (Notation for reals) We shall frequently write a typical element in $\mathbb{R}$ as "x". We shall denote by 0 and 1 the real numbers associated with the Cauchy sequences $(0)_{j\in\mathbb{Z}_{>0}}$ and $(1)_{j\in\mathbb{Z}_{>0}}$. •

Exercises

2.1.1 Show that the definitions of addition, multiplication, and division of rational numbers in Definition 2.1.3 are independent of representative.

2.1.2 Show that the order and absolute value on $\mathbb{Q}$ agree with those on $\mathbb{Z}$. That is to say, show the following:

(a) for $j, k \in \mathbb{Z}$, $j < k$ if and only if $i_{\mathbb{Z}}(j) < i_{\mathbb{Z}}(k)$;

(b) for $k \in \mathbb{Z}$, $|k| = |i_{\mathbb{Z}}(k)|$.

(Note that we see clearly here the abuse of notation that follows from using $<$ for both the order on $\mathbb{Z}$ and $\mathbb{Q}$ and from using $|\cdot|$ as the absolute value both on $\mathbb{Z}$ and $\mathbb{Q}$. It is expected that the reader can understand where the notational abuse occurs.)

2.1.3 Show that the set of rational numbers is countable using an argument along the following lines.

1. Construct a doubly infinite grid in the plane with a point at each integer coordinate. Note that every rational number $q = \frac{n}{m}$ is represented by the grid point (n, m).
2. Start at the "centre" of the grid with the rational number 0 being assigned to the grid point $(0, 0)$, and construct a spiral which passes through each grid point. Note that this spiral should hit every grid point exactly once.
3. Use this spiral to infer the existence of a bijection from $\mathbb{Q}$ to $\mathbb{Z}_{>0}$.

The following exercise leads you through Cantor's famous "diagonal argument" for showing that the set of real numbers is uncountable.

2.1.4 Fill in the gaps in the following construction, justifying all steps.

1. Let $\{x_j \mid j \in \mathbb{Z}_{>0}\}$ be a countable subset of $(0, 1)$.
2. Construct a doubly infinite table for which the kth column of the jth row contains the kth term in the decimal expansion for x_j.
3. Construct $\bar{x} \in (0, 1)$ by declaring the kth term in the decimal expansion for $\bar{x}$ to be different from the kth term in the decimal expansion for x_k.
4. Show that $\bar{x}$ is not an element of the set $\{x_j \mid j \in \mathbb{Z}_{>0}\}$.
   Hint: *Be careful to understand that a real number might have different decimal expansions.*

2.1.5 Show that for any $x \in \mathbb{R}$ and $\epsilon \in \mathbb{R}_{>0}$ there exists $k \in \mathbb{Z}_{>0}$ and an odd integer j such that $|x - \frac{j}{2^k}| < \epsilon$.

Section 2.2

Properties of the set of real numbers

In this section we present some of the well known properties as the real numbers, both algebraic and (referring ahead to the language of Chapter **??**) topological.

Do I need to read this section? Many of the properties given in Sections 2.2.1, 2.2.2 and 2.2.3 will be well known to any student with a high school education. However, these may be of value as a starting point in understanding some of the abstract material in Chapters **??** and **??**. Similarly, the material in Section 2.2.4 is "obvious." However, since this material will be assumed knowledge, it might be best for the reader to at least skim the section, to make sure there is nothing new in it for them. •

2.2.1 Algebraic properties of $\mathbb{R}$

In this section we define addition, multiplication, order, and absolute value for $\mathbb{R}$, mirroring the presentation for $\mathbb{Q}$ in Section 2.1.1. Here, however, the definitions and verifications are not just trivialities, as they are for $\mathbb{Q}$.

First we define addition and multiplication. We do this by defining these operations first on elements of $\mathrm{CS}(\mathbb{Q})$, and then showing that the operations depend only on equivalence class. The following is the key step in doing this.

2.2.1 Proposition (Addition, multiplication, and division of Cauchy sequences) *Let* $(q_j)_{j\in\mathbb{Z}_{>0}}, (r_j)_{j\in\mathbb{Z}_{>0}} \in \mathrm{CS}(\mathbb{Q})$. *Then the following statements hold.*

(i) *The sequence* $(q_j + r_j)_{j\in\mathbb{Z}_{>0}}$ *is a Cauchy sequence which we denote by* $(q_j)_{j\in\mathbb{Z}_{>0}} + (r_j)_{j\in\mathbb{Z}_{>0}}$.

(ii) *The sequence* $(q_j \cdot r_j)_{j\in\mathbb{Z}_{>0}}$ *is a Cauchy sequence which we denote by* $(q_j)_{j\in\mathbb{Z}_{>0}} \cdot (r_j)_{j\in\mathbb{Z}_{>0}}$.

(iii) *If, for all* $j \in \mathbb{Z}_{>0}$, $q_j \neq 0$ *and if the sequence* $(q_j)_{j\in\mathbb{Z}_{>0}}$ *does not converge to* 0, *then* $(q_j^{-1})_{j\in\mathbb{Z}_{>0}}$ *is a Cauchy sequence.*

Furthermore, if $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}, (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \in \mathrm{CS}(\mathbb{Q})$ *satisfy*

$$(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \sim (q_j)_{j\in\mathbb{Z}_{>0}}, \quad (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \sim (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}},$$

then

(iv) $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} + (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} = (q_j)_{j\in\mathbb{Z}_{>0}} + (r_j)_{j\in\mathbb{Z}_{>0}}$,

(v) $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \cdot (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} = (q_j)_{j\in\mathbb{Z}_{>0}} \cdot (r_j)_{j\in\mathbb{Z}_{>0}}$, *and*

(vi) *if, for all* $j \in \mathbb{Z}_{>0}$, $q_j, \tilde{q}_j \neq 0$ *and if the sequences* $(q_j)_{j\in\mathbb{Z}_{>0}}, (\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}$ *do not converge to* 0, *then* $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \sim (q_j)_{j\in\mathbb{Z}_{>0}}$.

Proof (i) Let $\epsilon \in \mathbb{Q}_{>0}$ and let $N \in \mathbb{Z}_{>0}$ have the property that $|q_j - q_k|, |r_j - r_k| < \frac{\epsilon}{2}$ for all $j, k \geq N$. Then, using the triangle inequality,

$$|(q_j + r_j) - (q_k + r_k)| \leq |q_j - q_k| + |r_j - r_k| = \epsilon, \qquad j, k \geq N.$$

(ii) Let $M \in \mathbb{Q}_{>0}$ have the property that $|q_j|, |r_j| < M$ for all $j \in \mathbb{Z}_{>0}$. For $\epsilon \in \mathbb{Q}_{>0}$ let $N \in \mathbb{Z}_{>0}$ have the property that $|q_j - q_k|, |r_j - r_k| < \frac{\epsilon}{2M}$ for all $j, k \geq N$. Then, using the triangle inequality,

$$|(q_j \cdot r_j) - (q_k \cdot r_k)| = |q_j(r_j - r_k) - r_k(q_k - q_j)| \leq |q_j||r_j - r_k| + |r_k||q_k - q_j| < \epsilon, \qquad j, k \geq N.$$

(iii) We claim that if $(q_j)_{j \in \mathbb{Z}_{>0}}$ satisfies the conditions stated, then there exists $\delta \in \mathbb{Q}_{>0}$ such that $|q_k| \geq \delta$ for all $k \in \mathbb{Z}_{>0}$. Indeed, since $(q_j)_{j \in \mathbb{Z}_{>0}}$ does not converge to zero, choose $\epsilon \in \mathbb{Q}_{>0}$ such that, for all $N \in \mathbb{Z}_{>0}$, there exists $j \geq N$ for which $|q_j| \geq \epsilon$. Next take $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < \frac{\epsilon}{2}$ for $j, k \geq N$. Then there exists $\tilde{N} \geq N$ such that $|q_{\tilde{N}}| \geq \epsilon$. For any $j \geq N$ we then have

$$|q_j| = |q_{\tilde{N}} - (q_{\tilde{N}} - q_j)| \geq ||q_{\tilde{N}}| - |q_{\tilde{N}} - q_j|| \geq \epsilon - \tfrac{\epsilon}{2} = \tfrac{\epsilon}{2},$$

where we have used Exercise 2.2.7. The claim follows by taking δ to be the smallest of the numbers $\frac{\epsilon}{2}, |q_1|, \ldots, |q_N|$.

Now let $\epsilon \in \mathbb{Q}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < \delta^2\epsilon$ for $j, k \geq N$. Then

$$|q_j^{-1} - q_k^{-1}| = \left|\frac{q_k - q_j}{q_j q_k}\right| < \frac{\delta^2\epsilon}{\delta^2} = \epsilon, \qquad j, k \geq N.$$

(iv) For $\epsilon \in \mathbb{Q}_{>0}$ let $N \in \mathbb{Z}_{>0}$ have the property that $|\tilde{q}_j - q_j|, |\tilde{r}_j - r_j| < \frac{\epsilon}{2}$. Then, using the triangle inequality,

$$|(\tilde{q}_j + \tilde{r}_j) - (q_k + r_k)| \leq |\tilde{q}_j - q_k| + |\tilde{r}_k - r_k| < \epsilon, \qquad j, k \geq N.$$

(v) Let $M \in \mathbb{Q}_{>0}$ have the property that $|\tilde{q}_j|, |r_j| < M$ for all $j \in \mathbb{Z}_{>0}$. Then, for $\epsilon \in \mathbb{Q}_{>0}$, take $N \in \mathbb{Z}_{>0}$ such that $|\tilde{r}_j - r_k|, |\tilde{q}_j - q_k| < \frac{\epsilon}{2M}$ for $j, k \geq N$. We then use the triangle inequality to give

$$|(\tilde{q}_j \cdot \tilde{r}_j) - (q_k \cdot r_k)| = |\tilde{q}_j(\tilde{r}_j - r_k) - r_k(q_k - \tilde{q}_j)| < \epsilon, \qquad j, k \geq N.$$

(vi) Let $\delta \in \mathbb{Q}_{>0}$ satisfy $|q_j|, |\tilde{q}_j| \geq \delta$ for all $j \in \mathbb{Z}_{>0}$. Then, for $\epsilon \in \mathbb{Q}_{>0}$, choose $N \in \mathbb{Z}_{>0}$ such that $|\tilde{q}_j - q_j| < \delta^2\epsilon$ for $j \geq N$. Then we have

$$|\tilde{q}_j^{-1} - q_j^{-1}| = \left|\frac{q_j - \tilde{q}_j}{q_j \tilde{q}_j}\right| < \frac{\delta^2\epsilon}{\delta^2}, \qquad j \geq N,$$

so completing the proof. ∎

The requirement, in parts (iii) and (vi), that the sequence $(q_j)_{j \in \mathbb{Z}_{>0}}$ have no zero elements is not really a restriction in the same way as is the requirement that the sequence not converge to zero. The reason for this is that, as we showed in the proof, if the sequence does not converge to zero, then there exists $\epsilon \in \mathbb{Q}_{>0}$ and $N \in \mathbb{Z}_{>0}$ such that $|q_j| > \epsilon$ for $j \geq N$. Thus the tail of the sequence is guaranteed to have no zero elements, and the tail of the sequence is all that matters for the equivalence class.

Now that we have shown how to add and multiply Cauchy sequences in $\mathbb{Q}$, and that this addition and multiplication depends only on equivalence classes under the notion of equivalence given in Definition 2.1.16, we can easily define addition and multiplication in $\mathbb{R}$.

2.2.2 Definition (Addition, multiplication, and division in $\mathbb{R}$) Define the operations of ***addition***, ***multiplication***, and ***division*** in $\mathbb{R}$ by

(i) $[(q_j)_{j\in\mathbb{Z}_{>0}}] + [(r_j)_{j\in\mathbb{Z}_{>0}}] = [(q_j)_{j\in\mathbb{Z}_{>0}} + (r_j)_{j\in\mathbb{Z}_{>0}}]$,

(ii) $[(q_j)_{j\in\mathbb{Z}_{>0}}] \cdot [(r_j)_{j\in\mathbb{Z}_{>0}}] = [(q_j)_{j\in\mathbb{Z}_{>0}} \cdot (r_j)_{j\in\mathbb{Z}_{>0}}]$,

(iii) $[(q_j)_{j\in\mathbb{Z}_{>0}}]/[(r_j)_{j\in\mathbb{Z}_{>0}}] = [(q_j/r_j)_{j\in\mathbb{Z}_{>0}} + (r_j)_{j\in\mathbb{Z}_{>0}}]$,

respectively, where, in the definition of division, we require that the sequence $(r_j)_{j\in\mathbb{Z}_{>0}}$ have no zero elements, and that it not converge to 0. We will sometimes omit the "$\cdot$" when writing multiplication. •

Similarly to what we have done previously with $\mathbb{Z}$ and $\mathbb{Q}$, we let $-x = [(-1)_{j\in\mathbb{Z}_{>0}}] \cdot x$. For $x \in \mathbb{R}\setminus\{0\}$, we also denote by x^{-1} the real number corresponding to a Cauchy sequence $(\frac{1}{q_j})_{j\in\mathbb{Z}_{>0}}$, where $x = [(q_j)_{j\in\mathbb{Z}_{>0}}]$.

As with integers and rational numbers, we can define powers of real numbers. For $x \in \mathbb{R}\setminus\{0\}$ and $k \in \mathbb{Z}_{\geq 0}$ we define $x^k \in \mathbb{R}$ inductively by $x^0 = 1$ and $x^{k^+} = x^k \cdot x$. As usual, we call x^k the kth ***power*** of x. For $k \in \mathbb{Z}\setminus\mathbb{Z}_{\geq 0}$, we take $x^k = (x^{-k})^{-1}$. For real numbers, the notion of the power of a number can be extended. Let us show how this is done. In the statement of the result, we use the notion of positive real numbers which are not defined until Definition 2.2.8. Also, in our proof, we refer ahead to properties of $\mathbb{R}$ that are not considered until Section 2.3. However, it is convenient to state the construction here.

2.2.3 Proposition ($x^{1/k}$) *For* $x \in \mathbb{R}_{>0}$ *and* $k \in \mathbb{Z}_{>0}$, *there exists a unique* $y \in \mathbb{R}_{>0}$ *such that* $y^k = x$. *We denote the number* y *by* $x^{1/k}$.

Proof Let $S_x = \{y \in \mathbb{R} \mid y^k < x\}$. Since $x \geq 0$, $0 \in S$ so $S \neq \emptyset$. We next claim that $\max\{1, x\}$ is an upper bound for S_x. First suppose that $x < 1$. Then, for $y \in S_x$, $y^k < x < 1$, and so 1 is an upper bound for S_x. If $x \geq 1$ and $y \in S_x$, then we claim that $y \leq x$. Indeed, if $y > x$ then $y^k > x^k > x$, and so $y \notin S_x$. This shows that S_x is upper bounded by x in this case. Now we know that S_x has a least upper bound by Theorem 2.3.7. Let y denote this least upper bound.

We shall now show that $y^k = x$. Suppose that $y^k \neq x$. From Corollary 2.2.9 we have $y^k < x$ or $y^k > x$.

Suppose first that $y^k < x$. Then, for $\epsilon \in \mathbb{R}_{>0}$ we have

$$(y+\epsilon)^k = \epsilon^k + a_{k-1}y\epsilon^{k-1} + \cdots + a_1 y^{k-1}\epsilon + y^k$$

for some numbers $a_1, \ldots, a_{k-1}$ (these are the binomial coefficients of Exercise 2.2.1). If $\epsilon \leq 1$ then $\epsilon^k \leq \epsilon$ for $k \in \mathbb{Z}_{>0}$. Therefore, if $\epsilon \leq 1$ we have

$$(y+\epsilon)^k \leq \epsilon(1 + a_{k-1}y + \cdots + a_1 y^{k-1}) + y^k.$$

Now, if $\epsilon < \min\{1, \frac{x-y^k}{1+a_{k-1}y+\cdots+a_a y^{k-1}}\}$, then $(y+\epsilon)^k < x$, contradicting the fact that y is an upper bound for S_x.

Now suppose that $y^k > x$. Then, for $\epsilon \in \mathbb{R}_{>0}$, we have

$$(y-\epsilon)^k = (-1)^k\epsilon^k + (-1)^{k-1}a_{k-1}y\epsilon^{k-1} + \cdots - a_1 y^{k-1}\epsilon + y^k.$$

The sum on the right involves terms that are positive and negative. This sum will be greater than the corresponding sum with the positive terms involving powers of ϵ removed. That is to say,

$$(y-\epsilon)^k > y^k - a_1 y^{k-1}\epsilon - a_3 y^{k-3}\epsilon^3 + \cdots .$$

For $\epsilon \leq 1$ we again gave $\epsilon^k \leq \epsilon$ for $k \in \mathbb{Z}_{>0}$. Therefore

$$(y-\epsilon)^k > y^k - (a_1 y^{k-1} + a_3 y^{k-3} + \cdots)\epsilon.$$

Thus, if $\epsilon < \min\{1, \frac{y^k - x}{a_1 y^{k-1} + a_3 y^{k-3} + \dots}\}$ we have $(y-\epsilon)^k > x$, contradicting the fact that y is the least upper bound for S_x.

We are forced to conclude that $y^k = x$, so giving the result. ■

If $x \in \mathbb{R}_{>0}$ and $q = \frac{j}{k} \in \mathbb{Q}$ with $j \in \mathbb{Z}$ and $k \in \mathbb{Z}_{>0}$, we define $x^q = (x^{1/k})^j$.

Let us record the basic properties of addition and multiplication, mirroring analogous results for $\mathbb{Q}$. The properties all follow easily from the similar properties for $\mathbb{Q}$, along with Proposition 2.2.1 and the definition of addition and multiplication in $\mathbb{R}$.

2.2.4 Proposition (Properties of addition and multiplication in $\mathbb{R}$) *Addition and multiplication in $\mathbb{R}$ satisfy the following rules:*

(i) $x_1 + x_2 = x_2 + x_1$, $x_1, x_2 \in \mathbb{R}$ (***commutativity*** *of addition);*

(ii) $(x_1 + x_2) + x_3 = x_1 + (x_2 + x_3)$, $x_1, x_2, x_3 \in \mathbb{R}$ (***associativity*** *of addition);*

(iii) $x + 0 = x$, $t \in \mathbb{R}$ (***additive identity***);

(iv) $x + (-x) = 0$, $x \in \mathbb{R}$ (***additive inverse***);

(v) $x_1 \cdot x_2 = x_2 \cdot x_1$, $x_1, x_2 \in \mathbb{R}$ (***commutativity*** *of multiplication);*

(vi) $(x_1 \cdot x_2) \cdot x_3 = x_1 \cdot (x_2 \cdot x_3)$, $x_1, x_2, x_3 \in \mathbb{R}$ (***associativity*** *of multiplication);*

(vii) $x \cdot 1 = x$, $x \in \mathbb{R}$ (***multiplicative identity***);

(viii) $x \cdot x^{-1} = 1$, $x \in \mathbb{R} \setminus \{0\}$ (***multiplicative inverse***);

(ix) $y \cdot (x_1 + x_2) = y \cdot x_1 + y \cdot x_2$, $y, x_1, x_2 \in \mathbb{R}$ (***distributivity***);

(x) $x^{k_1} \cdot x^{k_2} = x^{k_1 + k_2}$, $x \in \mathbb{R}$, $k_1, k_2 \in \mathbb{Z}_{\geq 0}$.

Moreover, if we define $i_{\mathbb{Q}} \colon \mathbb{Q} \to \mathbb{R}$ *by* $i_{\mathbb{Q}}(q) = [(q)_{j \in \mathbb{Z}_{>0}}]$, *then addition and multiplication in $\mathbb{R}$ agrees with that in $\mathbb{Q}$:*

$$i_{\mathbb{Q}}(q_1) + i_{\mathbb{Q}}(q_2) = i_{\mathbb{Q}}(q_1 + q_2), \quad i_{\mathbb{Q}}(q_1) \cdot i_{\mathbb{Q}}(q_2) = i_{\mathbb{Q}}(q_1 \cdot q_2).$$

As we have done in the past with $\mathbb{Z} \subseteq \mathbb{Q}$, we will often regard $\mathbb{Q}$ as a subset of $\mathbb{R}$ without making explicit mention of the inclusion $i_{\mathbb{Q}}$. Note that this also allows us to think of both $\mathbb{Z}_{\geq 0}$ and $\mathbb{Z}$ as subsets of $\mathbb{R}$, since $\mathbb{Z}_{\geq 0}$ is regarded as a subset of $\mathbb{Z}$, and since $\mathbb{Z} \subseteq \mathbb{Q}$. Of course, this is nothing surprising. Indeed, perhaps the more surprising thing is that it is not actually the case that the definitions do not precisely give $\mathbb{Z}_{\geq 0} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$!

Now is probably a good time to mention that an element of $\mathbb{R}$ that is not in the image of $i_{\mathbb{Q}}$ is called ***irrational***. Also, one can show that the set $\mathbb{Q}$ of rational numbers is countable (Exercise 2.1.3), but that the set $\mathbb{R}$ of real numbers is uncountable (Exercise 2.1.4). Note that it follows that the set of irrational numbers is uncountable, since an uncountable set cannot be a union of two countable sets.

2.2.2 The total order on $\mathbb{R}$

Next we define in $\mathbb{R}$ a natural total order. To do so requires a little work. The approach we take is this. On the set CS($\mathbb{Q}$) of Cauchy sequences in $\mathbb{Q}$ we define a partial order that is *not* a total order. We then show that, for any two Cauchy sequences, in each equivalence class in CS($\mathbb{Q}$) with respect to the equivalence relation of Definition 2.1.16, there exists representatives that can be compared using the order. In this way, while the order on the set of Cauchy sequences is not a total order, there is induced a total order on the set of equivalence classes.

First we define the partial order on the set of Cauchy sequences.

2.2.5 Definition (Partial order on CS($\mathbb{Q}$)) The partial order $\preceq$ on CS($\mathbb{Q}$) is defined by

$$(q_j)_{j\in\mathbb{Z}_{>0}} \preceq (r_j)_{j\in\mathbb{Z}_{>0}} \iff q_j \leq r_j,\ j \in \mathbb{Z}_{>0}. \qquad \bullet$$

This partial order is clearly not a total order. For example, the Cauchy sequences $(\frac{1}{j})_{j\in\mathbb{Z}_{>0}}$ and $(\frac{(-1)^j}{j})_{j\in\mathbb{Z}_{>0}}$ are not comparable with respect to this order. However, what is true is that equivalence classes of Cauchy sequences *are* comparable. We refer the reader to Definition 2.1.16 for the definition of the equivalence relation we denote by $\sim$ in the following result.

2.2.6 Proposition *Let* $(\mathrm{q_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}, (\mathrm{r_j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \in \mathrm{CS}(\mathbb{Q})$ *and suppose that* $(\mathrm{q_j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \not\sim (\mathrm{r_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$. *The following two statements hold:*

(i) There exists $(\tilde{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}, (\tilde{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \in \mathrm{CS}(\mathbb{Q})$ *such that*

(a) $(\tilde{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \sim (\mathrm{q_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *and* $(\tilde{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \sim (\mathrm{r_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$, *and*

(b) either $(\tilde{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \prec (\tilde{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *or* $(\tilde{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \prec (\tilde{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$.

(ii) There does not exist $(\tilde{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}, (\bar{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}, (\tilde{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}, (\bar{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \in \mathrm{CS}(\mathbb{Q})$ *such that*

(a) $(\tilde{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \sim (\bar{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \sim (\mathrm{q_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *and* $(\tilde{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \sim (\bar{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \sim (\mathrm{r_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$, *and*

(b) one of the following two statements holds:

I. $(\tilde{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \prec (\tilde{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *and* $(\bar{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \prec (\bar{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$;

II. $(\tilde{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \prec (\tilde{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *and* $(\bar{\mathrm{q}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}} \prec (\bar{\mathrm{r}}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$.

Proof (i) We begin with a useful lemma.

1 Lemma *With the given hypotheses, there exists* $\delta \in \mathbb{Q}_{>0}$ *and* $\mathrm{N} \in \mathbb{Z}_{>0}$ *such that* $|\mathrm{q_j} - \mathrm{r_j}| \geq \delta$ *for all* $\mathrm{j} \geq \mathrm{N}$.

Proof Since $(q_j - r_j)_{j\in\mathbb{Z}_{>0}}$ does not converge to zero, choose $\epsilon \in \mathbb{Q}_{>0}$ such that, for all $N \in \mathbb{Z}_{>0}$, there exists $j \geq N$ such that $|q_j - r_j| \geq \epsilon$. Now take $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k|, |r_k - r_k| \leq \frac{\epsilon}{4}$ for $j, k \geq N$. Then, by our assumption about ϵ, there exists $\tilde{N} \geq N$ such that $|q_{\tilde{N}} - r_{\tilde{N}}| \geq \epsilon$. Then, for any $j \geq N$, we have

$$\begin{aligned} |q_j - r_j| &= |(q_{\tilde{N}} - r_{\tilde{N}}) - (q_{\tilde{N}} - r_{\tilde{N}}) - (q_j - r_j)| \\ &\geq \|q_{\tilde{N}} - r_{\tilde{N}}| - |(q_{\tilde{N}} - r_{\tilde{N}}) - (q_j - r_j)\| \geq \epsilon - \tfrac{\epsilon}{2}. \end{aligned}$$

The lemma follows by taking $\delta = \frac{\epsilon}{2}$. ▼

Now take N and δ as in the lemma. Then take $\tilde{N} \in \mathbb{Z}_{>0}$ such that $|q_j - q_k|, |r_j - r_k| < \frac{\delta}{2}$ for $j, k \geq \tilde{N}$. Then, using the triangle inequality,

$$|(q_j - r_j) - (q_k - r_k)| \leq \delta, \qquad j, k \geq \tilde{N}.$$

Now take K to be the larger of N and $\tilde{N}$. We then have either $q_K - r_K \geq \delta$ or $r_K - q_K \geq \delta$. First suppose that $q_K - r_K \geq \delta$ and let $j \geq K$. Either $q_j - r_j \geq \delta$ or $r_j - q_j \geq \delta$. If the latter, then

$$q_j - r_j \leq -\delta \quad \implies \quad (q_j - r_k) - (q_K - r_K) \leq 2\delta,$$

contradicting the definition of K. Therefore, we must have $q_j - r_j \geq \delta$ for all $j \geq K$. A similar argument when $r_K - q_K \geq \delta$ shows that $r_j - q_j \geq \delta$ for all $j \geq K$. For $j \in \mathbb{Z}_{>0}$ we then define

$$\tilde{q}_j = \begin{cases} q_K, & j < K, \\ q_j, & j \geq K, \end{cases} \qquad \tilde{r}_j = \begin{cases} r_K, & j < K, \\ r_j, & j \geq K, \end{cases},$$

and we note that the sequences $(\tilde{q}_j)_{j \in \mathbb{Z}_{>0}}$ and $(\tilde{r}_j)_{j \in \mathbb{Z}_{>0}}$ satisfy the required conditions.

(ii) Suppose that

1. $(q_j)_{j \in \mathbb{Z}_{>0}} \asymp (r_j)_{j \in \mathbb{Z}_{>0}}$,
2. $(\tilde{q}_j)_{j \in \mathbb{Z}_{>0}} \sim (\bar{q}_j)_{j \in \mathbb{Z}_{>0}} \sim (q_j)_{j \in \mathbb{Z}_{>0}}$,
3. $(\tilde{r}_j)_{j \in \mathbb{Z}_{>0}} \sim (\bar{r}_j)_{j \in \mathbb{Z}_{>0}} \sim (r_j)_{j \in \mathbb{Z}_{>0}}$, and
4. $(\tilde{q}_j)_{j \in \mathbb{Z}_{>0}} \prec (\tilde{r}_j)_{j \in \mathbb{Z}_{>0}}$.

From the previous part of the proof we know that there exists $\delta \in \mathbb{Q}_{>0}$ and $N \in \mathbb{Z}_{>0}$ such that $\tilde{q}_j - \tilde{r}_j \geq \delta$ for $j \geq N$. Then take $\tilde{N} \in \mathbb{Z}_{>0}$ such that $|\tilde{q}_j - \bar{q}_j|, |\tilde{r}_j - \bar{r}_j| < \frac{\delta}{4}$ for $j \geq \tilde{N}$. This implies that for $j \geq \tilde{N}$ we have

$$|(\tilde{q}_j - \tilde{r}_j) - (\bar{q}_j - \bar{r}_j)| < \tfrac{\delta}{2}.$$

Therefore,

$$(\bar{q}_j - \bar{r}_j) > (\tilde{q}_j - \tilde{r}_j) - \tfrac{\delta}{2}, \qquad j \geq \tilde{N}.$$

If additionally $j \geq N$, then we have

$$(\bar{q}_j - \bar{r}_j) > \delta - \tfrac{\delta}{2} = \tfrac{\delta}{2}.$$

This shows the impossibility of $(\bar{r}_j)_{j \in \mathbb{Z}_{>0}} \prec (\bar{q}_j)_{j \in \mathbb{Z}_{>0}}$. A similar argument shows that $(\tilde{r}_j)_{j \in \mathbb{Z}_{>0}} \prec (\tilde{q}_j)_{j \in \mathbb{Z}_{>0}}$ bars the possibility that $(\bar{q}_j)_{j \in \mathbb{Z}_{>0}} \prec (\bar{r}_j)_{j \in \mathbb{Z}_{>0}}$. ■

Using the preceding result, the following definition then makes sense.

2.2.7 Definition (Order on $\mathbb{R}$) The total order on $\mathbb{R}$ is defined by $x \leq y$ if and only if there exists $(q_j)_{j \in \mathbb{Z}_{>0}}, (r_j)_{j \in \mathbb{Z}_{>0}} \in \mathrm{CS}(\mathbb{Q})$ such that

(i) $x = [(q_j)_{j \in \mathbb{Z}_{>0}}]$ and $y = [(r_j)_{j \in \mathbb{Z}_{>0}}]$ and

(ii) $(q_j)_{j \in \mathbb{Z}_{>0}} \preceq (r_j)_{j \in \mathbb{Z}_{>0}}$. •

Note that we have used the symbol "$\leq$" for the total order on $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. This is justified since, if we think of $\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$, then the various total orders agree (Exercises 2.1.2 and 2.2.5).

We have the usual language and notation we associate with various kinds of numbers.

2.2.8 Definition (Positive and negative real numbers) A real number x is:

(i) ***positive*** if $0 < x$;
(ii) ***negative*** if $x < 0$;
(iii) ***nonnegative*** if $0 \leq x$;
(iv) ***nonpositive*** if $x \leq 0$.

The set of positive real numbers is denoted by $\mathbb{R}_{>0}$, the set of nonnegative real numbers is denoted by $\mathbb{R}_{\geq 0}$, the set of negative real numbers is denoted by $\mathbb{R}_{<0}$, and the set of nonpositive real numbers is denoted by $\mathbb{R}_{\leq 0}$. •

Now is a convenient moment to introduce some simple notation and concepts that are associated with the natural total order on $\mathbb{R}$. The ***signum function*** is the map $\operatorname{sign}\colon \mathbb{R} \to \{-1, 0, 1\}$ defined by

$$\operatorname{sign}(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0. \end{cases}$$

For $x \in \mathbb{R}$, $\lceil x \rceil$ is the ***ceiling*** of x which is the smallest integer not less than x. Similarly, $\lfloor x \rfloor$ is the ***floor*** of x which is the largest integer less than or equal to x. In Figure 2.1 we show the ceiling and floor functions.

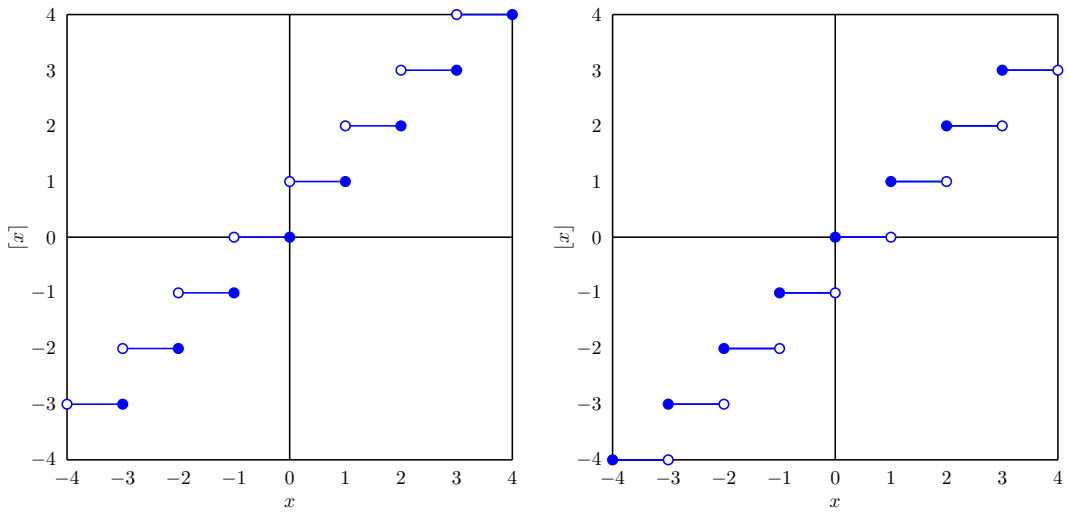


Figure 2.1 The ceiling function (left) and floor function (right)

A consequence of our definition of order is the following extension of the Trichotomy Law to $\mathbb{R}$.

2.2.9 Corollary (Trichotomy Law for $\mathbb{R}$) *For $x, y \in \mathbb{R}$, exactly one of the following possibilities holds:*

(i) $x < y$;
(ii) $y < x$;

(iii) $x = y$.

As with integers and rational numbers, addition and multiplication of real numbers satisfy the expected properties with respect to the total order.

2.2.10 Proposition (Relation between addition and multiplication and <) *For* $x, y, z \in \mathbb{R}$*, the following statements hold:*

(i) *if* $x < y$ *then* $x + z < y + z$*;*

(ii) *if* $x < y$ *and if* $z > 0$ *then* $z \cdot x < z \cdot y$*;*

(iii) *if* $x < y$ *and if* $z < 0$ *then* $z \cdot y < z \cdot x$*;*

(iv) *if* $0 < x, y$ *then* $0 < x \cdot y$*;*

(v) *if* $x < y$ *and if either*

(a) $0 < x, y$ *or*

(b) $x, y < 0$,

then $y^{-1} < x^{-1}$.

Proof These statements all follow from the similar statements for $\mathbb{Q}$, along with Proposition 2.2.6. We leave the straightforward verifications to the reader as Exercise 2.2.4. ■

2.2.3 The absolute value function on $\mathbb{R}$

In this section we generalise the absolute value function on $\mathbb{Q}$. As we shall see in subsequent sections, this absolute value function is essential for providing much of the useful structure of the set of real numbers.

The definition of the absolute value is given as usual.

2.2.11 Definition (Real absolute value function) The ***absolute value function*** on $\mathbb{R}$ is the map from $\mathbb{R}$ to $\mathbb{R}_{\geq 0}$, denoted by $x \mapsto |x|$, defined by

$$|x| = \begin{cases} x, & 0 < x, \\ 0, & x = 0, \\ -x, & x < 0. \end{cases}$$ •

Note that we have used the symbol "$|\cdot|$" for the absolute values on $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. This is justified since, if we think of $\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$, then the various absolute value functions agree (Exercises 2.1.2 and 2.2.5).

The real absolute value function has the expected properties. The proof of the following result is straightforward, and so omitted.

2.2.12 Proposition (Properties of absolute value on $\mathbb{R}$) *The following statements hold:*

(i) $|x| \geq 0$ *for all* $x \in \mathbb{R}$*;*

(ii) $|x| = 0$ *if and only if* $x = 0$*;*

(iii) $|x \cdot y| = |x| \cdot |y|$ *for all* $x, y \in \mathbb{R}$*;*

(iv) $|x + y| \leq |x| + |y|$ *for all* $x, y \in \mathbb{R}$ *(**triangle inequality**);*

(v) $|x^{-1}| = |x|^{-1}$ *for all* $x \in \mathbb{R} \setminus \{0\}$.

2.2.4 Properties of $\mathbb{Q}$ as a subset of $\mathbb{R}$

In this section we give some seemingly obvious, and indeed not difficult to prove, properties of the rational numbers as a subset of the real numbers.

The first property bears the name of Archimedes,[2] but Archimedes actually attributes this to Eudoxus.[3] In any case, it is an Ancient Greek property.

2.2.13 Proposition (Archimedean property of $\mathbb{R}$) *Let $\epsilon \in \mathbb{R}_{>0}$. Then, for any $\mathrm{x} \in \mathbb{R}$ there exists $\mathrm{k} \in \mathbb{Z}_{>0}$ such that $\mathrm{k} \cdot \epsilon > \mathrm{x}$.*

Proof Let $(q_j)_{j\in\mathbb{Z}_{>0}}$ and $(e_j)_{j\in\mathbb{Z}_{>0}}$ be Cauchy sequences in $\mathbb{Q}$ such that $x = [(q_j)_{j\in\mathbb{Z}_{>0}}]$ and $\epsilon = [(e_j)_{j\in\mathbb{Z}_{>0}}]$. By Proposition 2.1.14 there exists $M \in \mathbb{R}_{>0}$ such that $|q_j| < M$ for all $j \in \mathbb{Z}_{>0}$, and by Proposition 2.2.6 we may suppose that $e_j > \delta$ for $j \in \mathbb{Z}_{>0}$, for some $\delta \in \mathbb{Q}_{>0}$. Let $k \in \mathbb{Z}_{>0}$ satisfy $k > \frac{M+1}{\delta}$ (why is this possible?). Then we have

$$k \cdot e_j > \frac{M+1}{\delta} \cdot \delta = M + 1 \geq q_j + 1, \qquad j \in \mathbb{Z}_{>0}.$$

Now consider the sequence $(k \cdot e_j - q_j)_{j\in\mathbb{Z}_{>0}}$. This is a Cauchy sequence by Proposition 2.2.1 since it is a sum of products of Cauchy sequences. Moreover, our computations show that each term in the sequence is larger than 1. Also, this Cauchy sequence has the property that $[(k \cdot e_j - q_j)_{j\in\mathbb{Z}_{>0}}] = k \cdot \epsilon - x$. This shows that $k \cdot \epsilon - x \in \mathbb{R}_{>0}$, so giving the result. ■

The Archimedean property roughly says that there are no real numbers which are greater all rational numbers. The next result says that there are no real numbers that are smaller than all rational numbers.

2.2.14 Proposition (There is no smallest positive real number) *If $\epsilon \in \mathbb{R}_{>0}$ then there exists $\mathrm{q} \in \mathbb{Q}_{>0}$ such that $\mathrm{q} < \epsilon$.*

Proof Since $\epsilon^{-1} \in \mathbb{R}_{>0}$ let $k \in \mathbb{Z}_{>0}$ satisfy $k \cdot 1 > \epsilon^{-1}$ by Proposition 2.2.13. Then taking $q = k^{-1} \in \mathbb{Q}_{>0}$ gives $q < \epsilon$. ■

Using the preceding two results, it is then easy to see that arbitrarily near any real number lies a rational number.

2.2.15 Proposition (Real numbers are well approximated by rational numbers I) *If $\mathrm{x} \in \mathbb{R}$ and if $\epsilon \in \mathbb{R}_{>0}$, then there exists $\mathrm{q} \in \mathbb{Q}$ such that $|\mathrm{x} - \mathrm{q}| < \epsilon$.*

Proof If $x = 0$ then the result follows by taking $q = 0$. Let us next suppose that $x > 0$. If $x < \epsilon$ then the result follows by taking $q = 0$, so we assume that $x \geq \epsilon$. Let $\delta \in \mathbb{Q}_{>0}$ satisfy $\delta < \epsilon$ by Proposition 2.2.14. Then use Proposition 2.2.13 to choose $k \in \mathbb{Z}_{>0}$ to satisfy $k \cdot \delta > x$. Moreover, since $x > 0$, we will assume that k is the smallest such

[2] Archimedes of Syracuse (287 BC–212 BC) was a Greek mathematician and physicist (although in that era such classifications of scientific aptitude were less rigid than they are today). Much of his mathematical work was in the area of geometry, but many of Archimedes' best known achievements were in physics (e.g., the Archimedean Principle in fluid mechanics). The story goes that when the Romans captured Syracuse in 212 BC, Archimedes was discovered working on some mathematical problem, and struck down in the act by a Roman soldier.

[3] Eudoxus of Cnidus (408 BC–355 BC) was a Greek mathematician and astronomer. His mathematical work was concerned with geometry and numbers.

number. Since $x \geq \epsilon$, $k \geq 2$. Thus $(k-1) \cdot \delta \leq x$ since k is the smallest natural number for which $k \cdot \delta > x$. Now we compute

$$0 \leq x - (k-1) \cdot \delta < k \cdot \delta - (k-1) \cdot \delta = \delta < \epsilon.$$

It is now easy to check that the result holds by taking $q = (k-1) \cdot \delta$. The situation when $x < 0$ is easily shown to follow from the situation when $x > 0$. ■

The following stronger result is also useful, and can be proved along the same lines as Proposition 2.2.15, using the Archimedean property of $\mathbb{R}$. The reader is asked to do this as Exercise 2.2.3.

2.2.16 Corollary (Real numbers are well approximated by rational numbers II) *If* $\mathrm{x}, \mathrm{y} \in \mathbb{R}$ *with* $\mathrm{x} < \mathrm{y}$*, then there exists* $\mathrm{q} \in \mathbb{Q}$ *such that* $\mathrm{x} < \mathrm{q} < \mathrm{y}$.

One can also show that irrational numbers have the same property.

2.2.17 Proposition (Real numbers are well approximated by irrational numbers) *If* $\mathrm{x} \in \mathbb{R}$ *and if* $\epsilon \in \mathbb{R}_{>0}$*, then there exists* $\mathrm{y} \in \mathbb{R} \setminus \mathbb{Q}$ *such that* $|\mathrm{x} - \mathrm{y}| < \epsilon$.

Proof By Corollary 2.2.16 choose $q_1, q_2 \in \mathbb{Q}$ such that $x - \epsilon < q_1 < q_2 < x + \epsilon$. Then the number

$$y = q_1 + \frac{q_2 - q_1}{\sqrt{2}}$$

is irrational and satisfies $q_1 < y < q_2$. Therefore, $x - \epsilon < y < x + \epsilon$, or $|x - y| < \epsilon$. ■

It is also possible to state a result regarding the approximation of a collection of real numbers by rational numbers of a certain form. The following result gives one such result.

2.2.18 Theorem (Dirichlet Simultaneous Approximation Theorem) *If* $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{k} \in \mathbb{R}$ *and if* $\mathrm{N} \in \mathbb{Z}_{>0}$*, then there exists* $\mathrm{m} \in \{1, \ldots, \mathrm{N}^\mathrm{k}\}$ *and* $\mathrm{m}_1, \ldots, \mathrm{m}_\mathrm{k} \in \mathbb{Z}$ *such that*

$$\max\{|\mathrm{m}\mathrm{x}_1 - \mathrm{m}_1|, \ldots, |\mathrm{m}\mathrm{x}_\mathrm{k} - \mathrm{m}_\mathrm{k}|\} < \frac{1}{\mathrm{N}}.$$

Proof Let

$$C = [0,1)^k \subseteq \mathbb{R}^k$$

be the "cube" in $\mathbb{R}^k$. For $j \in \{1, \ldots, N\}$ denote $I_j = [\frac{j-1}{N}, \frac{j}{N})$ and note that the sets

$$\{I_{j_1} \times \cdots \times I_{j_k} \subseteq C \mid \ j_1, \ldots, j_k \in \{1, \ldots, N\}\}$$

form a partition of the cube C into N^k "subcubes." Now consider the $N^k + 1$ points

$$\{(lx_1, \ldots, lx_k) \mid \ l \in \{0, 1, \ldots, N^k\}\}$$

in $\mathbb{R}^k$. If $\lfloor x \rfloor$ denotes the floor of $x \in \mathbb{R}$ (i.e., the largest integer less than or equal to x), then

$$\{(lx_1 - \lfloor lx_1 \rfloor, \ldots, lx_k - \lfloor lx_k \rfloor) \mid \ l \in \{0, 1, \ldots, N^k\}\}$$

is a collection of $N^k + 1$ numbers in C. Since C is partitioned into the N^k cubes, it must be that at least two of these $N^k + 1$ points lie in the same cube. Let these points correspond

to $l_1, l_2 \in \{0, 1, \dots, n^k\}$ with $l_2 > l_1$. Then, letting $m = l_2 - l_2$ and $m_j = \lfloor l_2 x_j \rfloor - \lfloor l_1 x_j \rfloor$, $j \in \{1, \dots, k\}$, we have

$$|mx_j - m_j| = |l_2 - \lfloor l_2 x_j \rfloor - (l_1 x_j - \lfloor l_1 x_j \rfloor)| < \frac{1}{N}$$

for every $j \in \{1, \dots, k\}$, which is the result since $m \in \{1, \dots, N^k\}$. ■

2.2.19 Remark (Dirichlet's "pigeonhole principle") The proof of the preceding theorem is a clever application of the so-called "pigeonhole principle," whose use seems to have been pioneered by Dirichlet. The idea behind this principle is simple. One uses the problem data to define elements $x_1, \dots, x_m$ of some set S. One then constructs a partition $(S_1, \dots, S_k)$ of S with the property that, if any $x_{j_1}, x_{j_2} \in S_l$ for some $l \in \{1, \dots, k\}$ and some $j_1, j_2 \in \{1, \dots, m\}$, then the desired result holds. If $k > m$ this is automatically satisfied. •

Note that the previous result gives an arbitrarily accurate simultaneous approximation of the numbers $x_1, \dots, x_j$ by rational numbers with the same denominator since we have

$$\left| x_j - \frac{m_j}{m} \right| < \frac{1}{mN^k} \leq \frac{1}{N^{k+1}}.$$

By choosing N large, our simultaneous approximations can be made as good as desired.

Let us now ask a somewhat different sort of question. Given a fixed set $a_1, \dots, a_k \in \mathbb{R}$, what are the conditions on these numbers such that, given *any* set $x_1, \dots, x_k \in \mathbb{R}$, we can find another number $b \in \mathbb{R}$ such that the approximations $|ba_j - x_j|$, $j \in \{1, \dots, k\}$, are arbitrarily close to integer multiples of a certain number. The exact reason why this is interesting is not immediately clear, but becomes clear in Theorem **??** when we talk about the geometry of the unit circle in the complex plane. In any event, the following result addresses this approximation question, making reference to the notion of linear independence which we discuss in Section **??**. In the statement of the theorem, we think of $\mathbb{R}$ as being a $\mathbb{Q}$-vector space.

2.2.20 Theorem (Kronecker Approximation Theorem) *For* $\mathrm{a}_1, \dots, \mathrm{a}_k \in \mathbb{R}$ *and* $\Delta \in \mathbb{R}_{>0}$ *the following statements hold:*

(i) if $\{\mathrm{a}_1, \dots, \mathrm{a}_k\}$ *are linearly over* $\mathbb{Q}$ *then, for any* $\mathrm{x}_1, \dots, \mathrm{x}_k \in \mathbb{R}$, *for any* $\epsilon \in \mathbb{R}_{>0}$ *and for any* $\mathrm{N} \in \mathbb{Z}_{>0}$, *there exists* $\mathrm{b} \in \mathbb{R}$ *with* $\mathrm{b} > \mathrm{N}$ *and integers* $\mathrm{m}_1, \dots, \mathrm{m}_k$ *such that*

$$\max\{|\mathrm{ba}_1 - \mathrm{x}_1 - \mathrm{m}_1\Delta|, \dots, |\mathrm{ba}_k - \mathrm{x}_k - \mathrm{m}_k\Delta|\} < \epsilon;$$

(ii) if $\{\Delta, \mathrm{a}_1, \dots, \mathrm{a}_k\}$ *are linearly over* $\mathbb{Q}$ *then, for any* $\mathrm{x}_1, \dots, \mathrm{x}_k \in \mathbb{R}$, *for any* $\epsilon \in \mathbb{R}_{>0}$, *and for any* $\mathrm{N} \in \mathbb{Z}_{>0}$, *there exists* $\mathrm{b} \in \mathbb{Z}$ *with* $\mathrm{b} > \mathrm{N}$ *and integers* $\mathrm{m}_1, \dots, \mathrm{m}_k$ *such that*

$$\max\{|\mathrm{ba}_1 - \mathrm{x}_1 - \mathrm{m}_1\Delta|, \dots, |\mathrm{ba}_k - \mathrm{x}_k - \mathrm{m}_k\Delta|\} < \epsilon.$$

Proof Let us first suppose that $\Delta = 1$.

We prove the two assertions together, using induction on k.

First we prove (i) for $k = 1$. Thus suppose that $\{a_1\} \neq \{0\}$. Let $x_1 \in \mathbb{R}$, let $\epsilon \in \mathbb{R}_{>0}$, and let $N \in \mathbb{Z}_{>0}$. If m_1 is an integer greater than N and if $b = a_1^{-1}(x_1 + m_1)$, then we have $ba_1 - x_1 - m_1 = 0$, giving the result in this case.

Next we prove that if (i) holds for $k = r$ then (ii) also holds for $k = r$. Thus suppose that $\{1, a_1, \dots, a_r\}$ are linearly independent over $\mathbb{Q}$. Let $x_1, \dots, x_r \in \mathbb{R}$, let $\epsilon \in \mathbb{R}_{>0}$, and let $N \in \mathbb{Z}_{>0}$. By the Dirichlet Simultaneous Approximation Theorem, let $m, m'_1, \dots, m'_r \in \mathbb{Z}$ with $m \in \mathbb{Z}_{>0}$ be such that

$$|ma_j - m'_j| < \frac{\epsilon}{2}, \qquad j \in \{1, \dots, r\}.$$

We claim that $\{ma_1 - m'_1, \dots, ma_r - m'_r\}$ are linearly independent over $\mathbb{Q}$. Indeed, suppose that

$$q_1(ma_1 - m'_1) + \cdots + q_r(ma_r - m'_r) = 0$$

for some $q_1, \dots, q_r \in \mathbb{Q}$. Then we have

$$(mq_1)a_1 + \cdots + (mq_r)a_r) - (m'_1q_1 + \cdots + m'_rq_r)1 = 0.$$

By linear independence of $\{1, a_1, \dots, a_r\}$ over $\mathbb{Q}$ it follows that $mq_j = 0$, $j \in \{1, \dots, r\}$, and so $q_j = 0$, $j \in \{1, \dots, r\}$, giving the desired linear independence. Since $\{ma_1 - m'_1, \dots, ma_r - m'_r\}$ are linearly independent over $\mathbb{Q}$, we may use our assumption that (i) holds for $k = r$ to give the existence of $b' \in \mathbb{R}$ with $b' > N + 1$ and integers $m''_1, \dots, m''_r$ such that

$$|b'(ma_j - m'_j) - x_j - m''_j| < \frac{\epsilon}{2}, \qquad j \in \{1, \dots, r\}.$$

Now let $b = \lfloor b' \rfloor m > N$ and $m_j = m''_j + \lfloor b' \rfloor m'_j$, $j \in \{1, \dots, k\}$. Using the triangle inequality we have

$$\begin{aligned}
|ba_j - x_j - m_j| &= |\lfloor b'm \rfloor a_j - x_j - (m''_j + \lfloor b' \rfloor m'_j)| \\
&= |\lfloor b' \rfloor (ma_j - m'_j) - x_j - m''_j| \\
&= |(\lfloor b' \rfloor - b')(ma_j - m'_j) + b'(ma_j - m'_j) - x_j - m''_j| \\
&\leq |(\lfloor b' \rfloor - b')(ma_j - m'_j)| + |b'(ma_j - m'_j) - x_j - m''_j| < \epsilon,
\end{aligned}$$

as desired.

Now we prove that (ii) with $k = r$ implies (i) with $k = r + 1$. Thus let $a_1, \dots, a_{r+1}$ be linearly independent over $\mathbb{Q}$. Let $x_1, \dots, x_{r+1} \in \mathbb{R}$, let $\epsilon \in \mathbb{R}_{>0}$, and let $N \in \mathbb{Z}_{>0}$. Note that linear independence implies that $a_{r+1} \neq 0$ (see Proposition **??**(**??**)). We claim that $\{1, \frac{a_1}{a_{r+1}}, \dots, \frac{a_r}{a_{r+1}}\}$ are linearly independent over $\mathbb{Q}$. Since (ii) holds for $k = r$ there exists $b' \in \mathbb{Z}$ with $b' > N$ and integers $m'_1, \dots, m'_r$ such that

$$\left| b' \frac{a_j}{a_{r+1}} - \left(x_j - x_{r+1} \frac{a_j}{a_{r+1}} \right) - m'_j \right| < \epsilon, \qquad j \in \{1, \dots, r\}.$$

Rewriting this as

$$\left| \left(\frac{b' + x_{r+1}}{a_{r+1}} \right) a_j - x_j - m'_j \right| < \epsilon, \qquad j \in \{1, \dots, r\},$$

and noting that

$$\Big(\frac{b' + x_{r+1}}{a_{r+1}}\Big)a_{r+1} - x_{r+1} - b' = 0,$$

which gives (i) by taking

$$b = \frac{b' + x_{r+1}}{a_{r+1}},\ m_1 = m_1',\ \ldots,\ m_r = m_r',\ m_{r+1} = b'.$$

The above induction arguments give the theorem with $\Delta = 1$. Now let us relax the assumption that $\Delta = 1$. Thus let $\Delta \in \mathbb{R}_{>0}$. Let us define $a_j' = \Delta^{-1}a_j$, $j \in \{1, \ldots, k\}$. We claim that $\{a_1', \ldots, a_k'\}$ is linearly independent over $\mathbb{Q}$ if $\{a_1, \ldots, a_k\}$ is linearly independent over $\mathbb{Q}$. Indeed, suppose that

$$q_1a_1' + \cdots + q_ka_k' = 0$$

for some $q_1, \ldots, q_k \in \mathbb{Q}$. Multiplying by Δ and using the linear independence of $\{a_1, \ldots, a_k\}$ immediately gives $q_j = 0$, $j \in \{1, \ldots, k\}$. We also claim that $\{1, a_1', \ldots, a_k'\}$ is linearly independent over $\mathbb{Q}$ if $\{\Delta, a_1, \ldots, a_k\}$ is linearly independent over $\mathbb{Q}$. Indeed, suppose that

$$q_0\,1 + q_1a_1' + \cdots + q_ka_k' = 0$$

for some $q_0, q_1, \ldots, q_k \in \mathbb{Q}$. Multiplying by Δ and using the linear independence of $\{\Delta, a_1, \ldots, a_k\}$ immediately gives $q_j = 0$, $j \in \{1, \ldots, k\}$. Let $x_1, \ldots, x_k \in \mathbb{R}$, $\epsilon \in \mathbb{R}_{>0}$, and $N \in \mathbb{Z}$. Define $x_j' = \Delta^{-1}x_j$, $j \in \{1, \ldots, k\}$. Since the theorem holds for $\Delta = 1$, there exists $b > N$ (with $b \in \mathbb{R}$ for part (i) and $b \in \mathbb{Z}$ for part (ii)) such that

$$|ba_j' - x_j' - m_1| < \frac{\epsilon}{\Delta}, \qquad j \in \{1, \ldots, k\}.$$

Multiplying the inequality by Δ gives the result. ■

2.2.5 The extended real line

It is sometimes convenient to be able to talk about the concept of "infinity" in a somewhat precise way. We do so by using the following idea.

2.2.21 Definition (Extended real line) The *extended real line* is the set $\mathbb{R} \cup \{-\infty\} \cup \{\infty\}$, and we denote this set by $\overline{\mathbb{R}}$. •

Note that in this definition the symbols "$-\infty$" and "∞" are to simply be thought of as labels given to the elements of the singletons $\{-\infty\}$ and $\{\infty\}$. That they somehow correspond to our ideas of what "infinity" means is a consequence of placing some additional structure on $\overline{\mathbb{R}}$, as we now describe.

First we define "arithmetic" in $\overline{\mathbb{R}}$. We can also define some rules for arithmetic in $\overline{\mathbb{R}}$.

2.2.22 Definition (Addition and multiplication in $\overline{\mathbb{R}}$) For $x, y \in \overline{\mathbb{R}}$, define

$$x + y = \begin{cases} x + y, & x, y \in \mathbb{R}, \\ \infty, & x \in \mathbb{R},\ y = \infty, \text{ or } x = \infty,\ y \in \mathbb{R}, \\ \infty, & x = y = \infty, \\ -\infty, & x = -\infty,\ y \in \mathbb{R} \text{ or } x \in \mathbb{R},\ y = -\infty, \\ -\infty, & x = y = -\infty. \end{cases}$$

The operations $\infty + (-\infty)$ and $(-\infty) + \infty$ are undefined. Also define

$$x \cdot y = \begin{cases} x \cdot y, & x, y \in \mathbb{R}, \\ \infty, & x \in \mathbb{R}_{>0},\ y = \infty, \text{ or } x = \infty,\ y \in \mathbb{R}_{>0}, \\ \infty, & x \in \mathbb{R}_{<0},\ y = -\infty, \text{ or } x = -\infty,\ y \in \mathbb{R}_{<0}, \\ \infty, & x = y = \infty, \text{ or } x = y = -\infty, \\ -\infty, & x \in \mathbb{R}_{>0},\ y = -\infty, \text{ or } x = -\infty,\ y \in \mathbb{R}_{>0}, \\ -\infty, & x \in \mathbb{R}_{<0},\ y = \infty, \text{ or } x = \infty,\ y \in \mathbb{R}_{<0}, \\ -\infty, & x = \infty,\ y = -\infty \text{ or } x = -\infty,\ y = \infty, \\ 0, & x = 0, y \in \{-\infty, \infty\} \text{ or } x \in \{-\infty, \infty\},\ y = 0. \end{cases}$$

•

2.2.23 Remarks (Algebra in $\overline{\mathbb{R}}$)

1. The above definitions of addition and multiplication on $\overline{\mathbb{R}}$ *do not* make this a field. Thus, in some sense, the operations are simply notation, since they do not have the usual properties we associate with addition and multiplication.
2. Note we *do* allow multiplication between 0 and $-\infty$ and ∞. This convention is not universally agreed upon, but it will be useful for us to do adopt this convention in Chapter **??**. •

2.2.24 Definition (Order on $\overline{\mathbb{R}}$) For $x, y \in \overline{\mathbb{R}}$, write

$$x \le y \iff \begin{cases} x = y, & \text{or} \\ x, y \in \mathbb{R},\ x \le y, & \text{or} \\ x \in \mathbb{R},\ y = \infty, & \text{or} \\ x = -\infty,\ y \in \mathbb{R}, & \text{or} \\ x = -\infty,\ y = \infty. & \end{cases}$$

•

This is readily verified to be a total order on $\overline{\mathbb{R}}$, with $-\infty$ being the least element and ∞ being the greatest element of $\overline{\mathbb{R}}$. As with $\mathbb{R}$, we have the notation

$$\overline{\mathbb{R}}_{>0} = \{x \in \overline{\mathbb{R}} \mid x > 0\}, \quad \overline{\mathbb{R}}_{\ge 0} = \{x \in \overline{\mathbb{R}} \mid x \ge 0\}.$$

Finally, we can extend the absolute value on $\mathbb{R}$ to $\overline{\mathbb{R}}$.

2.2.25 Definition (Extended real absolute value function) The ***extended real absolute function*** is the map from $\overline{\mathbb{R}}$ to $\overline{\mathbb{R}}_{\geq 0}$, denoted by $x \mapsto |x|$, and defined by

$$|x| = \begin{cases} |x|, & x \in \mathbb{R}, \\ \infty, & x = \infty, \\ \infty, & x = -\infty. \end{cases}$$ •

2.2.6 sup and inf

We recall from Definition 1.5.11 the notation $\sup S$ and $\inf S$ for the least upper bound and greatest lower bound, respectively, associated to a partial order. This construction applies, in particular to the partially ordered set $(\overline{\mathbb{R}}, \leq)$. Note that if $A \subseteq \mathbb{R}$ then we might possibly have $\sup(A) = \infty$ and/or $\inf(A) = -\infty$. In brief section we give a few properties of sup and inf.

The following property of sup and inf is often useful.

2.2.26 Lemma (Property of sup and inf) *Let* $\mathrm{A} \subseteq \mathbb{R}$ *be such that* $\inf(\mathrm{A}), \sup(\mathrm{A}) \in \mathbb{R}$ *and let* $\epsilon \in \mathbb{R}_{>0}$. *Then there exists* $\mathrm{x}_+, \mathrm{x}_- \in \mathrm{A}$ *such that*

$$\mathrm{x}_+ + \epsilon > \sup(\mathrm{A}), \quad \mathrm{x}_- - \epsilon < \inf(\mathrm{A}).$$

Proof We prove the assertion for sup, as the assertion for inf follows along similar lines, of course. Suppose that there is no $x_+ \in A$ such that $x_+ + \epsilon > \sup(A)$. Then $x \leq \sup(A) - \epsilon$ for every $x \in A$, and so $\sup(A) - \epsilon$ is an upper bound for A. But this contradicts $\sup(A)$ being the least upper bound. ■

Let us record and prove the properties of interest for sup.

2.2.27 Proposition (Properties of sup) *For subsets* $\mathrm{A}, \mathrm{B} \subseteq \mathbb{R}$ *and for* $\mathrm{a} \in \mathbb{R}_{>0}$, *the following statements hold:*

(i) *if* $\mathrm{A} + \mathrm{B} = \{\mathrm{x} + \mathrm{y} \mid \mathrm{x} \in \mathrm{A},\ \mathrm{y} \in \mathrm{B}\}$, *then* $\sup(\mathrm{A} + \mathrm{B}) = \sup(\mathrm{A}) + \sup(\mathrm{B})$;

(ii) *if* $-\mathrm{A} = \{-\mathrm{x} \mid \mathrm{x} \in \mathrm{A}\}$, *then* $\sup(-\mathrm{A}) = -\inf(\mathrm{A})$;

(iii) *if* $\mathrm{aA} = \{\mathrm{ax} \mid \mathrm{x} \in \mathrm{A}\}$, *then* $\sup(\mathrm{aA}) = \mathrm{a}\sup(\mathrm{A})$;

(iv) *if* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval, if* $\mathrm{A} \subseteq \mathbb{R}$, *if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is strictly monotonically (see Definition 3.1.27), and if* $\mathrm{f}(\mathrm{A}) = \{\mathrm{f}(\mathrm{x}) \mid \mathrm{x} \in \mathrm{A}\}$, *then* $\sup(\mathrm{f}(\mathrm{A})) = \mathrm{f}(\sup(\mathrm{A}))$.

Proof (i) Let $x \in A$ and $y \in B$ so that $x + y \in A + B$. Then $x + y \leq \sup A + \sup B$ which implies that $\sup A + \sup B$ is an upper bound for $A + B$. Since $\sup(A + B)$ is the least upper bound this implies that $\sup(A + B) \leq \sup A + \sup B$. Now let $\epsilon \in \mathbb{R}_{>0}$ and let $x \in A$ and $y \in B$ satisfy $\sup A - x < \frac{\epsilon}{2}$ and $\sup B - y < \frac{\epsilon}{2}$. Then

$$\sup A + \sup B - (x + y) < \epsilon.$$

Thus, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $x + y \in A + B$ such that $\sup A + \sup B - (x + y) < \epsilon$. Therefore, $\sup A + \sup B \leq \sup(A + B)$.

(ii) Let $x \in -A$. Then $\sup(-A) \geq x$ or $-\sup(-A) \leq -x$. Thus $-\sup(-A)$ is a lower bound for A and so $\inf(A) \geq -\sup(-A)$. Next let $\epsilon \in \mathbb{R}_{>0}$ and let $x \in -A$ satisfy $x + \epsilon > \sup(-A)$. Then $-x - \epsilon < -\sup(-A)$. Thus, for every $\epsilon \in \mathbb{R}_{>0}$, there exists $y \in A$ such that $y - (-\sup(-A)) < \epsilon$. Thus $-\sup(-A) \geq \inf(A)$, giving this part of the result.

(iii) Let $x \in A$ and note that since $\sup(A) \geq x$, we have $a \sup(A) \geq ax$. Thus $a \sup(A)$ is an upper bound for aA, and so we must have $\sup(aA) \leq a \sup(A)$. Now let $\epsilon \in \mathbb{R}_{>0}$ and let $x \in A$ be such that $x + \frac{\epsilon}{a} > \sup(A)$. Then $ax + \epsilon > a \sup(A)$. Thus, given $\epsilon \in \mathbb{R}_{>0}$ there exists $y \in aA$ such that $a \sup(A) - ax < \epsilon$. Thus $a \sup(A) \leq \sup(aA)$.

(iv) ***missing stuff*** ■

For inf the result is, of course, quite similar. We leave the proof, which mirrors the above proof for sup, to the reader.

2.2.28 Proposition (Properties of inf) *For subsets* $A, B \subseteq \mathbb{R}$ *and for* $a \in \mathbb{R}_{\geq 0}$, *the following statements hold:*

(i) *if* $A + B = \{x + y \mid x \in A,\ y \in B\}$, *then* $\inf(A + B) = \inf(A) + \inf(B)$;

(ii) *if* $-A = \{-x \mid x \in A\}$, *then* $\inf(-A) = -\sup(A)$;

(iii) *if* $aA = \{ax \mid x \in A\}$, *then* $\inf(aA) = a \inf(A)$;

(iv) *if* $I \subseteq \mathbb{R}$ *is an interval, if* $A \subseteq \mathbb{R}$, *if* $f : I \to \mathbb{R}$ *is strictly monotonically (see Definition 3.1.27), and if* $f(A) = \{f(x) \mid x \in A\}$, *then* $\inf(f(A)) = f(\inf(A))$.

If $S \subseteq \mathbb{R}$ is a *finite* set, then both $\sup S$ and $\inf S$ are elements of S. In this case we might denote $\max S = \sup S$ and $\min S = \inf S$.

2.2.7 Notes

The Archimedean property of $\mathbb{R}$ seems obvious. The lack of the Archimedean property would mean that there exists t for which $t > N$ for every natural number N. This property is actually possessed by certain fields used in so-called "nonstandard analysis," and we refer the interested reader to [Robinson 1974].

Theorem 2.2.18 is due to Dirichlet [1842], and the proof is a famous use of the "pigeonhole principle." Theorem 2.2.20 is due to [Kronecker 1899], and the proof we give is from [Kueh 1986].

Exercises

2.2.1 Prove the ***Binomial Theorem*** which states that, for $x, y \in \mathbb{R}$ and $k \in \mathbb{Z}_{>0}$,

$$(x + y)^k = \sum_{j=0}^{k} B_{k,j} x^j y^{k-j},$$

where

$$B_{k,j} = \binom{k}{j} \triangleq \frac{k!}{j!(k-j)!}, \qquad j, k \in \mathbb{Z}_{>0},\ j \leq k,$$

are the ***binomial coefficients***, and $k! = 1 \cdot 2 \cdot \cdots \cdot k$ is the ***factorial*** of k. We take the convention that $0! = 1$.

2.2.2 Let $q \in \mathbb{Q} \setminus \{0\}$ and $x \in \mathbb{R} \setminus \mathbb{Q}$. Show the following:

(a) $q + x$ is irrational;

(b) qx is irrational;

(c) $\frac{x}{q}$ is irrational;

(d) $\frac{q}{x}$ is irrational.

2.2.3 Prove Corollary 2.2.16.

2.2.4 Prove Proposition 2.2.10.

2.2.5 Show that the order and absolute value on $\mathbb{R}$ agree with those on $\mathbb{Q}$. That is to say, show the following:

(a) for $q, r \in \mathbb{Q}$, $q < r$ if and only if $i_{\mathbb{Q}}(q) < i_{\mathbb{Q}}(r)$;

(b) for $q \in \mathbb{Q}$, $|q| = |i_{\mathbb{Q}}(q)|$.

(Note that we see clearly here the abuse of notation that follows from using $<$ for both the order on $\mathbb{Z}$ and $\mathbb{Q}$ and from using $|\cdot|$ as the absolute value both on $\mathbb{Z}$ and $\mathbb{Q}$. It is expected that the reader can understand where the notational abuse occurs.)

2.2.6 Do the following:

(a) show that if $x \in \mathbb{R}_{>0}$ satisfies $x < 1$, then $x^k < x$ for each $k \in \mathbb{Z}_{>0}$ satisfying $k \geq 2$;

(b) show that if $x \in \mathbb{R}_{>0}$ satisfies $x > 1$, then $x^k > x$ for each $k \in \mathbb{Z}_{>0}$ satisfying $k \geq 2$.

2.2.7 Show that, for $t, s \in \mathbb{R}$, $||t| - |s|| \leq |t - s|$.

2.2.8 Show that if $s, t \in \mathbb{R}$ satisfy $s < t$, then there exists $q \in \mathbb{Q}$ such that $s < q < t$.

Section 2.3

Sequences in $\mathbb{R}$

In our construction of the real numbers, sequences played a key rôle, inasmuch as Cauchy sequences of rational numbers were integral to our definition of real numbers. In this section we study sequences of real numbers. In particular, in Theorem 2.3.5 we prove the result, absolutely fundamental in analysis, that $\mathbb{R}$ is "complete," meaning that Cauchy sequences of real numbers converge.

Do I need to read this section? If you do not already know the material in this section, then it ought to be read. It is also worth the reader spending some time over the idea that Cauchy sequences of real numbers converge, as compared to rational numbers where this is not the case. The same idea will arise in more abstract settings in Chapter **??**, and so it will pay to understand it well in the simplest case. •

2.3.1 Definitions and properties of sequences

In this section we consider the extension to $\mathbb{R}$ of some of the ideas considered in Section 2.1.2 concerning sequences in $\mathbb{Q}$. As we shall see, it is via sequences, and other equivalent properties, that the nature of the difference between $\mathbb{Q}$ and $\mathbb{R}$ is spelled out quite clearly.

We begin with definitions, generalising in a trivial way the similar definitions for $\mathbb{Q}$.

2.3.1 Definition (Cauchy sequence, convergent sequence, bounded sequence, monotone sequence) Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}$. The sequence:

(i) is a ***Cauchy sequence*** if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|x_j - x_k| < \epsilon$ for $j, k \geq N$;

(ii) ***converges to*** $\mathbf{s_0}$ if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|x_j - s_0| < \epsilon$ for $j \geq N$;

(iii) ***diverges*** if it does not converge to any element in $\mathbb{R}$;

(iv) is ***bounded above*** if there exists $M \in \mathbb{R}$ such that $x_j < M$ for each $j \in \mathbb{Z}_{>0}$;

(v) is ***bounded below*** if there exists $M \in \mathbb{R}$ such that $x_j > M$ for each $j \in \mathbb{Z}_{>0}$;

(vi) is ***bounded*** if there exists $M \in \mathbb{R}_{>0}$ such that $|x_j| < M$ for each $j \in \mathbb{Z}_{>0}$;

(vii) is ***monotonically increasing*** if $x_{j+1} \geq x_j$ for $j \in \mathbb{Z}_{>0}$;

(viii) is ***strictly monotonically increasing*** if $x_{j+1} > x_j$ for $j \in \mathbb{Z}_{>0}$;

(ix) is ***monotonically decreasing*** if $x_{j+1} \leq x_j$ for $j \in \mathbb{Z}_{>0}$;

(x) is ***strictly monotonically decreasing*** if $x_{j+1} < x_j$ for $j \in \mathbb{Z}_{>0}$;

(xi) is ***constant*** if $x_j = x_1$ for every $j \in \mathbb{Z}_{>0}$;

(xii) is ***eventually constant*** if there exists $N \in \mathbb{Z}_{>0}$ such that $x_j = x_N$ for every $j \geq N$. •

Associated with the notion of convergence is the notion of a limit. We also, for convenience, wish to allow sequences with infinite limits. This makes for some rather subtle use of language, so the reader should pay attention to this.

2.3.2 Definition (Limit of a sequence) Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence.

(i) If $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to s_0, then the sequence has s_0 as a ***limit***, and we write $\lim_{j\to\infty} x_j = s_0$.

(ii) If, for every $M \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $x_j > M$ (resp. $x_k < -M$) for $j \geq N$, then the sequence ***diverges to ∞*** (resp. ***diverges to $-\infty$***), and we write $\lim_{j\to\infty} x_j = \infty$ (resp. $\lim_{j\to\infty} x_j = -\infty$);

(iii) If $\lim_{j\to\infty} x_j \in \mathbb{R}$, then the limit of the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ ***exists***.

(iv) If the limit of the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ does not exist, does not diverge to ∞, or does not diverge to $-\infty$, then the sequence is ***oscillatory***. •

The reader can prove in Exercise 2.3.1 that limits, if they exist, are unique.

That convergent sequences are Cauchy, and that Cauchy sequences are bounded follows in exactly the same manner as the analogous results, stated as Propositions 2.1.13 and 2.1.14, for $\mathbb{Q}$. Let us state the results here for reference.

2.3.3 Proposition (Convergent sequences are Cauchy) *If a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to x_0, then it is a Cauchy sequence.*

2.3.4 Proposition (Cauchy sequences are bounded) *If $(x_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ then it is bounded.*

Moreover, what is true for $\mathbb{R}$, and that is not true for $\mathbb{Q}$, is that every Cauchy sequence converges.

2.3.5 Theorem (Cauchy sequences in $\mathbb{R}$ converge) *If $(x_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ then there exists $s_0 \in \mathbb{R}$ such that $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to s_0.*

Proof For $j \in \mathbb{Z}_{>0}$ choose $q_j \in \mathbb{Q}_{>0}$ such that $|x_j - q_j| < \frac{1}{j}$, this being possible by Proposition 2.2.15. For $\epsilon \in \mathbb{R}_{>0}$ let $N_1 \in \mathbb{Z}_{>0}$ satisfy $|x_j - x_k| < \frac{\epsilon}{2}$ for $j, k \geq N_1$. By Proposition 2.2.13 let $N_2 \in \mathbb{Z}_{>0}$ satisfy $N_2 \cdot 1 > 4\epsilon^{-1}$, and let N be the larger of N_1 and N_2. Then, for $j, k \geq N$, we have

$$|q_j - q_k| = |q_j - x_j + x_j - x_k + x_k - q_k| \leq |x_j - q_j| + |x_j - x_k| + |x_k - q_k| < \tfrac{1}{j} + \tfrac{\epsilon}{2} + \tfrac{1}{k} < \epsilon.$$

Thus $(q_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence, and so we define $s_0 = [(q_j)_{j\in\mathbb{Z}_{>0}}]$.

Now we show that $(q_j)_{j\in\mathbb{Z}_{>0}}$ converges to s_0. Let $\epsilon \in \mathbb{R}_{>0}$ and take $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < \frac{\epsilon}{2}$, $j, k \geq N$, and rewrite this as

$$\tfrac{\epsilon}{2} < q_j - q_k + \epsilon, \quad \tfrac{\epsilon}{2} < -q_k + q_k + \epsilon, \qquad j, k \geq N. \tag{2.4}$$

For $j_0 \geq N$ consider the sequence $(q_j - q_{j_0} + \epsilon)_{j\in\mathbb{Z}_{>0}}$. This is a Cauchy sequence by Proposition 2.2.1. Moreover, by Proposition 2.2.6, $[(q_j - q_{j_0} + \epsilon)_{j\in\mathbb{Z}_{>0}}] > 0$, using the first of the inequalities in (2.4). Thus we have $s_0 - q_{j_0} + \epsilon > 0$, or

$$-\epsilon < s_0 - q_{j_0}, \qquad j_0 \geq N.$$

Arguing similarly, but using the second of the inequalities (2.4), we determine that

$$s_0 - q_{j_0} < \epsilon, \qquad j_0 \geq N.$$

This gives $|s_0 - q_j| < \epsilon$ for $j \geq N$, so showing that $(q_j)_{j \in \mathbb{Z}_{>0}}$ converges to s_0.

Finally, we show that $(x_j)_{j \in \mathbb{Z}_{>0}}$ converges to s_0. Let $\epsilon \in \mathbb{R}_{>0}$ and take $N_1 \in \mathbb{Z}_{>0}$ such that $|s_0 - q_j| < \frac{\epsilon}{2}$ for $j \geq N_1$. Also choose $N_2 \in \mathbb{Z}_{>0}$ such that $N_2 \cdot 1 > 2\epsilon^{-1}$ by Proposition 2.2.13. If N is the larger of N_1 and N_2, then we have

$$|s_0 - x_j| = |s_0 - q_j + q_j - x_j| \leq |s_0 - q_j| + |q_j - x_j| < \tfrac{\epsilon}{2} + \tfrac{1}{j} < \epsilon,$$

for $j \geq N$, so giving the result. ■

2.3.6 Remark (Completeness of $\mathbb{R}$) The property of $\mathbb{R}$ that Cauchy sequences are convergent gives, in the more general setting of Section **??**, $\mathbb{R}$ the property of being ***complete***. Completeness is an extremely important concept in analysis. We shall say some words about this in Section **??**; for now let us just say that the subject of calculus would not exist, but for the completeness of $\mathbb{R}$. •

2.3.2 Some properties equivalent to the completeness of $\mathbb{R}$

Using the fact that Cauchy sequences converge, it is easy to prove two other important features of $\mathbb{R}$, both of which seem obvious intuitively.

2.3.7 Theorem (Bounded subsets of $\mathbb{R}$ have a least upper bound) *If $S \subseteq \mathbb{R}$ is nonempty and possesses an upper bound with respect to the standard total order $\leq$, then S possesses a least upper bound with respect to the same total order.*

Proof Since S has an upper bound, there exists $y \in \mathbb{R}$ such that $x \leq y$ for all $x \in S$. Now choose some $x \in S$. We then define two sequences $(x_j)_{j \in \mathbb{Z}_{>0}}$ and $(y_j)_{j \in \mathbb{Z}_{>0}}$ recursively as follows:

1. define $x_1 = x$ and $y_1 = y$;
2. suppose that x_j and y_j have been defined;
3. if there exists $z \in S$ with $\frac{1}{2}(x_j + y_j) < z \leq y_j$, take $x_{j+1} = z$ and $y_{j+1} = y_j$;
4. if there is no $z \in S$ with $\frac{1}{2}(x_j + y_j) < z \leq y_j$, take $x_{j+1} = x_j$ and $y_{j+1} = \frac{1}{2}(x_j + y_j)$.

A lemma characterises these sequences.

1 Lemma *The sequences $(x_j)_{j \in \mathbb{Z}_{>0}}$ and $(y_j)_{j \in \mathbb{Z}_{>0}}$ have the following properties:*

(i) $x_j \in S$ *for* $j \in \mathbb{Z}_{>0}$;

(ii) $x_{j+1} \geq x_j$ *for* $j \in \mathbb{Z}_{>0}$;

(iii) y_j *is an upper bound for* S *for* $j \in \mathbb{Z}_{>0}$;

(iv) $y_{j+1} \leq y_j$ *for* $j \in \mathbb{Z}_{>0}$;

(v) $0 \leq y_j - x_j \leq \frac{1}{2^j}(y - x)$ *for* $j \in \mathbb{Z}_{>0}$.

Proof We prove the result by induction on j. The result is obviously true for = 0. Now suppose the result true for $j \in \{1, \ldots, k\}$.

First take the case where there exists $z \in S$ with $\frac{1}{2}(x_k + y_k) < z \leq y_k$, so that $x_{k+1} = z$ and $y_{k+1} = y_k$. Clearly $x_{k+1} \in S$ and $y_{k+1} \geq y_k$. Since $y_k \geq x_k$ by the induction

hypotheses, $\frac{1}{2}(x_k + y_k) \geq x_k$ giving $x_{k+1} = z \geq x_k$. By the induction hypotheses, y_{k+1} is an upper bound for S. By definition of x_{k+1} and y_{k+1},

$$y_{k+1} - x_{k+1} = y_k - z \geq 0$$

and

$$y_{k+1} - x_{k+1} = y_k - z = y_k - \tfrac{1}{2}(y_k - x_k) = \tfrac{1}{2}(y_k - x_k),$$

giving $y_{k+1} - x_{k+1} \leq \frac{1}{2^{k+1}}(y - x)$ by the induction hypotheses.

Now we take the case where there is no $z \in S$ with $\frac{1}{2}(x_j + y_j) < z \leq y_j$, so that $x_{k+1} = x_k$ and $y_{k+1} = \frac{1}{2}(x_k + y_k)$. Clearly $x_{k+1} \geq x_k$ and $x_{k+1} \in S$. If y_{k+1} were not an upper bound for S, then there exists $a \in S$ such that $a > y_{k+1}$. By the induction hypotheses, y_k is an upper bound for S so $a \leq y_k$. But this means that $\frac{1}{2}(y_k + x_k) < a \leq y_k$, contradicting our assumption concerning the nonexistence of $z \in S$ with $\frac{1}{2}(x_j + y_j) < z \leq y_j$. Thus y_{k+1} is an upper bound for S. Since $x_k \leq y_k$ by the induction hypotheses,

$$y_{k+1} = \tfrac{1}{2}(y_k + x_k) \leq y_k.$$

Also

$$y_{k+1} - x_{k+1} = \tfrac{1}{2}(y_k - x_k)$$

by the induction hypotheses. This completes the proof. ▼

The following lemma records a useful fact about the sequences $(x_j)_{j \in \mathbb{Z}_{>0}}$ and $(y_j)_{j \in \mathbb{Z}_{>0}}$.

2 Lemma *Let* $(\mathrm{x_j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *and* $(\mathrm{y_j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *be sequences in* $\mathbb{R}$ *satisfying:*

(i) $\mathrm{x_{j+1}} \geq \mathrm{x_j}$, $\mathrm{j} \in \mathbb{Z}_{>0}$;

(ii) $\mathrm{y_{j+1}} \leq \mathrm{y_j}$, $\mathrm{j} \in \mathbb{Z}_{>0}$;

(iii) the sequence $(\mathrm{y_j} - \mathrm{x_j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *converges to* 0.

Then $(\mathrm{x_j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *and* $(\mathrm{y_j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *converge, and converge to the same limit.*

Proof First we claim that $x_j \leq y_k$ for all $j, k \in \mathbb{Z}_{>0}$. Indeed, suppose not. Then there exists $j, k \in \mathbb{Z}_{>0}$ such that $x_j > y_k$. If N is the larger of j and k, then we have $y_N \leq y_k < x_j \leq x_N$. This implies that

$$x_m - y_m \geq x_j - y_m \geq x_j - y_k > 0, \qquad m \geq N,$$

which contradicts the fact that $(y_j - x_j)_{j \in \mathbb{Z}_{>0}}$ converges to zero.

Now, for $\epsilon \in \mathbb{R}_{>0}$ let $N \in \mathbb{Z}_{>0}$ satisfy $|y_j - x_j| < \epsilon$ for $j \geq N$, or, simply, $y_j - x_j < \epsilon$ for $j \geq N$. Now let $j, k \geq N$, and suppose that $j \geq k$. Then

$$0 \leq x_j - x_k \leq x_j - y_k < \epsilon.$$

Similarly, if $j \leq k$ we have $0 \leq x_k - x_j < \epsilon$. In other words, $|x_j - x_k| < \epsilon$ for $j, k \geq N$. Thus $(x_j)_{j \in \mathbb{Z}_{>0}}$ is a Cauchy sequence. In like manner one shows that $(y_j)_{j \in \mathbb{Z}_{>0}}$ is also a Cauchy sequence. Therefore, by Theorem 2.3.5, these sequences converge, and let us denote their limits by s_0 and t_0, respectively. However, since $(x_j)_{j \in \mathbb{Z}_{>0}}$ and $(y_j)_{j \in \mathbb{Z}_{>0}}$ are equivalent Cauchy sequences in the sense of Definition 2.1.16, it follows that $s_0 = t_0$. ▼

Using Lemma 1 we easily verify that the sequences $(x_j)_{j\in\mathbb{Z}_{>0}}$ and $(y_j)_{j\in\mathbb{Z}_{>0}}$ satisfy the hypotheses of Lemma 2. Therefore these sequences converge to a common limit, which we denote by s. We claim that s is a least upper bound for S. First we show that it is an upper bound. Suppose that there is $x \in S$ such that $x > s$ and define $\epsilon = x - s$. Since $(y_j)_{j\in\mathbb{Z}_{>0}}$ converges to s, there exists $N \in \mathbb{Z}_{>0}$ such that $|s - y_j| < \epsilon$ for $j \geq N$. Then, for $j \geq N$,

$$y_j - s < \epsilon = x - s,$$

implying that $y_j < x$, and so contradicting Lemma 1.

Finally, we need to show that s is a least upper bound. To see this, let b be an upper bound for S and suppose that $b < s$. Define $\epsilon = s - b$, and choose $N \in \mathbb{Z}_{>0}$ such that $|s - x_j| < \epsilon$ for $j \geq N$. Then

$$s - x_j < \epsilon = s - b,$$

implying that $b < x_j$ for $j \geq N$. This contradicts the fact, from Lemma 1, that $x_j \in S$ and that b is an upper bound for S. ■

As we shall explain more fully in Aside 2.3.9, the least upper bound property of the real numbers as stated in the preceding theorem is actually *equivalent* to the completeness of $\mathbb{R}$. In fact, the least upper bound property forms the basis for an alternative definition of the real numbers using ***Dedekind cuts***.[4] Here the idea is that one defines a real number as being a splitting of the rational numbers into two halves, one corresponding to the rational numbers less than the real number one is defining, and the other corresponding to the rational numbers greater than the real number one is defining. Historically, Dedekind cuts provided the first rigorous construction of the real numbers. We refer to Section 2.3.9 for further discussion. We also comment, as we discuss in Aside 2.3.9, that any construction of the real numbers with the property of completeness, or an equivalent, will produce something that is "essentially" the real numbers as we have defined them.

Another consequence of Theorem 2.3.5 is the following.

2.3.8 Theorem (Bounded, monotonically increasing sequences in $\mathbb{R}$ converge) *If $(x_j)_{j\in\mathbb{Z}_{>0}}$ is a bounded, monotonically increasing sequence in $\mathbb{R}$, then it converges.*

Proof The subset $(x_j)_{j\in\mathbb{Z}_{>0}}$ of $\mathbb{R}$ has an upper bound, since it is bounded. By Theorem 2.3.7 let b be the least upper bound for this set. We claim that $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to b. Indeed, let $\epsilon \in \mathbb{R}_{>0}$. We claim that there exists some $N \in \mathbb{Z}_{>0}$ such that $b - x_N < \epsilon$ since b is a least upper bound. Indeed, if there is no such N, then $b \geq x_j + \epsilon$ for all $j \in \mathbb{Z}_{>0}$ and so $b - \frac{\epsilon}{2}$ is an upper bound for $(x_j)_{j\in\mathbb{Z}_{>0}}$ that is smaller than b. Now, with N chosen so that $b - x_N < \epsilon$, the fact that $(x_j)_{j\in\mathbb{Z}_{>0}}$ is monotonically increasing implies that $|b - x_j| < \epsilon$ for $j \geq N$, as desired. ■

It turns out that Theorems 2.3.5, 2.3.7, and 2.3.8 are equivalent. But to make sense of this requires one to step outside the concrete representation we have given for the real numbers to a more axiomatic one. This can be skipped, so we present it as an aside.

[4] After Julius Wihelm Richard Dedekind (1831–1916), the German mathematician, did work in the areas of analysis, ring theory, and set theory. His rigorous mathematical style has had a strong influence on modern mathematical presentation.

2.3.9 Aside (Complete ordered fields) An *ordered field* is a field $\mathbb{F}$ (see Definition **??** for the definition of a field) equipped with a total order satisfying the conditions

1. if $x < y$ then $x + z < y + z$ for $x, y, z \in \mathbb{F}$ and
2. if $0 < x, y$ then $0 < x \cdot y$.

Note that in an ordered field one can define the absolute value exactly as we have done for $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. There are many examples of ordered fields, of which $\mathbb{Q}$ and $\mathbb{R}$ are two that we have seen. However, if one adds to the conditions for an ordered field an additional condition, then this turns out to essentially uniquely specify the set of real numbers. (We say "essentially" since the uniqueness is up to a bijection that preserves the field structure as well as the order.) This additional structure comes in various forms, of which three are as stated in Theorems 2.3.5, 2.3.7, and 2.3.8. To be precise, we have the following theorem.

Theorem *If $\mathbb{F}$ is an ordered field, then the following statements are equivalent:*

(i) every Cauchy sequence converges;

(ii) each set possessing an upper bound possesses a least upper bound;

(iii) each bounded, monotonically increasing sequence converges.

We have almost proved this theorem with our arguments above. To see this, note that in the proof of Theorem 2.3.7 we use the fact that Cauchy sequences converge. Moreover, the argument can easily be adapted from the special case of $\mathbb{R}$ to a general ordered field. This gives the implication (i) $\implies$ (ii) in the theorem above. In like manner, the proof of Theorem 2.3.8 gives the implication (ii) $\implies$ (iii), since the proof is again easily seen to be valid for a general ordered field. The argument for the implication (iii) $\implies$ (i) is outlined in Exercise 2.3.5. An ordered field satisfying any one of the three equivalent conditions (i), (ii), and (iii) is called a ***complete ordered field***. Thus there is essentially only one complete ordered field, and it is $\mathbb{R}$. ♠

2.3.3 Tests for convergence of sequences

There is generally no algorithmic way, other than checking the definition, to ascertain when a sequence converges. However, there are a few simple results that are often useful, and here we state some of these.

2.3.10 Proposition (Squeezing Principle) *Let $(x_j)_{j\in\mathbb{Z}_{>0}}$, $(y_j)_{j\in\mathbb{Z}_{>0}}$, and $(z_j)_{j\in\mathbb{Z}_{>0}}$ be sequences in $\mathbb{R}$ satisfying*

(i) $x_j \leq z_j \leq y_j$ for all $j \in \mathbb{Z}_{>0}$ and

(ii) $\lim_{j\to\infty} x_j = \lim_{j\to\infty} y_j = \alpha$.

Then $\lim_{j\to\infty} z_j = \alpha$.

Proof Let $\epsilon \in \mathbb{R}_{>0}$ and let $N_1, N_2 \in \mathbb{Z}_{>0}$ have the property that $|x_j - \alpha| < \frac{\epsilon}{3}$ for $j \geq N_1$ and $|y_j - \alpha| < \frac{\epsilon}{3}$. Then, for $j \geq \max\{N_1, N_2\}$,

$$|x_j - y_j| = |x_j - \alpha + \alpha - y_j| \leq |x_j - \alpha| + |y_j - \alpha| < \tfrac{2\epsilon}{3},$$

using the triangle inequality. Then, for $j \geq \max\{N_1, N_2\}$, we have

$$|z_j - \alpha| = |z_j - x_j + x_j - \alpha| \leq |z_j - x_j| + |x_j - \alpha| \leq |y_j - x_j| + |x_j - \alpha| = \epsilon,$$

again using the triangle inequality. ■

The next test for convergence of a series is sometimes useful.

2.3.11 Proposition (Ratio Test for sequences) *Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = \alpha$. If $\alpha < 1$ then the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to 0, and if $\alpha > 1$ then the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ diverges.*

Proof For $\alpha < 1$, define $\beta = \frac{1}{2}(\alpha + 1)$. Then $\alpha < \beta < 1$. Now take $N \in \mathbb{Z}_{>0}$ such that

$$\left|\left|\frac{x_{j+1}}{x_j}\right| - \alpha\right| < \tfrac{1}{2}(1 - \alpha), \qquad j > N.$$

This implies that

$$\left|\frac{x_{j+1}}{x_j}\right| < \beta.$$

Now, for $j > N$,

$$|x_j| < \beta|x_{j-1}| < \beta^2|x_{j-1}| < \cdots < \beta^{j-N}|x_N|.$$

Clearly the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to 0 if and only if the sequence obtained by replacing the first N terms by 0 also converges to 0. If this latter sequence is denoted by $(y_j)_{j\in\mathbb{Z}_{>0}}$, then we have

$$0 \leq y_j \leq \frac{|x_N|}{\beta^N}\beta^j.$$

The sequence $(\frac{|x_N|}{\beta^N}\beta^j)_{j\in\mathbb{Z}_{>0}}$ converges to 0 since $\beta < 1$, and so this part of the result follows from the Squeezing Principle.

For $\alpha > 1$, there exists $N \in \mathbb{Z}_{>0}$ such that, for all $j \geq N$, $x_j \neq 0$. Consider the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ which is 0 for the first N terms, and satisfies $y_j = x_j^{-1}$ for the remaining terms. We then have $\left|\frac{y_{j+1}}{y_j}\right| < \alpha^{-1} < 1$, and so, from the first part of the proof, the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ converges to 0. Thus the sequence $(|y_j|)_{j\in\mathbb{Z}_{>0}}$ converges to ∞, which prohibits the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ from converging. ■

In Exercise 2.3.3 the reader can explore the various possibilities for the ratio test when $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$.

2.3.4 lim sup and lim inf

Recall from Section 2.2.6 the notions of sup and inf for subsets of $\mathbb{R}$. Associated with the least upper bound and greatest lower bound properties of $\mathbb{R}$ is a useful notion that weakens the usual idea of convergence. In order for us to make a sensible definition, we first prove a simple result.

2.3.12 Proposition (Existence of lim sup and lim inf) *For any sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* $\mathbb{R}$*, the limits*

$$\lim_{N\to\infty}\Big(\sup\{x_j \mid j \geq N\}\Big), \quad \lim_{N\to\infty}\Big(\inf\{x_j \mid j \geq N\}\Big)$$

exist, diverge to ∞*, or diverge to* $-\infty$*.*

Proof Note that the sequences $(\sup\{x_j \mid j \geq N\})_{N\in\mathbb{Z}_{>0}}$ and $(\inf\{x_j \mid j \geq N\})_{N\in\mathbb{Z}_{>0}}$ in $\overline{\mathbb{R}}$ are monotonically decreasing and monotonically increasing, respectively, with respect to the natural order on $\overline{\mathbb{R}}$. Moreover, note that a monotonically increasing sequence in $\overline{\mathbb{R}}$ is either bounded by some element of $\mathbb{R}$, or it is not. If the sequence is upper bounded by some element of $\mathbb{R}$, then by Theorem 2.3.8 it either converges or is the sequence $(-\infty)_{j\in\mathbb{Z}_{>0}}$. If it is not bounded by some element in $\mathbb{R}$, then either it diverges to ∞, or it is the sequence $(\infty)_{j\in\mathbb{Z}_{>0}}$ (this second case cannot arise in the specific case of the monotonically increasing sequence $(\sup\{x_j \mid j \geq N\})_{N\in\mathbb{Z}_{>0}}$. In all cases, the limit $\lim_{N\to\infty}\Big(\sup\{x_j \mid j \geq N\}\Big)$ exists or diverges to ∞. A similar argument for holds for $\lim_{N\to\infty}\Big(\inf\{x_j \mid j \geq N\}\Big)$. ■

2.3.13 Definition (lim sup and lim inf) For a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in $\mathbb{R}$ denote

$$\limsup_{j\to\infty} x_j = \lim_{N\to\infty}\Big(\sup\{x_j \mid j \geq N\}\Big),$$
$$\liminf_{j\to\infty} x_j = \lim_{N\to\infty}\Big(\inf\{x_j \mid j \geq N\}\Big).$$ •

Before we get to characterising lim sup and lim inf, we give some examples to illustrate all the cases that can arise.

2.3.14 Examples (lim sup and lim inf)

1. Consider the sequence $(x_j = (-1)^j)_{j\in\mathbb{Z}_{>0}}$. Here we have $\limsup_{j\to\infty} x_j = 1$ and $\liminf_{j\to\infty} x_j = -1$.
2. Consider the sequence $(x_j = j)_{j\in\mathbb{Z}_{>0}}$. Here $\limsup_{j\to\infty} x_j = \liminf_{j\to\infty} = \infty$.
3. Consider the sequence $(x_j = -j)_{j\in\mathbb{Z}_{>0}}$. Here $\limsup_{j\to\infty} x_j = \liminf_{j\to\infty} = -\infty$.
4. Define
$$x_j = \begin{cases} j, & j \text{ even}, \\ 0, & j \text{ odd}. \end{cases}$$
We then have $\limsup_{j\to\infty} x_j = \infty$ and $\liminf_{j\to\infty} x_j = 0$.
5. Define
$$x_j = \begin{cases} -j, & j \text{ even}, \\ 0, & j \text{ odd}. \end{cases}$$
We then have $\limsup_{j\to\infty} x_j = 0$ and $\liminf_{j\to\infty} = -\infty$.
6. Define
$$x_j = \begin{cases} j, & j \text{ even}, \\ -j, & j \text{ odd}. \end{cases}$$
We then have $\limsup_{j\to\infty} x_j = \infty$ and $\liminf_{j\to\infty} = -\infty$. •

There are many ways to characterise lim sup and lim inf, and we shall indicate but a few of these.

2.3.15 Proposition (Characterisation of lim sup) *For a sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* $\mathbb{R}$ *and* $\alpha \in \mathbb{R}$, *the following statements are equivalent:*

(i) $\alpha = \limsup_{j\to\infty} x_j$;

(ii) $\alpha = \inf\{\sup\{x_j \mid j \geq k\} \mid k \in \mathbb{Z}_{>0}\}$;

(iii) for each $\epsilon \in \mathbb{R}_{>0}$ *the following statements hold:*

(a) there exists $N \in \mathbb{Z}_{>0}$ *such that* $x_j < \alpha + \epsilon$ *for all* $j \geq N$;

(b) for an infinite number of $j \in \mathbb{Z}_{>0}$ *it holds that* $x_j > \alpha - \epsilon$.

Proof (i) $\iff$ (ii) Let $y_k = \sup\{x_j \mid j \geq k\}$ and note that the sequence $(y_k)_{k\in\mathbb{Z}_{>0}}$ is monotonically decreasing. Therefore, the sequence $(y_k)_{k\in\mathbb{Z}_{>0}}$ converges if and only if it is lower bounded. Moreover, if it converges, it converges to $\inf(y_k)_{k\in\mathbb{Z}_{>0}}$. Putting this all together gives the desired implications.

(i) $\implies$ (iii) Let y_k be as in the preceding part of the proof. Since $\lim_{k\to\infty} y_k = \alpha$, for each $\epsilon \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that $|y_k - \alpha| < \epsilon$ for $k \geq N$. In particular, $y_N < \alpha + \epsilon$. Therefore, $x_j < \alpha + \epsilon$ for all $j \geq N$, so (iii a) holds. We also claim that, for every $\epsilon \in \mathbb{R}_{>0}$ and for every $N \in \mathbb{Z}_{>0}$, there exists $j \geq N$ such that $x_j > y_N - \epsilon$. Indeed, if $x_j \leq y_N - \epsilon$ for every $j \geq N$, then this contradicts the definition of y_N. Since $y_N \geq \alpha$ we have $x_j > y_N - \epsilon \geq \alpha - \epsilon$ for some j. Since N is arbitrary, (iii b) holds.

(iii) $\implies$ (i) Condition (iii a) means that there exists $N \in \mathbb{Z}_{>0}$ such that $y_k < \alpha + \epsilon$ for all $k \geq N$. Condition (iii b) implies that $y_k > \alpha - \epsilon$ for all $k \in \mathbb{Z}_{>0}$. Combining these conclusions shows that $\lim_{k\to\infty} y_k = \alpha$, as desired. ■

The corresponding result for lim inf is the following. The proof follows in the same manner as the result for lim sup.

2.3.16 Proposition (Characterisation of lim inf) *For a sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* $\mathbb{R}$ *and* $\alpha \in \mathbb{R}$, *the following statements are equivalent:*

(i) $\alpha = \liminf_{j\to\infty} x_j$;

(ii) $\alpha = \sup\{\inf\{x_j \mid j \geq k\} \mid k \in \mathbb{Z}_{>0}\}$;

(iii) for each $\epsilon \in \mathbb{R}_{>0}$ *the following statements hold:*

(a) there exists $N \in \mathbb{Z}_{>0}$ *such that* $x_j > \alpha - \epsilon$ *for all* $j \geq N$;

(b) for an infinite number of $j \in \mathbb{Z}_{>0}$ *it holds that* $x_j < \alpha + \epsilon$.

Finally, we characterise the relationship between lim sup, lim inf, and lim.

2.3.17 Proposition (Relationship between lim sup, lim inf, and lim) *For a sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *and* $s_0 \in \mathbb{R}$, *the following statements are equivalent:*

(i) $\lim_{j\to\infty} x_j = s_0$;

(ii) $\limsup_{j\to\infty} x_j = \liminf_{j\to\infty} x_j = s_0$.

Proof (i) $\implies$ (ii) Let $\epsilon \in \mathbb{R}_{>0}$ and take $N \in \mathbb{Z}_{>0}$ such that $|x_j - s_0| < \epsilon$ for all $j \geq N$. Then $x_j < s_0 + \epsilon$ and $x_j > s_0 - \epsilon$ for all $j \geq N$. The current implication now follows from Propositions 2.3.15 and 2.3.16.

(ii) $\implies$ (i) Let $\epsilon \in \mathbb{R}_{>0}$. By Propositions 2.3.15 and 2.3.16 there exists $N_1, N_2 \in \mathbb{Z}_{>0}$ such that $x_j - s_0 < \epsilon$ for $j \geq N_1$ and $s_0 - x_j < \epsilon$ for $j \geq N_2$. Thus $|x_j - s_0| < \epsilon$ for $j \geq \max\{N_1, N_2\}$, giving this implication. ■

2.3.5 Multiple sequences

It will be sometimes useful for us to be able to consider sequences indexed, not by a single index, but by multiple indices. We consider the case here of two indices, and extensions to more indices are done by induction.

2.3.18 Definition (Double sequence) A ***double sequence*** in $\mathbb{R}$ is a family of elements of $\mathbb{R}$ indexed by $\mathbb{Z}_{>0} \times \mathbb{Z}_{>0}$. We denote a double sequence by $(x_{jk})_{j,k\in\mathbb{Z}_{>0}}$, where x_{jk} is the image of $(j,k) \in \mathbb{Z}_{>0} \times \mathbb{Z}_{>0}$ in $\mathbb{R}$. •

It is not *a priori* obvious what it might mean for a double sequence to converge, so we should carefully say what this means.

2.3.19 Definition (Convergence of double sequences) Let $s_0 \in \mathbb{R}$. A double sequence $(x_{jk})_{j,k\in\mathbb{Z}_{>0}}$:

(i) ***converges to*** $\mathbf{s_0}$, and we write $\lim_{j,k\to\infty} x_{jk} = s_0$, if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|s_0 - x_{jk}| < \epsilon$ for $j,k \geq N$;

(ii) has s_0 as a ***limit*** if it converges to s_0.

(iii) is ***convergent*** if it converges to some member of $\mathbb{R}$;

(iv) ***diverges*** if it does not converge;

(v) ***diverges to*** $\boldsymbol{\infty}$ (resp. ***diverges to*** $\boldsymbol{-\infty}$), and we write $\lim_{j,k\to\infty} x_{jk} = \infty$ (resp. $\lim_{j,k\to\infty} x_{jk} = -\infty$) if, for each $M \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $x_{jk} > M$ (resp. $x_{jk} < -M$) for $j,k \geq N$;

(vi) has a limit that ***exists*** if $\lim_{j,k\to\infty} x_{jk} \in \mathbb{R}$;

(vii) is ***oscillatory*** if the limit of the sequence does not exist, does not diverge to ∞, or does not diverge to $-\infty$. •

Note that the definition of convergence requires that one check both indices at the same time. Indeed, if one thinks, as it is useful to do, of a double sequence as assigning a real number to each point in an infinite grid defined by the set $\mathbb{Z}_{>0} \times \mathbb{Z}_{>0}$, convergence means that the values on the grid can be made arbitrarily small outside a sufficiently large square (see Figure 2.2). It is useful, however, to have means of computing limits of double sequences by computing limits of sequences in the usual sense. Our next results are devoted to this.

2.3.20 Proposition (Computation of limits of double sequences I) *Suppose that for the double sequence* $(\mathrm{x_{jk}})_{\mathrm{j,k}\in\mathbb{Z}_{>0}}$ *it holds that*

(i) the double sequence is convergent and

(ii) for each $\mathrm{j} \in \mathbb{Z}_{>0}$*, the limit* $\lim_{\mathrm{k}\to\infty} \mathrm{x_{jk}}$ *exists.*

Then the limit $\lim_{\mathrm{j}\to\infty}(\lim_{\mathrm{k}\to\infty} \mathrm{x_{jk}})$ *exists and is equal to* $\lim_{\mathrm{j,k}\to\infty} \mathrm{x_{jk}}$.

Proof Let $s_0 = \lim_{j,k\to\infty} x_{jk}$ and denote $s_j = \lim_{k\to\infty} x_{jk}$, $j \in \mathbb{Z}_{>0}$. For $\epsilon \in \mathbb{R}_{>0}$ take $N \in \mathbb{Z}_{>0}$ such that $|x_{jk} - s_0| < \frac{\epsilon}{2}$ for $j,k \geq N$. Also take $N_j \in \mathbb{Z}_{>0}$ such that $|x_{jk} - s_j| < \frac{\epsilon}{2}$ for $k \geq N_j$. Next take $j \geq N$ and let $k \geq \max\{N, N_j\}$. We then have

$$|s_j - s_0| = |s_j - x_{jk} + x_{jk} - s_0| \leq |s_j - x_{jk}| + |x_{jk} - s_0| < \epsilon,$$

using the triangle inequality. ■

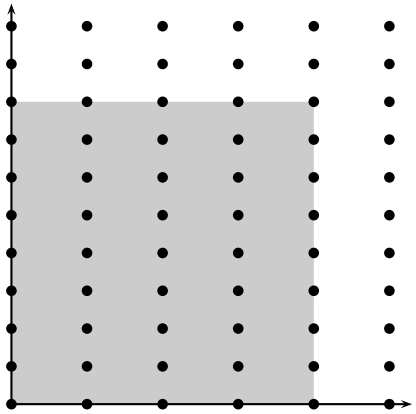


Figure 2.2 Convergence of a double sequence: by choosing the square large enough, the values at the unshaded grid points can be arbitrarily close to the limit

2.3.21 Proposition (Computation of limits of double sequences II) *Suppose that for the double sequence* $(x_{jk})_{j,k\in\mathbb{Z}_{>0}}$ *it holds that*

(i) the double sequence is convergent,

(ii) for each $j \in \mathbb{Z}_{>0}$*, the limit* $\lim_{k\to\infty} x_{jk}$ *exists, and*

(iii) for each $k \in \mathbb{Z}_{>0}$*, the limit* $\lim_{j\to\infty} x_{jk}$ *exists.*

Then the limits $\lim_{j\to\infty}(\lim_{k\to\infty} x_{jk})$ *and* $\lim_{k\to\infty}(\lim_{j\to\infty} x_{jk})$ *exist and are equal to* $\lim_{j,k\to\infty} x_{jk}$.

Proof This follows from two applications of Proposition 2.3.20. ■

Let us give some examples that illustrate the idea of convergence of a double sequence.

2.3.22 Examples (Double sequences)

1. It is easy to check that the double sequence $(\frac{1}{j+k})_{j,k\in\mathbb{Z}_{>0}}$ converges to 0. Indeed, for $\epsilon \in \mathbb{R}_{>0}$, if we take $N \in \mathbb{Z}_{>0}$ such that $\frac{1}{2N} < \epsilon$, it follows that $\frac{1}{j+k} < \epsilon$ for $j, k \geq N$.
2. The double sequence $(\frac{j}{j+k})_{j,k\in\mathbb{Z}_{>0}}$ does not converge. To see this we should find $\epsilon \in \mathbb{R}_{>0}$such that, for any $N \in \mathbb{Z}_{>0}$, there exists $j, k \geq N$ for which $\frac{j}{j+k} \geq \epsilon$. Take $\epsilon = \frac{1}{2}$ and let $N \in \mathbb{Z}_{>0}$. Then, if $j, k \geq N$ satisfy $j \geq 2k$, we have $\frac{j}{j+k} \geq \epsilon$.

   Note that for this sequence, the limits $\lim_{j\to\infty} \frac{j}{j+k}$ and $\lim_{k\to\infty} \frac{j}{j+k}$ exist for each fixed k and j, respectively. This cautions about trying to use these limits to infer convergence of the double sequence.
3. The double sequence $(\frac{(-1)^j}{k})_{j,k\in\mathbb{Z}_{>0}}$ is easily seen to converge to 0. However, the limit $\lim_{j\to\infty} \frac{(-1)^j}{k}$ does not exist for any fixed k. Therefore, one needs condition (ii) in Proposition 2.3.20 and conditions (ii) and (iii) in Proposition 2.3.21 in order for the results to be valid. •

2.3.6 Algebraic operations on sequences

It is of frequent interest to add, multiply, or divide sequences and series. In such cases, one would like to ensure that convergence of the sequences or series is sufficient to ensure convergence of the sum, product, or quotient. In this section we address this matter.

2.3.23 Proposition (Algebraic operations on sequences) *Let* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(y_j)_{j\in\mathbb{Z}_{>0}}$ *be sequences converging to* s_0 *and* t_0*, respectively, and let* $\alpha \in \mathbb{R}$*. Then the following statements hold:*

(i) the sequence $(\alpha x_j)_{j\in\mathbb{Z}_{>0}}$ *converges to* αs_0*;*

(ii) the sequence $(x_j + y_j)_{j\in\mathbb{Z}_{>0}}$ *converges to* $s_0 + t_0$*;*

(iii) the sequence $(x_j y_j)_{j\in\mathbb{Z}_{>0}}$ *converges to* $s_0 t_0$*;*

(iv) if, for all $j \in \mathbb{Z}_{>0}$, $y_j \neq 0$ *and if* $s_0 \neq 0$*, then the sequence* $(\frac{x_j}{y_j})_{j\in\mathbb{Z}_{>0}}$ *converges to* $\frac{s_0}{t_0}$*.*

Proof (i) The result is trivially true for $a = 0$, so let us suppose that $a \neq 0$. Let $\epsilon \in \mathbb{R}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|x_j - s_0| < \frac{\epsilon}{|\alpha|}$. Then, for $j \geq N$,

$$|\alpha x_j - \alpha s_0| = |\alpha||x_j - s_0| < \epsilon.$$

(ii) Let $\epsilon \in \mathbb{R}_{>0}$ and take $N_1, N_2 \in \mathbb{Z}_{>0}$ such that

$$|x_j - s_0| < \tfrac{\epsilon}{2}, \quad j \geq N_1, \qquad |y_j - t_0| < \tfrac{\epsilon}{2}, \quad j \geq N_2.$$

Then, for $j \geq \max\{N_1, N_2\}$,

$$|x_j + y_j - (s_0 + t_0)| \leq |x_j - s_0| + |y_j - t_0| = \epsilon,$$

using the triangle inequality.

(iii) Let $\epsilon \in \mathbb{R}_{>0}$ and define $N_1, N_2, N_3 \in \mathbb{Z}_{>0}$ such that

$$\begin{aligned}
&|x_j - s_0| < 1, \qquad j \geq N_1, \quad \Longrightarrow \quad |x_j| < |s_0| + 1, \qquad j \geq N_1,\\
&|x_j - s_0| < \frac{\epsilon}{2(|t_0| + 1)}, \qquad j \geq N_2,\\
&|y_j - t_0| < \frac{\epsilon}{2(|s_0| + 1)}, \qquad j \geq N_2.
\end{aligned}$$

Then, for $j \geq \max\{N_1, N_2, N_3\}$,

$$\begin{aligned}
|x_j y_j - s_0 t_0| &= |x_j y_j - x_j t_0 + x_j t_0 - s_0 t_0|\\
&= |x_j(y_j - t_0) + t_0(x_j - s_0)|\\
&\leq |x_j||y_j - t_0| + |t_0||x_j - s_0|\\
&\leq (|s_0| + 1)\frac{\epsilon}{2(|s_0| + 1)} + (|t_0| + 1)\frac{\epsilon}{2(|t_0| + 1)} = \epsilon.
\end{aligned}$$

(iv) It suffices using part (iii) to consider the case where $x_j = 1$, $j \in \mathbb{Z}_{>0}$. For $\epsilon \in \mathbb{R}_{>0}$ take $N_1.N_2 \in \mathbb{Z}_{>0}$ such that

$$\begin{aligned}
&|y_j - t_0| < \frac{|t_0|}{2}, \qquad j \geq N_1, \quad \Longrightarrow \quad |y_j| > \frac{|t_0|}{2}, \qquad j \geq N_1,\\
&|y_j - t_0| < \frac{|t_0|^2\epsilon}{2}, \qquad j \geq N_2.
\end{aligned}$$

Then, for $j \geq \max\{N_1, N_2\}$,

$$\left|\frac{1}{y_j} - \frac{1}{t_0}\right| = \left|\frac{y_j - t_0}{y_j t_0}\right| \leq \frac{|t_0|^2\epsilon}{2}\frac{2}{|t_0|}\frac{1}{|t_0|} = \epsilon,$$

as desired. ■

As we saw in the statement of Proposition 2.2.1, the restriction in part (iv) that $y_j \neq 0$ for all $j \in \mathbb{Z}_{>0}$ is not a real restriction. The salient restriction is that the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ not converge to 0.

2.3.7 Convergence using $\mathbb{R}$-nets

Up to this point in this section we have talked about convergence of sequences. However, in practice it is often useful to take limits of more general objects where the index set is not $\mathbb{Z}_{>0}$, but a subset of $\mathbb{R}$. In Section 1.6.4 we introduced a generalisation of sequences called nets. In this section we consider particular cases of nets, called $\mathbb{R}$-nets, that arise commonly when dealing with real numbers and subsets of real numbers. These will be particularly useful when considering the relationships between limits and functions. As we shall see, this slightly more general notion of convergence can be reduced to standard convergence of sequences. We comment that the notions of convergence in this section can be generalised to general nets, and we refer the reader to ***missing stuff*** for details.

Our objective is to understand what is meant by an expression like $\lim_{x\to a} \phi(a)$, where $\phi\colon A \to \mathbb{R}$ is a map from a subset A of $\mathbb{R}$ to $\mathbb{R}$. We will mainly be interested in subsets A of a rather specific form. However, we consider the general case so as to cover all situations that might arise.

2.3.24 Definition ($\mathbb{R}$-directed set) A $\mathbb{R}$-***directed set*** is a pair $D = (A, \preceq)$ where the partial order $\preceq$ is defined by $x \preceq y$ if either

(i) $x \leq y$,

(ii) $x \geq y$, or

(iii) there exists $x_0 \in \mathbb{R}$ such that $|x - x_0| \leq |y - x_0|$ (we abbreviate this relation as $x \preceq_{x_0} y$). •

Note that if $D = (A, \preceq)$ is a $\mathbb{R}$-directed set, then it is indeed a directed set because, corresponding to the three cases of the definition,

1. if $x, y \in A$, then $z = \max\{x, y\}$ has the property that $x \preceq z$ and $y \preceq z$ (for the first case in the definition),
2. if $x, y \in A$, then $z = \min\{x, y\}$ has the property that $x \preceq z$ and $y \preceq z$ (for the second case in the definition), or
3. if $x, y \in A$ then, taking z to satisfy $|z - x_0| = \min\{|x - x_0|, |y - x_0|\}$, we have $x \preceq z$ and $y \preceq z$ (for the third case of the definition).

Let us give some examples to illustrate the sort of phenomenon one can see for $\mathbb{R}$-directed sets.

2.3.25 Examples ($\mathbb{R}$-directed sets)

1. Let us take the $\mathbb{R}$-directed set $([0,1],\leq)$. Here we see that, for any $x, y \in [0,1]$, we have $x \leq 1$ and $y \leq 1$.
2. Next take the $\mathbb{R}$-directed set $([0,1),\leq)$. Here, there is no element z of $[0,1)$ for which $x \leq z$ and $y \leq z$ for every $x, y \in [0,1)$. However, it obviously holds that $x \leq 1$ and $y \leq 1$ for every $x, y \in [0,1)$.
3. Next we consider the $\mathbb{R}$ directed set $([0,\infty),\geq)$. Here we see that, for any $x, y \in [0,\infty)$, $x \geq 0$ and $y \geq 0$.
4. Next we consider the $\mathbb{R}$ directed set $((0,\infty),\geq)$. Here we see that there is no element $z \in (0,\infty)$ such that, for every $x, y \in (0,\infty)$, $x \geq z$ and $y \geq z$. However, it is true that $x \geq 0$ and $y \geq 0$ for every $x, y \in (0,\infty)$.
5. Now we take the $\mathbb{R}$-directed set $([0,\infty),\leq)$. Here we see that there is no element $z \in [0,\infty)$ such that $x \leq z$ and $y \leq z$ for every $x, y \in [0,\infty)$. Moreover, there is also no element $z \in \mathbb{R}$ for which $x \leq z$ and $y \leq z$ for every $x, y \in [0,\infty)$.
6. Next we take the $\mathbb{R}$-directed set $(\mathbb{Z},\leq)$. As in the preceding example, there is no element $z \in [0,\infty)$ such that $x \leq z$ and $y \leq z$ for every $x, y \in [0,\infty)$. Moreover, there is also no element $z \in \mathbb{R}$ for which $x \leq z$ and $y \leq z$ for every $x, y \in [0,\infty)$.
7. Now consider the $\mathbb{R}$-directed set $(\mathbb{R},\leq_0)$. Note that $0 \in \mathbb{R}$ has the property that, for any $x, y \in \mathbb{R}$, $x \leq_0 0$ and $y \leq_0 0$.
8. Similar to the preceding example, consider the $\mathbb{R}$-directed set $(\mathbb{R}\setminus\{0\},\leq_0)$. Here there is no element $z \in \mathbb{R}\setminus\{0\}$ such that $x \leq_0 z$ and $y \leq_0 z$ for every $x, y \in \mathbb{R}\setminus\{0\}$. However, we clearly have $x \leq_0 0$ and $y \leq_0 0$ for every $x, y \in \mathbb{R}\setminus\{0\}$. •

The examples may seem a little silly, but this is just because the notion of a $\mathbb{R}$-directed set is, in and of itself, not so interesting. What is more interesting is the following notion.

2.3.26 Definition ($\mathbb{R}$-net, convergence in $\mathbb{R}$-nets) If $D = (A,\leq)$ is a $\mathbb{R}$-directed set, a ***$\mathbb{R}$-net*** in D is a map $\phi\colon A \to \mathbb{R}$. A $\mathbb{R}$-net $\phi\colon A \to \mathbb{R}$ in a $\mathbb{R}$-directed set $D = (A,\leq)$

(i) ***converges*** to $s_0 \in \mathbb{R}$ if, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $x \in A$ such that $|\phi(y) - s_0| < \epsilon$ for any $y \in A$ satisfying $x \leq y$,

(ii) has s_0 as a ***limit*** if it converges to s_0, and we write $s_0 = \lim_D \phi$,

(iii) ***diverges*** if it does not converge,

(iv) ***diverges to ∞*** ((resp. ***diverges to $-\infty$***, and we write $\lim_D \phi = \infty$ (resp. $\lim_D \phi = -\infty$), if, for each $M \in \mathbb{R}_{>0}$, there exists $x \in A$ such that $\phi(y) > M$ (resp. $\phi(y) < -M$) for every $y \in A$ for which $x \leq y$,

(v) has a limit that ***exists*** if $\lim_D \phi \in \mathbb{R}$, and

(vi) is ***oscillatory*** if the limit of the $\mathbb{R}$-net does not exist, does not diverge to ∞, and does not diverge to $-\infty$. •

2.3.27 Notation (Limits of $\mathbb{R}$-nets) The importance $\mathbb{R}$-nets can now be illustrated by showing how they give rise to a collection of convergence phenomenon. Let us look at various cases for convergence of a $\mathbb{R}$-net in a $\mathbb{R}$-directed set $D = (A, \preceq)$.

(i) $\preceq = \leq$: Here there are two subcases to consider.
  (a) $\sup A = x_0 < \infty$: In this case we write $\lim_D \phi = \lim_{x \uparrow x_0} \phi(x)$.
  (b) $\sup A = \infty$: In this case we write $\lim_D \phi = \lim_{x \to \infty} \phi(x)$.

(ii) $\preceq = \geq$: Again we have two subcases.
  (a) $\inf A = x_0 > -\infty$: In this case we write $\lim_D \phi = \lim_{x \downarrow x_0} \phi(x)$.
  (b) $\inf A = -\infty$: In this case we write $\lim_D \phi = \lim_{x \to -\infty} \phi(x)$.

(iii) $\preceq = \leq_{x_0}$: There are three subcases here that we wish to distinguish.
  (a) $\sup A = x_0$: Here we denote $\lim_D \phi = \lim_{x \uparrow x_0} \phi(x)$.
  (b) $\inf A = x_0$: Here we denote $\lim_D \phi = \lim_{x \downarrow x_0} \phi(x)$.
  (c) $x_0 \notin \{\inf A, \sup A\}$: Here we denote $\lim_D \phi = \lim_{x \to x_0} \phi(x)$. •

In the case when the directed set is an interval, we have the following notation that unifies the various limit notations for this special often encountered case.

2.3.28 Notation (Limit in an interval) Let $I \subseteq \mathbb{R}$ be an interval, let $\phi \colon I \to \mathbb{R}$ be a map, and let $a \in I$. We define $\lim_{x \to_I a} \phi(x)$ by

(i) $\lim_{x \to_I a} \phi(x) = \lim_{x \uparrow a} \phi(x)$ if $a = \sup I$,
(ii) $\lim_{x \to_I a} \phi(x) = \lim_{x \downarrow a} \phi(x)$ if $a = \inf I$, and
(iii) $\lim_{x \to_I a} \phi(x) = \lim_{x \to a} \phi(x)$ otherwise. •

We expect that most readers will be familiar with the idea here, even if the notation is not conventional. Let us also give the notation a precise characterisation in terms of limits of sequences in the case when the point x_0 is in the closure of the set A.

2.3.29 Proposition (Convergence in $\mathbb{R}$-nets in terms of sequences) *Let $(\mathrm{A}, \preceq)$ be a $\mathbb{R}$-directed set and let $\phi \colon \mathrm{A} \to \mathbb{R}$ be a $\mathbb{R}$-net in $(\mathrm{A}, \preceq)$. Then, corresponding to the cases and subcases of Notation 2.3.27, we have the following statements:*

(i) (a) if $\mathrm{x}_0 \in \mathrm{cl}(\mathrm{A})$, the following statements are equivalent:
  I. $\lim_{\mathrm{x} \uparrow \mathrm{x}_0} \phi(\mathrm{x}) = \mathrm{s}_0$;
  II. $\lim_{\mathrm{j} \to \infty} \phi(\mathrm{x}_\mathrm{j}) = \mathrm{s}_0$ *for every sequence* $(\mathrm{x}_\mathrm{j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{\mathrm{j} \to \infty} \mathrm{x}_\mathrm{j} = \mathrm{x}_0$;

(b) the following statements are equivalent:
  I. $\lim_{\mathrm{x} \to \infty} \phi(\mathrm{x}) = \mathrm{s}_0$;
  II. $\lim_{\mathrm{j} \to \infty} \phi(\mathrm{x}_\mathrm{j}) = \mathrm{s}_0$ *for every sequence* $(\mathrm{x}_\mathrm{j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{\mathrm{j} \to \infty} \mathrm{x}_\mathrm{j} = \infty$;

(ii) (a) if $\mathrm{x}_0 \in \mathrm{cl}(\mathrm{A})$, the following statements are equivalent:
  I. $\lim_{\mathrm{x} \downarrow \mathrm{x}_0} \phi(\mathrm{x}) = \mathrm{s}_0$;

II. $\lim_{j\to\infty} \phi(x_j) = s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{j\to\infty} x_j = x_0$*;*

(b) *the following statements are equivalent:*

I. $\lim_{x\to-\infty} \phi(x) = s_0$*;*

II. $\lim_{j\to\infty} \phi(x_j) = s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{j\to\infty} x_j = -\infty$*;*

(iii) (a) *if* $x_0 \in \mathrm{cl}(A)$*, the following statements are equivalent:*

I. $\lim_{x\uparrow x_0} \phi(x) = s_0$*;*

II. $\lim_{j\to\infty} \phi(x_j) = s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{j\to\infty} x_j = x_0$*;*

(b) *if* $x_0 \in \mathrm{cl}(A)$*, the following statements are equivalent:*

I. $\lim_{x\downarrow x_0} \phi(x) = s_0$*;*

II. $\lim_{j\to\infty} \phi(x_j) = s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{j\to\infty} x_j = x_0$*;*

(c) *the following statements are equivalent:*

I. $\lim_{x\to\infty} \phi(x) = s_0$*;*

II. $\lim_{j\to\infty} \phi(x_j) = s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{j\to\infty} x_j = \infty$*;*

Proof These statements are all proved in essentially the same way, so let us prove just, say, part (i a).

First suppose that $\lim_{x\uparrow x_0} \phi(x) = s_0$, and let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in A converging to x_0. Let $\epsilon \in \mathbb{R}_{>0}$ and choose $x \in A$ such that $|\phi(y) - s_0| < \epsilon$ whenever $y \in A$ satisfies $x \leq y$. Then, since $\lim_{j\to\infty} x_j = x_0$, there exists $N \in \mathbb{Z}_{>0}$ such that $x \leq x_j$ for all $j \geq N$. Clearly, $|\phi(x_j) - s_0| < \epsilon$, so giving convergence of $(\phi(x_j))_{j\in\mathbb{Z}_{>0}}$ to s_0 for every sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in A converging to x_0.

For the converse, suppose that $\lim_{x\uparrow x_0} \phi(x) \neq s_0$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that, for any $x \in A$, we have a $y \in A$ with $x \leq y$ for which $|\phi(y) - s_0| \geq \epsilon$. Since $x_0 \in \mathrm{cl}(A)$ it follows that, for any $j \in \mathbb{Z}_{>0}$, there exists $x_j \in \mathsf{B}(\frac{1}{j}, x_0) \cap A$ such that $|\phi(x_j) - s_0| \geq \epsilon$. Thus the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in A converging to x_0 has the property that $(\phi(x_j))_{j\in\mathbb{Z}_{>0}}$ does not converge to s_0. ■

Of course, similar conclusions hold when "convergence to s_0" is replaced with "divergence," "convergence to ∞," "convergence to $-\infty$," or "oscillatory." We leave the precise statements to the reader.

Let us give some examples to illustrate that this is all really nothing new.

2.3.30 Examples (Convergence in $\mathbb{R}$-nets)

1. Consider the $\mathbb{R}$-directed set $([0,\infty), \leq)$ and the corresponding $\mathbb{R}$-net ϕ defined by $\phi(x) = \frac{1}{1+x^2}$. This $\mathbb{R}$-net then converges to 0. Let us verify this using the formal definition of convergence of a $\mathbb{R}$-net. For $\epsilon \in \mathbb{R}_{>0}$ choose $x > 0$ such that $x^2 = \frac{1}{\epsilon} > \frac{1}{\epsilon} - 1$. Then, if $x \leq y$, we have

$$\left|\frac{1}{1+y^2} - 0\right| < \frac{1}{1+x^2} < \epsilon,$$

giving convergence to $\lim_{x\to\infty}\phi(x) = 0$ as stated.

2. Next consider the $\mathbb{R}$-directed set $((0,1],\geq)$ and the corresponding $\mathbb{R}$-net ϕ defined by $\phi(x) = x\sin\frac{1}{x}$. We claim that this $\mathbb{R}$-net converges to 0. To see this, let $\epsilon\in\mathbb{R}_{>0}$ and let $x\in(0,\epsilon)$. Then we have, for $x\geq y$,

$$\left|y\sin\tfrac{1}{y} - 0\right| = y \leq x < \epsilon,$$

giving $\lim_{x\downarrow 0}\phi(x) = 0$ as desired.

3. Consider the $\mathbb{R}$-directed set $([0,\infty),\leq)$ and the associated $\mathbb{R}$-net ϕ defined by $\phi(x) = x$. In this case we have $\lim_{x\to\infty}\phi(x) = \infty$.
4. Consider the $\mathbb{R}$-directed set $([0,\infty),\leq)$ and the associated $\mathbb{R}$-net ϕ defined by $\phi(x) = x\sin x$. In this case, due to the oscillatory nature of sin, $\lim_{x\to\infty}\phi(x)$ does not exist, nor does it diverge to either ∞ or $-\infty$.
5. Take the $\mathbb{R}$-directed set $(\mathbb{R}\setminus\{0\},\leq_0)$. Define the $\mathbb{R}$-net ϕ by $\phi(x) = x$. Clearly, $\lim_{x\to 0}\phi(x) = 0$. •

There are also generalisations of lim sup and lim inf to $\mathbb{R}$-nets. We let $D = (A,\leq)$ be a $\mathbb{R}$-directed set and let $\phi\colon A\to\mathbb{R}$ be a $\mathbb{R}$-net in this $\mathbb{R}$-directed set. We denote by $\sup_D\phi, \inf_D\phi\colon A\to\mathbb{R}$ the $\mathbb{R}$-nets in D given by

$$\sup_D\phi(x) = \sup\{\phi(y)\mid x\leq y\},\quad \inf_D\phi(x) = \inf\{\phi(y)\mid x\leq y\}.$$

Then we define

$$\limsup_D\phi = \lim_D\sup_D\phi,\quad \liminf_D\phi = \lim_D\inf_D\phi.$$

These allow us to talk of limits in cases where limits in the usual sense to not exist. Let us consider this via an example.

2.3.31 Example (lim sup and lim inf in $\mathbb{R}$-nets) We consider the $\mathbb{R}$-directed set $D = ([0,\infty),\leq)$ and let ϕ be the $\mathbb{R}$-net defined by $\phi(x) = \mathrm{e}^{-x} + \sin x$.[5] We claim that $\limsup_D\phi = 1$ and that $\liminf_D\phi = -1$. Let us prove the first claim, and leave the second as an exercise. We then have

$$\sup_D\phi(x) = \sup\{\mathrm{e}^{-y} + \sin y\mid x\leq y\} = \mathrm{e}^{-x} + 1.$$

First note that $\sup_D\phi(x)\geq 1$ for every $x\in[0,\infty)$, and so $\limsup_D\phi\geq 1$. Now let $\epsilon\in\mathbb{R}_{>0}$ and take $x > \log\epsilon$. Then, for any $y\geq x$,

$$\sup_D\phi(y) = \mathrm{e}^{-y} + 1 \leq 1 + \epsilon.$$

Therefore, $\limsup_D\phi\leq 1$, and so $\limsup_D\phi = 1$, as desired. •

[5] We have not yet defined e^{-x} or $\sin x$. The reader who is unable to go on without knowing what these functions really are can skip ahead to Section 3.6.

2.3.8 A first glimpse of Landau symbols

In this section we introduce for the first time the so-called Landau symbols. These provide commonly used notation for when two functions behave "asymptotically" the same. Given our development of $\mathbb{R}$-nets in the preceding section, it is easy for us to be fairly precise here. We also warn the reader that the Landau symbols often get used in an imprecise or vague way. We shall try to avoid such usage.

We begin with the definition.

2.3.32 Definition (Landau symbols "O" and "o") Let $D = (A, \preceq)$ be a $\mathbb{R}$-directed set and let $\phi : A \to \mathbb{R}$.

(i) Denote by $O_D(\phi)$ the functions $\psi : A \to \mathbb{R}$ for which there exists $x_0 \in A$ and $M \in \mathbb{R}_{>0}$ such that $|\psi(x)| \leq M|\phi(x)|$ for $x \in A$ satisfying $x_0 \preceq x$.

(ii) Denote by $o_D(\phi)$ the functions $\psi : A \to \mathbb{R}$ such that, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $x_0 \in A$ such that $|\psi(x)| < \epsilon|\phi(x)|$ for $x \in A$ satisfying $x_0 \preceq x$.

If $\psi \in O_D(\phi)$ (resp. $\psi \in o_D(\phi)$) then we say that ψ is ***big oh of ϕ*** (resp. ***little oh of ϕ***). •

It is very common to see simply $O(\phi)$ and $o(\phi)$ in place of $O_D(\phi)$ and $o_D(\phi)$. This is because the most common situation for using this notation is in the case when $\sup A = \infty$ and $\preceq = \leq$. In such cases, the notation indicates means, essentially, that $\psi \in O(\phi)$ if ψ has "size" no larger than ϕ for large values of the argument and that $\psi \in o(\phi)$ if ψ is "small" compared to ϕ for large values of the argument. However, we shall use the Landau symbols in other cases, so we allow the possibility of explicitly including the $\mathbb{R}$-directed set in our notation for the sake of clarity.

It is often the case that the comparison function ϕ is positive on A. In such cases, one can give a somewhat more concrete characterisation of O_D and o_D.

2.3.33 Proposition (Alternative characterisation of Landau symbols) *Let $\mathrm{D} = (\mathrm{A}, \preceq)$ be a $\mathbb{R}$-directed set, and let $\phi : \mathrm{A} \to \mathbb{R}_{>0}$ and $\psi : \mathrm{A} \to \mathbb{R}$. Then*

(i) $\psi \in \mathrm{O}_\mathrm{D}(\phi)$ if and only if $\limsup_\mathrm{D} \frac{\psi}{\phi} < \infty$ and

(ii) $\psi \in \mathrm{o}_\mathrm{D}(\phi)$ if and only if $\lim_\mathrm{D} \frac{\psi}{\phi} = 0$.

Proof We leave this as Exercise 2.3.6. ■

Let us give some common examples of where the Landau symbols are used. Some examples will make use of ideas we have not yet discussed, but which we imagine are familiar to most readers.

2.3.34 Examples (Landau symbols)

1. Let $I \subseteq \mathbb{R}$ be an interval for which $x_0 \in I$ and let $f : I \to \mathbb{R}$. Consider the $\mathbb{R}$-directed set $D = (I \setminus \{x_0\}, \preceq_{x_0})$ and the $\mathbb{R}$-net ϕ in D given by $\phi(x) = 1$. Define $g_{f,x_0} : I \to \mathbb{R}$ by $g_{f,x_0}(x) = f(x_0)$. We claim that f is continuous at x_0 if and only if

$f - g_{f,x_0} \in o_D(\phi)$. Indeed, by Theorem 3.1.3 we have that f is continuous at x_0 if and only if

$$\begin{aligned} & \lim_{x \to_I x_0} f(x) = f(x_0) \\ \implies \quad & \lim_{x \to_I x_0} (f(x) - g_{f,x_0}(x)) = 0 \\ \implies \quad & \lim_{x \to_I x_0} \frac{(f(x) - g_{f,x_0}(x))}{\phi(x)} = 0 \\ \implies \quad & f - g_{f,x_0} \in o_D(\phi). \end{aligned}$$

The idea is that f is continuous at x_0 if and only if f is "approximately constant" near x_0.

2. Let $I \subseteq \mathbb{R}$ be an interval for which $x_0 \in I$ and let $f\colon I \to \mathbb{R}$. For $L \in \mathbb{R}$ define $g_{f,x_0,L}\colon I \setminus \{x_0\} \to \mathbb{R}$ by

$$g_{x_0,L}(x) = f(x_0) + L(x - x_0).$$

Consider the $\mathbb{R}$-directed set $D = (I \setminus \{x_0\}, \leq_{x_0})$, and define $\phi\colon I \setminus \{x_0\} \to \mathbb{R}_{>0}$ by $\phi(x) = |x - x_0|$. Then we claim that f is differentiable at x_0 with derivative $f'(x_0) = L$ if and only if $f - g_{f,x_0,L} \in o_D(\phi)$. Indeed, by definition, f is differentiable at x_0 with derivative $f'(x_0) = L$ if and only if, then

$$\begin{aligned} & \lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0} = L \\ \iff \quad & \lim_{x \to_I x_0} \frac{1}{x - x_0}\Big(f(x) - g_{f,x_0,L}(x)\Big) = 0 \\ \iff \quad & \lim_{x \to_I x_0} \frac{1}{|x - x_0|}\Big(f(x) - g_{f,x_0,L}(x)\Big) = 0 \\ \iff \quad & f(x) - g_{f,x_0,L}(x) \in o_D(\phi), \end{aligned}$$

using Proposition 2.3.33. The idea is that f is differentiable at x_0 if and only if f is "nearly linear" at x_0.

3. We can generalise the preceding two examples. Let $I \subseteq \mathbb{R}$ be an interval, let $x_0 \in I$, and consider the $\mathbb{R}$-directed set $(I \setminus \{x_0\}, \leq_{x_0})$. For $m \in \mathbb{Z}_{\geq 0}$ define the $\mathbb{R}$-net ϕ_m in D by $\phi_m(x) = |x - x_0|^m$. We shall say that a function $f\colon I \to \mathbb{R}$ ***vanishes to order*** $\mathbf{m}$ ***at*** $\mathbf{x_0}$ if $f \in O_D(\phi_m)$. Moreover, f is m-times differentiable at x_0 with $f^{(j)}(x_0)alpha_j$, $j \in \{0, 1, \ldots, m\}$, if and only if $f - g_{f,x_0,\boldsymbol{\alpha}} \in o_D(\phi_m)$, where

$$g_{f,x_0,\boldsymbol{\alpha}}(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_m x^m.$$

4. One of the common places where Landau symbols are used is in the analysis of the complexity of algorithms. An algorithm, loosely speaking, takes some input data, performs operations on the data, and gives an outcome. A very simple example of an algorithm is the multiplication of two square matrices, and we will use this simple example to illustrate our discussion. It is assumed that the size of the input data is measured by an integer N. For example, for

the multiplication of square matrices, this integer is the size of the matrices. The complexity of an algorithm is then determined by the number of steps, denoted by, say, $\psi(N)$, of a certain type in the algorithm. For example, for the multiplication of square matrices, this number is normally taken to be the number of multiplications that are needed, and this is easily seen to be no more than N^2. To describe the complexity of the algorithm, one finds uses Landau symbols in the following way. First of all, we use the $\mathbb{R}$-directed set $D = (\mathbb{Z}_{>0}, \leq)$. If $\phi\colon \mathbb{Z}_{>0} \to \mathbb{R}_{>0}$ is such that $\psi \in O_D(\phi)$, then we say the algorithm ***is* O(ϕ)**. For example, matrix multiplication is $O(N^2)$.

In Theorem **??** we show that the computational complexity of the so-called Cooley–Tukey algorithm for computing the FFT is $O(N \log N)$.

Since we are talking about computational complexity of algorithms, it is a good time to make mention of an important problem in the theory of computational complexity. This discussion is limited to so-called decision algorithms, where the outcome is an affirmative or negative declaration about some problem, e.g., is the determinant of a matrix bounded by some number. For such an algorithm, a ***verification algorithm*** is an algorithm that checks whether given input data does indeed give an affirmative answer. Denote by ***P*** the class of algorithms that are $O(N^m)$ for some $m \in \mathbb{Z}_{>0}$. Such algorithms are known as ***polynomial time*** algorithms. Denote by ***NP*** the class of algorithms for which there exists a verification algorithm that is $O(N^m)$ for some $m \in \mathbb{Z}_{>0}$. An important unresolved question is, "Does P=NP?" •

2.3.9 Notes

Citation for Dedekind cuts.

Exercises

2.3.1 Show that if $(x_j)_{j\in\mathbb{Z}_{>0}}$ is a sequence in $\mathbb{R}$ and if $\lim_{j\to\infty} x_j = x_0$ and $\lim_{j\to\infty} x_j = x_0'$, then $x_0 = x_0'$.

2.3.2 Answer the following questions:

(a) find a subset $S \subseteq \mathbb{Q}$ that possesses an upper bound in $\mathbb{Q}$, but which has no least element;

(b) find a bounded monotonic sequence in $\mathbb{Q}$ that does not converge in $\mathbb{Q}$.

2.3.3 Do the following.

(a) Find a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which converges in $\mathbb{R}$.

(b) Find a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which diverges to ∞.

(c) Find a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which diverges to $-\infty$.

(d) Find a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which is oscillatory.

2.3.4 *missing stuff*

In the next exercise you will show that the property that a bounded, monotonically increasing sequence converges implies that Cauchy sequences converge. This completes the argument needed to prove the theorem stated in Aside 2.3.9 concerning characterisations of complete ordered fields.

2.3.5 Assume that every bounded, monotonically increasing sequence in $\mathbb{R}$ converges, and using this show that every Cauchy sequence in $\mathbb{R}$ converges using an argument as follows.

1. Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a Cauchy sequence.
2. Let $I_0 = [a, b]$ be an interval that contains all elements of $(x_j)_{j\in\mathbb{Z}_{>0}}$ (why is this possible?)
3. Split $[a, b]$ into two equal length closed intervals, and argue that in at least one of these there is an infinite number of points from the sequence. Call this interval I_1 and let $x_{k_i} \in (x_j)_{j\in\mathbb{Z}_{>0}} \cap I_1$.
4. Repeat the process for I_1 to find an interval I_2 which contains an infinite number of points from the sequence. Let $x_{k_2} \in (x_j)_{j\in\mathbb{Z}_{>0}} \cap I_2$.
5. Carry on doing this to arrive at a sequence $(x_{k_j})_{j\in\mathbb{Z}_{>0}}$ of points in $\mathbb{R}$ and a sequence $(I_j)_{j\in\mathbb{Z}_{>0}}$.
6. Argue that the sequence of left endpoints of the intervals $(I_j)_{j\in\mathbb{Z}_{>0}}$ is a bounded monotonically increasing sequence, and that the sequence of right endpoints is a bounded monotonically decreasing sequence. and so both converge.
7. Show that they converge to the same number, and that the sequence $(x_{k_j})_{j\in\mathbb{Z}_{>0}}$ also converges to this limit.
8. Show that the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to this limit.

2.3.6 Prove Proposition 2.3.33.

Section 2.4

Series in $\mathbb{R}$

From a sequence $(x_j)_{j\in\mathbb{R}}$ in $\mathbb{R}$, one can consider, in principle, the infinite sum $\sum_{j=1}^{\infty} x_j$. Of course, such a sum *a priori* makes no sense. However, as we shall see in Chapter **??**, such infinite sums are important for characterising certain discrete-time signal spaces. Moreover, such sums come up frequently in many places in analysis. In this section we outline some of the principle properties of these sums.

Do I need to read this section? Most readers will probably have seen much of the material in this section in their introductory calculus course. What might be new for some readers is the fairly careful discussion in Theorem 2.4.5 of the difference between convergence and absolute convergence of series. Since absolute convergence will be of importance to us, it might be worth understanding in what ways it is different from convergence. The material in Section 2.4.7 can be regarded as optional until it is needed during the course of reading other material in the text. •

2.4.1 Definitions and properties of series

A ***series*** in $\mathbb{R}$ is an expression of the form

$$S = \sum_{j=1}^{\infty} x_j, \tag{2.5}$$

where $x_j \in \mathbb{R}$, $j \in \mathbb{Z}_{>0}$. Of course, the problem with this "definition" is that the expression (2.5) is meaningless as an element of $\mathbb{R}$ unless it possesses additional features. For example, if $x_j = 1$, $j \in \mathbb{Z}_{>0}$, then the sum is infinite. Also, if $x_j = (-1)^j$, $j \in \mathbb{Z}_{>0}$, then it is not clear what the sum is: perhaps it is 0 or perhaps it is 1. Therefore, to be precise, a series is prescribed by the sequence of numbers $(x_j)_{j\in\mathbb{Z}_{>0}}$, and is represented in the form (2.5) in order to distinguish it from the sequence with the same terms.

If the expression (2.5) is to have meaning as a number, we need some sort of condition placed on the terms in the series.

2.4.1 Definition (Convergence and absolute convergence of series) Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}$ and consider the series

$$S = \sum_{j=1}^{\infty} x_j.$$

The corresponding sequence of ***partial sums*** is the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ defined by

$$S_k = \sum_{j=1}^{k} x_j.$$

Let $s_0 \in \mathbb{R}$. The series:

(i) ***converges to*** $\mathbf{s_0}$, and we write $\sum_{j=1}^{\infty} x_j = s_0$, if the sequence of partial sums converges to s_0;

(ii) has s_0 as a ***limit*** if it converges to s_0;

(iii) is ***convergent*** if it converges to some member of $\mathbb{R}$;

(iv) ***converges absolutely***, or is ***absolutely convergent***, if the series

$$\sum_{j=1}^{\infty} |x_j|$$

converges;

(v) ***converges conditionally***, or is ***conditionally convergent***, if it is convergent, but not absolutely convergent;

(vi) ***diverges*** if it does not converge;

(vii) ***diverges to*** ∞ (resp. ***diverges to*** $-\infty$), and we write $\sum_{j=1}^{\infty} x_j = \infty$ (resp. $\sum_{j=1}^{\infty} x_j = -\infty$), if the sequence of partial sums diverges to ∞ (resp. diverges to $-\infty$);

(viii) has a limit that ***exists*** if $\lim_{j\to\infty} S_j \in \mathbb{R}$;

(ix) is ***oscillatory*** if the sequence of partial sums is oscillatory. •

Let us consider some examples of series in $\mathbb{R}$.

2.4.2 Examples (Series in $\mathbb{R}$)

1. First we consider the ***geometric series*** $\sum_{j=1}^{\infty} x^{j-1}$ for $x \in \mathbb{R}$. We claim that this series converges if and only if $|x| < 1$. To prove this we claim that the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ of partial sums is defined by

$$S_k = \begin{cases} \frac{1-x^{k+1}}{1-x}, & x \neq 1, \\ k, & x = 1. \end{cases}$$

The conclusion is obvious for $x = 1$, so we can suppose that $x \neq 1$. The conclusion is obvious for $k = 1$, so suppose it true for $j \in \{1, \dots, k\}$. Then

$$S_{k+1} = \sum_{j=1}^{k+1} x^j = x^{k+1} + \frac{1-x^{k+1}}{1-x} = \frac{x^{k+1} - x^{k+2} + 1 - x^{k+1}}{1-x} = \frac{1-x^{k+2}}{1-x},$$

as desired. It is clear, then, that if $x = 1$ then the series diverges to ∞. If $x = -1$ then the series is directly checked to be oscillatory; the sequence of partial sums is $\{1, 0, 1, \dots\}$. For $x > 1$ we have

$$\lim_{k\to\infty} S_k = \lim_{k\to\infty} \frac{1-x^{k+1}}{1-x} = \infty,$$

showing that the series diverges to ∞ in this case. For $x < -1$ it is easy to see that the sequence of partial sums is oscillatory, but increasing in magnitude.

This leaves the case when $|x| < 1$. Here, since the sequence $(x^{k+1})_{k\in\mathbb{Z}_{>0}}$ converges to zero, the sequence of partial sums also converges, and converges to $\frac{1}{1-x}$. (We have used the results concerning the swapping of limits with algebraic operations as described in Section 2.3.6.)

2. We claim that the series $\sum_{j=1}^{\infty} \frac{1}{j}$ diverges to ∞. To show this, we show that the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ is not upper bounded. To show this, we shall show that $S_{2^k} \geq 1 + \frac{1}{2}k$ for all $k \in \mathbb{Z}_{>0}$. This is true directly when $k = 1$. Next suppose that $S_{2^j} \geq 1 + \frac{1}{2}j$ for $j \in \{1, \dots, k\}$. Then

$$\begin{aligned} S_{2^{k+1}} &= S_{2^k} + \frac{1}{2^k+1} + \frac{1}{2^k+2} + \cdots + \frac{1}{2^{k+1}} \\ &\geq 1 + \frac{1}{2}k + \underbrace{\frac{1}{2^{k+1}} + \cdots + \frac{1}{2^{k+1}}}_{2^k \text{ terms}} \\ &= 1 + \frac{1}{2}k + \frac{2^k}{2^{k+1}} = 1 + \frac{1}{2}(k+1). \end{aligned}$$

Thus the sequence of partial sums is indeed unbounded, and since it is monotonically increasing, it diverges to ∞, as we first claimed.

3. We claim that the series $S = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j}$ converges. To see this, we claim that, for any $m \in \mathbb{Z}_{>0}$, we have

$$S_2 \leq S_4 \leq \cdots \leq S_{2m} \leq S_{2m-1} \leq \cdots \leq S_3 \leq S_1.$$

That $S_2 \leq S_4 \leq \cdots \leq S_{2m}$ follows since $S_{2k} - S_{2k-2} = \frac{1}{2k-1} - \frac{1}{2k} > 0$ for $k \in \mathbb{Z}_{>0}$. That $S_{2m} \leq S_{2m-1}$ follows since $S_{2m-1} - S_{2m} = \frac{1}{2m}$. Finally, $S_{2m-1} \leq \cdots \leq S_3 \leq S_1$ since $S_{2k-1} - S_{2k+1} = \frac{1}{2k} - \frac{1}{2k+1} > 0$ for $k \in \mathbb{Z}_{>0}$. Thus the sequences $(S_{2k})_{k\in\mathbb{Z}_{>0}}$ and $(S_{2k-1})_{k\in\mathbb{Z}_{>0}}$ are monotonically increasing and monotonically decreasing, respectively, and their tails are getting closer and closer together since $\lim_{m\to\infty} S_{2m-1} - S_{2m} = \frac{1}{2m} = 0$. By Lemma 2 from the proof of Theorem 2.3.7, it follows that the sequences $(S_{2k})_{k\in\mathbb{Z}_{>0}}$ and $(S_{2k-1})_{k\in\mathbb{Z}_{>0}}$ converge and converge to the same limit. Therefore, the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ converges as well to the same limit. One can moreover show that the limit of the series is $\log 2$, where log denotes the natural logarithm.

Note that we have now shown that the series $\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j}$ converges, but does not converge absolutely; therefore, it is conditionally convergent.

4. We next consider the ***harmonic series*** $\sum_{j=1}^{\infty} j^{-k}$ for $k \in \mathbb{Z}_{\geq 0}$. For $k = 1$ this agrees with our example of part 2. We claim that this series converges if and only if $k > 1$. We have already considered the case of $k = 1$. For $k < 1$ we have $j^{-k} \geq j^{-1}$ for $j \in \mathbb{Z}_{>0}$. Therefore,

$$\sum_{j=1}^{\infty} j^{-k} \geq \sum_{j=1}^{\infty} j^{-1} = \infty,$$

showing that the series diverges to ∞.

For $k > 1$ we note that the sequence of partial sums is monotonically increasing. Thus, so show convergence of the series it suffices by Theorem 2.3.8 to show that the sequence of partial sums is bounded above. Let $N \in \mathbb{Z}_{>0}$ and take $j \in \mathbb{Z}_{>0}$ such that $N < 2^j - 1$. Then the Nth partial sum satisfies

$$\begin{aligned} S_N \leq S_{2^j-1} &= 1 + \frac{1}{2^k} + \frac{1}{3^k} + \cdots + \frac{1}{(2^j-1)^k} \\ &= 1 + \underbrace{\Big(\frac{1}{2^k} + \frac{1}{3^k}\Big)}_{\text{2 terms}} + \underbrace{\Big(\frac{1}{4^k} + \cdots + \frac{1}{7^k}\Big)}_{\text{4 terms}} + \cdots + \underbrace{\Big(\frac{1}{(2^{j-1})^k} + \cdots + \frac{1}{(2^j-1)^k}\Big)}_{2^{j-1}\text{ terms}} \\ &< 1 + \frac{2}{2^k} + \frac{4}{4^k} + \cdots + \frac{2^{j-1}}{(2^{j-1})^k} \\ &= 1 + \frac{1}{2^{k-1}} + \Big(\frac{1}{2^{k-1}}\Big)^2 + \cdots + \Big(\frac{1}{2^{k-1}}\Big)^{j-1}. \end{aligned}$$

Now we note that the last expression on the right-hand side is bounded above by the sum $\sum_{j=1}^{\infty}(2^{k-1})^{j-1}$, which is a convergent geometric series as we saw in part 1. This shows that S_N is bounded above by this sum for all N, so showing that the harmonic series converges for $k > 1$.

5. The series $\sum_{j=1}^{\infty}(-1)^{j+1}$ does not converge, and also does not diverge to ∞ or $-\infty$. Therefore, it is oscillatory. •

Let us next explore relationships between the various notions of convergence. First we relate the notions of convergence and absolute convergence in the only possible way, given that the series $\sum_{j=1} \frac{(-1)^{j+1}}{j}$ has been shown to be convergent, but not absolutely convergent.

2.4.3 Proposition (Absolutely convergent series are convergent) *If a series $\sum_{j=1}^{\infty} x_j$ is absolutely convergent, then it is convergent.*

Proof Denote

$$s_k = \sum_{j=1}^{k} x_j, \quad \sigma_k = \sum_{j=1}^{k} |x_j|,$$

and note that $(\sigma_k)_{k\in\mathbb{Z}_{>0}}$ is a Cauchy sequence since the series $\sum_{j=1}^{\infty} x_j$ is absolutely convergent. Thus let $\epsilon \in \mathbb{R}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|\sigma_k - \sigma_l| < \epsilon$ for $k, l \geq N$. For $m > k$ we then have

$$|s_m - s_k| = \Big|\sum_{j=k+1}^{m} x_j\Big| \leq \sum_{j=k+1}^{m} |x_j| = |\sigma_m - \sigma_k| < \epsilon,$$

where we have used Exercise 2.4.3. Thus, for $m > k \geq N$ we have $|s_m - s_k| < \epsilon$, showing that $(s_k)_{k\in\mathbb{Z}_{>0}}$ is a Cauchy sequence, and so convergent by Theorem 2.3.5. ■

The following result is often useful.

2.4.4 Proposition (Swapping summation and absolute value) *For a sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$, *if the series* $S = \sum_{j=1}^{\infty} x_j$ *is absolutely convergent, then*

$$\left|\sum_{j=1}^{\infty} x_j\right| \leq \sum_{j=1}^{\infty} |x_j|.$$

Proof Define

$$S_m^1 = \left|\sum_{j=1}^{m} x_j\right|, \quad S_m^2 = \sum_{j=1}^{m} |x_j|, \qquad m \in \mathbb{Z}_{>0}.$$

By Exercise 2.4.3 we have $S_m^1 \leq S_m^2$ for each $m \in \mathbb{Z}_{>0}$. Moreover, by Proposition 2.4.3 the sequences $(S_m^1)_{m\in\mathbb{Z}_{>0}}$ and $(S_m^2)_{m\in\mathbb{Z}_{>0}}$ converge. It is then clear (why?) that

$$\lim_{m\to\infty} S_m^1 \leq \lim_{m\to\infty} S_m^2,$$

which is the result. ■

It is not immediately clear on a first encounter why the notion of absolute convergence is useful. However, as we shall see in Chapter **??**, it is the notion of absolute convergence that will be of most use to us in our characterisation of discrete signal spaces. The following result indicates why mere convergence of a series is perhaps not as nice a notion as one would like, and that absolute convergence is in some sense better behaved.***missing stuff***

2.4.5 Theorem (Convergence and rearrangement of series) *For a series* $S = \sum_{j=1}^{\infty} x_j$, *the following statements hold:*

(i) *if* S *is conditionally convergent then, for any* $s_0 \in \mathbb{R}$, *there exists a bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ *such that the series* $S_\phi = \sum_{j=1}^{\infty} x_{\phi(j)}$ *converges to* s_0*;*

(ii) *if* S *is conditionally convergent then there exists a bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ *such that the series* $S_\phi = \sum_{j=1}^{\infty} x_{\phi(j)}$ *diverges to* ∞*;*

(iii) *if* S *is conditionally convergent then there exists a bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ *such that the series* $S_\phi = \sum_{j=1}^{\infty} x_{\phi(j)}$ *diverges to* $-\infty$*;*

(iv) *if* S *is conditionally convergent then there exists a bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ *such that the limit of the partial sums for the series* $S_\phi = \sum_{j=1}^{\infty} x_{\phi(j)}$ *is oscillating;*

(v) *if* S *is absolutely convergent then, for any bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$, *the series* $S_\phi = \sum_{j=1}^{\infty} x_{\phi(j)}$ *converges to the same limit as the series* S.

Proof We shall be fairly "descriptive" concerning the first four parts of the proof. More precise arguments can be tediously fabricated from the ideas given. We shall use the fact, given as Exercise 2.4.1, that if a series is conditionally convergent, then the two series formed by the positive terms and the negative terms diverge.

(i) First of all, rearrange the terms in the series so that the positive terms are arranged in decreasing order, and the negative terms are arranged in increasing order. We suppose that $s_0 \geq 0$, as a similar argument can be fabricated when $s_0 < 0$. Take as the first elements of the rearranged sequence the enough of the first few positive terms in the sequence so that their sum exceeds s_0. As the next terms, take enough of the first few negative terms in the series such that their sum, combined with the already

chosen positive terms, is less than s_0. Now repeat this process. Because the series was initially rearranged so that the positive and negative terms are in descending and ascending order, respectively, one can show that the construction we have given yields a sequence of partial sums that starts greater than s_0, then monotonically decreases to a value less than s_0, then monotonically increases to a value greater than s_0, and so on. Moreover, at the end of each step, the values get closer to s_0 since the sequence of positive and negative terms both converge to zero. An argument like that used in the proof of Proposition 2.3.10 can then be used to show that the resulting sequence of partial sums converges to s_0.

(ii) To get the suitable rearrangement, proceed as follows. Partition the negative terms in the sequence into disjoint finite sets S_j^-, $j \in \mathbb{Z}_{>0}$. Now partition the positive terms in the sequence as follows. Define S_1^+ to be the first N_1 positive terms in the sequence, where N_1 is sufficiently large that the sum of the elements of S_1^+ exceeds by at least 1 in absolute value the sum of the elements from S_1^-. This is possible since the series of positive terms in the sequence diverges to ∞. Now define S_2^+ by taking the next N_2 positive terms in the sequence so that the sum of the elements of S_2^+ exceeds by at least 1 in absolute value the sum of the elements from S_2^-. Continue in this way, defining $S_3^+, S_4^+, \ldots$. The rearrangement of the terms in the series is then made by taking the first collection of terms to be the elements of S_1^+, the second collection to be the elements of S_1^-, the third collection to be the elements of S_2^+, and so on. One can verify that the resulting sequence of partial sums diverges to ∞.

(iii) The argument here is entirely similar to the previous case.

(iv) This result follows from part (i) in the following way. Choose an oscillating sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$. For y_1, by part (i) one can find a finite number of terms from the original series whose sum is as close as desired to y_1. These will form the first terms in the rearranged series. Next, the same argument can be applied to the remaining elements of the series to yield a finite number of terms in the series that are as close as desired to y_2. One carries on in this way, noting that since the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ is oscillating, so too will be the sequence of partial sums for the rearranged series.

(v) Let $y_j = x_{\phi(j)}$ for $j \in \mathbb{Z}_{>0}$. Then define sequences $(x_j^+)_{j\in\mathbb{Z}_{>0}}$, $(x_j^-)_{j\in\mathbb{Z}_{>0}}$, $(y_j^+)_{j\in\mathbb{Z}_{>0}}$, and $(y_j^-)_{j\in\mathbb{Z}_{>0}}$ by

$$x_j^+ = \max\{x_j, 0\}, \quad x_j^- = \max\{-x_j, 0\},$$
$$y_j^+ = \max\{y_j, 0\}, \quad y_j^- = \max\{-y_j, 0\}, \qquad j \in \mathbb{Z}_{>0},$$

noting that $|x_j| = \max\{x_j^-, x_j^+\}$ and $|y_j| = \max\{y_j^-, y_j^+\}$ for $j \in \mathbb{Z}_{>0}$. By Proposition 2.4.8 it follows that the series

$$S^+ = \sum_{j=1}^{\infty} x_j^+, \quad S^- = \sum_{j=1}^{\infty} x_j^-, \quad S_\phi^+ = \sum_{j=1}^{\infty} y_j^+, \quad S_\phi^- = \sum_{j=1}^{\infty} y_j^-$$

converge. We claim that for each $k \in \mathbb{Z}_{>0}$ we have

$$\sum_{j=1}^{k} x_j^+ \leq \sum_{j=1}^{\infty} y_j^+.$$

To see this, we need only note that there exists $N \in \mathbb{Z}_{>0}$ such that

$$\{x_1^+, \ldots, x_k^+\} \subseteq \{y_1^+, \ldots, y_N^+\}.$$

With N having this property,

$$\sum_{j=1}^{k} x_j^+ \leq \sum_{j=1}^{N} y_j^+ \leq \sum_{j=1}^{\infty} y_j^+,$$

as desired. Therefore,

$$\sum_{j=1}^{\infty} x_j^+ \leq \sum_{j=1}^{\infty} y_j^+.$$

Reversing the argument gives

$$\sum_{j=1}^{\infty} y_j^+ \leq \sum_{j=1}^{\infty} x_j^+ \quad \Longrightarrow \quad \sum_{j=1}^{\infty} x_j^+ = \sum_{j=1}^{\infty} y_j^+.$$

A similar argument also gives

$$\sum_{j=1}^{\infty} x_j^- = \sum_{j=1}^{\infty} y_j^-.$$

This then gives

$$\sum_{j=1}^{\infty} y_j = \sum_{j=1}^{\infty} y_j^+ - \sum_{j=1}^{\infty} y_j^- = \sum_{j=1}^{\infty} x_j^+ - \sum_{j=1}^{\infty} x_j^- = \sum_{j=1}^{\infty} x_j,$$

as desired. ∎

The theorem says, roughly, that absolute convergence is necessary and sufficient to ensure that the limit of a series be independent of rearrangement of the terms in the series. Note that the necessity portion of this statement, which is parts (i)–(iv) of the theorem, comes in a rather dramatic form which suggests that conditional convergence behaves maximally poorly with respect to rearrangement.

2.4.2 Tests for convergence of series

In this section we give some of the more popular tests for convergence of a series. It is infeasible to expect an easily checkable general condition for convergence. However, in some cases the tests we give here are sufficient.

First we make a simple general observation that is very often useful; it is merely a reflection that the convergence of a series depends only on the tail of the series. We shall often make use of this result without mention.

2.4.6 Proposition (Convergence is unaffected by changing a finite number of terms) *Let $\sum_{j=1} x_j$ and $\sum_{j=1}^{\infty} y_j$ be series in $\mathbb{R}$ and suppose that there exists $K \in \mathbb{Z}$ and $N \in \mathbb{Z}_{>0}$ such that $x_j = y_{j+K}$ for $j \geq N$. Then the following statements hold:*

(i) the series $\sum_{j=1}^{\infty} x_j$ converges if and only if the series $\sum_{j=1}^{\infty} y_j$ converges;

(ii) the series $\sum_{j=1}^{\infty} x_j$ diverges if and only if the series $\sum_{j=1}^{\infty} y_j$ diverges;

(iii) the series $\sum_{j=1}^{\infty} x_j$ diverges to ∞ if and only if the series $\sum_{j=1}^{\infty} y_j$ diverges to ∞;

(iv) the series $\sum_{j=1}^{\infty} x_j$ diverges to $-\infty$ if and only if the series $\sum_{j=1}^{\infty} y_j$ diverges to $-\infty$.

The next convergence result is also a more or less obvious one.

2.4.7 Proposition (Sufficient condition for a series to diverge) *If the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ does not converge to zero, then the series $\sum_{j=1}^{\infty} x_j$ diverges.*

Proof Suppose that the series $\sum_{j=1}^{\infty} x_j$ converges to s_0 and let $(S_k)_{k\in\mathbb{Z}_{>0}}$ be the sequence of partial sums. Then $x_k = S_k - S_{k-1}$. Then

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} S_k - \lim_{k\to\infty} S_{k-1} = s_0 - s_0 = 0_V,$$

as desired. ■

Note that Example 2.4.2–2 shows that the converse of this result is false. That is to say, for a series to converge, it is not sufficient that the terms in the series go to zero. For this reason, checking the convergence of a series numerically becomes something that must be done carefully, since the blind use of the computer with a prescribed numerical accuracy will suggest the false conclusion that a series converges if and only if the terms in the series go to zero as the index goes to infinity.

Another more or less obvious result is the following.

2.4.8 Proposition (Comparison Test) *Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ and $(y_j)_{j\in\mathbb{Z}_{>0}}$ be sequences of nonnegative numbers for which there exists $\alpha \in \mathbb{R}_{>0}$ satisfying $y_j \leq \alpha x_j$, $j \in \mathbb{Z}_{>0}$. Then the following statements hold:*

(i) the series $\sum_{j=1}^{\infty} y_j$ converges if the series $\sum_{j=1}^{\infty} x_j$ converges;

(ii) the series $\sum_{j=1}^{\infty} x_j$ diverges if the series $\sum_{j=1}^{\infty} y_j$ diverges.

Proof We shall show that, if the series $\sum_{j=1}^{\infty} x_j$ converges, then the sequence $(T_k)_{k\in\mathbb{Z}_{>0}}$ of partial sums for the series $\sum_{j=1}^{\infty} y_j$ is a Cauchy sequence. Since the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ for $\sum_{j=1}^{\infty} x_j$ is convergent, it is Cauchy. Therefore, for $\epsilon \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that whenever $k, m \geq N$, with $k > m$ without loss of generality,

$$S_k - S_m = \sum_{j=m+1}^{k} x_j < \epsilon\alpha^{-1}.$$

Then, for $k, m \geq N$ with $k > m$ we have

$$T_k - T_m = \sum_{j=m+1}^{k} y_j \leq \alpha \sum_{j=m+1}^{k} x_j < \epsilon,$$

showing that $(T_k)_{k\in\mathbb{Z}_{>0}}$ is a Cauchy sequence, as desired.

The second statement is the contrapositive of the first. ■

Now we can get to some less obvious results for convergence of series. The first result concerns series where the terms alternate sign.

2.4.9 Proposition (Alternating Test) *Let* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *be a sequence in* $\mathbb{R}$ *satisfying*

(i) $x_j > 0$ *for* $j \in \mathbb{Z}_{>0}$,

(ii) $x_{j+1} \leq x_j$ *for* $j \in \mathbb{Z}_{>0}$, *and*

(iii) $\lim_{j\to\infty} x_j = 0$.

Then the series $\sum_{j=1}^{\infty}(-1)^{j+1}x_j$ *converges.*

Proof The proof is a straightforward generalisation of that given for Example 2.4.2–3, and we leave for the reader the simple exercise of verifying that this is so. ■

Our next result is one that is often useful.

2.4.10 Proposition (Ratio Test for series) *Let* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *be a nonzero sequence in* $\mathbb{R}$ *with* $\sum_{j=1}^{\infty} x_j$ *the corresponding series. Then the following statements hold:*

(i) if $\limsup_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| < 1$, *then the series converges absolutely;*

(ii) if there exists $N \in \mathbb{Z}_{>0}$ *such that* $\left|\frac{x_{j+1}}{x_j}\right| > 1$ *for all* $j \geq N$, *then the series diverges.*

Proof (i) By Proposition 2.3.15 there exists $\beta \in (0,1)$ and $N \in \mathbb{Z}_{>0}$ such that $\left|\frac{x_{j+1}}{x_j}\right| < \beta$ for $j \geq N$. Then

$$\left|\frac{x_j}{x_N}\right| = \left|\frac{x_{N+1}}{x_N}\right|\left|\frac{x_{N+2}}{x_{N+1}}\right|\cdots\left|\frac{x_j}{x_{j-1}}\right| < \beta^{j-N}, \qquad j > N,$$

implying that

$$|x_j| < \frac{|x_N|}{\beta^N}\beta^j.$$

Since $\beta < 1$, the geometric series $\sum_{j=1}^{\infty}\beta^j$ converges. The result for $\alpha < 1$ now follows by the Comparison Test.

(ii) The sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ cannot converge to 0 in this case, and so this part of the result follows from Proposition 2.4.7. ■

The following simpler test is often stated as the Ratio Test.

2.4.11 Corollary (Weaker version of the Ratio Test) *If* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *is a nonzero sequence in* $\mathbb{R}$ *for which* $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = \alpha$, *then the series* $\sum_{j=1}^{\infty} x_j$ *converges absolutely if* $\alpha < 1$ *and diverges if* $\alpha > 1$.

2.4.12 Remark (Nonzero assumption in Ratio Test) In the preceding two results we asked that the terms in the series be nonzero. This is not a significant limitation. Indeed, one can enumerate the nonzero terms in the series, and then apply the ratio test to this. •

Our next result has a similar character to the previous one.

2.4.13 Proposition (Root Test) *Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence for which $\limsup_{j\to\infty}|x_j|^{1/j} = \alpha$. Then the series $\sum_{j=1}^{\infty} x_j$ converges absolutely if $\alpha < 1$ and diverges if $\alpha > 1$.*

Proof First take $\alpha < 1$ and define $\beta = \frac{1}{2}(\alpha + 1)$. Then, just as in the proof of Proposition 2.4.10, $\alpha < \beta < 1$. By Proposition 2.3.15 there exists $N \in \mathbb{Z}_{>0}$ such that $|x_j|^{1/j} < \beta$ for $j \geq N$. Thus $|x_j| < \beta^j$ for $j \geq N$. Note that $\sum_{j=N+1}^{\infty} \beta^j$ converges by Example 2.4.2–1. Now $\sum_{j=0}^{\infty}|x_j|$ converges by the Comparison Test.

Next take $\alpha > 1$. In this case we have $\lim_{j\to\infty}|x_j| \neq 0$, and so we conclude divergence from Proposition 2.4.7. ■

The following obvious corollary is often stated as the Root Test.

2.4.14 Corollary (Weaker version of Root Test) *Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence for which $\lim_{j\to\infty}|x_j|^{1/j} = \alpha$. Then the series $\sum_{j=1}^{\infty} x_j$ converges absolutely if $\alpha < 1$ and diverges if $\alpha > 1$.*

The Ratio Test and the Root Test are related, as the following result indicates.

2.4.15 Proposition (Root Test implies Ratio Test) *If $(p_j)_{j\in\mathbb{Z}_{\geq 0}}$ is a sequence in $\mathbb{R}_{>0}$ then*

$$\liminf_{j\to\infty} \frac{p_{j+1}}{p_j} \leq \liminf_{j\to\infty} p_j^{1/j}$$
$$\limsup_{j\to\infty} p_j^{1/j} \leq \limsup_{j\to\infty} \frac{p_{j+1}}{p_j}.$$

In particular, for a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$, if $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right|$ exists, then $\lim_{j\to\infty}|x_j|^{1/j} = \lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right|$.

Proof For the first inequality, let $\alpha = \liminf_{j\to\infty} \frac{p_{j+1}}{p_j}$. First consider the case where $\alpha = \infty$. Then, given $M \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $\frac{p_{j+1}}{p_j} > M$ for $j \geq N$. Then we have

$$\left|\frac{p_j}{p_N}\right| = \left|\frac{p_{N+1}}{p_N}\right|\left|\frac{p_{N+1}}{p_{N+1}}\right|\cdots\left|\frac{p_j}{p_{j-1}}\right| > M^{j-N}, \qquad j > N.$$

This gives

$$p_j > \frac{p_N}{M^N} M^j, \qquad j > N.$$

Thus $p_j^{1/j} > (\frac{p_N}{M^N})^{1/j} M$. Since $\lim_{j\to\infty}(p_N\beta^{-N})^{1/j} = 1$ (cf. the definition of P_a in Section 3.6.3), we have $\liminf_{j\to\infty} p_j^{1/j} > M$, giving the desired conclusion in this case, since M is arbitrary. Next consider the case when $\alpha \in \mathbb{R}_{>0}$ and let $\beta < \alpha$. By Proposition 2.3.16 there exists $N \in \mathbb{Z}_{>0}$ such that $\frac{p_{j+1}}{p_j} \geq \beta$ for $j \geq N$. Performing just the same computation as above gives $p_j \geq \beta^{j-N} p_N$ for $j \geq N$. Therefore, $p_j^{1/j} \geq (p_N\beta^{-N})^{1/j}\beta$. Since $\lim_{j\to\infty}(p_N\beta^{-N})^{1/j} = 1$ we have $\liminf_{j\to\infty} p_j^{1/j} \geq \beta$. The first inequality follows since $\beta < \alpha$ is arbitrary.

Now we prove the second inequality. Let $\alpha = \limsup_{j\to\infty} \frac{p_{j+1}}{p_j}$. If $\alpha = \infty$ then the second inequality in the statement of the result is trivial. If $\alpha \in \mathbb{R}_{>0}$ then let $\beta > \alpha$ and note that there exists $N \in \mathbb{Z}_{>0}$ such that $\frac{p_{j+1}}{p_j} \leq \beta$ for $j \geq N$ by Proposition 2.3.15. In particular, just as in the proof of Proposition 2.4.10, $p_j \leq \beta^{j-N} p_N$ for $j \geq N$. Therefore,

$p_j^{1/j} \leq (p_N\beta^{-N})^{1/j}\beta$. Since $\lim_{j\to\infty}(p_N\beta^{-N})^{1/j} = 1$ we then have $\liminf_{j\to\infty} p_j^{1/j} \leq \beta$. the second inequality follows since $\beta > \alpha$ is arbitrary.

The final assertion follows immediately from the two inequalities using Proposition 2.3.17. ■

In Exercises 2.4.6 and 2.4.7 the reader can explore the various possibilities for the ratio test and root test when $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and $\lim_{j\to\infty}|x_j|^{1/j} = 1$, respectively.

The final result we state in this section can be thought of as the summation version of integration by parts.

2.4.16 Proposition (Abel's[6] partial summation formula) *For sequences* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(y_j)_{j\in\mathbb{Z}_{>0}}$ *of real numbers, denote* $S_k = \sum_{j=1}^k x_j$. *Then*

$$\sum_{j=1}^k x_j y_j = S_k y_{k+1} - \sum_{j=1}^k S_j(y_{j+1} - y_j) = S_k y_1 + \sum_{j=1}^k (S_k - S_j)(y_{j+1} - y_j).$$

Proof Let $S_0 = 0$ by convention. Since $x_j = S_j - S_{j-1}$ we have

$$\sum_{j=1}^n x_j y_j = \sum_{j=1}^n (S_j - S_{j-1})y_j = \sum_{j=1}^n S_j y_j - \sum_{j=1}^n S_{j-1}y_j.$$

Trivially,

$$\sum_{j=1}^n S_{j-1}y_j = \sum_{j=1}^n S_j y_{j+1} - S_n y_{n+1}.$$

This gives the first equality of the lemma. The second follows from a substitution of

$$y_{n+1} = \sum_{j=1}^n (y_{j+1} - y_j) + y_1$$

into the first equality. ■

2.4.3 e and π

In this section we consider two particular convergent series whose limits are among the most important of "physical constants."

2.4.17 Definition (e) $e = \sum_{j=0}^\infty \frac{1}{j!}$. •

Note that the series defining e indeed converges, for example, by the Ratio Test. Another common representation of e as a limit is the following.

[6] Niels Henrik Abel (1802–1829) was a Norwegian mathematician who worked in the area of analysis. An important theorem of Abel, one that is worth knowing for people working in application areas, is a theorem stating that there is no expression for the roots of a quintic polynomial in terms of the coefficients that involves only the operations of addition, subtraction, multiplication, division and taking roots.

2.4.18 Proposition (Alternative representations of e) *We have*

$$\mathrm{e} = \lim_{\mathrm{j}\to\infty}\Big(1+\tfrac{1}{\mathrm{j}}\Big)^{\mathrm{j}} = \lim_{\mathrm{j}\to\infty}\Big(1+\tfrac{1}{\mathrm{j}}\Big)^{\mathrm{j}+1}.$$

Proof First note that if the limit $\lim_{j\to\infty}\big(1+\frac{1}{j}\big)^j$ exists, then, by Proposition 2.3.23,

$$\lim_{j\to\infty}\Big(1+\tfrac{1}{j}\Big)^{j+1} = \lim_{j\to\infty}\Big(1+\tfrac{1}{j}\Big)\Big(1+\tfrac{1}{j}\Big)^{j} = \lim_{j\to\infty}\Big(1+\tfrac{1}{j}\Big)^{j}.$$

Thus we will only prove that $\mathrm{e} = \lim_{j\to\infty}\big(1+\frac{1}{j}\big)^j$.

Let

$$S_k = \sum_{j=0}^{k}\frac{1}{k!}, \quad A_k = \Big(1+\tfrac{1}{k}\Big)^k, \quad B_k = \Big(1+\tfrac{1}{k}\Big)^{k+1},$$

be the kth partial sum of the series for e and the kth term in the proposed sequence for e. By the Binomial Theorem (Exercise 2.2.1) we have

$$A_k = \Big(1+\tfrac{1}{k}\Big)^k = \sum_{j=0}^{k}\binom{k}{j}\frac{1}{k^j}.$$

Moreover, the exact form for the binomial coefficients can directly be seen to give

$$A_k = \sum_{j=0}^{k}\frac{1}{j!}\Big(1-\frac{1}{k}\Big)\Big(1-\frac{2}{k}\Big)\dots\Big(1-\frac{j-1}{k}\Big).$$

Each coefficient of $\frac{1}{j!}$, $j\in\{0,1,\dots,k\}$ is then less than 1. Thus $A_k \le S_k$ for each $k\in\mathbb{Z}_{\ge 0}$. Therefore, $\limsup_{k\to\infty} A_k \le \limsup_{k\to\infty} S_k$. For $m\le k$ the same computation gives

$$A_k \ge \sum_{j=0}^{m}\frac{1}{j!}\Big(1-\frac{1}{k}\Big)\Big(1-\frac{2}{k}\Big)\dots\Big(1-\frac{j-1}{k}\Big).$$

Fixing m and letting $k\to\infty$ gives

$$\liminf_{k\to\infty} A_k \ge \sum_{j=0}^{m}\frac{1}{j!} = S_m.$$

Thus $\liminf_{k\to\infty} A_k \ge \liminf_{m\to\infty} S_m$, which gives the result when combined with our previous estimate $\limsup_{k\to\infty} A_k \le \limsup_{k\to\infty} S_k$. ■

It is interesting to note that the series representation of e allows us to conclude that e is irrational.

2.4.19 Proposition (Irrationality of e) $\mathrm{e} \in \mathbb{R} \setminus \mathbb{Q}$.

Proof Suppose that $\mathrm{e} = \frac{l}{m}$ for $l, m \in \mathbb{Z}_{>0}$. We compute

$$(m-1)!l = m!\mathrm{e} = m! \sum_{j=0}^{\infty} \frac{1}{j!} = \sum_{j=0}^{m} \frac{m!}{j!} + \sum_{j=m+1}^{\infty} \frac{m!}{j!},$$

which then gives

$$\sum_{j=m+1}^{\infty} \frac{m!}{j!} = (m-1)!l - \sum_{j=0}^{m} \frac{m!}{j!},$$

which implies that $\sum_{j=m+1}^{\infty} \frac{m!}{j!} \in \mathbb{Z}_{>0}$. We then compute, using Example 2.4.2–1,

$$0 < \sum_{j=m+1}^{\infty} \frac{m!}{j!} < \sum_{j=m+1}^{\infty} \frac{1}{(m+1)^{j-m}} = \sum_{j=1}^{\infty} \frac{1}{(m+1)^{j}} = \frac{\frac{1}{m+1}}{1-\frac{1}{m+1}} = \frac{1}{m} \leq 1.$$

Thus $\sum_{j=m+1}^{\infty} \frac{m!}{j!} \in \mathbb{Z}_{>0}$, being an integer, must equal 1, and, moreover, $m = 1$. Thus we have

$$\sum_{j=2}^{\infty} \frac{1}{j!} = \mathrm{e} - 2 = 1 \quad \Longrightarrow \quad \mathrm{e} = 3.$$

Next let

$$\alpha = \sum_{j=1}^{\infty} \Big(\frac{1}{2^{j-1}} - \frac{1}{j!} \Big),$$

noting that this series for α converges, and converges to a positive number since each term in the series is positive. Then, using Example 2.4.2–1,

$$\alpha = (2 - (\mathrm{e} - 1)) \quad \Longrightarrow \quad \mathrm{e} = 3 - \alpha.$$

Thus $\mathrm{e} < 3$, and we have arrived at a contradiction. ■

Next we turn to the number π. Perhaps the best description of π is that it is the ratio of the circumference of a circle with the diameter of the circle. Indeed, the use of the Greek letter "p" (i.e., π) has its origins in the word "perimeter." However, to make sense of this definition, one must be able to talk effectively about circles, what the circumference means, etc. This is more trouble than it is worth for us at this point. Therefore, we give a more analytic description of π, albeit one that, at this point, is not very revealing of what the reader probably already knows about it.

2.4.20 Definition (π) $\pi = 4 \sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1}$. •

By the Alternating Test, this series representation for π converges.

We can also fairly easily show that π is irrational, although our proof uses some facts about functions on $\mathbb{R}$ that we will not discuss until Chapter 3.

2.4.21 Proposition (Irrationality of π) $\pi \in \mathbb{R} \setminus \mathbb{Q}$.

Proof In Section 3.6.4 we will give a definition of the trigonometric functions, sin and cos, and prove that, on $(0, \pi)$, sin is positive, and that $\sin 0 = \sin \pi = 0$. We will also prove the rules of differentiation for trigonometric functions necessary for the proof we now present.

Note that if π is rational, then π^2 is also rational. Therefore, it suffices to show that π^2 is irrational.

Let us suppose that $\pi^2 = \frac{l}{m}$ for $l, m \in \mathbb{Z}_{>0}$. For $k \in \mathbb{Z}_{>0}$ define $f_k \colon [0,1] \to \mathbb{R}$ by

$$f_k(x) = \frac{x^k(1-x)^k}{k!},$$

noting that $\operatorname{image}(f) \subseteq [0, \frac{1}{k!}]$. It is also useful to write

$$f_k(x) = \frac{1}{k!} \sum_{j=k}^{2k} c_j x^j,$$

where we observe that c_j, $j \in \{k, k+1, \dots, 2k\}$ are integers. Define $g_j \colon [0,1] \to \mathbb{R}$ by

$$g_k(x) = k^j \sum_{j=0}^{k} (-1)^j \pi^{2(k-j)} f^{(2j)}(x).$$

A direct computation shows that

$$f_k^{(j)}(0) = 0, \qquad j < k, \ j > 2k,$$

and that

$$f_k^{(j)}(0) = \frac{j!}{k!} c_j, \qquad j \in \{k, k+1, \dots, 2k\},$$

is an integer. Thus f and all of its derivatives take integer values at $x = 0$, and therefore also at $x = 1$ since $f_k(x) = f_k(1-x)$. One also verifies directly that $g_k(0)$ and $g_k(1)$ are integers.

Now we compute

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}(g_k'(x) \sin \pi x - \pi g_k(x) \cos \pi x) &= (g_k''(x) + \pi^2 g_k(x)) \sin \pi x \\ &= m^k \pi^{2k+2} f(x) \sin \pi x = \pi^2 l^k f(x) \sin \pi x, \end{aligned}$$

using the definition of g_k and the fact that $\pi^2 = \frac{l}{m}$. By the Fundamental Theorem of Calculus we then have, after a calculation,

$$\pi l^k \int_0^1 f(x) \sin \pi x \, \mathrm{d}x = g_k(0) + g_k(1) \in \mathbb{Z}_{>0}.$$

But we then have, since the integrand in the above integral is nonnegative,

$$0 < \pi l^k \int_0^1 f(x) \sin \pi x \, \mathrm{d}x < \frac{\pi l^k}{k!}$$

given the bounds on f_k. Note that $\lim_{k \to \infty} \frac{l^k}{k!} = 0$. Since the above computations hold for any k, if we take k sufficiently large that $\frac{\pi l^k}{k!} < 1$, we arrive at a contradiction. ■

2.4.4 Doubly infinite series

We shall frequently encounter series whose summation index runs not from 1 to ∞, but from $-\infty$ to ∞. Thus we call a family $(x_j)_{j\in\mathbb{Z}}$ of elements of $\mathbb{R}$ a ***doubly infinite sequence*** in $\mathbb{R}$, and a sum of the form $\sum_{j=-\infty}^{\infty} x_j$ a ***doubly infinite series***. A little care need to be shown when defining convergence for such series, and here we give the appropriate definitions.

2.4.22 Definition (Convergence and absolute convergence of doubly infinite series) Let $(x_j)_{j\in\mathbb{Z}}$ be a doubly infinite sequence and let $S = \sum_{j=-\infty}^{\infty} x_j$ be the corresponding doubly infinite series. The sequence of ***single partial sums*** is the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ where

$$S_k = \sum_{j=-k}^{k} x_j,$$

and the sequence of ***double partial sums*** is the double sequence $(S_{k,l})_{k,l\in\mathbb{Z}_{>0}}$ defined by

$$S_{k,l} = \sum_{j=-k}^{l} x_j.$$

Let $s_0 \in \mathbb{R}$. The doubly infinite series:

(i) ***converges to*** $\mathbf{s_0}$ if the double sequence of partial sums converges to s_0;

(ii) has s_0 as a ***limit*** if it converges to s_0;

(iii) is ***convergent*** if it converges to some element of $\mathbb{R}$;

(iv) ***converges absolutely***, or is ***absolutely convergent***, if the doubly infinite series

$$\sum_{j=-\infty}^{\infty} |x_j|$$

converges;

(v) ***converges conditionally***, or is ***conditionally convergent***, if it is convergent, but not absolutely convergent;

(vi) ***diverges*** if it does not converge;

(vii) ***diverges to*** $\boldsymbol{\infty}$ (resp. ***diverges to*** $\boldsymbol{-\infty}$), and we write $\sum_{j=-\infty}^{\infty} x_j = \infty$ (resp. $\sum_{j=-\infty}^{\infty} x_j = -\infty$), if the sequence of double partial sums diverges to ∞ (resp. diverges to $-\infty$);

(viii) has a limit that ***exists*** if $\sum_{j=-\infty}^{\infty} x_j \in \mathbb{R}$;

(ix) is ***oscillatory*** if the limit of the double sequence of partial sums is oscillatory. •

2.4.23 Remark (Partial sums versus double partial sums) Note that the convergence of the sequence of partial sums is not a very helpful notion, in general. For example, the series $\sum_{j=-\infty}^{\infty} j$ possesses a sequence of partial sums that is identically zero, and so the sequence of partial sums obviously converges to zero. However, it is not likely that one would wish this doubly infinite series to qualify as convergent. Thus partial sums are not a particularly good measure of convergence. However, there are situations—for example, the convergence of Fourier series (see Chapter **??**)—where the standard notion of convergence of a doubly infinite series is made using the partial sums. However, in these cases, there is additional structure on the setup that makes this a reasonable thing to do. •

The convergence of a doubly infinite series has the following useful, intuitive characterisation.

2.4.24 Proposition (Characterisation of convergence of doubly infinite series) *For a doubly infinite series* $\mathrm{S} = \sum_{\mathrm{j}=-\infty}^{\infty} \mathrm{x}_\mathrm{j}$*, the following statements are equivalent:*

(i) S *converges;*

(ii) the two series $\sum_{\mathrm{j}=0}^{\infty} \mathrm{x}_\mathrm{j}$ *and* $\sum_{\mathrm{j}=1}^{\infty} \mathrm{x}_{-\mathrm{j}}$ *converge.*

Proof For $k, l \in \mathbb{Z}_{>0}$, denote

$$S_{k,l} = \sum_{-k}^{l} x_j, \quad S_k^+ = \sum_{j=0}^{k} x_j, \quad S_k^- = \sum_{-k}^{-1} x_j,$$

so that $S_{k,l} = S_k^- + S_l^+$.

(i) $\implies$ (ii) Let $\epsilon \in \mathbb{R}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|S_{j,k} - s_0| < \frac{\epsilon}{2}$ for $j, k \geq N$. Now let $j, k \geq N$, choose some $l \geq N$, and compute

$$|S_j^+ - S_k^+| \leq |S_j^+ + S_l^- - s_0| + |S_k^+ + S_l^- - s_0| < \epsilon.$$

Thus $(S_j^+)_{j \in \mathbb{Z}_{>0}}$ is a Cauchy sequence, and so is convergent. A similar argument shows that $(S_j^-)_{j \in \mathbb{Z}_{>0}}$ is also a Cauchy sequence.

(ii) $\implies$ (i) Let s^+ be the limit of $\sum_{j=0}^{\infty} x_j$ and let s^- be the limit of $\sum_{j=1}^{\infty} x_{-j}$. For $\epsilon \in \mathbb{R}_{>0}$ define $N^+, N^- \in \mathbb{Z}_{>0}$ such that $|S_j^+ - s^+| < \frac{\epsilon}{2}$, $j \geq N^+$, and $|S_j^- - s^-| < \frac{\epsilon}{2}$, $j \leq -N^-$. Then, for $j, k \geq \max\{N^-, N^+\}$,

$$|S_{j,k} - (s^+ + s^-)| = |S_k^+ - s_+ + S_j^- - s_-| \leq |S_k^+ - s_+| + |S_j^- - s_-| < \epsilon,$$

thus showing that S converges. ■

Thus convergent doubly infinite series are really just combinations of convergent series in the sense that we have studied in the preceding sections. Thus, for example, one can use the tests of Section 2.4.2 to check for convergence of a doubly infinite series by applying them to both "halves" of the series. Also, the relationships between convergence and absolute convergence for series also hold for doubly infinite series. And a suitable version of Theorem 2.4.5 also holds for doubly infinite series. These facts are so straightforward that we will assume them in the sequel without explicit mention; they all follow directly from Proposition 2.4.24.

2.4.5 Multiple series

Just as we considered multiple sequences in Section 2.3.5, we can consider multiple series. As we did with sequences, we content ourselves with double series.

2.4.25 Definition (Double series) A ***double series*** in $\mathbb{R}$ is a sum of the form $\sum_{j,k=1}^{\infty} x_{jk}$ where $(x_{jk})_{j,k\in\mathbb{Z}_{>0}}$ is a double sequence in $\mathbb{R}$. •

While our definition of a series was not entirely sensible since it was not really identifiable as anything unless it had certain convergence properties, for double series, things are even worse. In particular, it is not clear what $\sum_{j,k=1}^{\infty} x_{jk}$ means. Does it mean $\sum_{j=1}^{\infty}\left(\sum_{k=1}^{\infty} x_{jk}\right)$? Does it mean $\sum_{k=1}^{\infty}\left(\sum_{j=1}^{\infty} x_{jk}\right)$? Or does it mean something different from both of these? The only way to rectify our poor mathematical manners is to define convergence for double series as quickly as possible.

2.4.26 Definition (Convergence and absolute convergence of double series) Let $(x_{jk})_{j,k\in\mathbb{Z}_{>0}}$ be a double sequence in $\mathbb{R}$ and consider the double series

$$S = \sum_{j,k=1}^{\infty} x_{jk}.$$

The corresponding sequence of ***partial sums*** is the double sequence $(S_{jk})_{j,k\in\mathbb{Z}_{>0}}$ defined by

$$S_{jk} = \sum_{l=1}^{j}\sum_{m=1}^{k} x_{lm}.$$

Let $s_0 \in \mathbb{R}$. The double series:

(i) ***converges to*** $\mathbf{s_0}$, and we write $\sum_{j,k=1}^{\infty} x_{jk} = s_0$, if the double sequence of partial sums converges to s_0;

(ii) has s_0 as a ***limit*** if it converges to s_0;

(iii) is ***convergent*** if it converges to some member of $\mathbb{R}$;

(iv) ***converges absolutely***, or is ***absolutely convergent***, if the series

$$\sum_{j,k=1}^{\infty} |x_{jk}|$$

converges;

(v) ***converges conditionally***, or is ***conditionally convergent***, if it is convergent, but not absolutely convergent;

(vi) ***diverges*** if it does not converge;

(vii) ***diverges to*** ∞ (resp. ***diverges to*** $-\infty$), and we write $\sum_{j,k=1}^{\infty} x_{jk} = \infty$ (resp. $\sum_{j,k=1}^{\infty} x_{jk} = -\infty$), if the double sequence of partial sums diverges to ∞ (resp. diverges to $-\infty$);

(viii) has a limit that ***exists*** if $\sum_{j,k=1}^{\infty} x_{jk} \in \mathbb{R}$;

(ix) is ***oscillatory*** if the sequence of partial sums is oscillatory. •

Note that the definition of the partial sums, S_{jk}, $j, k \in \mathbb{Z}_{>0}$, for a double series is unambiguous since

$$\sum_{l=1}^{j} \sum_{m=1}^{k} x_{lm} = \sum_{m=1}^{k} \sum_{l=1}^{j} x_{lm},$$

this being valid for finite sums. The idea behind convergence of double series, then, has an interpretation that can be gleaned from that in Figure 2.2 for double sequences.

Let us state a result, derived from similar results for double sequences, that allows the computation of limits of double series by computing one limit at a time.

2.4.27 Proposition (Computation of limits of double series I) *Suppose that for the double series* $\sum_{j,k=1}^{\infty} x_{jk}$ *it holds that*

(i) the double series is convergent and

(ii) for each $j \in \mathbb{Z}_{>0}$, *the series* $\sum_{k=1}^{\infty} x_{jk}$ *converges.*

Then the series $\sum_{j=1}^{\infty}(\sum_{k=1}^{\infty} x_{jk})$ *converges and its limit is equal to* $\sum_{j,k=1}^{\infty} x_{jk}$.

Proof This follows directly from Proposition 2.3.20. ■

2.4.28 Proposition (Computation of limits of double series II) *Suppose that for the double series* $\sum_{j,k=1}^{\infty} x_{jk}$ *it holds that*

(i) the double series is convergent,

(ii) for each $j \in \mathbb{Z}_{>0}$, *the series* $\sum_{k=1}^{\infty} x_{jk}$ *converges, and*

(iii) for each $k \in \mathbb{Z}_{>0}$, *the limit* $\sum_{j=1}^{\infty} x_{jk}$ *converges.*

Then the series $\sum_{j=1}^{\infty}(\sum_{k=1}^{\infty} x_{jk})$ *and* $\sum_{k=1}^{\infty}(\sum_{j=1}^{\infty} x_{jk})$ *converge and their limits are both equal to* $\sum_{j,k=1}^{\infty} x_{jk}$.

Proof This follows directly from Proposition 2.3.21. ■

missing stuff

2.4.6 Algebraic operations on series

In this section we consider the manner in which series interact with algebraic operations. The results here mirror, to some extent, the results for sequences in Section 2.3.6. However, the series structure allows for different ways of thinking about the product of sequences. Let us first give these definitions. For notational convenience, we use sums that begin at 0 rather than 1. This clearly has no affect on the definition of a series, or on any of its properties.

2.4.29 Definition (Products of series) Let $S = \sum_{j=0}^{\infty} x_j$ and $T = \sum_{j=0}^{\infty} y_j$ be series in $\mathbb{R}$.

(i) The ***product*** of S and T is the double series $\sum_{j,k=0}^{\infty} x_j y_k$.

(ii) The ***Cauchy product*** of S and T is the series $\sum_{k=0}^{\infty}\Big(\sum_{j=0}^{k} x_j y_{k-j}\Big)$. •

Now we can state the basic results on algebraic manipulation of series.

2.4.30 Proposition (Algebraic operations on series) *Let* $\mathrm{S} = \sum_{\mathrm{j}=0}^{\infty} \mathrm{x_j}$ *and* $\mathrm{T} = \sum_{\mathrm{j}=0}^{\infty} \mathrm{y_j}$ *be series in* $\mathbb{R}$ *that converges to* s_0 *and* t_0*, respectively, and let* $\alpha \in \mathbb{R}$*. Then the following statements hold:*

(i) the series $\sum_{\mathrm{j}=0}^{\infty} \alpha \mathrm{x_j}$ *converges to* $\alpha \mathrm{s}_0$*;*

(ii) the series $\sum_{\mathrm{j}=0}^{\infty} (\mathrm{x_j} + \mathrm{y_j})$ *converges to* $\mathrm{s}_0 + \mathrm{t}_0$*;*

(iii) if S *and* T *are absolutely convergent, then the product of* S *and* T *is absolutely convergent and converges to* $\mathrm{s}_0\mathrm{t}_0$*;*

(iv) if S *and* T *are absolutely convergent, then the Cauchy product of* S *and* T *is absolutely convergent and converges to* $\mathrm{s}_0\mathrm{t}_0$*;*

(v) if S *or* T *are absolutely convergent, then the Cauchy product of* S *and* T *is convergent and converges to* $\mathrm{s}_0\mathrm{t}_0$*;*

(vi) if S *and* T *are convergent, and if the Cauchy product of* S *and* T *is convergent, then the Cauchy product of* S *and* T *converges to* $\mathrm{s}_0\mathrm{t}_0$*.*

Proof (i) Since $\sum_{j=0}^{k} \alpha x_j = \alpha \sum_{j=0}^{k} x_j$, this follows from part (i) of Proposition 2.3.23.

(ii) Since $\sum_{j=0}^{\infty}(x_j + y_j) = \sum_{j=0}^{k} x_j + \sum_{j=0}^{k} y_j$, this follows from part (ii) of Proposition 2.3.23.

(iii) and (iv) To prove these parts of the result, we first make a general argument. We note that $\mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$ is a countable set (e.g., by Proposition 1.7.16), and so there exists a bijection, in fact many bijections, $\phi \colon \mathbb{Z}_{>0} \to \mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$. For such a bijection ϕ, suppose that we are given a double sequence $(x_{jk})_{j,k \in \mathbb{Z}_{\geq 0}}$ and define a sequence $(x_j^{\phi})_{j \in \mathbb{Z}_{>0}}$ by $x_j^{\phi} = x_{kl}$ where $(k, l) = \phi(j)$. We then claim that, for any bijection $\phi \colon \mathbb{Z}_{>0} \to \mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$, the double series $A = \sum_{k,l=1}^{\infty} x_{kl}$ converges absolutely if and only if the series $A^{\phi} = \sum_{j=1}^{\infty} x_j^{\phi}$ converges absolutely.

Indeed, suppose that the double series $|A| = \sum_{k,l=1}^{\infty} |x_{kl}|$ converges to $\beta \in \mathbb{R}$. For $\epsilon \in \mathbb{R}_{>0}$ the set

$$\{(k, l) \in \mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0} \mid \; ||A|_{kl} - \beta| \geq \epsilon\}$$

is then finite. Therefore, there exists $N \in \mathbb{Z}_{>0}$ such that, if $(k, l) = \phi(j)$ for $j \geq N$, then $||A|_{kl} - \beta| < \epsilon$. It therefore follows that $||A^{\phi}|_j - \beta| < \epsilon$ for $j \geq N$, where $|A^{\phi}|$ denotes the series $\sum_{j=1}^{\infty} |x_j^{\phi}|$. This shows that the series $|A^{\phi}|$ converges to β.

For the converse, suppose that the series $|A^{\phi}|$ converges to β. Then, for $\epsilon \in \mathbb{R}_{>0}$ the set

$$\{j \in \mathbb{Z}_{>0} \mid \; ||A^{\phi}|_j - \beta| \geq \epsilon\}$$

is finite. Therefore, there exists $N \in \mathbb{Z}_{>0}$ such that

$$\{(k, l) \in \mathbb{Z}_{\geq 0} \mid \; k, l \geq N\} \cap \{(k, l) \in \mathbb{Z}_{\geq 0} \mid \; ||A^{\phi}|_{\phi^{-1}(k,l)} - \beta| \geq \epsilon\} = \emptyset.$$

It then follows that for $k, l \geq N$ we have $||A|_{kl} - \beta| < \epsilon$, showing that $|A|$ converges to β.

Thus we have shown that A is absolutely convergent if and only if A^ϕ is absolutely convergent for any bijection $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$. From part (v) of Theorem 2.4.5, and its generalisation to double series, we know that the limit of an absolutely convergent series or double series is independent of the manner in which the terms in the series are arranged.

Consider now a term in the product of S and T. It is easy to see that this term appears exactly once in the Cauchy product of S and T. Conversely, each term in the Cauchy product appears exactly one in the product. Thus the product and Cauchy product are simply rearrangements of one another. Moreover, each term in the product and the Cauchy product appears exactly once in the expression

$$\Big(\sum_{j=0}^{N} x_j\Big)\Big(\sum_{k=0}^{N} y_k\Big)$$

as we allow N to go to ∞. That is to say,

$$\sum_{j,k=0}^{\infty} x_j y_k = \sum_{k=0}^{\infty}\Big(\sum_{j=k}^{k} x_j y_{k-j}\Big) = \lim_{N\to\infty}\Big(\sum_{j=0}^{N} x_j\Big)\Big(\sum_{k=0}^{N} y_k\Big).$$

However, this last limit is exactly $s_0 t_0$, using part (iii) of Proposition 2.3.23.

(v) Without loss of generality, suppose that S converges absolutely. Let $(S_k)_{k\in\mathbb{Z}_{>0}}$, $(T_k)_{k\in\mathbb{Z}_{>0}}$, and $((ST)_k)_{k\in\mathbb{Z}_{>0}}$ be the sequences of partial sums for S, T, and the Cauchy product, respectively. Also define $\tau_k = T_k - t_0$, $k \in \mathbb{Z}_{\geq 0}$. Then

$$\begin{aligned}(ST)_k &= x_0 y_0 + (x_0 y_1 + x_1 y_0) + \cdots + (x_0 y_k + \cdots + x_k y_0)\\ &= x_0 T_k + x_1 T_{k-1} + \cdots + x_k T_0\\ &= x_0(t_0 + \tau_k) + x_1(t_0 + \tau_{k-1}) + \cdots + x_k(t_0 + \tau_0)\\ &= S_k t_0 + x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0.\end{aligned}$$

Since $\lim_{k\to\infty} S_k t_0 = s_0 t_0$ by part (i), this part of the result will follow if we can show that

$$\lim_{k\to\infty}(x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0) = 0. \tag{2.6}$$

Denote

$$\sigma = \sum_{j=0}^{\infty} |x_j|,$$

and for $\epsilon \in \mathbb{R}_{>0}$ choose $N_1 \in \mathbb{Z}_{>0}$ such that $|\tau_j| \leq \frac{\epsilon}{2\sigma}$ for $j \geq N_1$, this being possible since $(\tau_j)_{j\in\mathbb{Z}_{>0}}$ clearly converges to zero. Then, for $k \geq N_1$,

$$\begin{aligned}|x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0| &\leq |x_0\tau_k + \cdots + x_{k-N_1-1}\tau_{N_1-1}| + |x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0|\\ &\leq \tfrac{\epsilon}{2} + |x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0|.\end{aligned}$$

Since $\lim_{k\to\infty} x_k = 0$, choose $N_2 \in \mathbb{Z}_{>0}$ such that

$$|x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0| < \tfrac{\epsilon}{2}$$

for $k \geq N_2$. Then

$$\begin{aligned}\limsup_{k\to\infty}|x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0| &= \lim_{k\to\infty}\sup\{|x_0\tau_j + x_1\tau_{j-1} + \cdots + x_j\tau_0| \mid j \geq k\}\\ &\leq \lim_{k\to\infty}\sup\{\tfrac{\epsilon}{2} + |x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0| \mid j \geq k\}\\ &\leq \sup\{\tfrac{\epsilon}{2} + |x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0| \mid j \geq N_2\} \leq \epsilon.\end{aligned}$$

Thus

$$\limsup_{k\to\infty}|x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0| \leq 0,$$

and since clearly

$$\liminf_{k\to\infty}|x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0| \geq 0,$$

we infer that (2.6) holds by Proposition 2.3.17.

(vi) The reader can prove this as Exercise 3.5.3. ■

The reader is recommended to remember the Cauchy product when we talk about convolution of discrete-time signals in Section **??**.

missing stuff

2.4.7 Series with arbitrary index sets

It will be helpful on a few occasions to be able to sum series whose index set is not necessarily countable, and here we indicate how this can be done. This material should be considered optional until one comes to that point in the text where it is needed.

2.4.31 Definition (Sum of series for arbitrary index sets) Let A be a set and let $(x_a)_{a\in A}$ be a family of elements of $\overline{\mathbb{R}}$. Let $A_+ = \{a \in A \mid x_a \in [0,\infty]\}$ and $A_- = \{a \in A \mid x_a \in [-\infty, 0]\}$.

(i) If $x_a \in [0,\infty]$ for $a \in A$, then $\sum_{a\in A} x_a = \sup\{\sum_{a\in A'} x_a \mid A' \subseteq A \text{ is finite}\}$.

(ii) For a general family, $\sum_{a\in A} x_a = \sum_{a_+\in A_+} x_{a_+} - \sum_{a_-\in A_-}(-x_{a_-})$, provided that at least one of $\sum_{a_+\in A_+} x_{a_+}$ or $\sum_{a_-\in A_-}(-x_{a_-})$ is finite.

(iii) If both $\sum_{a_+\in A_+} x_{a_+}$ are $\sum_{a_-\in A_-}(-x_{a_-})$ are finite, then $(x_a)_{a\in A}$ is ***summable***. •

We should understand the relationship between this sort of summation and our existing notion of the sum of a series in the case where the index set is $\mathbb{Z}_{>0}$.

2.4.32 Proposition (A summable series with index set $\mathbb{Z}_{>0}$ is absolutely convergent) *A sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in $\mathbb{R}$ is summable if and only if the series $S = \sum_{j=1}^{\infty} x_j$ is absolutely convergent.*

Proof Consider the sequences $(x_j^+)_{j\in\mathbb{Z}_{>0}}$ and $(x_j^-)_{j\in\mathbb{Z}_{>0}}$ defined by

$$x_j^+ = \max\{x_j, 0\}, \quad x_j^- = \max\{-x_j, 0\}, \qquad j \in \mathbb{Z}_{>0}.$$

Then $(x_j)_{j\in\mathbb{Z}_{>0}}$ is summable if and only if both of the expressions

$$\sup\Big\{\sum_{j\in A'} x_j^+ \;\Big|\; A' \subseteq \mathbb{Z}_{>0} \text{ is finite}\Big\}, \quad \sup\Big\{\sum_{j\in A'} x_j^- \;\Big|\; A' \subseteq \mathbb{Z}_{>0} \text{ is finite}\Big\} \tag{2.7}$$

are finite.

First suppose that $(x_j)_{j\in\mathbb{Z}_{>0}}$ is summable. Therefore, if $(S_k^+)_{k\in\mathbb{Z}_{>0}}$ and $(S_k^-)_{k\in\mathbb{Z}_{>0}}$ are the sequences of partial sums

$$S_k^+ = \sum_{j=1}^{k} x_j^+, \quad S_k^- = \sum_{j=1}^{k} x_j^-,$$

then these sequences are increasing and so convergent by (2.7). Then, by Proposition 2.3.23,

$$\sum_{j=1}^{\infty} |x_j| = \sum_{j=1}^{\infty} x_j^+ + \sum_{j=1}^{\infty} x_j^-$$

giving absolute convergence of S.

Now suppose that S is absolutely convergent. Then the subsets $\{S_k^+ \mid k \in \mathbb{Z}_{>0}\}$ and $\{S_k^- \mid k \in \mathbb{Z}_{>0}\}$ are bounded above (as well as being bounded below by zero) so that both expressions

$$\sup\{S_k^+ \mid k \in \mathbb{Z}_{>0}\}, \quad \sup\{S_k^- \mid k \in \mathbb{Z}_{>0}\}$$

are finite. Then for any finite set $A' \subseteq \mathbb{Z}_{>0}$ we have

$$\sum_{j\in A'} x_j^+ \leq S^+_{\sup A'}, \quad \sum_{j\in A'} x_j^- \leq S^-_{\sup A'}.$$

From this we deduce that

$$\sup\Big\{\sum_{j\in A'} x_j^+ \;\Big|\; A' \subseteq \mathbb{Z}_{>0} \text{ is finite}\Big\} \leq \sup\{S_k^+ \mid k \in \mathbb{Z}_{>0}\},$$
$$\sup\Big\{\sum_{j\in A'} x_j^- \;\Big|\; A' \subseteq \mathbb{Z}_{>0} \text{ is finite}\Big\} \leq \sup\{S_k^- \mid k \in \mathbb{Z}_{>0}\},$$

which implies that $(x_j)_{j\in\mathbb{Z}_{>0}}$ is summable. ■

Now we can actually show that, for a summable family of real numbers, only countably many of them can be nonzero.

2.4.33 Proposition (A summable family has at most countably many nonzero members) *If* $(\mathrm{x_a})_{\mathrm{a}\in\mathrm{A}}$ *is summable, then the set* $\{\mathrm{a} \in \mathrm{A} \mid \mathrm{x_a} \neq 0\}$ *is countable.*

Proof Note that for any $k \in \mathbb{Z}_{>0}$, the set $\{a \in A \mid |x_a| \geq \frac{1}{k}\}$ must be finite if $(x_a)_{a\in A}$ is summable (why?). Thus, since

$$\{a \in A \mid |x_a| \neq 0\} = \cup_{k\in\mathbb{Z}_{>0}}\{a \in A \mid |x_a| \geq \tfrac{1}{k}\},$$

the set $\{a \in A \mid x_a \neq 0\}$ is a countable union of finite sets, and so is countable by Proposition 1.7.16. ■

A legitimate question is, since a summable family reduces to essentially being countable, why should we bother with the idea at all? The reason is simply that it will be notationally convenient in Section **??**.

2.4.8 Notes

The numbers e and π are not only irrational, but have the much stronger property of being ***transcendental***. This means that they are not the roots of any polynomial having rational coefficients (see Definition **??**). That e is transcendental was proved by Hermite[7] in 1873, and the that π is transcendental was proved by Lindemann[8] in 1882.

The proof we give for the irrationality of π is essentially that of Niven [1947]; this is the most commonly encountered proof, and is simpler than the original proof of Lambert[9] presented to the Berlin Academy in 1768.

Exercises

2.4.1 Let $S = \sum_{j=1}^{\infty} x_j$ be a series in $\mathbb{R}$, and, for $j \in \mathbb{Z}_{>0}$, define

$$x_j^+ = \max\{x_j, 0\}, \quad x_j^- = \max\{0, -x_j\}.$$

Show that, if S is conditionally convergent, then the series $S^+ = \sum_{j=1}^{\infty} x_j^+$ and $S^- = \sum_{j=1}^{\infty} x_j^-$ diverge to ∞.

2.4.2 In this exercise we consider more carefully the paradox of Zeno given in Exercise 1.9.2. Let us attach some symbols to the relevant data, so that we can say useful things. Suppose that the tortoise travels with constant velocity v_t and that Achilles travels with constant velocity v_a. Suppose that the tortoise gets a head start of t_0 seconds.

(a) Compute directly using elementary physics (i.e., time/distance/velocity relations) the time at which Achilles will overtake the tortoise, and the distance both will have travelled during that time.

(b) Consider the sequences $(d_j)_{j\in\mathbb{Z}_{>0}}$ and $(t_j)_{j\in\mathbb{Z}_{>0}}$ defined so that

1. d_1 is the distance travelled by the tortoise during the head start time t_0,
2. t_j, $j \in \mathbb{Z}_{>0}$, is the time it takes Achilles to cover the distance d_j,
3. d_j, $j \geq 2$, is the distance travelled by the tortoise in time t_{j-1}.

Find explicit expressions for these sequences in terms of t_0, v_t, and v_a.

(c) Show that the series $\sum_{j=1}^{\infty} d_j$ and $\sum_{j=1}^{\infty} t_j$ converge, and compute their limits.

(d) What is the relationship between the limits of the series in part (c) and the answers to part (a).

[7]Charles Hermite (1822–1901) was a French mathematician who made contributions to the fields of number theory, algebra, differential equations, and analysis.

[8]Carl Louis Ferdinand von Lindemann (1852–1939) was born in what is now Germany. His mathematical contributions were in the areas of analysis and geometry. He also was interested in physics.

[9]Johann Heinrich Lambert (1728–1777) was born in France. His mathematical work included contributions to analysis, geometry, and probability. He also made contributions to astronomical theory.

(e) Does this shed some light on how to resolve Zeno's paradox?

2.4.3 Show that

$$\left|\sum_{j=1}^{m} x_j\right| \leq \sum_{j=1}^{m} |x_j|$$

for any finite family $(x_1, \ldots, x_m) \subseteq \mathbb{R}$.

2.4.4 State the correct version of Proposition 2.4.4 in the case that $S = \sum_{j=1}^{\infty} x_j$ is not absolutely convergent, and indicate why it is not a very interesting result.

2.4.5 For a sum

$$S = \sum_{j=1}^{\infty} s_j,$$

answer the following questions.

(a) Show that if S converges then the sequence $(s_j)_{j \in \mathbb{Z}_{>0}}$ converges to 0.

(b) Is the converse of part (a) true? That is to say, if the sequence $(s_j)_{j \in \mathbb{Z}_{>0}}$ converges to zero, does S converge? If this is true, prove it. If it is not true, give a counterexample.

2.4.6 Do the following.

(a) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j \to \infty} \left|\frac{x_{j+1}}{x_j}\right| = 1$ and which converges in $\mathbb{R}$.

(b) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j \to \infty} \left|\frac{x_{j+1}}{x_j}\right| = 1$ and which diverges to ∞.

(c) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j \to \infty} \left|\frac{x_{j+1}}{x_j}\right| = 1$ and which diverges to $-\infty$.

(d) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j \to \infty} \left|\frac{x_{j+1}}{x_j}\right| = 1$ and which is oscillatory.

2.4.7 Do the following.

(a) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j \to \infty} |x_j|^{1/j} = 1$ and which converges in $\mathbb{R}$.

(b) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j \to \infty} |x_j|^{1/j} = 1$ and which diverges to ∞.

(c) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j \to \infty} |x_j|^{1/j} = 1$ and which diverges to $-\infty$.

(d) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j \to \infty} |x_j|^{1/j} = 1$ and which is oscillatory.

The next exercise introduces the notion of the decimal expansion of a real number. An ***infinite decimal expansion*** is a series in $\mathbb{Q}$ of the form

$$\sum_{j=0}^{\infty} \frac{a_j}{10^j}$$

where $a_0 \in \mathbb{Z}$ and where $a_j \in \{0, 1, \ldots, 9\}$, $j \in \mathbb{Z}_{>0}$. An infinite decimal expansion is ***eventually periodic*** if there exists $k, m \in \mathbb{Z}_{>0}$ such that $a_{j+k} = a_j$ for all $j \geq m$.

2.4.8 (a) Show that the sequence of partial sums for an infinite decimal expansion is a Cauchy sequence.

(b) Show that, for every Cauchy sequence $(q_j)_{j\in\mathbb{Z}_{>0}}$, there exists a sequence $(d_j)_{j\in\mathbb{Z}_{>0}}$ of partial sums for a decimal expansion having the property that $[(q_j)_{j\in\mathbb{Z}_{>0}}] = [(d_j)_{j\in\mathbb{Z}_{>0}}]$ (the equivalence relation is that in the Cauchy sequences in $\mathbb{Q}$ as defined in Definition 2.1.16).

(c) Give an example that shows that two distinct infinite decimal expansions can be equivalent.

(d) Show that if two distinct infinite decimal expansions are equivalent, and if one of them is eventually periodic, then the other is also eventually periodic.

The previous exercises show that every real number is the limit of a (not necessarily unique) infinite decimal expansion. The next exercises characterise the infinite decimal expansions that correspond to rational numbers. First you will show that an eventually periodic decimal expansion corresponds to a rational number. Let $\sum_{j=0}^{\infty} \frac{a_j}{10^j}$ be an eventually periodic infinite decimal expansion and let $k, m \in \mathbb{Z}_{>0}$ have the property that $a_{j+k} = a_j$ for $j \geq m$. Denote by $x \in \mathbb{R}$ the number to which the infinite decimal expansion converges.

(e) Show that

$$10^{m+k}x = \sum_{j=0}^{\infty} \frac{b_j}{10^j}, \quad 10^m x = \sum_{j=0}^{\infty} \frac{c_j}{10^j}$$

are decimal expansions, and give expressions for b_j and c_j, $j \in \mathbb{Z}_{>0}$, in terms of a_j, $j \in \mathbb{Z}_{>0}$. In particular, show that $b_j = c_j$ for $j \geq 1$.

(f) Conclude that $(10^{m+k} - 10^m)x$ is an integer, and so x is therefore rational.

Next you will show that the infinite decimal expansion of a rational number is eventually periodic. Thus let $q \in \mathbb{Q}$.

(g) Let $q = \frac{a}{b}$ for $a, b \in \mathbb{Z}$ and with $b > 0$. For $j \in \{0, 1, \dots, b\}$, let $r_j \in \{0, 1, \dots, b-1\}$ satisfy $\frac{10^j}{b} = s_j + \frac{r_j}{b}$ for $s_j \in \mathbb{Z}$, i.e., r_j is the remainder after dividing 10^j by b. Show that at least two of the numbers $\{r_0, r_1, \dots, r_b\}$ must agree, i.e., conclude that $r_m = r_{m+k}$ for $k, m \in \mathbb{Z}_{\geq 0}$ satisfying $0 \leq m < m + k \leq b$.
Hint: *There are only* b *possible values for these* b + 1 *numbers.*

(h) Show that b exactly divides $10^{m+k} - 10^k$ with k and m as above. Thus $bc = 10^{m+k} - 10^k$ for some $c \in \mathbb{Z}$.

(i) Show that

$$\frac{a}{b} = 10^{-m} \frac{ac}{10^k - 1},$$

and so write

$$q = 10^{-m}\Big(s + \frac{r}{10^k - 1}\Big)$$

for $s \in \mathbb{Z}$ and $r \in \{0, 1, \dots, 10^k - 1\}$, i.e., r is the remainder after dividing ac by $10^k - 1$.

(j) Argue that we can write

$$b = \sum_{j=1}^{k} b_j 10^j,$$

for $b_j \in \{0, 1, \ldots, 9\}$, $j \in \{1, \ldots, k\}$.

(k) With b_j, $j \in \{1, \ldots, k\}$ as above, define an infinite decimal expansion $\sum_{j=0}^{\infty} \frac{a_j}{10^j}$ by asking that $a_0 = 0$, that $a_j = b_j$, $j \in \{1, \ldots, k\}$, and that $a_{j+km} = a_j$ for $j, m \in \mathbb{Z}_{>0}$. Let $d \in \mathbb{R}$ be the number to which this decimal expansion converges. Show that $(10^k - 1)d = b$, so $d \in \mathbb{Q}$.

(l) Show that $10^m q = s + d$, and so conclude that $10^m q$ has the eventually periodic infinite decimal expansion $s + \sum_{j=1}^{\infty} \frac{a_j}{10^j}$.

(m) Conclude that q has an eventually periodic infinite decimal expansion, and then conclude from (d) that any infinite decimal expansion for q is eventually periodic.

Section 2.5

Subsets of $\mathbb{R}$

In this section we study in some detail the nature of various sorts of subsets of $\mathbb{R}$. The character of these subsets will be of some importance when we consider the properties of functions defined on $\mathbb{R}$, and/or taking values in $\mathbb{R}$. Our presentation also gives us an opportunity to introduce, in a fairly simple setting, some concepts that will appear later in more abstract settings, e.g., open sets, closed sets, compactness.

Do I need to read this section? Unless you know the material here, it is indeed a good idea to read this section. Many of the ideas are basic, but some are not (e.g., the Heine–Borel Theorem). Moreover, many of the not-so-basic ideas will appear again later, particularly in Chapter **??**, and if a reader does not understand the ideas in the simple case of $\mathbb{R}$, things will only get more difficult. Also, the ideas expressed here will be essential in understanding even basic things about signals as presented in Chapter **??**. •

2.5.1 Open sets, closed sets, and intervals

One of the basic building blocks in the understanding of the real numbers is the idea of an open set. In this section we define open sets and some related notions, and provide some simple properties associated to these ideas.

First, it is convenient to introduce the following ideas.

2.5.1 Definition (Open ball, closed ball) For $r \in \mathbb{R}_{>0}$ and $x_0 \in \mathbb{R}$,

(i) the ***open ball*** in $\mathbb{R}$ of radius r about x_0 is the set

$$\mathsf{B}(r, x_0) = \{x \in \mathbb{R} \mid |x - x_0| < r\},$$

and

(ii) the ***closed ball*** of radius r about x_0 is the set

$$\overline{\mathsf{B}}(r, x_0) = \{x \in \mathbb{R} \mid |x - x_0| \leq r\}.$$ •

These sets are simple to understand, and we depict them in Figure 2.3. With

Figure 2.3 An open ball (left) and a closed ball (right) in $\mathbb{R}$

the notion of an open ball, it is easy to give some preliminary definitions.

2.5.2 Definition (Open and closed sets in $\mathbb{R}$) A set $A \subseteq \mathbb{R}$ is:

(i) ***open*** if, for every $x \in A$, there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}(\epsilon, x) \subseteq A$ (the empty set is also open, by declaration);

(ii) ***closed*** if $\mathbb{R} \setminus A$ is open. •

A trivial piece of language associated with an open set is the notion of a neighbourhood.

2.5.3 Definition (Neighbourhood in $\mathbb{R}$) A ***neighbourhood*** of an element $x \in \mathbb{R}$ is an open set U for which $x \in U$. •

Some authors allow a "neighbourhood" to be a set A which contains a neighbourhood in our sense. Such authors will then frequently call what we call a neighbourhood an "open neighbourhood."

Let us give some examples of sets that are open, closed, or neither. The examples we consider here are important ones, since they are all examples of ***intervals***, which will be of interest at various times, and for various reasons, throughout these volumes. In particular, the notation we introduce here for intervals will be used a great deal.

2.5.4 Examples (Intervals)

1. For $a, b \in \mathbb{R}$ with $a < b$ the set
$$(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$$
is open. Indeed, let $x \in (a, b)$ and let $\epsilon = \frac{1}{2}\min\{b - x, x - a\}$. It is then easy to see that $\mathsf{B}(\epsilon, x) \subseteq (a, b)$. If $a \geq b$ we take the convention that $(a, b) = \emptyset$.
2. For $a \in \mathbb{R}$ the set
$$(a, \infty) = \{x \in \mathbb{R} \mid a < x\}$$
is open. For example, if $x \in (a, \infty)$ then, if we define $\epsilon = \frac{1}{2}(x - a)$, we have $\mathsf{B}(\epsilon, x) \subseteq (a, \infty)$.
3. For $b \in \mathbb{R}$ the set
$$(-\infty, b) = \{x \in \mathbb{R} \mid x < b\}$$
is open.
4. For $a, b \in \mathbb{R}$ with $a \leq b$ the set
$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$$
is closed. Indeed, $\mathbb{R} \setminus [a, b] = (-\infty, a) \cup (b, \infty)$. The sets $(-\infty, a)$ and (b, ∞) are both open, as we have already seen. Moreover, it is easy to see, directly from the definition, that the union of open sets is also an open set. Therefore, $\mathbb{R} \setminus [a, b]$ is open, and so $[a, b]$ is closed.
5. For $a \in \mathbb{R}$ the set
$$[a, \infty) = \{x \in \mathbb{R} \mid a \leq x\}$$
is closed since it complement in $\mathbb{R}$ is $(-\infty, a)$ which is open.

6. For $b \in \mathbb{R}$ the set
$$(-\infty, b] = \{x \in \mathbb{R} \mid \ x \leq b\}$$
is closed.
7. For $a, b \in \mathbb{R}$ with $a < b$ the set
$$(a, b] = \{x \in \mathbb{R} \mid \ a < x \leq b\}$$
is neither open nor closed. To see that it is not open, note that $b \in (a, b]$, but that any open ball about b will contain points not in $(a, b]$. To see that $(a, b]$ is not closed, note that $a \in \mathbb{R} \setminus (a, b]$, and that any open ball about a will contain points not in $\mathbb{R} \setminus (a, b]$.
8. For $a, b \in \mathbb{R}$ with $a < b$ the set
$$[a, b) = \{x \in \mathbb{R} \mid \ a \leq x < b\}$$
is neither open nor closed.
9. The set $\mathbb{R}$ is both open and closed. That it is open is clear. That it is closed follows since $\mathbb{R} \setminus \mathbb{R} = \emptyset$, and $\emptyset$ is, by convention, open. We will sometimes, although not often, write $\mathbb{R} = (-\infty, \infty)$. •

We shall frequently denote typical interval by I, and the set of intervals we denote by $\mathscr{I}$. If I and J are intervals with $J \subseteq I$, we will say that J is a ***subinterval*** of I. The expressions "open interval" and "closed interval" have their natural meanings as intervals that are, as subsets of $\mathbb{R}$, open and closed, respectively. An interval that is neither open nor closed will be called ***half-open*** or ***half-closed***. A ***left endpoint*** (resp. ***right endpoint***) for an interval I is a number $x \in \mathbb{R}$ such that $\inf I = x$ (resp. $\sup I = x$). An endpoint x, be it left or right, is ***open*** if $x \notin I$ and is ***closed*** if $x \in I$. If $\inf I = -\infty$ (resp. $\sup I = \infty$), then we saw that I is ***unbounded on the left*** (resp. ***unbounded on the right***). We will also use the interval notation to denote subsets of the extended real numbers $\overline{\mathbb{R}}$. Thus, we may write

1. $(a, \infty] = (a, \infty) \cup \{\infty\}$,
2. $[a, \infty] = [a, \infty) \cup \{\infty\}$,
3. $[-\infty, b) = (-\infty, b) \cup \{-\infty\}$,
4. $[-\infty, b] = (-\infty, b] \cup \{-\infty\}$, and
5. $[-\infty, \infty] = (-\infty, \infty) \cup \{-\infty, \infty\} = \overline{\mathbb{R}}$.

The following characterisation of intervals is useful.

2.5.5 Proposition (Characterisation of intervals) *A subset* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval if and only if, for each* $\mathrm{a}, \mathrm{b} \in \mathrm{I}$ *with* $\mathrm{a} < \mathrm{b}$, $[\mathrm{a}, \mathrm{b}] \subseteq \mathrm{I}$.

Proof It is clear from the definition that, if I is an interval, then, for each $a, b \in I$ with $a < b$, $[a, b] \subseteq I$. So suppose that, for each $a, b \in I$ with $a < b$, $[a, b] \subseteq I$. Let $A = \inf I$ and let $B = \sup I$. We have the following cases to consider.

1. $A = B$: Trivially I is an interval.

2. $A, B \in \mathbb{R}$ and $A \neq B$: Choose $a_1, b_1 \in I$ such that $a_1 < b_1$. Define $a_{j+1}, b_{j+1} \in I$, $j \in \mathbb{Z}_{>0}$, inductively as follows. Let a_{j+1} be a point in I to the left of $\frac{1}{2}(A + a_j)$ and let b_{j+1} be a point in I to the right of $\frac{1}{2}(b_j + B)$. These constructions make sense by definition of A and B. Note that $(a_j)_{j\in\mathbb{Z}_{>0}}$ is a monotonically decreasing sequence converging to A and that $(b_j)_{j\in\mathbb{Z}_{>0}}$ is a monotonically increasing sequence converging to B. Also,

$$\bigcup_{j\in\mathbb{Z}_{>0}} [a_j, b_j] \subseteq I.$$

   We also have either $\cup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = (A, B)$, $\cup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = [A, B)$, $\cup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = (A, B]$, or $\cup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = [A, B]$. Therefore we conclude that I is an interval with endpoints A and B.

3. $A = -\infty$ and $B \in \mathbb{R}$. Choose $a_1, b_1 \in I$ with $a_a < b_1 < B$. Define $a_{j+1}, b_{j+1} \in I$, $j \in \mathbb{Z}_{>0}$, inductively by asking that a_{j+1} be a point in I to the left of $a_j - 1$ and that b_{j+1} be a point in I to the right of $\frac{1}{2}(b_j + B)$. These constructions make sense by definition of A and B. Thus $(a_j)_{j\in\mathbb{Z}_{>0}}$ is a monotonically decreasing sequence in I diverging to $-\infty$ and $(b_j)_{j\in\mathbb{Z}_{>0}}$ is a monotonically increasing sequence in I converging to B. Thus

$$\bigcup_{j\in\mathbb{Z}_{>0}} [a_j, b_j] =\subseteq I.$$

   Note that either $\bigcup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = (-\infty, B)$ or $\bigcup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = (-\infty, B]$. This means that either $I = (-\infty, B)$ or $I = (-\infty, B]$.

4. $A \in \mathbb{R}$ and $B = \infty$: A construction entirely like the preceding one shows that either $I = (A, \infty)$ or $I = [A, \infty)$.

5. $A = -\infty$ and $B = \infty$: Choose $a_1, b_1 \in I$ with $a_1 < b_1$. Inductively define $a_{j+1}, b_{j+1} \in I$, $j \in \mathbb{Z}_{>0}$, by asking that a_{j+1} be a point in I to the left of a_j and that b_{j+1} be a point in I to the right of b_j. We then conclude that

$$\bigcup_{j\in\mathbb{Z}_{>0}} [a_j, b_j] = \mathbb{R} =\subseteq I,$$

   and so $I = \mathbb{R}$.

In all cases we have concluded that I is an interval. ■

The following property of open sets will be useful for us, and tells us a little about the character of open sets.

2.5.6 Proposition (Open sets in $\mathbb{R}$ are unions of open intervals) *If $\mathrm{U} \subseteq \mathbb{R}$ is a nonempty open set then U is a countable union of disjoint open intervals.*

Proof Let $x \in U$ and let I_x be the largest open interval containing x and contained in U. This definition of I_x makes sense since the union of open intervals containing x is also an open interval containing x. Now to each interval can be associated a rational number within the interval. Therefore, the number of intervals to cover U can be associated with a subset of $\mathbb{Q}$, and is therefore countable or finite. This shows that U is indeed a finite or countable union of open intervals. ■

2.5.2 Partitions of intervals

In this section we consider the idea of partitioning an interval of the form $[a, b]$. This is a construction that will be useful in a variety of places, but since we dealt with intervals in the previous section, this is an appropriate time to make the definition and the associated constructions.

2.5.7 Definition (Partition of an interval) A ***partition*** of an interval $[a, b]$ is a family $(I_1, \dots, I_k)$ of intervals such that

(i) $\text{int}(I_j) \neq \emptyset$ for $j \in \{1, \dots, k\}$,

(ii) $[a, b] = \cup_{j=1}^{k} I_j$, and

(iii) $I_j \cap I_l = \emptyset$ for $j \neq l$.

We denote by $\text{Part}([a, b])$ the set of partitions of $[a, b]$. •

We shall always suppose that a partition $(I_1, \dots, I_k)$ is totally ordered so that the left endpoint of I_{j+1} agrees with the right endpoint of I_j for each $j \in \{1, \dots, k-1\}$. That is to say, when we write a partition, we shall list the elements of the set according to this total order. Note that associated to a partition $(I_1, \dots, I_k)$ are the endpoints of the intervals. Thus there exists a family $(x_0, x_1, \dots, x_k)$ of $[a, b]$, ordered with respect to the natural total order on $\mathbb{R}$, such that, for each $j \in \{1, \dots, k\}$, x_{j-1} is the left endpoint of I_j and x_j is the right endpoint of I_j. Note that necessarily we have $x_0 = a$ and $x_k = b$. The set of endpoints of the intervals in a partition $P = (I_1, \dots, I_k)$ we denote by $\text{EP}(P)$. In Figure 2.4 we show a partition with all

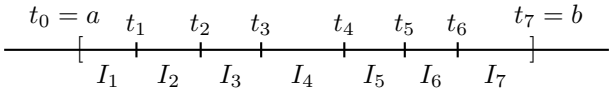


Figure 2.4 A partition

ingredients labelled. For a partition P with $\text{EP}(P) = (x_0, x_1, \dots, x_k)$, denote

$$|P| = \max\{|x_j - x_l| \mid j, l \in \{1, \dots, k\}\},$$

which is the ***mesh*** of P. Thus $|P|$ is the length of the largest interval of the partition.

It is often useful to be able to say one partition is finer than another, and the following definition makes this precise.

2.5.8 Definition (Refinement of a partition) If P_1 and P_2 are partitions of an interval $[a, b]$, then P_2 is a ***refinement*** of P_1 if $\text{EP}(P_1) \subseteq \text{EP}(P_2)$. •

Next we turn to a sometimes useful construction involving the addition of certain structure onto a partition. This construction is rarely used in the text, so may be skipped until it is encountered.

2.5.9 Definition (Tagged partition, δ-fine tagged partition) Let $[a, b]$ be an interval and let $\delta\colon [a, b] \to \mathbb{R}_{>0}$.

(i) A ***tagged partition*** of $[a, b]$ is a finite family of pairs $((c_1, I_1), \ldots, (c_k, I_k))$ where $(I_1, \ldots, I_k)$ is a partition and where c_j is contained in the union of I_j with its endpoints.

(ii) A tagged partition $((c_1, I_1), \ldots, (c_k, I_k))$ is ***δ-fine*** if the interval I_j, along with its endpoints, is a subset of $\mathsf{B}(\delta(c_j), c_j)$. •

The following result asserts that δ-fine tagged partitions always exist.

2.5.10 Proposition (δ-fine tagged partitions exist) *For any positive function $\delta\colon [a, b] \to \mathbb{R}_{>0}$, there exists a δ-fine tagged partition.*

Proof Let Δ be the set of all points $x \in (a, b]$ such that there exists a δ-fine tagged partition of $[a, x]$. Note that $(a, a + \delta(a)) \subseteq \Delta$ since, for each $x \in (a, a + \delta(a))$, $((a, [a, x]))$ is a δ-fine tagged partition of $[a, x]$. Let $b' = \sup \Delta$. We will show that $b' = b$ and that $b' \in \Delta$.

Since $b' = \sup \Delta$ there exists $b'' \in \Delta$ such that $b' - \delta(b') < b'' < b'$. Then there exists a δ-fine partition P' of $[a, b']$. Now $P' \cup ((b', (b'', b']))$ is δ-fine tagged partition of $[a, b']$. Thus $b' \in \Delta$.

Now suppose that $b' < b$ and choose $b'' < b$ such that $b' < b'' < b' + \delta(b')$. If P is a tagged partition of $[a, b']$ (this exists since $b' \in \Delta$), then $P \cup ((b', (b', b'']))$ is a δ-fine tagged partition of $[a, b'']$. This contradicts the fact that $b' = \sup \Delta$. Thus we conclude that $b' = b$. ■

2.5.3 Interior, closure, boundary, and related notions

Associated with the concepts of open and closed are a collection of useful concepts.

2.5.11 Definition (Accumulation point, cluster point, limit point in $\mathbb{R}$) Let $A \subseteq \mathbb{R}$. A point $x \in \mathbb{R}$ is:

(i) an ***accumulation point*** for A if, for every neighbourhood U of x, the set $A \cap (U \setminus \{x\})$ is nonempty;

(ii) a ***cluster point*** for A if, for every neighbourhood U of x, the set $A \cap U$ is infinite;

(iii) a ***limit point*** of A if there exists a sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ in A converging to x.

The set of accumulation points of A is called the ***derived set*** of A, and is denoted by $\operatorname{der}(A)$. •

2.5.12 Remark (Conventions concerning "accumulation point," "cluster point," and "limit point") There seems to be no agreed upon convention about what is meant by the three concepts of accumulation point, cluster point, and limit point. Some authors make no distinction between the three concepts at all. Some authors lump two together, but give the third a different meaning. As we shall see in Proposition 2.5.13 below, sometimes there is no need to distinguish between two of the concepts. However, in order to keep as clear as possible the transition to

the more abstract presentation of Chapter **??**, we have gone with the most pedantic interpretation possible for the concepts of "accumulation point," "cluster point," and "limit point." •

The three concepts of accumulation point, cluster point, and limit point are actually excessive for $\mathbb{R}$ since, as the next result shall indicate, two of them are exactly the same. However, in the more general setup of Chapter **??**, the concepts are no longer equivalent.

2.5.13 Proposition ("Accumulation point" equals "cluster point" in $\mathbb{R}$) *For a set* $A \subseteq \mathbb{R}$, $x \in \mathbb{R}$ *is an accumulation point for* A *if and only if it is a cluster point for* A.

Proof It is clear that a cluster point for A is an accumulation point for A. Suppose that x is not a cluster point. Then there exists a neighbourhood U of x for which the set $A \cap U$ is finite. If $A \cap U = \{x\}$, then clearly x is not an accumulation point. If $A \cap U \neq \{x\}$, then $A \cap (U \setminus \{x\}) \supseteq \{x_1, \dots, x_k\}$ where the points $x_1, \dots, x_k$ are distinct from x. Now let

$$\epsilon = \tfrac{1}{2} \min\{|x_1 - x|, \dots, |x_k - x|\}.$$

Clearly $A \cap (\mathsf{B}(\epsilon, x) \setminus \{x\})$ is then empty, and so x is not an accumulation point for A. ■

Now let us give some examples that illustrate the differences between accumulation points (or equivalently cluster points) and limit points.

2.5.14 Examples (Accumulation points and limit points)

1. For any subset $A \subseteq \mathbb{R}$ and for every $x \in A$, x is a limit point for A. Indeed, the constant sequence $(x_j = x)_{j \in \mathbb{Z}_{>0}}$ is a sequence in A converging to x. However, as we shall see in the examples to follow, it is not the case that all points in A are accumulation points.
2. Let $A = (0, 1)$. The set of accumulation points of A is then easily seen to be $[0, 1]$. The set of limit points is also $[0, 1]$.
3. Let $A = [0, 1)$. Then, as in the preceding example, both the set of accumulation points and the set of limit points are the set $[0, 1]$.
4. Let $A = [0, 1] \cup \{2\}$. Then the set of accumulation points is $[0, 1]$ whereas the set of limit points is A.
5. Let $A = \mathbb{Q}$. One can readily check that the set of accumulation points of A is $\mathbb{R}$ and the set of limit points of A is also $\mathbb{R}$. •

The following result gives some properties of the derived set.

2.5.15 Proposition (Properties of the derived set in $\mathbb{R}$) *For* $\mathrm{A}, \mathrm{B} \subseteq \mathbb{R}$ *and for a family of subsets* $(\mathrm{A}_i)_{i \in I}$ *of* $\mathbb{R}$, *the following statements hold:*

(i) $\operatorname{der}(\emptyset) = \emptyset$;

(ii) $\operatorname{der}(\mathbb{R}) = \mathbb{R}$;

(iii) $\operatorname{der}(\operatorname{der}(\mathrm{A})) = \operatorname{der}(\mathrm{A})$;

(iv) *if* $\mathrm{A} \subseteq \mathrm{B}$ *then* $\operatorname{der}(\mathrm{A}) \subseteq \operatorname{der}(\mathrm{B})$;

(v) $\operatorname{der}(\mathrm{A} \cup \mathrm{B}) = \operatorname{der}(\mathrm{A}) \cup \operatorname{der}(\mathrm{B})$;

(vi) $\text{der}(A \cap B) \subseteq \text{der}(A) \cap \text{der}(B)$.

Proof Parts (i) and (ii) follow directly from the definition of the derived set.

(iii) ***missing stuff***

(iv) Let $x \in \text{der}(A)$ and let U be a neighbourhood of x. Then the set $A \cap (U \setminus \{x\})$ is nonempty, implying that the set $B \cap (U \setminus \{x\})$ is also nonempty. Thus $x \in \text{der}(B)$.

(v) Let $x \in \text{der}(A \cup B)$ and let U be a neighbourhood of x. Then the set $U \cap ((A \cup B) \setminus \{x\})$ is nonempty. But

$$\begin{aligned} U \cap ((A \cup B) \setminus \{x\}) &= U \cap ((A \setminus \{x\}) \cup (B \setminus \{x\})) \\ &= (U \cap (A \setminus \{x\})) \cup (U \cap (B \setminus \{x\})). \end{aligned} \tag{2.8}$$

Thus it cannot be that both $U \cap (A \setminus \{x\})$ and $U \cap (B \setminus \{x\})$ are empty. Thus x is an element of either $\text{der}(A)$ or $\text{der}(B)$.

Now let $x \in \text{der}(A) \cup \text{der}(A)$. Then, using (2.8), $U \cap ((A \cup B) \setminus \{x\})$ is nonempty, and so $x \in \text{der}(A \cup B)$.

(vi) Let $x \in \text{der}(A \cap B)$ and let U be a neighbourhood of x. Then $U \cap ((A \cap B) \setminus \{x\}) \neq \emptyset$. We have

$$U \cap ((A \cap B) \setminus \{x\}) = U \cap ((A \setminus \{x\}) \cap (B \setminus \{x\}))$$

Thus the sets $U \cap (A \setminus \{x\})$ and $U \cap (B \setminus \{x\})$ are both nonempty, showing that $x \in \text{der}(A) \cap \text{der}(B)$. ■

Next we turn to characterising distinguished subsets of subsets of $\mathbb{R}$.

2.5.16 Definition (Interior, closure, and boundary in $\mathbb{R}$) Let $A \subseteq \mathbb{R}$.

(i) The ***interior*** of A is the set

$$\text{int}(A) = \cup\{U \mid U \subseteq A,\ U \text{ open}\}.$$

(ii) The ***closure*** of A is the set

$$\text{cl}(A) = \cap\{C \mid A \subseteq C,\ C \text{ closed}\}.$$

(iii) The ***boundary*** of A is the set $\text{bd}(A) = \text{cl}(A) \cap \text{cl}(\mathbb{R} \setminus A)$. •

In other words, the interior of A is the largest open set contained in A. Note that this definition makes sense since a union of open sets is open (Exercise 2.5.1). In like manner, the closure of A is the smallest closed set containing A, and this definition makes sense since an intersection of closed sets is closed (Exercise 2.5.1 again). Note that $\text{int}(A)$ is open and $\text{cl}(A)$ is closed. Moreover, since $\text{bd}(A)$ is the intersection of two closed sets, it too is closed (Exercise 2.5.1 yet again).

Let us give some examples of interiors, closures, and boundaries.

2.5.17 Examples (Interior, closure, and boundary)

1. Let $A = \text{int}(0, 1)$. Then $\text{int}(A) = (0, 1)$ since A is open. We claim that $\text{cl}(A) = [0, 1]$. Clearly $[0, 1] \subseteq \text{cl}(A)$ since $[0, 1]$ is closed and contains A. Moreover, the only smaller subsets contained in $[0, 1]$ and containing A are $[0, 1)$, $(0, 1]$, and $(0, 1)$, none of which are closed. We may then conclude that $\text{cl}(A) = [0, 1]$. Finally we claim that $\text{bd}(A) = \{0, 1\}$. To see this, note that we have $\text{cl}(A) = [0, 1]$ and $\text{cl}(\mathbb{R} \setminus A) = (-\infty, 0] \cup [1, \infty)$ (by an argument like that used to show that $\text{cl}(A) = [0, 1]$). Therefore, $\text{bd}(A) = \text{cl}(A) \cap \text{cl}(\mathbb{R} \setminus A) = \{0, 1\}$, as desired.

2. Let $A = [0,1]$. Then $\operatorname{int}(A) = (0,1)$. To see this, we note that $(0,1) \subseteq \operatorname{int}(A)$ since $(0,1)$ is open and contained in A. Moreover, the only larger sets contained in A are $[0,1)$, $(0,1]$, and $[0,1]$, none of which are open. Thus $\operatorname{int}(A) = (0,1)$, just as claimed. Since A is closed, $\operatorname{cl}(A) = A$. Finally we claim that $\operatorname{bd}(A) = \{0,1\}$. Indeed, $\operatorname{cl}(A) = [0,1]$ and $\operatorname{cl}(\mathbb{R} \setminus A) = (-\infty, 0] \cup [1, \infty)$. Therefore, $\operatorname{bd}(A) = \operatorname{cl}(A) \cap \operatorname{cl}(\mathbb{R} \setminus A) = \{0,1\}$, as claimed.
3. Let $A = (0,1) \cup \{2\}$. We have $\operatorname{int}(A) = (0,1)$, $\operatorname{cl}(A) = [0,1] \cup \{2\}$, and $\operatorname{bd}(A) = \{0,1,2\}$. We leave the simple details of these assertions to the reader.
4. Let $A = \mathbb{Q}$. One readily ascertains that $\operatorname{int}(A) = \emptyset$, $\operatorname{cl}(A) = \mathbb{R}$, and $\operatorname{bd}(A) = \mathbb{R}$. •

Now let us give a characterisation of interior, closure, and boundary that are often useful in practice. Indeed, we shall often use these characterisations without explicitly mentioning that we are doing so.

2.5.18 Proposition (Characterisation of interior, closure, and boundary in $\mathbb{R}$) *For $A \subseteq \mathbb{R}$, the following statements hold:*

(i) $x \in \operatorname{int}(A)$ *if and only if there exists a neighbourhood U of x such that $U \subseteq A$;*

(ii) $x \in \operatorname{cl}(A)$ *if and only if, for each neighbourhood U of x, the set $U \cap A$ is nonempty;*

(iii) $x \in \operatorname{bd}(A)$ *if and only if, for each neighbourhood U of x, the sets $U \cap A$ and $U \cap (\mathbb{R} \setminus A)$ are nonempty.*

Proof (i) Suppose that $x \in \operatorname{int}(A)$. Since $\operatorname{int}(A)$ is open, there exists a neighbourhood U of x contained in $\operatorname{int}(A)$. Since $\operatorname{int}(A) \subseteq A$, $U \subseteq A$.

Next suppose that $x \notin \operatorname{int}(A)$. Then, by definition of interior, for any open set U for which $U \subseteq A$, $x \notin U$.

(ii) Suppose that there exists a neighbourhood U of x such that $U \cap A = \emptyset$. Then $\mathbb{R} \setminus U$ is a closed set containing A. Thus $\operatorname{cl}(A) \subseteq \mathbb{R} \setminus U$. Since $x \notin \mathbb{R} \setminus U$, it follows that $x \notin \operatorname{cl}(A)$.

Suppose that $x \notin \operatorname{cl}(A)$. Then x is an element of the open set $\mathbb{R} \setminus \operatorname{cl}(A)$. Thus there exists a neighbourhood U of x such that $U \subseteq \mathbb{R} \setminus \operatorname{cl}(A)$. In particular, $U \cap A = \emptyset$.

(iii) This follows directly from part (ii) and the definition of boundary. ■

Now let us state some useful properties of the interior of a set.

2.5.19 Proposition (Properties of interior in $\mathbb{R}$) *For $A, B \subseteq \mathbb{R}$ and for a family of subsets $(A_i)_{i \in I}$ of $\mathbb{R}$, the following statements hold:*

(i) $\operatorname{int}(\emptyset) = \emptyset$;

(ii) $\operatorname{int}(\mathbb{R}) = \mathbb{R}$;

(iii) $\operatorname{int}(\operatorname{int}(A)) = \operatorname{int}(A)$;

(iv) *if* $A \subseteq B$ *then* $\operatorname{int}(A) \subseteq \operatorname{int}(B)$;

(v) $\operatorname{int}(A \cup B) \supseteq \operatorname{int}(A) \cup \operatorname{int}(B)$;

(vi) $\operatorname{int}(A \cap B) = \operatorname{int}(A) \cap \operatorname{int}(B)$;

(vii) $\operatorname{int}(\cup_{i \in I} A_i) \supseteq \cup_{i \in I} \operatorname{int}(A_i)$;

(viii) $\operatorname{int}(\cap_{i \in I} A_i) \subseteq \cap_{i \in I} \operatorname{int}(A_i)$.

Moreover, a set $A \subseteq \mathbb{R}$ is open if and only if $\operatorname{int}(A) = A$.

Proof Parts (i) and (ii) are clear by definition of interior. Part (v) follows from part (vii), so we will only prove the latter.

(iii) This follows since the interior of an open set is the set itself.

(iv) Let $x \in \operatorname{int}(A)$. Then there exists a neighbourhood U of x such that $U \subseteq A$. Thus $U \subseteq B$, and the result follows from Proposition 2.5.18.

(vi) Let $x \in \operatorname{int}(A) \cap \operatorname{int}(B)$. Since $\operatorname{int}(A) \cap \operatorname{int}(B)$ is open by Exercise 2.5.1, there exists a neighbourhood U of x such that $U \subseteq \operatorname{int}(A) \cap \operatorname{int}(B)$. Thus $U \subseteq A \cap B$. This shows that $x \in \operatorname{int}(A \cap B)$. This part of the result follows from part (viii).

(vii) Let $x \in \cup_{i\in I} \operatorname{int}(A_i)$. By Exercise 2.5.1 the set $\cup_{i\in I} \operatorname{int}(A_i)$ is open. Thus there exists a neighbourhood U of x such that $U \subseteq \cup_{i\in I} \operatorname{int}(A_i)$. Thus $U \subseteq \cup_{i\in I} A_i$, from which we conclude that $x \in \operatorname{int}(\cup_{i\in I} A_i)$.

(viii) Let $x \in \operatorname{int}(\cap_{i\in I} A_i)$. Then there exists a neighbourhood U of x such that $U \subseteq \cap_{i\in I} A_i$. It therefore follows that $U \subseteq A_i$ for each $i \in I$, and so that $x \in \operatorname{int}(A_i)$ for each $i \in I$.

The final assertion follows directly from Proposition 2.5.18. ■

Next we give analogous results for the closure of a set.

2.5.20 Proposition (Properties of closure in $\mathbb{R}$) *For* $\mathrm{A}, \mathrm{B} \subseteq \mathbb{R}$ *and for a family of subsets* $(\mathrm{A}_i)_{i\in I}$ *of* $\mathbb{R}$*, the following statements hold:*

(i) $\operatorname{cl}(\emptyset) = \emptyset$;

(ii) $\operatorname{cl}(\mathbb{R}) = \mathbb{R}$;

(iii) $\operatorname{cl}(\operatorname{cl}(\mathrm{A})) = \operatorname{cl}(\mathrm{A})$;

(iv) *if* $\mathrm{A} \subseteq \mathrm{B}$ *then* $\operatorname{cl}(\mathrm{A}) \subseteq \operatorname{cl}(\mathrm{B})$;

(v) $\operatorname{cl}(\mathrm{A} \cup \mathrm{B}) = \operatorname{cl}(\mathrm{A}) \cup \operatorname{cl}(\mathrm{B})$;

(vi) $\operatorname{cl}(\mathrm{A} \cap \mathrm{B}) \subseteq \operatorname{cl}(\mathrm{A}) \cap \operatorname{cl}(\mathrm{B})$;

(vii) $\operatorname{cl}(\cup_{i\in I} \mathrm{A}_i) \supseteq \cup_{i\in I} \operatorname{cl}(\mathrm{A}_i)$;

(viii) $\operatorname{cl}(\cap_{i\in I} \mathrm{A}_i) \subseteq \cap_{i\in I} \operatorname{cl}(\mathrm{A}_i)$.

Moreover, a set $\mathrm{A} \subseteq \mathbb{R}$ *is closed if and only if* $\operatorname{cl}(\mathrm{A}) = \mathrm{A}$.

Proof Parts (i) and (ii) follow immediately from the definition of closure. Part (vi) follows from part (viii), so we will only prove the latter.

(iii) This follows since the closure of a closed set is the set itself.

(iv) Suppose that $x \in \operatorname{cl}(A)$. Then, for any neighbourhood U of x, the set $U \cap A$ is nonempty, by Proposition 2.5.18. Since $A \subseteq B$, it follows that $U \cap B$ is also nonempty, and so $x \in \operatorname{cl}(B)$.

(v) Let $x \in \operatorname{cl}(A \cup B)$. Then, for any neighbourhood U of x, the set $U \cap (A \cup B)$ is nonempty by Proposition 2.5.18. By Proposition 1.1.4, $U \cap (A \cup B) = (U \cap A) \cup (U \cap B)$. Thus the sets $U \cap A$ and $U \cap B$ are not both nonempty, and so $x \in \operatorname{cl}(A) \cup \operatorname{cl}(B)$. That $\operatorname{cl}(A) \cup \operatorname{cl}(B) \subseteq \operatorname{cl}(A \cup B)$ follows from part (vii).

(vi) Let $x \in \operatorname{cl}(A \cap B)$. Then, for any neighbourhood U of x, the set $U \cap (A \cap B)$ is nonempty. Thus the sets $U \cap A$ and $U \cap B$ are nonempty, and so $x \in \operatorname{cl}(A) \cap \operatorname{cl}(B)$.

(vii) Let $x \in \cup_{i\in I} \operatorname{cl}(A_i)$ and let U be a neighbourhood of x. Then, for each $i \in I$, $U \cap A_i \neq \emptyset$. Therefore, $\cup_{i\in I}(U \cap A_i) \neq \emptyset$. By Proposition 1.1.7, $\cup_{i\in I}(U \cap A_i) = U \cap (\cup_{i\in I} A_i)$, showing that $U \cap (\cup_{i\in I} A_i) \neq \emptyset$. Thus $x \in \operatorname{cl}(\cup_{i\in I} A_i)$.

(viii) Let $x \in \operatorname{cl}(\cap_{i\in I} A_i)$ and let U be a neighbourhood of x. Then the set $U \cap (\cap_{i\in I} A_i)$ is nonempty. This means that, for each $i \in I$, the set $U \cap A_i$ is nonempty. Thus $x \in \operatorname{cl}(A_i)$ for each $i \in I$, giving the result. ■

Note that there is a sort of "duality" between int and cl as concerns their interactions with union and intersection. This is reflective of the fact that open and closed sets themselves have such a "duality," as can be seen from Exercise 2.5.1. We refer the reader to Exercise 2.5.4 to construct counterexamples to any missing opposite inclusions in Propositions 2.5.19 and 2.5.20.

Let us state some relationships between certain of the concepts we have thus far introduced.

2.5.21 Proposition (Joint properties of interior, closure, boundary, and derived set in $\mathbb{R}$) *For $A \subseteq \mathbb{R}$, the following statements hold:*

(i) $\mathbb{R} \setminus \operatorname{int}(A) = \operatorname{cl}(\mathbb{R} \setminus A)$;

(ii) $\mathbb{R} \setminus \operatorname{cl}(A) = \operatorname{int}(\mathbb{R} \setminus A)$.

(iii) $\operatorname{cl}(A) = A \cup \operatorname{bd}(A)$;

(iv) $\operatorname{int}(A) = A - \operatorname{bd}(A)$;

(v) $\operatorname{cl}(A) = \operatorname{int}(A) \cup \operatorname{bd}(A)$;

(vi) $\operatorname{cl}(A) = A \cup \operatorname{der}(A)$;

(vii) $\mathbb{R} = \operatorname{int}(A) \cup \operatorname{bd}(A) \cup \operatorname{int}(\mathbb{R} \setminus A)$.

Proof (i) Let $x \in \mathbb{R} \setminus \operatorname{int}(A)$. Since $x \notin \operatorname{int}(A)$, for every neighbourhood U of x it holds that $U \not\subset A$. Thus, for any neighbourhood U of x, we have $U \cap (\mathbb{R} \setminus A) \neq \emptyset$, showing that $x \in \operatorname{cl}(\mathbb{R} \setminus A)$.

Now let $x \in \operatorname{cl}(\mathbb{R} \setminus A)$. Then for any neighbourhood U of x we have $U \cap (\mathbb{R} \setminus A) \neq \emptyset$. Thus $x \notin \operatorname{int}(A)$, so $x \in \mathbb{R} \setminus A$.

(ii) The proof here strongly resembles that for part (i), and we encourage the reader to provide the explicit arguments.

(iii) This follows from part (v).

(iv) Clearly $\operatorname{int}(A) \subseteq A$. Suppose that $x \in A \cap \operatorname{bd}(A)$. Then, for any neighbourhood U of x, the set $U \cap (\mathbb{R} \setminus A)$ is nonempty. Therefore, no neighbourhood of x is a subset of A, and so $x \notin \operatorname{int}(A)$. Conversely, if $x \in \operatorname{int}(A)$ then there is a neighbourhood U of x such that $U \subseteq A$. The precludes the set $U \cap (\mathbb{R} \setminus A)$ from being nonempty, and so we must have $x \notin \operatorname{bd}(A)$.

(v) Let $x \in \operatorname{cl}(A)$. For a neighbourhood U of x it then holds that $U \cap A \neq \emptyset$. If there exists a neighbourhood V of x such that $V \subseteq A$, then $x \in \operatorname{int}(A)$. If there exists *no* neighbourhood V of x such that $V \subseteq A$, then for every neighbourhood V of x we have $V \cap (\mathbb{R} \setminus A) \neq \emptyset$, and so $x \in \operatorname{bd}(A)$.

Now let $x \in \operatorname{int}(A) \cup \operatorname{bd}(A)$. If $x \in \operatorname{int}(A)$ then $x \in A$ and so $x \in\subseteq \operatorname{cl}(A)$. If $x \in \operatorname{bd}(A)$ then it follows immediately from Proposition 2.5.18 that $x \in \operatorname{cl}(A)$.

(vi) Let $x \in \operatorname{cl}(A)$. If $x \notin A$ then, for every neighbourhood U of x, $U \cap A = U \cap (A \setminus \{x\}) \neq \emptyset$, and so $x \in \operatorname{der}(A)$.

If $x \in A \cup \operatorname{der}(A)$ then either $x \in A \subseteq \operatorname{cl}(A)$, or $x \notin A$. In this latter case, $x \in \operatorname{der}(A)$ and so the set $U \cap (A \setminus \{x\})$ is nonempty for each neighbourhood U of x, and we again conclude that $x \in \operatorname{cl}(A)$.

(vii) Clearly $\text{int}(A) \cap \text{int}(\mathbb{R} \setminus A) = \emptyset$ since $A \cap (\mathbb{R} \setminus A) = \emptyset$. Now let $x \in \mathbb{R} \setminus (\text{int}(A) \cup \text{int}(\mathbb{R} \setminus A))$. Then, for any neighbourhood U of x, we have $U \not\subset A$ and $U \not\subset (\mathbb{R} \setminus A)$. Thus the sets $U \cap (\mathbb{R} \setminus A)$ and $U \cap A$ must both be nonempty, from which we conclude that $x \in \text{bd}(A)$. ■

An interesting class of subset of $\mathbb{R}$ is the following.

2.5.22 Definition (Discrete subset of $\mathbb{R}$) A subset $A \subseteq \mathbb{R}$ is ***discrete*** if there exists $\epsilon \in \mathbb{R}_{>0}$ such that, for each $x, y \in A$, $|x - y| \geq \epsilon$. •

Let us give a characterisation of discrete sets.

2.5.23 Proposition (Characterisation of discrete sets in $\mathbb{R}$) *A discrete subset* $A \subseteq \mathbb{R}$ *is countable and has no accumulation points.*

Proof It is easy to show (Exercise 2.5.6) that if A is discrete and if $N \in \mathbb{Z}_{>0}$, then the set $A \cap [-N, N]$ is finite. Therefore

$$A = \cup_{N \in \mathbb{Z}_{>0}} A \cap [-N, N],$$

which gives A as a countable union of finite sets, implying that A is countable by Proposition 1.7.16. Now let $\epsilon \in \mathbb{R}_{>0}$ satisfy $|x - y| \geq \epsilon$ for $x, y \in A$. Then, if $x \in A$ then the set $A \cap \mathsf{B}(\frac{\epsilon}{2}, x)$ is empty, implying that x is not an accumulation point. If $x \notin A$ then $\mathsf{B}(\frac{\epsilon}{2}, x)$ can contain at most one point from A, which again prohibits x from being an accumulation point. ■

The notion of a discrete set is actually a more general one having to do with what is known as the discrete topology (cf. Example **??**–**??**). The reader can explore some facts about discrete subsets of $\mathbb{R}$ in Exercise 2.5.6.

2.5.4 Compactness

The idea of compactness is absolutely fundamental in much of mathematics. The reasons for this are not at all clear to a newcomer to analysis. Indeed, the definition we give for compactness comes across as extremely unmotivated. This might be particularly since for $\mathbb{R}$ (or more generally, in $\mathbb{R}^n$) compact sets have a fairly banal characterisation as sets that are closed and bounded (Theorem 2.5.27). However, the original definition we give for a compact set is the most useful one. The main reason it is useful is that it allows for certain pointwise properties to be automatically extended to the entire set. A good example of this is Theorem 3.1.24, where continuity of a function on a compact set is extended to uniform continuity on the set. This idea of uniformity is an important one, and accounts for much of the value of the notion of compactness. But we are getting ahead of ourselves.

As indicated in the above paragraph, we shall give a rather strange seeming definition of compactness. Readers looking for a quick and dirty definition of compactness, valid for subsets of $\mathbb{R}$, can refer ahead to Theorem 2.5.27. Our construction relies on the following idea.

2.5.24 Definition (Open cover of a subset of $\mathbb{R}$) Let $A \subseteq \mathbb{R}$.

(i) An ***open cover*** for A is a family $(U_i)_{i\in I}$ of open subsets of $\mathbb{R}$ having the property that $A \subseteq \cup_{i\in I} U_i$.

(ii) A ***subcover*** of an open cover $(U_i)_{i\in I}$ of A is an open cover $(V_j)_{j\in J}$ of A having the property that $(V_j)_{j\in J} \subseteq (U_i)_{i\in I}$. •

The following property of open covers of subsets of $\mathbb{R}$ is useful.

2.5.25 Lemma (Lindelöf[10] Lemma for $\mathbb{R}$) *If $(U_i)_{i\in I}$ is an open cover of $A \subseteq \mathbb{R}$, then there exists a countable subcover of A.*

Proof Let $\mathscr{B} = \{B(r, x) \mid x, r \in \mathbb{Q}\}$. Note that $\mathscr{B}$ is a countable union of countable sets, and so is countable by Proposition 1.7.16. Therefore, we can write $\mathscr{B} = (B(r_j, x_j))_{j\in\mathbb{Z}_{>0}}$. Now define

$$\mathscr{B}' = \{B(r_j, x_j) \mid B(r_j, x_j) \subseteq U_i \text{ for some } i \in I\}.$$

Let us write $\mathscr{B}' = (B(r_{j_k}, x_{j_k}))_{k\in\mathbb{Z}_{>0}}$. We claim that $\mathscr{B}'$ covers A. Indeed, if $x \in A$ then $x \in U_i$ for some $i \in I$. Since U_i is open there then exists $k \in \mathbb{Z}_{>0}$ such that $x \in B(r_{j_k}, x_{j_k}) \subseteq U_i$. Now, for each $k \in \mathbb{Z}_{>0}$, let $i_k \in I$ satisfy $B(r_{j_k}, x_{j_k}) \subseteq U_{i_k}$. Then the countable collection of open sets $(U_{i_k})_{k\in\mathbb{Z}_{>0}}$ clearly covers A since $\mathscr{B}'$ covers A. ■

Now we define the important notion of compactness, along with some other related useful concepts.

2.5.26 Definition (Bounded, compact, and totally bounded in $\mathbb{R}$) A subset $A \subseteq \mathbb{R}$ is:

(i) ***bounded*** if there exists $M \in \mathbb{R}_{>0}$ such that $A \subseteq \overline{B}(M, 0)$;

(ii) ***compact*** if every open cover $(U_i)_{i\in I}$ of A possesses a finite subcover;

(iii) ***precompact***[11] if $\mathrm{cl}(A)$ is compact;

(iv) ***totally bounded*** if, for every $\epsilon \in \mathbb{R}_{>0}$ there exists $x_1, \ldots, x_k \in \mathbb{R}$ such that $A \subseteq \cup_{j=1}^{k} B(\epsilon, x_j)$. •

The simplest characterisation of compact subsets of $\mathbb{R}$ is the following. We shall freely interchange our use of the word compact between the definition given in Definition 2.5.26 and the conclusions of the following theorem.

2.5.27 Theorem (Heine–Borel[12] Theorem in $\mathbb{R}$) *A subset $K \subseteq \mathbb{R}$ is compact if and only if it is closed and bounded.*

Proof Suppose that K is closed and bounded. We first consider the case when $K = [a, b]$. Let $\mathcal{O} = (U_i)_{i\in I}$ be an open cover for $[a, b]$ and let

$$S_{[a,b]} = \{x \in \mathbb{R} \mid x \leq b \text{ and } [a, x] \text{ has a finite subcover in } \mathcal{O}\}.$$

[10] Ernst Leonard Lindelöf (1870–1946) was a Finnish mathematician who worked in the areas of differential equations and complex analysis.

[11] What we call "precompact" is very often called "relatively compact." However, we shall use the term "relatively compact" for something different.

[12] Heinrich Eduard Heine (1821–1881) was a German mathematician who worked mainly with special functions. Félix Edouard Justin Emile Borel (1871–1956) was a French mathematician, and he worked mainly in the area of analysis.

Note that $S_{[a,b]} \neq \emptyset$ since $a \in S_{[a,b]}$. Let $c = \sup S_{[a,b]}$. We claim that $c = b$. Suppose that $c < b$. Since $c \in [a, b]$ there is some $\bar{i} \in I$ such that $c \in U_{\bar{i}}$. As $U_{\bar{i}}$ is open, there is some $\epsilon \in \mathbb{R}_{>0}$ sufficiently small that $\mathsf{B}(\epsilon, c) \subseteq U_{\bar{i}}$. By definition of c, there exists some $x \in (c - \epsilon, c)$ for which $x \in S_{[a,b]}$. By definition of $S_{[a,b]}$ there is a finite collection of open sets $U_{i_1}, \dots, U_{i_m}$ from $\mathcal{O}$ which cover $[a, x]$. Therefore, the finite collection $U_{i_1}, \dots, U_{i_m}, U_{\bar{i}}$ of open sets covers $[a, c+\epsilon)$. This then contradicts the fact that $c = \sup S_{[a,b]}$, so showing that $b = \sup S_{[a,b]}$. The result follows by definition of $S_{[a,b]}$.

Now suppose that K is a general closed and bounded set. Then $K \subseteq [a, b]$ for some suitable $a, b \in \mathbb{R}$. Suppose that $\mathcal{O} = (U_i)_{i \in I}$ is an open cover of K, and define a new open cover $\tilde{\mathcal{O}} = \mathcal{O} \cup (\mathbb{R} \setminus K)$. Note that $\cup_{i \in I} U_i \cup (\mathbb{R} \setminus K) = \mathbb{R}$ showing that $\tilde{\mathcal{O}}$ is an open cover for $\mathbb{R}$, and therefore also is an open cover for $[a, b]$. By the first part of the proof, there exists a finite subset of $\tilde{\mathcal{O}}$ which covers $[a, b]$, and therefore also covers K. We must show that this finite cover can be chosen so as not to include the set $\mathbb{R} \setminus K$ as this set is not necessarily in $\mathcal{O}$. However, if $[a, b]$ is covered by $U_{i_1}, \dots, U_{i_k}, \mathbb{R} \setminus K$, then one sees that K is covered by $U_{i_1}, \dots, U_{i_k}$, since $K \cap (\mathbb{R} \setminus K) = \emptyset$. Thus we have arrived at a finite subset of $\mathcal{O}$ covering K, as desired.

Now suppose that K is compact. Consider the following collection of open subsets: $\mathcal{O}_K = (\mathsf{B}(\epsilon, x))_{x \in K}$. Clearly this is an open cover of K. Thus there exists a finite collection of point $x_1, \dots, x_k \in K$ such that $(\mathsf{B}(\epsilon, x_j))_{j \in \{1,\dots,k\}}$ covers K. If we take

$$M = \max\{|x_1|, \dots, |x_k|\} + 2$$

then we easily see that $K \subseteq \overline{\mathsf{B}}(M, 0)$, so that K is bounded. Now suppose that K is not closed. Then $K \subset \mathrm{cl}(K)$. By part (vi) of Proposition 2.5.21 there exists an accumulation point x_0 of K that is not in K. Then, for any $j \in \mathbb{Z}_{>0}$ there exists a point $x_j \in K$ such that $|x_0 - x_j| < \frac{1}{j}$. Define

$$U_j = (-\infty, x_0 - \tfrac{1}{j}) \cup (x_0 + \tfrac{1}{j}, \infty),$$

noting that U_j is open, since it is the union of open sets (see Exercise 2.5.1). We claim that $(U_j)_{j \in \mathbb{Z}_{>0}}$ is an open cover of K. Indeed, we will show that $\cup_{j \in \mathbb{Z}_{>0}} U_j = \mathbb{R} \setminus \{x_0\}$. To see this, let $x \in \mathbb{R} \setminus \{x_0\}$ and choose $k \in \mathbb{Z}_{>0}$ such that $\frac{1}{k} < |x - x_0|$. Then it follows by definition of U_k that $x \in U_k$. Since $x_0 \notin K$, we then have $K \subseteq \cup_{j \in \mathbb{Z}_{>0}} U_j$. Next we show that there is no finite subset of $(U_j)_{j \in \mathbb{Z}_{>0}}$ that covers K. Indeed, consider a finite set $j_1, \dots, j_k \in \mathbb{Z}_{>0}$, and suppose without loss of generality that $j_1 < \cdots < j_k$. Then the point x_{j_k+1} satisfies $|x_0 - x_{j_k+1}| < \frac{1}{j_k+1} < \frac{1}{j_k}$, implying that $x_{j_k+1} \notin U_{j_k} \supseteq \cdots \supseteq U_{j_1}$. Thus, if K is not closed, we have constructed an open cover of K having no finite subcover. From this we conclude that if K is compact, then it is closed. ■

The Heine–Borel Theorem has the following useful corollary.

2.5.28 Corollary (Closed subsets of compact sets in $\mathbb{R}$ are compact) *If* $\mathrm{A} \subseteq \mathbb{R}$ *is compact and if* $\mathrm{B} \subseteq \mathrm{A}$ *is closed, then* B *is compact.*

Proof Since A is bounded by the Heine–Borel Theorem, B is also bounded. Thus B is also compact, again by the Heine–Borel Theorem. ■

In Chapter **??** we shall encounter many of the ideas in this section in the more general setting of topological spaces. Many of the ideas for $\mathbb{R}$ transfer directly to this more general setting. However, with compactness, some care must be exercised. In particular, it is *not* true that, in a general topological space, a subset is compact

if and only if it is closed and bounded. Indeed, in a general topological space, the notion of bounded is not defined. It is not an uncommon error for newcomers to confuse "compact" with "closed and bounded" in situations where this is not the case.

missing stuff

The following result is another equivalent characterisation of compact subsets of $\mathbb{R}$, and is often useful.

2.5.29 Theorem (Bolzano–Weierstrass[13] Theorem in $\mathbb{R}$) *A subset $\mathrm{K} \subseteq \mathbb{R}$ is compact if and only if every sequence in K has a subsequence which converges in K.*

Proof First suppose that K is compact. Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in K. Since K is bounded by Theorem 2.5.27, the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ is bounded. We next show that there exists either a monotonically increasing, or a monotonically decreasing, subsequence of $(x_j)_{j\in\mathbb{Z}_{>0}}$. Define

$$D = \{j \in \mathbb{Z}_{>0} \mid x_k > x_j,\ k > j\}$$

If the set D is infinite, then we can write $D = (j_k)_{k\in\mathbb{Z}_{>0}}$. By definition of D, it follows that $x_{j_{k+1}} > x_{j_k}$ for each $k \in \mathbb{Z}_{>0}$. Thus the subsequence $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$ is monotonically increasing. If the set D is finite choose $j_1 > \sup D$. Then there exists $j_2 > j_1$ such that $x_{j_2} \leq x_{j_1}$. Since $j_2 > \sup D$, there then exists $j_3 > j_2$ such that $x_{j_3} \leq x_{j_2}$. By definition of D, this process can be repeated inductively to yield a monotonically decreasing subsequence $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$. It now follows from Theorem 2.3.8 that the sequence $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$, be it monotonically increasing or monotonically decreasing, converges.

Next suppose that every sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in K possesses a convergent subsequence. Let $(U_i)_{i\in I}$ be an open cover of K, and by Lemma 2.5.25 choose a countable subcover which we denote by $(U_j)_{j\in\mathbb{Z}_{>0}}$. Now suppose that every finite subcover of $(U_j)_{j\in\mathbb{Z}_{>0}}$ does not cover K. This means that, for every $k \in \mathbb{Z}_{>0}$, the set $C_k = K \setminus \left(\cup_{j=1}^{k} U_j\right)$ is nonempty. Thus we may define a sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$ in $\mathbb{R}$ such that $x_k \in C_k$. Since the sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$ is in K, it possesses a convergent subsequence $(x_{k_m})_{m\in\mathbb{Z}_{>0}}$, by hypotheses. Let x be the limit of this subsequence. Since $x \in K$ and since $K = \cup_{j\in\mathbb{Z}_{>0}} U_j$, $x \in U_l$ for some $l \in \mathbb{Z}_{>0}$. Since the sequence $(x_{k_m})_{m\in\mathbb{Z}_{>0}}$ converges to x, it follows that there exists $N \in \mathbb{Z}_{>0}$ such that $x_{k_m} \in U_l$ for $m \geq N$. But this contradicts the definition of the sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$, forcing us to conclude that our assumption is wrong that there is no finite subcover of K from the collection $(U_j)_{j\in\mathbb{Z}_{>0}}$. ■

The following property of compact intervals of $\mathbb{R}$ is useful.

2.5.30 Theorem (Lebesgue[14] number for compact intervals) *Let $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ be a compact interval. Then for any open cover $(\mathrm{U}_\alpha)_{\alpha\in\mathrm{A}}$ of $[\mathrm{a}, \mathrm{b}]$, there exists $\delta \in \mathbb{R}_{>0}$, called the*

[13] Bernard Placidus Johann Nepomuk Bolzano (1781–1848) was a Czechoslovakian philosopher, mathematician, and theologian who made mathematical contributions to the field of analysis. Karl Theodor Wilhelm Weierstrass (1815–1897) is one of the greatest of all mathematicians. He made significant contributions to the fields of analysis, complex function theory, and the calculus of variations.

[14] Henri Léon Lebesgue (1875–1941) was a French mathematician. His work was in the area of analysis. The Lebesgue integral is considered to be one of the most significant contributions to mathematics in the past century or so.

***Lebesgue number** of* I, *such that, for each* $x \in [a,b]$, *there exists* $\alpha \in A$ *such that* $B(\delta, x) \cap I \subseteq U_\alpha$.

Proof Suppose there exists an open cover $(U_\alpha)_{\alpha\in A}$ such that, for all $\delta \in \mathbb{R}_{>0}$, there exists $x \in [a,b]$ such that none of the sets U_α, $\alpha \in A$, contains $B(\delta, x) \cap I$. Then there exists a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in I such that

$$\{\alpha \in A \mid B(\tfrac{1}{j}, x_j) \subseteq U_\alpha\} = \emptyset$$

for each $j \in \mathbb{Z}_{>0}$. By the Bolzano–Weierstrass Theorem there exists a subsequence $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$ that converges to a point, say x, in $[a,b]$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ and $\alpha \in A$ such that $B(\epsilon, x) \subseteq U_\alpha$. Now let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $|x_{j_k} - x| < \frac{\epsilon}{2}$ for $k \geq N$ and such that $\frac{1}{j_N} < \frac{\epsilon}{2}$. Now let $k \geq N$. Then, if $y \in B(\frac{1}{j_k}, x_{j_k})$ we have

$$|y - x| = |y - x_{j_k} + x_{j_k} - x| \leq |y - x_{j_k}| + |x - x_{j_k}| < \epsilon.$$

Thus we arrive at the contradiction that $B(\frac{1}{j_k}, x_{j_k}) \subseteq U_\alpha$. ■

The following result is sometimes useful.

2.5.31 Proposition (Countable intersections of nested compact sets are nonempty) *Let* $(K_j)_{j\in\mathbb{Z}_{>0}}$ *be a collection of compact subsets of* $\mathbb{R}$ *satisfying* $K_{j+1} \subseteq K_j$. *Then* $\cap_{j\in\mathbb{Z}_{>0}} K_j$ *is nonempty.*

Proof It is clear that $K = \cap_{j\in\mathbb{Z}_{>0}} K_j$ is bounded, and moreover it is closed by Exercise 2.5.1. Thus K is compact by the Heine–Borel Theorem. Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence for which $x_j \in K_j$ for $j \in \mathbb{Z}_{>0}$. This sequence is thus a sequence in K_1 and so, by the Bolzano–Weierstrass Theorem, has a subsequence $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$ converging to $x \in K_1$. The sequence $(x_{j_{k+1}})_{k\in\mathbb{Z}_{>0}}$ is then a sequence in K_2 which is convergent, so showing that $x \in K_2$. Similarly, one shows that $x \in K_j$ for all $j \in \mathbb{Z}_{>0}$, giving the result. ■

Finally, let us indicate the relationship between the notions of relative compactness and total boundedness. We see that for $\mathbb{R}$ these concepts are the same. This may not be true in general.***missing stuff***

2.5.32 Proposition ("Precompact" equals "totally bounded" in $\mathbb{R}$) *A subset of* $\mathbb{R}$ *is precompact if and only if it is totally bounded.*

Proof Let $A \subseteq \mathbb{R}$.

First suppose that A is precompact. Since $A \subseteq \text{cl}(A)$ and since $\text{cl}(A)$ is bounded by the Heine–Borel Theorem, it follows that A is bounded. It is then easy to see that A is totally bounded.

Now suppose that A is totally bounded. For $\epsilon \in \mathbb{R}_{>0}$ let $x_1, \ldots, x_k \in \mathbb{R}$ have the property that $A \subseteq \cup_{j=1}^k B(\epsilon, x_j)$. If

$$M_0 = \max\{|x_j - x_l| \mid j, l \in \{i, \ldots, k\}\} + 2\epsilon,$$

then it is easy to see that $A \subseteq B(M, 0)$ for any $M > M_0$. Then $\text{cl}(A) \subseteq \overline{B}(M, 0)$ by part (iv) of Proposition 2.5.20, and so $\text{cl}(A)$ is bounded. Since $\text{cl}(A)$ is closed, it follows from the Heine–Borel Theorem that A is precompact. ■

missing stuff missing stuff

2.5.5 Connectedness

The idea of a connected set will come up occasionally in these volumes. Intuitively, a set is connected if it cannot be "broken in two." We will study it more systematically in ***missing stuff***, and here we only give enough detail to effectively characterise connected subsets of $\mathbb{R}$.

2.5.33 Definition (Connected subset of $\mathbb{R}$) Subsets $A, B \subseteq \mathbb{R}$ are ***separated*** if $A \cap \mathrm{cl}(B) = \emptyset$ and $\mathrm{cl}(A) \cap B = \emptyset$. A subset $S \subseteq \mathbb{R}$ is ***disconnected*** if $S = A \cup B$ for nonempty separated subsets A and B. A subset $S \subseteq \mathbb{R}$ is ***connected*** if it is not disconnected. •

Rather than give examples, let us simply immediately characterise the connected subsets of $\mathbb{R}$, since this renders all examples trivial to understand.

2.5.34 Theorem (Connected subsets of $\mathbb{R}$ are intervals and vice versa) *A subset $S \subseteq \mathbb{R}$ is connected if and only if S is an interval.*

Proof Suppose that S is not an interval. Then, by Proposition 2.5.5, there exists $a, b \in S$ with $a < b$ and $c \in (a, b)$ such that $c \notin S$. Let $A_c = S \cap (-\infty, c)$ and $B_c = S \cap (c, \infty)$, and note that both A_c and B_c are nonempty. Also, since $c \notin S$, $S = A_c \cup B_c$. Since $(-\infty, c) \cap [c, \infty) = \emptyset$ and $(-\infty, c] \cap (c, \infty) = \emptyset$, A_c and B_c are separated. That S is not connected follows.

Now suppose that S is not connected, and write $S = A \cup B$ for nonempty separated sets A and B. Without loss of generality, let $a \in A$ and $b \in B$ have the property that $a < b$. Note that $A \cap [a, b]$ is bounded so that $c = \sup A \cap [a, b]$ exists in $\mathbb{R}$. Then $c \in \mathrm{cl}(A \cap [a, b]) \subseteq \mathrm{cl}(A) \cap [a, b]$. In other words, $c \in \mathrm{cl}(A)$. Since $\mathrm{cl}(A) \cap B = \emptyset$, $c \notin B$. If $c \notin A$ then $c \notin S$, and so S is not connected by Proposition 2.5.5. If $c \in A$ then, since $A \cap \mathrm{cl}(B) = \emptyset$, $c \notin \mathrm{cl}(B)$. In this case there exists an open interval containing c that does not intersect $\mathrm{cl}(B)$. In particular, there exists $d > c$ such that $d \notin B$. Since $d > c$ we also have $d \notin A$, and so $d \notin S$. Again we conclude that S is not an interval by Proposition 2.5.5. ■

Let us consider a few examples.

2.5.35 Examples (Connected subsets of sets)

1. If $D \subseteq \mathbb{R}$ is a discrete set as given in Definition 2.5.22. From Theorem 2.5.34 we see that the only subsets of D that are connected are singletons.
2. Note that it also follows from Theorem 2.5.34 that the only connected subsets of $\mathbb{Q} \subseteq \mathbb{R}$ are singletons. However, $\mathbb{Q}$ is not discrete. •

2.5.6 Sets of measure zero

The topic of this section will receive a full treatment in the context of measure theory as presented in Chapter **??**. However, it is convenient here to talk about a simple concepts from measure theory, one which formalises the idea of a set being "small." We shall only give here the definition and a few examples. The reader should look ahead to Chapter **??** for more detail.

2.5.36 Definition (Set of measure zero in $\mathbb{R}$) A subset $A \subseteq \mathbb{R}$ has ***measure zero***, or is ***of measure zero***, if

$$\inf\Big\{\sum_{j=1}^{\infty}|b_j - a_j| \;\Big|\; A \subseteq \bigcup_{j\in\mathbb{Z}_{>0}}(a_j, b_j)\Big\} = 0. \qquad \bullet$$

The idea, then, is that one can cover a set A with open intervals, each of which have some length. One can add all of these lengths to get a total length for the intervals used to cover A. Now, if one can make this total length arbitrarily small, then the set has measure zero.

2.5.37 Notation ("Almost everywhere" and "a.e.") We give here an important piece of notation associated to the notion of a set of measure zero. Let $A \subseteq \mathbb{R}$ and let $P\colon A \to \{\text{true}, \text{false}\}$ be a property defined on A (see the prelude to the Principle of Transfinite Induction, Theorem 1.5.14). The property P holds ***almost everywhere***, ***a.e.***, or ***for almost every*** $\mathbf{x \in A}$ if the set $\{x \in A \mid P(x) = \text{false}\}$ has measure zero. $\bullet$

This is best illustrated with some examples.

2.5.38 Examples (Sets of measure zero)

1. Let $A = \{x_1, \ldots, x_k\}$ for some distinct $x_1, \ldots, x_k \in \mathbb{R}$. We claim that this set has measure zero. Note that for any $\epsilon \in \mathbb{R}_{>0}$ the intervals $(x_j - \frac{\epsilon}{4k}, x_j + \frac{\epsilon}{4k})$, $j \in \{1, \ldots, k\}$, clearly cover A. Now consider the countable collection of open intervals

$$((x_j - \tfrac{\epsilon}{4k}, x_j + \tfrac{\epsilon}{4k}))_{j\in\{1,\ldots,k\}} \cup ((0, \tfrac{\epsilon}{2^{j+1}}))_{j\in\mathbb{Z}_{>0}}$$

obtained by adding to the intervals covering A a collection of intervals around zero. The total length of these intervals is

$$\sum_{j=1}^{k}|(x_j + \tfrac{\epsilon}{4k}) - (x_j - \tfrac{\epsilon}{4k})| + \frac{\epsilon}{2}\sum_{j=1}^{\infty}\frac{1}{2^j} = \frac{\epsilon}{2} + \frac{\epsilon}{2},$$

using the fact that $\sum_{j=1}^{\infty}\frac{\epsilon}{2^j} = 1$ (by Example 2.4.2–1). Since $\inf\{2k\epsilon \mid \epsilon \in \mathbb{R}_{>0}\} = 0$, our claim that A has zero measure is validated.

2. Now let $A = \mathbb{Q}$ be the set of rational numbers. To show that A has measure zero, note that from Exercise 2.1.3 that A is countable. Thus we can write the elements of A as $(q_j)_{j\in\mathbb{Z}_{>0}}$. Now let $\epsilon \in \mathbb{R}_{>0}$ and for $j \in \mathbb{Z}_{>0}$ define $a_j = q_j - \frac{\epsilon}{2^j}$ and $b_j = q_j + \frac{\epsilon}{2^j}$. Then the collection (a_j, b_j), $j \in \mathbb{Z}_{>0}$, covers A. Moreover,

$$\sum_{j=1}^{\infty}|b_j - a_j| = \sum_{j=1}^{\infty}\frac{2\epsilon}{2^j} = 2\epsilon,$$

using the fact, shown in Example 2.4.2–1, that the series $\sum_{j=1}^{\infty}\frac{1}{2^j}$ converges to 1. Now, since $\inf\{2\epsilon \mid \epsilon \in \mathbb{R}_{>0}\} = 0$, it follows that A indeed has measure zero.

3. Let $A = \mathbb{R} \setminus \mathbb{Q}$ be the set of irrational numbers. We claim that this set does not have measure zero. To see this, let $k \in \mathbb{Z}_{>0}$ and consider the set $A_k = A \cap [-k, k]$. Now let $\epsilon \in \mathbb{R}_{>0}$. We claim that if $((a_j, b_j))_{j \in \mathbb{Z}_{>0}}$, is a collection of open intervals for which $A_k \subseteq \cup_{j \in \mathbb{Z}_{>0}} (a_j, b_j)$, then

$$\sum_{j=1}^{\infty} |b_j - a_j| \geq 2k - \epsilon. \tag{2.9}$$

To see this, let $((c_l, d_l))_{l \in \mathbb{Z}_{>0}}$ be a collection of intervals such that $\mathbb{Q} \cap [-k, k] \subseteq \cup_{l \in \mathbb{Z}_{>0}} (c_l, d_l)$ and such that

$$\sum_{l=1}^{\infty} |d_l - c_l| < \epsilon.$$

Such a collection of intervals exists since we have already shown that $\mathbb{Q}$, and therefore $\mathbb{Q} \cap [-k, k]$, has measure zero (see Exercise 2.5.7). Now note that

$$[-k, k] \subseteq \Big(\bigcup_{j \in \mathbb{Z}_{>0}} (a_j, b_j) \Big) \cup \Big(\bigcup_{l \in \mathbb{Z}_{>0}} (c_l, d_l) \Big),$$

so that

$$\Big(\sum_{j=1}^{\infty} |b_j - a_j| \Big) + \Big(\sum_{l=1}^{\infty} |d_l - c_l| \Big) \geq 2k.$$

From this we immediately conclude that (2.9) does indeed hold. Moreover, (2.9) holds for every $k \in \mathbb{Z}_{>0}$, for every $\epsilon \in \mathbb{R}_{>0}$, and for every open cover $((a_j, b_j))_{j \in \mathbb{Z}_{>0}}$ of A_k. Thus,

$$\inf \Big\{ \sum_{l=1}^{\infty} |\tilde{b}_l - \tilde{a}_l| \;\Big|\; A \subseteq \bigcup_{l \in \mathbb{Z}_{>0}} (\tilde{a}_l, \tilde{b}_l) \Big\}$$
$$\geq \inf \Big\{ \sum_{j=1}^{\infty} |b_j - a_j| \;\Big|\; A_k \subseteq \bigcup_{j \in \mathbb{Z}_{>0}} (a_j, b_j) \Big\} \geq 2k - \epsilon$$

for every $k \in \mathbb{Z}_{>0}$ and for every $\epsilon \in \mathbb{R}_{>0}$. This precludes A from having measure zero. •

The preceding examples suggest sets of measure zero are countable. This is not so, and the next famous example gives an example of an uncountable set with measure zero.

2.5.39 Example (An uncountable set of measure zero: the middle-thirds Cantor set) In this example we construct one of the standard "strange" sets used in real analysis to exhibit some of the characteristics that can possibly be attributed to subsets of $\mathbb{R}$. We shall also use this set in a construction in Example 3.2.27 to give an example of a continuous monotonically increasing function whose derivative is zero almost everywhere.

Let $C_0 = [0, 1]$. Then define

$$\begin{aligned} C_1 &= [0, \tfrac{1}{3}] \cup [\tfrac{2}{3}, 1], \\ C_2 &= [0, \tfrac{1}{9}] \cup [\tfrac{2}{9}, \tfrac{1}{3}] \cup [\tfrac{2}{3}, \tfrac{7}{9}] \cup [\tfrac{8}{9}, 1], \\ &\vdots \end{aligned}$$

so that C_k is a collection of 2^k disjoint closed intervals each of length 3^{-k} (see Figure 2.5). We define $C = \cap_{k \in \mathbb{Z}_{>0}} C_k$, which we call the ***middle-thirds Cantor set***.

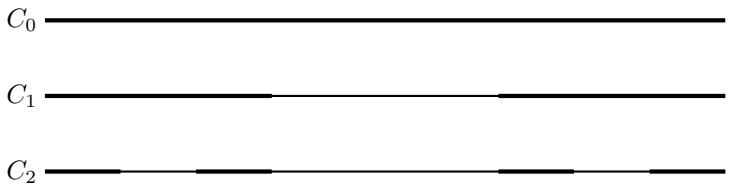


Figure 2.5 The first few sets used in the construction of the middle-thirds Cantor set

Let us give some of the properties of C.

1 Lemma C *has the same cardinality as* $[0, 1]$.

Proof Note that each of the sets C_k, $k \in \mathbb{Z}_{\geq 0}$, is a collection of disjoint closed intervals. Let us write $C_k = \cup_{j=1}^{2^k} I_{k,j}$, supposing that the intervals $I_{k,j}$ are enumerated such that the right endpoint of $I_{k,j}$ lies to the left of the left endpoint of $I_{k,j+1}$ for each $k \in \mathbb{Z}_{\geq 0}$ and $j \in \{1, \ldots, 2^k\}$. Now note that each interval $I_{k+1,j}$, $k \in \mathbb{Z}_{\geq 0}$, $j \in \{1, \ldots, 2^{k+1}\}$ comes from assigning two intervals to each of the intervals $I_{k,j}$, $k \in \mathbb{Z}_{\geq 0}$, $j \in \{1, \ldots, 2^k\}$. Assign to an interval $I_{k+1,j}$, $k \in \mathbb{Z}_{\geq 0}$, $j \in \{1, \ldots, 2^k\}$, the number 0 (resp. 1) if it the left (resp. right) interval coming from an interval $I_{k,j'}$ of C_k. In this way, each interval in C_k, $k \in \mathbb{Z}_{\geq 0}$, is assigned a 0 or a 1 in a unique manner. Since, for each point in $x \in C$, there is exactly one $j \in \{1, \ldots, 2^k\}$ such that $x \in I_{k,j}$. Therefore, for each point in C there is a unique decimal expansion $0.n_1n_2n_3 \ldots$ where $n_k \in \{0, 1\}$. Moreover, for every such decimal expansion, there is a corresponding point in C. However, such decimal expansions are exactly binary decimal expansions for points in $[0, 1]$. In other words, there is a bijection from C to $[0, 1]$. ▼

2 Lemma C *is a set of measure zero.*

Proof Let $\epsilon \in \mathbb{R}_{>0}$. Note that each of the sets C_k can be covered by a finite number of closed intervals whose lengths sum to $\left(\frac{2}{3}\right)^k$. Therefore, each of the sets C_k can be covered by open intervals whose lengths sum to $\left(\frac{2}{3}\right)^k + \frac{\epsilon}{2}$. Choosing k sufficiently large that $\left(\frac{2}{3}\right)^k < \frac{\epsilon}{2}$ we see that C is contained in the union of a finite collection of open intervals whose lengths sum to ϵ. Since ϵ is arbitrary, it follows that C has measure zero. ▼

This example thus shows that sets of measure zero, while "small" in some sense, can be "large" in terms of the number of elements they possess. Indeed, in terms of cardinality, C has the same size as $[0,1]$, although their measures differ by as much as possible. •

2.5.7 Cantor sets

The remainder of this section is devoted to a characterisation of certain sorts of exotic sets, perhaps the simplest example of which is the middle-thirds Cantor set of Example 2.5.39. This material is only used occasionally, and so can be omitted until the reader feels they need/want to understand it.

The qualifier "middle-thirds" in Example 2.5.39 makes one believe that there might be a general notion of a "Cantor set." This is indeed the case.

2.5.40 Definition (Cantor set) Let $I \subseteq \mathbb{R}$ be a closed interval. A subset $A \subseteq I$ is a ***Cantor set*** if

(i) A is closed,

(ii) $\text{int}(A) = \emptyset$, and

(iii) every point of A is an accumulation point of A. •

We leave it to the reader to verify in Exercise 2.5.10 that the middle-thirds Cantor set is a Cantor set, according to the previous definition.

One might wonder whether all Cantor sets have the properties of having the cardinality of an interval and of having measure zero. To address this, we give a result and an example. The result shows that all Cantor sets are uncountable.

2.5.41 Proposition (Cantor sets are uncountable) *If $\mathrm{A} \subseteq \mathbb{R}$ is a nonempty set having the property that each of its points is an accumulation point, then A is uncountable. In particular, Cantor sets are uncountable.*

Proof Any finite set has no accumulation points by Proposition 2.5.13. Therefore A must be either countably infinite or uncountable. Suppose that A is countable and write $A = (x_j)_{j \in \mathbb{Z}_{>0}}$. Let $y_1 \in A \setminus \{x_1\}$. For $r_1 < |x_1 - y_1|$ we have $x_1 \notin \overline{\mathsf{B}}(r_1, y_1)$. We note that y_1 is an accumulation point for $A \setminus \{x_1, x_2\}$; this follows immediately from Proposition 2.5.13. Thus there exists $y_2 \in A \setminus \{x_1, x_2\}$ such that $y_2 \in \mathsf{B}(r_1, y_1)$ and such that $y_2 \neq y_1$. If $r_2 < \min\{|x_2 - y_2|, r_1 - |y_2 - y_2|\}$ then $x_2 \notin \overline{\mathsf{B}}(r_2, y_2)$ and $\overline{\mathsf{B}}(r_2, y_2) \subseteq \mathsf{B}(r_1, y_1)$ by a simple application of the triangle inequality. Continuing in this way we define a sequence $(\overline{\mathsf{B}}(r_j, y_j))_{j \in \mathbb{Z}_{>0}}$ of closed balls having the following properties:

1. $\overline{\mathsf{B}}(r_{j+1}, y_{j+1}) \subseteq \overline{\mathsf{B}}(r_j, y_j)$ for each $j \in \mathbb{Z}_{>0}$;
2. $x_j \notin \overline{\mathsf{B}}(r_j, y_j)$ for each $j \in \mathbb{Z}_{>0}$.

Note that $(\overline{\mathsf{B}}(r_j, y_j) \cap A)_{j \in \mathbb{Z}_{>0}}$ is a nested sequence of compact subsets of A, and so by Proposition 2.5.31, $\cap_{j \in \mathbb{Z}_{>0}} (\overline{\mathsf{B}}(r_j, y_j) \cap A)$ is a nonempty subset of A. However, for any $j \in \mathbb{Z}_{>0}$, $x_j \notin \cap_{j \in \mathbb{Z}_{>0}} (\overline{\mathsf{B}}(r_j, y_j) \cap A)$, and so we arrive, by contradiction, to the conclusion that A is not countable. ■

The following example shows that Cantor sets may not have measure zero.

2.5.42 Example (A Cantor set not having zero measure) We will define a subset of $[0,1]$ that is a Cantor set, but does not have measure zero. The construction mirrors closely that of Example 2.5.39.

We let $\epsilon \in (0,1)$. Let $C_{\epsilon,0} = [0,1]$ and define $C_{\epsilon,1}$ by deleting from $C_{\epsilon,0}$ an open interval of length $\frac{\epsilon}{2}$ centered at the midpoint of $C_{\epsilon,0}$. Note that $C_{\epsilon,1}$ consists of two disjoint closed intervals whose lengths sum to $1 - \frac{\epsilon}{2}$. Next define $C_{\epsilon,2}$ by deleting from $C_{\epsilon,1}$ two open intervals, each of length $\frac{\epsilon}{8}$, centered at the midpoints of each of the intervals comprising $C_{\epsilon,1}$. Note that $C_{\epsilon,2}$ consists of four disjoint closed intervals whose lengths sum to $1 - \frac{\epsilon}{4}$. Proceed in this way, defining a sequence of sets $(C_{\epsilon,k})_{k\in\mathbb{Z}_{>0}}$, where $C_{\epsilon,k}$ consists of 2^k disjoint closed intervals whose lengths sum to $1 - \sum_{j=1}^{k} \frac{\epsilon}{2^j} = 1 - \epsilon$. Take $C_\epsilon = \cap_{k\in\mathbb{Z}_{>0}} C_{\epsilon,k}$.

Let us give the properties of C_ϵ in a series of lemmata.

1 Lemma *C_ϵ is a Cantor set.*

Proof That C_ϵ is closed follows from Exercise 2.5.1 and the fact that it is the intersection of a collection of closed sets. To see that $\text{int}(C_\epsilon) = \emptyset$, let $I \subseteq [0,1]$ be an open interval and suppose that $I \subseteq C_\epsilon$. This means that $I \subseteq C_{\epsilon,k}$ for each $k \in \mathbb{Z}_{>0}$. Note that the sets $C_{\epsilon,k}$, $k \in \mathbb{Z}_{>0}$, are unions of closed intervals, and that for any $\delta \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that the lengths of the intervals comprising $C_{\epsilon,k}$ are less than δ for $k \geq N$. Thus the length of I must be zero, and so $I = \emptyset$. Thus C_ϵ contains no nonempty open intervals, and so must have an empty interior. To see that every point of C_ϵ is an accumulation point of C_ϵ, we note that all points in C_ϵ are endpoints for one of the closed intervals comprising $C_{\epsilon,k}$ for some $k \in \mathbb{Z}_{>0}$. Moreover, it is clear that every neighbourhood of a point in C_ϵ must contain another endpoint from one of the closed intervals comprising $C_{\epsilon,k}$ for some $k \in \mathbb{Z}_{>0}$. Indeed, were this not the case, this would imply the existence of a nonempty open interval contained in C_ϵ, and we have seen that there can be no such interval. ▼

2 Lemma *C_ϵ is uncountable.*

Proof This can be proved in exactly the same manner as the middle-thirds Cantor set was shown to be uncountable. ▼

3 Lemma *C_ϵ does not have measure zero.*

Proof Once one knows the basic properties of Lebesgue measure, it follows immediately that C_ϵ has, in fact, measure $1 - \epsilon$. However, since we have not yet defined measure, let us prove that C_ϵ does not have measure zero, using only the definition of a set of measure zero. Let $((a_j, b_j))_{j\in\mathbb{Z}_{>0}}$ be a countable collection of open intervals having the property that

$$C_\epsilon \subseteq \bigcup_{j\in\mathbb{Z}_{>0}} (a_j, b_j).$$

Since C_ϵ is closed, it is compact by Corollary 2.5.28. Therefore, there exists a finite collection $((a_{j_l}, b_{j_l}))_{l\in\{1,\dots,m\}}$ of intervals having the property that

$$C_\epsilon \subseteq \bigcup_{l=1}^{m} (a_{j_l}, b_{j_l}). \tag{2.10}$$

We claim that there exists $k \in \mathbb{Z}_{>0}$ such that

$$C_{\epsilon,k} \subseteq \bigcup_{l=1}^{m} (a_{j_l}, b_{j_l}). \tag{2.11}$$

Indeed, suppose that, for each $k \in \mathbb{Z}_{>0}$ there exists $x_k \in C_{\epsilon,k}$ such that $x_k \notin \cup_{l=1}^{m}(a_{j_l}, b_{j_l})$. The sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$ is then a sequence in the compact set $C_{\epsilon,1}$, and so by the Bolzano–Weierstrass Theorem, possesses a subsequence $(x_{k_r})_{r\in\mathbb{Z}_{>0}}$ converging to $x \in C_{\epsilon,1}$. But the sequence $(x_{k_{r+1}})_{r\in\mathbb{Z}_{>0}}$ is then a convergent sequence in $C_{\epsilon,2}$, so $x \in C_{\epsilon,2}$. Continuing in this way, $x \in \cap_{k\in\mathbb{Z}_{>0}} C_{\epsilon,k}$. Moreover, the sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$ is also a sequence in the closed set $[0,1] - \cup_{l=1}^{m}(a_{j_l}, b_{j_l})$, and so we conclude that $x \in [0,1] - \cup_{l=1}^{m}(a_{j_l}, b_{j_l})$. Thus we contradict the condition (2.10), and so there indeed must be a $k \in \mathbb{Z}_{>0}$ such that (2.11) holds. However, this implies that any collection of open intervals covering C_ϵ must have lengths which sum to at least $1 - \epsilon$. Thus C_ϵ cannot have measure zero. ▼

Cantor sets such as C_ϵ are sometimes called ***fat Cantor sets***, reflecting the fact that they do not have measure zero. Note, however, that they are not *that* fat, since they have an empty interior! •

2.5.8 Notes

Some uses of δ-fine tagged partitions in real analysis can be found in the paper of Gordon [1998].

Exercises

2.5.1 For an arbitrary collection $(U_a)_{a\in A}$ of open sets and an arbitrary collection $(C_b)_{b\in B}$ of closed sets, do the following:

(a) show that $\cup_{a\in A} U_a$ is open;

(b) show that $\cap_{b\in B} C_b$ is closed;

For open sets U_1 and U_2 and closed sets C_1 and C_2, do the following:

(c) show that $U_1 \cap U_2$ is open;

(d) show that $C_1 \cup C_2$ is closed.

2.5.2 Show that a set $A \subseteq \mathbb{R}$ is closed if and only if it contains all of its limit points.

2.5.3 For $A \subseteq \mathbb{R}$, show that $\operatorname{bd}(A) = \operatorname{bd}(\mathbb{R} \setminus A)$.

2.5.4 Find counterexamples to the following statements (cf. Propositions 2.5.15, 2.5.19, and 2.5.20):

(a) $\operatorname{int}(A \cup B) \subseteq \operatorname{int}(A) \cup \operatorname{int}(B)$;

(b) $\operatorname{int}(\cup_{i\in I} A_i) \subseteq \cup_{i\in I} \operatorname{int}(A_i)$;

(c) $\operatorname{int}(\cap_{i\in I} A_i) \supseteq \cap_{i\in I} \operatorname{int}(A_i)$;

(d) $\operatorname{cl}(A \cap B) \supseteq \operatorname{cl}(A) \cap \operatorname{cl}(B)$;

(e) $\operatorname{cl}(\cup_{i\in I} A_i) \subseteq \cup_{i\in I} \operatorname{cl}(A_i)$;

(f) $\operatorname{cl}(\cap_{i\in I} A_i) \supseteq \cap_{i\in I} \operatorname{cl}(A_i)$.

Hint: *No fancy sets are required. Intervals will suffice in all cases.*

2.5.5 For each of the following statements, prove the statement if it is true, and give a counterexample if it is not:

(a) $\text{int}(A_1 \cup A_2) = \text{int}(A_1) \cup \text{int}(A_2)$;

(b) $\text{int}(A_1 \cap A_2) = \text{int}(A_1) \cap \text{int}(A_2)$;

(c) $\text{cl}(A_1 \cup A_2) = \text{cl}(A_1) \cup \text{cl}(A_2)$;

(d) $\text{cl}(A_1 \cap A_2) = \text{cl}(A_1) \cap \text{cl}(A_2)$;

(e) $\text{bd}(A_1 \cup A_2) = \text{bd}(A_1) \cup \text{bd}(A_2)$;

(f) $\text{bd}(A_1 \cap A_2) = \text{bd}(A_1) \cap \text{bd}(A_2)$.

2.5.6 Do the following:

(a) show that any finite subset of $\mathbb{R}$ is discrete;

(b) show that a discrete bounded set is finite;

(c) find a set $A \subseteq \mathbb{R}$ that is countable and has no accumulation points, but that is not discrete.

2.5.7 Show that if $A \subseteq \mathbb{R}$ has measure zero and if $B \subseteq A$, then B has measure zero.

2.5.8 Show that any countable subset of $\mathbb{R}$ has measure zero.

2.5.9 Let $(Z_j)_{j \in \mathbb{Z}_{>0}}$ be a family of subsets of $\mathbb{R}$ that each have measure zero. Show that $\cup_{j \in \mathbb{Z}_{>0}} Z_j$ also has measure zero.

2.5.10 Show that the set C constructed in Example 2.5.39 is a Cantor set.

This version: 2018/01/09

Chapter 3

Functions of a real variable

In the preceding chapter we endowed the set $\mathbb{R}$ with a great deal of structure. In this chapter we employ this structure to endow functions whose domain and range is $\mathbb{R}$ with some useful properties. These properties include the usual notions of continuity and differentiability given in first-year courses on calculus. The theory of the Riemann integral is also covered here, and it can be expected that students will have at least a functional familiarity with this. However, students who have had the standard engineering course (at least in North American universities) dealing with these topics will find the treatment here a little different than what they are used to. Moreover, there are also topics covered that are simply not part of the standard undergraduate curriculum, but which still fit under the umbrella of "functions of a real variable." These include a detailed discussion of functions of bounded variation, an introductory treatment of absolutely continuous functions, and a generalisation of the Riemann integral called the Riemann–Stieltjes integral.

Do I need to read this chapter? For readers having had a good course in analysis, this chapter can easily be bypassed completely. It can be expected that all other readers will have some familiarity with the material in this chapter, although not perhaps with the level of mathematical rigour we undertake. This level of mathematical rigour is not necessarily needed, if all one wishes to do is deal with $\mathbb{R}$-valued functions defined on $\mathbb{R}$ (as is done in most engineering undergraduate programs). However, we will wish to use the ideas introduced in this chapter, particularly those from Section 3.1, in contexts far more general than the simple one of $\mathbb{R}$-valued functions. Therefore, it will be helpful, at least, to understand the simple material in this chapter in the rigorous manner in which it is presented.

As for the more advanced material, such as is contained in Sections **??**, **??**, and **??**, it is probably best left aside on a first reading. The reader will be warned when this material is needed in the presentation.

Some of what we cover in this chapter, particularly notions of continuity, differentiability, and Riemann integrability, will be covered in more generality in Chapter 4. Aggressive readers may want to skip this material here and proceed directly to the more general case. •

Contents

Section 3.1

Continuous $\mathbb{R}$-valued functions on $\mathbb{R}$

The notion of continuity is one of the most important in all of mathematics. Here we present this important idea in its simplest form: continuity for functions whose domain and range are subsets of $\mathbb{R}$.

Do I need to read this section? Unless you are familiar with this material, it is probably a good idea to read this section fairly carefully. It builds on the structure of $\mathbb{R}$ built up in Chapter 2 and uses this structure in an essential way. It is essential to understand this if one is to understand the more general ideas of continuity that will arise in Chapter **??**. This section also provides an opportunity to improve one's facility with the $\epsilon - \delta$ formalism. •

3.1.1 Definition and properties of continuous functions

In this section we will deal with functions defined on an interval $I \subseteq \mathbb{R}$. This interval might be open, closed, or neither, and bounded, unbounded, or neither. In this section, we shall reserve the letter I to denote such a general interval. It will also be convenient to say that a subset $A \subseteq I$ is ***open*** if $A = U \cap I$ for an open subset U of $\mathbb{R}$.[1] For example, if $I = [0, 1]$, then the subset $[0, \frac{1}{2})$ is an open subset of I, but not an open subset of $\mathbb{R}$. We will be careful to explicitly say that a subset is open *in* I if this is what we mean. *There is a chance for confusion here, so the reader is advised to be alert!*

Let us give the standard definition of continuity.

3.1.1 Definition (Continuous function) Let $I \subseteq \mathbb{R}$ be an interval. A map $f : I \to \mathbb{R}$ is:

(i) ***continuous at*** $\mathbf{x_0} \in \mathbf{I}$ if, for every $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(x_0)| < \epsilon$ whenever $x \in I$ satisfies $|x - x_0| < \delta$;

(ii) ***continuous*** if it is continuous at each $x_0 \in I$;

(iii) ***discontinuous at*** $\mathbf{x_0} \in \mathbf{I}$ if it is not continuous at x_0;

(iv) ***discontinuous*** if it is not continuous. •

The idea behind the definition of continuity is this: one can make the values of a continuous function as close as desired by making the points at which the function is evaluated sufficiently close. Readers not familiar with the definition should be prepared to spend some time embracing it. An often encountered oversimplification of continuity is illustrated in Figure 3.1. The idea is supposed to be that the function whose graph is shown on the left is continuous because its graph has no "gaps," whereas the function on the right is discontinuous because

[1] This is entirely related to the notion of relative topology which we will discuss in Section 4.2.8 for sets of multiple real variables and in Definition **??** within the general context of topological spaces.

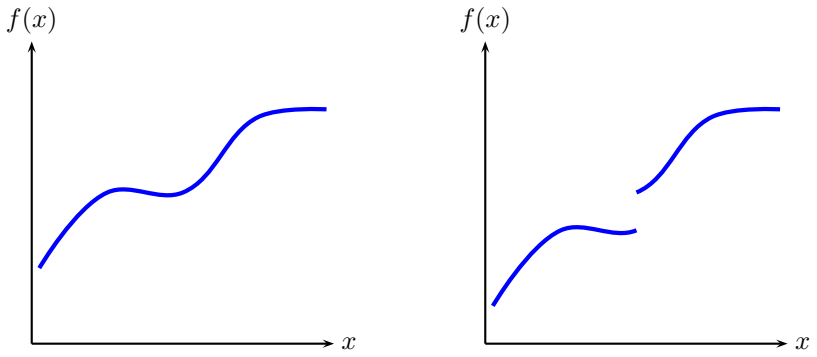


Figure 3.1 Probably not always the best way to envision continuity versus discontinuity

its graph does have a "gap." As we shall see in Example 3.1.2–4 below, it is possible for a function continuous at a point to have a graph with lots of "gaps" in a neighbourhood of that point. Thus the "graph gap" characterisation of continuity is a little misleading.

Let us give some examples of functions that are continuous or not. More examples of discontinuous functions are given in Example 3.1.9 below. We suppose the reader to be familiar with the usual collection of "standard functions," at least for the moment. We shall consider some such functions in detail in Section 3.6.

3.1.2 Examples (Continuous and discontinuous functions)

1. For $\alpha \in \mathbb{R}$, define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = \alpha$. Since $|f(x) - f(x_0)| = 0$ for all $x, x_0 \in \mathbb{R}$, it follows immediately that f is continuous.
2. Define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = x$. For $x_0 \in \mathbb{R}$ and $\epsilon \in \mathbb{R}_{>0}$ take $\delta = \epsilon$. It then follows that if $|x - x_0| < \delta$ then $|f(x) - f(x_0)| < \epsilon$, giving continuity of f.
3. Define $f\colon \mathbb{R} \to \mathbb{R}$ by
$$f(x) = \begin{cases} x \sin \frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$
We claim that f is continuous. We first note that the functions $f_1, f_2\colon \mathbb{R} \to \mathbb{R}$ defined by
$$f_1(x) = x, \quad f_2(x) = \sin x$$
are continuous. Indeed, f_1 is continuous from part 2 and in Section 3.6 we will prove that f_2 is continuous. The function $f_3\colon \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by $f_3(x) = \frac{1}{x}$ is continuous on any interval not containing 0 by Proposition 3.1.15 below. It then follows from Propositions 3.1.15 and 3.1.16 below that f is continuous at x_0, provided that $x_0 \neq 0$. To show continuity at $x = 0$, let $\epsilon \in \mathbb{R}_{>0}$ and take $\delta = \epsilon$. Then, provided that $|x| < \delta$,
$$|f(x) - f(0)| = \left|x \sin \tfrac{1}{x}\right| \leq |x| < \epsilon,$$
using the fact that image(sin) $\subseteq [-1, 1]$. This shows that f is continuous at 0, and so is continuous.

4. Define $f\colon \mathbb{R} \to \mathbb{R}$ by
$$f(x) = \begin{cases} x, & x \in \mathbb{Q}, \\ 0, & \text{otherwise.} \end{cases}$$
We claim that f is continuous at $x_0 = 0$ and discontinuous everywhere else.
To see that f is continuous at $x_0 = 0$, let $\epsilon \in \mathbb{R}_{>0}$ and choose $\delta = \epsilon$. Then, for $|x - x_0| < \delta$ we have either $f(x) = x$ or $f(x) = 0$. In either case, $|f(x) - f(x_0)| < \epsilon$, showing that f is indeed continuous at $x_0 = 0$. Note that this is a function whose continuity at $x_0 = 0$ is not subject to an interpretation like that of Figure 3.1 since the graph of f has an uncountable number of "gaps" near 0.
Next we show that f is discontinuous at x_0 for $x_0 \neq 0$. We have two possibilities.
   (a) $x_0 \in \mathbb{Q}$: Let $\epsilon < \frac{1}{2}|x_0|$. For any $\delta \in \mathbb{R}_{>0}$ the set $\mathsf{B}(\delta, x_0)$ will contain points $x \in \mathbb{R}$ for which $f(x) = 0$. Thus for any $\delta \in \mathbb{R}_{>0}$ the set $\mathsf{B}(\delta, x_0)$ will contain points x such that $|f(x) - f(x_0)| = |x_0| > \epsilon$. This shows that f is discontinuous at nonzero rational numbers.
   (b) $x_0 \in \mathbb{R} \setminus \mathbb{Q}$: Let $\epsilon = \frac{1}{2}|x_0|$. For any $\delta \in \mathbb{R}_{>0}$ we claim that the set $\mathsf{B}(\delta, x_0)$ will contain points $x \in \mathbb{R}$ for which $|f(x)| > \epsilon$ (why?). It then follows that for any $\delta \in \mathbb{R}_{>0}$ the set $\mathsf{B}(\delta, x_0)$ will contain points x such that $|f(x) - f(x_0)| = |f(x)| > \epsilon$, so showing that f is discontinuous at all irrational numbers.
5. Let $I = (0, \infty)$ and on I define the function $f\colon I \to \mathbb{R}$ by $f(x) = \frac{1}{x}$. It follows from Proposition 3.1.15 below that f is continuous on I.
6. Next take $I = [0, \infty)$ and define $f\colon I \to \mathbb{R}$ by
$$f(x) = \begin{cases} \frac{1}{x}, & x \in \mathbb{R}_{>0}, \\ 0, & x = 0. \end{cases}$$
In the previous example we saw that f is continuous at all points in $(0, \infty)$. However, at $x = 0$ the function is discontinuous, as is easily verified. •

The following alternative characterisations of continuity are sometimes useful. The first of these, part (ii) in the theorem, will also be helpful in motivating the general definition of continuity given for topological spaces in Section **??**. The reader will wish to recall from Notation 2.3.28 the notation $\lim_{x \to_I x_0} f(x)$ for taking limits in intervals.

3.1.3 Theorem (Alternative characterisations of continuity) *For a function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *defined on an interval* I *and for* $\mathrm{x}_0 \in \mathrm{I}$*, the following statements are equivalent:*

(i) f *is continuous at* x_0*;*

(ii) for every neighbourhood V *of* $\mathrm{f}(\mathrm{x}_0)$ *there exists a neighbourhood* U *of* x_0 *in* I *such that* $\mathrm{f}(\mathrm{U}) \subseteq \mathrm{V}$*;*

(iii) $\lim_{\mathrm{x} \to_{\mathrm{I}} \mathrm{x}_0} \mathrm{f}(\mathrm{x}) = \mathrm{f}(\mathrm{x}_0)$.

Proof (i) $\implies$ (ii) Let $V \subseteq \mathbb{R}$ be a neighbourhood of $f(x_0)$. Let $\epsilon \in \mathbb{R}_{>0}$ be defined such that $\mathsf{B}(\epsilon, f(x_0)) \subseteq V$, this being possible since V is open. Since f is continuous at x_0,

there exists $\delta \in \mathbb{R}_{>0}$ such that, if $x \in \mathsf{B}(\delta, x_0) \cap I$, then we have $f(x) \in \mathsf{B}(\epsilon, f(x_0))$. This shows that, around the point x_0, we can find an open set in I whose image lies in V.

(ii) $\implies$ (iii) Let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in I converging to x_0 and let $\epsilon \in \mathbb{R}_{>0}$. By hypothesis there exists a neighbourhood U of x_0 in I such that $f(U) \subseteq \mathsf{B}(\epsilon, f(x_0))$. Thus there exists $\delta \in \mathbb{R}_{>0}$ such that $f(\mathsf{B}(\delta, x_0) \cap I) \subseteq \mathsf{B}(\epsilon, f(x_0))$ since U is open in I. Now choose $N \in \mathbb{Z}_{>0}$ sufficiently large that $|x_j - x_0| < \delta$ for $j \geq N$. It then follows that $|f(x_j) - f(x_0)| < \epsilon$ for $j \geq N$, so giving convergence of $(f(x_j))_{j \in \mathbb{Z}_{>0}}$ to $f(x_0)$, as desired, after an application of Proposition 2.3.29.

(iii) $\implies$ (i) Let $\epsilon \in \mathbb{R}_{>0}$. Then, by definition of $\lim_{x \to_I x_0} f(x) = f(x_0)$, there exists $\delta \in \mathbb{R}_{>0}$ such that, for $x \in \mathsf{B}(\delta, x_0) \cap I$, $|f(x) - f(x_0)| < \epsilon$, which is exactly the definition of continuity of f at x_0. ■

3.1.4 Corollary *For an interval* $\mathrm{I} \subseteq \mathbb{R}$, *a function* $\mathrm{f} \colon \mathrm{I} \to \mathbb{R}$ *is continuous if and only if* $\mathrm{f}^{-1}(\mathrm{V})$ *is open in* I *for every open subset* V *of* $\mathbb{R}$.

Proof Suppose that f is continuous. If $V \cap \text{image}(f) = \emptyset$ then clearly $f^{-1}(V) = \emptyset$ which is open. So assume that $V \cap \text{image}(f) \neq \emptyset$ and let $x \in f^{-1}(V)$. Since f is continuous at x and since V is a neighbourhood of $f(x)$, there exists a neighbourhood U of x such that $f(U) \subseteq V$. Thus $U \subseteq f^{-1}(V)$, showing that $f^{-1}(V)$ is open.

Now suppose that $f^{-1}(V)$ is open for each open set V and let $x \in \mathbb{R}$. If V is a neighbourhood of $f(x)$ then $f^{-1}(V)$ is open. Then there exists a neighbourhood U of x such that $U \subseteq f^{-1}(V)$. By Proposition 1.3.5 we have $f(U) \subseteq f(f^{-1}(V)) \subseteq V$, thus showing that f is continuous. ■

The reader can explore these alternative representations of continuity in Exercise 3.1.9.

A stronger notion of continuity is sometimes useful. As well, the following definition introduces for the first time the important notion of "uniform."

3.1.5 Definition (Uniform continuity) Let $I \subseteq \mathbb{R}$ be an interval. A map $f \colon I \to \mathbb{R}$ is ***uniformly continuous*** if, for every $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x_1) - f(x_2)| < \epsilon$ whenever $x_1, x_2 \in I$ satisfy $|x_1 - x_2| < \delta$. •

3.1.6 Remark (On the idea of "uniformly") In the preceding definition we have encountered for the first time the idea of a property holding "uniformly." This is an important idea that comes up often in mathematics. Moreover, it is an idea that is often useful in applications of mathematics, since the absence of a property holding "uniformly" can have undesirable consequences. Therefore, we shall say some things about this here.

In fact, the comparison of continuity versus uniform continuity is a good one for making clear the character of something holding "uniformly." Let us compare the definitions.

1. One defines continuity of a function at a point x_0 by asking that, for each $\epsilon \in \mathbb{R}_{>0}$, one can find $\delta \in \mathbb{R}_{>0}$ such that if x is within δ of x_0, then $f(x)$ is within ϵ of $f(x_0)$. Note that δ will generally depend on ϵ, and most importantly for our discussion here, on x_0. Often authors explicitly write $\delta(\epsilon, x_0)$ to denote this dependence of δ on both ϵ and x_0.

2. One defines uniform continuity of a function on the interval I by asking that, for each $\epsilon \in \mathbb{R}_{>0}$, one can find $\delta \in \mathbb{R}_{>0}$ such that if x_1 and x_2 are within δ of one another, then $f(x_1)$ and $f(x_2)$ are within ϵ of one another. Here, the number δ depends *only* on ϵ. Again, to reflect this, some authors explicitly write $\delta(\epsilon)$, or state explicitly that δ is independent of x.

The idea of "uniform" then is that a property, in this case the existence of $\delta \in \mathbb{R}_{>0}$ with a certain property, holds for the entire set I, and not just for a single point. •

Let us give an example to show that uniformly continuous is not the same as continuous.

3.1.7 Example (Uniform continuity versus continuity) Let us give an example of a function that is continuous, but not uniformly continuous. Define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. We first show that f is continuous at each point $x_0 \in \mathbb{R}$. Let $\epsilon \in \mathbb{R}_{>0}$ and choose δ such that $2|x_0|\delta + \delta^2 < \epsilon$ (why is this possible?). Then, provided that $|x - x_0| < \delta$, we have

$$\begin{aligned} |f(x) - f(x_0)| = |x^2 - x_0^2| &= |x - x_0||x + x_0| \\ &\leq |x - x_0|(|x| + |x_0|) \leq |x - x_0|(2|x_0| + |x - x_0|) \\ &\leq \delta(2|x_0| + \delta) < \epsilon. \end{aligned}$$

Thus f is continuous.

Now let us show that f is not uniformly continuous. We will show that there exists $\epsilon \in \mathbb{R}_{>0}$ such that there is no $\delta \in \mathbb{R}_{>0}$ for which $|x - x_0| < \delta$ ensures that $|f(x) - f(x_0)| < \epsilon$ *for all* x_0. Let us take $\epsilon = 1$ and let $\delta \in \mathbb{R}_{>0}$. Then define $x_0 \in \mathbb{R}$ such that $\frac{\delta}{2}\left|2x_0 + \frac{\delta}{2}\right| > 1$ (why is this possible?). We then note that $x = x_0 + \frac{\delta}{2}$ satisfies $|x - x_0| < \delta$, but that

$$|f(x) - f(x_0)| = |x^2 - x_0^2| = |x - x_0||x + x_0| = \tfrac{\delta}{2}\left|2x_0 + \tfrac{\delta}{2}\right| > 1 = \epsilon.$$

This shows that f is not uniformly continuous. •

3.1.2 Discontinuous functions[2]

It is often useful to be specific about the nature of a discontinuity of a function that is not continuous. The following definition gives names to all possibilities. The reader may wish to recall from Section 2.3.7 the discussion concerning taking limits using an index set that is a subset of $\mathbb{R}$.

3.1.8 Definition (Types of discontinuity) Let $I \subseteq \mathbb{R}$ be an interval and suppose that $f\colon I \to \mathbb{R}$ is discontinuous at $x_0 \in I$. The point x_0 is:

(i) a ***removable discontinuity*** if $\lim_{x \to_I x_0} f(x)$ exists;

(ii) a ***discontinuity of the first kind***, or a ***jump discontinuity***, if the limits $\lim_{x \downarrow x_0} f(x)$ and $\lim_{x \uparrow x_0} f(x)$ exist;

[2]This section is rather specialised and technical and so can be omitted until needed. However, the material is needed at certain points in the text.

(iii) a ***discontinuity of the second kind***, or an ***essential discontinuity***, if at least one of the limits $\lim_{x\downarrow x_0} f(x)$ and $\lim_{x\uparrow x_0} f(x)$ does not exist.

The set of all discontinuities of f is denoted by D_f. •

In Figure 3.2 we depict the various sorts of discontinuity. We can also illustrate

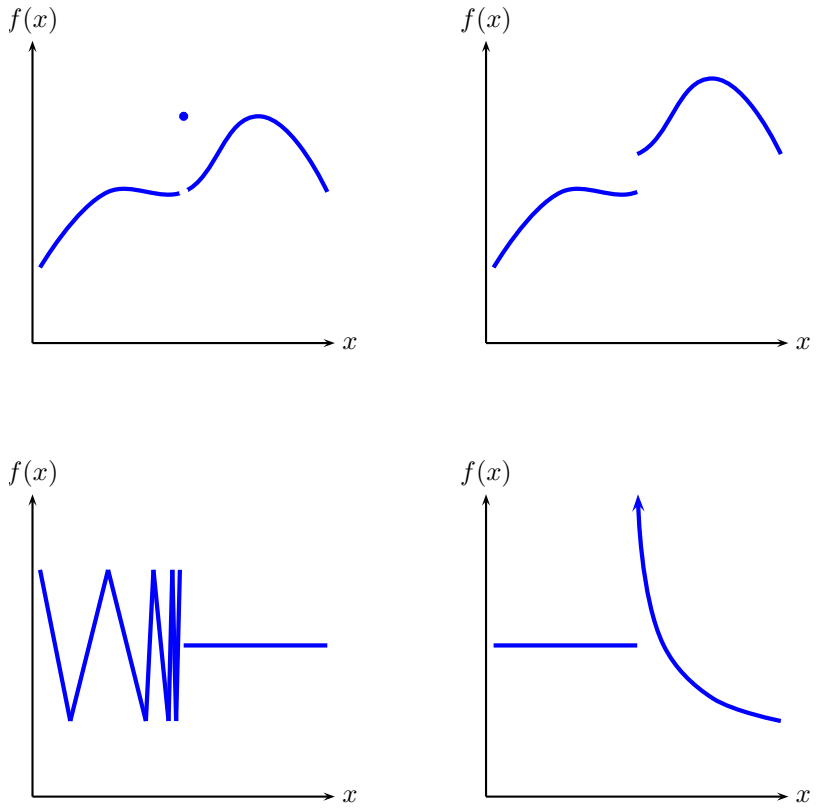


Figure 3.2 A removable discontinuity (top left), a jump discontinuity (top right), and two essential discontinuities (bottom)

these with explicit examples.

3.1.9 Examples (Types of discontinuities)

1. Let $I = [0,1]$ and let $f\colon I \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} x, & x \in (0,1], \\ 1, & x = 0. \end{cases}$$

It is clear that f is continuous for all $x \in (0,1]$, and is discontinuous at $x = 0$. However, since we have $\lim_{x\to_I 0} f(x) = 0$ (note that the requirement that this limit be taken in I amounts to the fact that the limit is given by $\lim_{x\downarrow 0} f(x) = 0$), it follows that the discontinuity is removable.

Note that one might be tempted to also say that the discontinuity is a jump discontinuity since the limit $\lim_{x\downarrow 0} f(x)$ exists and since the limit $\lim_{x\uparrow 0} f(x)$

cannot be defined here since 0 is a left endpoint for I. However, we do require that both limits exist at a jump discontinuity, which has as a consequence the fact that jump discontinuities can only occur at interior points of an interval.

2. Let $I = [-1, 1]$ and define $f : I \to \mathbb{R}$ by $f(x) = \text{sign}(x)$. We may easily see that f is continuous at $x \in [-1, 1] \setminus \{0\}$, and is discontinuous at $x = 0$. Then, since we have $\lim_{x \downarrow 0} f(x) = 1$ and $\lim_{x \uparrow 0} f(x) = -1$, it follows that the discontinuity at 0 is a jump discontinuity.
3. Let $I = [-1, 1]$ and define $f : I \to \mathbb{R}$ by

$$f(x) = \begin{cases} \sin \frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

Then, by Proposition 3.1.15 (and accepting continuity of sin), f is continuous at $x \in [-1, 1] \setminus \{0\}$. At $x = 0$ we claim that we have an essential discontinuity. To see this we note that, for any $\epsilon \in \mathbb{R}_{>0}$, the function f restricted to $[0, \epsilon)$ and $(-\epsilon, 0]$ takes all possible values in set $[-1, 1]$. This is easily seen to preclude existence of the limits $\lim_{x \downarrow 0} f(x)$ and $\lim_{x \uparrow 0} f(x)$.

4. Let $I = [-1, 1]$ and define $f : I \to \mathbb{R}$ by

$$f(x) = \begin{cases} \frac{1}{x}, & x \in (0, 1], \\ 0, & x \in [-1, 0]. \end{cases}$$

Then f is continuous at $x \in [-1, 1] \setminus \{0\}$ by Proposition 3.1.15. At $x = 0$ we claim that f has an essential discontinuity. Indeed, we have $\lim_{x \downarrow} f(x) = \infty$, which precludes f having a removable or jump discontinuity at $x = 0$. •

The following definition gives a useful quantitative means of measuring the discontinuity of a function.

3.1.10 Definition (Oscillation) Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be a function. The ***oscillation*** of f is the function $\omega_f : I \to \mathbb{R}$ defined by

$$\omega_f(x) = \inf\{\sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\} \mid \delta \in \mathbb{R}_{>0}\}. \quad \bullet$$

Note that the definition makes sense since the function

$$\delta \mapsto \sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\}$$

is monotonically increasing (see Definition 3.1.27 for a definition of monotonically increasing in this context). In particular, if f is bounded (see Definition 3.1.20 below) then ω_f is also bounded. The following result indicates in what way ω_f measures the continuity of f.

3.1.11 Proposition (Oscillation measures discontinuity) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$ *and a function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$, f *is continuous at* $\mathrm{x} \in \mathrm{I}$ *if and only if* $\omega_{\mathrm{f}}(\mathrm{x}) = 0$.

Proof Suppose that f is continuous at x and let $\epsilon \in \mathbb{R}_{>0}$. Choose $\delta \in \mathbb{R}_{>0}$ such that if $y \in \mathsf{B}(\delta, x) \cap I$ then $|f(y) - f(x)| < \frac{\epsilon}{2}$. Then, for $x_1, x_2 \in \mathsf{B}(\delta, x)$ we have

$$|f(x_1) - f(x_2)| \leq |f(x_1) - f(x)| + |f(x) - f(x_2)| < \epsilon.$$

Therefore,

$$\sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\} < \epsilon.$$

Since ϵ is arbitrary this gives

$$\inf\{\sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\} \mid \delta \in \mathbb{R}_{>0}\} = 0,$$

meaning that $\omega_f(x) = 0$.

Now suppose that $\omega_f(x) = 0$. For $\epsilon \in \mathbb{R}_{>0}$ let $\delta \in \mathbb{R}_{>0}$ be chosen such that

$$\sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\} < \epsilon.$$

In particular, $|f(y) - f(x)| < \epsilon$ for all $y \in \mathsf{B}(\delta, x) \cap I$, giving continuity of f at x. ■

Let us consider a simple example.

3.1.12 Example (Oscillation for a discontinuous function) We let $I = [-1, 1]$ and define $f\colon I \to \mathbb{R}$ by $f(x) = \operatorname{sign}(x)$. It is then easy to see that

$$\omega_f(x) = \begin{cases} 0, & x \neq 0, \\ 2, & x = 0. \end{cases}$$ •

We close this section with a technical property of the oscillation of a function. This property will be useful during the course of some proofs in the text.

3.1.13 Proposition (Closed preimages of the oscillation of a function) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a function. Then, for every* $\alpha \in \mathbb{R}_{\geq 0}$, *the set*

$$\mathrm{A}_\alpha = \{\mathrm{x} \in \mathrm{I} \mid \omega_{\mathrm{f}}(\mathrm{x}) \geq \alpha\}$$

is closed in I.

Proof The result where $\alpha = 0$ is clear, so we assume that $\alpha \in \mathbb{R}_{>0}$. For $\delta \in \mathbb{R}_{>0}$ define

$$\omega_f(x, \delta) = \sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\}$$

so that $\omega_f(x) = \lim_{\delta \to 0} \omega_f(x, \delta)$. Let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in A_α converging to $x \in \mathbb{R}$ and let $(\epsilon_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $(0, \alpha)$ converging to zero. Let $j \in \mathbb{Z}_{>0}$. We claim that there exists points $y_j, z_j \in \mathsf{B}(\epsilon_j, x_j) \cap I$ such that $|f(y_j) - f(z_j)| \geq \alpha - \epsilon_j$. Suppose otherwise so that for every $y, z \in \mathsf{B}(\epsilon_j, x_j) \cap I$ we have $|f(y) - f(z)| < \alpha - \epsilon_j$. It then follows that $\lim_{\delta \to 0} \omega_f(x_j, \delta) \leq \alpha - \epsilon_j < \alpha$, contradicting the fact that $x_j \in A_\alpha$. We claim that $(y_j)_{j \in \mathbb{Z}_{>0}}$ and $(z_j)_{j \in \mathbb{Z}_{>0}}$ converge to x. Indeed, let $\epsilon \in \mathbb{R}_{>0}$ and choose $N_1 \in \mathbb{Z}_{>0}$ sufficiently large that $\epsilon_j < \frac{\epsilon}{2}$ for $j \geq N_1$ and choose $N_2 \in \mathbb{Z}_{>0}$ such that $|x_j - x| < \frac{\epsilon}{2}$ for $j \geq N_2$. Then, for $j \geq \max\{N_1, N_2\}$ we have

$$|y_j - x| \leq |y_j - x_j| + |x_j - x| < \epsilon.$$

Thus $(y_j)_{j\in\mathbb{Z}_{>0}}$ converges to x, and the same argument, and therefore the same conclusion, also applies to $(z_j)_{j\in\mathbb{Z}_{>0}}$.

Thus we have sequences of points $(y_j)_{j\in\mathbb{Z}_{>0}}$ and $(z_j)_{j\in\mathbb{Z}_{>0}}$ in I converging to x and a sequence $(\epsilon_j)_{j\in\mathbb{Z}_{>0}}$ in $(0,\alpha)$ converging to zero for which $|f(y_j) - f(z_j)| \geq \alpha - \epsilon_j$. We claim that this implies that $\omega_f(x) \geq \alpha$. Indeed, suppose that $\omega_f(x) < \alpha$. There exists $N \in \mathbb{Z}_{>0}$ such that $\alpha - \epsilon_j > \alpha - \omega_f(x)$ for every $j \geq N$. Therefore,

$$|f(y_j) - f(z_j)| \geq \alpha - \epsilon_j > \alpha - \omega_f(x)$$

for every $j \geq N$. This contradicts the definition of $\omega_f(x)$ since the sequences $(y_j)_{j\in\mathbb{Z}_{>0}}$ and $(z_j)_{j\in\mathbb{Z}_{>0}}$ converge to x.

Now we claim that the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to x. Let $\epsilon \in \mathbb{R}_{>0}$ and let $N_1 \in \mathbb{Z}_{>0}$ be large enough that $|x - y_j| < \frac{\epsilon}{2}$ for $j \geq N_1$ and let $N_2 \in \mathbb{Z}_{>0}$ be large enough that $\epsilon_j < \frac{\epsilon}{2}$ for $j \geq N_2$. Then, for $j \geq \max\{N_1, N_2\}$ we have

$$|x - x_j| \leq |x - y_j| + |y_j - x_j| < \epsilon,$$

as desired.

This shows that every sequence in A_α converges to a point in A_α. It follows from Exercise 2.5.2 that A_α is closed. ■

The following corollary is somewhat remarkable, in that it shows that the set of discontinuities of a function cannot be arbitrary.

3.1.14 Corollary (Discontinuities are the countable union of closed sets) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a function. Then the set*

$$\mathrm{D_f} = \{\mathrm{x} \in \mathrm{I} \mid \ \mathrm{f} \textit{ is not continuous at } \mathrm{x}\}$$

is the countable union of closed sets.

Proof This follows immediately from Proposition 3.1.13 after we note that

$$D_f = \cup_{k\in\mathbb{Z}_{>0}}\{x \in I \mid \ \omega_f(x) \geq \tfrac{1}{k}\}.$$ ■

missing stuff

3.1.3 Continuity and operations on functions

Let us consider how continuity behaves relative to simple operations on functions. To do so, we first note that, given an interval I and two functions $f, g\colon I \to \mathbb{R}$, one can define two functions $f + g, fg\colon I \to \mathbb{R}$ by

$$(f + g)(x) = f(x) + g(x), \qquad (fg)(x) = f(x)g(x),$$

respectively. Moreover, if $g(x) \neq 0$ for all $x \in I$, then we define

$$\Big(\frac{f}{g}\Big)(x) = \frac{f(x)}{g(x)}.$$

Thus one can add and multiply $\mathbb{R}$-valued functions using the operations of addition and multiplication in $\mathbb{R}$.

3.1.15 Proposition (Continuity, and addition and multiplication) *For an interval* $I \subseteq \mathbb{R}$, *if* $f, g \colon I \to \mathbb{R}$ *are continuous at* $x_0 \in I$, *then both* $f + g$ *and* fg *are continuous at* x_0. *If additionally* $g(x) \neq 0$ *for all* $x \in I$, *then* $\frac{f}{g}$ *is continuous at* x_0.

Proof To show that $f + g$ and fg are continuous at x_0 if f and g are continuous at x_0, let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in I converging to x_0. Then, by Theorem 3.1.3 the sequences $(f(x_j))_{j \in \mathbb{Z}_{>0}}$ and $(g(x_j))_{j \in \mathbb{Z}_{>0}}$ converge to $f(x_0)$ and $g(x_0)$, respectively. Then, by Proposition 2.3.23, the sequences $(f(x_j) + g(x_j))_{j \in \mathbb{Z}_{>0}}$ and $(f(x_j)g(x_j))_{j \in \mathbb{Z}_{>0}}$ converge to $f(x_0) + g(x_0)$ and $f(x_0)g(x_0)$, respectively. Then $\lim_{j \to \infty}(f + g)(x_j) = (f + g)(x_0)$ and $\lim_{j \to \infty}(fg)(x_j) = (fg)(x_0)$, and the result follows by Proposition 2.3.29 and Theorem 3.1.3.

Now suppose that $g(x) \neq 0$ for every $x \in I$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that $|g(x_0)| > 2\epsilon$. By Theorem 3.1.3 take $\delta \in \mathbb{R}_{>0}$ such that $g(\mathsf{B}(\delta, x_0)) \subseteq \mathsf{B}(\epsilon, g(x_0))$. Thus g is nonzero on the ball $\mathsf{B}(\delta, x_0)$. Now let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $\mathsf{B}(\delta, x_0)$ converging to x_0. Then, as above, the sequences $(f(x_j))_{j \in \mathbb{Z}_{>0}}$ and $(g(x_j))_{j \in \mathbb{Z}_{>0}}$ converge to $f(x_0)$ and $g(x_0)$, respectively. We can now employ Proposition 2.3.23 to conclude that the sequence $\left(\frac{f(x_j)}{g(x_j)}\right)_{j \in \mathbb{Z}_{>0}}$ converges to $\frac{f(x_0)}{g(x_0)}$, and the last part of the result follows by Proposition 2.3.29 and Theorem 3.1.3. ■

3.1.16 Proposition (Continuity and composition) *Let* $I, J \subseteq \mathbb{R}$ *be intervals and let* $f \colon I \to J$ *and* $f \colon J \to \mathbb{R}$ *be continuous at* $x_0 \in I$ *and* $f(x_0) \in J$, *respectively. Then* $g \circ f \colon I \to \mathbb{R}$ *is continuous at* x_0.

Proof Let W be a neighbourhood of $g \circ f(x_0)$. Since g is continuous at $f(x_0)$ there exists a neighbourhood V of $f(x_0)$ such that $g(V) \subseteq W$. Since f is continuous at x_0 there exists a neighbourhood U of x_0 such that $f(U) \subseteq V$. Clearly $g \circ f(U) \subseteq W$, and the result follows from Theorem 3.1.3. ■

3.1.17 Proposition (Continuity and restriction) *If* $I, J \subseteq \mathbb{R}$ *are intervals for which* $J \subseteq I$, *and if* $f \colon I \to \mathbb{R}$ *is continuous at* $x_0 \in J \subseteq I$, *then* $f|J$ *is continuous at* x_0.

Proof This follows immediately from Theorem 3.1.3, also using Proposition 1.3.5, after one notes that open subsets of J are of the form $U \cap I$ where U is an open subset of I. ■

Note that none of the proofs of the preceding results use the definition of continuity, but actually use the alternative characterisations of Theorem 3.1.3. Thus these alternative characterisations, while less intuitive initially (particularly the one involving open sets), they are in fact quite useful.

Let us finally consider the behaviour of continuity with respect to the operations of selection of maximums and minimums.

3.1.18 Proposition (Continuity and min and max) *If* $I \subseteq \mathbb{R}$ *is an interval and if* $f, g \colon I \to \mathbb{R}$ *are continuous functions, then the functions*

$$I \ni x \mapsto \min\{f(x), g(x)\} \in \mathbb{R}, \qquad I \ni x \mapsto \max\{f(x), g(x)\} \in \mathbb{R}$$

are continuous.

Proof Let $x_0 \in I$ and let $\epsilon \in \mathbb{R}_{>0}$. Let us first assume that $f(x_0) > g(x_0)$. That is to say, assume that $(f-g)(x_0) \in \mathbb{R}_{>0}$. Continuity of f and g ensures that there exists $\delta_1 \in \mathbb{R}_{>0}$ such that if $x \in \mathsf{B}(\delta_1, x_0) \cap I$ then $(f-g)(x) \in \mathbb{R}_{>0}$. That is, if $x \in \mathsf{B}(\delta_1, x_0) \cap I$ then

$$\min\{f(x), g(x)\} = g(x), \quad \max\{f(x), g(x)\} = f(x).$$

Continuity of f ensures that there exists $\delta_2 \in \mathbb{R}_{>0}$ such that if $x \in \mathsf{B}(\delta_2, x_0) \cap I$ then $|f(x) - f(x_0)| < \epsilon$. Similarly, continuity of f ensures that there exists $\delta_3 \in \mathbb{R}_{>0}$ such that if $x \in \mathsf{B}(\delta_3, x_0) \cap I$ then $|g(x) - g(x_0)| < \epsilon$. Let $\delta_4 = \min\{\delta_1, \delta_2\}$. If $x \in \mathsf{B}(\delta_4, x_0) \cap I$ then

$$|\min\{f(x), g(x)\} - \min\{f(x_0), g(x_0)\}| = |g(x) - g(x_0)| < \epsilon$$

and

$$|\max\{f(x), g(x)\} - \max\{f(x_0), g(x_0)\}| = |f(x) - f(x_0)| < \epsilon.$$

This gives continuity of the two functions in this case. Similarly, swapping the rôle of f and g, if $f(x_0) < g(x_0)$ one can arrive at the same conclusion. Thus we need only consider the case when $f(x_0) = g(x_0)$. In this case, by continuity of f and g, choose $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(x_0)| < \epsilon$ and $|g(x) - g(x_0)| < \epsilon$ for $x \in \mathsf{B}(\delta, x_0) \cap I$. Then let $x \in \mathsf{B}(\delta, x_0) \cap I$. If $f(x) \geq g(x)$ then we have

$$|\min\{f(x), g(x)\} - \min\{f(x_0), g(x_0)\}| = |g(x) - g(x_0)| < \epsilon$$

and

$$|\max\{f(x), g(x)\} - \max\{f(x_0), g(x_0)\}| = |f(x) - f(x_0)| < \epsilon.$$

This gives the result in this case, and one similarly gets the result when $f(x) < g(x)$. ■

3.1.4 Continuity, and compactness and connectedness

In this section we will consider some of the relationships that exist between continuity, and compactness and connectedness. We see here for the first time some of the benefits that can be drawn from the notion of continuity. Moreover, if one studies the proofs of the results in this section, one can see that we use the actual definition of compactness (rather than the simpler alternative characterisation of compact sets as being closed and bounded) to great advantage.

The first result is a simple and occasionally useful one.

3.1.19 Proposition (The continuous image of a compact set is compact) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is a compact interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is continuous, then* image(f) *is compact.*

Proof Let $(U_a)_{a\in A}$ be an open cover of image(f). Then $(f^{-1}(U_a))_{a\in A}$ is an open cover of I, and so there exists a finite subset $(a_1, \ldots, a_k) \subseteq A$ such that $\cup_{j=1}^{k} f^{-1}(U_{a_k}) = I$. It is then clear that $(f(f^{-1}(U_{a_1})), \ldots, f(f^{-1}(U_{a_k})))$ covers image(f). Moreover, by Proposition 1.3.5, $f(f^{-1}(U_{a_j})) \subseteq U_{a_j}$, $j \in \{1, \ldots, k\}$. Thus $(U_{a_1}, \ldots, U_{a_k})$ is a finite subcover of $(U_a)_{a\in A}$. ■

A useful feature that a function might possess is that of having bounded values.

3.1.20 Definition (Bounded function) For an interval I, a function $f\colon I \to \mathbb{R}$ is:

(i) ***bounded*** if there exists $M \in \mathbb{R}_{>0}$ such that image$(f) \subseteq \overline{\mathsf{B}}(M, 0)$;

(ii) ***locally bounded*** if $f|J$ is bounded for every compact interval $J \subseteq I$;

(iii) ***unbounded*** if it is not bounded. •

3.1.21 Remark (On "locally") This is our first encounter with the qualifier "locally" assigned to a property, in this case, of a function. This concept will appear frequently, as for example in this chapter with the notion of "locally bounded variation" (Definition **??**) and "locally absolutely continuous" (Definition **??**). The idea in all cases is the same; that a property holds "locally" if it holds on every compact subset. •

For continuous functions it is sometimes possible to immediately assert boundedness simply from the property of the domain.

3.1.22 Theorem (Continuous functions on compact intervals are bounded) *If* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *is a compact interval, then a continuous function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is bounded.*

Proof Let $x \in I$. As f is continuous, there exists $\delta \in \mathbb{R}_{>0}$ so that $|f(y) - f(x)| < 1$ provided that $|y - x| < \delta$. In particular, if $x \in I$, there is an open interval I_x in I with $x \in I_x$ such that $|f(y)| \leq |f(x)| + 1$ for all $x \in I_x$. Thus f is bounded on I_x. This can be done for each $x \in I$, so defining a family of open sets $(I_x)_{x \in I}$. Clearly $I \subseteq \cup_{x \in I} I_x$, and so, by Theorem 2.5.27, there exists a finite collection of points $x_1, \ldots, x_k \in I$ such that $I \subseteq \cup_{j=1}^{k} I_{x_j}$. Obviously for any $x \in I$,

$$|f(x)| \leq 1 + \max\{f(x_1), \ldots, f(x_k)\},$$

thus showing that f is bounded. ■

In Exercise 3.1.7 the reader can explore cases where the theorem does not hold. Related to the preceding result is the following.

3.1.23 Theorem (Continuous functions on compact intervals achieve their extreme values) *If* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *is a compact interval and if* $\mathrm{f}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is continuous, then there exist points* $\mathrm{x}_{\min}, \mathrm{x}_{\max} \in [\mathrm{a}, \mathrm{b}]$ *such that*

$$\mathrm{f}(\mathrm{x}_{\min}) = \inf\{\mathrm{f}(\mathrm{x}) \mid \mathrm{x} \in [\mathrm{a}, \mathrm{b}]\}, \quad \mathrm{f}(\mathrm{x}_{\max}) = \sup\{\mathrm{f}(\mathrm{x}) \mid \mathrm{x} \in [\mathrm{a}, \mathrm{b}]\}.$$

Proof It suffices to show that f achieves its maximum on I since if f achieves its maximum, then $-f$ will achieve its minimum. So let $M = \sup\{f(x) \mid x \in I\}$, and suppose that there is no point $x_{\max} \in I$ for which $f(x_{\max}) = M$. Then $f(x) < M$ for each $x \in I$. For a given $x \in I$ we have

$$f(x) = \tfrac{1}{2}(f(x) + f(x)) < \tfrac{1}{2}(f(x) + M).$$

Continuity of f ensures that there is an open interval I_x containing x such that, for each $y \in I_x \cap I$, $f(y) < \frac{1}{2}(f(x) + M)$. Since $I \subseteq \cup_{x \in I} I_x$, by the Heine–Borel theorem, there exists a finite number of points $x_1, \ldots, x_k$ such that $I \subseteq \cup_{j=1}^{k} I_{x_j}$. Let $m = \max\{f(x_1), \ldots, f(x_k)\}$ so that, for each $y \in I_{x_j}$, and for each $j \in \{1, \ldots, k\}$, we have

$$f(y) < \tfrac{1}{2}(f(x_j) + M) < \tfrac{1}{2}(m + M),$$

which shows that $\frac{1}{2}(m + M)$ is an upper bound for f. However, since f attains the value m on I, we have $m < M$ and so $\frac{1}{2}(m + M) < M$, contradicting the fact that M is the least upper bound. Thus our assumption that f cannot attain the value M on I is false. ■

The theorem tells us that a continuous function on a bounded interval actually *attains* its maximum and minimum value *on the interval*. You should understand that this is not the case if I is neither closed nor bounded (see Exercise 3.1.8).

Our next result gives our first connection between the concepts of uniformity and compactness. This is something of a theme in analysis where continuity is involved. A good place to begin to understand the relationship between compactness and uniformity is the proof of the following theorem, since it is one of the simplest instances of the phenomenon.

3.1.24 Theorem (Heine–Cantor Theorem) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval. If* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is continuous, then it is uniformly continuous.*

Proof Let $x \in [a, b]$ and let $\epsilon \in \mathbb{R}_{>0}$. Since f is continuous, then there exists $\delta_x \in \mathbb{R}_{>0}$ such that, if $|y - x| < \delta_x$, then $|f(y) - f(x)| < \frac{\epsilon}{2}$. Now define an open interval $I_x = (x - \frac{1}{2}\delta_x, x + \frac{1}{2}\delta_x)$. Note that $[a, b] \subseteq \cup_{x \in [a,b]} I_x$, so that the open sets $(I_x)_{x \in [a,b]}$ cover $[a, b]$. By definition of compactness, there then exists a finite number of open sets from $(I_x)_{x \in [a,b]}$ that cover $[a, b]$. Denote this finite family by $(I_{x_1}, \ldots, I_{x_k})$ for some $x_1, \ldots, x_k \in [a, b]$. Take $\delta = \frac{1}{2} \min\{\delta_{x_1}, \ldots, \delta_{x_k}\}$. Now let $x, y \in [a, b]$ satisfy $|x - y| < \delta$. Then there exists $j \in \{1, \ldots, k\}$ such that $x \in I_{x_j}$ since the sets $I_{x_1}, \ldots, I_{x_k}$ cover $[a, b]$. We also have

$$|y - x_j| = |y - x + x - x_j| \leq |y - x| + |x - x_j| < \tfrac{1}{2}\delta_{x_j} + \tfrac{1}{2}\delta_{x_j} = \delta_{x_j},$$

using the triangle inequality. Therefore,

$$\begin{aligned} |f(y) - f(x)| = |f(y) - f(x_j) + f(x_j) - f(x)| \\ \leq |f(y) - f(x_j)| + |f(x_j) - f(x)| < \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon, \end{aligned}$$

again using the triangle inequality. Since this holds for *any* $x \in [a, b]$, it follows that f is uniformly continuous. ■

Next we give a standard result from calculus that is frequently useful.

3.1.25 Theorem (Intermediate Value Theorem) *Let* I *be an interval and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be continuous. If* $\mathrm{x}_1, \mathrm{x}_2 \in \mathrm{I}$ *then, for any* $\mathrm{y} \in [\mathrm{f}(\mathrm{x}_1), \mathrm{f}(\mathrm{x}_2)]$, *there exists* $\mathrm{x} \in \mathrm{I}$ *such that* $\mathrm{f}(\mathrm{x}) = \mathrm{y}$.

Proof Since otherwise the result is obviously true, we may suppose that $y \in (f(x_1), f(x_2))$. Also, since we may otherwise replace f with $-f$, we may without loss of generality suppose that $x_1 < x_2$. Now define $S = \{x \in [x_1, x_2] \mid f(x) \leq y\}$ and let $x_0 = \sup S$. We claim that $f(x_0) = y$. Suppose not. Then first consider the case where $f(x_0) > y$, and define $\epsilon = f(x_0) - y$. Then there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(x_0)| < \epsilon$ for $|x - x_0| < \delta$. In particular, $f(x_0 - \delta) > y$, contradicting the fact that $x_0 = \sup S$. Next suppose that $f(x_0) < y$. Let $\epsilon = y - f(x_0)$ so that there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(x_0)| < \epsilon$ for $|x - x_0| < \delta$. In particular, $f(x_0 + \delta) < y$, contradicting again the fact that $x_0 = \sup S$. ■

In Figure 3.3 we give the idea of the proof of the Intermediate Value Theorem. There is also a useful relationship between continuity and connected sets.

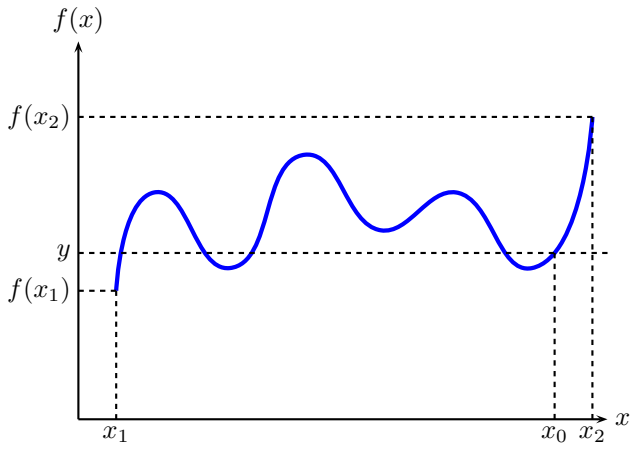


Figure 3.3 Illustration of the Intermediate Value Theorem

3.1.26 Proposition (The continuous image of a connected set is connected) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval, if* $\mathrm{S} \subseteq \mathrm{I}$ *is connected, and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is continuous, then* $\mathrm{f}(\mathrm{S})$ *is connected.*

Proof Suppose that $f(S)$ is not connected. Then there exist nonempty separated sets A and B such that $f(S) = A \cup B$. Let $C = S \cap f^{-1}(A)$ and $D = S \cap f^{-1}(B)$. By Propositions 1.1.4 and 1.3.5 we have

$$\begin{aligned} C \cup D &= (S \cap f^{-1}(A)) \cup (S \cap f^{-1}(B)) \\ &= S \cap (f^{-1}(A) \cup f^{-1}(B)) = S \cap f^{-1}(A \cup B) = S. \end{aligned}$$

By Propositions 2.5.20 and 1.3.5, and since $f^{-1}(\mathrm{cl}(A))$ is closed, we have

$$\mathrm{cl}(C) = \mathrm{cl}(f^{-1}(A)) \subseteq \mathrm{cl}(f^{-1}(\mathrm{cl}(A)) = f^{-1}(\mathrm{cl}(A)).$$

We also clearly have $D \subseteq f^{-1}(B)$. Therefore, by Proposition 1.3.5,

$$\mathrm{cl}(C) \cap D \subseteq f^{-1}(\mathrm{cl}(A)) \cap f^{-1}(B) = f^{-1}(\mathrm{cl}(A) \cap B) = \emptyset.$$

We also similarly have $C \cap \mathrm{cl}(D) = \emptyset$. Thus S is not connected, which gives the result. ■

3.1.5 Monotonic functions and continuity

In this section we consider a special class of functions, namely those that are "increasing" or "decreasing."

3.1.27 Definition (Monotonic function) For $I \subseteq \mathbb{R}$ an interval, a function $f\colon I \to \mathbb{R}$ is:

(i) ***monotonically increasing*** if, for every $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) \leq f(x_2)$;

(ii) ***strictly monotonically increasing*** if, for every $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) < f(x_2)$;

(iii) ***monotonically decreasing*** if, for every $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) \geq f(x_2)$;

(iv) ***strictly monotonically decreasing*** if, for every $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) > f(x_2)$;

(v) ***constant*** if there exists $\alpha \in \mathbb{R}$ such that $f(x) = \alpha$ for every $x \in I$. •

Let us see how monotonicity can be used to make some implications about the continuity of a function. In Theorem 3.2.26 below we will explore some further properties of monotonic functions.

3.1.28 Theorem (Characterisation of monotonic functions I) *If* $I \subseteq \mathbb{R}$ *is an interval and if* $f\colon I \to \mathbb{R}$ *is either monotonically increasing or monotonically decreasing, then the following statements hold:*

(i) the limits $\lim_{x\downarrow x_0} f(x)$ *and* $\lim_{x\uparrow x_0} f(x)$ *exist whenever they make sense as limits in* I*;*

(ii) the set on which f *is discontinuous is countable.*

Proof We can assume without loss of generality (why?), we assume that $I = [a, b]$ and that f is monotonically increasing.

(i) First let us consider limits from the left. Thus let $x_0 > a$ and consider $\lim_{x\uparrow x_0} f(x)$. For any increasing sequence $(x_j)_{j\in\mathbb{Z}_{>0}} \subseteq [a, x_0)$ converging to x_0 the sequence $(f(x_j))_{j\in\mathbb{Z}_{>0}}$ is bounded and increasing. Therefore it has a limit by Theorem 2.3.8. In a like manner, one shows that right limits also exist.

(ii) Define

$$j(x_0) = \lim_{x\downarrow x_0} f(x) - \lim_{x\uparrow x_0} f(x)$$

as the jump at x_0. This is nonzero if and only if x_0 is a point of discontinuity of f. Let A_f be the set of points of discontinuity of f. Since f is monotonically increasing and defined on a compact interval, it is bounded and we have

$$\sum_{x\in A_f} j(x) \leq f(b) - f(a). \tag{3.1}$$

Now let $n \in \mathbb{Z}_{>0}$ and denote

$$A_n = \left\{x \in [a, b] \;\middle|\; j(x) > \tfrac{1}{n}\right\}.$$

The set A_n must be finite by (3.1). We also have

$$A_f = \bigcup_{n\in\mathbb{Z}_{>0}} A_n,$$

meaning that A_f is a countable union of finite sets. Thus A_f is itself countable. ■

Sometimes the following "local" characterisation of monotonicity is useful.

3.1.29 Proposition (Monotonicity is "local") *A function* $f\colon I \to \mathbb{R}$ *defined on an interval* I *is*

(i) monotonically increasing if and only if, for every $x \in I$*, there exists a neighbourhood* U *of* x *such that* $f|U \cap I$ *is monotonically increasing;*

(ii) strictly monotonically increasing if and only if, for every $x \in I$*, there exists a neighbourhood* U *of* x *such that* $f|U \cap I$ *is strictly monotonically increasing;*

(iii) monotonically decreasing if and only if, for every $x \in I$*, there exists a neighbourhood* U *of* x *such that* $f|U \cap I$ *is monotonically decreasing;*

(iv) strictly monotonically decreasing if and only if, for every $x \in I$*, there exists a neighbourhood* U *of* x *such that* $f|U \cap I$ *is strictly monotonically decreasing.*

Proof We shall only prove the first assertion as the other follow from an identical sort of argument. Also, the "only if" assertion is clear, so we need only prove the "if" assertion.

Let $x_1, x_2 \in I$ with $x_1 < x_2$. By hypothesis, for $x \in [x_1, x_2]$, there exists $\epsilon_x \in \mathbb{R}_{>0}$ such that, if we define $U_x = (x - \epsilon, x + \epsilon)$, then $f|U_x \cap I$ is monotonically increasing. Note that $(U_x)_{x \in [x_1, x_2]}$ covers $[x_1, x_2]$ and so, by the Heine–Borel Theorem, there exists $\xi_1, \ldots, \xi_k \in [x_1, x_2]$ such that $[x_1, x_2] \subseteq \cup_{j=1}^{k} U_{\xi_j}$. We can assume that $\xi_1, \ldots, \xi_k$ are ordered so that $x_1 \in U_{\xi_1}$, that $U_{\xi_{j+1}} \cap U_{\xi_j} \neq \emptyset$, and such that $x_2 \in U_{\xi_k}$. We have that $f|U_{\xi_1} \cap I$ is monotonically increasing. Since $f|U_{\xi_2} \cap I$ is monotonically increasing and since $U_{\xi_1} \cap U_{\xi_2} \neq \emptyset$, we deduce that $f|(U_{\xi_1} \cup U_{\xi_2}) \cap I$ is monotonically increasing. We can continue this process to show that

$$f|(U_{\xi_1} \cup \cdots \cup U_{\xi_k}) \cap I$$

is monotonically increasing, which is the result. ■

In thinking about the graph of a continuous monotonically increasing function, it will not be surprising that there might be a relationship between monotonicity and invertibility. In the next result we explore the precise nature of this relationship.

3.1.30 Theorem (Strict monotonicity and continuity implies invertibility) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{f} \colon \mathrm{I} \to \mathbb{R}$ *be continuous and strictly monotonically increasing (resp. strictly monotonically decreasing). If* $\mathrm{J} = \mathrm{image}(\mathrm{f})$ *then the following statements hold:*

(i) J *is an interval;*

(ii) there exists a continuous, strictly monotonically increasing (resp. strictly monotonically decreasing) inverse $\mathrm{g} \colon \mathrm{J} \to \mathrm{I}$ *for* f.

Proof We suppose f to be strictly monotonically increasing; the case where it is strictly monotonically decreasing is handled similarly (or follows by considering $-f$, which is strictly monotonically increasing if f is strictly monotonically decreasing).

(i) This follows from Theorem 2.5.34 and Proposition 3.1.26, where it is shown that intervals are the only connected sets, and that continuous images of connected sets are connected.

(ii) Since f is strictly monotonically increasing, if $f(x_1) = f(x_2)$, then $x_1 = x_2$. Thus f is injective as a map from I to J. Clearly $f \colon I \to J$ is also surjective, and so is invertible. Let $y_1, y_2 \in J$ and suppose that $y_1 < y_2$. Then $f(g(y_1)) < f(g(y_2))$, implying that $g(y_1) < g(y_2)$. Thus g is strictly monotonically increasing. It remains to show that the inverse g is continuous. Let $y_0 \in J$ and let $\epsilon \in \mathbb{R}_{>0}$. First suppose that $y_0 \in \mathrm{int}(J)$. Let $x_0 = g(y_0)$ and, supposing ϵ sufficiently small, define $y_1 = f(x_0 - \epsilon)$ and $y_2 = f(x_0 + \epsilon)$. Then let $\delta = \min\{y_0 - y_1, y_2 - y_0\}$. If $y \in \mathsf{B}(\delta, y_0)$ then $y \in (y_1, y_2)$, and since g is strictly monotonically increasing

$$x_0 - \epsilon = g(y_1) < g(y) < g(y_2) = x_0 + \epsilon.$$

Thus $g(y) \in \mathsf{B}(\epsilon, y_0)$, giving continuity of g at x_0. An entirely similar argument can be given if y_0 is an endpoint of J. ■

3.1.6 Convex functions and continuity

In this section we see for the first time the important notion of convexity, here in a fairly simple setting.

Let us first define what we mean by a convex function.

3.1.31 Definition (Convex function) For an interval $I \subseteq \mathbb{R}$, a function $f : I \to \mathbb{R}$ is:

(i) ***convex*** if

$$f((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2)$$

for every $x_1, x_2 \in I$ and $s \in [0,1]$;

(ii) ***strictly convex*** if

$$f((1-s)x_1 + sx_2) < (1-s)f(x_1) + sf(x_2)$$

for every distinct $x_1, x_2 \in I$ and for every $s \in (0,1)$;

(iii) ***concave*** if $-f$ is convex;

(iv) ***strictly concave*** if $-f$ is strictly convex. •

Let us give some examples of convex functions.

3.1.32 Examples (Convex functions)

1. A constant function $x \mapsto c$, defined on any interval, is both convex and concave in a trivial way. It is neither strictly convex nor strictly concave.
2. A linear function $x \mapsto ax+b$, defined on any interval, is both convex and concave. It is neither strictly convex nor strictly concave.
3. The function $x \mapsto x^2$, defined on any interval, is strictly convex. Let us verify this. For $s \in (0,1)$ and for $x, y \in \mathbb{R}$ we have, using the triangle inequality,

$$((1-s)x + sy)^2 \leq |(1-s)x + sy|^2 < (1-s)^2x^2 + s^2y^2 \leq (1-s)x^2 + sy^2.$$

4. We refer to Section 3.6.1 for the definition of exponential function $\exp : \mathbb{R} \to \mathbb{R}$. We claim that exp is strictly convex. This can be verified explicitly with some effort. However, it follows easily from the fact, proved as Proposition 3.2.30 below, that a function like exp that is twice continuously differentiable with a positive second-derivative is strictly convex. (Note that $\exp'' = \exp$.)
5. We claim that the function log defined in Section 3.6.2 is strictly concave as a function on $\mathbb{R}_{>0}$. Here we compute $\log''(x) = -\frac{1}{x^2}$, which gives strict convexity of $-\log$ (and hence strict concavity of log) by Proposition 3.2.30 below.
6. For $x_0 \in \mathbb{R}$, the function $n_{x_0} : \mathbb{R} \to \mathbb{R}$ defined by $n_{x_0} = |x - x_0|$ is convex. Indeed, if $x_1, x_2 \in \mathbb{R}$ and $s \in [0,1]$ then

$$\begin{aligned} n_{x_0}((1-s)x_1 + sx_2) &= |(1-s)x_1 + sx_2 - x_0| = |(1-s)(x_1 - x_0) + s(x_2 - x_0)| \\ &\leq (1-s)|x_1 - x_0| + s|x_2 - x_0| = (1-s)n_{x_0}(x_1) + sn_{x_0}(x_2), \end{aligned}$$

using the triangle inequality. •

Let us give an alternative and insightful characterisation of convex functions. For an interval $I \subseteq \mathbb{R}$ define

$$E_I = \{(x, y) \in I^2 \mid s < t\}$$

and, for $a, b \in I$, denote

$$L_b = \{a \in I \mid (a, b) \in E_I\}, \quad R_a = \{b \in I \mid (a, b) \in E_I\}.$$

Now, for $f : I \to \mathbb{R}$ define $s_f : E_I \to \mathbb{R}$ by

$$s_f(a, b) = \frac{f(b) - f(a)}{b - a}.$$

With this notation at hand, we have the following result.

3.1.33 Lemma (Alternative characterisation of convexity) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$*, a function* $\mathrm{f} : \mathrm{I} \to \mathbb{R}$ *is (strictly) convex if and only if, for every* $\mathrm{a}, \mathrm{b} \in \mathrm{I}$*, the functions*

$$\mathrm{L_b} \ni \mathrm{a} \mapsto \mathrm{s_f(a, b)} \in \mathbb{R}, \quad \mathrm{R_a} \ni \mathrm{b} \mapsto \mathrm{s_f(a, b)} \in \mathbb{R} \tag{3.2}$$

are (strictly) monotonically increasing.

Proof First suppose that f is convex. Let $a, b, c \in I$ satisfy $a < b < c$. Define $s \in (0, 1)$ by $s = \frac{b-a}{c-a}$ and note that the definition of convexity using this value of s gives

$$f(b) \leq \frac{c - b}{c - a} f(a) + \frac{b - a}{c - a} f(c).$$

Simple rearrangement gives

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

and so we have

$$\frac{f(b) - f(a)}{b - a} \leq \frac{f(c) - f(a)}{c - a}, \quad \frac{f(c) - f(a)}{c - a} \leq \frac{f(c) - f(b)}{c - b}.$$

In other words, $s_f(a, b) \leq s_f(a, c)$ and $s_f(a, c) \leq s_f(b, c)$. Since this holds for every $a, b, c \in I$ with $a < b < c$, we conclude that the functions (3.2) are monotonically increasing, as stated. If f is strictly convex, then the inequalities in the above computation are strict, and one concludes that the functions (3.2) are strictly monotonically increasing.

Next suppose that the functions (3.2) are monotonically increasing and let $a, c \in I$ with $a < c$ and let $s \in (0, 1)$. Define $b = (1 - s)a + sc$. A rearrangement of the inequality $s_f(a, b) \leq s_f(a, c)$ gives

$$\begin{aligned} & f(b) \leq \frac{c - b}{c - a} f(a) + \frac{b - a}{c - a} f(c) \\ \implies \quad & f((1 - s)a + sc) \leq (1 - s) f(a) + s f(c), \end{aligned}$$

showing that f is convex since $a, c \in I$ with $a < c$ and $s \in (0, 1)$ are arbitrary in the above computation. If the functions (3.2) are strictly monotonically increasing, then the inequalities in the preceding computations are strict, and so one deduces that f is strictly convex. ■

In Figure 3.4 we depict what the lemma is telling us about convex functions. The idea is that the slope of the line connecting the points $(a, f(a))$ and $(b, f(b))$ in the plane is nondecreasing in a and b.

The following inequality for convex functions is very often useful.

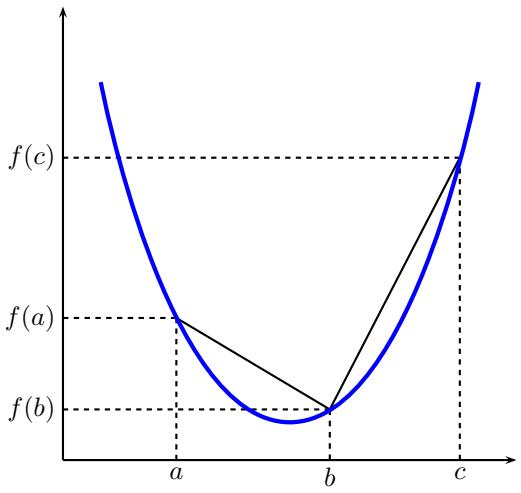


Figure 3.4 A characterisation of a convex function

3.1.34 Theorem (Jensen's inequality) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$*, for a convex function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$*, for* $\mathrm{x}_1, \dots, \mathrm{x}_k \in \mathrm{I}$*, and for* $\lambda_1, \dots, \lambda_k \in \mathbb{R}_{\geq 0}$*, we have*

$$\mathrm{f}\Big(\frac{\lambda_1}{\sum_{\mathrm{j}=1} \lambda_{\mathrm{j}}}\mathrm{x}_1 + \cdots + \frac{\lambda_k}{\sum_{\mathrm{j}=1}^{\mathrm{k}} \lambda_{\mathrm{j}}}\mathrm{x}_k\Big) \leq \frac{\lambda_1}{\sum_{\mathrm{j}=1}^{\mathrm{k}} \lambda_{\mathrm{j}}}\mathrm{f}(\mathrm{x}_1) + \cdots + \frac{\lambda_k}{\sum_{\mathrm{j}=1}^{\mathrm{k}} \lambda_{\mathrm{j}}}\mathrm{f}(\mathrm{x}_k).$$

Moreover, if f *is strictly convex and if* $\lambda_1, \dots, \lambda_k \in \mathbb{R}_{>0}$*, than we have equality in the preceding expression if and only if* $\mathrm{x}_1 = \cdots = \mathrm{x}_k$.

Proof We first comment that, with $\lambda_1, \dots, \lambda_k$ and $x_1, \dots, x_k$ as stated,

$$\frac{\lambda_1}{\sum_{j=1}^k \lambda_j}x_1 + \cdots + \frac{\lambda_k}{\sum_{j=1}^k \lambda_j}x_k \in I.$$

This is because intervals are convex, something that will become clear in Section **??**.

It is clear that we can without loss of generality, by replacing λ_j with

$$\lambda'_m = \frac{\lambda_m}{\sum_{j=1}^k \lambda_j}, \qquad m \in \{1, \dots, k\},$$

if necessary, that we can assume that $\sum_{j=1}^k \lambda_j = 1$.

We first note that if $x_1 = \cdots = x_k$ then the inequality in the statement of the theorem is an equality, no matter what the character of f.

The proof is by induction on k, the result being obvious when $k = 1$. So suppose the result is true when $k = m$ and let $x_1, \dots, x_{m+1} \in I$ and let $\lambda_1, \dots, \lambda_{m+1} \in \mathbb{R}_{\geq 0}$ satisfy $\sum_{j=1}^{m+1} \lambda_j = 1$. Without loss of generality (by reindexing if necessary), suppose that $\lambda_{m+1} \in [0, 1)$. Note that

$$\frac{\lambda_1}{1 - \lambda_{m+1}} + \cdots + \frac{\lambda_m}{1 - \lambda_{m+1}} = 1$$

so that, by the induction hypothesis,

$$f\Big(\frac{\lambda_1}{1 - \lambda_{m+1}}x_1 + \cdots + \frac{\lambda_m}{1 - \lambda_{m+1}}x_m\Big) \leq \frac{\lambda_1}{1 - \lambda_{m+1}}f(x_1) + \cdots + \frac{\lambda_m}{1 - \lambda_{m+1}}f(x_m).$$

Now, by convexity of f,

$$f\Big((1-\lambda_{m+1})\Big(\frac{\lambda_1}{1-\lambda_{m+1}}x_1+\cdots+\frac{\lambda_m}{1-\lambda_{m+1}}x_m\Big)+\lambda_{m+1}x_{m+1}\Big)$$
$$\leq (1-\lambda_{m+1})f\Big(\frac{\lambda_1}{1-\lambda_{m+1}}x_1+\cdots+\frac{\lambda_m}{1-\lambda_{m+1}}x_m\Big)+\lambda_{m+1}f(x_{m+1}).$$

The desired inequality follows by combining the previous two equations.

To prove the final assertion of the theorem, suppose that f is strictly convex, that $\lambda_1,\dots,\lambda_k\in\mathbb{R}_{>0}$ satisfy $\sum_{j=1}^k\lambda_j=1$, and that the inequality in the theorem is equality. We prove by induction that $x_1=\cdots=x_k$. For $k=1$ the assertion is obvious. Let us prove the assertion for $k=2$. Thus suppose that

$$f((1-\lambda)x_1+\lambda x_2)=(1-\lambda)f(x_1)+\lambda f(x_2)$$

for $x_1,x_2\in I$ and for $\lambda\in(0,1)$. If $x_1\neq x_2$ then we have, by definition of strict convexity,

$$f((1-\lambda)x_1+\lambda x_2)<(1-\lambda)f(x_1)+\lambda f(x_2),$$

contradicting our hypotheses. Thus we must have $x_1=x_2$. Now suppose the assertion is true for $k=m$ and let $x_1,\dots,x_{m+1}\in I$, let $\lambda_1,\dots,\lambda_{m+1}\in\mathbb{R}_{>0}$ satisfy $\sum_{j=1}^{m+1}\lambda_j=1$, and suppose that

$$f(\lambda_1x_1+\cdots+\lambda_{m+1}x_{m+1})=\lambda_1f(x_1)+\cdots+\lambda_{m+1}f(x_{m+1}).$$

Since none of $\lambda_1,\dots,\lambda_{m+1}$ are zero we must have $\lambda_{m+1}\in(0,1)$. Now note that

$$f(\lambda_1x_1+\cdots+\lambda_{m+1}x_{m+1})=f\Big((1-\lambda_{m+1})\Big(\frac{\lambda_1}{1-\lambda_{m+1}}x_1+\cdots+\frac{\lambda_m}{1-\lambda_{m+1}}x_m\Big)+\lambda_{m+1}x_{m+1}\Big)\quad(3.3)$$

and that

$$\lambda_1f(x_1)+\cdots+\lambda_{m+1}f(x_{m+1})$$
$$=(1-\lambda_{m+1})f\Big(\frac{\lambda_1}{1-\lambda_{m+1}}x_1+\cdots+\frac{\lambda_m}{1-\lambda_{m+1}}x_m\Big)+\lambda_{m+1}f(x_{m+1}).$$

Therefore, by assumption,

$$f\Big((1-\lambda_{m+1})\Big(\frac{\lambda_1}{1-\lambda_{m+1}}x_1+\cdots+\frac{\lambda_m}{1-\lambda_{m+1}}x_m\Big)+\lambda_{m+1}x_{m+1}\Big)$$
$$=(1-\lambda_{m+1})f\Big(\frac{\lambda_1}{1-\lambda_{m+1}}x_1+\cdots+\frac{\lambda_m}{1-\lambda_{m+1}}x_m\Big)+\lambda_{m+1}f(x_{m+1}).\quad(3.4)$$

Since the assertion we are proving holds for $k=2$ this implies that

$$x_{m+1}=\frac{\lambda_1}{1-\lambda_{m+1}}x_1+\cdots+\frac{\lambda_m}{1-\lambda_{m+1}}x_m.\quad(3.5)$$

Now suppose that the numbers $x_1,\dots,x_m$ are not all equal. Then, by the induction hypothesis,

$$f\Big(\frac{\lambda_1}{1-\lambda_{m+1}}x_1+\cdots+\frac{\lambda_m}{1-\lambda_{m+1}}x_m\Big)<\frac{\lambda_1}{1-\lambda_{m+1}}f(x_1)+\cdots+\frac{\lambda_m}{1-\lambda_{m+1}}f(x_m)$$

since

$$\frac{\lambda_1}{1-\lambda_{m+1}} + \cdots + \frac{\lambda_m}{1-\lambda_{m+1}} = 1.$$

Therefore, combining (3.3) and (3.4)

$$f(\lambda_1 x_1 + \cdots + \lambda_{m+1} x_{m+1}) < \lambda_1 f(x_1) + \cdots + \lambda_{m+1} f(x_{m+1}),$$

contradicting our hypotheses. Thus we must have $x_1 = \cdots = x_m$. From (3.5) we then conclude that $x_1 = \cdots = x_{m+1}$, as desired. ∎

An interesting application of Jensen's inequality is the derivation of the so-called arithmetic/geometric mean inequalities. If $x_1, \ldots, x_k \in \mathbb{R}_{>0}$, their ***arithmetic mean*** is

$$\frac{1}{k}(x_1 + \cdots + x_k)$$

and their ***geometric mean*** is

$$(x_1 \cdots x_k)^{1/k}.$$

We first state a result which relates generalisations of the arithmetic and geometric means.

3.1.35 Corollary (Weighted arithmetic/geometric mean inequality) *Let* $x_1, \ldots, x_k \in \mathbb{R}_{\geq 0}$ *and suppose that* $\lambda_1, \ldots, \lambda_k \in \mathbb{R}_{>0}$ *satisfy* $\sum_{j=1}^{k} \lambda_j = 1$. *Then*

$$x_1^{\lambda_1} \cdots x_k^{\lambda_k} \leq \lambda_1 x_1 + \cdots + \lambda_k x_k,$$

and equality holds if and only if $x_1 = \cdots = x_k$.

Proof Since the inequality obviously holds if any of $x_1, \ldots, x_k$ are zero, let us suppose that these numbers are all positive. By Example 3.1.32–5, $-\log$ is convex. Thus Jensen's inequality gives

$$-\log(\lambda_1 x_1 + \cdots + \lambda_k x_k) \leq -\lambda_1 \log(x_1) - \cdots - \lambda_k \log(x_k) = -\log(x_1^{\lambda_1} \cdots x_k^{\lambda_k}).$$

Since $-\log$ is strictly monotonically decreasing by Proposition 3.6.6(ii), the result follows. Moreover, since $-\log$ is strictly convex by Proposition 3.2.30, the final assertion of the corollary follows from the final assertion of Theorem 3.1.34. ∎

The corollary gives the following inequality as a special case.

3.1.36 Corollary (Arithmetic/geometric mean inequality) *If* $x_1, \ldots, x_k \in \mathbb{R}_{\geq 0}$ *then*

$$(x_1 \cdots x_k)^{1/k} \leq \frac{x_1 + \cdots + x_k}{k},$$

and equality holds if and only if $x_1 = \cdots = x_k$.

Let us give some properties of convex functions. Further properties of convex function are give in Proposition 3.2.29

3.1.37 Proposition (Properties of convex functions I) *For an interval* $I \subseteq \mathbb{R}$ *and for a convex function* $f\colon I \to \mathbb{R}$*, the following statements hold:*

(i) if I *is open, then* f *is continuous;*

(ii) for any compact interval $K \subseteq \operatorname{int}(I)$*, there exists* $L \in \mathbb{R}_{>0}$ *such that*

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|, \qquad x_1, x_2 \in K.$$

Proof (ii) Let $K = [a, b] \subseteq \operatorname{int}(I)$ and let $a', b' \in I$ satisfy $a' < a$ and $b' > b$, this being possible since $K \subseteq \operatorname{int}(I)$. Now let $x_1, x_2 \in K$ and note that, by Lemma 3.1.33,

$$s_f(a', a) \leq s_f(x_1, x_2) \leq s_f(b, b')$$

since $a' < x_1$, $a \leq x_2$, $x_1 \leq b$, and $x_2 < b'$. Thus, taking $L = \max\{s_f(a', a), s_f(b, b')\}$, we have

$$-L \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq L,$$

which gives the result.

(i) This follows from part (ii) easily. Indeed let $x \in I$ and let K be a compact subinterval of I such that $x \in \operatorname{int}(K)$, this being possible since I is open. If $\epsilon \in \mathbb{R}_{>0}$, let $\delta = \frac{\epsilon}{L}$. It then immediately follows that if $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$. ■

Let us give some an example that illustrates that openness is necessary in the first part of the preceding result.

3.1.38 Example (A convex discontinuous function) Let $I = [0, 1]$ and define $f\colon [0, 1] \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1, & x = 1, \\ 0, & x \in [0, 1). \end{cases}$$

If $x_1, x_2 \in [0, 1)$ and if $s \in [0, 1]$ then

$$0 = f((1 - s)x_1 + sx_2) = (1 - s)f(x_1) + sf(x_2).$$

If $x_1 \in [0, 1)$, if $x_2 = 1$, and if $s \in (0, 1)$ then

$$0 = f((1 - s)x_1 + sx_2) \leq (1 - s)f(x_1) + sf(x_2) = s,$$

showing that f is convex as desired. Note that f is not continuous, but that its discontinuity is on the boundary, as must be the case since convex functions on open sets are continuous. •

Let us also present some operations that preserve convexity.

3.1.39 Proposition (Convexity and operations on functions) *For an interval* $I \subseteq \mathbb{R}$ *and for convex functions* $f, g\colon I \to \mathbb{R}$*, the following statements hold:*

(i) the function $I \ni x \mapsto \max\{f(x), g(x)\}$ *is convex;*

(ii) the function af *is convex if* $a \in \mathbb{R}_{\geq 0}$*;*

(iii) the function $f + g$ *is convex;*

(iv) *if* $\mathrm{J} \subseteq \mathbb{R}$ *is an interval, if* f *takes values in* J*, and if* $\phi\colon \mathrm{J} \to \mathbb{R}$ *is convex and monotonically increasing, then* $\phi \circ \mathrm{f}$ *is convex;*

(v) *if* $\mathrm{x}_0 \in \mathrm{I}$ *is a local minimum for* f *(see Definition 3.2.15). then* x_0 *is a minimum for* f.

Proof (i) Let $x_1, x_2 \in I$ and let $s \in [0,1]$. Then, by directly applying the definition of convexity to f and g, we have

$$\max\{f((1-s)x_1 + sx_2), g((1-s)x_1 + sx_2)\} \\ \leq (1-s)\max\{f(x_1), g(x_1)\} + s\max\{f(x_2), g(x_2)\}.$$

(ii) This follows immediately from the definition of convexity.

(iii) For $x_1, x_2 \in I$ and for $s \in [0,1]$ we have

$$f((1-s)x_1 + sx_2) + g((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2) + (1-s)g(x_1) + sg(x_2) \\ = (1-s)(f(x_1) + g(x_1)) + s(f(x_2 + g(x_2)),$$

by applying the definition of convexity to f and g.

(iv) For $x_1, x_2 \in I$ and for $s \in [0,1]$, convexity of f gives

$$f((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2)$$

and so monotonicity of ϕ gives

$$\phi \circ f((1-s)x_1 + sx_2) \leq \phi((1-s)f(x_1) + sf(x_2)).$$

Now convexity of ϕ gives

$$\phi \circ f((1-s)x_1 + sx_2) \leq (1-s)\phi \circ f(x_1) + s\phi \circ f(x_2),$$

as desired.

(v) Suppose that x_0 is a local minimum for f, i.e., there is a neighbourhood $U \subseteq I$ of x_0 such that $f(x) \geq f(x_0)$ for all $x \in U$. Now let $x \in I$ and note that

$$s \mapsto (1-s)x_0 + sx$$

is continuous and $\lim_{s\to 0}(1-s)x_0 + sx = x_0$. Therefore, there exists $s_0 \in (0,1]$ such that $(1-s)x_0 + sx \in U$ for all $s \in (0, s_0)$. Thus

$$f(x_0) \leq f((1-s)x_0 + sx) \leq (1-s)f(x_0) + sf(x)$$

for $s \in (0, s_0)$. Simplification gives $f(x_0) \leq f(x)$ and so x_0 is a minimum for f. ∎

3.1.7 Piecewise continuous functions

It is often of interest to consider functions that are not continuous, but which possess only jump discontinuities, and only "few" of these. In order to do so, it is convenient to introduce some notation. For and interval $I \subseteq \mathbb{R}$, a function $f\colon I \to \mathbb{R}$, and $x \in I$ define

$$f(x-) = \lim_{\epsilon \downarrow 0} f(x-\epsilon), \quad f(x+) = \lim_{\epsilon \downarrow 0} f(x+\epsilon),$$

allowing that these limits may not be defined (or even make sense if $x \in \mathrm{bd}(I)$).

We then have the following definition, recalling our notation concerning partitions of intervals given in and around Definition 2.5.7.

3.1.40 Definition (Piecewise continuous function) A function $f\colon [a,b] \to \mathbb{R}$ is ***piecewise continuous*** if there exists a partition $P = (I_1, \dots, I_k)$, with $\mathrm{EP}(P) = (x_0, x_1, \dots, x_k)$, of $[a,b]$ with the following properties:

(i) $f|\operatorname{int}(I_j)$ is continuous for each $j \in \{1, \dots, k\}$;

(ii) for $j \in \{1, \dots, k-1\}$, the limits $f(x_j+)$ and $f(x_j-)$ exist;

(iii) the limits $f(a+)$ and $f(b-)$ exist. •

Let us give a couple of examples to illustrate some of the things that can happen with piecewise continuous functions.

3.1.41 Examples (Piecewise continuous functions)

1. Let $I = [-1,1]$ and define $f_1, f_2, f_3\colon I \to \mathbb{R}$ by

$$f_1(x) = \operatorname{sign}(x),$$
$$f_2(x) = \begin{cases} \operatorname{sign}(x), & x \neq 0, \\ 1, & x = 0, \end{cases}$$
$$f_2(x) = \begin{cases} \operatorname{sign}(x), & x \neq 0, \\ -1, & x = 0. \end{cases}$$

   One readily verifies that all of these functions are piecewise continuous with a single discontinuity at $x = 0$. Note that the functions do not have the same value at the discontinuity. Indeed, the definition of piecewise continuity is unconcerned with the value of the function at discontinuities.

2. Let $I = [-1,1]$ and define $f\colon I \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

   This function is, by definition, piecewise continuous with a single discontinuity at $x = 0$. This shows that the definition of piecewise continuity includes functions, not just with jump discontinuities, but with removable discontinuities. •

Exercises

3.1.1

Oftentimes, a continuity novice will think that the definition of continuity at x_0 of a function $f\colon I \to \mathbb{R}$ is as follows: for every $\epsilon \in \mathbb{R}_{>0}$ there exists $\delta \in \mathbb{R}_{>0}$ such that if $|f(x) - f(x_0)| < \epsilon$ then $|x - x_0| < \delta$. Motivated by this, let us call a function ***fresh-from-high-school continuous*** if it has the preceding property at each point $x \in I$.

3.1.2 Answer the following two questions.

(a) Find an interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ such that f is continuous but not fresh-from-high-school continuous.

(b) Find an interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ such that f is fresh-from-high-school continuous but not continuous.

3.1.3 Let $I \subseteq \mathbb{R}$ be an interval and let $f, g\colon I \to \mathbb{R}$ be functions.

(a) Show that $D_{fg} \subseteq D_f \cup D_g$.

(b) Show that it is not generally true that $D_f \cap D_g \subseteq D_{fg}$.

(c) Suppose that f is bounded. Show that if $x \in (D_f \cap (I \setminus D_g)) \cap (I \setminus D_{fg})$, then $g(x) = 0$.***missing stuff***

3.1.4 Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function. For $x \in I$ and $\delta \in \mathbb{R}_{>0}$ define

$$\omega_f(x, \delta) = \sup\{|f(x_1), f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\}.$$

Show that, if $y \in \mathsf{B}(\delta, x)$, then $\omega_f(y, \frac{\delta}{2}) \leq \omega_f(x, \delta)$.

3.1.5 Recall from Theorem 3.1.24 that a continuous function defined on a compact interval is uniformly continuous. Show that this assertion is generally false if the interval is not compact.

3.1.6 Give an example of an interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ that is locally bounded but not bounded.

3.1.7 Answer the following three questions.

(a) Find a bounded interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ such that f is continuous but not bounded.

(b) Find a compact interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ such that f is bounded but not continuous.

(c) Find a closed but unbounded interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ such that f is continuous but not bounded.

3.1.8 Answer the following two questions.

(a) For $I = [0, 1)$ find a bounded, continuous function $f\colon I \to \mathbb{R}$ that does not attain its maximum on I.

(b) For $I = [0, \infty)$ find a bounded, continuous function $f\colon I \to \mathbb{R}$ that does not attain its maximum on I.

3.1.9 Explore your understanding of Theorem 3.1.3 and its Corollary 3.1.4 by doing the following.

(a) For the continuous function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$, verify Theorem 3.1.3 by (1) determining $f^{-1}(I)$ for a general open interval I and (2) showing that this is sufficient to ensure continuity.
Hint: *For the last part, consider using Proposition 2.5.6 and part (iv) of Proposition 1.3.5.*

(b) For the discontinuous function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \operatorname{sign}(x)$, verify Theorem 3.1.3 by (1) finding an open subset $U \subseteq \mathbb{R}$ for which $f^{-1}(U)$ is not open and (2) finding a sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ converging to $x_0 \in \mathbb{R}$ for which $(f(x_j))_{j \in \mathbb{Z}_{>0}}$ does not converge to $f(x_0)$.

3.1.10 Find a continuous function $f\colon I \to \mathbb{R}$ defined on some interval I and a sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ such that the sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ does not converge but the sequence $(f(x_j))_{j \in \mathbb{Z}_{>0}}$ does converge.

3.1.11 Let $I \subseteq \mathbb{R}$ be an interval and let $f, g\colon I \to \mathbb{R}$ be convex.

(a) Is it true that $x \mapsto \min\{f(x), g(x)\}$ is convex?

(b) Is it true that $f - g$ is convex?

3.1.12 Let $U \subseteq \mathbb{R}$ be open and suppose that $f\colon U \to \mathbb{R}$ is continuous and has the property that

$$\{x \in U \mid \ f(x) \neq 0\}$$

has measure zero. Show that $f(x) = 0$ for *all* $x \in U$.

Section 3.2

Differentiable $\mathbb{R}$-valued functions on $\mathbb{R}$

In this section we deal systematically with another topic with which most readers are at least somewhat familiar: differentiation. However, as with everything we do, we do this here is a manner that is likely more thorough and systematic than that seen by some readers. We do suppose that the reader has had that sort of course where one learns the derivatives of the standard functions, and learns to apply some of the standard rules of differentiation, such as we give in Section 3.2.3.

Do I need to read this section? If you are familiar with, or perhaps even if you only think you are familiar with, the meaning of "continuously differentiable," then you can probably forgo the details of this section. However, if you have not had the benefit of a rigorous calculus course, then the material here might at least be interesting. •

3.2.1 Definition of the derivative

The definition we give of the derivative is as usual, with the exception that, as we did when we talked about continuity, we allow functions to be defined on general intervals. In order to do this, we recall from Section 2.3.7 the notation $\lim_{x\to_I x_0} f(x)$.

3.2.1 Definition (Derivative and differentiable function) Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be a function.

(i) The function f is ***differentiable at*** $\mathbf{x_0 \in I}$ if the limit

$$\lim_{x\to_I x_0} \frac{f(x) - f(x_0)}{x - x_0} \tag{3.6}$$

exists.

(ii) If the limit (3.6) exists, then it is denoted by $f'(x_0)$ and called the ***derivative*** of f at x_0.

(iii) If f is differentiable at each point $x \in I$, then f is ***differentiable***.

(iv) If f is differentiable and if the function $x \mapsto f'(x)$ is continuous, then f is ***continuously differentiable***, or of ***class*** $\mathbf{C^1}$. •

3.2.2 Notation (Alternative notation for derivative) In applications where $\mathbb{R}$-valued functions are clearly to be thought of as functions of "time," we shall sometimes write $\dot{f}$ rather than f' for the derivative.

Sometimes it is convenient to write the derivative using the convention $f'(x) = \frac{df}{dx}$. This notation for derivative suffers from the same problems as the notation "$f(x)$" to denote a function as discussed in Notation 1.3.2. That is to say, one

cannot really use $\frac{\mathrm{d}f}{\mathrm{d}x}$ as a substitute for f', but only for $f'(x)$. Sometimes one can kludge one's way around this with something like $\frac{\mathrm{d}f}{\mathrm{d}x}\big|_{x=x_0}$ to specify the derivative at x_0. But this still leaves unresolved the matter of what is the rôle of "x" in the expression $\frac{\mathrm{d}f}{\mathrm{d}x}\big|_{x=x_0}$. For this reason, we will generally (but not exclusively) stick to f', or sometimes $\dot{f}$. For notation for the derivative for multivariable functions, we refer to Definition 4.4.2. •

Let us consider some examples that illustrate the definition.

3.2.3 Examples (Derivative)

1. Take $I = \mathbb{R}$ and define $f : I \to \mathbb{R}$ by $f(x) = x^k$ for $k \in \mathbb{Z}_{>0}$. We claim that f is continuously differentiable, and that $f'(x) = kx^{k-1}$. To prove this we first note that
$$(x - x_0)(x^{k-1} + x^{k-1}x_0 + \cdots + xx_0^{k-2} + x_0^{k-1}) = x^k - x_0^k,$$
as can be directly verified. Then we compute
$$\begin{aligned}\lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} &= \lim_{x\to x_0} \frac{x^k - x_0^k}{x - x_0} \\ &= \lim_{x\to x_0} (x^{k-1} + x^{k-1}x_0 + \cdots + xx_0^{k-2} + x_0^{k-1}) = kx_0^{k-1},\end{aligned}$$
as desired. Since f' is obviously continuous, we obtain that f is continuously differentiable, as desired.
2. Let $I = [0, 1]$ and define $f : I \to \mathbb{R}$ by
$$f(x) = \begin{cases} x, & x \neq 0, \\ 1, & x = 0. \end{cases}$$
From Example 1 we know that f is continuously differentiable at points in $(0, 1]$. We claim that f is not differentiable at $x = 0$. This will follow from Proposition 3.2.7 below, but let us show this here directly. We have
$$\lim_{x\to_I 0} \frac{f(x) - f(0)}{x - 0} = \lim_{x\downarrow 0} \frac{x - 1}{x} = -\infty.$$
Thus the limit does not exist, and so f is not differentiable at $x = 0$, albeit in a fairly stupid way.
3. Let $I = [0, 1]$ and define $f : I \to \mathbb{R}$ by $f(x) = \sqrt{x(1 - x)}$. We claim that f is differentiable at points in $(0, 1)$, but is not differentiable at $x = 0$ or $x = 1$. Providing that one believes that the function $x \mapsto \sqrt{x}$ is differentiable on $\mathbb{R}_{>0}$ (see Section 3.6*missing stuff*), then the continuous differentiability of f on $(0, 1)$ follows from the results of Section 3.2.3. Moreover, the derivative of f at $x \in (0, 1)$ can be explicitly computed as
$$f'(x) = \frac{1 - 2x}{2\sqrt{x(1 - x)}}.$$

To show that f is not differentiable at $x = 0$ we compute

$$\lim_{x \to_I 0} \frac{f(x) - f(0)}{x - 0} = \lim_{x \downarrow 0} \frac{\sqrt{1 - x}}{\sqrt{x}} = \infty.$$

Similarly, at $x = 1$ we compute

$$\lim_{x \to_I 1} \frac{f(x) - f(1)}{x - 1} = \lim_{x \uparrow 1} \frac{-\sqrt{x}}{\sqrt{x - 1}} = -\infty.$$

Since neither of these limits are elements of $\mathbb{R}$, it follows that f is not differentiable at $x = 0$ or $x = 1$.

4. Let $I = \mathbb{R}$ and define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

We first claim that f is differentiable. The differentiability of f at points $x \in \mathbb{R} \setminus \{0\}$ will follow from our results in Section 3.2.3 concerning differentiability, and algebraic operations along with composition. Indeed, using these rules for differentiation we compute that for $x \neq 0$ we have

$$f'(x) = 2x \sin \tfrac{1}{x} - \cos \tfrac{1}{x}.$$

Next let us prove that f is differentiable at $x = 0$ and that $f'(0) = 0$. We have

$$\lim_{x \to 0} \frac{f(x) - f(x)}{x - 0} = \lim_{x \to 0} x \sin \tfrac{1}{x}.$$

Now let $\epsilon \in \mathbb{R}_{>0}$. Then, for $\delta = \epsilon$ we have

$$\left| x \sin \tfrac{1}{x} - 0 \right| < \epsilon$$

since $\left| \sin \frac{1}{x} \right| \leq 1$. This shows that $f'(0) = 0$, as claimed. This shows that f is differentiable.

However, we claim that f is not *continuously* differentiable. Clearly there are no problems away from $x = 0$, again by the results of Section 3.2.3. But we note that f' is discontinuous at $x = 0$. Indeed, we note that f is the sum of two functions, one ($x \sin \frac{1}{x}$) of which goes to zero as x goes to zero, and the other ($-\cos \frac{1}{x}$) of which, when evaluated in any neighbourhood of $x = 0$, takes all possible values in the interval $[-1, 1]$. This means that in any sufficiently small neighbourhood of $x = 0$, the function f' will take all possible values in the interval $[-\frac{1}{2}, \frac{1}{2}]$. This precludes the limit $\lim_{x \to 0} f'(x)$ from existing, and so precludes f' from being continuous at $x = 0$ by Theorem 3.1.3. •

Let us give some intuition about the derivative. Given an interval I and functions $f, g : I \to \mathbb{R}$, we say that f and g are ***tangent*** at $x_0 \in \mathbb{R}$ if

$$\lim_{x \to_I x_0} \frac{f(x) - g(x)}{x - x_0} = 0.$$

In Figure 3.5 we depict the idea of two functions being tangent. Using this idea, we can give the following interpretation of the derivative.

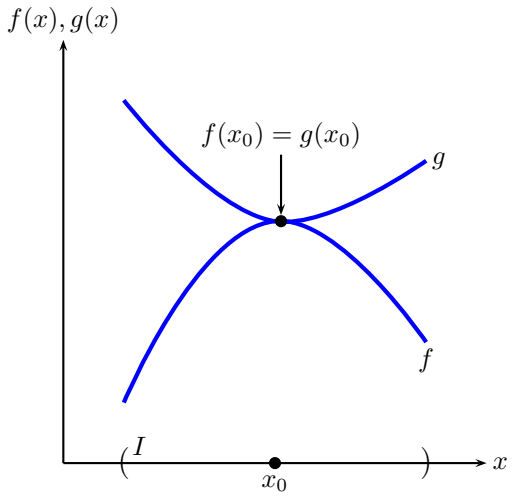


Figure 3.5 Functions that are tangent

3.2.4 Proposition (Derivative and linear approximation) *Let* $\mathrm{I} \subseteq \mathbb{R}$, *let* $\mathrm{x}_0 \in \mathrm{I}$, *and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a function. Then there exists at most one number* $\alpha \in \mathbb{R}$ *such that* f *is tangent at* x_0 *with the function* $\mathrm{x} \mapsto \mathrm{f}(\mathrm{x}_0) + \alpha(\mathrm{x} - \mathrm{x}_0)$. *Moreover, such a number* α *exists if and only if* f *is differentiable at* x_0, *in which case* $\alpha = \mathrm{f}'(\mathrm{x}_0)$.

Proof Suppose there are two such numbers α_1 and α_2. Thus

$$\lim_{x \to_I x_0} \frac{f(x) - (f(x_0) + \alpha_j(x - x_0))}{x - x_0} = 0, \qquad j \in \{1, 2\}, \tag{3.7}$$

We compute

$$\begin{aligned} |\alpha_1 - \alpha_2| &= \frac{|\alpha_1(x - x_0) - \alpha_2(x - x_0)|}{|x - x_0|} \\ &= \frac{|-f(x) + f(x_0) + \alpha_1(x - x_0) + f(x) - f(x_0) - \alpha_2(x - x_0)|}{|x - x_0|} \\ &\leq \frac{|f(x) - f(x_0) - \alpha_1(x - x_0)|}{|x - x_0|} + \frac{|f(x) - f(x_0) - \alpha_2(x - x_0)|}{|x - x_0|}. \end{aligned}$$

Since α_1 and α_2 satisfy (3.7), as we let $x \to x_0$ the right-hand side goes to zero showing that $|\alpha_1 - \alpha_2| = 0$. This proves the first part of the result.

Next suppose that there exists $\alpha \in \mathbb{R}$ such that

$$\lim_{x \to_I x_0} \frac{f(x) - (f(x_0) + \alpha(x - x_0))}{x - x_0} = 0.$$

It then immediately follows that

$$\lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0} = \alpha.$$

Thus f is differentiable at x_0 with derivative equal to α. Conversely, if f is differentiable

at x_0 then we have

$$f'(x_0) = \lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0},$$
$$\implies \lim_{x \to_I x_0} \frac{f(x) - f(x_0) - f'(x_0)(x - x_0)}{x - x_0} = 0,$$

which completes the proof. ■

The idea, then, is that the derivative serves, as we are taught in first-year calculus, as the best linear approximation to the function, since the function $x \mapsto f(x_0) + \alpha(x - x_0)$ is a linear function with slope α passing through $f(x_0)$.

We may also define derivatives of higher-order. Suppose that $f : I \to \mathbb{R}$ is differentiable, so that the function $f' : I \to \mathbb{R}$ can be defined. If the limit

$$\lim_{x \to_I x_0} \frac{f'(x) - f'(x_0)}{x - x_0}$$

exists, then we say that f is ***twice differentiable at*** $\mathbf{x_0}$. We denote the limit by $f''(x_0)$, and call it the ***second derivative*** of f at x_0. If f is differentiable at each point $x \in I$ then f is ***twice differentiable***. If additionally the map $x \mapsto f''(x)$ is continuous, then f is ***twice continuously differentiable***, or of ***class*** $\mathbf{C^2}$. Clearly this process can be continued inductively. Let us record the language coming from this iteration.

3.2.5 Definition (Higher-order derivatives) Let $I \subseteq \mathbb{R}$ be an interval, let $f : I \to \mathbb{R}$ be a function, let $r \in \mathbb{Z}_{>0}$, and suppose that f is $(r-1)$ times differentiable with g the corresponding $(r-1)$st derivative.

(i) The function f is **r *times differentiable at*** $\mathbf{x_0} \in \mathbf{I}$ if the limit

$$\lim_{x \to_I x_0} \frac{g(x) - g(x_0)}{x - x_0} \tag{3.8}$$

exists.

(ii) If the limit (3.8) exists, then it is denoted by $f^{(r)}(x_0)$ and called the **r*th derivative*** of f at x_0.

(iii) If f is r times differentiable at each point $x \in I$, then f is **r *times differentiable***.

(iv) If f is r times differentiable and if the function $x \mapsto f^{(r)}(x)$ is continuous, then f is **r *times continuously differentiable***, or of ***class*** $\mathbf{C^r}$.

If f is of class C^r for each $r \in \mathbb{Z}_{>0}$, then f is ***infinitely differentiable***, or of ***class*** $\mathbf{C^\infty}$. •

3.2.6 Notation (Class $\mathbf{C^0}$) A continuous function will sometimes be said to be of ***class*** $\mathbf{C^0}$, in keeping with the language used for functions that are differentiable to some order. •

3.2.2 The derivative and continuity

In this section we simply do two things. We show that differentiable functions are continuous (Proposition 3.2.7), and we (dramatically) show that the converse of this is not true (Example 3.2.9).

3.2.7 Proposition (Differentiable functions are continuous) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is a function differentiable at* $\mathrm{x}_0 \in \mathrm{I}$*, then* f *is continuous at* x_0*.*

Proof Using Propositions 2.3.23 and 2.3.29 the limit

$$\lim_{x \to_I x_0} \Big(\frac{f(x) - f(x_0)}{x - x_0}\Big)(x - x_0)$$

exists, and is equal to the product of the limits

$$\lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0}, \qquad \lim_{x \to_I x_0} (x - x_0),$$

i.e., is equal to zero. We therefore can conclude that

$$\lim_{x \to_I x_0} (f(x) - f(x_0)) = 0,$$

and the result now follows from Theorem 3.1.3. ∎

If the derivative is bounded, then there is more that one can say.

3.2.8 Proposition (Functions with bounded derivative are uniformly continuous) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is differentiable with* $\mathrm{f}'\colon \mathrm{I} \to \mathbb{R}$ *bounded, then* f *is uniformly continuous.*

Proof Let

$$M = \sup\{f'(t) \mid t \in I\}.$$

Then, for every $x, y \in I$, by the Mean Value Theorem, Theorem 3.2.19 below, there exists $z \in [x, y]$ such that

$$f(x) - f(y) = f'(z)(x - y) \quad \Longrightarrow \quad |f(x) - f(y)| \leq M\|x - y\|.$$

Now let $\epsilon \in \mathbb{R}_{>0}$ and let $x \in I$. Define $\delta = \frac{\epsilon}{M}$ and note that if $y \in I$ satisfies $|x - y| < \delta$ then we have

$$|f(x) - f(y)| \leq M\|x - y\| \leq \epsilon,$$

giving the desired uniform continuity. ∎

Of course, it is not true that a continuous function is differentiable; we have an example of this as Example 3.2.3–3. However, things are much worse than that, as the following example indicates.

3.2.9 Example (A continuous but nowhere differentiable function) For $k \in \mathbb{Z}_{>0}$ define $g_k\colon \mathbb{R} \to \mathbb{R}$ as shown in Figure 3.6. Thus g_k is periodic with period $4 \cdot 2^{-2^k}$.[3] We then define

$$f(x) = \sum_{k=1}^{\infty} 2^{-k} g_k(x).$$

[3]We have not yet defined what is meant by a periodic function, although this is likely clear. In case it is not, a function $f\colon \mathbb{R} \to \mathbb{R}$ is ***periodic*** with period $T \in \mathbb{R}_{>0}$ if $f(x + T) = f(x)$ for every $x \in \mathbb{R}$. Periodic functions will be discussed in some detail in Section **??**.

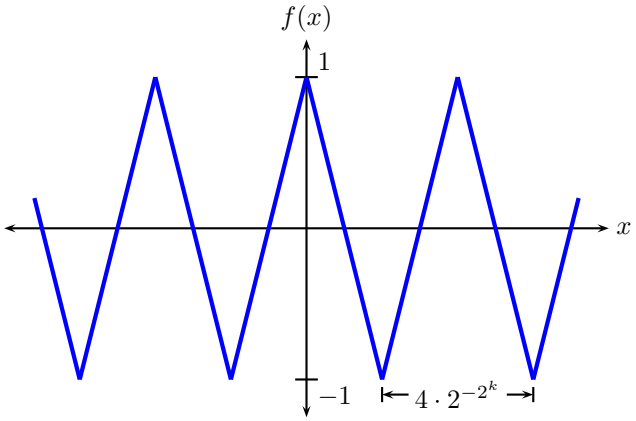


Figure 3.6 The function g_k

Since g_k is bounded in magnitude by 1, and since the sum $\sum_{k=1}^{\infty} 2^{-k}$ is absolutely convergent (Example 2.4.2–4), for each x the series defining f converges, and so f is well-defined. We claim that f is continuous but is nowhere differentiable.

It is easily shown by the Weierstrass M-test (see Theorem 3.4.15 below) that the series converges uniformly, and so defines a continuous function in the limit by Theorem 3.4.8. Thus f is continuous.

Now let us show that f is nowhere differentiable. Let $x \in \mathbb{R}$, $k \in \mathbb{Z}_{>0}$, and choose $h_k \in \mathbb{R}$ such that $|h| = 2^{-2^k}$ and such that x and $x + h_k$ lie on the line segment in the graph of g_k (this is possible since h_k is small enough, as is easily checked). Let us prove a few lemmata for this choice of x and h_k.

1 Lemma $g_l(x + h_k) = g(x)$ *for* $l > k$.

Proof This follows since g_l is periodic with period $4 \cdot 2^{-2^l}$, and so is therefore also periodic with period 2^{-2^k} since

$$\frac{4 \cdot 2^{-2^l}}{2^{-2^k}} = 4 \cdot 2^{-2^l - 2^k} \in \mathbb{Z}$$

for $l > k$. ▼

2 Lemma $|g_k(x + h_k) - g_k(x)| = 1$.

Proof This follows from the fact that we have chosen h_k such that x and $x + h_k$ lie on the same line segment in the graph of g_k, and from the fact that $|h_k|$ is one-quarter the period of g_k (cf. Figure 3.6). ▼

3 Lemma $\left|\sum_{j=1}^{k-1} 2^{-j} g_j(x + h_k) - \sum_{j=1}^{k-1} 2^{-j} g_j(x)\right| \leq 2^k 2^{-2^{k-1}}$.

Proof We note that if x and $x + h_k$ are on the same line segment in the graph of g_k, then they are also on the same line segment of the graph of g_j for $j \in \{1, \ldots, k\}$.

Using this fact, along with the fact that the slope of the line segments of the function g_j have magnitude 2^{2^j}, we compute

$$\begin{aligned}\Big|\sum_{j=1}^{k-1} 2^{-j} g_j(x+h_k) - \sum_{j=1}^{k-1} 2^{-j} g_j(x)\Big| & \\ \leq (k-1)\max\{|2^{-j}g_j(x+h_k) - 2^{-j}g_j(x)| \mid j \in \{1,\dots,k\}\} & \\ = (k-1)2^{2^{k-1}}2^{-2^k} < 2^k 2^{-2^{k-1}}. &\end{aligned}$$

The final inequality follows since $k-1 < 2^k$ for $k \geq 1$ and since $2^{2^{k-1}}2^{-2^k} = 2^{-2^{k-1}}$. ▼

Now we can assemble these lemmata to give the conclusion that f is not differentiable at x. Let $x \in \mathbb{R}$, let $\epsilon \in \mathbb{R}_{>0}$, choose $k \in \mathbb{Z}_{>0}$ such that $2^{-2^k} < \epsilon$, and choose h_k as above. We then have

$$\begin{aligned}\Big|\frac{f(x+h_k)-f(x)}{h_k}\Big| &= |\frac{\sum_{j=1}^{\infty} 2^{-j} g_j(x+h_k) - \sum_{j=1}^{\infty} 2^{-j} g_j(x)}{h_k}| \\ = |\frac{\sum_{j=1}^{k-1} 2^{-j} g_j(x+h_k) - \sum_{j=1}^{k-1} 2^{-j} g_j(x)}{h_k} &+ \frac{2^{-k}(g_k(x+h_k)-g_k(x))}{h_k}| \\ \geq 2^{-k}2^{2^k} - 2^k 2^{-2^{k-1}}.&\end{aligned}$$

Since $\lim_{k\to\infty}(2^{-k}2^{2^k} - 2^k 2^{-2^{k-1}}) = \infty$, it follows that any neighbourhood of x will contain a point y for which $\frac{f(y)-f(x)}{y-x}$ will be as large in magnitude as desired. This precludes f from being differentiable at x. Now, since x was arbitrary in our construction, we have shown that f is nowhere differentiable as claimed.

In Figure 3.7 we plot the function

$$f_k(x) = \sum_{j=1}^{k} 2^{-j} g_j(x)$$

for $j \in \{1,2,3,4\}$. Note that, to the resolution discernible by the eye, there is no difference between f_3 and f_4. However, if we were to magnify the scale, we would see the effects that lead to the limit function not being differentiable. •

3.2.3 The derivative and operations on functions

In this section we provide the rules for using the derivative in conjunction with the natural algebraic operations on functions as described at the beginning of Section 3.1.3. Most readers will probably be familiar with these ideas, at least inasmuch as how to use them in practice.

3.2.10 Proposition (The derivative, and addition and multiplication) *Let* $I \subseteq \mathbb{R}$ *be an interval and let* $f, g\colon I \to \mathbb{R}$ *be functions differentiable at* $x_0 \in I$. *Then the following statements hold:*

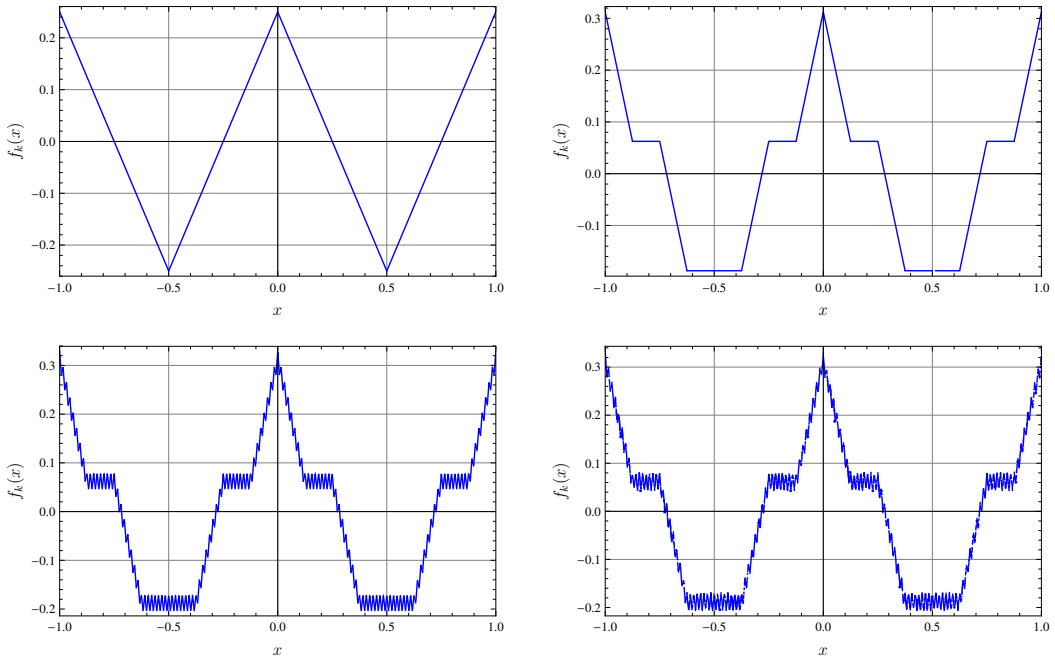


Figure 3.7 The first four partial sums for f

(i) $f + g$ *is differentiable at* x_0 *and* $(f + g)'(x_0) = f'(x_0) + g'(x_0)$*;*

(ii) fg *is differentiable at* x_0 *and* $(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0)$ *(**product rule** or **Leibniz'**[4] **rule**);*

(iii) if additionally $g(x_0) \neq 0$*, then* $\frac{f}{g}$ *is differentiable at* x_0 *and*

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2} \quad \textit{(\textbf{quotient rule})}.$$

Proof (i) We have

$$\frac{(f+g)(x) - (f+g)(x_0)}{x - x_0} = \frac{f(x) - f(x_0)}{x - x_0} + \frac{g(x) - g(x_0)}{x - x_0}.$$

Now we may apply Propositions 2.3.23 and 2.3.29 to deduce that

$$\lim_{x \to_I x_0} \frac{(f+g)(x) - (f+g)(x_0)}{x - x_0} = \lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0} + \lim_{x \to_I x_0} \frac{g(x) - g(x_0)}{x - x_0} = f'(x_0) + g'(x_0),$$

as desired.

[4]Gottfried Wilhelm von Leibniz (1646–1716) was born in Leipzig (then a part of Saxony), and was a lawyer, philosopher, and mathematician. His main mathematical contributions were to the development of calculus, where he had a well-publicised feud over priority with Newton, and algebra. His philosophical contributions, mainly in the area of logic, were also of some note.

(ii) Here we note that

$$\begin{aligned}\frac{(fg)(x)-(fg)(x_0)}{x-x_0} &= \frac{f(x)g(x)-f(x)g(x_0)+f(x)g(x_0)-f(x_0)g(x_0)}{x-x_0}\\ &= f(x)\frac{g(x)-g(x_0)}{x-x_0}+g(x_0)\frac{f(x)-f(x_0)}{x-x_0}.\end{aligned}$$

Since f is continuous at x_0 by Proposition 3.2.7, we may apply Propositions 2.3.23 and 2.3.29 to conclude that

$$\lim_{x\to_I x_0}\frac{(fg)(x)-(fg)(x_0)}{x-x_0} = f'(x_0)g(x_0)+f(x_0)g'(x_0),$$

just as claimed.

(iii) By using part (ii), it suffices to consider the case where f is defined by $f(x) = 1$ (why?). Note that if $g(x_0) \neq 0$, then there is a neighbourhood of x_0 to which the restriction of g is nowhere zero. Thus, without loss of generality, we suppose that $g(x) \neq 0$ for all $x \in I$. But in this case we have

$$\lim_{x\to_I x_0}\frac{\frac{1}{g(x)}-\frac{1}{g(x_0)}}{x-x_0} = \lim_{x\to_I x_0}\frac{1}{g(x)g(x_0)}\frac{g(x_0)}{x-x_0} = -\frac{g'(x_0)}{g(x_0)^2},$$

giving the result in this case. We have used Propositions 2.3.23 and 2.3.29 as usual. ■

The following generalisation of the product rule will be occasionally useful.

3.2.11 Proposition (Higher-order product rule) *Let* $I \subseteq \mathbb{R}$ *be an interval, let* $x_0 \in I$, *let* $r \in \mathbb{Z}_{>0}$, *and suppose that* $f, g\colon I \to \mathbb{R}$ *are of class* C^{r-1} *and are* r*-times differentiable at* x_0. *Then* fg *is* r*-times differentiable at* x_0, *and*

$$(fg)^{(r)}(x_0) = \sum_{j=0}^{r}\binom{r}{j}f^{(j)}(x_0)g^{(r-j)}(x_0),$$

where

$$\binom{r}{j} = \frac{r!}{j!(r-j)!}.$$

Proof The result is true for $r = 1$ by Proposition 3.2.10. So suppose the result true for $k \in \{1, \dots, r\}$. We then have

$$\begin{aligned}\frac{(fg)^{(r)}(x)-(fg)^{(r)}(x_0)}{x-x_0} &= \frac{\sum_{j=0}^{r}\binom{r}{j}f^{(j)}(x)g^{(r-j)}(x)-\sum_{j=0}^{r}\binom{r}{j}f^{(j)}(x_0)g^{(r-j)}(x_0)}{x-x_0}\\ &= \sum_{j=0}^{r}\binom{r}{j}\frac{f^{(j)}(x)g^{(r-j)}(x)-f^{(j)}(x_0)g^{(r-j)}(x_0)}{x-x_0}.\end{aligned}$$

Now we note that

$$\lim_{x\to_I x_0}\frac{f^{(j)}(x)g^{(r-j)}(x)-f^{(j)}(x_0)g^{(r-j)}(x_0)}{x-x_0} = f^{(j+1)}(x_0)g^{(r-j)}(x_0)+f^{(j)}(x_0)g^{(r-j+1)}(x_0).$$

Therefore,

$$\begin{aligned}
&\lim_{x\to_I x_0} \frac{(fg)^{(r)}(x) - (fg)^{(r)}(x_0)}{x - x_0} \\
&= \sum_{j=0}^{r} \binom{r}{j} \left(f^{(j+1)}(x_0) g^{(r-j)}(x_0) + f^{(j)}(x_0) g^{(r-j+1)}(x_0) \right) \\
&= f(x_0) g^{(r+1)}(x_0) + \sum_{j=0}^{r} \binom{r}{j} f^{(j+1)}(x_0) g^{(r-j)}(x_0) + \sum_{j=1}^{r} \binom{r}{j} f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&= f(x_0) g^{(r+1)}(x_0) + \sum_{j=1}^{r+1} \binom{r}{j-1} f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&\quad + \sum_{j=1}^{r} \binom{r}{j} f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&= f^{(r+1)}(x_0) g(x_0) + f(x_0) g^{(r+1)}(x_0) \\
&\quad + \sum_{j=1}^{r} \left(\binom{r}{j} + \binom{r}{j-1} \right) f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&= f^{(r+1)}(x_0) g(x_0) + f(x_0) g^{(r+1)}(x_0) + \sum_{j=1}^{r} \binom{r+1}{j} f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&= \sum_{j=0}^{r+1} \binom{r+1}{j} f^{(j)}(x_0) g^{(r-j)}(x_0).
\end{aligned}$$

In the penultimate step we have used ***Pascal's***[5] ***Rule*** which states that

$$\binom{r}{j} + \binom{r}{j-1} = \binom{r+1}{j}.$$

We leave the direct proof of this fact to the reader. ■

The preceding two results had to do with differentiability at a point. For convenience, let us record the corresponding results when we consider the derivative, not just at a point, but on the entire interval.

3.2.12 Proposition (Class C^r, and addition and multiplication) *Let $I \subseteq \mathbb{R}$ be an interval and let $f, g \colon I \to \mathbb{R}$ be functions of class C^r. Then the following statements hold:*

(i) $f + g$ *is of class* C^r*;*

(ii) fg *is of class* C^r*;*

(iii) if additionally $g(x) \neq 0$ *for all* $x \in I$*, then* $\frac{f}{g}$ *is of class* C^r*.*

Proof This follows directly from Propositions 3.2.10 and 3.2.11, along with the fact, following from Proposition 3.1.15, that the expressions for the derivatives of sums, products, and quotients are continuous, as they are themselves sums, products, and quotients. ■

[5]Blaise Pascal (1623–1662) was a French mathematician and philosopher. Much of his mathematical work was on analytic geometry and probability theory.

The following rule for differentiating the composition of functions is one of the more useful of the rules concerning the behaviour of the derivative.

3.2.13 Theorem (Chain Rule) *Let* $\mathrm{I}, \mathrm{J} \subseteq \mathbb{R}$ *be intervals and let* $\mathrm{f}\colon \mathrm{I} \to \mathrm{J}$ *and* $\mathrm{g}\colon \mathrm{J} \to \mathbb{R}$ *be functions for which* f *is differentiable at* $\mathrm{x}_0 \in \mathrm{I}$ *and* g *is differentiable at* $\mathrm{f}(\mathrm{x}_0) \in \mathrm{J}$. *Then* $\mathrm{g} \circ \mathrm{f}$ *is differentiable at* x_0, *and* $(\mathrm{g} \circ \mathrm{f})'(\mathrm{x}_0) = \mathrm{g}'(\mathrm{f}(\mathrm{x}_0))\mathrm{f}'(\mathrm{x}_0)$.

Proof Let us define $h\colon J \to \mathbb{R}$ by

$$h(y) = \begin{cases} \frac{g(y)-g(f(x_0))}{y-f(x_0)}, & g(y) \neq g(f(x_0)), \\ g'(f(x_0)), & g(y) = g(f(x_0)). \end{cases}$$

We have

$$\frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = \frac{(g \circ f)(x) - (g \circ f)(x_0)}{f(x) - f(x_0)} \frac{f(x) - f(x_0)}{x - x_0} = h(f(x)) \frac{f(x) - f(x_0)}{x - x_0},$$

provided that $f(x) \neq f(x_0)$. On the other hand, if $f(x) = f(x_0)$, we immediately have

$$\frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = h(f(x)) \frac{f(x) - f(x_0)}{x - x_0}$$

since both sides of this equation are zero. Thus we simply have

$$\frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = h(f(x)) \frac{f(x) - f(x_0)}{x - x_0}$$

for all $x \in I$. Note that h is continuous at $f(x_0)$ by Theorem 3.1.3 since

$$\lim_{y \to_I f(x_0)} h(y) = g'(x_0) = h(x_0),$$

using the fact that g is differentiable at x_0. Now we can use Propositions 2.3.23 and 2.3.29 to ascertain that

$$\lim_{x \to_I x_0} \frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = \lim_{x \to_I x_0} h(f(x)) \frac{f(x) - f(x_0)}{x - x_0} = g'(f(x_0)) f'(x_0),$$

as desired. ∎

The derivative behaves as one would expect when restricting a differentiable function.

3.2.14 Proposition (The derivative and restriction) *If* $\mathrm{I}, \mathrm{J} \subseteq \mathbb{R}$ *are intervals for which* $\mathrm{J} \subseteq \mathrm{I}$, *and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is differentiable at* $\mathrm{x}_0 \in \mathrm{J} \subseteq \mathrm{I}$, *then* $\mathrm{f}|\mathrm{J}$ *is differentiable at* x_0.

Proof This follows since if the limit

$$\lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists, then so too does the limit

$$\lim_{x \to_J x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

provided that $J \subseteq I$. ∎

missing stuff

3.2.4 The derivative and function behaviour

From the behaviour of the derivative of a function, it is often possible to deduce some important features of the function itself. One of the most important of these concerns maxima and minima of a function. Let us define these concepts precisely.

3.2.15 Definition (Local maximum and local minimum) Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be a function. A point $x_0 \in I$ is a:

(i) ***local maximum*** if there exists a neighbourhood U of x_0 such that $f(x) \leq f(x_0)$ for every $x \in U$;

(ii) ***strict local maximum*** if there exists a neighbourhood U of x_0 such that $f(x) < f(x_0)$ for every $x \in U \setminus \{x_0\}$;

(iii) ***local minimum*** if there exists a neighbourhood U of x_0 such that $f(x) \geq f(x_0)$ for every $x \in U$;

(iv) ***strict local minimum*** if there exists a neighbourhood U of x_0 such that $f(x) > f(x_0)$ for every $x \in U \setminus \{x_0\}$. •

Now we have the standard result that relates derivatives to maxima and minima.

3.2.16 Theorem (Derivatives, and maxima and minima) *For $\mathrm{I} \subseteq \mathbb{R}$ an interval, $\mathrm{f} : \mathrm{I} \to \mathbb{R}$ a function, and $\mathrm{x}_0 \in \mathrm{int}(\mathrm{I})$, the following statements hold:*

(i) if f is differentiable at x_0 and if x_0 is a local maximum or a local minimum for f, then $\mathrm{f}'(\mathrm{x}_0) = 0$;

(ii) if f is twice differentiable at x_0, and if x_0 is a local maximum (resp. local minimum) for f, then $\mathrm{f}''(\mathrm{x}_0) \leq 0$ (resp. $\mathrm{f}''(\mathrm{x}_0) \geq 0$);

(iii) if f is twice differentiable at x_0, and if $\mathrm{f}'(\mathrm{x}_0) = 0$ and $\mathrm{f}''(\mathrm{x}_0) \in \mathbb{R}_{<0}$ (resp. $\mathrm{f}''(\mathrm{x}_0) \in \mathbb{R}_{>0}$), then x_0 is a strict local maximum (resp. strict local minimum) for f.

Proof (i) We will prove the case where x_0 is a local minimum, since the case of a local maximum is similar. If x_0 is a local minimum, then there exists $\epsilon \in \mathbb{R}_{>0}$ such that $f(x) \geq f(x_0)$ for all $x \in \mathsf{B}(\epsilon, x_0)$. Therefore, $\frac{f(x)-f(x_0)}{x-x_0} \geq 0$ for $x \geq x_0$ and $\frac{f(x)-f(x_0)}{x-x_0} \leq 0$ for $x \leq x_0$. Since the limit $\lim_{x \to x_0} \frac{f(x)-f(x_0)}{x-x_0}$ exists, it must be equal to both limits

$$\lim_{x \downarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}, \quad \lim_{x \uparrow x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

However, since the left limit is nonnegative and the right limit is nonpositive, we conclude that $f'(x_0) = 0$.

(ii) We shall show that if f is twice differentiable at x_0 and $f''(x_0)$ is not less than or equal to zero, then x_0 is not a local maximum. The statement concerning the local minimum is argued in the same way. Now, if f is twice differentiable at x_0, and if $f''(x_0) \in \mathbb{R}_{>0}$, then x_0 is a local minimum by part (iii), which prohibits it from being a local maximum.

(iii) We consider the case where $f''(x_0) \in \mathbb{R}_{>0}$, since the other case follows in the same manner. Choose $\epsilon \in \mathbb{R}_{>0}$ such that, for $x \in \mathsf{B}(\epsilon, x_0)$,

$$\left| \frac{f'(x) - f'(x_0)}{x - x_0} - f''(x_0) \right| < \tfrac{1}{2} f''(x_0),$$

this being possible since $f''(x_0) > 0$ and since f is twice differentiable at x_0. Since $f''(x_0) > 0$ it follows that, for $x \in \mathsf{B}(\epsilon, x_0)$,

$$\frac{f'(x) - f'(x_0)}{x - x_0} > 0,$$

from which we conclude that $f'(x) > 0$ for $x \in (x_0, x_0 + \epsilon)$ and that $f'(x) < 0$ for $x \in (x_0 - \epsilon, x_0)$. Now we prove a technical lemma.

1 Lemma *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an open interval, let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a continuous function that is differentiable, except possibly at* $\mathrm{x}_0 \in \mathrm{I}$. *If* $\mathrm{f}'(\mathrm{x}) > 0$ *for every* $\mathrm{x} > \mathrm{x}_0$ *and if* $\mathrm{f}'(\mathrm{x}) < 0$ *for every* $\mathrm{x} < \mathrm{x}_0$, *then* x_0 *is a strict local minimum for* f.

Proof We will use the Mean Value Theorem (Theorem 3.2.19) which we prove below. Note that our proof of the Mean Value Theorem depends on part (i) of the present theorem, but not on part that we are now proving. Let $x \in I \setminus \{x_0\}$. We have two cases.

1. $x > x_0$: By the Mean Value Theorem there exists $a \in (x, x_0)$ such that $f(x) - f(x_0) = (x - x_0)f'(a)$. Since $f'(a) > 0$ it then follows that $f(x) > f(x_0)$.
2. $x < x_0$: A similar argument as in the previous case again gives $f(x) > f(x_0)$.

Combining these conclusions, we see that $f(x) > f(x_0)$ for all $x \in I$, and so x_0 is a strict local maximum for f. ▼

The lemma now immediately applies to the restriction of f to $\mathsf{B}(\epsilon, x_0)$, and so gives the result. ■

Let us give some examples that illustrate the value and limitations of the preceding result.

3.2.17 Examples (Derivatives, and maxima and minima)

1. Let $I = \mathbb{R}$ and define $f\colon I \to \mathbb{R}$ by $f(x) = x^2$. Note that f is infinitely differentiable, so Theorem 3.2.16 can be applied freely. We compute $f'(x) = 2x$, and so $f'(x) = 0$ if and only if $x = 0$. Therefore, the only local maxima and local minima must occur at $x = 0$. To check whether a local maxima, a local minima, or neither exists at $x = 0$, we compute the second derivative which is $f''(x) = 2$. This is positive at $x = 0$ (and indeed everywhere), so we may conclude that $x = 0$ is a strict local maximum for f from part (iii) of the theorem.
   Applying the same computations to $g(x) = -x^2$ shows that $x = 0$ is a strict local maximum for g.
2. Let $I = \mathbb{R}$ and define $f\colon I \to \mathbb{R}$ by $f(x) = x^3$. We compute $f'(x) = 3x^2$, from which we ascertain that all maxima and minima must occur, if at all, at $x = 0$. However, since $f''(x) = 6x$, $f''(0) = 0$, and we cannot conclude from Theorem 3.2.16 whether there is a local maximum, a local minimum, or neither at $x = 0$. In fact, one can see "by hand" that $x = 0$ is neither a local maximum nor a local minimum for f.
   The same arguments apply to the functions $g(x) = x^4$ and $h(x) = -x^4$ to show that when the second derivative vanishes, it is possible to have all possibilities—a local maximum, a local minimum, or neither—at a point where both f' and f'' are zero.

3. Let $I = [-1, 1]$ and define $f : I \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1 - x, & x \in [0, 1], \\ 1 + x, & x \in [-1, 0). \end{cases}$$

"By hand," one can check that f has a strict local maximum at $x = 0$, and strict local minima at $x = -1$ and $x = 1$. However, we can detect none of these using Theorem 3.2.16. Indeed, the local minima at $x = -1$ and $x = 1$ occur at the boundary of I, and so the hypotheses of the theorem do not apply. This, indeed, is why we demand that x_0 lie in $\operatorname{int}(I)$ in the theorem statement. For the local maximum at $x = 0$, the theorem does not apply since f is not differentiable at $x = 0$. However, we do note that Lemma 1 (with modifications to the signs of the derivative in the hypotheses, and changing "minimum" to "maximum" in the conclusions) in the proof of the theorem *does* apply, since f is differentiable at points in $(-1, 0)$ and $(0, 1)$, and for $x > 0$ we have $f'(x) < 0$ and for $x < 0$ we have $f'(x) > 0$. The lemma then allows us to conclude that f has a strict local maximum at $x = 0$. •

Next let us prove a simple result that, while not always of great value itself, leads to the important Mean Value Theorem below.

3.2.18 Theorem (Rolle's[6] Theorem) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{f} : \mathrm{I} \to \mathbb{R}$ *be continuous, and suppose that for* $\mathrm{a}, \mathrm{b} \in \mathrm{I}$ *it holds that* $\mathrm{f}|(\mathrm{a}, \mathrm{b})$ *is differentiable and that* $\mathrm{f}(\mathrm{a}) = \mathrm{f}(\mathrm{b})$. *Then there exists* $\mathrm{c} \in (\mathrm{a}, \mathrm{b})$ *such that* $\mathrm{f}'(\mathrm{c}) = 0$.

Proof Since $f|[a, b]$ is continuous, by Theorem 3.1.23 there exists $x_1, x_2 \in [a, b]$ such that $\operatorname{image}(f|[a, b]) = [f(x_1), f(x_2)]$. We have three cases to consider.

1. $x_1, x_2 \in \operatorname{bd}([a, b])$: In this case it holds that f is constant since $f(a) = f(b)$. Thus the conclusions of the theorem hold for any $c \in (a, b)$.
2. $x_1 \in \operatorname{int}([a, b])$: In this case, f has a local minimum at x_1, and so by Theorem 3.2.16(i) we conclude that $f'(x_1) = 0$.
3. $x_2 \in \operatorname{int}([a, b])$: In this case, f has a local maximum at x_2, and so by Theorem 3.2.16(i) we conclude that $f'(x_2) = 0$. ■

Rolle's Theorem has the following generalisation, which is often quite useful, since it establishes links between the values of a function and the values of its derivative.

3.2.19 Theorem (Mean Value Theorem) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{f} : \mathrm{I} \to \mathbb{R}$ *be continuous, and suppose that for* $\mathrm{a}, \mathrm{b} \in \mathrm{I}$ *it holds that* $\mathrm{f}|(\mathrm{a}, \mathrm{b})$ *is differentiable. Then there exists* $\mathrm{c} \in (\mathrm{a}, \mathrm{b})$ *such that*

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

Proof Define $g : I \to \mathbb{R}$ by

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

[6] Michel Rolle (1652–1719) was a French mathematician whose primary contributions were to algebra.

Using the results of Section 3.2.3 we conclude that g is continuous and differentiable on (a, b). Moreover, direct substitution shows that $g(b) = g(a)$. Thus Rolle's Theorem allows us to conclude that there exists $c \in (a, b)$ such that $g'(c) = 0$. However, another direct substitution shows that $g'(c) = f'(c) - \frac{f(b)-f(a)}{b-a}$. ■

In Figure 3.8 we give the intuition for Rolle's Theorem, the Mean Value Theorem,

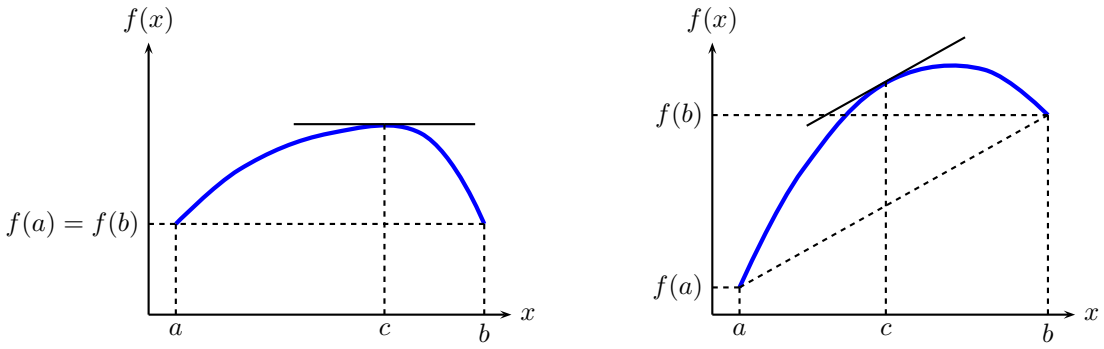


Figure 3.8 Illustration of Rolle's Theorem (left) and the Mean Value Theorem (right)

and the relationship between the two results.

Another version of the Mean Value Theorem relates the values of two functions with the values of their derivatives.

3.2.20 Theorem (Cauchy's Mean Value Theorem) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{f}, \mathrm{g} \colon \mathrm{I} \to \mathbb{R}$ *be continuous, and suppose that for* $\mathrm{a}, \mathrm{b} \in \mathrm{I}$ *it holds that* $\mathrm{f}|(\mathrm{a}, \mathrm{b})$ *and* $\mathrm{g}|(\mathrm{a}, \mathrm{b})$ *are differentiable, and that* $\mathrm{g}'(\mathrm{x}) \neq 0$ *for each* $\mathrm{x} \in (\mathrm{a}, \mathrm{b})$. *Then there exists* $\mathrm{c} \in (\mathrm{a}, \mathrm{b})$ *such that*

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

Proof Note that $g(b) \neq g(a)$ by Rolle's Theorem, since $g'(x) \neq 0$ for $x \in \mathrm{int}(a, b)$. Let

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

and define $h \colon I \to \mathbb{R}$ by $h(x) = f(x) - \alpha g(x)$. Using the results of Section 3.2.3, one verifies that h is continuous on I and differentiable on (a, b). Moreover, one can also verify that $h(a) = h(b)$. Thus Rolle's Theorem implies the existence of $c \in (a, b)$ for which $h'(c) = 0$. A simple computation verifies that $h'(c) = 0$ is equivalent to the conclusion of the theorem. ■

We conclude this section with the useful L'Hôpital's Rule. This rule for finding limits is sufficiently useful that we state and prove it here in an unusual level of generality.

3.2.21 Theorem (L'Hôpital's[7] Rule) *Let $I \subseteq \mathbb{R}$ be an interval, let $x_0 \in \mathbb{R}$, and let $f, g\colon I \to \mathbb{R}$ be differentiable functions with $g'(x) \neq 0$ for all $x \in I - \{x_0\}$. Then the following statements hold.*

(i) Suppose that x_0 is an open right endpoint for I and suppose that either

(a) $\lim_{x\uparrow x_0} f(x) = 0$ and $\lim_{x\uparrow x_0} g(x) = 0$ or

(b) $\lim_{x\uparrow x_0} f(x) = \infty$ and $\lim_{x\uparrow x_0} g(x) = \infty$,

and suppose that $\lim_{x\uparrow x_0} \frac{f'(x)}{g'(x)} = s_0 \in \overline{\mathbb{R}}$. Then $\lim_{x\uparrow x_0} \frac{f(x)}{g(x)} = s_0$.

(ii) Suppose that x_0 is an left right endpoint for I and suppose that either

(a) $\lim_{x\downarrow x_0} f(x) = 0$ and $\lim_{x\downarrow x_0} g(x) = 0$ or

(b) $\lim_{x\uparrow x_0} f(x) = \infty$ and $\lim_{x\downarrow x_0} g(x) = \infty$,

and suppose that $\lim_{x\downarrow x_0} \frac{f'(x)}{g'(x)} = s_0 \in \overline{\mathbb{R}}$. Then $\lim_{x\downarrow x_0} \frac{f(x)}{g(x)} = s_0$.

(iii) Suppose that $x_0 \in \operatorname{int}(I)$ and suppose that either

(a) $\lim_{x\to x_0} f(x) = 0$ and $\lim_{x\to x_0} g(x) = 0$ or

(b) $\lim_{x\to x_0} f(x) = \infty$ and $\lim_{x\to x_0} g(x) = \infty$,

and suppose that $\lim_{x\to x_0} \frac{f'(x)}{g'(x)} = s_0 \in \overline{\mathbb{R}}$. Then $\lim_{x\to x_0} \frac{f(x)}{g(x)} = s_0$.

The following two statements which are independent of x_0 (thus we ask that $g'(x) \neq 0$ for all $x \in I$) also hold.

(iv) Suppose that I is unbounded on the right and suppose that either

(a) $\lim_{x\to\infty} f(x) = 0$ and $\lim_{x\to\infty} g(x) = 0$ or

(b) $\lim_{x\to\infty} f(x) = \infty$ and $\lim_{x\to\infty} g(x) = \infty$,

and suppose that $\lim_{x\to\infty} \frac{f'(x)}{g'(x)} = s_0 \in \overline{\mathbb{R}}$. Then $\lim_{x\to\infty} \frac{f(x)}{g(x)} = s_0$.

(v) Suppose that I is unbounded on the left and suppose that either

(a) $\lim_{x\to-\infty} f(x) = 0$ and $\lim_{x\to-\infty} g(x) = 0$ or

(b) $\lim_{x\to-\infty} f(x) = \infty$ and $\lim_{x\to-\infty} g(x) = \infty$,

and suppose that $\lim_{x\to-\infty} \frac{f'(x)}{g'(x)} = s_0 \in \overline{\mathbb{R}}$. Then $\lim_{x\to-\infty} \frac{f(x)}{g(x)} = s_0$.

Proof (i) First suppose that $\lim_{x\uparrow x_0} f(x) = 0$ and $\lim_{x\uparrow x_0} g(x) = 0$ and that $s_0 \in \mathbb{R}$. We may then extend f and g to be defined at x_0 by taking their values at x_0 to be zero, and the resulting function will be continuous by Theorem 3.1.3. We may now apply Cauchy's Mean Value Theorem to assert that for $x \in I$ there exists $c_x \in (x, x_0)$ such that

$$\frac{f'(c_x)}{g'(c_x)} = \frac{f(x_0) - f(x)}{g(x_0) - g(x)} = \frac{f(x)}{g(x)}.$$

Now let $\epsilon \in \mathbb{R}_{>0}$ and choose $\delta \in \mathbb{R}_{>0}$ such that $\left|\frac{f'(x)}{g'(x)} - s_0\right| < \epsilon$ for $x \in \mathsf{B}(\delta, x_0) \cap I$. Then, for $x \in \mathsf{B}(\delta, x_0) \cap I$ we have

$$\left|\frac{f(x)}{g(x)} - s_0\right| = \left|\frac{f'(c_x)}{g'(c_x)} - s_0\right| < \epsilon$$

[7] Guillaume François Antoine Marquis de L'Hôpital (1661–1704) was one of the early developers of calculus.

since $c_x \in \mathsf{B}(\delta, x_0) \cap I$. This shows that $\lim_{x \uparrow x_0} \frac{f(x)}{g(x)} = s_0$, as claimed.

Now suppose that $\lim_{x \uparrow x_0} f(x) = \infty$ and $\lim_{x \uparrow x_0} g(x) = \infty$ and that $s_0 \in \mathbb{R}$. Let $\epsilon \in \mathbb{R}_{>0}$ and choose $\delta_1 \in \mathbb{R}_{>0}$ such that $\left|\frac{f'(x)}{g'(x)} - s_0\right| < \frac{\epsilon}{2(1+|s_0|)}$ for $x \in \mathsf{B}(\delta_1, x_0) \cap I$. For $x \in \mathsf{B}(\delta_1, x_0) \cap I$, by Cauchy's Mean Value Theorem there exists $c_x \in \mathsf{B}(\delta_1, x_0) \cap I$ such that

$$\frac{f'(c_x)}{g'(c_x)} = \frac{f(x) - f(x - \delta_1)}{g(x) - g(x - \delta_1)}.$$

Therefore,

$$\left|\frac{f(x) - f(x - \delta_1)}{g(x) - g(x - \delta_1)} - s_0\right| < \frac{\epsilon}{2(1 + |s_0|)}$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$. Now define

$$h(x) = \frac{1 - \frac{f(x-\delta_1)}{f(x)}}{1 - \frac{g(x-\delta_1)}{g(x)}}$$

and note that

$$\frac{f(x) - f(x - \delta_1)}{g(x) - g(x - \delta_1)} = h(x)\frac{f(x)}{g(x)}.$$

Therefore we have

$$\left|h(x)\frac{f(x)}{g(x)} - s_0\right| < \frac{\epsilon}{2(1 + |s_0|)}$$

for $x \in \mathsf{B}(\delta_1, x_0) \cap I$. Note also that $\lim_{x \uparrow x_0} h(x) = 1$. Thus we can choose $\delta_2 \in \mathbb{R}_{>0}$ such that $|h(x) - 1| < \frac{\epsilon}{2(1+|s_0|)}$ and $h(x) > \frac{1}{2}$ for $x \in \mathsf{B}(\delta_2, x_0) \cap I$. Then define $\delta = \min\{\delta_1, \delta_2\}$. For $x \in \mathsf{B}(\delta, x_0) \cap I$ we then have

$$\begin{aligned}\left|h(x)\Big(\frac{f(x)}{g(x)} - s_0\Big)\right| &= \left|h(x)\frac{f(x)}{g(x)} - h(x)s_0\right| \\ &\leq \left|h(x)\frac{f(x)}{g(x)} - s_0\right| + |(1 - h(x))s_0| \\ &< \frac{\epsilon}{2(1 + |s_0|)} + \frac{\epsilon}{2(1 + |s_0|)}|s_0| = \frac{\epsilon}{2}.\end{aligned}$$

Then, finally,

$$\left|\frac{f(x)}{g(x)} - s_0\right| < \frac{\epsilon}{2h(x)} < \epsilon,$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$.

Now we consider the situation when $s_0 \in \{-\infty, \infty\}$. We shall take only the case of $s_0 = \infty$ since the other follows in a similar manner. We first take the case where $\lim_{x \uparrow x_0} f(x) = 0$ and $\lim_{x \uparrow x_0} g(x) = 0$. In this case, for $x \in I$, from the Cauchy Mean Value Theorem we can find $c_x \in (x, x_0)$ such that

$$\frac{f'(c_x)}{g'(c_x)} = \frac{f(x)}{g(x)}.$$

Now for $M \in \mathbb{R}_{>0}$ we choose $\delta \in \mathbb{R}_{>0}$ such that for $x \in \mathsf{B}(\delta, x_0) \cap I$ we have $\frac{f'(x)}{g'(x)} > M$. Then we immediately have

$$\frac{f(x)}{g(x)} = \frac{f'(c_x)}{g'(c_x)} > M$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$ since $c_x \in \mathsf{B}(\delta, x_0)$, which gives the desired conclusion.

The final case we consider in this part of the proof is that where $s_0 = \infty$ and $\lim_{x \uparrow x_0} f(x) = \infty$ and $\lim_{x \uparrow x_0} g(x) = \infty$. For $M \in \mathbb{R}_{>0}$ choose $\delta_1 \in \mathbb{R}_{>0}$ such that $\frac{f'(x)}{g'(x)} > 2M$ provided that $x \in \mathsf{B}(\delta_1, x_0) \cap I$. Then, using Cauchy's Mean Value Theorem, for $x \in \mathsf{B}(\delta_1, x_0) \cap I$ there exists $c_x \in \mathsf{B}(\delta_1, x_0)$ such that

$$\frac{f'(c_x)}{g'(c_x)} = \frac{f(x) - f(x - \delta_1)}{g(x) - g(x - \delta_1)}.$$

Therefore,

$$\frac{f(x) - f(x - \delta_1)}{g(x) - g(x - \delta_1)} > 2M$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$. As above, define

$$h(x) = \frac{1 - \frac{f(x-\delta_1)}{f(x)}}{1 - \frac{g(x-\delta_1)}{g(x)}}$$

and note that

$$\frac{f(x) - f(x - \delta_1)}{g(x) - g(x - \delta_1)} = h(x)\frac{f(x)}{g(x)}.$$

Therefore

$$h(x)\frac{f(x)}{g(x)} > 2M$$

for $x \in \mathsf{B}(\delta_1, x_0)$. Now take $\delta_2 \in \mathbb{R}_{>0}$ such that, if $x \in \mathsf{B}(\delta_2, x_0) \cap I$, then $h(x) \in [\frac{1}{2}, 2]$, this being possible since $\lim_{x \uparrow x_0} h(x) = 1$. It then follows that

$$\frac{f(x)}{g(x)} > \frac{2M}{h(x)} > M$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$ where $\delta = \min\{\delta_1, \delta_2\}$.

(ii) This follows in the same manner as part (i).

(iii) This follows from parts (i) and (ii).

(iv) Let us define $\phi \colon (0, \infty) \to (0, \infty)$ by $\phi(x) = \frac{1}{x}$. Then define $\tilde{I} = \phi(I)$, noting that $\tilde{I}$ is an interval having 0 as an open left endpoint. Now define $\tilde{f}, \tilde{g} \colon \tilde{I} \to \mathbb{R}$ by $\tilde{f} = f \circ \phi$ and $\tilde{g} = g \circ \phi$. Using the Chain Rule (Theorem 3.2.13 below) we compute

$$\tilde{f}'(\tilde{x}) = f'(\phi(\tilde{x}))\phi'(\tilde{x}) = -\frac{f'(\frac{1}{\tilde{x}})}{\tilde{x}^2}$$

and similarly $\tilde{g}'(\tilde{x}) = -\frac{f'(\frac{1}{\tilde{x}})}{\tilde{x}^2}$. Therefore, for $\tilde{x} \in \tilde{I}$,

$$\frac{f'(\frac{1}{\tilde{x}})}{g'(\frac{1}{\tilde{x}})} = \frac{\tilde{f}'(\tilde{x})}{\tilde{g}'(\tilde{x})}.$$

and so, using part (ii) (it is easy to see that the hypotheses are verified),

$$\begin{aligned} & \lim_{\tilde{x}\downarrow 0} \frac{f'(\frac{1}{\tilde{x}})}{g'(\frac{1}{\tilde{x}})} = \lim_{\tilde{x}\downarrow 0} \frac{\tilde{f}'(\tilde{x})}{\tilde{g}'(\tilde{x})} \\ \implies & \lim_{x\to\infty} \frac{f'(x)}{g'(x)} = \lim_{\tilde{x}\downarrow 0} \frac{\tilde{f}(\tilde{x})}{\tilde{g}(\tilde{x})} \\ \implies & \lim_{x\to\infty} \frac{f'(x)}{g'(x)} = \lim_{x\to\infty} \frac{f(x)}{g(x)}, \end{aligned}$$

which is the desired conclusion.

(v) This follows in the same manner as part (iv). ∎

3.2.22 Examples (Uses of L'Hôpital's Rule)

1. Let $I = \mathbb{R}$ and define $f, g\colon I \to \mathbb{R}$ by $f(x) = \sin x$ and $g(x) = x$. Note that f and g satisfy the hypotheses of Theorem 3.2.21 with $x_0 = 0$. Therefore we may compute
$$\lim_{x\to 0} \frac{f(x)}{g(x)} = \lim_{x\to 0} \frac{f'(x)}{g'(x)} = \frac{\cos 0}{1} = 1.$$
2. Let $I = [0, 1]$ and define $f, g\colon I \to \mathbb{R}$ by $f(x) = \sin x$ and $g(x) = x^2$. We can verify that f and g satisfy the hypotheses of L'Hôpital's Rule with $x_0 = 0$. Therefore we compute
$$\lim_{x\downarrow 0} \frac{f(x)}{g(x)} = \lim_{x\downarrow 0} \frac{f'(x)}{g'(x)} = \lim_{x\downarrow 0} \frac{\cos x}{2x} = \infty.$$
3. Let $I = \mathbb{R}_{>0}$ and define $f, g\colon I \to \mathbb{R}$ by $f(x) = \mathrm{e}^x$ and $g(x) = -x$. Note that $\lim_{x\to\infty} f(x) = \infty$ and that $\lim_{x\to\infty} g(x) = -\infty$. Thus f and g do not quite satisfy the hypotheses of part (iv) of Theorem 3.2.21 since $\lim_{x\to\infty} g(x) \neq \infty$. However, the problem is a superficial one, as we now illustrate. Define $\tilde{g}(x) = -g(x) = x$. Then f and $\tilde{g}$ *do* satisfy the hypotheses of Theorem 3.2.21(iv). Therefore,
$$\lim_{x\to\infty} \frac{f(x)}{\tilde{g}(x)} = \lim_{x\to\infty} \frac{f'(x)}{\tilde{g}'(x)} = \lim_{x\to\infty} \frac{\mathrm{e}^x}{1} = \infty,$$
and so
$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} -\frac{f(x)}{\tilde{g}(x)} = -\infty.$$
4. Consider the function $h\colon \mathbb{R} \to \mathbb{R}$ defined by $h(x) = \frac{x}{\sqrt{1+x^2}}$. We wish to determine $\lim_{x\to\infty} h(x)$, if this limit indeed exists. We will try to use L'Hôpital's Rule with $f(x) = x$ and $g(x) = \sqrt{1+x^2}$. First, one should check that f and g satisfy the hypotheses of the theorem taking $x_0 = 0$. One can check that f and g are differentiable on I and that $g'(x)$ is nonzero for $x \in I\backslash\{x_0\}$. Moreover, $\lim_{x\to 0} f(x) = 0$ and $\lim_{x\to 0} g(x) = 0$. Thus it only remains to check that $\lim_{x\to 0} \frac{f'(x)}{g'(x)} \in \overline{\mathbb{R}}$. To this end, one can easily compute that
$$\frac{f'(x)}{g'(x)} = \frac{g(x)}{f(x)},$$

which immediately implies that an application of L'Hôpital's Rule is destined to fail. However, the actual limit $\lim_{x\to\infty} h(x)$ does exist, however, and is readily computed, using the definition of limit, to be 1. Thus the converse of L'Hôpital's Rule does not hold. •

3.2.5 Monotonic functions and differentiability

In Section 3.1.5 we considered the notion of monotonicity, and its relationship with continuity. In this section we see how monotonicity is related to differentiability.

For functions that are differentiable, the matter of deciding on their monotonicity properties is straightforward.

3.2.23 Proposition (Monotonicity for differentiable functions) *For $I \subseteq \mathbb{R}$ an interval and $f\colon I \to \mathbb{R}$ a differentiable function, the following statements hold:*

(i) *f is constant if and only if $f'(x) = 0$ for all $x \in I$;*

(ii) *f is monotonically increasing if and only $f'(x) \geq 0$ for all $x \in I$;*

(iii) *f is strictly monotonically increasing if and only $f'(x) > 0$ for all $x \in I$;*

(iv) *f is monotonically decreasing if and only if $f'(x) \leq 0$ for all $x \in I$.*

(v) *f is strictly monotonically decreasing if and only if $f'(x) < 0$ for all $x \in I$.*

Proof In each case the "only if" assertions follow immediately from the definition of the derivative. To prove the "if" assertions, let $x_1, x_2 \in I$ with $x_1 < x_2$. By the Mean Value Theorem there exists $c \in [x_1, x_2]$ such that $f(x_1) - f(x_2) = f'(c)(x_1 - x_2)$. The result follows by considering the three cases of $f'(c) = 0$, $f'(c) \leq 0$, $f'(c) > 0$, $f'(c) \leq 0$, and $f'(c) < 0$, respectively. ■

The previous result gives the relationship between the derivative and monotonicity. Combining this with Theorem 3.1.30 which relates monotonicity with invertibility, we obtain the following characterisations of the derivative of the inverse function.

3.2.24 Theorem (Inverse Function Theorem for $\mathbb{R}$) *Let $I \subseteq J$ be an interval, let $x_0 \in I$, and let $f\colon I \to J = \text{image}(f)$ be a continuous, strictly monotonically increasing function that is differentiable at x_0 and for which $f'(x_0) \neq 0$. Then $f^{-1}\colon J \to I$ is differentiable at $f(x_0)$ and the derivative is given by*

$$(f^{-1})'(f(x_0)) = \frac{1}{f'(x_0)}.$$

Proof From Theorem 3.1.30 we know that f is invertible. Let $y_0 = f(x_0)$, let $y_1 \in J$, and define $x_1 \in I$ by $f(x_1) = y_1$. Then, if $x_1 \neq x_0$,

$$\frac{f^{-1}(y_1) - f^{-1}(y_0)}{y_1 - y_0} = \frac{x_1 - x_0}{f(x_1) - f(x_0)}.$$

Therefore,

$$(f^{-1})'(y_0) = \lim_{y_1 \to_J y_0} \frac{f^{-1}(y_1) - f^{-1}(y_0)}{y_1 - y_0} = \lim_{x_1 \to_I x_0} \frac{x_1 - x_0}{f(x_1) - f(x_0)} = \frac{1}{f'(x_0)},$$

as desired. ■

3.2.25 Corollary (Alternate version of Inverse Function Theorem) *Let* $I \subseteq \mathbb{R}$ *be an interval, let* $x_0 \in I$*, and let* $f\colon I \to \mathbb{R}$ *be a function of class* C^1 *such that* $f'(x_0) \neq 0$*. Then there exists a neighbourhood* U *of* x_0 *in* I *and a neighbourhood* V *of* $f(x_0)$ *such that* $f|U$ *is invertible, and such that* $(f|U)^{-1}$ *is differentiable, and the derivative is given by*

$$((f|U)^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}$$

for each $y \in V$*.*

Proof Since f' is continuous and is nonzero at x_0, there exists a neighbourhood U of x_0 such that $f'(x)$ has the same sign as $f'(x_0)$ for all $x \in U$. Thus, by Proposition 3.2.23, $f|U$ is either strictly monotonically increasing (if $f'(x_0) > 0$) or strictly monotonically decreasing (if $f'(x_0) < 0$). The result now follows from Theorem 3.2.24. ■

For general monotonic functions, Proposition 3.2.23 turns out to be "almost" enough to characterise them. To understand this, we recall from Section 2.5.6 the notion of a subset of $\mathbb{R}$ of measure zero. With this recollection having been made, we have the following characterisation of general monotonic functions.

3.2.26 Theorem (Characterisation of monotonic functions II) *If* $I \subseteq \mathbb{R}$ *is an interval and if* $f\colon I \to \mathbb{R}$ *is either monotonically increasing (resp. monotonically decreasing), then* f *is differentiable almost everywhere, and* $f'(x) \geq 0$ *(resp.* $f'(x) \leq 0$*) at all points* $x \in I$ *where* f *is differentiable.*

Proof We first prove a technical lemma.

1 Lemma *If* $g\colon [a,b] \to \mathbb{R}$ *has the property that, for each* $x \in [a,b]$*, the limits* $g(x+)$ *and* $g(x-)$ *exist whenever they are defined as limits in* $[a,b]$*. If we define*

$$S = \{x \in [a,b] \mid \textit{there exists } x' > x \textit{ such that } g(x') > \max\{g(x-), g(x), g(x+)\}\},$$

then S *is a disjoint union of a countable collection* $\{I_\alpha \mid \alpha \in A\}$ *of intervals that are open as subsets of* $[a,b]$ *(cf. the beginning of Section 3.1.1).*

Proof Let $x \in S$. We have three cases.

1. There exists $x' > x$ such that $g(x') > g(x-)$, and $g(x-) \geq g(x)$ and $g(x-) \geq g(x+)$: Define $g_{x,-}, g_{x,+}\colon [a,b] \to \mathbb{R}$ by

$$g_{x,-}(y) = \begin{cases} g(y), & y \neq 1, \\ g(x-), & y = x, \end{cases} \qquad g_{x,+}(y) = \begin{cases} g(y), & y \neq 1, \\ g(x+), & y = x. \end{cases}$$

Since the limit $g(x-)$ exists, $g_{x,-}|[a,x]$ is continuous at x by Theorem 3.1.3. Since $g(x') > g_{x,-}(x)$, there exists $\epsilon_1 \in \mathbb{R}_{>0}$ such that $g(x') > g_{x,-}(y) = g(y)$ for all $y \in (x - \epsilon_1, x)$. Now note that $g(x') > g(x-) \geq g_{x,+}(x)$. Arguing similarly to what we have done, there exists $\epsilon_2 \in \mathbb{R}_{>0}$ such that $g(x') > g_{x,+}(y) = g(y)$ for all $y \in (x, x+\epsilon_2)$. Let $\epsilon = \min\{\epsilon_1, \epsilon_2\}$. Since $g(x') > g(x-) \geq g(x)$, it follows that $g(x') > g(y)$ for all $y \in (x - \epsilon, x + \epsilon)$, so we can conclude that S is open.

2. There exists $x' > x$ such that $g(x') > g(x)$, and $g(x) \geq g(x-)$ and $g(x) \geq g(x+)$: Define $g_{x,-}$ and $g_{x,+}$ as above. Then, since $g(x') > g(x) \geq g(x-)$ and $g(x') > g(x) \geq g(x+)$, we can argue as in the previous case that there exists $\epsilon \in \mathbb{R}_{>0}$ such that $g(x') > g(y)$ for all $y \in (x - \epsilon, x + \epsilon)$. Thus S is open.

3. There exists $x' > x$ such that $g(x') > g(x+)$, and $g(x+) \geq g(x)$ and $g(x+) \geq g(x-)$: Here we can argue in a manner entirely similar to the first case that S is open.

The preceding arguments show that S is open, and so by Proposition 2.5.6 it is a countable union of open intervals. ▼

Now define

$$\Lambda_l(x) = \limsup_{h \downarrow 0} \frac{f(x-h) - f(x)}{-h} \qquad \lambda_l(x) = \liminf_{h \downarrow 0} \frac{f(x-h) - f(x)}{-h}$$
$$\Lambda_r(x) = \limsup_{h \downarrow 0} \frac{f(x+h) - f(x)}{h} \qquad \lambda_r(x) = \liminf_{h \downarrow 0} \frac{f(x+h) - f(x)}{h}.$$

If f is differentiable at x then these four numbers will be finite and equal. We shall show that

1. $\Lambda_r(x) < \infty$ and
2. $\Lambda_r(x) \leq \lambda_l(x)$

for almost every $x \in [a, b]$. Since the relations

$$\lambda_l \leq \Lambda_l \leq \lambda_r \leq \Lambda_r$$

hold due to monotonicity of f, the differentiability of f for almost all x will then follow.

For 1, if $M \in \mathbb{R}_{>0}$ denote

$$S_M = \{x \in [a, b] \mid \Lambda_r(x) > M\}.$$

Thus, for $x_0 \in S_M$, there exists $x > x_0$ such that

$$\frac{f(x) - f(x_0)}{x - x_0} > M.$$

Defining $g_M(x) = f(x) - Mx$ this asserts that $g_M(x) > g_M(x_0)$. The function g_M satisfies the hypotheses of Lemma 1 by part (i). This means that S_M is contained in a finite or countable disjoint union of intervals $\{I_\alpha \mid \alpha \in A\}$, open in $[a, b]$, for which

$$g_M(a_\alpha) \leq \max\{g_M(b_\alpha-), g_M(b_\alpha), g_M(b_\alpha+)\}, \qquad \alpha \in A,$$

where a_α and b_α are the left and right endpoints, respectively, for I_α, $\alpha \in A$. In particular, $g_M(a_\alpha) \leq g_M(b_\alpha)$. A trivial manipulation then gives

$$M(b_\alpha - a_\alpha) \leq f(b_\alpha) - f(a_\alpha), \qquad \alpha \in A.$$

We have

$$M \sum_{\alpha \in A} |b_\alpha - a_\alpha| \leq \sum_{\alpha \in A} |f(b_\alpha) - f(a_\alpha)| \leq f(b) - f(a)$$

since f is monotonically increasing. Since f is bounded, this shows that as $M \to \infty$ the length of the open intervals $\{(a_\alpha, b_\alpha) \mid \alpha \in A\}$ covering S_M must go to zero. This shows that the set of points where 1 holds has zero measure.

Now we turn to 2. Let $0 < m < M$, define $g_m(x) = -f(x) + mx$ and $g_M(x) = f(x) - Mx$. Also define

$$S_m = \{x \in [a, b] \mid \lambda_l(x) < m\}.$$

For $x_0 \in S_m$ there exists $x < x_0$ such that

$$\frac{f(x) - f(x_0)}{x - x_0} < m,$$

which is equivalent to $g_m(x) > g_m(x_0)$. Therefore, by Lemma 1, note that S_m is contained in a finite or countable disjoint union of intervals $\{I_\alpha \mid \alpha \in A\}$, open in $[a, b]$. Denote by a_α and b_α the left and right endpoints, respectively, for I_α for $\alpha \in A$. For $\alpha \in A$ denote

$$S_{\alpha,M} = \{x \in [a_\alpha, b_\alpha] \mid \Lambda_r(x) > M\},$$

and arguing as we did in the proof that 1 holds almost everywhere, denote by $\{I_{\alpha,\beta} \mid \beta \in B_\alpha\}$ the countable collection of subintervals, open in $[a, b]$, of (a_α, b_α) that contain $S_{\alpha,M}$. Denote by $a_{\alpha,\beta}$ and $b_{\alpha,\beta}$ the left and right endpoints, respectively, of $I_{\alpha,\beta}$ for $\alpha \in A$ and $\beta \in B_\alpha$. Note that the relations

$$\begin{aligned} &g_m(a_\alpha) \le \max\{g_m(b_\alpha-), g_m(b_\alpha), g_m(b_\alpha+)\}, \qquad \alpha \in A, \\ &g_M(a_{\alpha,\beta}) \le \max\{g_M(b_{\alpha,\beta}-), g_M(b_{\alpha,\beta}), g_M(b_{\alpha,\beta}+)\}, \qquad \alpha \in A,\ \beta \in B_\alpha \end{aligned}$$

hold. We then may easily compute

$$\begin{aligned} &f(b_\alpha) - f(a_\alpha) \le m(b_\alpha - a_\alpha), \qquad \alpha \in A, \\ &f(b_{\alpha,\beta}) - f(a_{\alpha,\beta}) \ge M(b_{\alpha,\beta} - b_{\alpha,\beta}), \qquad \alpha \in A,\ \beta \in A_\alpha. \end{aligned}$$

Therefore, for each $\alpha \in A$,

$$M \sum_{\beta \in A_\alpha} |b_{\alpha,\beta} - a_{\alpha,\beta}| \le \sum_{\beta \in A_\alpha} |f(b_{\alpha,\beta} - a_{\alpha,\beta})| \le f(b_\alpha) - f(a_\alpha) \le m(b_\alpha - a_\alpha).$$

This then gives

$$M \sum_{\alpha \in A} \sum_{\beta \in A_\alpha} |b_{\alpha,\beta} - a_{\alpha,\beta}| \le m \sum_{\alpha \in A} |b_\alpha - a_\alpha|,$$

or $\Sigma_2 \le \frac{m}{M} \Sigma_1$, where

$$\Sigma_1 = \sum_{\alpha \in A} \sum_{\beta_\alpha \in K_\alpha} |b_{\alpha,\beta} - a_{\alpha,\beta}|, \quad \Sigma_2 = \sum_{\alpha \in A} |b_\alpha - a_\alpha|.$$

Now, this process can be repeated, defining

$$S_{\alpha,\beta,m} = \{x \in [a_{\alpha,\beta}, b_{\alpha,\beta}] \mid \lambda_l(x) < m\},$$

and so on. We then generate a sequence of finite or countable disjoint intervals of total length Σ_α and satisfying

$$\Sigma_{2\alpha} \le \frac{m}{M} \Sigma_{2\alpha-1} \le \Big(\frac{m}{M}\Big)^\alpha \Sigma_1, \qquad \alpha \in A.$$

It therefore follows that $\lim_{\alpha \to \infty} \Sigma_\alpha = 0$. Thus the set of points

$$S_{M,m} = \{x \in [a, b] \mid m < \lambda_l(x) \text{ and } \Lambda_r(x) > M\}$$

is contained in a set of zero measure provided that $m < M$. Now note that

$$\{x \in [a,b] \mid \lambda_l(x) \geq \Lambda_r(x)\} \subseteq \bigcup \{S_{M,m} \mid m, M \in \mathbb{Q},\ m < M\}.$$

The union on the left is a countable union of sets of zero measure, and so has zero measure itself (by Exercise 2.5.9). This shows that f is differentiable on a set whose complement has zero measure.

To show that $f'(x) \geq 0$ for all points x at which f is differentiable, suppose the converse. Thus suppose that there exists $x \in [a,b]$ such that $f'(x) < 0$. This means that for ϵ sufficiently small and positive,

$$\frac{f(x+\epsilon) - f(x)}{\epsilon} < 0 \quad \implies \quad f(x+\epsilon) - f(x) < 0,$$

which contradicts the fact that f is monotonically increasing. This completes the proof of the theorem. ■

Let us give two examples of functions that illustrate the surprisingly strange behaviour that can arise from monotonic functions. These functions are admittedly degenerate, and not something one is likely to encounter in applications. However, they do show that one cannot strengthen the conclusions of Theorem 3.2.26.

Our first example is one of the standard "peculiar" monotonic functions, and its construction relies on the middle-thirds Cantor set constructed in Example 2.5.39.

3.2.27 Example (A continuous increasing function with an almost everywhere zero derivative) Let C_k, $k \in \mathbb{Z}_{>0}$, be the sets, comprised of collections of disjoint closed intervals, used in the construction of the middle-thirds Cantor set of Example 2.5.39. Note that, for $x \in [0,1]$, the set $[0,x] \cap C_k$ consists of a finite number of intervals. Let $g_k \colon [0,1] \to [0,1]$ be defined by asking that $g_{C,k}(x)$ be the sum of the lengths of the intervals comprising $[0,x] \cap C_k$. Then define $f_{C,k} \colon [0,1] \to [0,1]$ by $f_{C,k}(x) = \left(\frac{3}{2}\right)^k g_{C,k}(x)$. Thus $f_{C,k}$ is a function that is constant on the complement to the closed intervals comprising C_k, and is linear on those same closed intervals, with a slope determined in such a way that the function is continuous. We then define $f_C \colon [0,1] \to [0,1]$ by $f_C(x) = \lim_{k\to\infty} f_{C,k}(x)$. In Figure 3.9 we depict f_C. The reader new to this function should take the requisite moment or two to understand our definition of f_C, perhaps by sketching a couple of the functions $f_{C,k}$, $k \in \mathbb{Z}_{>0}$.

Let us record some properties of the function f_C, which is called the ***Cantor function*** or the ***Devil's staircase***.

1 Lemma f_C *is continuous.*

Proof We prove this by showing that the sequence of functions $(f_{C,k})_{k\in\mathbb{Z}_{>0}}$ converges uniformly, and then using Theorem 3.4.8 to conclude that the limit function is continuous. Note that the functions $f_{C,k}$ and $f_{C,k+1}$ differ only on the closed intervals comprising C_k. Moreover, if $J_{k,j}$, $k \in \mathbb{Z}_{\geq 0}$, $j \in \{1, \ldots, 2^k - 1\}$, denotes the set of open intervals forming $[0,1] \setminus C_k$, numbered from left to right, then the value of $f_{C,k}$ on $J_{k,j}$ is $j2^{-k}$. Therefore,

$$\sup\{|f_{C,k+1}(x) - f_{C,k}(x)| \mid x \in [0,1]\} < 2^{-k}, \qquad k \in \mathbb{Z}_{\geq 0}.$$

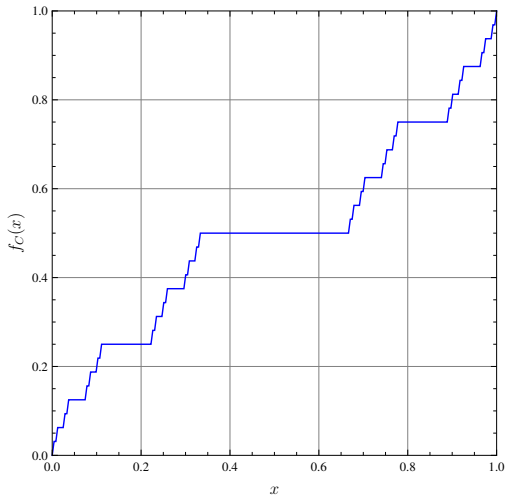


Figure 3.9 A depiction of the Cantor function

This implies that $(f_{C,k})_{k\in\mathbb{Z}_{>0}}$ is uniformly convergent as in Definition 3.4.4. Thus Theorem 3.4.8 gives continuity of f_C, as desired. ▼

2 Lemma f_C *is differentiable at all points in* $[0,1] \setminus C$*, and its derivative, where it exists, is zero.*

Proof Since C is constructed as an intersection of the closed sets C_k, and since such intersections are themselves closed by Exercise 2.5.1, it follows that $[0,1] \setminus C$ is open. Thus if $x \in [0,1] \setminus C$, there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}(\epsilon, x) \subseteq [0,1] \setminus C$. Since $\mathsf{B}(\epsilon, x)$ contains no endpoints for intervals from the sets C_k, $k \in \mathbb{Z}_{>0}$, it follows that $f_{C,k}|\mathsf{B}(\epsilon, x)$ is constant for sufficiently large k. Therefore $f_C|\mathsf{B}(\epsilon, x)$ is constant, and it then follows that f_C is differentiable at x, and that $f_C'(x) = 0$. ▼

In Example 2.5.39 we showed that C has measure zero. Thus we have a continuous, monotonically increasing function from $[0,1]$ to $[0,1]$ whose derivative is almost everywhere zero. It is perhaps not *a priori* obvious that such a function can exist, since one's first thought might be that zero derivative implies a constant function. The reasons for the failure of this rule of thumb in this example will not become perfectly clear until we examine the notion of absolute continuity in Section **??**. •

The second example of a "peculiar" monotonic function is not quite as standard in the literature, but is nonetheless interesting since it exhibits somewhat different oddities than the Cantor function.

3.2.28 Example (A strictly increasing function, discontinuous on the rationals, with an almost everywhere zero derivative) We define a strictly monotonically increasing function $f_{\mathbb{Q}} \colon \mathbb{R} \to \mathbb{R}$ as follows. Let $(q_j)_{j\in\mathbb{Z}_{>0}}$ be an enumeration of the rational numbers and for $x \in \mathbb{R}$ define

$$I(x) = \{j \in \mathbb{Z}_{>0} \mid q_j < x\}.$$

Now define

$$f_{\mathbb{Q}}(x) = \sum_{j \in I(x)} \frac{1}{2^j}.$$

Let us record the properties of $f_{\mathbb{Q}}$ in a series of lemmata.

1 Lemma $\lim_{x \to -\infty} f_{\mathbb{Q}}(x) = 0$ *and* $\lim_{x \to \infty} f_{\mathbb{Q}}(x) = 1$.

Proof Recall from Example 2.4.2–1 that $\sum_{j=1}^{\infty} \frac{1}{2^j} = 1$. Let $\epsilon \in \mathbb{R}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $\sum_{j=N+1}^{\infty} \frac{1}{2^j} < \epsilon$. Now choose $M \in \mathbb{R}_{>0}$ such that $\{q_1, \dots, q_N\} \subseteq [-M, M]$. Then, for $x < M$ we have

$$f_{\mathbb{Q}}(x) = \sum_{j \in I(x)} \frac{1}{2^j} = \sum_{j=1}^{\infty} \frac{1}{2^j} - \sum_{j \in \mathbb{Z}_{>0} \setminus I(x)} \frac{1}{2^j} \leq \sum_{j=1}^{\infty} \frac{1}{2^j} - \sum_{j=1}^{N} \frac{1}{2^j} < \epsilon.$$

Also, for $x > M$ we have

$$f_{\mathbb{Q}}(x) = \sum_{j \in I(x)} \frac{1}{2^j} \geq \sum_{j=1}^{N} \frac{1}{2^j} > 1 - \epsilon.$$

Thus $\lim_{x \to -\infty} f_{\mathbb{Q}}(x) = 0$ and $\lim_{x \to \infty} f_{\mathbb{Q}}(x) = 1$. ▼

2 Lemma $f_{\mathbb{Q}}$ *is strictly monotonically increasing.*

Proof Let $x, y \in \mathbb{R}$ with $x < y$. Then, by Corollary 2.2.16, there exists $q \in \mathbb{Q}$ such that $x < q < y$. Let $j_0 \in \mathbb{Z}_{>0}$ have the property that $q = q_{j_0}$. Then

$$f_{\mathbb{Q}}(y) = \sum_{j \in I(y)} \frac{1}{2^j} \geq \sum_{j \in I(x)} \frac{1}{2^j} + \frac{1}{2^{j_0}} > f_{\mathbb{Q}}(x),$$

as desired. ▼

3 Lemma $f_{\mathbb{Q}}$ *is discontinuous at each point in* $\mathbb{Q}$.

Proof Let $q \in \mathbb{Q}$ and let $x > q$. Let $j_0 \in \mathbb{Z}_{>0}$ satisfy $q = q_{j_0}$. Then

$$f_{\mathbb{Q}}(x) = \sum_{j \in I(x)} \frac{1}{2^j} \geq \frac{1}{2^{j_0}} + \sum_{j \in I(q)} \frac{1}{2^j} = \frac{1}{2^{j_0}} + \sum_{j \in I(q)} \frac{1}{2^j}.$$

Therefore, $\lim_{x \downarrow q} f_{\mathbb{Q}}(x) \geq \frac{1}{2^{j_0}} + f_{\mathbb{Q}}(q)$, implying that $f_{\mathbb{Q}}$ is discontinuous at q by Theorem 3.1.3. ▼

4 Lemma $\mathrm{f}_{\mathbb{Q}}$ *is continuous at each point in* $\mathbb{R} \setminus \mathbb{Q}$.

Proof Let $x \in \mathbb{R} \setminus \mathbb{Q}$ and let $\epsilon \in \mathbb{R}_{>0}$. Take $N \in \mathbb{Z}_{>0}$ such that $\sum_{j=N+1}^{\infty} \frac{1}{2^j} < \epsilon$ and define $\delta \in \mathbb{R}_{>0}$ such that $\mathsf{B}(\delta, x) \cap \{q_1, \ldots, q_N\} = \emptyset$ (why is this possible?). Now let

$$I(\delta, x) = \{j \in \mathbb{Z}_{>0} \mid q_j \in \mathsf{B}(\delta, x)\}$$

and note that, for $y \in \mathsf{B}(\delta, x)$ with $x < y$, we have

$$\begin{aligned} f_{\mathbb{Q}}(y) - f_{\mathbb{Q}}(x) &= \sum_{j \in I(y)} \frac{1}{2^j} - \sum_{j \in I(x)} \frac{1}{2^j} \leq \sum_{j \in I(\delta, x)} \frac{1}{2^j} = \sum_{j=1}^{\infty} \frac{1}{2^j} - \sum_{\mathbb{Z}_{>0} \setminus I(\delta, x)} \frac{1}{2^j} \\ &\leq \sum_{j=1}^{\infty} \frac{1}{2^j} - \sum_{j=1}^{N} \frac{1}{2^j} = \sum_{j=N+1}^{\infty} \frac{1}{2^j} < \epsilon. \end{aligned}$$

A similar argument holds for $y < x$ giving $f_{\mathbb{Q}}(x) - f_{\mathbb{Q}}(y) < \epsilon$ in this case. Thus $|f_{\mathbb{Q}}(y) - f_{\mathbb{Q}}(x)| < \epsilon$ for $|y - x| < \delta$, thus showing continuity of f at x. ▼

5 Lemma *The set* $\{\mathrm{x} \in \mathbb{R} \mid \mathrm{f}'_{\mathbb{Q}}(\mathrm{x}) \neq 0\}$ *has measure zero.*

Proof The proof relies on some concepts from Section 3.4. For $k \in \mathbb{Z}_{>0}$ define $f_{\mathbb{Q},k} \colon \mathbb{R} \to \mathbb{R}$ by

$$f_{\mathbb{Q},k}(x) = \sum_{j \in I(x) \cap \{1, \ldots, k\}} \frac{1}{2^j}.$$

Note that $(f_{\mathbb{Q},k})_{k \in \mathbb{Z}_{>0}}$ is a sequence of monotonically increasing functions with the following properties:

1. $\lim_{k \to \infty} f_{\mathbb{Q},k}(x) = f_{\mathbb{Q}}(x)$ for each $x \in \mathbb{R}$;
2. the set $\{x \in \mathbb{R} \mid f'_{\mathbb{Q},k}(x) \neq 0\}$ is finite for each $k \in \mathbb{Q}$.

The result now follows from Theorem 3.4.25. ▼

Thus we have an example of a strictly monotonically increasing function whose derivative is zero almost everywhere. Note that this function also has the feature that in any neighbourhood of a point where it is differentiable, there lie points where it is not differentiable. This is an altogether peculiar function. •

3.2.6 Convex functions and differentiability

Let us now return to our consideration of convex functions introduced in Section 3.1.6. Here we discuss the differentiability properties of convex functions. The following notation for a function $f \colon I \to \mathbb{R}$ will be convenient:

$$f'(x+) = \lim_{\epsilon \downarrow 0} \frac{f(x + \epsilon) - f(x)}{\epsilon}, \quad f'(x-) = \lim_{\epsilon \downarrow 0} \frac{f(x) - f(x - \epsilon)}{\epsilon},$$

provided that these limits exist.

With this notation, convex functions have the following properties.

3.2.29 Proposition (Properties of convex functions II) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$ *and for a convex function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$*, the following statements hold:*

(i) if I *is open then the limits* $\mathrm{f}'(\mathrm{x}+)$ *and* $\mathrm{f}'(\mathrm{x}-)$ *exist and* $\mathrm{f}'(\mathrm{x}-) \leq \mathrm{f}'(\mathrm{x}+)$ *for each* $\mathrm{x} \in \mathrm{I}$*;*

(ii) if I *is open then the functions*

$$\mathrm{I} \ni \mathrm{x} \mapsto \mathrm{f}'(\mathrm{x}+), \quad \mathrm{I} \ni \mathrm{x} \mapsto \mathrm{f}'(\mathrm{x}-)$$

are monotonically increasing, and strictly monotonically increasing if f *is strictly convex;*

(iii) if I *is open and if* $\mathrm{x}_1, \mathrm{x}_2 \in \mathrm{I}$ *satisfy* $\mathrm{x}_1 < \mathrm{x}_2$*, then* $\mathrm{f}'(\mathrm{x}_1+) \leq \mathrm{f}'(\mathrm{x}_2-)$*;*

(iv) f *is differentiable except at a countable number of points in* I*.*

Proof (i) Since I is open there exists $\epsilon_0 \in \mathbb{R}_{>0}$ such that $[x, x+\epsilon_0) \subseteq I$. Let $(\epsilon_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $(0, \epsilon_0)$ converging to 0 and such that $\epsilon_{j+1} < \epsilon_j$ for every $j \in \mathbb{Z}_{>0}$. Then the sequence $(s_f(x, x+\epsilon_j))_{j\in\mathbb{Z}_{>0}}$ is monotonically decreasing. This means that, by Lemma 3.1.33,

$$\frac{f(x+\epsilon_{j+1}) - f(x)}{\epsilon_{j+1}} \leq \frac{f(x+\epsilon_j) - f(x)}{\epsilon_j}$$

for each $j \in \mathbb{Z}_{>0}$. Moreover, if $x' \in I$ satisfies $x' < x$ then we have $s_f(x', x) \leq s_f(x, x+\epsilon_j)$ for each $j \in \mathbb{Z}_{>0}$. Thus the sequence $(\epsilon_j^{-1}(f(x+\epsilon_j) - f(x)))_{j\in\mathbb{Z}_{>0}}$ is decreasing and bounded from below. Thus it must converge, cf. Theorem 2.3.8.

The proof for the existence of the other asserted limit follows that above, *mutatis mutandis*.

To show that $f'(x-) \leq f'(x+)$, note that, for all ϵ sufficiently small,

$$\frac{f(x) - f(x-\epsilon)}{\epsilon} = s_f(x-\epsilon, x) \leq s_f(x, x+\epsilon) = \frac{f(x+\epsilon) - f(x)}{\epsilon}.$$

Taking limits as $\epsilon \downarrow 0$ gives the desired inequality.

(ii) For $x_1, x_2 \in I$ with $x_1 < x_2$ we have

$$f'(x_1+) = \lim_{\epsilon\downarrow 0} s_f(x_1, x_1+\epsilon) \leq \lim_{\epsilon\downarrow 0} s_f(x_2, x_2+\epsilon) = f'(x_2+),$$

using Lemma 3.1.33. A similar computation, *mutatis mutandis*, shows that the other function in this part of the result is also monotonically increasing. Moreover, if f is strictly convex that the inequalities above can be replaced with strict inequalities by (3.2). From this we conclude that $x \mapsto f'(x_+)$ and $x \mapsto f'(x_-)$ are strictly monotonically increasing.

(iii) For $\epsilon \in \mathbb{R}_{>0}$ sufficiently small we have

$$x_1 + \epsilon < x_2 - \epsilon.$$

For all such sufficiently small ϵ we have

$$\frac{f(x_1+\epsilon) - f(x_1)}{\epsilon} = s_f(x_1, x_1+\epsilon) \leq s_f(x_2-\epsilon, x_2) = \frac{f(x_2) - f(x_2-\epsilon)}{\epsilon}$$

by Lemma 3.1.33. Taking limits as $\epsilon \downarrow 0$ gives this part of the result.

(iv) Let A_f be the set of points in I where f is not differentiable. Note that

$$\frac{f(x) - f(x-\epsilon)}{\epsilon} = s_f(x-\epsilon, x) \leq s_f(x, x+\epsilon) = \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

by Lemma 3.1.33. Therefore, if $x \in A_f$, then $f'(x-) < f'(x+)$. We define a map $\phi\colon A_f \to \mathbb{Q}$ as follows. If $x \in A_f$ we use the Axiom of Choice and Corollary 2.2.16 to select $\phi(x) \in \mathbb{Q}$ such that $f'(x-) < \phi(x) < f'(x+)$. We claim that ϕ is injective. Indeed, if $x, y \in A_f$ are distinct (say $x < y$) then, using parts (ii) and (iii),

$$f'(x-) < \phi(x) < f'(x+) < f'(y-) < \phi(y) < f'(y+).$$

Thus $\phi(x) < \phi(y)$ and so ϕ is injective as desired. Thus A_f must be countable. ∎

For functions that are sufficiently differentiable, it is possible to conclude convexity from properties of the derivative.

3.2.30 Proposition (Convexity and derivatives) *For an interval* $I \subseteq \mathbb{R}$ *and for a function* $f\colon I \to \mathbb{R}$ *the following statements hold:*

(i) for each $x_1, x_2 \in I$ *with* $x_1 \neq x_2$ *we have*

$$f(x_2) \geq f(x_1) + f'(x_1+)(x_2 - x_1), \quad f(x_2) \geq f(x_1) + f'(x_1-)(x_2 - x_1);$$

(ii) if f *is differentiable, then* f *is convex if and only if* f' *is monotonically increasing;*

(iii) if f *is differentiable, then* f *is strictly convex if and only if* f' *is strictly monotonically increasing;*

(iv) if f *is twice continuously differentiable, then it is convex if and only if* $f''(x) \geq 0$ *for every* $x \in I$*;*

(v) if f *is twice continuously differentiable, then it is strictly convex if and only if* $f''(x) > 0$ *for every* $x \in I$.

Proof (i) Suppose that $x_1 < x_2$. Then, for $\epsilon \in \mathbb{R}_{>0}$ sufficiently small,

$$\frac{f(x_1+\epsilon) - f(x_1)}{\epsilon} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

by Lemma 3.1.33. Thus, taking limits as $\epsilon \downarrow 0$,

$$f'(x_1+) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1},$$

and rearranging gives

$$f(x_2) \geq f(x_1) + f'(x_1+)(x_2 - x_1).$$

Since we also have $f'(x_1-) \leq f'(x_1+)$ by Proposition 3.2.29(i), we have both of the desired inequalities in this case.

Now suppose that $x_2 < x_1$. Again, for $\epsilon \in \mathbb{R}_{>0}$ sufficiently small, we have

$$\frac{f(x_1+\epsilon) - f(x_1)}{\epsilon} \geq \frac{f(x_1) - f(x_2)}{x_1 - x_2},$$

and taking the limit as $\epsilon \downarrow 0$ gives

$$f'(x_1+) \geq \frac{f(x_1) - f(x_2)}{x_1 - x_2}.$$

Rearranging gives

$$f(x_2) \geq f(x_1) + f'(x_1+)(x_2 - x_1)$$

and since $f'(x_1-) \leq f'(x_1+)$ the desired inequalities follow in this case.

(ii) From Proposition 3.2.29(ii) we deduce that if f is convex and differentiable then f' is monotonically increasing. Conversely, suppose that f is differentiable and that f' is monotonically increasing. Let $x_1, x_2 \in I$ satisfy $x_1 < x_2$ and let $s \in (0,1)$. By the Mean Value Theorem there exists $c_1, c_2 \in I$ satisfying

$$x_1 < c_1 < (1-s)x_1 + sx_2 < d_1 < x_2$$

such that

$$\frac{f((1-s)x_1 + sx_2) - f(x_1)}{(1-s)x_1 + sx_2 - x_1} = f'(c_1) \leq f'(c_2) = \frac{f(x_2) - f((1-s)x_1 + sx_2)}{x_2 - ((1-s)x_1 + sx_2)}. \tag{3.9}$$

Rearranging, we get

$$\frac{f((1-s)x_1 + sx_2) - f(x_1)}{s(x_2 - x_1)} \leq \frac{f(x_2) - f((1-s)x_1 + sx_2)}{(1-s)(x_2 - x_1)},$$

and further rearranging gives

$$f((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2),$$

and so f is convex.

(iii) If f is strictly convex, then from Proposition 3.2.29 we conclude that f' is strictly monotonically increasing. Next suppose that f' is strictly monotonically decreasing and let $x_1, x_2 \in I$ satisfy $x_1 < x_2$ and let $s \in (0,1)$. The proof that f is strictly convex follows as in the preceding part of the proof, noting that, in (3.9), we have $f'(c_1) < f'(c_2)$. Carrying this strict inequality through the remaining computations shows that

$$f((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2),$$

giving strict convexity of f.

(iv) If f'' is nonnegative, then f' is monotonically increasing by Proposition 3.2.23. The result now follows from part (ii).

(iv) If f'' is positive, then f' is strictly monotonically increasing by Proposition 3.2.23. The result now follows from part (iii). ■

Let us consider a few examples illustrating how convexity and differentiability are related.

3.2.31 Examples (Convex functions and differentiability)

1. The convex function $n_{x_0}\colon \mathbb{R} \to \mathbb{R}$ defined by $n_{x_0}(x) = |x - x_0|$ is differentiable everywhere except for $x = x_0$. But at $x = x_0$ the derivatives from the left and right exist. Moreover, $f'(x) = -1$ for $x < x_0$ and $f'(x) = 1$ for $x > x_0$. Thus we see that the derivative is monotonically increasing, although it is not defined everywhere.
2. As we showed in Proposition 3.2.29(iv), a convex function is differentiable except at a countable set of points. Let us show that this conclusion cannot be improved. Let $C \subseteq \mathbb{R}$ be a countable set. We shall construct a convex function $f\colon \mathbb{R} \to \mathbb{R}$ whose derivative exists on $\mathbb{R} \setminus C$ and does not exist on C. In case C is finite, we write $C = \{x_1, \dots, x_k\}$. Then one verifies that the function f defined by

$$f(x) = \sum_{j=1}^{k} |x - x_j|$$

is verified to be convex, being a finite sum of convex functions (see Proposition 3.1.39). It is clear that f is differentiable at points in $\mathbb{R} \setminus C$ and is not differentiable at points in C. Now suppose that C is not finite. Let us write $C = \{x_j\}_{j \in \mathbb{Z}_{>0}}$, i.e., enumerate the points in C. Let us define $c_j = (2^j \max\{1, |x_j|\})^{-1}$, $j \in \mathbb{Z}_{>0}$, and define $f\colon \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \sum_{j=1}^{\infty} c_j |x - x_j|.$$

We shall prove that this function is well-defined, convex, differentiable at points in $\mathbb{R} \setminus C$, and not differentiable at points in C. In proving this, we shall make reference to some results we have not yet proved.
First let us show that f is well-defined.

1 Lemma *For every compact subset* $\mathrm{K} \subseteq \mathbb{R}$*, the series*

$$\sum_{\mathrm{j}=1}^{\infty} \mathrm{c_j} |\mathrm{x} - \mathrm{x_j}|$$

converges uniformly on K *(see Section 3.4.2 for uniform convergence).*

Proof Let $K \subseteq \mathbb{R}$ and let $R \in \mathbb{R}_{>0}$ be large enough that $K \subseteq [-R, R]$. Then, for $x \in K$ we have

$$|c_j|x - x_j|| \leq c_j(|x| + |x_j|) \leq \frac{R+1}{2^j}.$$

By the Weierstrass M-test (Theorem 3.4.15 below) and Example 2.4.2–1 the lemma follows. ▼

It follows immediately from the lemma that the series defining f converges pointwise, and so f is well-defined, and is moreover convex by Theorem 3.4.26.
Now we show that f is differentiable at points in $\mathbb{R} \setminus C$.

2 Lemma *The function* f *is differentiable at every point in* $\mathbb{R} \setminus C$.

Proof Let us denote $g_j(x) = c_j|x - x_j|$. Let $x_0 \in \mathbb{R} \setminus C$ and define, for each $j \in \mathbb{Z}_{>0}$,

$$h_{j,x_0} = \begin{cases} \frac{g_j(x)-g_j(x_0)}{x-x_0}, & x \neq x_0, \\ g_j'(x_0), & x = x_0, \end{cases}$$

noting that the functions g_j, $j \in \mathbb{Z}_{>0}$, are differentiable at points in $\mathbb{R} \setminus C$.
Let $j \in \mathbb{Z}$. We claim that if $x_0 \neq x_j$ then

$$|h_{j,x_0}(x)| \leq \frac{3}{2^j} \tag{3.10}$$

for all $x \in \mathbb{R}$. We consider three cases.

(a) $x = x_0$: Note that g_j is differentiable at $x = x_0$ and that $|g_j'(x_0)| = c_j \leq \frac{1}{2^j} < \frac{3}{2^j}$. Thus the estimate (3.10) holds when $x = x_0$.

(b) $x \neq x_0$ and $(x - x_j)(x_0 - x_j) > 0$: We have

$$|h_{j,x_0}(x)| = c_j \left| \frac{(x - x_j) - (x_0 - x_j)}{x - x_0} \right| = a_j \leq \frac{1}{2^j} < \frac{3}{2^j},$$

giving (3.10) in this case.

(c) $x \neq x_0$ and $(x - x_j)(x_0 - x_j) < 0$: We have

$$|h_{j,x_0}(x)| = c_j \left| \frac{(x - x_j) - (x_j - x_0)}{x - x_0} \right| = c_j \left| 1 + \frac{2(x_0 - x_j)}{x_0 - x} \right| \leq \frac{1}{2^j} \left| 1 + \frac{2(x_0 - x_j)}{x_0 - x} \right|.$$

Since $(x - x_j)$ and $x_0 - x_j$ have opposite sign, this implies that either (1) $x < x_j$ and $x_0 > x_j$ or (2) $x > x_j$ and $x_0 < x_j$. In either case, $|x_0 - x_j| < |x_0 - x|$. This, combined with our estimate above, gives (3.10) in this case.

Now, given (3.10), we can use the Weierstrass M-test (Theorem 3.4.15 below) and Example 2.4.2–1 to conclude that $\sum_{j=1}^{\infty} h_{j,x_0}$ converges uniformly on $\mathbb{R}$ for each $x_0 \in \mathbb{R} \setminus C$.
Now we prove that f is differentiable at $x_0 \in \mathbb{R} \setminus C$. If $x \neq x_0$ then the definition of the functions h_{j,x_0}, $j \in \mathbb{Z}_{>0}$, gives

$$\frac{f(x) - f(x_0)}{x - x_0} = \sum_{j=1}^{\infty} h_{j,x_0}(x),$$

the latter sum making sense since we have shown that it converges uniformly. Moreover, since the functions g_j, $j \in \mathbb{Z}_{>0}$, are differentiable at x_0, it follows that, for each $j \in \mathbb{Z}_{>0}$,

$$\lim_{x \to x_0} h_{j,x_0}(x) = \lim_{x \to x_0} \frac{g_j(x) - g_j(x_0)}{x - x_0} = g_j'(x_0) = h_{j,x_0}(x_0).$$

That is, h_{j,x_0} is continuous at x_0. It is clear that h_{j,x_0} is continuous at all $x \neq x_0$. Thus, since $\sum_{j=1}^{\infty} h_{j,x_0}$ converges uniformly, the limit function is continuous by Theorem 3.4.8. Thus we have

$$\lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x\to x_0} \sum_{j=1}^{\infty} h_{j,x_0}(x) = \sum_{j=1}^{\infty} h_{j,x_0}(x_0) = \sum_{j=1}^{\infty} g_j'(x_0).$$

This gives the desired differentiability since the last series converges. ▼

Finally, we show that f is not differentiable at points in C.

3 Lemma *The function* f *is not differentiable at every point in* C.

Proof For $k \in \mathbb{Z}_{>0}$, let us write

$$f(x) = g_k(x) + \underbrace{\sum_{\substack{j=1\\ j\neq k}} g_j(x)}_{f_j(x)}.$$

The arguments from the proof of the preceding lemma can be applied to show that the function f_j defined by the sum on the right is differentiable at x_k. Since g_k is not differentiable at x_k, we conclude that f cannot be differentiable at x_k by Proposition 3.2.10. ▼

This shows that the conclusions of Proposition 3.2.29(iv) cannot generally be improved. •

3.2.7 Piecewise differentiable functions

In Section 3.1.7 we considered functions that were piecewise continuous. In this section we consider a class of piecewise continuous functions that have additional properties concerning their differentiability. We let $I \subseteq \mathbb{R}$ be an interval with $f\colon I \to \mathbb{R}$ a function. In Section 3.1.7 we defined the notation $f(x-)$ and $f(x+)$. Here we also define

$$f'(x-) = \lim_{\epsilon\downarrow 0} \frac{f(x-\epsilon) - f(x-)}{-\epsilon}, \quad f'(x+) = \lim_{\epsilon\downarrow 0} \frac{f(x+\epsilon) - f(x+)}{\epsilon}.$$

These limits, of course, may fail to exist, or even to make sense if $x \in \mathrm{bd}(I)$.

Now, recalling the notion of a partition from Definition 2.5.7, we make the following definition.

3.2.32 Definition (Piecewise differentiable function) A function $f\colon [a,b] \to \mathbb{R}$ is ***piecewise differentiable*** if there exists a partition $P = (I_1, \ldots, I_k)$, with $\mathrm{EP}(P) = (x_0, x_1, \ldots, x_k)$, of $[a,b]$ with the following properties:

(i) $f|\operatorname{int}(I_j)$ is differentiable for each $j \in \{1, \ldots, k\}$;

(ii) for $j \in \{1, \ldots, k-1\}$, the limits $f(x_j+)$, $f(x_j-)$, $f'(x_j+)$, and $f'(x_j-)$ exist;

(iii) the limits $f(a+)$, $f(b-)$, $f'(a+)$, and $f'(b-)$ exist. •

It is evident that a piecewise differentiable function is piecewise continuous. It is not surprising that the converse is not true, and a simple example of this will be given in the following collection of examples.

3.2.33 Examples (Piecewise differentiable functions)

1. Let $I = [-1, 1]$ and define $f : I \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1 + x, & x \in [-1, 0], \\ 1 - x, & (0, 1]. \end{cases}$$

One verifies that f is differentiable on $(-1, 0)$ and $(0, 1)$. Moreover, we compute the limits

$$f(-1+) = 0, \quad f'(-1+) = 1, \quad f(1-) = 0, \quad f'(1-) = -1,$$
$$f(0-) = 1, \quad f(0+) = 1, \quad f'(0-) = 1, \quad f'(0+) = -1.$$

Thus f is piecewise differentiable. Note that f is also continuous.

2. Let $I = [-1, 1]$ and define $f : I \to \mathbb{R}$ by $f(x) = \operatorname{sign}(x)$. On $(-1, 0)$ and $(0, 1)$ we note that f is differentiable. Moreover, we compute

$$f(-1+) = -1, \quad f'(-1+) = 0, \quad f(1-) = 1, \quad f'(1-) = 0,$$
$$f(0-) = -1, \quad f(0+) = 1, \quad f'(0-) = 0, \quad f'(0+) = 0.$$

Note that it is important here to *not* compute the limits $f'(0-)$ and $f'(0+)$ using the formulae

$$\lim_{\epsilon \downarrow 0} \frac{f(0 - \epsilon) - f(0)}{-\epsilon}, \quad \lim_{\epsilon \downarrow 0} \frac{f(0 + \epsilon) - f(0)}{\epsilon}.$$

Indeed, these limits do not exist, where as the limits $f'(0-)$ and $f'(0+)$ do exist. In any event, f is piecewise differentiable, although it is not continuous.

3. Let $I = [0, 1]$ and define $f : I \to \mathbb{R}$ by $f(x) = \sqrt{x(1 - x)}$. On $(0, 1)$, f is differentiable. Also, the limits $f(0+)$ and $f(1-)$ exist. However, the limits $f'(0+)$ and $f'(1-)$ do not exist, as we saw in Example 3.2.3–3. Thus f is not piecewise differentiable. However, it is continuous, and therefore piecewise continuous, on $[0, 1]$. •

3.2.8 Notes

It was Weierstrass who first proved the existence of a continuous but nowhere differentiable function. The example Weierstrass gave was

$$\tilde{f}(x) = \sum_{j=0}^{\infty} b^n \cos(a^n \pi x),$$

where $b \in (0, 1)$ and a satisfies $ab > \frac{3}{2}\pi + 1$. It requires a little work to show that this function is nowhere differentiable. The example we give as Example 3.2.9 is fairly simple by comparison, and is taken from the paper of McCarthy [1953].

Example 3.2.31–2 if from [Siksek and El-Sedy 2004]

Exercises

3.2.1 Let $I \subseteq \mathbb{R}$ be an interval and let $f, g\colon I \to \mathbb{R}$ be differentiable. Is it true that the functions

$$I \ni x \mapsto \min\{f(x), g(x)\} \in \mathbb{R}, \qquad I \ni x \mapsto \max\{f(x), g(x)\} \in \mathbb{R},$$

are differentiable? If it is true provide a proof, if it is not true, give a counterexample.

Section 3.3

The Riemann integral

Opposite to the derivative, in a sense made precise by Theorem 3.3.30, is the notion of integration. In this section we describe a "simple" theory of integration, called Riemann integration,[8] that typically works insofar as computations go. In Chapter **??** we shall see that the Riemann integration suffers from a defect somewhat like the defect possessed by rational numbers. That is to say, just like there are sequences of rational numbers that seem like they should converge (i.e., are Cauchy) but do not, there are sequences of functions possessing a Riemann integral which do not converge to a function possessing a Riemann integral (see Example **??**). This has some deleterious consequences for developing a general theory based on the Riemann integral, and the most widely used fix for this is the Lebesgue integral of Chapter **??**. However, for now let us stick to the more pedestrian, and more easily understood, Riemann integral.

As we did with differentiation, we suppose that the reader has had the sort of calculus course where they learn to compute integrals of common functions. Indeed, while we do not emphasise the art of computing integrals, we do not intend this to mean that this art should be ignored. The reader should know the basic integrals and the basic tricks and techniques for computing them. ***missing stuff***

Do I need to read this section? The best way to think of this section is as a setup for the general developments of Chapter **??**. Indeed, we begin Chapter **??** with essentially a deconstruction of what we do in this section. For this reason, this chapter should be seen as preparatory to Chapter **??**, and so can be skipped until one wants to learn Lebesgue integration in a serious way. At that time, a reader may wish to be prepared by understanding the slightly simpler Riemann integral. •

3.3.1 Step functions

Our discussion begins by our considering intervals that are compact. In Section 3.3.4 we consider the case of noncompact intervals.

In a theme that will be repeated when we consider the Lebesgue integral in Chapter **??**, we first introduce a simple class of functions whose integral is "obvious." These functions are then used to approximate a more general class of functions which are those that are considered "integrable." For the Riemann integral, the simple class of functions are defined as being constant on the intervals forming a partition. We recall from Definition 2.5.7 the notion of a partition and from the

[8] After Georg Friedrich Bernhard Riemann, 1826–1866. Riemann made important and long lasting contributions to real analysis, geometry, complex function theory, and number theory, to name a few areas. The presently unsolved Riemann Hypothesis is one of the outstanding problems in modern mathematics.

discussion surrounding the definition the notion of the endpoints associated with a partition.

3.3.1 Definition (Step function) Let $I = [a, b]$ be a compact interval. A function $f : I \to \mathbb{R}$ is a ***step function*** if there exists a partition $P = (I_1, \ldots, I_k)$ of I such that

(i) $f|\operatorname{int}(I_j)$ is a constant function for each $j \in \{1, \ldots, k\}$,

(ii) $f(a+) = f(a)$ and $f(b-) = f(b)$, and

(iii) for each $x \in \operatorname{EP}(P) \setminus \{a, b\}$, either $f(x-) = f(x)$ or $f(x+) = f(x)$. •

In Figure 3.10 we depict a typical step function. Note that at discontinuities

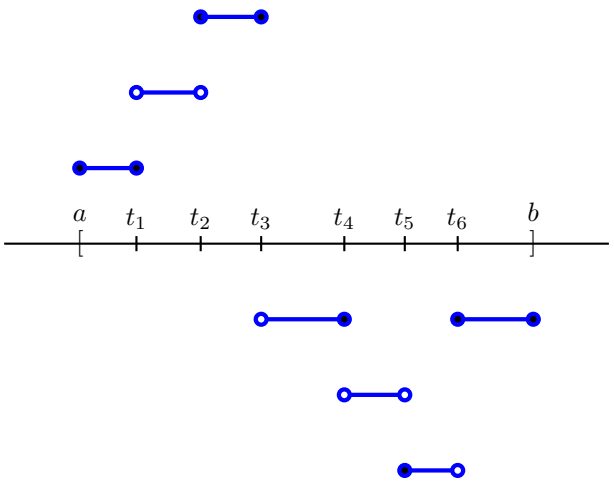


Figure 3.10 A step function

we allow the function to be continuous from either the right or the left. In the development we undertake, it does not really matter which it is.

The idea of the integral of a function is that it measures the "area" below the graph of a function. If the value of the function is negative, then the area is taken to be negative. For step functions, this idea of the area under the graph is clear, so we simply define this to be the integral of the function.

3.3.2 Definition (Riemann integral of a step function) Let $I = [a, b]$ and let $f : I \to \mathbb{R}$ be a step function defined using the partition $P = (I_1, \ldots, I_k)$ with endpoints $\operatorname{EP}(P) = (x_0, x_1, \ldots, x_k)$. Suppose that the value of f on $\operatorname{int}(I_j)$ is c_j for $j \in \{1, \ldots, k\}$. The ***Riemann integral*** of f is

$$A(f) = \sum_{j=1}^{k} c_j(x_j - x_{j-1}).$$ •

The notation $A(f)$ is intended to suggest "area."

3.3.2 The Riemann integral on compact intervals

Next we define the Riemann integral of a function that is not necessarily a step function. We do this by approximating a function by step functions.

3.3.3 Definition (Lower and upper step functions) Let $I = [a, b]$ be a compact interval, let $f : I \to \mathbb{R}$ be a bounded function, and let $P = (I_1, \dots, I_k)$ be a partition of I.

(i) The ***lower step function*** associated to f and P is the function $s_-(f, P) : I \to \mathbb{R}$ defined according to the following:
   (a) if $x \in I$ lies in the interior of an interval I_j, $j \in \{1, \dots, k\}$, then $s_-(f, P)(x) = \inf\{f(x) \mid x \in \operatorname{cl}(I_j)\}$;
   (b) $s_-(f, P)(a) = s_-(f, P)(a+)$ and $s_-(f, P)(b) = s_-(f, P)(b-)$;
   (c) for $x \in \operatorname{EP}(P) \setminus \{a, b\}$, $s_-(f, P)(x) = s_-(f, P)(x+)$.

(ii) The ***upper step function*** associated to f and P is the function $s_+(f, P) : I \to \mathbb{R}$ defined according to the following:
   (a) if $x \in I$ lies in the interior of an interval I_j, $j \in \{1, \dots, k\}$, then $s_+(f, P)(x) = \sup\{f(x) \mid x \in \operatorname{cl}(I_j)\}$;
   (b) $s_+(f, P)(a) = s_+(f, P)(a+)$ and $s_+(f, P)(b) = s_+(f, P)(b-)$;
   (c) for $x \in \operatorname{EP}(P) \setminus \{a, b\}$, $s_+(f, P)(x) = s_+(f, P)(x+)$. •

Note that both the lower and upper step functions are well-defined since f is bounded. Note also that at the middle endpoints for the partition, we ask that the lower and upper step functions be continuous from the right. This is an arbitrary choice. Finally, note that for each $x \in [a, b]$ we have

$$s_-(f, P)(x) \leq f(x) \leq s_+(f, P)(x).$$

That is to say, for any bounded function f, we have defined two step functions, one bounding f from below and one bounding f from above.

Next we associate to the lower and upper step functions their integrals, which we hope to use to define the integral of the function f.

3.3.4 Definition (Lower and upper Riemann sums) Let $I = [a, b]$ be a compact interval, let $f : I \to \mathbb{R}$ be a bounded function, and let $P = (I_1, \dots, I_k)$ be a partition of I.

(i) The ***lower Riemann sum*** associated to f and P is $A_-(f, P) = A(s_-(f, P))$.

(ii) The ***upper Riemann sum*** associated to f and P is $A_+(f, P) = A(s_+(f, P))$. •

Now we define the best approximations of the integral of f using the lower and upper Riemann sums.

3.3.5 Definition (Lower and upper Riemann integral) Let $I = [a, b]$ be a compact interval and let $f : I \to \mathbb{R}$ be a bounded function.

(i) The ***lower Riemann integral*** of f is

$$I_-(f) = \sup\{A_-(f, P) \mid P \in \operatorname{Part}(I)\}.$$

(ii) The ***upper Riemann integral*** of f is

$$I_+(f) = \inf\{A_+(f,P) \mid P \in \mathrm{Part}(I)\}.$$ •

Note that since f is bounded, it follows that the sets

$$\{A_-(f,P) \mid P \in \mathrm{Part}(I)\}, \quad \{A_+(f,P) \mid P \in \mathrm{Part}(I)\}$$

are bounded (why?). Therefore, the lower and upper Riemann integral always exist. So far, then, we have made a some constructions that apply to *any* bounded function. That is to say, for any bounded function, it is possible to define the lower and upper Riemann integral. What is not clear is that these two things should be equal. In fact, they are *not* generally equal, which leads to the following definition.

3.3.6 Definition (Riemann integrable function on a compact interval) A bounded function $f\colon [a,b] \to \mathbb{R}$ on a compact interval is ***Riemann integrable*** if $I_-(f) = I_+(f)$. We denote

$$\int_a^b f(x)\,\mathrm{d}x = I_-(f) = I_+(f),$$

which is the ***Riemann integral*** of f. The function f is called the ***integrand***. •

3.3.7 Notation (Swapping limits of integration) In the expression $\int_a^b f(x)\,\mathrm{d}x$, "$a$" is the ***lower limit of integration*** and "b" is the ***upper limit of integration***. We have tacitly assumed that $a < b$ in our constructions to this point. However, we can consider the case where $b < a$ by adopting the convention that

$$\int_b^a f(x)\,\mathrm{d}x = -\int_a^b f(x)\,\mathrm{d}x.$$ •

Let us provide an example which illustrates that, in principle, it is possible to use the definition of the Riemann integral to perform computations, even though this is normally tedious. A more common method for computing integrals is to use the Fundamental Theorem of Calculus to "reverse engineer" the process.

3.3.8 Example (Computing a Riemann integral) Let $I = [0,1]$ and define $f\colon I \to \mathbb{R}$ by $f(x) = x$. Let $P = (I_1, \ldots, I_k)$ be a partition with $s_-(f,P)$ and $s_+(f,P)$ the associated lower and upper step functions, respectively. Let $\mathrm{EP}(P) = (x_0, x_1, \ldots, x_k)$ be the endpoints of the intervals of the partition. One can then see that, for $j \in \{1, \ldots, k\}$, $s_-(f,P)|\operatorname{int}(I_j) = x_{j-1}$ and $s_+(f,P)|\operatorname{int}(I_j) = x_j$. Therefore,

$$A_-(f,P) = \sum_{j=1}^{k} x_{j-1}(x_j - x_{j-1}), \quad A_+(f,P) = \sum_{j=1}^{k} x_j(x_j - x_{j-1}).$$

We claim that $I_-(f) \geq \frac{1}{2}$ and that $I_+(f) \leq \frac{1}{2}$, and note that, once we prove this, it follows that f is Riemann integrable and that $I_-(f) = I_+(f) = \frac{1}{2}$ (why?).

For $k \in \mathbb{Z}_{>0}$ consider the partition P_k with endpoints $\mathrm{EP}(P_k) = \{\frac{j}{k} \mid j \in \{0, 1, \ldots, k\}\}$. Then, using the formula $\sum_{j=1}^{l} j = \frac{1}{2}l(l+1)$, we compute

$$A_-(f, P_k) = \sum_{j=1}^{k} \frac{j-1}{k^2} = \frac{k(k-1)}{2k^2}, \qquad A_+(f, P_k) = \sum_{j=1}^{k} \frac{j}{k^2} = \frac{k(k+1)}{2k^2}.$$

Therefore,

$$\lim_{k\to\infty} A_-(f, P_k) = \tfrac{1}{2}, \qquad \lim_{k\to\infty} A_+(f, P_k) = \tfrac{1}{2}.$$

This shows that $I_-(f) \geq \frac{1}{2}$ and that $I_+(f) \leq \frac{1}{2}$, as desired. •

3.3.3 Characterisations of Riemann integrable functions on compact intervals

In this section we provide some insightful characterisations of the notion of Riemann integrability. First we provide four equivalent characterisations of the Riemann integral. Each of these captures, in a slightly different manner, the notion of the Riemann integral as a limit. It will be convenient to introduce the language that a ***selection*** from a partition $P = (I_1, \ldots, I_k)$ is a family $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)$ of points such that $\xi_j \in \mathrm{cl}(I_j)$, $j \in \{1, \ldots, k\}$.

3.3.9 Theorem (Riemann, Darboux,[9] and Cauchy characterisations of Riemann integrable functions) *For a compact interval* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *and a bounded function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$*, the following statements are equivalent:*

(i) f *is Riemann integrable;*

(ii) for every $\epsilon \in \mathbb{R}_{>0}$*, there exists a partition* P *such that* $\mathrm{A}_+(\mathrm{f}, \mathrm{P}) - \mathrm{A}_-(\mathrm{f}, \mathrm{P}) < \epsilon$ *(**Riemann's condition**);*

(iii) there exists $\mathrm{I}(\mathrm{f}) \in \mathbb{R}$ *such that, for every* $\epsilon \in \mathbb{R}_{>0}$ *there exists* $\delta \in \mathbb{R}_{>0}$ *such that, if* $\mathrm{P} = (\mathrm{I}_1, \ldots, \mathrm{I}_k)$ *is a partition for which* $|\mathrm{P}| < \delta$ *and if* $(\xi_1, \ldots, \xi_k)$ *is a selection from* P*, then*

$$\left| \sum_{j=1}^{k} \mathrm{f}(\xi_j)(\mathrm{x}_j - \mathrm{x}_{j-1}) - \mathrm{I}(\mathrm{f}) \right| < \epsilon,$$

where $\mathrm{EP}(\mathrm{P}) = (\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_k)$ *(**Darboux' condition**);*

(iv) for each $\epsilon \in \mathbb{R}_{>0}$ *there exists* $\delta \in \mathbb{R}_{>0}$ *such that, for any partitions* $\mathrm{P} = (\mathrm{I}_1, \ldots, \mathrm{I}_k)$ *and* $\mathrm{P}' = (\mathrm{I}'_1, \ldots, \mathrm{I}'_{k'})$ *with* $|\mathrm{P}|, |\mathrm{P}'| < \delta$ *and for any selections* $(\xi_1, \ldots, \xi_k)$ *and* $(\xi'_1, \ldots, \xi'_{k'})$ *from* P *and* P′*, respectively, we have*

$$\left| \sum_{j=1}^{k} \mathrm{f}(\xi_j)(\mathrm{x}_j - \mathrm{x}_{j-1}) - \sum_{j=1}^{k'} \mathrm{f}(\xi'_j)(\mathrm{x}'_j - \mathrm{x}'_{j-1}) \right| < \epsilon,$$

where $\mathrm{EP}(\mathrm{P}) = (\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_k)$ *and* $\mathrm{EP}(\mathrm{P}') = (\mathrm{x}'_0, \mathrm{x}'_1, \ldots, \mathrm{x}'_{k'})$ *(**Cauchy's condition**).*

Proof First let us prove a simple lemma about lower and upper Riemann sums and refinements of partitions.

[9] Jean Gaston Darboux (1842–1917) was a French mathematician. His made important contributions to analysis and differential geometry.

1 Lemma *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$, *let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be bounded, and let* P_1 *and* P_2 *be partitions of* I *with* P_2 *a refinement of* P_1. *Then*

$$\mathrm{A}_-(\mathrm{f}, \mathrm{P}_2) \geq \mathrm{A}_-(\mathrm{f}, \mathrm{P}_1), \quad \mathrm{A}_+(\mathrm{f}, \mathrm{P}_2) \leq \mathrm{A}_+(\mathrm{f}, \mathrm{P}_1).$$

Proof Let $x_1, x_2 \in \mathrm{EP}(P_1)$ and denote by $y_1, \dots, y_l$ the elements of $\mathrm{EP}(P_2)$ that satisfy

$$x_1 \leq y_1 < \cdots < y_l \leq x_2.$$

Then

$$\begin{aligned}\sum_{j=1}^{l} (y_j - y_{j-1}) \inf\{f(y) \mid\ y \in [y_j, y_{j-1}]\} &\geq \sum_{j=1}^{l} (y_j - y_{j-1}) \inf\{f(x) \mid\ x \in [x_1, x_2]\} \\ &= (x_2 - x_1) \inf\{f(x) \mid\ x \in [x_1, x_2]\}.\end{aligned}$$

Now summing over all consecutive pairs of endpoints for P_1 gives $A_-(f, P_2) \geq A_-(f, P_1)$. A similar argument gives $A_+(f, P_2) \leq A_+(f, P_1)$. ▼

The following trivial lemma will also be useful.

2 Lemma $\mathrm{I}_-(\mathrm{f}) \leq \mathrm{I}_+(\mathrm{f})$.

Proof Since, for any two partitions P_1 and P_2, we have

$$s_-(f, P_1) \leq f(x) \leq s_+(f, P_2),$$

it follows that

$$\sup\{A_-(f, P) \mid\ P \in \mathrm{Part}(I)\} \leq \inf\{A_+(f, P) \mid\ P \in \mathrm{Part}(I)\},$$

which is the result. ▼

(i) $\Longrightarrow$ (ii) Suppose that f is Riemann integrable and let $\epsilon \in \mathbb{R}_{>0}$. Then there exists partitions P_- and P_+ such that

$$A_-(f, P_-) > I_-(f) - \tfrac{\epsilon}{2}, \quad A_+(f, P_+) < I_+(f) + \tfrac{\epsilon}{2}.$$

Now let P be a partition that is a refinement of both P_1 and P_2 (obtained, for example, by asking that $\mathrm{EP}(P) = \mathrm{EP}(P_1) \cup \mathrm{EP}(P_2)$). By Lemma 1 it follows that

$$A_+(f, P) - A_-(f, P) \leq A_+(f, P_+) - A_-(f, P_-) < I_+(f) + \tfrac{\epsilon}{2} - I_-(f) + \tfrac{\epsilon}{2} = \epsilon.$$

(ii) $\Longrightarrow$ (i) Now suppose that $\epsilon \in \mathbb{R}_{>0}$ and let P be a partition such that $A_+(f, P) - A_-(f, P) < \epsilon$. Since we additionally have $I_-(f) \leq I_+(f)$ by Lemma 2, it follows that

$$A_-(f, P) \leq I_-(f) \leq I_+(f) \leq A_+(f, P),$$

from which we deduce that

$$0 \leq I_+(f) - I_-(f) < \epsilon.$$

Since ϵ is arbitrary, we conclude that $I_-(f) = I_+(f)$, as desired.

(i) $\Longrightarrow$ (iii) We first prove a lemma about partitions of compact intervals.

3 Lemma *If* $\mathrm{P} = (\mathrm{I}_1, \ldots, \mathrm{I}_k)$ *is a partition of* $[a, b]$ *and if* $\epsilon \in \mathbb{R}_{>0}$*, then there exists* $\delta \in \mathbb{R}_{>0}$ *such that, if* $\mathrm{P}' = (\mathrm{I}'_1, \ldots, \mathrm{I}'_{k'})$ *is a partition with* $|\mathrm{P}'| < \delta$ *and if*

$$\{j'_1, \ldots, j'_r\} = \{j' \in \{1, \ldots, k'\} \mid \operatorname{cl}(\mathrm{I}'_{j'}) \not\subset \operatorname{cl}(\mathrm{I}_j) \text{ for any } j \in \{1, \ldots, k\}\},$$

then

$$\sum_{l=1}^{r} |x_{j'_l} - x_{j'_l - 1}| < \epsilon,$$

where $\mathrm{EP}(\mathrm{P}') = (x_0, x_1, \ldots, x_{k'})$.

Proof Let $\epsilon \in \mathbb{R}_{>0}$ and take $\delta = \frac{\epsilon}{k+1}$. Let $P' = (I'_1, \ldots, I'_{k'})$ be a partition with endpoints $(x_0, x_1, \ldots, x_{k'})$ and satisfying $|P'| < \delta$. Define

$$K_1 = \{j' \in \{1, \ldots, k'\} \mid \operatorname{cl}(I'_{j'}) \not\subset \operatorname{cl}(I_j) \text{ for any } j \in \{1, \ldots, k\}\}.$$

If $j' \in K_1$ then $I'_{j'}$ is not contained in any interval of P and so $I'_{j'}$ must contain at least one endpoint from P. Since P has $k+1$ endpoints we obtain $\operatorname{card}(K_1) \leq k+1$. Since the intervals $I'_{j'}$, $j' \in K_1$, have length at most δ we have

$$\sum_{j' \in K_1} (x_{j'} - x_{j'-1}) \leq (k+1)\delta \leq \epsilon,$$

as desired. ▼

Now let $\epsilon \in \mathbb{R}_{>0}$ and define $M = \sup\{|f(x)| \mid x \in I\}$. Denote by $I(f)$ the Riemann integral of f. Choose partitions P_- and P_+ such that

$$I(f) - A_-(f, P_-) < \tfrac{\epsilon}{2}, \quad A_+(f, P_+) - I(f) < \tfrac{\epsilon}{2}.$$

If $P = (I_1, \ldots, I_k)$ is chosen such that $\mathrm{EP}(P) = \mathrm{EP}(P_-) \cup \mathrm{EP}(P_+)$, then

$$I(f) - A_-(f, P) < \tfrac{\epsilon}{2}, \quad A_+(f, P) - I(f) < \tfrac{\epsilon}{2}.$$

By Lemma 3 choose $\delta \in \mathbb{R}_{>0}$ such that if P' is any partition for which $|P'| < \delta$ then the sum of the lengths of the intervals of P' not contained in some interval of P does not exceed $\frac{\epsilon}{2M}$. Let $P' = (I'_1, \ldots, I'_{k'})$ be a partition with endpoints $(x_0, x_1, \ldots, x_{k'})$ and satisfying $|P'| < \delta$. Denote

$$K_1 = \{j' \in \{1, \ldots, k'\} \mid I'_{j'} \not\subset I_j \text{ for some } j \in \{1, \ldots, k\}\}$$

and $K_2 = \{1, \ldots, k'\} \setminus K_1$. Let $(\xi_1, \ldots, \xi_{k'})$ be a selection of P'. Then we compute

$$\begin{aligned} \sum_{j=1}^{k'} f(\xi_j)(x_j - x_{j-1}) &= \sum_{j \in K_1} f(\xi_j)(x_j - x_{j-1}) + \sum_{j \in K_2} f(\xi_j)(x_j - x_{j-1}) \\ &\leq A_+(f, P) + M\frac{\epsilon}{2M} < I(f) + \epsilon. \end{aligned}$$

In like manner we show that

$$\sum_{j=1}^{k'} f(\xi_j)(x_j - x_{j-1}) > I(f) - \epsilon.$$

This gives

$$\left|\sum_{j=1}^{k'} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| < \epsilon,$$

as desired.

(iii) $\implies$ (ii) Let $\epsilon \in \mathbb{R}_{>0}$ and let $P = (I_1, \dots, I_k)$ be a partition for which

$$\left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| < \frac{\epsilon}{4}$$

for every selection $(\xi_1, \dots, \xi_k)$ from P. Now particularly choose a selection such that

$$|f(\xi_j) - \sup\{f(x) \mid x \in \mathrm{cl}(I_j)\}| < \frac{\epsilon}{4k(x_j - x_{j-1})}.$$

Then

$$\begin{aligned}|A_+(f,P) - I(f)| &\le \left|A_+(f,P) - \sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1})\right| + \left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| \\ &< \sum_{j=1}^{k} \frac{\epsilon}{4k(x_j - x_{j-1})}(x_j - x_{j-1}) + \frac{\epsilon}{4} < \frac{\epsilon}{2}.\end{aligned}$$

In like manner one shows that $|A_-(f,P) - I(f)| < \frac{\epsilon}{2}$. Therefore,

$$|A_+(f,P) - A_-(f,P)| \le |A_+(f,P) - I(f)| + |I(f) - A_-(f,P)| < \epsilon,$$

as desired.

(iii) $\implies$ (iv) Let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta \in \mathbb{R}_{>0}$ have the property that, whenever $P = (I_1, \dots_k)$ is a partition satisfying $|P| < \delta$ and $(\xi_1, \dots, \xi_k)$ is a selection from P, it holds that

$$\left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| < \frac{\epsilon}{2}.$$

Now let $P = (I_1, \dots, I_k)$ and $P' = (I'_1, \dots, I'_{k'})$ be two partitions with $|P|, |P'| < \delta$, and let $(\xi_1, \dots, \xi_k)$ and $(\xi'_1, \dots, \xi'_{k'})$ selections from P and P', respectively. Then we have

$$\begin{aligned}&\left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - \sum_{j=1}^{k'} f(\xi'_j)(x'_j - x'_{j-1})\right| \\ &\qquad \le \left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| + \left|\sum_{j=1}^{k'} f(\xi'_j)(x'_j - x'_{j-1}) - I(f)\right| < \epsilon,\end{aligned}$$

which gives this part of the result.

(iv) $\implies$ (iii) Let $(P_j = (I_{j,1}, \dots, I_{j,k_j}))_{j \in \mathbb{Z}_{>0}}$ be a sequence of partitions for which $\lim_{j\to\infty} |P_j| = 0$. Then, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that

$$\left|\sum_{j=1}^{k_l} f(\xi_{l,j})(x_{l,j} - x_{l,j-1}) - \sum_{j=1}^{k_m} f(\xi_{m,j})(x_{m,j} - x_{m,j-1})\right| < \epsilon,$$

for $l, m \geq N$, where $\boldsymbol{\xi}_j = (\xi_{j,1}, \ldots, \xi_{j,k_j})$, is a selection from P_j, $j \in \mathbb{Z}_{>0}$, and where $\mathrm{EP}(P_j) = (x_{j,0}, x_{j,1}, \ldots, x_{j,k_j})$, $j \in \mathbb{Z}_{>0}$. If we define

$$A(f, P_j, \boldsymbol{\xi}_j) = \sum_{r=1}^{k_j} f(\xi_r)(x_{j,r} - x_{j,r-1}),$$

then the sequence $(A(f, P_j, \boldsymbol{\xi}_j))_{j \in \mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ for any choices of points $\boldsymbol{\xi}_j$, $j \in \mathbb{Z}_{>0}$. Denote the resulting limit of this sequence by $I(f)$. We claim that $I(f)$ is the Riemann integral of f. To see this, let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta \in \mathbb{R}_{>0}$ be such that

$$\left| \sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - \sum_{j=1}^{k'} f(\xi'_j)(x'_j - x'_{j-1}) \right| < \frac{\epsilon}{2}$$

for any two partitions P and P' satisfying $|P|, |P'| < \delta$ and for any selections $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ from P and P', respectively. Now let $N \in \mathbb{Z}_{>0}$ satisfy $|P_j| < \delta$ for every $j \geq N$. Then, if P is any partition with $|P| < \delta$ and if $\boldsymbol{\xi}$ is any selection from P, we have

$$|A(f, P, \boldsymbol{\xi}) - I(f)| \leq |A(f, P, \boldsymbol{\xi}) - A(f, P_N, \boldsymbol{\xi}_N)| + |A(f, P_N, \boldsymbol{\xi}_N) - I(f)| < \epsilon,$$

for any selection $\boldsymbol{\xi}_N$ of P_N. This shows that $I(f)$ is indeed the Riemann integral of f, and so gives this part of the theorem. ■

A consequence of the proof is that, of course, the quantity $I(f)$ in part (iii) of the theorem is nothing other than the Riemann integral of f.

Many of the functions one encounters in practice are, in fact, Riemann integrable. However, not all functions are Riemann integrable, as the following simple examples shows.

3.3.10 Example (A function that is not Riemann integrable) Let $I = [0, 1]$ and let $f : I \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 1, & x \in \mathbb{Q} \cap I \\ 0, & x \notin \mathbb{Q} \cap I. \end{cases}$$

Thus f takes the value 1 at all rational points, and is zero elsewhere. Now let $s_+, s_- : I \to \mathbb{R}$ be any step functions satisfying $s_-(x) \leq f(x) \leq s_+(x)$ for all $x \in I$. Since any nonempty subinterval of I contains infinitely many irrational numbers, it follows that $s_-(x) \leq 0$ for every $x \in I$. Since every nonempty subinterval of I contains infinitely many rational numbers, it follows that $s_+(x) \geq 1$ for every $x \in I$. Therefore, $A(s_+) - A(s_-) \geq 1$. It follows from Theorem 3.3.9 that f is not Riemann integrable. While this example may seem pointless and contrived, it will be used in Examples 4.5.71 and **??** to exhibit undesirable features of the Riemann integral. •

The following result provides an interesting characterisation of Riemann integrable functions, illustrating precisely the sorts of functions whose Riemann integrals may be computed.

3.3.11 Theorem (Riemann integrable functions are continuous almost everywhere, and vice versa) *For a compact interval* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$*, a bounded function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is Riemann integrable if and only if the set*

$$\mathrm{D_f} = \{\mathrm{x} \in \mathrm{I} \mid \ \mathrm{f} \textit{ is discontinuous at } \mathrm{x}\}$$

has measure zero.

Proof Recall from Definition 3.1.10 the notion of the oscillation ω_f for a function f, and that $\omega_f(x) = 0$ if and only if f is continuous at x. For $k \in \mathbb{Z}_{>0}$ define

$$D_{f,k} = \left\{x \in I \;\middle|\; \omega_f(x) \geq \tfrac{1}{k}\right\}.$$

Then Proposition 3.1.11 implies that $D_f = \cup_{k \in \mathbb{Z}_{>0}} D_{f,k}$. By Exercise 2.5.9 we can assert that D_f has measure zero if and only if each of the sets $D_{f,k}$ has measure zero, $k \in \mathbb{Z}_{>0}$.

Now suppose that $D_{f,k}$ does not have measure zero for some $k \in \mathbb{Z}_{>0}$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that, if a family $((a_j, b_j))_{j \in \mathbb{Z}_{>0}}$ of open intervals has the property that

$$D_{f,k} \subseteq \bigcup_{j \in \mathbb{Z}_{>0}} (a_j, b_j),$$

then

$$\sum_{j=1}^{\infty} |b_j - a_j| \geq \epsilon.$$

Now let P be a partition of I and denote $\mathrm{EP}(P) = (x_0, x_1, \ldots, x_m)$. Now let $\{j_1, \ldots, j_l\} \subseteq \{1, \ldots, m\}$ be those indices for which $j_r \in \{j_1, \ldots, j_l\}$ implies that $D_{f,k} \cap (x_{j_r-1}, x_{j_r}) \neq \emptyset$. Note that it follows that the set $\bigcup_{r=1}^{l} (x_{j_r-1}, x_{j_r})$ covers $D_{f,k}$ with the possible exception of a finite number of points. It then follows that one can enlarge the length of each of the intervals (x_{j_r-1}, x_{j_r}), $r \in \{1, \ldots, l\}$, by $\frac{\epsilon}{2l}$, and the resulting intervals will cover $D_{f,k}$. The enlarged intervals will have total length at least ϵ, which means that

$$\sum_{r=1}^{l} |x_{j_r} - x_{j_r-1}| \geq \frac{\epsilon}{2}.$$

Moreover, for each $r \in \{1, \ldots, l\}$,

$$\sup\{f(x) \mid \ x \in [x_{j_r-1}, x_{j_r}]\} - \inf\{f(x) \mid \ x \in [x_{j_r-1}, x_{j_r}]\} \geq \tfrac{1}{k}$$

since $D_{f,k} \cap (x_{j_r-1}, x_{j_r}) \neq \emptyset$ and by definition of $D_{f,k}$ and ω_f. It now follows that

$$\begin{aligned}
A_+(f,P) - A_-(f,P) &= \sum_{j=1}^{m} (x_j - x_{j-1}) \Big(\sup\{f(x) \mid \ x \in [x_{j-1}, x_j]\} \\
&\qquad - \inf\{f(x) \mid \ x \in [x_{j-1}, x_j]\} \Big) \\
&\geq \sum_{r=1}^{l} (x_{j_r} - x_{j_r-1}) \Big(\sup\{f(x) \mid \ x \in [x_{j_r-1}, x_{j_r}]\} \\
&\qquad - \inf\{f(x) \mid \ x \in [x_{j_r-1}, x_{j_r}]\} \Big) \\
&\geq \tfrac{\epsilon}{2k}.
\end{aligned}$$

Since this must hold for every partition, it follows that f is not Riemann integrable.

Now suppose that D_f has measure zero. Since f is bounded, let $M = \sup\{|f(x)| \mid x \in I\}$. Let $\epsilon \in \mathbb{R}_{>0}$ and for brevity define $\epsilon' = \frac{\epsilon}{b-a+2}$. Choose a sequence $((a_j, b_j))_{j\in\mathbb{Z}_{>0}}$ of open intervals such that

$$D_f \subseteq \bigcup_{j\in\mathbb{Z}_{>0}} I_j, \quad \sum_{j=1}^{\infty} |b_j - a_j| < \frac{\epsilon'}{M}.$$

Define $\delta\colon I \to \mathbb{R}_{>0}$ such that the following properties hold:

1. if $x \notin D_f$ then $\delta(x)$ is taken such that, if $y \in I \cap \mathsf{B}(\delta(x), x)$, then $|f(y) - f(x)| < \frac{\epsilon'}{2}$;
2. if $x \in D_f$ then $\delta(x)$ is taken such that $\mathsf{B}(\delta(x), x) \subseteq I_j$ for some $j \in \mathbb{Z}_{>0}$.

Now, by Proposition 2.5.10, let $((c_1, I_1), \ldots, (c_k, I_k))$ be a δ-fine tagged partition with $P = (I_1, \ldots, I_k)$ the associated partition. Now partition the set $\{1, \ldots, k\}$ into two sets K_1 and K_2 such that $j \in K_1$ if and only if $c_j \notin D_f$. Then we compute

$$\begin{aligned}
A_+(f,P) - A_-(f,P) &= \sum_{j=1}^{k} (x_j - x_{j-1})\Big(\sup\{f(x) \mid x \in [x_{j-1}, x_j]\} \\
&\quad - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\Big) \\
&= \sum_{j\in K_1} (x_j - x_{j-1})\Big(\sup\{f(x) \mid x \in [x_{j-1}, x_j]\} \\
&\quad - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\Big) \\
&\quad + \sum_{j\in K_2} (x_j - x_{j-1})\Big(\sup\{f(x) \mid x \in [x_{j-1}, x_j]\} \\
&\quad - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\Big) \\
&\leq \sum_{j\in K_1} \epsilon'(x_j - x_{j-1}) + \sum_{j\in K_2} 2M(x_j - x_{j-1}) \\
&\leq \epsilon'(b-a) + 2M \sum_{j=1}^{\infty} |b_j - a_j| \\
&< \epsilon'(b - a + 2) = \epsilon.
\end{aligned}$$

This part of the result now follows by Theorem 3.3.9. ■

The theorem indicates why the function of Example 3.3.10 is not Riemann integrable. Indeed, the function in that example is discontinuous at *all* points in $[0, 1]$ (why?). The theorem also has the following obvious corollary which illustrates why so many functions in practice are Riemann integrable.

3.3.12 Corollary (Continuous functions are Riemann integrable) *If* $\mathrm{f}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is continuous, then it is Riemann integrable.*

By virtue of Theorem **??**, we also have the following result, giving another large class of Riemann integrable functions, distinct from those that are continuous.

3.3.13 Corollary (Functions of bounded variation are Riemann integrable) *If* $\mathrm{f}\colon [\mathrm{a},\mathrm{b}] \to \mathbb{R}$ *has bounded variation, then* f *is Riemann integrable.*

3.3.4 The Riemann integral on noncompact intervals

Up to this point in this section we have only considered the Riemann integral for bounded functions defined on compact intervals. In this section we extend the notion of the Riemann integral to allow its definition for unbounded functions and for general intervals. There are complications that arise in this situation that do not arise in the case of a compact interval in that one has two possible notions of what one might call a Riemann integrable function. In all cases, we use the existing definition of the Riemann integral for compact intervals as our basis, and allow the other cases as limits.

3.3.14 Definition (Positive Riemann integrable function on a general interval) Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}_{\geq 0}$ be a function whose restriction to every compact subinterval of I is Riemann integrable.

(i) If $I = [a,b]$ then the Riemann integral of f is as defined in the preceding section.

(ii) If $I = (a,b]$ then define

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{r_a \downarrow a} \int_{r_a}^b f(x)\,\mathrm{d}x.$$

(iii) If $I = [a,b)$ then define

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{r_b \uparrow b} \int_a^{r_b} f(x)\,\mathrm{d}x.$$

(iv) If $I = (a,b)$ then define

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{r_a \downarrow a} \int_{r_a}^c f(x)\,\mathrm{d}x + \lim_{r_b \uparrow b} \int_c^{r_b} f(x)\,\mathrm{d}x$$

for some $c \in (a,b)$.

(v) If $I = (-\infty,b]$ then define

$$\int_{-\infty}^b f(x)\,\mathrm{d}x = \lim_{R \to \infty} \int_{-R}^b f(x)\,\mathrm{d}x.$$

(vi) If $I = (-\infty,b)$ then define

$$\int_{-\infty}^b f(x)\,\mathrm{d}x = \lim_{R \to \infty} \int_{-R}^c f(x)\,\mathrm{d}x + \lim_{r_b \uparrow b} \int_c^{r_b} f(x)\,\mathrm{d}x$$

for some $c \in (-\infty,b)$.

(vii) If $I = [a, \infty)$ then define

$$\int_a^\infty f(x)\,\mathrm{d}x = \lim_{R\to\infty} \int_a^R f(x)\,\mathrm{d}x.$$

(viii) If $I = (a, \infty)$ then define

$$\int_a^\infty f(x)\,\mathrm{d}x = \lim_{r_a \downarrow a} \int_{r_a}^c f(x)\,\mathrm{d}x + \lim_{R\to\infty} \int_c^R f(x)\,\mathrm{d}x$$

for some $c \in (a, \infty)$.

(ix) If $I = \mathbb{R}$ then define

$$\int_{-\infty}^\infty f(x)\,\mathrm{d}x = \lim_{R\to\infty} \int_{-R}^c f(x)\,\mathrm{d}x + \lim_{R\to\infty} \int_c^R f(x)\,\mathrm{d}x$$

for some $c \in \mathbb{R}$.

If, for a given I and f, the appropriate of the above limits exists, then f is ***Riemann integrable*** on I, and the ***Riemann integral*** is the value of the limit. Let us denote by

$$\int_I f(x)\,\mathrm{d}x$$

the Riemann integral. •

One can easily show that where, in the above definitions, one must make a choice of c, the definition is independent of this choice (cf. Proposition 3.3.26).

The above definition is intended for functions taking nonnegative values. For more general functions we have the following definition.

3.3.15 Definition (Riemann integrable function on a general interval) Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be a function whose restriction to any compact subinterval of I is Riemann integrable. Define $f_+, f_- : I \to \mathbb{R}_{\geq 0}$ by

$$f_+(x) = \max\{0, f(x)\}, \quad f_-(x) = -\min\{0, f(x)\}$$

so that $f = f_+ - f_-$. The function f is ***Riemann integrable*** if both f_+ and f_- are Riemann integrable, and the ***Riemann integral*** of f is

$$\int_I f(x)\,\mathrm{d}x = \int_I f_+(x)\,\mathrm{d}x - \int_I f_-(x)\,\mathrm{d}x.$$ •

At this point, if I is compact, we have potentially competing definitions for the Riemann integral of a bounded function $I : f \to \mathbb{R}$. One definition is the direct one of Definition 3.3.6. The other definition involves computing the Riemann integral, as per Definition 3.3.6, of the positive and negative parts of f, and then take the difference of these. Let us resolve the equivalence of these two notions.

3.3.16 Proposition (Consistency of definition of Riemann integral on compact intervals) *Let* $\mathrm{I} = [\mathrm{a},\mathrm{b}]$, *let* $\mathrm{f}\colon [\mathrm{a},\mathrm{b}] \to \mathbb{R}$, *and let* $\mathrm{f}_+, \mathrm{f}_-\colon [\mathrm{a},\mathrm{b}] \to \mathbb{R}_{\geq 0}$ *be the positive and negative parts of* f. *Then the following two statements are equivalent:*

(i) f *is integrable as per Definition 3.3.6 with Riemann integral* $\mathrm{I}(\mathrm{f})$*;*

(ii) f_+ *and* f_- *are Riemann integrable as per Definition 3.3.6 with Riemann integrals* $\mathrm{I}(\mathrm{f}_+)$ *and* $\mathrm{I}(\mathrm{f}_-)$.

Moreover, if one, and therefore both, of parts (i) and (ii) hold, then $\mathrm{I}(\mathrm{f}) = \mathrm{I}(\mathrm{f}_+) - \mathrm{I}(\mathrm{f}_-)$.

Proof We shall refer ahead to the results of Section 3.3.5.

(i) $\implies$ (ii) Define continuous functions $g_+, g_-\colon \mathbb{R} \to \mathbb{R}$ by

$$g_+(x) = \max\{0, x\}, \quad g_-(x) = -\min\{0, x\}$$

so that $f_+ = g_+ \circ f$ and $f_- = g_- \circ f$. By Proposition 3.3.23 (noting that the proof of that result is valid for the Riemann integral as per Definition 3.3.6) it follows that f_+ and f_- are Riemann integrable as per Definition 3.3.6.

(ii) $\implies$ (i) Note that $f = f_+ - f_-$. Also note that the proof of Proposition 3.3.22 is valid for the Riemann integral as per Definition 3.3.6. Therefore, f is Riemann integrable as per Definition 3.3.6.

Now we show that $I(f) = I(f_+) - I(f_-)$. This, however, follows immediately from Proposition 3.3.22. ■

It is not uncommon to see the general integral as we have defined it called the ***improper Riemann integral***.

The preceding definitions may appear at first to be excessively complicated. The following examples illustrate the rationale behind the care taken in the definitions.

3.3.17 Examples (Riemann integral on a general interval)

1. Let $I = (0,1]$ and let $f(x) = x^{-1}$. Then, if $r_a \in (0,1)$, we compute the proper Riemann integral
$$\int_{r_a}^{1} f(x)\,\mathrm{d}x = -\log r_a,$$
where log is the natural logarithm. Since $\lim_{r_a \downarrow} \log r_a = -\infty$ this function is not Riemann integrable on $(0,1]$.
2. Let $I = (0,1]$ and let $f(x) = x^{-1/2}$. Then, if $r_a \in (0,1)$, we compute the proper Riemann integral
$$\int_{r_a}^{1} f(x)\,\mathrm{d}x = 2 - 2\sqrt{r_a}.$$
In this case the function is Riemann integrable on $(0,1]$ and the value of the Riemann integral is 2.
3. Let $I = \mathbb{R}$ and define $f(x) = (1+x^2)^{-1}$. In this case we have
$$\begin{aligned}\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,\mathrm{d}x &= \lim_{R\to\infty} \int_{-R}^{0} \frac{1}{1+x^2}\,\mathrm{d}x + \lim_{R\to\infty} \int_{0}^{R} \frac{1}{1+x^2}\,\mathrm{d}x \\ &= \lim_{R\to\infty} \arctan R + \lim_{R\to\infty} \arctan R = \pi.\end{aligned}$$
Thus this function is Riemann integrable on $\mathbb{R}$ and has a Riemann integral of π.

4. The next example we consider is $I = \mathbb{R}$ and $f(x) = x(1 + x^2)^{-1}$. In this case we compute

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,\mathrm{d}x &= \lim_{R\to\infty} \int_{-R}^{0} \frac{x}{1+x^2}\,\mathrm{d}x + \lim_{R\to\infty} \int_{0}^{R} \frac{x}{1+x^2}\,\mathrm{d}x \\ &= \lim_{R\to\infty} \frac{1}{2}\log(1+R^2) - \lim_{R\to\infty} \frac{1}{2}\log(1+R^2).\end{aligned}$$

   Now, it is not permissible to say here that $\infty - \infty = 0$. Therefore, we are forced to conclude that f is not Riemann integrable on $\mathbb{R}$.

5. To make the preceding example a little more dramatic, and to more convincingly illustrate why we should not cancel the infinities, we take $I = \mathbb{R}$ and $f(x) = x^3$. Here we compute

$$\int_{-\infty}^{\infty} x^3\,\mathrm{d}x = \lim_{R\to\infty} \frac{1}{4}R^4 - \lim_{R\to\infty} \frac{1}{4}R^4.$$

   In this case again we must conclude that f is not Riemann integrable on $\mathbb{R}$. Indeed, it seems unlikely that one *would* wish to conclude that such a function was Riemann integrable since it is so badly behaved as $|t| \to \infty$. However, if we reject this function as being Riemann integrable, we must also reject the function of Example 4, even though it is not as ill behaved as the function here. •

Note that the above constructions involved first separating a function into its positive and negative parts, and then integrating these separately. However, there is not *a priori* reason why we could not have defined the limits in Definition 3.3.14 directly, and not just for positive functions. One can do this in fact. However, as we shall see, the two ensuing constructions of the integral are not equivalent.

3.3.18 Definition (Conditionally Riemann integrable functions on a general interval) Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function whose restriction to any compact subinterval of I is Riemann integrable. Then f is ***conditionally Riemann integrable*** if the limit in the appropriate of the nine cases of Definition 3.3.14 exists. This limit is called the ***conditional Riemann integral*** of f. If f is conditionally integrable we write

$$C\int_I f(x)\,\mathrm{d}x$$

as the conditional Riemann integral. •

missing stuff

Before we explain the differences between conditionally integrable and integrable functions via examples, let us provide the relationship between the two notions.

3.3.19 Proposition (Relationship between integrability and conditional integrability) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$, *then the following statements hold:*

(i) if f *is Riemann integrable then it is conditionally Riemann integrable;*

(ii) *if* I *is additionally compact then, if* f *is conditionally Riemann integrable it is Riemann integrable.*

Proof In the proof it is convenient to make use of the results from Section 3.3.5.

(i) Let f_+ and f_- be the positive and negative parts of f. Since f is Riemann integrable, then so are f_+ and f_- by Definition 3.3.15. Moreover, since Riemann integrability and conditional Riemann integrability are clearly equivalent for nonnegative functions, it follows that f_+ and f_- are conditionally Riemann integrable. Therefore, by Proposition 3.3.22, it follows that $f = f_+ - f_-$ is conditionally Riemann integrable.

(ii) This follows from Definition 3.3.15 and Proposition 3.3.16. ■

Let us show that conditional Riemann integrability and Riemann integrability are not equivalent.

3.3.20 Example (A conditionally Riemann integrable function that is not Riemann integrable) Let $I = [1, \infty)$ and define $f(x) = \frac{\sin x}{x}$. Let us first show that f is conditionally Riemann integrable. We have, using integration by parts (Proposition 3.3.28),

$$\begin{aligned}\int_1^\infty \frac{\sin x}{x}\,\mathrm{d}x &= \lim_{R\to\infty} \int_1^R \frac{\sin x}{x}\,\mathrm{d}x = \lim_{R\to\infty}\Big(-\frac{\cos x}{x}\Big|_1^R - \int_1^R \frac{\cos x}{x^2}\,\mathrm{d}x\Big)\\ &= \cos 1 - \lim_{R\to\infty} \int_1^R \frac{\cos x}{x^2}\,\mathrm{d}x.\end{aligned}$$

We claim that the last limit exists. Indeed,

$$\Big|\int_1^R \frac{\cos x}{x^2}\,\mathrm{d}x\Big| \le \int_1^R \frac{|\cos x|}{x^2}\,\mathrm{d}x \le \int_1^R \frac{1}{x^2}\,\mathrm{d}x = 1 - \frac{1}{R},$$

and the limit as $R \to \infty$ is then 1. This shows that the limit defining the conditional integral is indeed finite, and so f is conditionally Riemann integrable on $[1, \infty)$.

Now let us show that this function is not Riemann integrable. By Proposition 3.3.25, f is Riemann integrable if and only if $|f|$ is Riemann integrable. For $R > 0$ let $N_R \in \mathbb{Z}_{>0}$ satisfy $R \in [N_R\pi, (N_R + 1)\pi]$. We then have

$$\begin{aligned}\int_1^R \Big|\frac{\sin x}{x}\Big|\,\mathrm{d}x &\ge \int_\pi^{N_R\pi} \Big|\frac{\sin x}{x}\Big|\,\mathrm{d}x\\ &\ge \sum_{j=1}^{N_R-1} \frac{1}{j\pi} \int_{j\pi}^{(j+1)\pi} |\sin x|\,\mathrm{d}x = \frac{2}{\pi}\sum_{j=1}^{N_R-1}\frac{1}{j}.\end{aligned}$$

By Example 2.4.2–2, the last sum diverges to ∞ as $N_R \to \infty$, and consequently the integral on the left diverges to ∞ as $R \to \infty$, giving the assertion. •

3.3.21 Remark ("Conditional Riemann integral" versus "Riemann integral") The previous example illustrates that one needs to exercise some care when talking about the Riemann integral. Adding to the possible confusion here is the fact that there is no established convention concerning what is intended when one says "Riemann integral." Many authors use "Riemann integrability" where we use "conditional Riemann integrability" and then use "absolute Riemann integrability" where we use "Riemann integrability." There is a good reason to do this.

1. One can think of integrals as being analogous to sums. When we talked about convergence of sums in Section 2.4 we used "convergence" to talk about that concept which, for the Riemann integral, is analogous to "conditional Riemann integrability" in our terminology. We used the expression "absolute convergence" for that concept which, for the Riemann integral, is analogous to "Riemann integrability" in our terminology. Thus the alternative terminology of "Riemann integrability" for "conditional Riemann integrability" and "absolute Riemann integrability" for "Riemann integrability" is more in alignment with the (more or less) standard terminology for sums.

However, there is also a good reason to use the terminology we use. However, the reasons here have to do with terminology attached to the Lebesgue integral that we discuss in Chapter **??**. However, here is as good a place as any to discuss this.

2. For the Lebesgue integral, the most natural notion of integrability is analogous to the notion of "Riemann integrability" in our terminology. That is, the terminology "Lebesgue integrability" is a generalisation of "Riemann integrability." The notion of "conditional Riemann integrability" is not much discussed for the Lebesgue integral, so there is not so much an established terminology for this. However, if there were an established terminology it would be "conditional Lebesgue integrability."

In Table 3.1 we give a summary of the preceding discussion, noting that apart

Table 3.1 "Conditional" versus "absolute" terminology. In the top row we give our terminology, in the second row we give the alternative terminology for the Riemann integral, in the third row we give the analogous terminology for sums, and in the fourth row we give the terminology for the Lebesgue integral.

	Riemann integrable	conditionally Riemann integrable
Alternative	absolutely Riemann integrable	Riemann integrable
Sums	absolutely convergent	convergent
Lebesgue integral	Lebesgue integrable	conditionally Lebesgue integrable

from overwriting some standard conventions, there is no optimal way to choose what language to use. Our motivation for the convention we use is that it is best that "Lebesgue integrability" should generalise "Riemann integrability." But it is necessary to understand what one is reading and what is intended in any case. •

3.3.5 The Riemann integral and operations on functions

In this section we consider the interaction of integration with the usual algebraic and other operations on functions. We will consider both Riemann integrability and conditional Riemann integrability. If we wish to make a statement that we intend to hold for both notions, we shall write "(conditionally) Riemann integrable" to connote this. We will also write

$$(C)\int_I f(x)\,dx$$

to denote either the Riemann integral or the conditional Riemann integral in cases where we wish for both to apply. The reader should also keep in mind that Riemann integrability and conditional Riemann integrability agree for compact intervals.

3.3.22 Proposition (Algebraic operations and the Riemann integral) *Let* $I \subseteq \mathbb{R}$ *be an interval, let* $f, g\colon I \to \mathbb{R}$ *be (conditionally) Riemann integrable functions, and let* $c \in \mathbb{R}$. *Then the following statements hold:*

(i) $f + g$ *is (conditionally) Riemann integrable and*

$$(C)\int_I (f+g)(x)\,dx = (C)\int_I f(x)\,dx + (C)\int_I g(x)\,dx;$$

(ii) cf *is (conditionally) Riemann integrable and*

$$(C)\int_I (cf)(x)\,dx = c(C)\int_I f(x)\,dx;$$

(iii) if I *is additionally compact, then* fg *is Riemann integrable;*

(iv) if I *is additionally compact and if there exists* $\alpha \in \mathbb{R}_{>0}$ *such that* $g(x) \geq \alpha$ *for each* $x \in I$, *then* $\frac{f}{g}$ *is Riemann integrable.*

Proof (i) We first suppose that $I = [a, b]$ is a compact interval. Let $\epsilon \in \mathbb{R}_{>0}$ and by Theorem 3.3.9 we let P_f and P_g be partitions of $[a, b]$ such that

$$A_+(f, P_f) - A_-(f, P_f) < \tfrac{\epsilon}{2}, \quad A_+(g, P_g) - A_-(g, P_g) < \tfrac{\epsilon}{2},$$

and let P be a partition for which $(x_0, x_1, \ldots, x_k) = \mathrm{EP}(P) = \mathrm{EP}(P_f) \cup \mathrm{EP}(P_g)$. Then, using Proposition 2.2.27,

$$\sup\{f(x) + g(x) \mid x \in [x_{j-1}, x_j]\} = \sup\{f(x) \mid x \in [x_{j-1}, x_j]\} + \sup\{g(x) \mid x \in [x_{j-1}, x_j]\}$$

and

$$\inf\{f(x) + g(x) \mid x \in [x_{j-1}, x_j]\} = \inf\{f(x) \mid x \in [x_{j-1}, x_j]\} + \inf\{g(x) \mid x \in [x_{j-1}, x_j]\}$$

for each $j \in \{1, \ldots, k\}$. Thus

$$A_+(f + g, P) - A_-(f + g, P) \leq A_+(f, P) + A_+(g, P) - A_-(f, P) - A_-(g, P) < \epsilon,$$

using Lemma 1 from the proof of Theorem 3.3.9. This shows that $f + g$ is Riemann integrable by Theorem 3.3.9.

Now let P_f and P_g be any two partitions and let P satisfy $(x_0, x_1, \dots, x_k) = \mathrm{EP}(P) = \mathrm{EP}(P_f) \cup \mathrm{EP}(P_g)$. Then

$$A_+(f, P_f) + A_+(g, P_g) \geq A_+(f, P) + A_+(g, P) \geq A_+(f + g, P) \geq I_+(f + g).$$

We then have

$$I_+(f + g) \leq A_+(f, P_f) + A_+(g, P_g) \quad \Longrightarrow \quad I_+(f + g) \leq I_+(f) + I_+(g).$$

In like fashion we obtain the estimate

$$I_-(f + g) \geq I_-(f) + I_-(g).$$

Combining this gives

$$I_-(f) + I_-(g) \leq I_-(f + g) = I_+(f + g) \leq I_+(f) + I_+(g),$$

which implies equality of these four terms since $I_-(f) = I_+(f)$ and $I_-(g) = I_+(g)$. This gives this part of the result when I is compact. The result follows for general intervals from the definition of the Riemann integral for such intervals, and by applying Proposition 2.3.23.

(ii) As in part (i), the result will follow if we can prove it when I is compact. When $c = 0$ the result is trivial, so suppose that $c \neq 0$. First consider the case $c > 0$. For $\epsilon \in \mathbb{R}_{>0}$ let P be a partition for which $A_+(f, P) - A_-(f, P) < \frac{\epsilon}{c}$. Since $A_-(cf, P) = cA_-(f, P)$ and $A_+(cf, P) = cA_+(f, P)$ (as is easily checked), we have $A_+(cf, P) - A_-(cf, P) < \epsilon$, showing that cf is Riemann integrable. The equalities $A_-(cf, P) = cA_-(f, P)$ and $A_+(cf, P) = cA_+(f, P)$ then directly imply that $I_-(cf) = cI_-(f)$ and $I_+(cf) = cI_+(f)$, giving the result for $c > 0$. For $c < 0$ a similar argument holds, but asking that P be a partition for which $A_+(f, P) - A_-(f, P) < -\frac{\epsilon}{c}$.

(iii) First let us show that if I is compact then f^2 is Riemann integrable if f is Riemann integrable. This, however, follows from Proposition 3.3.23 by taking $g \colon I \to \mathbb{R}$ to be $g(x) = x^2$. To show that a general product fg of Riemann integrable functions on a compact interval is Riemann integrable, we note that

$$fg = \tfrac{1}{2}((f + g)^2 - f^2 - g^2).$$

By part (i) and using the fact that the square of a Riemann integrable function is Riemann integrable, the function on the right is Riemann integrable, so giving the result.

(iv) That $\frac{1}{g}$ is Riemann integrable follows from Proposition 3.3.23 by taking $g \colon I \to \mathbb{R}$ to be $g(x) = \frac{1}{x}$. ■

In parts (iii) and (iv) we asked that the interval be compact. It is simple to find counterexamples which indicate that compactness of the interval is generally necessary (see Exercise 3.3.3).

We now consider the relationship between composition and Riemann integration.

3.3.23 Proposition (Function composition and the Riemann integral) *If* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *is a compact interval, if* $\mathrm{f}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is a Riemann integrable function satisfying* $\mathrm{image}(\mathrm{f}) \subseteq [\mathrm{c}, \mathrm{d}]$*, and if* $\mathrm{g}\colon [\mathrm{c}, \mathrm{d}] \to \mathbb{R}$ *is continuous, then* $\mathrm{g} \circ \mathrm{f}$ *is Riemann integrable.*

Proof Denote $M = \sup\{|g(y)| \mid y \in [c, d]\}$. Let $\epsilon \in \mathbb{R}_{>0}$ and write $\epsilon' = \frac{\epsilon}{2M+d-c}$. Since g is uniformly continuous by the Heine–Cantor Theorem, let $\delta \in \mathbb{R}$ be chosen such that $0 < \delta < \epsilon'$ and such that, $|y_1 - y_2| < \delta$ implies that $|g(y_1) - g(y_2)| < \epsilon'$. Then choose a partition P of $[a, b]$ such that $A_+(f, P) - A_-(f, P) < \delta^2$. Let $(x_0, x_1, \ldots, x_k)$ be the endpoints of P and define

$$
\begin{aligned}
A &= \{j \in \{1, \ldots, k\} \mid \sup\{f(x) \mid x \in [x_{j-1}, x_j]\} - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\} < \delta\},\\
B &= \{j \in \{1, \ldots, k\} \mid \sup\{f(x) \mid x \in [x_{j-1}, x_j]\} - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\} \geq \delta\}.
\end{aligned}
$$

For $j \in A$ we have $|f(\xi_1) - f(\xi_2)| < \delta$ for every $\xi_1, \xi_2 \in [x_{j-1}, x_j]$ which implies that $|g \circ f(\xi_1) - g \circ f(\xi_2)| < \epsilon'$ for every $\xi_1, \xi_2 \in [x_{j-1}, x_j]$. For $j \in B$ we have

$$
\begin{aligned}
\delta \sum_{j \in B} (x_j - x_{j-1}) \leq & \sum_{j \in B} \Big(\sup\{f(x) \mid x \in [x_{j-1}, x_j]\} \\
& \quad - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\} \Big) (x_j - x_{j-1}) \\
& \leq A_+(f, P) - A_-(f, P) < \delta^2.
\end{aligned}
$$

Therefore we conclude that

$$
\sum_{j \in B} (x_j - x_{j-1}) \leq \epsilon'.
$$

Thus

$$
\begin{aligned}
A_+(g \circ f, P) - A_-(g \circ f, P) = & \sum_{j=1}^{k} \Big(\sup\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\} \\
& \quad - \inf\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\} \Big) (x_j - x_{j-1}) \\
= & \sum_{j \in A} \Big(\sup\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\} \\
& \quad - \inf\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\} \Big) (x_j - x_{j-1}) \\
& + \sum_{j \in B} \Big(\sup\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\} \\
& \quad - \inf\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\} \Big) (x_j - x_{j-1}) \\
< & \epsilon'(d - c) + 2\epsilon' M < \epsilon,
\end{aligned}
$$

giving the result by Theorem 3.3.9. ■

The Riemann integral also has the expected properties relative to the partial order and the absolute value function on $\mathbb{R}$.

3.3.24 Proposition (Riemann integral and total order on $\mathbb{R}$) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *be (conditionally) Riemann integrable functions for which* $\mathrm{f}(\mathrm{x}) \leq \mathrm{g}(\mathrm{x})$ *for each* $\mathrm{x} \in \mathrm{I}$*. Then*

$$
(\mathrm{C}) \int_{\mathrm{I}} \mathrm{f}(\mathrm{x})\, \mathrm{dx} \leq (\mathrm{C}) \int_{\mathrm{I}} \mathrm{g}(\mathrm{x})\, \mathrm{dx}.
$$

Proof Note that by part (i) of Proposition 3.3.22 it suffices to take $f = 0$ and then show that $\int_I g(x)\,dx \geq 0$. In the case where $I = [a, b]$ we have

$$\int_a^b g(x)\,dx \geq (b-a)\inf\{g(x) \mid x \in [a,b]\} \geq 0,$$

which gives the result in this case. The result for general intervals follows from the definition, and the fact the a limit of nonnegative numbers is nonnegative. ■

3.3.25 Proposition (Riemann integral and absolute value on $\mathbb{R}$) *Let* I *be an interval, let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$*, and define* $|\mathrm{f}|\colon \mathrm{I} \to \mathbb{R}$ *by* $|\mathrm{f}|(\mathrm{x}) = |\mathrm{f}(\mathrm{x})|$*. Then the following statements hold:*

(i) if f *is Riemann integrable then* $|\mathrm{f}|$ *is Riemann integrable;*

(ii) if I *is compact and if* f *is conditionally Riemann integrable then* $|\mathrm{f}|$ *is conditionally Riemann integrable.*

Moreover, if the hypotheses of either part hold then

$$\left|\int_{\mathrm{I}} \mathrm{f}(\mathrm{x})\,d\mathrm{x}\right| \leq \int_{\mathrm{I}} |\mathrm{f}|(\mathrm{x})\,d\mathrm{x}.$$

Proof (i) If f is Riemann integrable then f_+ and f_- are Riemann integrable. Since $|f| = f_+ + f_-$ it follows from Proposition 3.3.22 that $|f|$ is Riemann integrable.

(ii) When I is compact, the statement follows since conditional Riemann integrability is equivalent to Riemann integrability.

The inequality in the statement of the proposition follows from Proposition 3.3.24 since $f(x) \leq |f(x)|$ for all $x \in I$. ■

We comment that the preceding result is, in fact, not true if one removes the condition that I be compact. We also comment that the converse of the result is false, in that the Riemann integrability of $|f|$ does not imply the Riemann integrability of f. The reader is asked to sort this out in Exercise 3.3.4.

The Riemann integral also behaves well upon breaking an interval into two intervals that are disjoint except for a common endpoint.

3.3.26 Proposition (Breaking the Riemann integral in two) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{I} = \mathrm{I}_1 \cup \mathrm{I}_2$*, where* $\mathrm{I}_1 \cap \mathrm{I}_2 = \{\mathrm{c}\}$*, where* c *is the right endpoint of* I_1 *and the left endpoint of* I_2*. Then* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is (conditionally) Riemann integrable if and only if* $\mathrm{f}|_{\mathrm{I}_1}$ *and* $\mathrm{f}|_{\mathrm{I}_2}$ *are (conditionally) Riemann integrable. Furthermore, we have*

$$(\mathrm{C})\int_{\mathrm{I}} \mathrm{f}(\mathrm{x})\,d\mathrm{x} = (\mathrm{C})\int_{\mathrm{I}_1} \mathrm{f}(\mathrm{x})\,d\mathrm{x} + (\mathrm{C})\int_{\mathrm{I}_2} \mathrm{f}(\mathrm{x})\,d\mathrm{x}.$$

Proof We first consider the case where $I_1 = [a, c]$ and $I_2 = [c, b]$.

Let us suppose that f is Riemann integrable and let $(x_0, x_1, \ldots, x_k)$ be endpoints of a partition of $[a, b]$ for which $A_+(f, P) - A_-(f, P) < \epsilon$. If $c \in (x_0, x_1, \ldots, x_k)$, say $c = x_j$, then we have

$$A_-(f,P) = A_-(f|I_1, P_1) + A_-(f|I_2, P_2), \quad A_+(f,P) = A_+(f|I_1, P_1) + A_+(f|I_2, P_2),$$

where $\mathrm{EP}(P_1) = (x_0, x_1, \dots, x_j)$ are the endpoints of a partition of $[a, c]$ and $\mathrm{EP}(P_2) = (x_j, \dots, x_k)$ is a partition of $[c, b]$. From this we directly deduce that

$$A_+(f|I_1, P_1) - A_-(f|I_1, P_1) < \epsilon, \quad A_+(f|I_2, P_2) - A_-(f|I_2, P_2) < \epsilon. \tag{3.11}$$

If c is not an endpoint of P, then one can construct a new partition P' of $[a, b]$ with c as an extra endpoint. By Lemma 1 of Theorem 3.3.9 we have $A_+(f, P') - A_-(f, P') < \epsilon$. The argument then proceeds as above to show that (3.11) holds. Thus $f|I_1$ and $f|I_2$ are Riemann integrable by Theorem 3.3.9.

To prove the equality of the integrals in the statement of the proposition, we proceed as follows. Let P_1 and P_2 be partitions of I_1 and I_2, respectively. From these construct a partition $P(P_1, P_2)$ of I by asking that $\mathrm{EP}(P(P_1, P_2)) = \mathrm{EP}(P_1) \cup \mathrm{EP}(P_2)$. Then

$$A_+(f|I_1, P_1) + A_+(f|I_2, P_2) = A_+(f, P(P_1, P_2)).$$

Thus

$$\inf\{A_+(f|I_1, P_1) \mid P_1 \in \mathrm{Part}(I_1)\} + \inf\{A_+(f|I_2, P_2) \mid P_2 \in \mathrm{Part}(I_2)\} \geq \inf\{A_+(f, P) \mid P \in \mathrm{Part}(I)\}. \tag{3.12}$$

Now let P be a partition of I and construct partitions $P_1(P)$ and $P_2(P)$ of I_1 and I_2 respectively by adding defining, if necessary, a new partition P' of I with c as the (say) jth endpoint, and then defining $P_1(P)$ such that $\mathrm{EP}(P_1(P))$ are the first $j + 1$ endpoints of P' and then defining $P_2(P)$ such that $\mathrm{EP}(P_2(P))$ are the last $k - j$ endpoints of P'. By Lemma 1 of Theorem 3.3.9 we then have

$$A_+(f, P) \geq A_+(f, P') = A_+(f|I_1, P_1(P)) + A_+(f|I_2, P_2(P)).$$

This gives

$$\inf\{A_+(f, P) \mid P \in \mathrm{Part}(I)\} \geq \inf\{A_+(f|I_1, P_1) \mid P_1 \in \mathrm{Part}(I_1)\} + \inf\{A_+(f|I_2, P_2) \mid P_2 \in \mathrm{Part}(I_2)\}.$$

Combining this with (3.12) gives

$$\inf\{A_+(f, P) \mid P \in \mathrm{Part}(I)\} = \inf\{A_+(f|I_1, P_1) \mid P_1 \in \mathrm{Part}(I_1)\} + \inf\{A_+(f|I_2, P_2) \mid P_2 \in \mathrm{Part}(I_2)\},$$

which is exactly the desired result.

The result for a general interval follows from the general definition of the Riemann integral, and from Proposition 2.3.23. ∎

The next result gives a useful tool for evaluating integrals, as well as a being a result of some fundamental importance.

3.3.27 Proposition (Change of variables for the Riemann integral) *Let* $[a, b]$ *be a compact interval and let* $u\colon [a, b] \to \mathbb{R}$ *be differentiable with* u' *Riemann integrable. Suppose that* $\mathrm{image}(u) \subseteq [c, d]$ *and that* $f\colon [c, d] \to \mathbb{R}$ *is Riemann integrable and that* $f = F'$ *for some differentiable function* $F\colon [c, d] \to \mathbb{R}$. *Then*

$$\int_a^b f \circ u(x) u'(x)\, dx = \int_{u(a)}^{u(b)} f(y)\, dy.$$

Proof Let $G\colon [a,b] \to \mathbb{R}$ be defined by $G = F \circ u$. Then $G' = (f \circ u)u'$ by the Chain Rule. Moreover, G' is Riemann integrable by Propositions 3.3.22 and 3.3.23. Thus, twice using Theorem 3.3.30 below,

$$\int_a^b f \circ u(x)u'(x)\,dx = G(b) - G(a) = F \circ u(b) - F \circ u(a) = \int_{u(a)}^{u(b)} f(y)\,dy,$$

as desired. ■

As a final result in this section, we prove the extremely valuable integration by parts formula.

3.3.28 Proposition (Integration by parts for the Riemann integral) *If* $[a,b]$ *is a compact interval and if* $f, g\colon [a,b] \to \mathbb{R}$ *are differentiable functions with* f' *and* g' *Riemann integrable, then*

$$\int_a^b f(x)g'(x)\,dx + \int_a^b f'(x)g(x)\,dx = f(b)g(b) - f(a)g(a).$$

Proof By Proposition 3.2.10 it holds that fg is differentiable and that $(fg)' = f'g + fg'$. Thus, by Proposition 3.3.22, fg is differentiable with Riemann integrable derivative. Therefore, by Theorem 3.3.30 below,

$$\int_a^b (fg)(x)\,dx = f(b)g(b) - f(a)g(a),$$

and the result follows directly from the formula for the product rule. ■

3.3.6 The Fundamental Theorem of Calculus and the Mean Value Theorems

In this section we begin to explore the sense in which differentiation and integration are inverses of one another. This is, in actuality, and somewhat in contrast to the manner in which one considers this question in introductory calculus courses, a quite complicated matter. Indeed, we will not fully answer this question until Section **??**, after we have some knowledge of the Lebesgue integral. Nevertheless, in this section we give some simple results, and some examples which illustrate the value and the limitations of these results. We also present the Mean Value Theorems for integrals.

The following language is often used in conjunction with the Fundamental Theorem of Calculus.

3.3.29 Definition (Primitive) If $I \subseteq \mathbb{R}$ is an interval and if $f\colon I \to \mathbb{R}$ is a function, a ***primitive*** for f is a function $F\colon I \to \mathbb{R}$ such that $F' = f$. •

Note that primitives are not unique since if one adds a constant to a primitive, the resulting function is again a primitive.

The basic result of this section is the following.

3.3.30 Theorem (Fundamental Theorem of Calculus for Riemann integrals) *For a compact interval* $I = [a, b]$*, the following statements hold:*

(i) if $f\colon I \to \mathbb{R}$ *is Riemann integrable with primitive* $F\colon I \to \mathbb{R}$*, then*

$$\int_a^b f(x)\,dx = F(b) - F(a);$$

(ii) if $f\colon I \to \mathbb{R}$ *is Riemann integrable, and if* $F\colon I \to \mathbb{R}$ *is defined by*

$$F(x) = \int_a^x f(\xi)\,d\xi,$$

then

(a) F *is continuous and*

(b) at each point $x \in I$ *for which* f *is continuous,* F *is differentiable and* $F'(x) = f(x)$.

Proof (i) Let $(P_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of partitions for which $\lim_{j\to\infty}|P_j| = 0$. Denote by $(x_{j,0}, x_{j,1}, \dots, x_{j,k_j})$ the endpoints of P_j, $j \in \mathbb{Z}_{>0}$. By the Mean Value Theorem, for each $j \in \mathbb{Z}_{>0}$ and for each $r \in \{1, \dots, k_r\}$, there exists $\xi_{j,r} \in [x_{j,r-1}, x_{j,r}]$ such that $F(x_{j,r}) - F(x_{j,r-1}) = f(\xi_{j,r})(x_{j,r} - x_{j,r-1})$. Since f is Riemann integrable we have

$$\begin{aligned}\int_a^b f(x)\,dx &= \lim_{j\to\infty} \sum_{r=1}^{k_j} f(\xi_{j,r})(x_{j,r} - x_{j,r-1})\\ &= \lim_{j\to\infty} \sum_{r=1}^{k_j} (F(x_{j,r}) - F(x_{j,r-1}))\\ &= \lim_{j\to\infty} (F(b) - F(a)) = F(b) - F(a),\end{aligned}$$

as desired.

(ii) Let $x \in (a, b)$ and note that, for h sufficiently small,

$$F(x + h) - F(x) = \int_x^{x+h} f(\xi)\,d\xi,$$

using Proposition 3.3.26. By Proposition 3.3.24 it follows that

$$h \inf\{f(y) \mid y \in [a, b]\} \le \int_x^{x+h} f(\xi)\,d\xi \le h \sup\{f(y) \mid y \in [a, b]\},$$

provided that $h > 0$. This shows that

$$\lim_{h\downarrow 0} \int_x^{x+h} f(\xi)\,d\xi = 0.$$

A similar argument can be fashioned for the case when $h < 0$ to show also that

$$\lim_{h\uparrow 0} \int_x^{x+h} f(\xi)\,d\xi = 0,$$

so showing that F is continuous at point in (a, b). A slight modification to this argument shows that F is also continuous at a and b.

Now suppose that f is continuous at x. Let $h > 0$. Again using Proposition 3.3.24 we have

$$h \inf\{f(y) \mid y \in [x, x+h]\} \leq \int_x^{x+h} f(\xi)\,\mathrm{d}\xi \leq h \sup\{f(y) \mid y \in [x, x+h]\}$$
$$\implies \quad \inf\{f(y) \mid y \in [x, x+h]\} \leq \frac{F(x+h) - F(x)}{h} \leq \sup\{f(y) \mid y \in [x, x+h]\}.$$

Continuity of f at x gives

$$\lim_{h \downarrow 0} \inf\{f(y) \mid y \in [x, x+h]\} = f(x), \quad \lim_{h \downarrow 0} \sup\{f(y) \mid y \in [x, x+h]\} = f(x).$$

Therefore,

$$\lim_{h \downarrow 0} \frac{F(x+h) - F(x)}{h} = f(x).$$

A similar argument can be made for $h < 0$ to give

$$\lim_{h \uparrow 0} \frac{F(x+h) - F(x)}{h} = f(x),$$

so proving this part of the theorem. ■

Let us give some examples that illustrate what the Fundamental Theorem of Calculus says and does not say.

3.3.31 Examples (Fundamental Theorem of Calculus)

1. Let $I = [0, 1]$ and define $f : I \to \mathbb{R}$ by

$$f(x) = \begin{cases} x, & x \in [0, \frac{1}{2}], \\ 1 - x, & x \in (\frac{1}{2}, 1]. \end{cases}$$

Then

$$F(x) \triangleq \int_0^x f(\xi)\,\mathrm{d}\xi = \begin{cases} \frac{1}{2}x^2, & x \in [0, \frac{1}{2}], \\ -\frac{1}{2}x^2 + x - \frac{1}{8}, & x \in (\frac{1}{2}, 1]. \end{cases}$$

Then, for any $x \in [a, b]$, we see that

$$\int_0^x f(\xi)\,\mathrm{d}\xi = F(x) - F(0).$$

This is consistent with part (i) of Theorem 3.3.30, whose hypotheses apply since f is continuous, and so Riemann integrable.

2. Let $I = [0, 1]$ and define $f : I \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1, & x \in [0, \frac{1}{2}], \\ -1, & x \in (\frac{1}{2}, 1]. \end{cases}$$

Then

$$F(x) \triangleq \int_0^x f(\xi)\,\mathrm{d}\xi = \begin{cases} x, & x \in [0, \frac{1}{2}], \\ 1 - x, & x \in (\frac{1}{2}, 1]. \end{cases}$$

Then, for any $x \in [a, b]$, we see that

$$\int_0^x f(\xi)\,\mathrm{d}\xi = F(x) - F(0).$$

In this case, we have the conclusions of part (i) of Theorem 3.3.30, and indeed the hypotheses hold, since f is Riemann integrable.

3. Let I and f be as in Example 1 above. Then f is Riemann integrable, and we see that F is continuous, as per part (ii) of Theorem 3.3.30, and that F is differentiable, also as per part (ii) of Theorem 3.3.30.
4. Let I and f be as in Example 2 above. Then f is Riemann integrable, and we see that F is continuous, as per part (ii) of Theorem 3.3.30. However, f is not continuous at $x = \frac{1}{2}$, and we see that, correspondingly, F is not differentiable at $x = \frac{1}{2}$.
5. The next example we consider is one with which, at this point, we can only be sketchy about the details. Consider the Cantor function $f_C \colon [0, 1] \to \mathbb{R}$ of Example 3.2.27. Note that f_C' is defined and equal to zero, except at points in the Cantor set C; thus except at points forming a set of measure zero. It will be clear when we discuss the Lebesgue integral in Section **??** that this ensures that $\int_0^x f_C'(\xi)\,\mathrm{d}\xi = 0$ for every $x \in [0, 1]$, where the integral in this case is the Lebesgue integral. (By defining f_C' arbitrarily on C, we can also use the Riemann integral by virtue of Theorem 3.3.11.) This shows that the conclusions of part (i) of Theorem 3.3.30 can fail to hold, even when the derivative of F is defined almost everywhere.
6. The last example we give is the most significant, in some sense, and is also the most complicated. The example we give is of a function $F \colon [0, 1] \to \mathbb{R}$ that is differentiable with bounded derivative, but whose derivative $f = F'$ is not Riemann integrable. Thus f possesses a primitive, but is not Riemann integrable.
   To define F, let $G \colon \mathbb{R}_{>0} \to \mathbb{R}$ be the function

$$G(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

   For $c > 0$ let $x_c > 0$ be defined by

$$x_c = \sup\{x \in \mathbb{R}_{>0} \mid G'(x) = 0,\ x \leq c\},$$

   and define $G_c \colon (0, c] \to \mathbb{R}$ by

$$G_c(x) = \begin{cases} G(x), & x \in (0, x_c], \\ G(x_c), & x \in (x_c, x]. \end{cases}$$

Now, for $\epsilon \in (0, \frac{1}{2})$, let $C_\epsilon \subseteq [0,1]$ be a fat Cantor set as constructed in Example 2.5.42. Define F as follows. If $x \in C_\epsilon$ we take $F(x) = 0$. If $x \notin C_\epsilon$, then, since C_ϵ is closed, by Proposition 2.5.6 x lies in some open interval, say (a, b). Then take $c = \frac{1}{2}(b-a)$ and define

$$F(x) = \begin{cases} G_c(x-a), & x \in (a, \frac{1}{2}(a+b)), \\ G_c(b-x), & x \in [\frac{1}{2}(a+b), b). \end{cases}$$

Note that $F|(a,b)$ is designed so that its derivative will oscillate wildly in the limit as the endpoints of (a,b) are approached, but be nicely behaved at all points in (a,b). This is, as we shall see, the key feature of F.
Let us record some properties of F in a sequence of lemmata.

1 Lemma *If* x ∈ C_ϵ, *then* F *is differentiable at* x *and* F′(x) = 0.

Proof Let $y \in [0,1] \setminus \{x\}$. If $y \in C_\epsilon$ then

$$\frac{f(y) - f(x)}{y - x} = 0.$$

If $y \notin C_\epsilon$, then y must lie in an open interval, say (a,b). Let d be the endpoint of (a,b) nearest y and let $c = \frac{1}{2}(b-a)$. Then

$$\begin{aligned} \left|\frac{f(y) - f(x)}{y - x}\right| = \frac{f(y)}{y - x} &\leq \frac{f(y)}{y - d} = \frac{G_c(|y-d|)}{y-d} \\ &\leq \frac{|y-d|^2}{y-d} = |y-d| \leq |y-x|. \end{aligned}$$

Thus

$$\lim_{y \to x} \frac{f(y) - f(x)}{y - x} = 0,$$

giving the lemma. ▼

2 Lemma *If* x ∉ C_ϵ, *then* F *is differentiable at* x *and* |F′(x)| ≤ 3.

Proof By definition of F for points not in C_ϵ we have

$$|F'(x)| \leq \left|2y \sin \tfrac{1}{y} - \cos \tfrac{1}{y}\right| \leq 3,$$

for some $y \in [0,1]$. ▼

3 Lemma $C_\epsilon \subseteq D_{F'}$.

Proof By construction of C_ϵ, if $x \in C_\epsilon$ then there exists a sequence $((a_j, b_j))_{j\in\mathbb{Z}_{>0}}$ of open intervals in $[0,1] \setminus C_\epsilon$ having the property that $\lim_{j\to\infty} a_j = \lim_{j\to\infty} b_j = x$. Note that $\limsup_{y\downarrow 0} g'(y) = 1$. Therefore, by the definition of F on the open intervals (a_j, b_j), $j \in \mathbb{Z}_{>0}$, it holds that $\limsup_{y\downarrow a_j} F'(y) = \limsup_{y\uparrow b_j} F'(y) = 1$. Therefore, $\limsup_{y\to x} F'(y) = 1$. Since $F'(x) = 0$, it follows that F' is discontinuous at x. ▼

Since F' is discontinuous at all points in C_ϵ, and since C_ϵ does not have measure zero, it follows from Theorem 3.3.11 that F' is not Riemann integrable. Therefore, the function $f = F'$ possesses a primitive, namely F, but is not Riemann integrable. •

Finally we state two results that, like the Mean Value Theorem for differentiable functions, relate the integral to the values of a function.

3.3.32 Proposition (First Mean Value Theorem for Riemann integrals) *Let* $[a,b]$ *be a compact interval and let* $f, g\colon [a,b] \to \mathbb{R}$ *be functions with* f *continuous and with* g *nonnegative and Riemann integrable. Then there exists* $c \in [a,b]$ *such that*

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx$$

Proof Let

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

Since g is nonnegative we have

$$mg(x) \le f(x)g(x) \le Mg(x), \qquad x \in [a,b],$$

from which we deduce that

$$m\int_a^b g(x)\,dx \le \int_a^b f(x)g(x)\,dx \le M\int_a^b g(x)\,dx.$$

Continuity of f and the Intermediate Value Theorem gives $c \in [a,b]$ such that the result holds. ■

3.3.33 Proposition (Second Mean Value Theorem for Riemann integrals) *Let* $[a,b]$ *be a compact interval and let* $f, g\colon [a,b] \to \mathbb{R}$ *be functions with*

(i) g *Riemann integrable and having the property that there exists* G *such that* $g = G'$, *and*

(ii) f *differentiable with Riemann integrable, nonnegative derivative.*

Then there exists $c \in [a,b]$ *so that*

$$\int_a^b f(x)g(x)\,dx = f(a)\int_a^c g(x)\,dx + f(b)\int_c^b g(x)\,dx.$$

Proof Without loss of generality we may suppose that

$$G(x) = \int_a^x g(\xi)\,\mathrm{d}\xi,$$

since all we require is that $G' = g$. We then compute

$$\begin{aligned}\int_a^b f(x)g(x)\,\mathrm{d}x &= \int_a^b f(x)G'(x)\,\mathrm{d}x = f(b)G(b) - \int_a^b f'(x)G(x)\,\mathrm{d}x\\ &= f(b)G(b) - G(c)\int_a^b f'(x)\,\mathrm{d}x,\end{aligned}$$

for some $c \in [a,b]$, using integration by parts and Proposition 3.3.32. Now using Theorem 3.3.30,

$$\int_a^b f(x)g(x)\,\mathrm{d}x = f(b)G(b) - G(c)(f(b) - f(a)),$$

which gives the desired result after using the definition of G and after some rearrangement. ■

3.3.7 The Cauchy principal value

In Example 3.3.17 we explored some of the nuances of the improper Riemann integral. There we saw that for integrals that are defined using limits, one often needs to make the definitions in a particular way. The principal value integral is intended to relax this, and enable one to have a meaningful notion of the integral in cases where otherwise one might not. To motivate our discussion we consider an example.

3.3.34 Example Let $I = [-1,2]$ and consider the function $f : I \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} \frac{1}{x}, & x \neq 0\\ 0, & \text{otherwise.}\end{cases}$$

This function has a singularity at $x = 0$, and the integral $\int_{-1}^2 f(x)\,\mathrm{d}x$ is actually divergent. However, for $\epsilon \in \mathbb{R}_{>0}$ note that

$$\int_{-1}^{-\epsilon} \frac{1}{x}\,\mathrm{d}x + \int_{\epsilon}^2 \frac{1}{x}\,\mathrm{d}x = -\log x|_{\epsilon}^1 + \log x|_{\epsilon}^2 = \log 2.$$

Thus we can devise a way around the singularity in this case, the reason being that the singular behaviour of the function on either side of the function "cancels" that on the other side. •

With this as motivation, we give a definition.

3.3.35 Definition (Cauchy principal value) Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$ be a function. Denote $a = \inf I$ and $b = \sup I$, allowing that $a = -\infty$ and $b = \infty$.

(i) If, for $x_0 \in \operatorname{int}(I)$, there exists $\epsilon_0 \in \mathbb{R}_{>0}$ such that the functions $f|(a, x_0 - \epsilon]$ and $f|[x_0 + \epsilon, b)$ are Riemann integrable for all $\epsilon \in (0, \epsilon_0]$, then the ***Cauchy principal value*** for f is defined by

$$\operatorname{pv} \int_I f(x)\,\mathrm{d}x = \lim_{\epsilon \to 0} \Big(\int_a^{x_0 - \epsilon} f(x)\,\mathrm{d}x + \int_{x_0 + \epsilon}^b f(x)\,\mathrm{d}x \Big).$$

(ii) If $a = -\infty$ and $b = \infty$ and if for each $R \in \mathbb{R}_{>0}$ the function $f|[-R, R]$ is Riemann integrable, then the ***Cauchy principal value*** for f is defined by

$$\operatorname{pv} \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \lim_{R \to \infty} \int_{-R}^{R} f(x)\,\mathrm{d}x. \qquad \bullet$$

3.3.36 Remarks

1. If f is Riemann integrable on I then the Cauchy principal value is equal to the Riemann integral.
2. The Cauchy principal value is allowed to be infinite by the preceding definition, as the following examples will show.
3. It is not standard to define the Cauchy principal value in part (ii) of the definition. In many texts where the Cauchy principal value is spoken of, it is part (i) that is being used. However, we will find the definition from part (ii) useful. $\bullet$

3.3.37 Examples (Cauchy principal value)

1. For the example of Example 3.3.34 we have
$$\operatorname{pv} \int_{-1}^{2} \frac{1}{x}\,\mathrm{d}x = \log 2.$$
2. For $I = \mathbb{R}$ and $f(x) = x(1 + x^2)^{-1}$ we have
$$\operatorname{pv} \int_{-\infty}^{\infty} \frac{x}{1 + x^2}\,\mathrm{d}x = \lim_{R \to \infty} \int_{-R}^{R} \frac{x}{1 + x^2}\,\mathrm{d}x = \lim_{R \to \infty} \Big(\frac{1}{2} \log(1 + R^2) - \frac{1}{2} \log(1 + R^2) \Big) = 0.$$
Note that in Example 3.3.17–4 we showed that this function was not Riemann integrable.
3. Next we consider $I = \mathbb{R}$ and $f(x) = |x|(1 + x^2)$. In this case we compute
$$\operatorname{pv} \int_{-\infty}^{\infty} \frac{|x|}{1 + x^2}\,\mathrm{d}x = \lim_{R \to \infty} \int_{-R}^{R} \frac{|x|}{1 + x^2}\,\mathrm{d}x = \lim_{R \to \infty} \Big(\frac{1}{2} \log(1 + R^2) + \frac{1}{2} \log(1 + R^2) \Big) = \infty.$$
We see then that there is no reason why the Cauchy principal value may not be infinite. $\bullet$

3.3.8 Notes

The definition we give for the Riemann integral is actually that used by Darboux, and the condition given in part (iii) of Theorem 3.3.9 is the original definition of Riemann. What Darboux showed was that the two definitions are equivalent. It is not uncommon to instead use the Darboux definition as the standard definition because, unlike the definition of Riemann, it does not rely on an arbitrary selection of a point from each of the intervals forming a partition.

Exercises

3.3.1 Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function that is Riemann integrable and satisfies $f(x) \geq 0$ for all $x \in I$. Show that $\int_I f(x)\,\mathrm{d}x \geq 0$.

3.3.2 Let $I \subseteq \mathbb{R}$ be an interval, let $f, g\colon I \to \mathbb{R}$ be functions, and define $D_{f,g} = \{x \in I \mid f(x) \neq g(x)\}$.

(a) Show that, if $D_{f,g}$ is finite and f is Riemann integrable, then g is Riemann integrable and $\int_I f(x)\,\mathrm{d}x = \int_I g(x)\,\mathrm{d}x$.

(b) Is it true that, if $D_{f,g}$ is countable and f is Riemann integrable, then g is Riemann integrable and $\int_I f(x)\,\mathrm{d}x = \int_I g(x)\,\mathrm{d}x$? If it is true, give a proof; if it is not true, give a counterexample.

3.3.3 Do the following:

(a) find an interval I and functions $f, g\colon I \to \mathbb{R}$ such that f and g are both Riemann integrable, but fg is not Riemann integrable;

(b) find an interval I and functions $f, g\colon I \to \mathbb{R}$ such that f and g are both Riemann integrable, but $g \circ f$ is not Riemann integrable.

3.3.4 Do the following:

(a) find an interval I and a conditionally Riemann integrable function $f\colon I \to \mathbb{R}$ such that $|f|$ is not Riemann integrable;

(b) find a function $f\colon [0,1] \to \mathbb{R}$ such that $|f|$ is Riemann integrable, but f is not Riemann integrable.

3.3.5 Show that, if $f\colon [a,b] \to \mathbb{R}$ is continuous, then there exists $c \in [a,b]$ such that

$$\int_a^b f(x)\,\mathrm{d}x = f(c)(b-a).$$

Section 3.4

Sequences and series of $\mathbb{R}$-valued functions

In this section we present for the first time the important topic of sequences and series of functions and their convergence. One of the reasons why convergence of sequences of functions is important is that is allows us to classify sets of functions. The idea of classifying sets of functions according to their possessing certain properties leads to the general idea of a "function space." Function spaces are important to understand when developing any systematic theory dealing with functions, since sets of general functions are simply too unstructured to allow much useful to be said. On the other hand, if one restricts the set of functions in the wrong way (e.g., by asking that they all be continuous), then one can end of with a framework with unpleasant properties. But this is getting a little ahead of the issue directly at hand, which is to consider convergence of sequences of functions.

Do I need to read this section? The material in this section is basic, particularly the concepts of pointwise convergence and uniform convergence and the distinction between them. However, it is possible to avoid reading this section until the material becomes necessary, as it will in Chapters **??**, **??**, **??**, and **??**, for example. •

3.4.1 Pointwise convergent sequences

The first type of convergence we deal with is probably what a typical first-year student, at least the rare one who understood convergence for summations of numbers, would proffer as a good candidate for convergence. As we shall see, it often leaves something to be desired.

In the discussion of pointwise convergence, one needs no assumptions on the character of the functions, as one is essentially talking about convergence of numbers.

3.4.1 Definition (Pointwise convergence of sequences) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of $\mathbb{R}$-valued functions on I.

(i) The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ ***converges pointwise*** to a function $f : I \to \mathbb{R}$ if, for each $x \in I$ and for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f(x) - f_j(x)| < \epsilon$ provided that $j \geq N$.

(ii) The function f in the preceding part of the definition is the ***limit function*** for the sequence.

(iii) The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ is ***pointwise Cauchy*** if, for each $x \in I$ and for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f_k(x)| < \epsilon$ provided that $j, k \geq N$.

•

Let us immediately establish the equivalence of pointwise convergent and pointwise Cauchy sequences. As is clear in the proof of the following result, the key fact is completeness of $\mathbb{R}$.

3.4.2 Theorem (Pointwise convergent equals pointwise Cauchy) *If $I \subseteq \mathbb{R}$ is an interval and if $(f_j)_{j\in\mathbb{Z}_{>0}}$ is a sequence of $\mathbb{R}$-valued functions on I then the following statements are equivalent:*

(i) there exists a function $f\colon I \to \mathbb{R}$ such that $(f_j)_{j\in\mathbb{Z}_{>0}}$ converges pointwise to f;

(ii) $(f_j)_{j\in\mathbb{Z}_{>0}}$ is pointwise Cauchy.

Proof This merely follows from the following facts.

1. If the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ converges to $f(x)$ then the sequence is Cauchy by Proposition 2.3.3.
2. If the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is Cauchy then there exists a number $f(x) \in \mathbb{R}$ such that $\lim_{j\to\infty} f_j(x) = f(x)$ by Theorem 2.3.5. ■

Based on the preceding theorem we shall switch freely between the notions of pointwise convergent and pointwise Cauchy sequences of functions.

Pointwise convergence is essentially the most natural form of convergence for a sequence of functions in that it depends in a trivial way on the basic notion of convergence of sequences in $\mathbb{R}$. However, as we shall see later in this section, and in Chapters **??** and **??**, other forms of convergence of often more useful.

3.4.3 Example (Pointwise convergence) Consider the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ of $\mathbb{R}$-valued functions defined on $[0, 1]$ by

$$f_j(x) = \begin{cases} 1, & x \in [0, \frac{1}{j}], \\ 0, & x \in (\frac{1}{j}, 1]. \end{cases}$$

Note that $f_j(0) = 1$ for every $j \in \mathbb{Z}_{>0}$, so that the sequence $(f_j(0))_{j\in\mathbb{Z}_{>0}}$ converges, trivially, to 1. For any $x_0 \in (0, 1]$, provided that $j > x_0^{-1}$, then $f_j(x_0) = 0$. Thus $(f_j(x_0))_{j\in\mathbb{Z}_{>0}}$ converges, as a sequence of real numbers, to 0 for each $x_0 \in (0, 1]$. Thus this sequence converges pointwise, and the limit function is

$$f(x) = \begin{cases} 1, & x = 0, \\ 0, & x \in (0, 1]. \end{cases}$$

If N is the smallest natural number with the property that $N > x_0^{-1}$, then we observe, trivially, that this number does indeed depend on x_0. As x_0 gets closer and closer to 0 we have to wait longer and longer in the sequence $(f_j(x_0))_{j\in\mathbb{Z}_{>0}}$ for the arrival of zero. •

3.4.2 Uniformly convergent sequences

Let us first say what we mean by uniform convergence.

3.4.4 Definition (Uniform convergence of sequences) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of $\mathbb{R}$-valued functions on I.

(i) The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ ***converges uniformly*** to a function $f\colon I \to \mathbb{R}$ if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f(x) - f_j(x)| < \epsilon$ for all $x \in I$, provided that $j \geq N$.

(ii) The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ is ***uniformly Cauchy*** if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f_k(x)| < \epsilon$ for all $x \in I$, provided that $j, k \geq N$. •

Let us immediately give the equivalence of the preceding notions of convergence.

3.4.5 Theorem (Uniformly convergent equals uniformly Cauchy) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$ *and a sequence of* $\mathbb{R}$*-valued functions* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *on* I *the following statements are equivalent:*

(i) there exists a function $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *such that* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *converges uniformly to* f*;*

(ii) $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *is uniformly Cauchy.*

Proof First suppose that $(f_j)_{j\in\mathbb{Z}_{>0}}$ is uniformly Cauchy. Then, for each $x \in I$ the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is Cauchy and so by Theorem 2.3.5 converges to a number that we denote by $f(x)$. This defines the function $f\colon I \to \mathbb{R}$ to which the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ converges pointwise. Let $\epsilon \in \mathbb{R}_{>0}$ and let $N_1 \in \mathbb{Z}_{>0}$ have the property that $|f_j(x)-f_k(x)| < \frac{\epsilon}{2}$ for $j, k \geq N_1$ and for each $x \in I$. Now let $x \in I$ and let $N_2 \in \mathbb{Z}_{>0}$ have the property that $|f_k(x) - f(x)| < \frac{\epsilon}{2}$ for $k \geq N_2$. Then, for $j \geq N_1$, we compute

$$|f_j(x) - f(x)| \leq |f_j(x) - f_k(x)| + |f_k(x) - f(x)| < \epsilon,$$

where $k \geq \max\{N_1, N_2\}$, giving the first implication.

Now suppose that, for $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f(x)| < \epsilon$ for all $j \geq N$ and for all $x \in I$. Then, for $\epsilon \in \mathbb{R}_{>0}$ let $N \in \mathbb{Z}_{>0}$ satisfy $|f_j(x) - f(x)| < \frac{\epsilon}{2}$ for $j \geq N$ and $x \in I$. Then, for $j, k \geq N$ and for $x \in I$, we have

$$|f_j(x) - f_k(x)| \leq |f_j(x) - f(x)| + |f_k(x) - f(x)| < \epsilon,$$

giving the sequence as uniformly Cauchy. ■

Compare this definition to that for pointwise convergence. They sound similar, but there is a fundamental difference. For pointwise convergence, the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is examined separately for convergence at each value of x. As a consequence of this, the value of N might depend on both ϵ and x. For uniform convergence, however, we ask that for a given ϵ, the convergence is tested over all of I. In Figure 3.11 we depict the idea behind uniform convergence. The distinction between uniform and pointwise convergence is subtle on a first encounter, and it is sometimes difficult to believe that pointwise convergence is possible without uniform convergence. However, this is indeed the case, and an example illustrates this readily.

3.4.6 Example (Uniform convergence) On $[0, 1]$ we consider the sequence of $\mathbb{R}$-valued functions defined by

$$f_j(x) = \begin{cases} 2jx, & x \in [0, \frac{1}{2j}], \\ -2jx + 2, & x \in (\frac{1}{2j}, \frac{1}{j}], \\ 0, & x \in (\frac{1}{j}, 1]. \end{cases}$$

In Figure 3.12 we graph f_j for $j \in \{1, 3, 10, 50\}$. The astute reader will see the point, but let's go through it just to make sure we see how this works.

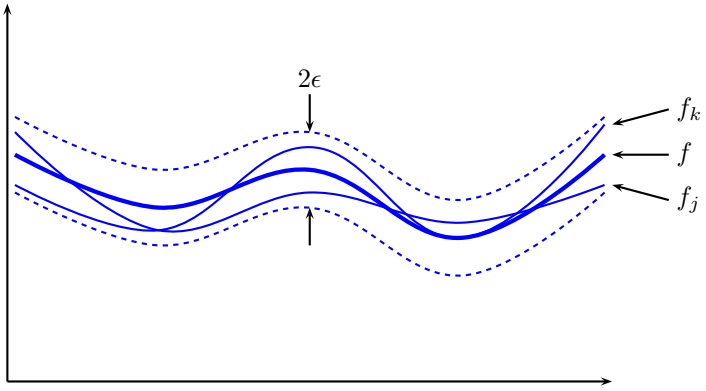


Figure 3.11 The idea behind uniform convergence

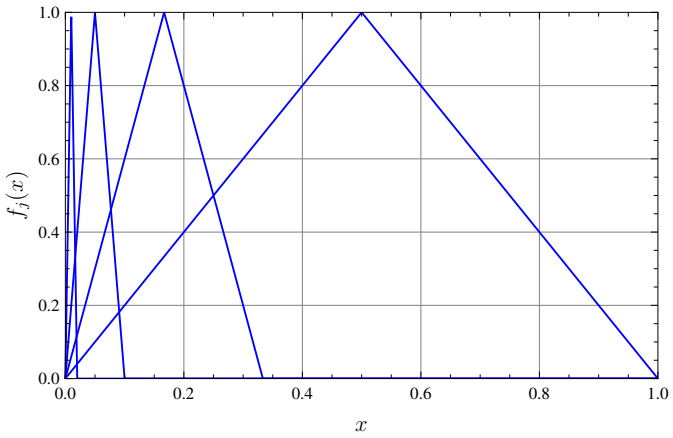


Figure 3.12 A sequence of functions converging pointwise, but not uniformly

First of all, we claim that the sequence converges pointwise to the limit function $f(x) = 0$, $x \in [0,1]$. Since $f_j(0) = 0$ for all $j \in \mathbb{Z}_{>0}$, obviously the sequence converges to 0 at $x = 0$. For $x \in (0,1]$, if $N \in \mathbb{Z}_{>0}$ satisfies $\frac{1}{N} < x$ then we have $f_j(x) = 0$ for $j \geq N$. Thus we do indeed have pointwise convergence.

We also claim that the sequence does not converge uniformly. Indeed, for any positive $\epsilon < 1$, we see that $f_j(\frac{1}{2j}) = 1 > \epsilon$ for every $j \in \mathbb{Z}_{>0}$. This prohibits our asserting the existence of $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f_k(x)| < \epsilon$ for every $x \in [0,1]$, provided that $j, k \geq N$. Thus convergence is indeed not uniform. •

As we say, this is perhaps subtle, at least until one comes to grips with, after which point it makes perfect sense. You should not stop thinking about this until it makes perfect sense. If you overlook this distinction between pointwise and uniform convergence, you will be missing one of the most important topics in the theory of frequency representations of signals.

3.4.7 Remark (On "uniformly" again) In Remark 3.1.6 we made some comments on

the notion of what is meant by "uniformly." Let us reinforce this here. In Definition 3.1.5 we introduced the notion of uniform continuity, which meant that the "δ" could be chosen so as to be valid on the entire domain. Here, with uniform convergence, the idea is that "N" can be chosen to be valid on the entire domain. Similar uses will occasionally be made of the word "uniformly" throughout the text, and it is hoped that the meaning should be clear from the context. •

Now we prove an important result concerning uniform convergence. The significance of this result is perhaps best recognised in a more general setting, such as that of Theorem **??**, where the idea of completeness is clear. However, even in the simple setting of our present discussion, the result is important enough.

3.4.8 Theorem (The uniform limit of bounded, continuous functions is bounded and continuous) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval with* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *a sequence of continuous bounded functions on* I *that converge uniformly. Then the limit function is continuous and bounded. In particular, a uniformly convergent sequence of continuous functions defined on a compact interval converges to a continuous limit function.*

Proof Let $x \in I$ define $f(x) = \lim_{j\to\infty} f_j(x)$. This pointwise limit exists since $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ (why?). We first claim that f is bounded. To see this, for $\epsilon \in \mathbb{R}_{>0}$, let $N \in \mathbb{Z}_{>0}$ have the property that $|f(x) - f_N(x)| < \epsilon$ for every $x \in I$. Then

$$|f(x)| \leq |f(x) - f_N(x)| + |f_N(x)| \leq \epsilon + \sup\{f_N(x) \mid x \in I\}.$$

Since the expression on the right is independent of x, this gives the desired boundedness of f.

Now we prove that the limit function f is continuous. Since $(f_j)_{j\in\mathbb{Z}_{>0}}$ is uniformly convergent, for any $\epsilon \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f(x)| < \frac{\epsilon}{3}$ for all $x \in I$ and $j \geq N$. Now fix $x_0 \in I$, and consider the $N \in \mathbb{Z}_{>0}$ just defined. By continuity of f_N, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $x \in I$ satisfies $|x - x_0| < \delta$, then $|f_N(x) - f_N(x_0)| < \frac{\epsilon}{3}$. Then, for $x \in I$ satisfying $|x - x_0| < \delta$, we have

$$\begin{aligned} |f(x) - f(x_0)| &= |(f(x) - f_N(x)) + (f_N(x) - f_N(x_0)) + (f_N(x_0) - f(x_0))| \\ &\leq |f(x) - f_N(x)| + |f_N(x) - f_N(x_0)| + |f_N(x_0) - f(x_0)| \\ &< \tfrac{\epsilon}{3} + \tfrac{\epsilon}{3} + \tfrac{\epsilon}{3} = \epsilon, \end{aligned}$$

where we have again used the triangle inequality. Since this argument is valid for any $x_0 \in I$, it follows that f is continuous. ■

Note that the hypothesis that the functions be bounded is essential for the conclusions to hold. As we shall see, the contrapositive of this result is often helpful. That is, it is useful to remember that if a sequence of continuous functions defined on a closed bounded interval converges to a *dis*continuous limit function, then the convergence is *not* uniform.

3.4.3 Dominated and bounded convergent sequences

Bounded convergence is a notion that is particularly useful when discussing convergence of function sequences on noncompact intervals.

3.4.9 Definition (Dominated and bounded convergence of sequences) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of $\mathbb{R}$-valued functions on I. For a function $g\colon I \to \mathbb{R}_{>0}$, the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ ***converges dominated by*** **g** if

(i) $f_j(x) \leq g(x)$ for every $j \in \mathbb{Z}_{>0}$ and for every $x \in I$ and

(ii) if, for each $x \in I$ and for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f_k(x)| < \epsilon$ for $j, k \geq N$.

If, moreover, g is a constant function, then a sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ that converges dominated by g ***converges boundedly***. •

It is clear that dominated convergence implies pointwise convergence. Indeed, bounded convergence is merely pointwise convergence with the extra hypothesis that all functions be bounded by the same positive function.

Let us give some examples that distinguish between the notions of convergence we have.

3.4.10 Examples (Pointwise, bounded, and uniform convergence)

1. The sequence of functions in Example 3.4.3 converges pointwise, boundedly, but not uniformly.
2. The sequence of functions in Example 3.4.6 converges pointwise, boundedly, but not uniformly.
3. Consider now a new sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ defined on $I = [0, 1]$ by

$$f_j(x) = \begin{cases} 2j^2x, & x \in [0, \frac{1}{2j}], \\ -2j^2x + 2j, & x \in (\frac{1}{2j}, \frac{1}{j}], \\ 0, & \text{otherwise.} \end{cases}$$

A few members of the sequence are shown in Figure 3.13. This sequence

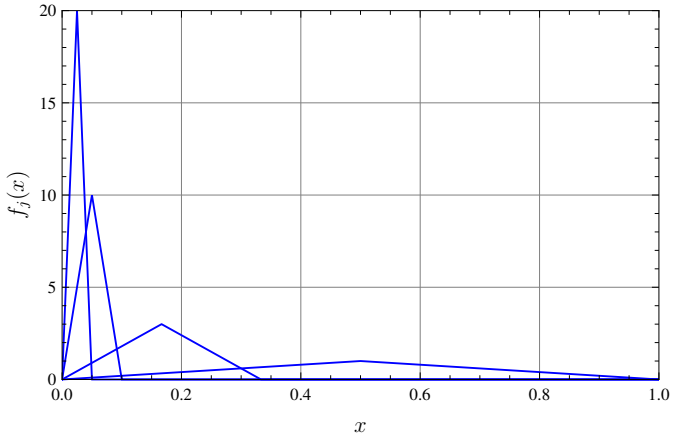


Figure 3.13 A sequence converging pointwise but not boundedly (shown are f_j, $j \in \{1, 5, 10, 20\}$)

converges pointwise to the zero function. Moreover, one can easily check that the convergence is dominated by the function $g\colon [0,1] \to \mathbb{R}$ defined by

$$g(x) = \begin{cases} \frac{1}{x}, & x \in (0,1], \\ 1, & x = 0. \end{cases}$$

The sequence converges neither boundedly nor uniformly.

4. On $I = \mathbb{R}$ consider the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ defined by $f_j(x) = x^2 + \frac{1}{j}$. This sequence clearly converges uniformly to $f\colon x \mapsto x^2$. However, it does not converge boundedly. Of course, the reason is simply that f is itself not bounded. We shall see that uniform convergence to a bounded function implies bounded convergence, in a certain sense. •

We have the following relationship between uniform and bounded convergence.

3.4.11 Proposition (Relationship between uniform and bounded convergence) *If a sequence* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *defined on an interval* I *converges uniformly to a bounded function* f*, then there exists* $\mathrm{N} \in \mathbb{Z}_{>0}$ *such that the sequence* $(\mathrm{f_{N+j}})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *converges boundedly to* f*.*

Proof Let $M \in \mathbb{R}_{>0}$ have the property that $|f(x)| < \frac{M}{2}$ for each $x \in I$. Since $(f_j)_{j\in\mathbb{Z}_{>0}}$ converges uniformly to f there exists $N \in \mathbb{Z}_{>0}$ such that $|f(x) - f_j(x)| < \frac{M}{2}$ for all $x \in I$ and for $j > N$. It then follows that

$$|f_j(x)| \leq |f(x) - f_j(x)| + |f(x)| < M$$

provided that $j > N$. From this the result follows since pointwise convergence of $(f_j)_{j\in\mathbb{Z}_{>0}}$ to f implies pointwise convergence of $(f_{N+j})_{j\in\mathbb{Z}_{>0}}$ to f. ■

3.4.4 Series of $\mathbb{R}$-valued functions

In the previous sections we considered the general matter of sequences of functions. Of course, this discussion carries over to ***series*** of functions, by which we mean expressions of the form $S(x) = \sum_{j=1}^{\infty} f_j(x)$. This is done in the usual manner by considering the partial sums. Let us do this formally.

3.4.12 Definition (Convergence of series) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of $\mathbb{R}$-valued functions on I. Let $F(x) = \sum_{j=1}^{\infty} f_j(x)$ be a series. The corresponding sequence of ***partial sums*** is the sequence $(F_k)_{k\in\mathbb{Z}_{>0}}$ of $\mathbb{R}$-valued functions on I defined by

$$S_k(x) = \sum_{j=1}^{k} f_j(x).$$

Let $g\colon I \to \mathbb{R}_{>0}$. The series:

(i) ***converges pointwise*** if the sequence of partial sums converges pointwise;

(ii) ***converges uniformly*** if the sequence of partial sums converges uniformly;

(iii) ***converges dominated by*** **g** if the sequence of partial sums converges dominated by g;

(iv) ***converges boundedly*** if the sequence of partial sums converges boundedly. •

A fairly simple extension of pointwise convergence of series is the following notion which is unique to series (as opposed to sequences).

3.4.13 Definition (Absolute convergence of series) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of $\mathbb{R}$-valued functions on I. The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ ***converges absolutely*** if, for each $x \in I$ and for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $||f_j(x)| - |f_k(x)|| < \epsilon$ provided that $j, k \geq N$. •

Thus an absolutely convergent sequence is one where, for each $x \in I$, the sequence $(|f_j(x)|)_{j\in\mathbb{Z}_{>0}}$ is Cauchy, and hence convergent. In other words, for each $x \in I$, the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is absolutely convergent. It is clear, then, that an absolutely convergent sequence of functions is pointwise convergent. Let us give some examples that illustrate the difference between pointwise and absolute convergence.

3.4.14 Examples (Absolute convergence)

1. The sequence of functions of Example 3.4.3 converges absolutely since the functions all take positive values.
2. For $j \in \mathbb{Z}_{>0}$, define $f_j\colon [0,1] \to \mathbb{R}$ by $f_j(x) = \frac{(-1)^{j+1}x}{j}$. Then, by Example 2.4.2–3, the series $S(x) = \sum_{j=1}^{\infty} f_j(x)$ is absolutely convergent if and only $x = 0$. But in Example 2.4.2–3 we showed that the series is pointwise convergent. •

3.4.5 Some results on uniform convergence of series

At various times in our development, we will find it advantageous to be able to refer to various standard results on uniform convergence, and we state these here.

Let us first recall the Weierstrass M-test.

3.4.15 Theorem (Weierstrass M-test) *If $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ is a sequence of $\mathbb{R}$-valued functions defined on an interval $\mathrm{I} \subseteq \mathbb{R}$ and if there exists a sequence of positive constants $(\mathrm{M_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ such that*

(i) $|\mathrm{f_j(x)}| \leq \mathrm{M_j}$ *for all* $\mathrm{x} \in \mathrm{I}$ *and for all* $\mathrm{j} \in \mathbb{Z}_{>0}$ *and*

(ii) $\sum_{\mathrm{j}=1}^{\infty} \mathrm{M_j} < \infty$,

then the series $\sum_{\mathrm{j}=1}^{\infty} \mathrm{f_j}$ converges uniformly and absolutely.

Proof For $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that, if $l \geq N$, we have

$$|M_l + \cdots + M_{l+k}| < \epsilon$$

for every $k \in \mathbb{Z}_{>0}$. Therefore, by the triangle inequality,

$$\left|\sum_{j=l}^{l+k} f_j(x)\right| \leq \sum_{j=l}^{l+k} |f_j(x)| \leq \sum_{j=l}^{l+k} M_j.$$

This shows that, for every $\epsilon \in \mathbb{R}_{>0}$, the tail of the series $\sum_{j=1}^{\infty} f_j$ can be made smaller than ϵ, and uniformly in x. This implies uniform and absolute convergence. ■

Next we present Abel's test.

3.4.16 Theorem (Abel's test) *Let* $(g_j)_{j\in\mathbb{Z}_{>0}}$ *be a sequence of* $\mathbb{R}$*-valued functions on an interval* $I \subseteq \mathbb{R}$ *for which* $g_{j+1}(x) \leq g_j(x)$ *for all* $j \in \mathbb{Z}_{>0}$ *and* $x \in I$*. Also suppose that there exists* $M \in \mathbb{R}_{>0}$ *such that* $g_j(x) \leq M$ *for all* $x \in I$ *and* $j \in \mathbb{Z}_{>0}$*. Then, if the series* $\sum_{j=1}^{\infty} f_j$ *converges uniformly on* I*, then so too does the series* $\sum_{j=1}^{\infty} g_j f_j$*.*

Proof Denote

$$F_k(x) = \sum_{j=1}^{k} f_j(x), \quad G_k(x) = \sum_{j=1}^{k} g_j(x) f_j(x)$$

as the partial sums. Using Abel's partial summation formula (Proposition 2.4.16), for $0 < k < l$ we write

$$G_l(x) - G_k(x) = (F_l(x) - F_k(x))G_1(x) + \sum_{j=k+1}^{l} (F_l(x) - F_j(x))(g_{j+1}(x) - g_j(x)).$$

An application of the triangle inequality gives

$$|G_l(x) - G_k(x)| = \big|(F_l(x) - F_k(x))\big||G_1(x)| + \sum_{j=k+1}^{l} \big|(F_l(x) - F_j(x))\big|(g_{j+1}(x) - g_j(x)),$$

since $|g_{j+1}(x) - g_j(x)| = g_{j+1}(x) - g_j(x)$. Now, given $\epsilon \in \mathbb{R}_{>0}$, let $N \in \mathbb{Z}_{>0}$ have the property that

$$\big|F_l(x) - F_k(x)\big| \leq \frac{\epsilon}{3M}$$

for all $k, l \geq N$. Then we have

$$\begin{aligned} |G_l(x) - G_k(x)| &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3M} \sum_{j=k+1}^{l} (g_{j+1}(x) - g_j(x)) \\ &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3M}(g_{k+1}(x) - g_{l+1}(x)) \\ &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3M}(|g_{k+1}(x)| + |g_{l+1}(x)|) \leq \epsilon. \end{aligned}$$

Thus the sequence $(G_j)_{j\in\mathbb{Z}_{>0}}$ is uniformly Cauchy, and hence uniformly convergent. ■

The final result on general uniform convergence we present is the Dirichlet test.[10]

3.4.17 Theorem (Dirichlet's test) *Let* $(f_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(g_j)_{j\in\mathbb{Z}_{>0}}$ *be sequences of* $\mathbb{R}$*-valued functions on an interval* I *and satisfying the following conditions:*

(i) there exists $M \in \mathbb{R}_{>0}$ *such that the partial sums*

$$F_k(x) = \sum_{j=1}^{k} f_j(x)$$

satisfy $|F_k(x)| \leq M$ *for all* $k \in \mathbb{Z}_{>0}$ *and* $x \in I$*;*

[10]Johann Peter Gustav Lejeune Dirichlet 1805–1859 was born in what is now Germany. His mathematical work was primarily in the areas of analysis, number theory and mechanics. For the purposes of these volumes, Dirichlet was gave the first rigorous convergence proof for the trigonometric series of Fourier. These and related results are presented in Section **??**.

(ii) $g_j(x) \geq 0$ *for all* $j \in \mathbb{Z}_{>0}$ *and* $x \in I$;
(iii) $g_{j+1}(x) \leq g_j(x)$ *for all* $j \in \mathbb{Z}_{>0}$ *and* $x \in I$;
(iv) the sequence $(g_j)_{j\in\mathbb{Z}_{>0}}$ *converges uniformly to the zero function.*
Then the series $\sum_{j=1}^{\infty} f_j g_j$ *converges uniformly on* I.

Proof We denote

$$F_k(x) = \sum_{j=1}^{k} f_j(x), \quad G_k(x) = \sum_{j=1}^{k} f_j(x)g_j(x).$$

We use again the Abel partial summation formula, Proposition 2.4.16, to write***missing stuff***

$$G_l(x) - G_k(x) = F_l(x)g_{l+1}(x) - F_k(x)g_{k+1}(x) - \sum_{j=k+1}^{l} F_j(x)(g_{l+1}(x) - g_l(x)).$$

Now we compute

$$\begin{aligned} |G_l(x) - G_k(x)| &\leq M(g_{l+1}(x) + g_{k+1}(x)) + M\sum_{j=k+1}^{l}(g_j(x) - g_{j+1}(x)) \\ &= 2Mg_{k+1}(x). \end{aligned}$$

Now, for $\epsilon \in \mathbb{R}_{>0}$, if one chooses $N \in \mathbb{Z}_{>0}$ such that $g_k(x) \leq \frac{\epsilon}{2M}$ for all $x \in I$ and $k \geq N$, then it follows that $|G_l(x) - G_k(x)| \leq \epsilon$ for $k, l \geq N$ and for all $x \in I$. From this we deduce that the sequence of partial sums $(G_j)_{j\in\mathbb{Z}_{>0}}$ is uniformly Cauchy, and hence uniformly convergent. ■

3.4.6 The Weierstrass Approximation Theorem

In this section we prove an important result in analysis. The theorem is one on approximating continuous functions with a certain class of easily understood functions. The idea, then, is that if one say something about the class of easily understood functions, it may be readily also ascribed to continuous functions. Let us first describe the class of functions we wish to use to approximate continuous functions.

3.4.18 Definition (Polynomial functions) A function $P\colon \mathbb{R} \to \mathbb{R}$ is a ***polynomial function*** if

$$P(x) = a_k x^k + \cdots + a_1 x + a_0$$

for some $a_0, a_1, \ldots, a_k \in \mathbb{R}$. The ***degree*** of the polynomial function P is the largest $j \in \{0, 1, \ldots, k\}$ for which $a_j \neq 0$. •

We shall have a great deal to say about polynomials in an algebraic setting in Section **??**. Here we will only think about the most elementary features of polynomials.

Our constructions are based on a special sort of polynomial. We recall the notation

$$\binom{m}{k} \triangleq \frac{m!}{k!(m-k)!}$$

which are the ***binomial coefficients***.

3.4.19 Definition (Bernstein polynomial, Bernstein approximation) For $m \in \mathbb{Z}_{\geq 0}$ and $k \in \{0, 1, \ldots, m\}$ the polynomial function

$$P_k^m(x) = \binom{m}{k} x^k (1-x)^{m-k}$$

is a ***Bernstein polynomial***. For a continuous function $f\colon [a,b] \to \mathbb{R}$ the **m*th Bernstein approximation*** of f is the function $B_m^{[a,b]} f\colon [a,b] \to \mathbb{R}$ defined by

$$B_m^{[a,b]} f(x) = \sum_{k=0}^{m} f(a + \tfrac{k}{m}(b-a)) P_k^m(\tfrac{x-a}{b-a}).$$ •

In Figure 3.14 we depict some of the Bernstein polynomials. The way to imagine

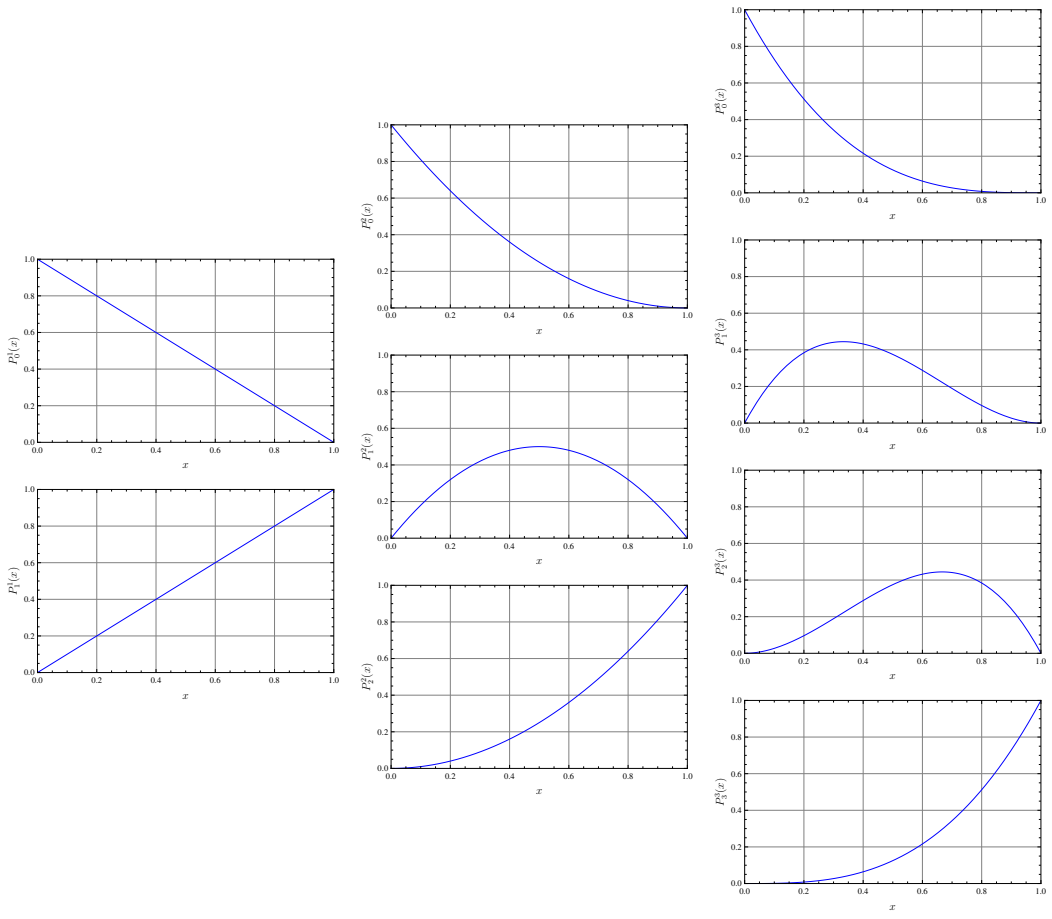


Figure 3.14 The Bernstein polynomials P_0^1 and P_1^1 (left), P_0^2, P_1^2, and P_2^2 (middle), and P_0^3, P_1^3, P_2^3, and P_3^3 (right)

the point of these functions is as follows. The polynomial P_k^m on the interval $[0,1]$

has a single maximum at $\frac{k}{m}$. By letting m vary over $\mathbb{Z}_{\geq 0}$ and letting $k \in \{0, 1, \ldots, m\}$, the points of the form $\frac{k}{m}$ will get arbitrarily close to any point in $[0, 1]$. The function $f(\frac{k}{m})P_k^m$ thus has a maximum at $\frac{k}{m}$ and the behaviour of f away from $\frac{k}{m}$ is thus (sort of) attenuated. In fact, for large m the behaviour of the function P_k^m becomes increasingly "focussed" at $\frac{k}{m}$. Thus, as m gets large, the function $f(\frac{k}{m})P_k^m$ starts looking like the function taking the value $f(\frac{k}{m})$ at $\frac{k}{m}$ and zero elsewhere. Now, using the identity

$$\sum_{k=0}^{m} \binom{m}{k} x^k (1-x)^m = 1 \tag{3.13}$$

which can be derived using the Binomial Theorem (see Exercise 2.2.1), this means that for large m, $B_m^{[0,1]} f(\frac{k}{m})$ approaches the value $f(\frac{k}{m})$. This is the idea of the Bernstein approximation.

That being said, let us prove some basic facts about Bernstein approximations.

3.4.20 Lemma (Properties of Bernstein approximations) *For continuous functions* $f, g\colon [a, b] \to \mathbb{R}$, *for* $\alpha \in \mathbb{R}$, *and for* $m \in \mathbb{Z}_{\geq 0}$, *the following statements hold:*

(i) $B_m^{[a,b]}(f + g) = B_m^{[a,b]} f + B_m^{[a,b]} g$;

(ii) $B_m^{[a,b]}(\alpha f) = \alpha B_m^{[a,b]} f$;

(iii) $B_m^{[a,b]} f(x) \geq 0$ *for all* $x \in [a, b]$ *if* $f(x) \geq 0$ *for all* $x \in [a, b]$;

(iv) $B_m^{[a,b]} f(x) \leq B_m^{[a,b]} g(x)$ *for all* $x \in [a, b]$ *if* $f(x) \leq g(x)$ *for all* $x \in [a, b]$;

(v) $|B_m^{[a,b]} f(x)| \leq B_m^{[a,b]} g(x)$ *for all* $x \in [a, b]$ *if* $|f(x)| \leq g(x)$ *for all* $x \in [a, b]$;

(vi) *for* $k, m \in \mathbb{Z}_{\geq 0}$ *we have*

$$(B_{m+k}^{[a,b]})^{(k)}(x) = \frac{(m+k)!}{m!} \frac{1}{(b-a)^k} \sum_{j=0}^{m} \Delta_h^k f(a + \tfrac{j}{k+m}(b-a)) P_j^m(\tfrac{x-a}{b-a}),$$

where $h = \frac{1}{k+m}$ *and where* $\Delta_h^k f\colon [a, b] \to \mathbb{R}$ *is defined by*

$$\Delta_h^k f(x) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(x + jh)$$

(vii)

(viii) *if we define* $f_0, f_1, f_2\colon [0, 1] \to \mathbb{R}$ *by*

$$f_0(x) = 1, \quad f_1(x) = x, \quad f_2(x) = x^2, \qquad x \in [0, 1],$$

then

$$B_m^{[0,1]} f_0(x) = 1, \quad B_m^{[0,1]} f_1(x) = x, \quad B_m^{[0,1]} f_2(x) = x^2 + \tfrac{1}{m}(x - x^2)$$

for $x \in [0, 1]$ *and* $m \in \mathbb{Z}_{\geq 0}$.

Proof Let $\hat{f}\colon [0,1] \to \mathbb{R}$ be defined by $\hat{f}(y) = f(a + \frac{y}{(}b - a))$. One can verify that if the lemma holds for $\hat{f}$ then it immediately follows for f, and so without loss of generality we suppose that $[a,b] = [0,1]$. We also abbreviate $B_m^{[0,1]} = B_m$.

(i)–(iv) These assertions follow directly from the definition of the Bernstein approximations.

(v) If $|f(x)| \leq g(x)$ for all $x \in [0,1]$ then

$$\begin{aligned} & -f(x) \leq g(x) \leq f(x), \qquad x \in [0,1] \\ \implies \quad & -B_m f(x) \leq B_m g(x) \leq B_m f(x), \qquad x \in [0,1], \end{aligned}$$

using the fourth assertion.

(vi) Note that

$$B_{m+k}(x) = \sum_{j=0}^{m+k} f(\tfrac{j}{m+k}) \binom{m+k}{j} x^j (1-x)^{m+k-j}.$$

Let $g_j(x) = x^j$ and $h_j(x) = (1-x)^{m+k-j}$ and compute

$$g_j^{(r)}(x) = \begin{cases} \frac{j!}{(j-r)!} x^{j-r}, & j-r \geq 0, \\ 0, & j-r < 0 \end{cases}$$

and

$$h_j^{(k-r)}(x) = \begin{cases} (-1)^{k-r} \frac{(m+k-j)!}{(m+r-j)!} (1-x)^{m+r-j}, & j-r \leq m, \\ 0, & j-r > m. \end{cases}$$

By Proposition 3.2.11,

$$(g_j h_j)^{(k)}(x) = \sum_{r=0}^{k} \binom{k}{r} g_j^{(r)}(x) h_j^{(k-r)}(x).$$

Also note that

$$\begin{aligned} \binom{m+k}{j} \frac{j!}{(j-r)!} \frac{(m+k-j)!}{(m+r-j)!} &= \frac{(m+k)!}{j!(m+k-j)!} \frac{j!}{(j-r)!} \frac{(m+k-j)!}{(m+r-j)!} \\ &= \frac{(m+k)!}{m!} \frac{m!}{(m-(j-r))!(j-r)!} = \frac{(m+k)!}{m!} \binom{m}{j-r}. \end{aligned}$$

Putting this all together we have

$$\begin{aligned} B_{m+k}^{(k)}(x) &= \sum_{j=0}^{m+k} \sum_{r=0}^{k} f(\tfrac{j}{m+k}) \binom{m+k}{j} \binom{k}{r} g_j^{(r)}(x) h_j^{(k-r)}(x) \\ &= \sum_{r=0}^{k} \sum_{l=-r}^{m+k-r} f(\tfrac{l+r}{m+k}) \binom{m+k}{l+r} \binom{k}{r} g_{l+r}^{(r)}(x) h_{l+r}^{(k-r)}(x) \\ &= \sum_{r=0}^{k} \sum_{l=0}^{m} (-1)^{k-r} \binom{k}{r} f(\tfrac{l+r}{m+k}) \binom{m}{l} x^l (1-x)^{n-l}, \end{aligned}$$

where we make the change of index $(l, r) = (j - r, r)$ in the second step and note that the derivatives of g_{l+r} and h_{l+r} vanish when $l < 0$ and $l > m$. Let $h = \frac{1}{m+k}$. Since

$$\Delta_h^k f(\tfrac{j}{m+k}) = \sum_{r=0}^{k} (-1)^{k-r} \binom{k}{r} f(\tfrac{j+r}{m+k})$$

this part of the result follows.

(vii)

(viii) It follows from (3.13) that $B_m f_0(x) = 1$ for every $x \in [0, 1]$. We also compute

$$\begin{aligned} B_m f_0(x) &= \sum_{k=0}^{m} \frac{k}{m} \frac{m!}{m!(m-k)!} x^k (1-x)^{m-k} \\ &= x \sum_{k=0}^{m-1} \frac{(m-1)!}{(k-1)!((m-1)-(k-1))!} x^k (1-x)^{m-1-k} \\ &= x(x + (1-x))^{m-1} = x, \end{aligned}$$

where we use the Binomial Theorem. To compute $B_m f_2$ we first compute

$$\begin{aligned} \frac{k^2}{m^2} \frac{m!}{k!(m-k)!} &= \frac{(k-1)+1}{m} \frac{(m-1)!}{(k-1)!(m-k)!} \\ &= \frac{(k-1)(n-1)}{n(n-1)} \frac{(m-1)!}{(k-1)!(m-k)!} + \frac{1}{m} \frac{(m-1)!}{(k-1)!(m-k)!} \\ &= \frac{m-1}{m} \binom{n-2}{k-2} + \frac{1}{m} \binom{n-1}{k-1}, \end{aligned}$$

where we adopt the convention that $\binom{j}{l} = 0$ if either j or l are zero. We now compute

$$\begin{aligned} B_m f_2(x) &= \sum_{k=0}^{m} \frac{k^2}{m^2} \binom{m}{k} x^k (1-x)^{m-k} \\ &= \frac{m-1}{m} \sum_{k=2}^{m} \binom{m-2}{k-2} x^k (1-x)^{m-k} + \frac{1}{m} \sum_{k=1}^{m} \binom{m-1}{k-1} x^k (1-x)^{m-k} \\ &= \frac{m-1}{m} x^2 (x + (1-x))^{m-2} + \frac{1}{m} x(x + (1-x))^{m-1} = \frac{m-1}{m} x^2 + \frac{1}{m} x, \end{aligned}$$

as desired. ∎

Now, heuristics aside, we state the main result in this section, a consequence of which is that every continuously function on a compact interval can be approximated arbitrarily well (in the sense that the maximum difference can be made as small as desired) by a polynomial function.

3.4.21 Theorem (Weierstrass Approximation Theorem) *Consider a compact interval* $[a, b] \subseteq \mathbb{R}$ *and let* $f\colon [a, b] \to \mathbb{R}$ *be continuous. Then the sequence* $(B_m^{[a,b]} f)_{m \in \mathbb{Z}_{>0}}$ *converges uniformly to* f *on* $[a, b]$.

Proof It is evident (why?) that we can take $[a,b] = [0,1]$ and then let us denote $B_m f = B_m^{[0,1]} f$ for simplicity.

Let $\epsilon \in \mathbb{R}_{>0}$. Since f is uniformly continuous by Theorem 3.1.24 there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(y)| \leq \frac{\epsilon}{2}$ whenever $|x - y| \leq \delta$. Let

$$M = \sup\{|f(x)| \mid x \in [0,1]\},$$

noting that $M < \infty$ by Theorem 3.1.23. Note then that if $|x - y| \leq \delta$ then

$$|f(x) - f(y)| \leq \tfrac{\epsilon}{2} \leq \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x - y)^2.$$

If $|x - y| > \delta$ then

$$|f(x) - f(y)| \leq 2M \leq 2M\Big(\tfrac{x-y}{\delta}\Big)^2 \leq \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x - y)^2.$$

That is to say, for every $x, y \in [0,1]$,

$$|f(x) - f(y)| \leq \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x - y)^2. \tag{3.14}$$

Now, fix $x_0 \in [0,1]$ and compute, using the lemma above (along with the notation f_0, f_1, and f_2 introduced in the lemma) and (3.14),

$$\begin{aligned} |B_m f(x) - f(x_0)| = |B_m(f - f(x_0)f_0)(x)| &\leq B_m\Big(\tfrac{\epsilon}{2} f_0 + \tfrac{2M}{\delta^2}(f_1 - x_0 f_0)^2\Big)(x) \\ &= \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x^2 + \tfrac{1}{m}(x - x^2) - 2x_0 x + x_0^2) \\ &= \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x - x_0)^2 + \tfrac{2M}{m\delta^2}(x - x^2), \end{aligned}$$

this holding for every $m \in \mathbb{Z}_{\geq 0}$. Now evaluate at $x = x_0$ to get

$$|B_m f(x_0) - f(x_0)| \leq \tfrac{\epsilon}{2} + \tfrac{2M}{m\delta^2}(x_0 - x_0^2) \leq \tfrac{\epsilon}{2} + \tfrac{M}{2m\delta^2},$$

using the fact that $x_0 - x_0^2 \leq \frac{1}{4}$ for $x_0 \in [0,1]$. Therefore, if $N \in \mathbb{Z}_{>0}$ is sufficiently large that $\frac{M}{2m\delta^2} < \frac{\epsilon}{2}$ for $m \geq N$ we have

$$|B_m f(x_0) - f(x_0)| < \epsilon,$$

and this holds for every $x_0 \in [0,1]$, giving us the desired uniform convergence. ■

For fun, let us illustrate the Bernstein approximations in an example.

3.4.22 Example (Bernstein approximation) Let us consider $f\colon [0,1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} x, & x \in [0, \frac{1}{2}], \\ 1 - x, & x \in (\frac{1}{2}, 1]. \end{cases}$$

In Figure 3.15 we show some Bernstein approximations to f. Note that the convergence is rather poor. One might wish to contrast the 100th approximation in Figure 3.15 with the 10 approximation of the same function using Fourier series depicted in Figure **??**. (If you have no clue what a Fourier series is, that is fine. We will get there in time.) •

We shall revisit the Weierstrass Approximation Theorem in Sections 4.5.2 and ***missing stuff***.

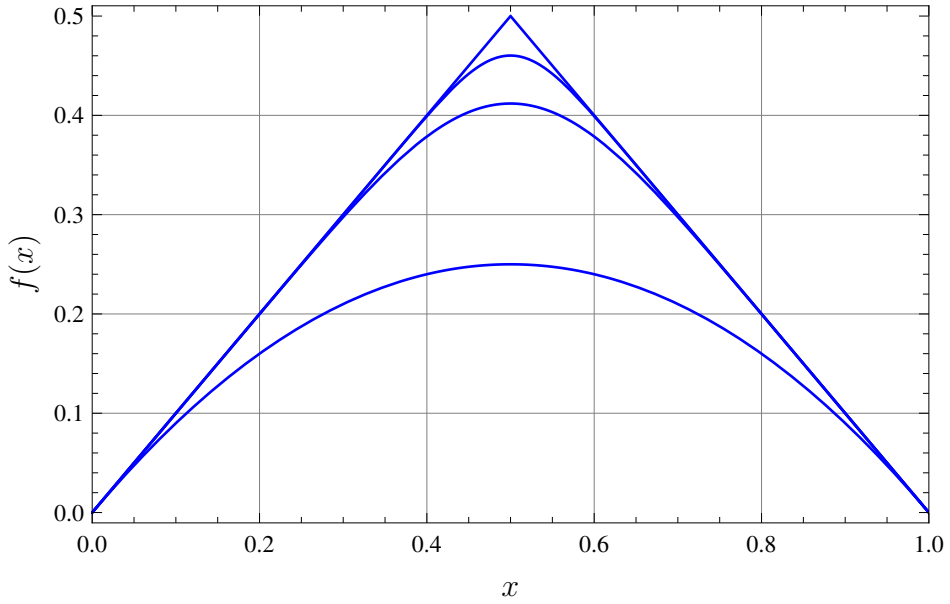


Figure 3.15 Bernstein approximations for $m \in \{2, 50, 100\}$

3.4.7 Swapping limits with other operations

In this section we give some basic result concerning the swapping of various function operations with limits. The first result we consider pertains to integration. When we consider Lebesgue integration in Chapter **??** we shall see that there are more powerful limit theorems available. Indeed, the *raison d'etre* for the Lebesgue integral is just these limit theorems, as these are not true for the Riemann integral. However, for the moment these theorems have value in that they apply in at least some cases, and indicate what *is* true for the Riemann integral.

3.4.23 Theorem (Uniform limits commute with Riemann integration) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $(\mathrm{f_j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *be a sequence of continuous* $\mathbb{R}$*-valued functions defined on* $[\mathrm{a}, \mathrm{b}]$ *that converge uniformly to* f. *Then*

$$\lim_{\mathrm{j} \to \infty} \int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f_j}(\mathrm{x})\, \mathrm{dx} = \int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\, \mathrm{dx}.$$

Proof As the functions $(f_j)_{j \in \mathbb{Z}_{>0}}$ are continuous and the convergence to f is uniform, f must be continuous by Theorem 3.4.8. Since the interval $[a, b]$ is compact, the functions f and f_j, $j \in \mathbb{Z}_{>0}$, are also bounded. Therefore, by part Proposition 3.3.25,***missing stuff***

$$\left| \int_a^b f(x)\, \mathrm{d}x \right| \leq M(b - a)$$

where $M = \sup\{|f(x)| \mid x \in [a, b]\}$. Let $\epsilon \in \mathbb{R}_{>0}$ and select $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f(x)| <$

$\frac{\epsilon}{b-a}$ for all $x \in [a,b]$, provided that $j \geq N$. Then

$$\left|\int_a^b f_j(x)\,\mathrm{d}x - \int_a^b f(x)\,\mathrm{d}x\right| = \left|\int_a^b (f_j(x) - f(x))\,\mathrm{d}x\right| \leq \frac{\epsilon}{b-a}(b-a) = \epsilon.$$

This is the desired result. ∎

Next we state a result that tells us when we may switch limits and differentiation.

3.4.24 Theorem (Uniform limits commute with differentiation) *Let* $\mathrm{I} = [\mathrm{a},\mathrm{b}]$ *be a compact interval and let* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence continuously differentiable* $\mathbb{R}$*-valued functions on* $[\mathrm{a},\mathrm{b}]$*, and suppose that the sequence converges pointwise to* f*. Also suppose that the sequence* $(\mathrm{f_j'})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *of derivatives converges uniformly to* g*. Then* f *is differentiable and* $\mathrm{f}' = \mathrm{g}$*.*

Proof Our hypotheses ensure that we may write, for each $j \in \mathbb{Z}_{>0}$,

$$f_j(x) = f_j(a) + \int_a^x f_j'(\xi)\,\mathrm{d}\xi.$$

for each $x \in [a,b]$. By Theorem 3.4.23, we may interchange the limit as $j \to \infty$ with the integral, and so we get

$$f(t) = f(a) + \int_a^x g(\xi)\,\mathrm{d}\xi.$$

Since g is continuous, being the uniform limit of continuous functions (by Theorem 3.4.8), the Fundamental Theorem of Calculus ensures that $f' = g$. ∎

The next result in this section has a somewhat different character than the rest. It actually says that it is possible to differentiate a sequence of monotonically increasing functions term-by-term, except on a set of measure zero. The interesting thing here is that only pointwise convergence is needed.

3.4.25 Theorem (Termwise differentiation of sequences of monotonic functions is a.e. valid) *Let* $\mathrm{I} = [\mathrm{a},\mathrm{b}]$ *be a compact interval, let* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence of monotonically increasing functions such that the series* $\mathrm{S} = \sum_{\mathrm{j}=1}^{\infty} \mathrm{f_j(x)}$ *converges pointwise to a function* f*. Then there exists a set* $\mathrm{Z} \subseteq \mathrm{I}$ *such that*

(i) Z *has measure zero and*

(ii) $\mathrm{f}'(\mathrm{x}) = \sum_{\mathrm{j}=1}^{\infty} \mathrm{f_j'(x)}$ *for all* $\mathrm{x} \in \mathrm{I} \setminus \mathrm{Z}$*.*

Proof Note that the limit function f is monotonically increasing. Denote by $Z_1 \subseteq [a,b]$ the set of points for which all of the functions f and f_j, $j \in \mathbb{Z}_{>0}$, do not possess derivatives. Note that by Theorem 3.2.26 it follows that Z_1 is a countable union of sets of measure zero. Therefore, by Exercise 2.5.9, Z_1 has measure zero. Now let $x \in I \setminus Z_1$ and let $\epsilon \in \mathbb{R}_{>0}$ be sufficiently small that $x + \epsilon \in [a,b]$. Then

$$\frac{f(x+\epsilon) - f(x)}{\epsilon} = \sum_{j=1}^{\infty} \frac{f_j(x+\epsilon) - f_j(x)}{\epsilon}.$$

Since $f_j(x+\epsilon) - f_j(x) \geq 0$, for any $k \in \mathbb{Z}_{>0}$ we have

$$\frac{f(x+\epsilon) - f(x)}{\epsilon} \geq \sum_{j=1}^{k} \frac{f_j(x+\epsilon) - f_j(x)}{\epsilon},$$

which then gives

$$f'(x) \geq \sum_{j=1}^{k} f_j'(x).$$

The sequence of partial sums for the series $\sum_{j=1}^{\infty} f_j'(x)$ is therefore bounded above. Moreover, by Theorem 3.2.26, it is increasing. Therefore, by Theorem 2.3.8 the series $\sum_{j=1}^{\infty} f_j'(x)$ converges for every $x \in I \setminus Z_1$.

Let us now suppose that $f(a) = 0$ and $f_j(a) = 0$, $j \in \mathbb{Z}_{>0}$. This can be done without loss of generality by replacing f with $f - f(a)$ and f_j with $f_j - f_j(a)$, $j \in \mathbb{Z}_{>0}$. With this assumption, for each $x \in [a,b]$ and $k \in \mathbb{Z}_{>0}$, we have $f(x) - S_k(x) \geq 0$ where $(S_k)_{k\in\mathbb{Z}_{>0}}$ is the sequence of partial sums for S. Choose a subsequence $(S_{k_l})_{l\in\mathbb{Z}_{>0}}$ of $(S_k)_{k\in\mathbb{Z}_{>0}}$ having the property that $0 \leq f(b) - S_{k_l}(b) \leq 2^{-l}$, this being possible since the sequence $(S_k(b))_{k\in\mathbb{Z}_{>0}}$ converges to $f(b)$. Note that

$$f(x) - S_{k_l}(x) = \sum_{j=k_l+1}^{\infty} f_j(x),$$

meaning that $f - S_{k_l}$ is a monotonically increasing function. Therefore, $0 \leq f(x) - S_{k_l}(x) \leq 2^{-l}$ for all $x \in [a,b]$. This shows that the series $\sum_{l=1}^{\infty}(f(x) - S_{k_l}(x))$ is a pointwise convergent sequence of monotonically increasing functions. Let g denote the limit function, and let $Z_2 \subseteq [a,b]$ be the set of points where all of the functions g and $f - S_{k_l}$, $l \in \mathbb{Z}_{>0}$, do not possess derivatives, noting that this set is, in the same manner as was Z_1, a set of measure zero. The argument above applies again to show that, for $x \in I \setminus Z_2$, the series $\sum_{l=1}^{\infty}(f'(x) - S_{k_l}'(x))$ converges. Thus, for $x \in I \setminus Z_2$, it follows that $\lim_{l\to\infty}(f'(x) - S_{k_l}'(x)) = 0$. Now, for $x \in I \setminus Z_1$, we know that $(S_k'(x))_{k\in\mathbb{Z}_{>0}}$ is a monotonically increasing sequence. Therefore, for $x \in I \setminus (Z_1 \cup Z_2)$, the sequence $(f'(x) - S_k'(x))_{k\in\mathbb{Z}_{>0}}$ must converge to zero. This gives the result by taking $Z = Z_1 \cup Z_2$. ■

As a final result, we indicate how convexity interacts with pointwise limits.

3.4.26 Theorem (The pointwise limit of convex functions is convex) *If $I \subseteq \mathbb{R}$ is convex and if $(f_j)_{j\in\mathbb{Z}_{>0}}$ is a sequence of convex functions converging pointwise to $f\colon I \to \mathbb{R}$, then f is convex.*

Proof Let $x_1, x_2 \in I$ and let $s \in [0,1]$. Then

$$\begin{aligned} f((1-s)x_1 + sx_2) &= \lim_{j\to\infty} f_j((1-s)x_1 + sx_2) \leq \lim_{j\to\infty}((1-s)f_j(x_1) + sf_j(x_2)) \\ &= (1-s)\lim_{j\to\infty} f_j(x_1) + s\lim_{j\to\infty} f_j(x_2) \\ &= (1-s)f(x_1) + sf(x_2), \end{aligned}$$

where we have used Proposition 2.3.23. ■

3.4.8 Notes

There are many proofs available of the Weierstrass Approximation Theorem, and the rather explicit proof we give is due to Bernstein [1912].

Exercises

3.4.1 Consider the sequence of functions $\{f_j\}_{j\in\mathbb{Z}_{>0}}$ defined on the interval $[0,1]$ by $f_j(x) = x^{1/2^j}$. Thus

$$f_1(x) = \sqrt{x},\ \ f_2(x) = \sqrt{f_1(x)} = \sqrt{\sqrt{x}},\ \ \ldots,\ \ f_j(x) = \sqrt{f_{j-1}(x)} = x^{1/2^j}, \ldots$$

(a) Sketch the graph of f_j for $j \in \{1,2,3\}$.
(b) Does the sequence of functions $(f_j)_{j\in\mathbb{Z}_{>0}}$ converge pointwise? If so, what is the limit function?
(c) Is the convergence of the sequence of functions $(f_j)_{j\in\mathbb{Z}_{>0}}$ uniform?
(d) Is it true that

$$\lim_{j\to\infty} \int_0^1 f_j(x)\,\mathrm{d}x = \int_0^1 \lim_{j\to\infty} f_j(x)\,\mathrm{d}x?$$

3.4.2 In each of the following exercises, you will be given a sequence of functions defined on the interval $[0,1]$. In each case, answer the following questions.

1. Sketch the first few functions in the sequence.
2. Does the sequence converge pointwise? If so, what is the limit function?
3. Does the sequence converge uniformly?

The sequences are as follows:

(a) $(f_j(x) = (x - \frac{1}{j^2})^2)_{j\in\mathbb{Z}_{>0}}$;
(b) $(f_j(x) = x - x^j)_{j\in\mathbb{Z}_{>0}}$.

3.4.3 Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of locally bounded functions on I converging pointwise to $f\colon I \to \mathbb{R}$. Show that there exists a function $g\colon I \to \mathbb{R}$ such that $(f_j)_{j\in\mathbb{Z}_{>0}}$ converges dominated by g.

Section 3.5

$\mathbb{R}$-power series

In Section 3.4.4 we considered the convergence of general series of functions. In this section we consider special series of functions where the functions in the series are given by $f_j(x) = a_j x^j$, $j \in \mathbb{Z}_{\geq 0}$. This class of series is important in a surprising number of ways. For example, as we shall see in Section 3.5.4, one can associate a power series to every function of class C^∞, and this power series sometimes approximates the function in some sense.

Do I need to read this section? The material in this section is of a somewhat technical character, and so can probably be skipped until it is needed. One of the main uses will occur in Section **??** when we explore the intimate relationship between power series and analytic functions in complex analysis. There will also be occasions throughout these volumes when it is convenient to use Taylor's Theorem. •

3.5.1 $\mathbb{R}$-formal power series

We begin with a discussion that is less analytical, and more algebraic in flavour. This discussion serves to separate the simpler algebraic features of power series from the more technical analytical features. A purely logical presentation of this material would certain present the material Section **??** before our present discussion. However, we have decided to make a small sacrifice in logic for the sake of organisation. Readers wishing to preserve the logical structure may wish to look ahead at this point to Section **??**.

Let us first give a formal definition of what we mean by a $\mathbb{R}$-formal power series, while at the same time defining the operations of addition and multiplication in this set.

3.5.1 Definition ($\mathbb{R}$-formal power series) A $\mathbb{R}$-*formal power series* is a sequence $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ in $\mathbb{R}$. If $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ and $B = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ are two $\mathbb{R}$-formal power series, then define $\mathbb{R}$-formal power series $A + B$ and $A \cdot B$ by

$$A + B = (a_j + b_j)_{j\in\mathbb{Z}_{\geq 0}}, \quad A \cdot B = \Big(\sum_{j=0}^{k} a_j b_{k-j}\Big)_{k\in\mathbb{Z}_{\geq 0}},$$

which are the ***sum*** and ***product*** of A and B, respectively. If $\alpha \in \mathbb{R}$ then αA denotes the $\mathbb{R}$-formal power series $(\alpha a_j)_{j\in\mathbb{Z}_{\geq 0}}$ which is the ***product*** of α and A. •

In order to distinguish between multiplication of two $\mathbb{R}$-formal power series and multiplication of a $\mathbb{R}$-formal power series by a real number, we shall call the latter ***scalar multiplication***. This is reflective of the idea of a vector space that we introduce in Section **??**. Note that the product of $\mathbb{R}$-formal power series is very

much related to the Cauchy product of series in Definition 2.4.29. As we shall see, this is not surprising given the natural manner of thinking about $\mathbb{R}$-formal power series.

Our definition of $\mathbb{R}$-formal power series is meant to be rigorous, but suffers from being at the same time obtuse. A less obtuse working definition is possible, and requires the following notion.

3.5.2 Definition (Indeterminate) The ***indeterminate*** in the set of $\mathbb{R}$-formal power series is the element $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$a_j = \begin{cases} 1, & j = 1, \\ 0, & \text{otherwise.} \end{cases}$$

If the indeterminate is denoted by the symbol ξ, then $\mathbb{R}[[\xi]]$ denotes the ***set of $\mathbb{R}$-formal power series in indeterminate ξ***. •

Now let us see what are the notational implications of introducing the indeterminate into the picture. A direct application of the definition of the product shows that, if the indeterminate is denoted by ξ and if $k \in \mathbb{Z}_{>0}$, then ξ^k (the k-fold product of ξ with itself) is the $\mathbb{R}$-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ given by

$$a_j = \begin{cases} 1, & j = k, \\ 0, & \text{otherwise.} \end{cases}$$

Let us adopt the convention that ξ^0 denotes the $\mathbb{R}$-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$a_j = \begin{cases} 1, & j = 0, \\ 0, & j \in \mathbb{Z}_{>0}. \end{cases}$$

Now let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ be an *arbitrary* $\mathbb{R}$-formal power series and, for $k \in \mathbb{Z}_{\geq 0}$, let A_k denote the $\mathbb{R}$-formal power series $(a_{k,j})_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$a_{k,j} = \begin{cases} a_j, & j \leq k, \\ 0, & j > k. \end{cases}$$

Note that, using the definition of

$$\begin{aligned} A_k &= (a_0, a_1, \dots, a_k, 0, \dots) \\ &= (a_0, 0, \dots, 0, 0, \dots) + (0, a_1, \dots, 0, 0, \dots) + \cdots + (0, 0, \dots, a_k, 0, \dots) \\ &= a_0\xi^1 + a_a\xi^1 + \cdots + a_k\xi^k. \end{aligned}$$

We would now like to write $A = \lim_{k\to\infty} A_k$, but the problem is that we do not really know what the limit means in this case. It certainly does not mean the limit thinking of the sum as one of real numbers; this limit will generally not exist. Thus we define what the limit means as follows.***missing stuff***

3.5.3 Definition (Limit of $\mathbb{R}$-formal power series) Let $(A_k = (a_{k,j})_{j\in\mathbb{Z}_{\geq 0}})_{k\in\mathbb{Z}_{\geq 0}}$ be a sequence of $\mathbb{R}$-formal power series and let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ be a $\mathbb{R}$-formal power series. The sequence $(A_k)_{k\in\mathbb{Z}_{\geq 0}}$ ***converges*** to A, and we write $A = \lim_{k\to\infty} A_k$, if, for each $j \in \mathbb{Z}_{\geq 0}$, there exists $N_j \in \mathbb{Z}_{\geq 0}$ such that $a_{k,j} = a_j$ for $k \geq N_j$. •

With this notion of convergence in the set of $\mathbb{R}$-formal power series we can prove what we want.

3.5.4 Proposition ($\mathbb{R}$-formal power series as limits of finite sums) *If* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *is a* $\mathbb{R}$*-formal power series, then*

$$\mathrm{A} = \lim_{\mathrm{k}\to\infty} \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{a_j}\xi^{\mathrm{j}}.$$

Proof Let $A_k = \sum_{j=0}^{k} a_j\xi^j$ and denote $A_k = (a_{k,j})_{j\in\mathbb{Z}_{\geq 0}}$. For $j \in \mathbb{Z}_{\geq 0}$ note that $a_{k,j} = a_j$ for $k \geq j$, which gives the condition that $(A_k)_{k\in\mathbb{Z}_{\geq 0}}$ converge to A by taking $N_j = j$ in the definition. ■

The upshot of the preceding exceedingly ponderous discussion is that we can write the $\mathbb{R}$-formal power $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ as

$$\sum_{j=0}^{\infty} a_j\xi^j,$$

and all of the symbols in this expression make exact sense. Moreover, with this representation of a $\mathbb{R}$-formal power series, addition is merely the addition of the coefficients of like powers of the indeterminate. Multiplication is to be interpreted as follows. Suppose that one wishes to find the coefficient of ξ^k in the product $A \cdot B$. One does this by writing, in indeterminate form, the first $k + 1$ terms in A and B, and multiplying them using the usual rules for multiplication of finite sums in $\mathbb{R}$. Thus we write

$$A_k = \sum_{j=0}^{k} a_j\xi^j, \quad B_k = \sum_{j=0}^{k} b_j\xi^j,$$

and compute

$$A_k \cdot B_k = \sum_{l=0}^{2k} \sum_{j=0}^{l} a_j b_{l-j}\xi^j$$

(this formula is easily proved, cf. Theorem **??**). One then can see that the coefficient of ξ^k in this expression is exactly the $(k + 1)$st term in the sequence $A \cdot B$.

Let us present the basic properties of the operations of addition and multiplication of $\mathbb{R}$-formal power series. To do this, we let $0_{\mathbb{R}[[\xi]]}$ denote the $\mathbb{R}$-formal power series $(0)_{j\in\mathbb{Z}_{\geq 0}}$ and we let $1_{\mathbb{R}[[\xi]]}$ denote the $\mathbb{R}$-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ given by

$$a_j = \begin{cases} 1, & j = 0, \\ 0, & j \in \mathbb{Z}_{>0}. \end{cases}$$

If $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ is a $\mathbb{R}$-formal power series, then we let $-A$ denote the $\mathbb{R}$-formal power series $(-a_j)_{j\in\mathbb{Z}_{\geq 0}}$. If $a_0 \neq 0$ then we define the $\mathbb{R}$-formal power series $A^{-1} = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ by inductively defining

$$\begin{aligned} b_0 &= \tfrac{1}{a_0},\\ b_1 &= \tfrac{1}{a_0}(-a_1 b_0),\\ &\vdots\\ b_k &= -\frac{1}{a_0}\sum_{j=1}^{k} a_j b_{k-j},\\ &\vdots \end{aligned}$$

With these definitions, the following result is straightforward to prove, and follows from our discussion of polynomials in Section **??**.

3.5.5 Proposition (Properties of addition and multiplication of $\mathbb{R}$-formal power series) *Let* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$, $\mathrm{B} = (\mathrm{b_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$, *and* $\mathrm{C} = (\mathrm{c_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *be* $\mathbb{R}$*-formal power series. Then the following statements hold:*

(i) $\mathrm{A} + \mathrm{B} = \mathrm{B} + \mathrm{A}$ (***commutativity*** *of addition);*

(ii) $(\mathrm{A} + \mathrm{B}) + \mathrm{C} = \mathrm{A} + (\mathrm{B} + \mathrm{C})$ (***associativity*** *of addition);*

(iii) $\mathrm{A} + 0_{\mathbb{R}[[\xi]]} = \mathrm{A}$ (***additive identity***);

(iv) $\mathrm{A} + (-\mathrm{A}) = 0_{\mathbb{R}[[\xi]]}$ (***additive inverse***);

(v) $\mathrm{A}\cdot\mathrm{B} = \mathrm{B}\cdot\mathrm{A}$ (***commutativity*** *of multiplication);*

(vi) $(\mathrm{A}\cdot\mathrm{B})\cdot\mathrm{C} = \mathrm{A}\cdot(\mathrm{B}\cdot\mathrm{C})$ (***associativity*** *of multiplication);*

(vii) $\mathrm{A}\cdot(\mathrm{B} + \mathrm{C}) = \mathrm{A}\cdot\mathrm{B} + \mathrm{A}\cdot\mathrm{C}$ (***left distributivity***);

(viii) $(\mathrm{A} + \mathrm{B})\cdot\mathrm{C} = \mathrm{A}\cdot\mathrm{C} + \mathrm{B}\cdot\mathrm{C}$ (***right distributivity***);

(ix) $\mathrm{A}\cdot 1_{\mathbb{R}[[\xi]]} = \mathrm{A}$ (***multiplicative identity***);

(x) *if* $\mathrm{a}_0 \neq 0$ *then* $\mathrm{A}\cdot\mathrm{A}^{-1} = 1_{\mathbb{R}[[\xi]]}$ (***multiplicative inverse***).

Proof With the exception of the multiplicative inverse, these properties all follow in the same manner as for polynomials as proved in Theorem **??**. The formula for the multiplicative inverse arises from writing down the elements in the equation $A\cdot A^{-1} = 1_{\mathbb{R}[[\xi]]}$, and solving recursively for the unknown elements of the sequence A^{-1}, starting with the zeroth term. ■

The preceding properties of addition and scalar multiplication can be summarised in the language of Section **??** by saying that $\mathbb{R}[[\xi]]$ is a ring. Note that the multiplicative inverse of a formal $\mathbb{R}$-power series does not always exist, even when $A \neq 0_{\mathbb{R}[[\xi]]}$.

For multiplication of a $\mathbb{R}$-formal power series by a real number, we have the following properties.

3.5.6 Proposition (Properties of scalar multiplication of $\mathbb{R}$-formal power series) *Let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ and $B = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ be $\mathbb{R}$-formal power series and let $\alpha, \beta \in \mathbb{R}$. Then the following statements hold:*

(i) $\alpha(\beta A) = (\alpha\beta)A$ *(**associativity**);*

(ii) $1 A = A$*;*

(iii) $\alpha(A + B) = \alpha A + \alpha B$ *(**distributivity**);*

(iv) $(\alpha + \beta)A = \alpha A + \beta B$ *(**distributivity** again).*

Proof These all follow directly from the definition of scalar multiplication and the properties of addition and multiplication in $\mathbb{R}$ as given in Proposition 2.2.4. ■

According to the terminology of Section **??**, the preceding result, along with the properties of addition from Proposition 3.5.5, ensure that $\mathbb{R}[[\xi]]$ is a $\mathbb{R}$-vector space. With the additional structure given by the product, we further see that $\mathbb{R}[[\xi]]$ is, in fact, a commutative and associative $\mathbb{R}$-algebra.***missing stuff***

In terms of our definition of convergence in $\mathbb{R}[[\xi]]$, one has the following properties of addition, multiplication, and scalar multiplication.

3.5.7 Proposition (Sums and products, and convergence in $\mathbb{R}[[\xi]]$) *Let $(A_k = (a_{k,j})_{j\in\mathbb{Z}_{\geq 0}})_{k\in\mathbb{Z}_{>0}}$ and $(B_k = (b_{k,j})_{j\in\mathbb{Z}_{\geq 0}})_{k\in\mathbb{Z}_{>0}}$ be sequences of $\mathbb{R}$-formal power series converging to the $\mathbb{R}$-formal power series $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ and $B = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$, respectively, and let $\alpha \in \mathbb{R}$. Then the following statements hold:*

(i) $\lim_{k\to\infty}(A_k + B_k) = A + B$*;*

(ii) $\lim_{k\to\infty}(A_k \cdot B_k) = A \cdot B$*;*

(iii) $\lim_{k\to\infty}(\alpha A_k) = \alpha A$*.*

Proof The first two conclusions follow from the definition of convergence of $\mathbb{R}$-formal power series, noting that the operations of addition and multiplication have the property that, if two $\mathbb{R}$-formal power series agree for sufficiently large values of the index, then so too do their sum and product. We leave the elementary, albeit slightly tedious, details to the reader. The final assertion follows trivially from the definition of convergence. ■

The first two parts of the previous result say that addition and multiplication are continuous, where continuity is as defined according to the notion of convergence in Definition 3.5.3.

One can also perform calculus for $\mathbb{R}$-formal power series without having to worry about the analytical problems concerning limits in $\mathbb{R}$. To do so, we simply "pretend" that an element of $\mathbb{R}[[\xi]]$ can be differentiated and integrated term-by-term with respect to ξ. After one is finished pretending, then one makes the following definition.

3.5.8 Definition (Differentiation and integration of $\mathbb{R}$-formal power series) Let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ be a $\mathbb{R}$-formal power series.

(i) The ***derivative*** of A is the $\mathbb{R}$-formal power series $A' = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by $b_j = (j + 1)a_{j+1}$, $j \in \mathbb{Z}_{\geq 0}$.

(ii) The ***integral*** of A is the $\mathbb{R}$-formal power series $\int A = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$b_j = \begin{cases} 0, & j = 0, \\ \frac{a_{j-1}}{j}, & j \in \mathbb{Z}_{>0}. \end{cases}$$ •

In terms of the indeterminate representation of a $\mathbb{R}$-formal power series, we have the following representation. If $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ is a $\mathbb{R}$-formal power series, then

$$A' = \Big(\sum_{j=0}^{\infty} a_j \xi^j\Big)' = \sum_{j=1}^{\infty} j a_j \xi^{j-1} = \sum_{j=0}^{\infty} (j+1) a_{j+1} \xi^j.$$

This is simply termwise differentiation with respect to the indeterminate. Note that in this case we can ignore the matter of whether it is valid to switch the sum and the derivative since we are not actually talking about functions. Similar statements hold, of course, for the integral of a $\mathbb{R}$-formal power series.

For this derivative operation, one has the usual rules.

3.5.9 Proposition (Properties of differentiation and integration of $\mathbb{R}$-formal power series) *Let* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *and* $\mathrm{B} = (\mathrm{b_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *be* $\mathbb{R}$*-formal power series and let* $\alpha \in \mathbb{R}$*. Then the following statements hold:*

(i) $(\mathrm{A} + \mathrm{B})' = \mathrm{A}' + \mathrm{B}'$*;*

(ii) $(\mathrm{A} \cdot \mathrm{B})' = \mathrm{A}' \cdot \mathrm{B} + \mathrm{A} \cdot \mathrm{B}'$*;*

(iii) $(\alpha\, \mathrm{A})' = \alpha\, \mathrm{A}'$*;*

(iv) $\int(\mathrm{A} + \mathrm{B}) = \int \mathrm{A} + \int \mathrm{B}$*;*

(v) $\int(\alpha\, \mathrm{A}) = \alpha \int \mathrm{A}$.

Proof The second statement is the only possibly nontrivial one, so it is the only thing we will prove. We note that

$$A \cdot B = \sum_{k=0}^{\infty} \Big(\sum_{j=0}^{k} a_j b_{k-j}\Big) \xi^k,$$

so that

$$\begin{aligned} (A \cdot B)' &= \sum_{k=1}^{\infty} \Big(\sum_{j=0}^{k} a_j b_{k-j}\Big) k \xi^{k-1} \\ &= \sum_{k=0}^{\infty} \Big(\sum_{j=0}^{k} (j+1) a_{j+1} b_{k-j}\Big) \xi^k + \sum_{k=0}^{\infty} \Big(\sum_{j=0}^{k} (j+1) a_{k-j} b_{j+1}\Big) \xi^k \\ &= A' \cdot B + A \cdot B', \end{aligned}$$

as desired. ■

The derivative also commutes with limits, as one would hope to be the case.

3.5.10 Proposition (Differentiation and integration, and convergence in $\mathbb{R}[[\xi]]$) *If $(A_k = (a_{k,j})_{j\in\mathbb{Z}_{\geq 0}})_{k\in\mathbb{Z}_{>0}}$ is a sequence in $\mathbb{R}[[\xi]]$ converging to A, then $A' = \lim_{k\to\infty} A'_k$ and $\int A = \lim_{k\to\infty} \int A_k$.*

Proof This is a more or less obvious result, given the definition of convergence of $\mathbb{R}$-formal power series. ■

Now that we have finished playing algebraic games, we turn to the matter of when a formal power series actually represents a function.

3.5.2 $\mathbb{R}$-convergent power series

The one thing that we did not do in the preceding section is think of $\mathbb{R}$-formal power series as functions. This is because not all $\mathbb{R}$-formal power series *can* be thought of as functions. For example, if $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ is the $\mathbb{R}$-formal power series defined by $a_j = j!$, $j \in \mathbb{Z}_{\geq 0}$, then the series $\sum_{j=1}^{\infty} a_j x^j$ diverges for any $x \in \mathbb{R} \setminus \{0\}$. In this section we address this matter by thinking of power series as being series of functions, just as we discussed in Section 3.4.4.

First we classify $\mathbb{R}$-formal power series according to the convergence properties possessed by the corresponding series of functions.

3.5.11 Proposition (Classification of $\mathbb{R}$-formal power series by convergence) *For each $\mathbb{R}$-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$, exactly one of the following statements holds:*

(i) the series $\sum_{j=0}^{\infty} a_j x^j$ converges absolutely for all $x \in \mathbb{R}$;

(ii) the series $\sum_{j=0}^{\infty} a_j x^j$ diverges for all $x \in \mathbb{R} \setminus \{0\}$;

(iii) there exists $R \in \mathbb{R}_{>0}$ such that the series $\sum_{j=0}^{\infty} a_j x^j$ converges absolutely for all $x \in \mathsf{B}(R, 0)$, and diverges for all $x \in \mathbb{R} \setminus \overline{\mathsf{B}}(R, 0)$.

Proof First let us prove a lemma.

1 Lemma *If the series $\sum_{j=0}^{\infty} a_j x_0^j$ converges for some $x_0 \in \mathbb{R}$, then the series $\sum_{j=0}^{\infty} a_j x^j$ converges absolutely for $x \in \mathsf{B}(|x_0|, 0)$.*

Proof Note that the sequence $(a_j x_0^j)_{j\in\mathbb{Z}_{\geq 0}}$ converges to zero, and so is bounded by Proposition 2.3.4. Thus let $M \in \mathbb{R}_{>0}$ have the property that $|a_j x_0^j| \leq M$ for each $j \in \mathbb{Z}_{\geq 0}$. Then, for $x \in \mathsf{B}(|x_0|, 0)$, we have

$$|a_j x^j| = |a_j x_0^j| \left|\frac{x}{x_0}\right|^j \leq M \left|\frac{x}{x_0}\right|^j, \qquad j \in \mathbb{Z}_{\geq 0}.$$

Since $\left|\frac{x}{x_0}\right| < 1$ the series $\sum_{j=0}^{\infty} M\left|\frac{x}{x_0}\right|$ converges as shown in Example 2.4.2–1. Therefore, by the Comparison Test, the series $\sum_{j=0}^{\infty} a_j x^j$ converges absolutely for $x \in \mathsf{B}(|x_0|, 0)$. ▼

Now let

$$R = \sup\left\{x \in \mathbb{R}_{\geq 0} \;\middle|\; \sum_{j=0}^{\infty} a_j x^j \text{ converges}\right\}.$$

We have three cases.

1. $R = \infty$: For $x \in \mathbb{R}$ choose $x_0 > 0$ such that $|x| < x_0$. By the lemma, the series $\sum_{j=0}^{\infty} a_j x^j$ converges absolutely. This is case (i) of the statement of the result.

2. $R = 0$: Let $x \in \mathbb{R} \setminus \{0\}$ and choose $x_0 > 0$ such that $|x| > x_0$. If $\sum_{j=0}^{\infty} a_j x^j$ converges, then by the lemma, the series $\sum_{j=0}^{\infty} a_j x_0^j$ converges absolutely, and so converges. But this contradicts the definition of R, so the series $\sum_{j=0}^{\infty} a_j x^j$ must diverge for every nonzero $x \in \mathbb{R}$. This is case (ii) of the statement of the result.
3. $R \in \mathbb{R}_{>0}$: If $x \in \mathsf{B}(R, 0)$ then, by the lemma, the series $\sum_{j=0}^{\infty} a_j x^j$ converges absolutely. If $x \in \mathbb{R} \setminus \overline{\mathsf{B}}(R, 0)$ then there exists $x_0 > R$ such that $|x| > x_0$. If the series $\sum_{j=0}^{\infty} a_j x^j$ converges, then by the lemma the series $\sum_{j=0}^{\infty} a_j x_0^j$ converges absolutely, and so converges. But this contradicts the definition of R. This is case (iii) of the statement of the result.

These three possibilities clearly are exhaustive and mutually exclusive. ■

Now we can sensibly define what we mean by a power series that converges.

3.5.12 Definition (ℝ-convergent power series) A $\mathbb{R}$-formal power series $(a_j)_{j \in \mathbb{Z}_{\geq 0}}$ is a ***$\mathbb{R}$-convergent power series*** if it falls into either case (i) or (iii) of Proposition 3.5.11. •

One can also say that a $\mathbb{R}$-formal power series that is not convergent has a zero radius of convergence, and sometimes it will be convenient to use this language.

Of course, one is interested in actually determining whether a given $\mathbb{R}$-formal power series is convergent or not. It turns out that this is actually possible, as the following result indicates.

3.5.13 Theorem (Cauchy–Hadamard[11] test for power series convergence) *Let* $(\mathrm{a_j})_{\mathrm{j} \in \mathbb{Z}_{\geq 0}}$ *be a* $\mathbb{R}$*-formal power series, and define* $\rho \in \overline{\mathbb{R}}_{\geq 0}$ *by* $\rho = \limsup_{\mathrm{j} \to \infty} |\mathrm{a_j}|^{1/\mathrm{j}}$. *Then define* $\mathrm{R} \in \overline{\mathbb{R}}_{\geq 0}$ *by*

$$\mathrm{R} = \begin{cases} \infty, & \rho = 0, \\ \frac{1}{\rho}, & \rho \in \mathbb{R}_{>0}, \\ 0, & \rho = \infty. \end{cases}$$

Then R *is the radius of convergence for* $(\mathrm{a_j})_{\mathrm{j} \in \mathbb{Z}_{\geq 0}}$.

Proof Let $x \in \mathbb{R}$. We have

$$\limsup_{j \to \infty} |a_j x^j|^{1/j} = \limsup_{j \to \infty} |x| |a_j|^{1/j} = |x| \rho.$$

Now, by the Root Test, $\sum_{j=0} a_j x^j$ converges if $|x|\rho < 1$ and diverges if $|x|\rho > 1$. From these statements, the result follows. ■

Note that in Proposition 3.5.11 we make no assertions about the convergence of power series for values of x whose magnitude us equal to the radius of convergence.

[11] Jacques Salomon Hadamard (1865–1963) was a French mathematician. He made significant contributions to the fields of complex analysis, number theory, differential equations, geometry and linear algebra.

3.5.14 Definition (Region of (absolute) convergence) Let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ be a $\mathbb{R}$-formal power series and consider the classification of Proposition 3.5.11. In case (i) the ***radius of convergence*** is ∞, and in case (iii) the ***radius of convergence*** is the positive number R asserted in the statement of the proposition. The ***region of absolute convergence*** is $\mathscr{R}_{\text{abs}}(A) = (-R, R)$, and the region of convergence is the largest interval $\mathscr{R}_{\text{conv}}(A) \subseteq \mathbb{R}$ on which the series $\sum_{j=0}^{\infty} a_j x^j$ converges. •

Note that the region of convergence could be either $(-R, R)$, $[-R, R)$, $(-R, R]$, or $[-R, R]$. The following examples show that all possibilities are realised.

3.5.15 Examples (Region of (absolute) convergence)

1. Consider the $\mathbb{R}$-formal power series $A = (a_j = \frac{1}{2^j j^2})_{j\in\mathbb{Z}_{>0}}$ (take $a_0 = 0$). We compute
$$\lim_{j\to\infty}\Big|\frac{a_{j+1}}{a_j}\Big| = \lim_{j\to\infty}\Big|\frac{2^j j^2}{2^{j+1}(j+1)^2}\Big| = \frac{1}{2}.$$
By Proposition 2.4.15 we conclude that the radius of convergence of the power series $\sum_{j=1}^{\infty} \frac{x^j}{2^j j^2}$ is 2. When $x = 2$ the series becomes $\sum_{j=1}^{\infty} \frac{1}{j^2}$, which we know converges by Example 2.4.2–4. When $x = -2$ the series becomes $\sum_{j=1}^{\infty} \frac{(-1)^j}{j^2}$, which again is convergent, this time by the Alternating Test. Thus $\mathscr{R}_{\text{abs}}(A) = (-2, 2)$, while $\mathscr{R}_{\text{conv}}(A) = [-2, 2]$.
2. Now consider the $\mathbb{R}$-formal power series $A = (a_j = \frac{1}{2^j j})_{j\in\mathbb{Z}_{>0}}$ (take $a_0 = 0$). We again use Proposition 2.4.15 and the computation
$$\lim_{j\to\infty}\Big|\frac{a_{j+1}}{a_j}\Big| = \lim_{j\to\infty}\Big|\frac{2^j j}{2^{j+1}(j+1)}\Big| = \frac{1}{2}$$
to deduce that this power series has radius of convergence 2. For $x = 2$ the series becomes $\sum_{j=1}^{\infty} \frac{1}{j}$ which diverges by Example 2.4.2–4, and for $x = -2$ the series becomes $\sum_{j=1}^{\infty} \frac{(-1)^j}{j}$ which converges by Example 2.4.2–3. Thus $\mathscr{R}_{\text{abs}}(A) = (-2, 2)$, while $\mathscr{R}_{\text{conv}}(A) = [-2, 2)$.
3. Now we define the $\mathbb{R}$-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ by
$$a_j = \begin{cases} 0, & j = 0, \\ 0, & j \text{ odd}, \\ \frac{2}{2^{-\frac{j}{2}} j}, & \text{otherwise.} \end{cases}$$
Thus the corresponding series is $\sum_{k=1}^{\infty} \frac{x^{2k}}{2^k k}$. We have
$$\limsup_{j\to\infty} |a_j|^{1/j} = \limsup_{k\to\infty}\Big|\frac{1}{2^k k}\Big|^{1/2k} = \frac{1}{\sqrt{2}}\lim_{k\to\infty}\Big(\frac{1}{k}\Big)^{1/2k} = \frac{1}{\sqrt{2}}.$$
Thus the radius of convergence is $\sqrt{2}$. For $x = \pm\sqrt{2}$ the series becomes $\sum_{k=1}^{\infty} \frac{1}{k}$ which diverges. Thus $\mathscr{R}_{\text{abs}}(A) = \mathscr{R}_{\text{conv}}(A) = (-\sqrt{2}, \sqrt{2})$. •

An important property of $\mathbb{R}$-convergent power series, is that, not only do they converge absolutely, they converge uniformly on any compact interval in the region of absolute convergence.

3.5.16 Theorem (Uniform convergence of $\mathbb{R}$-convergent power series) *If* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *is a* $\mathbb{R}$*-convergent power series, then the series* $\sum_{\mathrm{j}=0}^{\infty} \mathrm{a_j x^j}$ *converges uniformly on any compact interval* $\mathrm{J} \subseteq \mathscr{R}_{\mathrm{abs}}(\mathrm{A})$.

Proof It suffices to consider the case where $J = [-R_0, R_0]$ since any compact interval will be contained in an interval of this form. Let $x \in [-R_0, R_0]$. Since $\sum_{j=0}^{\infty} a_j R_0^j$ converges absolutely and since $|a_j x^j| \leq a_j R_0^j$, uniform convergence follows from the Weierstrass M-test. ■

The next result gives the value of the limit function at points in the boundary of the region of convergence.

3.5.17 Theorem (Continuous extension to region of convergence) *Let* $(\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *be a* $\mathbb{R}$*-convergent power series with radius of convergence* R. *If the series* $\sum_{\mathrm{j}=0}^{\infty} \mathrm{a_j R^j}$ *(resp.* $\sum_{\mathrm{j}=0}^{\infty} \mathrm{a_j}(-\mathrm{R})^{\mathrm{j}}$*) converges, then*

$$\lim_{\mathrm{x}\uparrow \mathrm{R}} \sum_{\mathrm{j}=0}^{\infty} \mathrm{a_j x^j} = \sum_{\mathrm{j}=0}^{\infty} \mathrm{a_j R^j} \qquad \Big(\textit{resp.} \lim_{\mathrm{x}\downarrow -\mathrm{R}} \sum_{\mathrm{j}=0}^{\infty} \mathrm{a_j x^j} = \sum_{\mathrm{j}=0}^{\infty} \mathrm{a_j}(-\mathrm{R^j})\Big).$$

Proof We shall only prove the theorem in the limit as x approaches R; the other case follows entirely similarly (or by a change of variable from x to $-x$). Denote by $f\colon \mathsf{B}(R,0) \to \mathbb{R}$ the limit function for the power series. Let $S_{-1} = 0$ and for $k \in \mathbb{Z}_{\geq 0}$ define

$$S_k = \sum_{j=0}^{k} a_j R^j.$$

We then directly have

$$\sum_{j=0}^{k} a_j x^j = \sum_{j=0}^{k} (S_j - S_{j-1})(\tfrac{x}{R})^j = (1 - \tfrac{x}{R}) \sum_{j=0}^{k-1} S_j (\tfrac{x}{R})^j + S_k (\tfrac{x}{R})^k.$$

For $x \in \mathsf{B}(R,0)$ we note that $\lim_{k\to\infty} S_k (\frac{x}{R})^k = 0$, and therefore

$$f(x) = \sum_{j=0}^{\infty} a_j x^j = (1 - \tfrac{x}{R}) \sum_{j=0}^{\infty} S_j (\tfrac{x}{R})^j.$$

If $S = \lim_{j\to\infty} S_j$, for $\epsilon \in \mathbb{R}_{>0}$ take $N \in \mathbb{Z}_{>0}$ such that $|S - S_j| < \frac{\epsilon}{2}$ for $j \geq N$. Note that, from Example 2.4.2–1, we have

$$(1 - \tfrac{x}{R}) \sum_{j=0}^{\infty} (\tfrac{x}{R})^j = 1$$

for $x \in \mathsf{B}(R,0)$. It therefore follows that for $x \in (0, R)$ we have

$$(1 - \tfrac{x}{R}) \sum_{j=N+1}^{\infty} |S_j - S| (\tfrac{x}{R})^j \leq \tfrac{\epsilon}{2} (1 - \tfrac{x}{R}) \sum_{j=N+1}^{\infty} (\tfrac{x}{R})^j < \tfrac{\epsilon}{2}. \tag{3.15}$$

Now let $\delta \in \mathbb{R}_{>0}$ have the property that for $x \in (R-\delta, R)$

$$(1 - \tfrac{x}{R}) \sum_{j=0}^{N} |S_j - S| < \tfrac{\epsilon}{2}.$$

It therefore follows that for $x \in (R-\delta, R)$ we also have

$$(1 - \tfrac{x}{R}) \sum_{j=0}^{N} |S_j - S| (\tfrac{x}{R})^j < \tfrac{\epsilon}{2}. \tag{3.16}$$

We therefore obtain, for $x \in (R-\delta, R)$,

$$\begin{aligned} |f(x) - S| &= \left|(1 - \tfrac{x}{R}) \sum_{j=0}^{\infty} (S_j - S)(\tfrac{x}{R})^j\right| \leq (1 - \tfrac{x}{R}) \sum_{j=0}^{\infty} |S_j - S| (\tfrac{x}{R})^j \\ &\leq (1 - \tfrac{x}{R}) \sum_{j=0}^{N} |S_j - S| (\tfrac{x}{R})^j + (1 - \tfrac{x}{R}) \sum_{j=N+1}^{\infty} |S_j - S| (\tfrac{x}{R})^j \\ &< \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon, \end{aligned}$$

using (3.15) and (3.16). It therefore follows that $\lim_{x \uparrow R} f(x) = S$, as desired. ■

The preceding two theorems have the following important corollary.

3.5.18 Corollary ($\mathbb{R}$-convergent power series have a continuous limit function) *If $\mathrm{A} = (\mathrm{a}_j)_{j \in \mathbb{Z}_{\geq 0}}$ is a $\mathbb{R}$-convergent power series, then the limit function on $\mathscr{R}_{\mathrm{conv}}(\mathrm{A})$ is continuous.*

Proof This follows immediately from the previous two theorems along with Theorem 3.4.8. ■

3.5.3 $\mathbb{R}$-convergent power series and operations on functions

In this section we explore how various operations on functions interact with power series. The results in this section have the usual mundane character of other similar sections in this chapter. However, it is worth noting that there is one rather spectacular conclusion that emerges, namely that the limit function of a $\mathbb{R}$-convergent power series is infinitely differentiable. The significance of this is perhaps not to be fully appreciated until we realise that, when this conclusion is extended to power series for complex functions, it allows the correspondence between analytic functions and power series.

But first some mundane things.***missing stuff***

3.5.19 Proposition (Addition and multiplication, and $\mathbb{R}$-convergent power series) *If $\mathrm{A} = (\mathrm{a}_j)_{j \in \mathbb{Z}_{\geq 0}}$ and $\mathrm{B} = (\mathrm{b}_j)_{j \in \mathbb{Z}_{\geq 0}}$ are $\mathbb{R}$-convergent power series, then the following statements hold:*

(i) $\mathscr{R}_{\mathrm{conv}}(\mathrm{A}+\mathrm{B}) \subseteq \mathscr{R}_{\mathrm{conv}}(\mathrm{A}) \cap \mathscr{R}_{\mathrm{conv}}(\mathrm{B})$*, and so, in particular,* $\mathrm{A}+\mathrm{B}$ *is a $\mathbb{R}$-convergent power series;*

(ii) $\mathscr{R}_{\text{conv}}(A \cdot B) \subseteq \mathscr{R}_{\text{conv}}(A) \cap \mathscr{R}_{\text{conv}}(B)$, *and so, in particular,* $A \cdot B$ *is a* $\mathbb{R}$*-convergent power series.*

Proof This follows immediately from Proposition 2.4.30. ∎

In the language of Section **??**, the preceding result says that the set of $\mathbb{R}$-convergent power series is a subring of the set of $\mathbb{R}$-formal power series. This in and of itself is not hugely interesting. However, the exact properties of the ring of $\mathbb{R}$-convergent power series *is* of quite some importance in the study of analytic functions; we refer the reader to Section 3.5.5 for further discussion and references.

3.5.20 Proposition (Differentiation and integration of $\mathbb{R}$-convergent power series) *If* $A = (a_j)_{j \in \mathbb{Z}_{\geq 0}}$ *is a* $\mathbb{R}$*-convergent power series, then the following statements hold:*

(i) $\mathscr{R}_{\text{abs}}(A') = \mathscr{R}_{\text{abs}}(A)$, *and so, in particular,* A' *is a* $\mathbb{R}$*-convergent power series;*

(ii) $\mathscr{R}_{\text{abs}}(\int A) = \mathscr{R}_{\text{abs}}(A)$, *and so, in particular,* $\int A$ *is a* $\mathbb{R}$*-convergent power series.*

Furthermore, if the series defined by A *converges to* $f \colon \mathscr{R}_{\text{abs}}(A) \to \mathbb{R}$, *then the series defined by* A' *converges to* f' *on* $\mathscr{R}_{\text{abs}}(A)$ *and the series defined by* $\int A$ *converges to the function* $x \mapsto \int_0^x f(\xi)\, d\xi$ *on* $\mathscr{R}_{\text{abs}}(A)$.

Proof That $\mathscr{R}_{\text{abs}}(A') = \mathscr{R}_{\text{abs}}(A)$ and $\mathscr{R}_{\text{abs}}(\int A) = \mathscr{R}_{\text{abs}}(A)$ follows since $\lim_{j\to\infty} j^{1/j} = \lim_{j\to\infty}(\frac{1}{j})^{1/j} = 1$ by Proposition 3.6.12, allowing us to conclude that

$$\limsup_{j\to\infty} |j a_j|^{1/j} = \limsup_{j\to\infty} |a_j|^{1/j}, \quad \limsup_{j\to\infty} \left|\frac{a_j}{j}\right|^{1/j} = \limsup_{j\to\infty} |a_j|^{1/j}.$$

That the series defined by A' and $\int A$ have the properties stated follows from Theorems 3.4.23 and 3.4.24, along with the definitions of A' and $\int A$. ∎

This gives the following remarkable corollary concerning the character of the limit function for $\mathbb{R}$-convergent power series.

3.5.21 Corollary (Limits of $\mathbb{R}$-convergent power series are infinitely differentiable) *If* $A = (a_j)_{j \in \mathbb{Z}_{\geq 0}}$ *is a* $\mathbb{R}$*-convergent power series converging to* $f \colon \mathscr{R}_{\text{abs}}(A) \to \mathbb{R}$, *then* f *is infinitely differentiable on* $\mathscr{R}_{\text{abs}}(A)$, *and* $a_j = \frac{f^{(j)}(0)}{j!}$.

Proof This follows simply by a repeated application of Proposition 3.5.20, and by performing term-by-term differentiation, and evaluating the resulting expressions at 0. ∎

3.5.4 Taylor series

In the preceding section we indicated how, for the special class of $\mathbb{R}$-formal power series that are convergent, one can construct a limit function that is infinitely differentiable. In this section we consider the possibility of "reversing" this operation, and producing a $\mathbb{R}$-formal power series from an infinitely differentiable function. Even in cases when a function is not infinitely differentiable, we shall attempt to approximate it using a truncated power series. What we shall see in this section is that the correspondence between functions and the power series which

purport to approximate them is a complicated one. Indeed, it is only for a special class of functions, those which we call "real analytic," that this correspondence as a useful one.

Let $I \subseteq \mathbb{R}$ be an interval and let $x_0 \in \text{int}(I)$. Suppose that $f\colon I \to \mathbb{R}$ is infinitely differentiable. If one takes as the final objective the idea that we wish to approximate f near x_0. If $x_0 = 0$ then we might like to write

$$f(x) = \sum_{j=0}^{\infty} a_j x^j.$$

For $x_0 \neq 0$ it makes sense to write this approximation as

$$f(x) = \sum_{j=0}^{\infty} a_j (x - x_0)^j.$$

Indeed, if we write our approximation in this way, and then believe that differentiation can be performed term-by-term on the right, we obtain

$$f(x_0) = a_0,\ f^{(1)}(x_0) = a_1,\ f^{(2)}(x_0) = 2a_2, \ldots, f^{(j)}(x_0) = j!a_j, \ldots$$

With this as motivation, we make the following definition.

3.5.22 Definition (Taylor polynomial and Taylor series) Let $I \subseteq \mathbb{R}$ be an interval, let $x_0 \in \text{int}(I)$, and let $f\colon I \to \mathbb{R}$ be r-times differentiable for $r \in \mathbb{Z}_{>0} \cup \{\infty\}$.

(i) For $k \leq r$, the ***Taylor polynomial*** of degree k for f about x_0 is the polynomial function $\mathscr{T}_k(f, x_0)$ defined by

$$\mathscr{T}_k(f, x_0)(x) = \sum_{j=0}^{k} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j.$$

(ii) If $r = \infty$ then the ***Taylor series*** for f about x_0 is the $\mathbb{R}$-formal power series $\mathscr{T}_\infty(f, x_0) = (\frac{f^{(j)}(x_0)}{j!})_{j \in \mathbb{Z}_{\geq 0}}$. •

Sometimes it can be tedious to compute the derivatives needed to explicitly exhibit the Taylor polynomial or the Taylor series. In some cases, the following result is helpful.

3.5.23 Proposition (Property of Taylor polynomial) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{r} \in \mathbb{Z}_{>0}$, *and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a function that is* r*-times differentiable with* $\mathrm{f}^{(\mathrm{r})}$ *locally bounded. If* $\mathrm{x}_0 \in \mathrm{I}$ *and if* $\mathrm{P}\colon \mathrm{I} \to \mathbb{R}$ *is a polynomial function of degree* $\mathrm{r} - 1$, *then* $\mathrm{P} = \mathscr{T}_{\mathrm{r}-1}(\mathrm{f}, \mathrm{x}_0)$ *if and only if*

$$\lim_{\mathrm{x} \to_{\mathrm{I}} \mathrm{x}_0} \frac{\mathrm{f}(\mathrm{x}) - \mathrm{P}(\mathrm{x})}{(\mathrm{x} - \mathrm{x}_0)^{\mathrm{r}-1}} = 0.$$

Proof We will use Taylor's Theorem stated below. Suppose that $P = \mathscr{T}_{r-1}(f, x_0)$. Then, by Taylor's Theorem, for x in a neighbourhood of x_0, we have

$$|f(x) - P(x)| \leq M|x - x_0|^r \quad \Longrightarrow \quad \lim_{x \to_I x_0} \left| \frac{f(x) - P(x)}{(x - x_0)^{r-1}} \right| \leq \lim_{x \to_I x_0} M|x - x_0| = 0.$$

Now suppose that

$$\lim_{x \to_I x_0} \frac{f(x) - P(x)}{(x - x_0)^{r-1}} = 0.$$

By Taylor's Theorem, write

$$f(x) = \mathscr{T}_{r-1}(f, x_0)(x) + R_r(f, x_0)(x),$$

where $R_r(f, x_0)(x)$ is a function defined in a neighbourhood of x_0 satisfying $|R_r(f, x_0)(x)| \leq M|x - x_0|^r$. Then, using Exercise 2.2.7,

$$\begin{aligned}
&\lim_{x \to_I x_0} \left| \frac{f(x) - P(x)}{(x - x_0)^{r-1}} \right| = 0, \\
\Longrightarrow \quad &\lim_{x \to_I x_0} \left| \frac{\mathscr{T}_{r-1}(f, x_0)(x) + R_r(f, x_0)(x) - P(x)}{(x - x_0)^{r-1}} \right| = 0, \\
\Longrightarrow \quad &\lim_{x \to_I x_0} \left| \left| \frac{\mathscr{T}_{r-1}(f, x_0)(x) - P(x)}{(x - x_0)^{r-1}} \right| - \left| \frac{R_r(f, x_0)(x)}{(x - x_0)^{r-1}} \right| \right| = 0.
\end{aligned}$$

Since

$$\lim_{x \to_I x_0} \left| \frac{R_r(f, x_0)(x)}{(x - x_0)^{r-1}} \right| = 0$$

by the properties of $R_r(f, x_0)$, we conclude that

$$\lim_{x \to_I x_0} \left| \frac{\mathscr{T}_{r-1}(f, x_0)(x) - P(x)}{(x - x_0)^{r-1}} \right| = 0.$$

If P and $\mathscr{T}_{r-1}(f, x_0)$ were distinct degree $r - 1$ polynomials, then we would either have

$$\lim_{x \to_I x_0} \left| \frac{\mathscr{T}_{r-1}(f, x_0)(x) - P(x)}{(x - x_0)^{r-1}} \right| = \alpha > 0, \quad \text{or} \quad \lim_{x \to_I x_0} \left| \frac{\mathscr{T}_{r-1}(f, x_0)(x) - P(x)}{(x - x_0)^{r-1}} \right| = \infty.$$

Thus the result follows. ∎

The way to interpret the result is that the Taylor polynomial of degree k about x_0 provides the best (in some sense) degree k polynomial approximation to f near x_0. In this sense, the Taylor polynomial can be thought of as the generalisation of the derivative, the derivative providing the best linear approximation of a function.

There are two fundamentally different sorts of questions arising from the notions of the Taylor polynomial and the Taylor series.

1. Is the Taylor series for an infinitely differentiable function a $\mathbb{R}$-convergent power series?
2. (a) Does the Taylor polynomial approximate f in some sense?
   (b) If f is infinitely differentiable and the Taylor series *is* a $\mathbb{R}$-convergent power series, does it approximate f in some sense?

Before we proceed to explore these questions in detail, let us give a definition which they immediately suggest.

3.5.24 Definition (Real analytic function) Let $I \subseteq \mathbb{R}$ be an interval, let $x_0 \in I$, and let $f\colon I \to \mathbb{R}$ be an infinitely differentiable function with Taylor series $\mathscr{T}_\infty(f, x_0) = (\frac{f^{(j)}(x_0)}{j!})_{j\in\mathbb{Z}_{\geq 0}}$. We say that f is ***real analytic*** at x_0 if $\mathscr{T}_\infty(f, x_0)$ is a $\mathbb{R}$-convergent power series, and if there exists a neighbourhood U of x_0 such that

$$f(x) = \sum_{j=0}^{\infty} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j$$

for all $x \in U$. •

Thus real analytic functions are exactly those that are perfectly approximated by their Taylor series. What is not clear at this time is whether "real analytic" is actually different than "infinitely differentiable." The following result addresses this in rather dramatic fashion.

3.5.25 Theorem (Borel's Theorem) *If* $(\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *is a* $\mathbb{R}$*-formal power series, then there exists an interval* $\mathrm{I} \subseteq \mathbb{R}$ *with* $0 \in \mathrm{int(I)}$ *and a function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *of class* C^∞ *such that* $\mathscr{T}_\infty(\mathrm{f}, 0) = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$.

Proof Define $\curlywedge\colon [-1, 1] \to \mathbb{R}$ by

$$\curlywedge(x) = \begin{cases} 0, & x \in \{-1, 1\}, \\ \mathrm{e}^{-\frac{1}{1-x^2}}\mathrm{e}, & x \in (-1, 1), \end{cases}$$

and note that

1. $\curlywedge$ is infinitely differentiable,
2. $\curlywedge(\pm 1) = 0$,
3. $\curlywedge(0) = 1$, and
4. $\curlywedge(x) \in (0, 1)$ for $|x| \in (0, 1)$.

(We refer the reader to Example 3.5.28–2 for the details concerning this function.) We take $I = [-1, 1]$ and, for $\epsilon \in (0, 1)$, define $g_\epsilon\colon I \to \mathbb{R}$ by

$$g_\epsilon(x) = \begin{cases} 0, & |x| \in [\epsilon, 1], \\ \curlywedge(1 + \frac{2x}{\epsilon}), & x \in (-\epsilon, -\frac{\epsilon}{2}), \\ \curlywedge(-1 + \frac{2x}{\epsilon}), & x \in (\frac{\epsilon}{2}, \epsilon), \\ 1, & |x| \in [0, \frac{\epsilon}{2}]. \end{cases}$$

Then, for $k \in \mathbb{Z}_{\geq 0}$, define $f_{\epsilon,k}\colon I \to \mathbb{R}$ inductively by taking $f_{\epsilon,0} = g_\epsilon$ and

$$f_{\epsilon,k}(x) = \int_0^x f_{\epsilon,k-1}(\xi)\,\mathrm{d}\xi.$$

Note that

1. $f_{\epsilon,k}^{(j)}(0) = 0$, $j \in \{0, 1, \ldots, k-1\}$,
2. $f_{\epsilon,k}^{(k)}(0) = 1$, and
3. $\left|f_{\epsilon,k}^{(j)}(x)\right| \leq \epsilon$ for $j \in \{0, 1, \ldots, k-1\}$ and $x \in I$.

Now let $(\epsilon_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}_{>0}$ for which the series $\sum_{j=0}^{\infty}|a_j|\epsilon_j$ converges. We claim that if

$$f(x) = \sum_{j=0}^{\infty} a_j f_{\epsilon_j,j}(x),$$

then f is well-defined and infinitely differentiable on $[-1,1]$, and has the property that $\mathscr{T}_\infty(f,0) = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$.

By our choice of the sequence $(\epsilon_j)_{j\in\mathbb{Z}_{\geq 0}}$, it follows from the Weierstrass M-test that f is well-defined by virtue of the absolute and uniform convergence of the series $\sum_{j=0}^{\infty} a_j f_{\epsilon_j,j}(x)$ for $x \in [-1,1]$. Moreover, the hypotheses of Theorem 3.4.24 hold, and so the series can be differentiated term-by-term. One may then directly verify that the Weierstrass M-test again ensures that the resulting differentiated series is again uniformly convergent. This argument may be repeated to show that f is infinitely differentiable, and the series for the kth derivative is the kth derivative of the series taken term-by-term. One now uses the properties of the functions $f_{\epsilon,j}$, $j \in \mathbb{Z}_{\geq 0}$, to directly verify that $\mathscr{T}_\infty(f,0) = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$. We leave the tedious, but direct, checking of the details of the assertions in this paragraph to the reader. ■

This result, therefore, rules out any sort of complete correspondence between a function and its Taylor series. Indeed, it even rules out the convergence of Taylor series.

It is clear, then, that a real analytic must have a rather specific character to its Taylor series. The following result precisely characterises this.

3.5.26 Theorem (Derivatives of real analytic functions) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an open interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is infinitely differentiable, then the following statements are equivalent:*

(i) f *is real analytic;*

(ii) for each $\mathrm{x}_0 \in \mathrm{I}$ *there exists a neighbourhood* $\mathrm{U} \subseteq \mathrm{I}$ *of* x_0 *and* $\mathrm{C}, \mathrm{r} \in \mathbb{R}_{>0}$ *such that*

$$|\mathrm{f}^{(\mathrm{m})}(\mathrm{x})| \leq \mathrm{C}\mathrm{m}!\mathrm{r}^{-\mathrm{m}}$$

for all $\mathrm{x} \in \mathrm{U}$ *and* $\mathrm{m} \in \mathbb{Z}_{\geq 0}$.

Proof First suppose that f is real analytic and let $x_0 \in I$. Let $\delta \in \mathbb{R}_{>0}$ be such that

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k, \qquad |x - x_0| < \delta.$$

This implies that, for each $\rho \in (0,\delta)$, the sequence $(a_k\rho^k)_{k\in\mathbb{Z}_{\geq 0}}$ is bounded, say by $C' \in \mathbb{R}_{>0}$. Therefore, by Corollary 3.5.21 we have

$$|f^{(m)}(x_0)| \leq C' m! \rho^{-m}.$$

Let us fix some $\rho \in (0,\delta)$.

By differentiating the power series for f term-by-term on $\mathsf{B}(x_0,\delta)$ we have

$$\frac{f^{(m)}(t)}{m!} = \frac{1}{m!}\sum_{k=0}^{\infty}(k+1)\cdots(k+m)a_{k+m}(x-x_0)^k = \sum_{k=0}^{\infty}\binom{k+m}{m}a_{k+m}(x-x_0)^k,$$

where

$$\binom{j}{l} = \frac{j!}{l!(j-l)!}$$

is the binomial coefficient defined for $j, l \in \mathbb{Z}_{\geq 0}$ with $j \geq l$. By Exercise 2.2.1 we have

$$2^j = (1+1)^j = \sum_{l=0}^{j} \binom{j}{l}.$$

Therefore,

$$\binom{j}{l} \leq 2^j, \qquad l \in \{0, 1, \dots, j\}.$$

Therefore, if $|x - x_0| < \frac{\rho}{3}$,

$$\left|\frac{f^{(m)}(x)}{m!}\right| \leq C' \rho^{-m} \sum_{k=0}^{\infty} \binom{k+m}{m} \rho^{-k} |x - x_0|^k \leq C' \Big(\frac{\rho}{2}\Big)^{-m} \sum_{k=0}^{\infty} \Big(\frac{2}{3}\Big)^k = 3C' \Big(\frac{\rho}{2}\Big)^{-m},$$

using Example 2.4.2–1. This gives the desired estimate, taking $C = 3C'$ and $r = \frac{\rho}{2}$.

Conversely suppose that for $x_0 \in I$, $|f^{(m)}(x)| \leq Cm!r^{-m}$ for some $C, r \in \mathbb{R}_{>0}$ and for each $m \in \mathbb{Z}_{\geq 0}$. Then, for $|x - x_0| < r$ we have

$$\sum_{k=0}^{\infty} \frac{|f^{(k)}(x_0)|}{k!} |x - x_0|^k \leq C \sum_{k=0}^{\infty} \Big(\frac{|x - x_0|}{r}\Big)^k < \infty$$

by Example 2.4.2–1. Thus the series

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

converges absolutely, and so converges, for each $x \in \mathsf{B}(x_0, r)$. Thus f is real analytic. ■

We now explore the question of how well a Taylor polynomial or Taylor series approximates the function generating it, under suitable hypotheses. We begin with the case where the function f is differentiable to finite order.

3.5.27 Theorem (Taylor's Theorem) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{r} \in \mathbb{Z}_{>0}$, *and let* $\mathrm{f} \colon \mathrm{I} \to \mathbb{R}$ *be a function that is* r*-times differentiable with* $\mathrm{f}^{(\mathrm{r})}$ *locally bounded. Then, if* $[\mathrm{a}, \mathrm{b}] \subseteq \mathrm{I}$ *is a compact interval, there exists* $\mathrm{c} \in [\mathrm{a}, \mathrm{b}]$ *such that*

$$\mathrm{f}(\mathrm{b}) = \mathscr{T}_{\mathrm{r}-1}(\mathrm{f}, \mathrm{a})(\mathrm{b}) + \frac{\mathrm{f}^{(\mathrm{r})}(\mathrm{c})}{\mathrm{r}!} (\mathrm{b} - \mathrm{a})^{\mathrm{r}}.$$

In particular, if $\mathrm{J} \subseteq \mathrm{I}$ *is a compact interval containing* x_0 *then there exists* $\mathrm{M} \in \mathbb{R}_{>0}$ *such that*

$$|\mathrm{f}(\mathrm{x}) - \mathscr{T}_{\mathrm{r}-1}(\mathrm{f}, \mathrm{x}_0)(\mathrm{x})| \leq \mathrm{M} |\mathrm{x} - \mathrm{x}_0|^{\mathrm{r}}$$

for all $\mathrm{x} \in \mathrm{J}$.

Proof Define $\alpha \in \mathbb{R}$ by asking that $f(b) = \mathscr{T}(f,a)(b) + \alpha(b-a)^r$. Now, if for $x \in [a,b]$ we define

$$g(x) = f(x) - \mathscr{T}_{r-1}(f,a)(x) - \alpha(x-a)^r,$$

then we have $g^{(r)}(x) = f^{(r)}(x) - r!\alpha$ since $\mathscr{T}_{r-1}(f,a)$ is a polynomial of degree $r-1$. We directly compute, using the definition of $\mathscr{T}(f,a)$, that $g^{(j)}(a) = 0$ for $j \in \{0,1,\dots,r-1\}$. We also directly have $g(b) = 0$. Therefore, there exists $c_1 \in [a,b]$ such that $g^{(1)}(c_1) = 0$ by the Mean Value Theorem applied to g. We similarly assert the existence of $c_2 \in [a,c_1]$ such that $g^{(2)}(c_2) = 0$, again by the Mean Value Theorem, but now applied to $g^{(1)}$. Continuing in this way we arrive at $c_r \in [a,c_{r-1}]$ such that $g^{(r)}(c_r) = 0$. Taking $c = c_r$, the result follows since $g^{(r)}(x) = f^{(r)}(x) - r!\alpha$. ■

One might be inclined to conjecture that, if f is of class C^∞, then increasing sequences of Taylor polynomials ought to better and better approximate a function. Of course, Theorem 3.5.25 immediately rules this out. The following examples serve to illustrate just how complicated is the correspondence between a function and its Taylor series.

3.5.28 Examples (Taylor series)

1. The first example we give is one of a function that is infinitely differentiable on $\mathbb{R}$, but whose Taylor series about 0 only converges in a bounded neighbourhood of 0.

   We define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = \frac{1}{1+x^2}$. This function, being the quotient of an infinitely differentiable function by a nonvanishing infinitely differentiable function is it self infinitely differentiable. To determine the Taylor series for f, let make an educated guess, and then check it using Proposition 3.5.23. By Example 2.4.2–1 we have, for $x^2 < 1$,

   $$\frac{1}{1+x^2} = \sum_{j=0}^{\infty} (-1)^j x^{2j}.$$

   Let us verify that this is actually the series associated to the Taylor series for f about 0. As we saw during the course of Example 2.4.2–1,

   $$\sum_{j=0}^{k} (-1)^j x^{2j} = \frac{1-(-x^2)^{k+1}}{1+x^2}.$$

   Therefore

   $$\left| \frac{1}{1+x^2} - \sum_{j=0}^{k} (-1)^j x^{2j} \right| = \frac{x^{2k+2}}{1+x^2}.$$

   Thus

   $$\lim_{x\to 0} \left| \frac{\frac{1}{1+x^2} - \sum_{j=0}^{k} (-1)^j x^{2j}}{x^{2k+1}} \right| = 0,$$

and we conclude from Proposition 3.5.23 that $\sum_{j=0}^{k}(-1)^j x^{2j} = \mathscr{T}_{2k+1}(f,0)$. Thus we do indeed have $\mathscr{T}_\infty(f,0) = (a_j)_{j\in\mathbb{Z}_{>0}}$ where

$$a_j = \begin{cases} 0, & j \text{ odd}, \\ (-1)^{j/2}, & j \text{ odd}. \end{cases}$$

By Example 2.4.2–1 the radius of convergence for the Taylor series is 1. Indeed, one easily sees that $\mathscr{R}_{\text{abs}}(\mathscr{T}_\infty(f,0)) = \mathscr{R}_{\text{conv}}(\mathscr{T}_\infty(f,0)) = (-1,1)$.
Thus we indeed have a function, infinitely differentiable on all of $\mathbb{R}$, whose Taylor series converges on a bounded interval. Note that this function *is* real analytic at 0. In fact, one can verify that the function is real analytic everywhere. But even this is not enough to ensure the global convergence of the Taylor series about a given point. In order to understand why the Taylor series for this function does not converge on all of $\mathbb{R}$, it is necessary to understand $\mathbb{C}$-power series, as we do in ***missing stuff***.

2. The next function we construct is one with a Taylor series whose radius of convergence is infinite, but which converges to the function only at one point. We define $f\colon \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} \mathrm{e}^{-\frac{1}{x^2}}, & x \neq 0, \\ 0, & x = 0, \end{cases}$$

and in Figure 3.16 we show the graph of f. We claim that $\mathscr{T}_\infty(f,0)$ is the zero

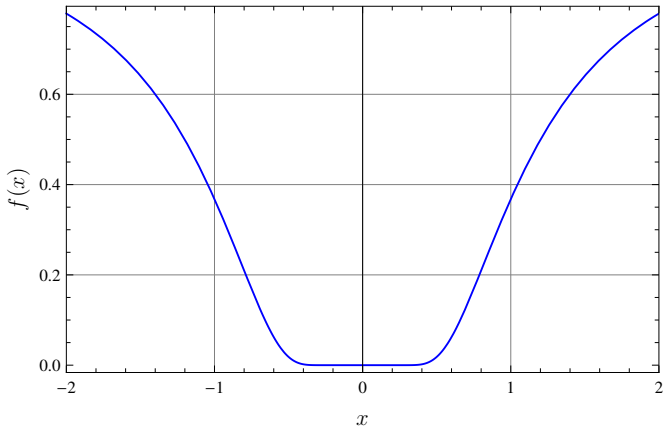


Figure 3.16 A function that is infinitely differentiable but not analytic

$\mathbb{R}$-formal power series. To prove this, we must compute the derivatives of f at $x = 0$. The following lemma is helpful in this regard.

1 Lemma *For* $\mathrm{j} \in \mathbb{Z}_{\geq 0}$ *there exists a polynomial* $\mathrm{p_j}$ *of degree at most* $2\mathrm{j}$ *such that*

$$\mathrm{f^{(j)}(x)} = \frac{\mathrm{p_j(x)}}{\mathrm{x^{3j}}}\mathrm{e}^{-\frac{1}{\mathrm{x}^2}}, \qquad \mathrm{x} \neq 0.$$

Proof We prove this by induction on j. Clearly the lemma holds for $j = 0$ by taking $p_0(x) = 1$. Now suppose the lemma holds for $j \in \{0, 1, \dots, k\}$. Thus

$$f^{(k)}(x) = \frac{p_k(x)}{x^{3k}} e^{-\frac{1}{x^2}}$$

for a polynomial p_k of degree at most $2k$. Then we compute

$$f^{(k+1)}(x) = \frac{x^3 p_k'(x) - 3kx^2 p_k(x) - 2p_k(x)}{x^{3(k+1)}} e^{-\frac{1}{x^2}}.$$

Using the rules for differentiation of polynomials, one easily checks that $x \mapsto x^3 p_k'(x) - 3kx^2 p_k(x) - 2p_k(x)$ is a polynomial whose degree is at most $2(k+1)$. ▼

From the lemma we infer the infinite differentiability of f on $\mathbb{R} \setminus \{0\}$. We now need to consider the derivatives at 0. For this we employ another lemma.

2 Lemma $\lim_{x \to 0} \frac{e^{-\frac{1}{x^2}}}{x^k} = 0$ *for all* $k \in \mathbb{Z}_{\geq 0}$.

Proof We note that

$$\lim_{x \downarrow 0} \frac{e^{-\frac{1}{x^2}}}{x^k} = \lim_{y \to \infty} \frac{y^k}{e^{y^2}}, \qquad \lim_{x \uparrow 0} \frac{e^{-\frac{1}{x^2}}}{x^k} = \lim_{y \to -\infty} \frac{y^k}{e^{y^2}}.$$

Using the properties of the exponential function as given in Section 3.6.1, we have

$$e^{y^2} = \sum_{j=0}^{\infty} \frac{y^{2j}}{j!}$$

In particular, $e^{y^2} \geq \frac{y^{2k}}{k!}$, and so

$$\left| \frac{y^k}{e^{y^2}} \right| \leq \left| \frac{k!}{y^k} \right|,$$

and so

$$\lim_{x \to 0} \frac{e^{-\frac{1}{x^2}}}{x^k} = 0,$$

as desired. ▼

Now, letting $p_k(x) = \sum_{j=0}^{2k} a_j x^j$, we may directly compute

$$\lim_{x \to 0} f^{(k)}(x) = \lim_{x \to 0} \sum_{j=0}^{2k} a_j x^{2j} \frac{e^{-\frac{1}{x^2}}}{x^{3k}} = \sum_{j=0}^{2k} a_j \lim_{x \to 0} \frac{e^{-\frac{1}{x^2}}}{x^{3k-j}} = 0.$$

Thus we arrive at the conclusion that f is infinitely differentiable on $\mathbb{R}$, and that f and all of its derivatives are zero at $x = 0$. Thus $\mathscr{T}_\infty(f, 0) = (0)_{j \in \mathbb{Z}_{\geq 0}}$. This is clearly a $\mathbb{R}$-convergent power series; it converges everywhere to the zero function. However, $f(x) \neq 0$ except when $x = 0$. Thus the Taylor series about

0 for f, while convergent everywhere, converges to f only at $x = 0$. This is therefore an example of a function that is infinitely differentiable at a point, but not real analytic there. This function may seem rather useless, but in actuality it is quite an important one. For example, we used it in the construction for the proof of Theorem 3.5.25. •

These examples, along with Borel's Theorem, indicate the intricate nature of the correspondence between a function and its Taylor series. For the correspondence to have any real meaning, the function must be analytic, and even then the correspondence is only local.

3.5.5 Notes

As we shall see in ***missing stuff***, there is, for $\mathbb{C}$-power series, a correspondence between convergent power series and holomorphic functions. This correspondence also applies to the real case, where "holomorphic" gets replaced with "real analytic." The ring-theoretic structure of the $\mathbb{R}$-convergent power series are of some importance. In particular, this ring possesses the property of being "Noetherian."[12]***missing stuff*** Because of the correspondence between convergent power series and analytic functions, the ring theoretic structure gets transfered, at least locally, to the set of analytic functions. This leads to some rather remarkable features of analytic functions as compared to, say, merely infinitely differentiable functions. We refer to [Krantz and Parks 2002] for a discussion of this in the real analytic case, and to [Hörmander 1966] for the holomorphic case.

Exercises

3.5.1 State and prove a version of the Fundamental Theorem of Calculus for $\mathbb{R}$-formal power series.

3.5.2 State and prove an integration by parts formula for $\mathbb{R}$-formal power series.

3.5.3 Prove part (vi) of Proposition 2.4.30 using Proposition 3.5.17.

[12]Noether

Section 3.6

Some $\mathbb{R}$-valued functions of interest

In this section we present, in a formal way, some of the special functions that will, and indeed already have, come up in these volumes.

Do I need to read this section? It is much more than likely the case that the reader has already encountered the functions we discuss in this section. However, it may be the case that the formal definitions and rigorous presentation of their properties will be new. This section, therefore, fits into the "read for pleasure" category. •

3.6.1 The exponential function

One of the most important functions in mathematics, particularly in applied mathematics, is the exponential function. This importance is nowhere to be found in the following definition, but hopefully at the end of their reading these volumes, the reader will have some appreciation for the exponential function.

3.6.1 Definition (Exponential function) The *exponential function*, denoted by $\exp\colon \mathbb{R} \to \mathbb{R}$, is given by

$$\exp(x) = \sum_{j=0}^{\infty} \frac{x^j}{j!}.$$ •

In Figure 3.17 we show the graphs of exp and its inverse log that we will be

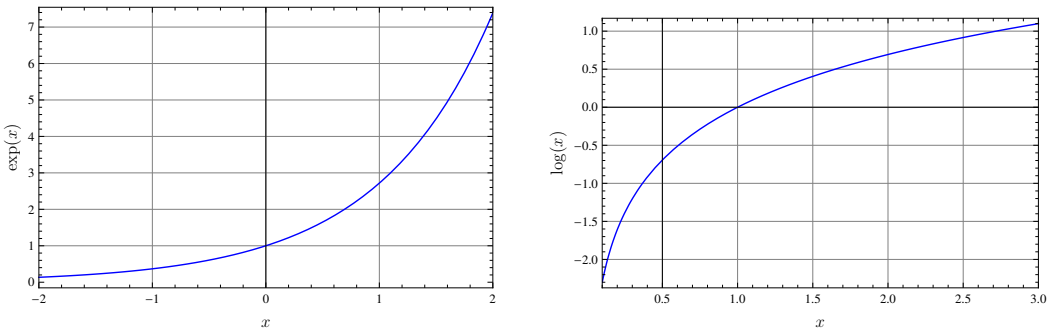


Figure 3.17 The function exp (left) and its inverse log (right)

discussing in the next section.

One can use Theorem 3.5.13, along with Proposition 2.4.15, to easily show that the power series for exp has an infinite radius of convergence, and so indeed defines a function on $\mathbb{R}$. Let us record some of the more immediate and useful properties of exp.

3.6.2 Proposition (Properties of the exponential function) *The exponential function enjoys the following properties:*

(i) exp *is infinitely differentiable;*

(ii) exp *is strictly monotonically increasing;*

(iii) $\exp(x) > 0$ *for all* $x \in \mathbb{R}$*;*

(iv) $\lim_{x\to\infty} \exp(x) = \infty$*;*

(v) $\lim_{x\to-\infty} \exp(x) = 0$*;*

(vi) $\exp(x + y) = \exp(x)\exp(y)$ *for all* $x, y \in \mathbb{R}$*;*

(vii) $\exp' = \exp$*;*

(viii) $\lim_{x\to\infty} x^k \exp(-x) = 0$ *for all* $k \in \mathbb{Z}_{>0}$.

Proof (i) This follows from Corollary 3.5.21, along with the fact that the radius of convergence of the power series for exp is infinite.

(vi) Using the Binomial Theorem and Proposition 2.4.30(iv) we compute

$$\begin{aligned}\exp(x)\exp(y) &= \Big(\sum_{j=0}^{\infty} \frac{x^j}{j!}\Big)\Big(\sum_{j=0}^{\infty} \frac{x^k}{k!}\Big) = \sum_{k=0}^{\infty}\sum_{j=0}^{k} \frac{x^j}{j!}\frac{y^{k-j}}{(k-j)!}\\ &= \sum_{k=0}^{\infty} \frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}x^j y^{k-j} = \sum_{k=0}^{\infty}\frac{(x+y)^k}{k!}.\end{aligned}$$

(viii) We have $\exp(-x) = \frac{1}{\exp(x)}$ by part (vi), and so we compute

$$\lim_{x\to\infty} x^k \exp(-x) = \lim_{x\to\infty} \frac{x^k}{\sum_{j=0}^{\infty}\frac{x^j}{j!}} \le \lim_{x\to\infty} \frac{(k+1)!x^k}{x^{k+1}} = 0.$$

(ii) From parts (i) and (viii) we know that exp has an everywhere positive derivative. Thus, from Proposition 3.2.23 we know that exp is strictly monotonically increasing.

(iii) Clearly $\exp(x) > 0$ for all $x \in \mathbb{R}_{\ge 0}$. From part (vi) we have

$$\exp(x)\exp(-x) = \exp(0) = 1.$$

Therefore, for $x \in \mathbb{R}_{<0}$ we have $\exp(x) = \frac{1}{\exp(-x)} > 0$.

(iv) We have

$$\lim_{x\to\infty} \exp(x) = \lim_{x\to\infty} \sum_{j=0}^{\infty} \frac{x^j}{j!} \ge \lim_{x\to\infty} x = \infty.$$

(v) By parts (vi) and (iv) we have

$$\lim_{x\to-\infty} \exp(x) = \lim_{x\to\infty} \frac{1}{\exp(-x)} = 0.$$

(vii) Using part (vi) and the power series representation for exp we compute

$$\exp'(x) = \lim_{h\to 0} \frac{\exp(x+h) - \exp(x)}{h} = \lim_{h\to 0} \frac{\exp(x)(\exp(h) - 1)}{h} = \exp(x). \qquad \blacksquare$$

One of the reasons for the importance of the function exp in applications can be directly seen from property (vii). From this one can see that exp is the solution to the "initial value problem"

$$y'(x) = y(x), \quad y(0) = 1. \tag{3.17}$$

Most readers will recognise this as the differential equation governing a scalar process which exhibits "exponential growth." It turns out that many physical processes can be modelled, or approximately modelled, by such an equation, or by a suitable generalisation of such an equation. Indeed, one could use the solution of (3.17) as the *definition* of the function exp. However, to be rigorous, one would then be required to show that this equation has a unique solution; this is not altogether difficult, but does take one off topic a little. Such are the constraints imposed by rigour.

In Section 2.4.3 we defined the constant e by

$$\mathrm{e} = \sum_{j=0}^{\infty} \frac{1}{j!}.$$

From this we see immediately that $\mathrm{e} = \exp(1)$. To explore the relationship between the exponential function exp and the constant e, we first prove the following result, which recalls from Proposition 2.2.3 and the discussion immediately following it, the definition of x^q for $x \in \mathbb{R}_{>0}$ and $q \in \mathbb{Q}$.

3.6.3 Proposition (exp(x) = e^x) $\exp(\mathrm{x}) = \sup\{\mathrm{e}^{\mathrm{q}} \mid \mathrm{q} \in \mathbb{Q},\ \mathrm{q} < \mathrm{x}\}$.

Proof First let us take the case where $x = q \in \mathbb{Q}$. Write $q = \frac{j}{k}$ for $j \in \mathbb{Z}$ and $k \in \mathbb{Z}_{>0}$. Then, by repeated application of part (vi) of Proposition 3.6.2 we have

$$\exp(q)^k = \exp(kq) = \exp(j) = \exp(j \cdot 1) = \exp(1)^j (\mathrm{e}^1)^j = \mathrm{e}^j.$$

By Proposition 2.2.3 this gives, by definition, $\exp(q) = \mathrm{e}^q$.

Now let $x \in \mathbb{R}$ and let $(q_j)_{j \in \mathbb{Z}_{>0}}$ be a monotonically increasing sequence in $\mathbb{Q}$ such that $\lim_{j\to\infty} q_j = x$. By Theorem 3.1.3 we have $\exp(x) = \lim_{j\to\infty} \exp(q_j)$. By part (ii) of Proposition 3.6.2 the sequence $(\exp(q_j))_{j \in \mathbb{Z}_{>0}}$ is strictly monotonically increasing. Therefore, by Theorem 2.3.8,

$$\lim_{j\to\infty} \exp(q_j) = \lim_{j\to\infty} \mathrm{e}^{q_j} = \sup\{\mathrm{e}^q \mid q < x\},$$

as desired. ∎

We shall from now on alternately use the notation e^x for $\exp(x)$, when this is more convenient.

3.6.2 The natural logarithmic function

From Proposition 3.6.2 we know that exp is a strictly monotonically increasing, continuous function. Therefore, by Theorem 3.1.30 we know that exp is an invertible function from $\mathbb{R}$ to image(exp). From parts (iii), (iv), and (v) of Proposition 3.6.2, as well as from Theorem 3.1.30 again, we know that $\text{image}(\exp) = \mathbb{R}_{>0}$. This then leads to the following definition.

3.6.4 Definition (Natural logarithmic function) The *natural logarithmic function*, denoted by $\log\colon \mathbb{R}_{>0} \to \mathbb{R}$, is the inverse of exp. •

We refer to Figure 3.17 for a depiction of the graph of log.

3.6.5 Notation (log versus ln) It is not uncommon to see the function that we denote by "log" written instead as "ln." In such cases, log is often used to refer to the base 10 logarithm (see Definition 3.6.13), since this convention actually sees much use in applications. However, we shall refer to the base 10 logarithm as $\log_{10}$. •

Now let us record the properties of log that follow immediately from its definition.

3.6.6 Proposition (Properties of the natural logarithmic function) *The natural logarithmic function enjoys the following properties:*

(i) $\log$ *is infinitely differentiable;*

(ii) $\log$ *is strictly monotonically increasing;*

(iii) $\log(x) = \int_1^x \frac{1}{\xi}\,\mathrm{d}\xi$ *for all* $x \in \mathbb{R}_{>0}$*;*

(iv) $\lim_{x\to\infty} \log(x) = \infty$*;*

(v) $\lim_{x\downarrow 0} \log(x) = -\infty$*;*

(vi) $\log(xy) = \log(x) + \log(y)$ *for all* $x, y \in \mathbb{R}_{>0}$*;*

(vii) $\lim_{x\to\infty} x^{-k} \log(x) = 0$ *for all* $k \in \mathbb{Z}_{>0}$.

Proof (iii) From the Chain Rule and using the fact that $\log \circ \exp(x) = x$ for all $x \in \mathbb{R}$ we have

$$\log'(\exp(x)) = \frac{1}{\exp(x)} \quad \Longrightarrow \quad \log'(y) = \frac{1}{y}$$

for all $y \in \mathbb{R}_{>0}$. Using the fact that $\log(1) = 0$ (which follows since $\exp(0) = 1$), we then apply the Fundamental Theorem of Calculus, this being valid since $y \mapsto \frac{1}{y}$ is Riemann integrable on any compact interval in $\mathbb{R}_{>0}$, we obtain $\log(x) = \int_1^y \frac{1}{\eta}\,\mathrm{d}\eta$, as desired.

(i) This follows from part (iii) using the fact that the function $x \mapsto \frac{1}{x}$ is infinitely differentiable on $\mathbb{R}_{>0}$.

(ii) This follows from Theorem 3.1.30.

(iv) We have

$$\lim_{x\to\infty} \log(x) = \lim_{y\to\infty} \log(\exp(y)) = \lim_{y\to\infty} y = \infty.$$

(v) We have

$$\lim_{x\downarrow 0} \log x = \lim_{y\to-\infty} \log(\exp(y)) = \lim_{y\to-\infty} y = -\infty.$$

(vi) For $x, y \in \mathbb{R}_{>0}$ write $x = \exp(a)$ and $y = \exp(b)$. Then

$$\log(xy) = \log(\exp(a)\exp(b)) = \log(\exp(a+b)) = a + b = \log(x) + \log(y).$$

(vii) We compute

$$\lim_{x\to\infty} \frac{\log x}{x^k} = \lim_{y\to\infty} \frac{\log\exp(y)}{\exp(y)^k} = \lim_{y\to\infty} \frac{y}{\exp(y)^k} \leq \lim_{y\to\infty} \frac{y}{(1 + y + \frac{1}{2}y^2)^k} = 0. \qquad \blacksquare$$

3.6.3 Power functions and general logarithmic functions

For $x \in \mathbb{R}_{>0}$ and $q \in \mathbb{Q}$ we had defined, in and immediately following Proposition 2.2.3, x^q by $(x^{1/k})^j$ if $q = \frac{j}{k}$ for $j \in \mathbb{Z}$ and $k \in \mathbb{Z}_{>0}$. In this section we wish to extend this definition to x^y for $y \in \mathbb{R}$, and to explore the properties of the resulting function of both x and y.

3.6.7 Definition (Power function) If $a \in \mathbb{R}_{>0}$ then the function $\mathsf{P}_a \colon \mathbb{R} \to \mathbb{R}$ is defined by $\mathsf{P}_a(x) = \exp(x \log(a))$. If $a \in \mathbb{R}$ then the function $\mathsf{P}^a \colon \mathbb{R}_{>0} \to \mathbb{R}$ is defined by $\mathsf{P}^a(x) = \exp(a \log(x))$. •

Let us immediately connect this (when seen for the first time rather nonintuitive) definition to what we already know.

3.6.8 Proposition ($\mathsf{P}_a(x) = a^x$) $\mathsf{P}_a(x) = \sup\{a^q \mid q \in \mathbb{Q},\ q < x\}$.

Proof Let us first take $x = q \in \mathbb{Q}$ and write $q = \frac{j}{k}$ for $j \in \mathbb{Z}$ and $k \in \mathbb{Z}_{>0}$. We have

$$\exp(q \log(a))^k = \exp\Big(\tfrac{j}{k} \log(a)\Big)^k = \exp(j \log(a)) = \exp(\log(a))^j = a^j.$$

Therefore, by Proposition 2.2.3 we have

$$\exp(q \log(a)) = a^q.$$

Now let $x \in \mathbb{R}$ and let $(q_j)_{j \in \mathbb{Z}_{>0}}$ be a strictly monotonically increasing sequence in $\mathbb{Q}$ converging to x. Since exp and log are continuous, by Theorem 3.1.3 we have

$$\lim_{j \to \infty} \exp(q_j \log(a)) = \exp(x \log(a)).$$

As we shall see in Proposition 3.6.10, the function $x \mapsto \mathsf{P}_a(x)$ is strictly monotonically increasing. Therefore the sequence $(\exp(q_j \log(a)))_{j \in \mathbb{Z}_{>0}}$ is strictly monotonically increasing. Thus

$$\lim_{j \to \infty} \exp(q_j \log(a)) = \sup\{\mathsf{P}_a(q) \mid q \in \mathbb{Q},\ q < x\},$$

as desired. ■

Clearly we also have the following result.

3.6.9 Corollary ($\mathsf{P}^a(x) = x^a$) $\mathsf{P}^a(x) = \sup\{x^q \mid q \in \mathbb{Q},\ q < a\}$.

As with the exponential function, we will use the notation a^x for $\mathsf{P}_a(x)$ and x^a for $\mathsf{P}^a(x)$ when it is convenient to do so.

Let us now record some of the properties of the functions P_a and P^a that follow from their definition. When possible, we state the result using both the notation $\mathsf{P}_a(x)$ and a^x (or P^a and x^a).

3.6.10 Proposition (Properties of P_a) *For $a \in \mathbb{R}_{>0}$, the function P_a enjoys the following properties:*

(i) P_a *is infinitely differentiable;*

(ii) P_a *is strictly monotonically increasing when* $a > 1$*, is strictly monotonically decreasing when* $a < 1$*, and is constant when* $a = 1$*;*

(iii) $\mathsf{P}_a(x) = a^x > 0$ *for all* $x \in \mathbb{R}$*;*

(iv) $\lim_{x\to\infty} \mathsf{P}_a(x) = \lim_{x\to\infty} a^x = \begin{cases} \infty, & a > 1, \\ 0, & a < 1, \\ 1, & a = 1; \end{cases}$

(v) $\lim_{x\to-\infty} \mathsf{P}_a(x) == \lim_{x\to-\infty} a^x = \begin{cases} 0, & a > 1, \\ \infty, & a < 1, \\ 1, & a = 1; \end{cases}$

(vi) $\mathsf{P}_a(x + y) = a^{x+y} = a^x a^y = \mathsf{P}_a(x)\mathsf{P}_a(y)$*;*

(vii) $\mathsf{P}'_a(x) = \log(a)\mathsf{P}_a(x)$*;*

(viii) if $a > 1$ *then* $\lim_{x\to\infty} x^k \mathsf{P}_a(-x) = \lim_{x\to\infty} x^k a^{-x} = 0$ *for all* $k \in \mathbb{Z}_{>0}$*;*

(ix) if $a < 1$ *then* $\lim_{x\to\infty} x^k \mathsf{P}_a(x) = \lim_{x\to\infty} x^k a^x = 0$ *for all* $k \in \mathbb{Z}_{>0}$.

Proof (i) Define $f, g\colon \mathbb{R} \to \mathbb{R}$ and $f(x) = x\log(a)$ and $g(x) = \exp(x)$. Then $\mathsf{P}_a = g \circ f$, and so is the composition of infinitely differentiable functions. This part of the result follows from Theorem 3.2.13.

(ii) Let $x_1 < x_2$. If $a > 1$ then $\log(a) > 0$ and so

$$x_1 \log(a) < x_2 \log(a) \quad \Longrightarrow \quad \exp(x_1 \log(a)) < \exp(x_2 \log(a))$$

since exp is strictly monotonically increasing. If $a < 1$ then $\log(a) < 0$ and so

$$x_1 \log(a) > x_2 \log(a) \quad \Longrightarrow \quad \exp(x_1 \log(a)) > \exp(x_2 \log(a)),$$

again since exp is strictly monotonically increasing. For $a = 1$ we have $\log(a) = 0$ so $\mathsf{P}_a(x) = 1$ for all $x \in \mathbb{R}$.

(iii) This follows since image(exp) $\subseteq \mathbb{R}_{>0}$.

(iv) For $a > 1$ we have

$$\lim_{x\to\infty} \mathsf{P}_a(x) = \lim_{x\to\infty} \exp(x\log(a)) = \lim_{y\to\infty} \exp(y) = \infty,$$

and for $a < 1$ we have

$$\lim_{x\to\infty} \mathsf{P}_a(x) = \lim_{x\to\infty} \exp(x\log(a)) = \lim_{y\to-\infty} \exp(y) = 0.$$

For $a = 1$ the result is clear since $\mathsf{P}_1(x) = 1$ for all $x \in \mathbb{R}$.

(v) For $a > 1$ we have

$$\lim_{x\to-\infty} \mathsf{P}_a(x) = \lim_{x\to-\infty} \exp(x\log(a)) = \lim_{y\to-\infty} \exp(y) = 0,$$

and for $a < 1$ we have

$$\lim_{x\to-\infty} \mathsf{P}_a(x) = \lim_{x\to-\infty} \exp(x\log(a)) = \lim_{y\to\infty} \exp(y) = \infty.$$

Again, for $a = 1$ the result is obvious.

(vi) We have

$$\mathsf{P}_a(x + y) = \exp((x + y)\log(a)) = \exp(x\log(a))\exp(y\log(a)) = \mathsf{P}_a(x)\mathsf{P}_a(y).$$

(vii) With f and g as in part (i), and using Theorem 3.2.13, we compute

$$\mathsf{P}_a'(x) = g'(f(x))f'(x) = \exp(x\log(a))\log(a) = \log(a)\mathsf{P}_a(x).$$

(viii) We compute

$$\lim_{x\to\infty} x^k \mathsf{P}_a(-x) = \lim_{x\to\infty} x^k \exp(-x\log(a)) = \lim_{y\to\infty} \Big(\frac{y}{\log(a)}\Big)^k \exp(-y) = 0,$$

using part (viii) of Proposition 3.6.2.

(ix) We have

$$\lim_{x\to\infty} x^k \mathsf{P}_a(x) = \lim_{x\to\infty} x^k \exp((-x)(-\log(a))) = 0$$

since $\log(a) < 0$. ∎

3.6.11 Proposition (Properties of P^{a}) *For* $\mathrm{a} \in \mathbb{R}$*, the function* P^{a} *enjoys the following properties:*

(i) P^{a} *is infinitely differentiable;*

(ii) P^{a} *is strictly monotonically increasing;*

(iii) $\mathsf{P}^{\mathrm{a}}(\mathrm{x}) = \mathrm{x}^{\mathrm{a}} > 0$ *for all* $\mathrm{x} \in \mathbb{R}_{>0}$*;*

(iv) $\lim_{\mathrm{x}\to\infty} \mathsf{P}^{\mathrm{a}}(\mathrm{x}) = \lim_{\mathrm{x}\to\infty} \mathrm{x}^{\mathrm{a}} = \begin{cases} \infty, & \mathrm{a} > 0, \\ 0, & \mathrm{a} < 0, \\ 1, & \mathrm{a} = 0; \end{cases}$

(v) $\lim_{\mathrm{x}\downarrow 0} \mathsf{P}^{\mathrm{a}}(\mathrm{x}) = \lim_{\mathrm{x}\downarrow 0} \mathrm{x}^{\mathrm{a}} = \begin{cases} 0, & \mathrm{a} > 0, \\ \infty, & \mathrm{a} < 0, \\ 1, & \mathrm{a} = 0; \end{cases}$

(vi) $\mathsf{P}^{\mathrm{a}}(\mathrm{xy}) = (\mathrm{xy})^{\mathrm{a}} = \mathrm{x}^{\mathrm{a}}\mathrm{y}^{\mathrm{a}} = \mathsf{P}^{\mathrm{a}}(\mathrm{x})\mathsf{P}^{\mathrm{a}}(\mathrm{y})$*;*

(vii) $(\mathsf{P}^{\mathrm{a}})'(\mathrm{x}) = \mathrm{a}\mathsf{P}^{\mathrm{a}-1}(\mathrm{x})$.

Proof (i) Define $f\colon \mathbb{R}_{>0} \to \mathbb{R}$, $g\colon \mathbb{R} \to \mathbb{R}$, and $h\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = \log(x)$, $g(x) = ax$, and $h(x) = \exp(x)$. Then $\mathsf{P}^a = h \circ g \circ f$. Since each of f, g, and h is infinitely differentiable, then so too is P^a by Theorem 3.2.13.

(ii) Let $x_1, x_2 \in \mathbb{R}_{>0}$ satisfy $x_1 < x_2$. Then

$$\mathsf{P}^a(x_1) = \exp(a\log(x_1)) < \exp(a\log(x_2)) = \mathsf{P}^a(x_2)$$

using the fact that both log and exp are strictly monotonically increasing.

(iii) This follows since image(exp) $\subseteq \mathbb{R}_{>0}$.

(iv) For $a > 0$ we have

$$\lim_{x\to\infty} \mathsf{P}^a(x) = \lim_{x\to\infty} \exp(a\log(x)) = \lim_{y\to\infty} \exp(y) = \infty,$$

and for $a < 0$ we have

$$\lim_{x\to\infty} \mathsf{P}^a(x) = \lim_{x\to\infty} \exp(a\log(x)) = \lim_{y\to-\infty} \exp(y) = 0.$$

For $a = 0$ we have $\mathsf{P}^a(x) = 1$ for all $x \in \mathbb{R}_{>0}$.

(v) For $a > 0$ we have

$$\lim_{x\downarrow 0} \mathsf{P}^a(x) = \lim_{x\downarrow 0} \exp(a\log(x)) = \lim_{y\to-\infty} \exp(y) = 0,$$

and for $a < 0$ we have

$$\lim_{x\downarrow 0} \mathsf{P}^a(x) = \lim_{x\downarrow 0} \exp(a\log(x)) = \lim_{y\to\infty} \exp(y) = \infty.$$

For $a = 1$, the result is trivial again.

(vi) We have

$$\mathsf{P}^a(xy) = \exp(a\log(xy)) = \exp(a(\log(x)+\log(y))) = \exp(a\log(x))\exp(a\log(y)) = \mathsf{P}^a(x)\mathsf{P}^a(y).$$

(vii) With f, g, and h as in part (i), and using the Chain Rule, we have

$$\begin{aligned}(\mathsf{P}^a)'(x) = h'(g(f(x)))g'(f(x))f'(x) &= a\exp(a\log(x))\tfrac{1}{x}\\ &= a\exp(a\log(x))\exp(-1\log(x)) = a\exp((a-1)\log(x)) = a\mathsf{P}^{a-1}(x),\end{aligned}$$

as desired, using part (vi) of Proposition 3.6.10. ■

The following result is also sometimes useful.

3.6.12 Proposition (Property of $\mathsf{P}_x(x^{-1})$) $\lim_{x\to\infty} \mathsf{P}_x(x^{-1}) = \lim_{x\to\infty} x^{1/x} = 1$.

Proof We have

$$\lim_{x\to\infty} \mathsf{P}_x(x^{-1}) = \lim_{x\to\infty} \exp(x^{-1}\log(x)) = \lim_{y\to 0} \exp(y) = 1,$$

using part (vii) of Proposition 3.6.6. ■

Now we turn to the process of inverting the power function. For the exponential function we required that $\log(e^x) = x$. Thus, if our inverse of P_a is denoted (for the moment) by f_a, then we expect that $f_a(a^x) = x$. This definition clearly has difficulties when $a = 1$, reflecting the fact that P_1 is not invertible. In all other case, since P_a is continuous, and either strictly monotonically increasing or strictly monotonically decreasing, we have the following definition, using Theorem 3.1.30.

3.6.13 Definition (Arbitrary base logarithm) For $a \in \mathbb{R}_{>0}\setminus\{1\}$, the function $\log_a : \mathbb{R}_{>0} \to \mathbb{R}$, called the ***base*** **a** ***logarithmic function***, is the inverse of P_a. When $a = 10$ we simply write $\log_{10} = \log$. •

The following result relates the logarithmic function for an arbitrary base to the natural logarithmic function.

3.6.14 Proposition (Characterisation of $\log_a$) $\log_a(x) = \dfrac{\log(x)}{\log(a)}$.

Proof Let $x \in \mathbb{R}_{>0}$ and write $x = a^y$ for some $y \in \mathbb{R}$. First suppose that $y \neq 0$. Then we have $\log(x) = y\log(a)$ and $\log_a(x) = y$, and the result follows by eliminating y from these two expressions. When $y = 0$ we have $x = a = a^1$. Therefore, $\log_a(x) = 1 = \frac{\log(x)}{\log(a)}$. ■

With this result we immediately have the following generalisation of Proposition 3.6.6. We leave the trivial checking of the details to the reader.

3.6.15 Proposition (Properties of $\log_a$) *For* $a \in \mathbb{R}_{>0} \setminus \{1\}$, *the function* $\log_a$ *enjoys the following properties:*

(i) $\log_a$ *is infinitely differentiable;*

(ii) $\log_a$ *is strictly monotonically increasing when* $a > 1$ *and is strictly monotonically decreasing when* $a < 1$;

(iii) $\log_a(x) = \frac{1}{\log(a)} \int_1^x \frac{1}{\xi}\, d\xi$ *for all* $x \in \mathbb{R}_{>0}$;

(iv) $\lim_{x\to\infty} \log_a(x) = \begin{cases} \infty, & a > 1, \\ -\infty, & a < 1; \end{cases}$

(v) $\lim_{x\downarrow 0} \log_a(x) = \begin{cases} -\infty, & a > 1, \\ \infty, & a < 1; \end{cases}$

(vi) $\log_a(xy) = \log_a(x) + \log_a(y)$ *for all* $x, y \in \mathbb{R}_{>0}$;

(vii) $\lim_{x\to\infty} x^{-k} \log_a(x) = 0$ *for all* $k \in \mathbb{Z}_{>0}$.

3.6.4 Trigonometric functions

Next we turn to describing the standard trigonometric functions. These functions are perhaps most intuitively introduced in terms of the concept of "angle" in plane geometry. However, to really do this properly would, at this juncture, require a significant expenditure of effort. Therefore, we define the trigonometric functions by their power series expansion, and then proceed to show that they have the expected properties. In the course of our treatment we will also see that the constant π introduced in Section 2.4.3 has the anticipated relationships to the trigonometric functions. Convenience in this section forces us to make a fairly serious logical jump in the presentation. While all constructions and theorems are stated in terms of real numbers, in the proofs we use complex numbers rather heavily.

3.6.16 Definition (sin and cos) The ***sine function***, denoted by $\sin\colon \mathbb{R} \to \mathbb{R}$, and the ***cosine function***, denoted by $\cos\colon \mathbb{R} \to \mathbb{R}$, are defined by

$$\sin(x) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1} x^{2j-1}}{(2j-1)!}, \quad \cos(x) = \sum_{j=0}^{\infty} \frac{(-1)^j x^{2j}}{(2j)!},$$

respectively. •

In Figure 3.18 we show the graphs of the functions sin and cos.

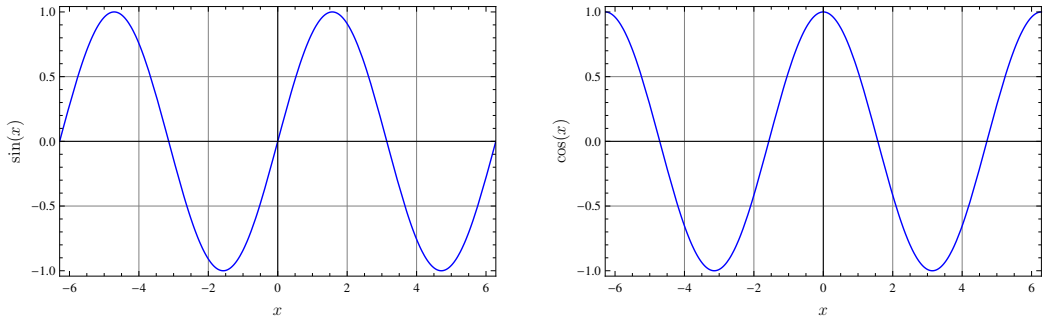


Figure 3.18 The functions sin (left) and cos (right)

3.6.17 Notation Following normal conventions, we shall frequently write $\sin x$ and $\cos x$ rather than the more correct $\sin(x)$ and $\cos(x)$. •

An application of Proposition 2.4.15 and Theorem 3.5.13 shows that the power series expansions for sin and cos are, in fact, convergent for all x, and so the functions are indeed defined with domain $\mathbb{R}$.

First we prove the existence of a number having the property that we know π to possess. In fact, we construct the number $\frac{\pi}{2}$, where π is as given in Section 2.4.3.

3.6.18 Theorem (Construction of π) *There exists a positive real number* p_0 *such that*

$$\mathrm{p}_0 = \inf\{\mathrm{x} \in \mathbb{R}_{>0} \mid \cos(\mathrm{x}) = 0\}.$$

Moreover, $\mathrm{p}_0 = \frac{\pi}{2}$.

Proof First we record the derivative properties for sin and cos.

1 Lemma *The functions* sin *and* cos *are infinitely differentiable and satisfy* $\sin' = \cos$ *and* $\cos' = -\sin$.

Proof This follows directly from Proposition 3.5.20 where it is shown that convergent power series can be differentiated term-by-term. ▼

Let us now perform some computations using complex variables that will be essential to many of the proofs in this section. We suppose the reader to be acquainted with the necessary elementary facts about complex numbers. The next observation is the most essential along these lines. We denote $\mathbb{S}^{\mathbb{C}}_1 = \{z \in \mathbb{C} \mid |z| = 1\}$, and recall that all points in $z \in \mathbb{S}^1_{\mathbb{C}}$ can be written as $z = \mathrm{e}^{\mathrm{i}x}$ for some $x \in \mathbb{R}$, and that, conversely, for any $x \in \mathbb{R}$ we have $\mathrm{e}^{\mathrm{i}x} \in \mathbb{S}^1_{\mathbb{C}}$.

2 Lemma $\mathrm{e}^{\mathrm{i}x} = \cos(\mathrm{x}) + \mathrm{i}\sin(\mathrm{x})$.

Proof This follows immediately from the $\mathbb{C}$-power series for the complex exponential function:

$$\mathrm{e}^z = \sum_{j=0}^{\infty} \frac{x^j}{j!}.$$

Substituting $z = \mathrm{i}x$, using the fact that $\mathrm{i}^{2j} = (-1)^j$ for all $j \in \mathbb{Z}_{>0}$, and using Proposition 2.4.30, we get the desired result. ▼

From the preceding lemma we then know that $\cos(x) = \operatorname{Re}(e^{ix})$ and that $\sin(x) = \operatorname{Im}(e^{ix})$. Therefore, since $e^{ix} \in \mathbb{S}^1_{\mathbb{C}}$, we have

$$\cos(x)^2 + \sin(x)^2 = 1. \tag{3.18}$$

Let us show that the set $\{x \in \mathbb{R}_{>0} \mid \cos(x) = 0\}$ is nonempty. Suppose that it is empty. Since $\cos(0) = 1$ and since cos is continuous, it must therefore be the case (by the Intermediate Value Theorem) that $\cos(x) > 0$ for all $x \in \mathbb{R}$. Therefore, by Lemma 1, $\sin'(x) > 0$ for all $x \in \mathbb{R}$, and so sin is strictly monotonically increasing by Proposition 3.2.23. Therefore, since $\sin(0) = 0$, $\sin(x) > 0$ for $x > 0$. Therefore, for $x_1, x_2 \in \mathbb{R}_{>0}$ satisfying $x_1 < x_2$, we have

$$\sin(x_1)(x_2 - x_1) < \int_{x_1}^{x_2} \sin(x)\,\mathrm{d}x = \cos(x_2) - \cos(x_1) \leq 2,$$

where we have used the fact that sin is strictly monotonically increasing, Lemma 1, the Fundamental Theorem of Calculus, and (3.18). We thus have arrive at the contradiction that $\limsup_{x_2 \to \infty} \sin(x_1)(x_2 - x_1) \leq 2$.

Since cos is continuous, the set $\{x \in \mathbb{R}_{>0} \mid \cos(x) = 0\}$ is closed. Therefore, $\inf\{x \in \mathbb{R}_{>0} \mid \cos(x) = 0\}$ is contained in this set, and this gives the existence of p_0. Note that, by (3.18), $\sin(p_0) \in \{-1, 1\}$. Since $\sin(0) = 0$ and since $\sin(x) = \cos(x) > 0$ for $x \in [0, p_0)$, we must have $\sin(p_0) = 1$.

The following property of p_0 will also be important.

3 Lemma $\cos(\frac{p_0}{2}) = \sin(\frac{p_0}{2}) = \frac{1}{\sqrt{2}}$.

Proof Let $x_0 = \cos(\frac{p_0}{2})$, $y_0 = \sin(\frac{p_0}{2})$, and $z_0 = x_0 + \mathrm{i}y_0$. Then, using Proposition **??**,

$$(e^{\mathrm{i}\frac{p_0}{2}})^2 = e^{\mathrm{i}p_0} = \mathrm{i}$$

since $\cos(p_0) = 0$ and $\sin(p_0) = 1$. Thus

$$(e^{\mathrm{i}\frac{p_0}{2}})^4 = \mathrm{i}^2 = -1,$$

again using Proposition **??**. Using the definition of complex multiplication we also have

$$(e^{\mathrm{i}\frac{p_0}{2}})^4 = (x_0 + \mathrm{i}y_0)^4 = x_0^4 - 6x_0^2y_0^2 + y_0^4 + 4\mathrm{i}x_0y_0(x_0^2 - y_0^2).$$

Thus, in particular, $x_0^2 - y_0^2 = 0$. Combining this with $x_0^2 + y_0^2 = 1$ we get $x_0^2 = y_0^2 = \frac{1}{2}$. Since both x_0 and y_0 are positive by virtue of $\frac{p_0}{2}$ lying in $(0, p_0)$, we must have $x_0 = y_0 = \frac{1}{\sqrt{2}}$, as claimed. ▼

Now we show, through a sequence of seemingly irrelevant computations, that $p_0 = \frac{\pi}{2}$. Define the function $\tan\colon (-p_0, p_0) \to \mathbb{R}$ by $\tan(x) = \frac{\sin(x)}{\cos(x)}$, noting that tan is well-defined since $\cos(-x) = \cos(x)$ and since $\cos(x) > 0$ for $x \in [0, p_0)$. We claim that tan is continuous and strictly monotonically increasing. We have, using the quotient rule,

$$\tan'(x) = \frac{\cos(x)^2 + \sin(x)^2}{\cos(x)^2} = \frac{1}{\cos(x)^2}.$$

Thus $\tan'(x) > 0$ for all $x \in (-p_0, p_0)$, and so tan is strictly monotonically increasing by Proposition 3.2.23. Since $\sin(p_0) = 1$ and (since $\sin(-x) = -\sin(x)$) since $\sin(-p_0) = -1$, we have

$$\lim_{x \uparrow p_0} \tan(x) = \infty, \quad \lim_{x \downarrow p_0} \tan(x) = -\infty.$$

This shows that tan is an invertible and differentiable mapping from $(-p_0, p_0)$ to $\mathbb{R}$. Moreover, since tan' is nowhere zero, the inverse, denoted by $\tan^{-1} : \mathbb{R} \to (-p_0, p_0)$, is also differentiable and the derivative of its inverse is given by

$$(\tan^{-1})'(x) = \frac{1}{\tan'(\tan^{-1}(x))},$$

as per Theorem 3.2.24. We further claim that

$$(\tan^{-1})'(x) = \frac{1}{1 + x^2}.$$

Indeed, our above arguments show that $(\tan^{-1})'(x) = (\cos(\tan^{-1}(x)))^2$. If $y = \tan^{-1}(x)$ then

$$\frac{\sin(y)}{\cos(y)} = x.$$

Since $\sin(y) > 0$ for $y \in (0, p_0)$, we have $\sin(y) = \sqrt{1 - \cos(y)}$ by (3.18). Therefore,

$$\frac{1 - \cos(y)^2}{\cos(y)^2} = x^2 \quad \Longrightarrow \quad \cos(y)^2 = \frac{1}{1 + x^2}$$

as desired.

By the Fundamental Theorem of Calculus we then have

$$\int_0^1 \frac{1}{1 + x^2} \, \mathrm{d}x = \tan^{-1}(1) - \tan^{-1}(0).$$

Since $\tan^{-1}(1) = \frac{p_0}{2}$ by Lemma 3 above and since $\tan^{-1}(0) = 0$ (and using part (v) of Proposition 3.6.19 below), we have

$$\int_0^1 \frac{1}{1 + x^2} \, \mathrm{d}x = \frac{p_0}{2}. \tag{3.19}$$

Now recall from Example 3.5.28–1 that we have

$$\frac{1}{1 + x^2} = \sum_{j=0}^{\infty} (-1)^j x^{2j},$$

with the series converging uniformly on any compact subinterval of $(-1, 1)$. Therefore, by Proposition 3.5.20, for $\epsilon \in (0, 1)$ we have

$$\begin{aligned} \int_0^{1-\epsilon} \frac{1}{1 + x^2} \, \mathrm{d}x &= \int_0^{1-\epsilon} \sum_{j=0}^{\infty} (-1)^j x^{2j} \, \mathrm{d}x \\ &= \sum_{j=0}^{\infty} (-1)^j \int_0^{1-\epsilon} x^{2j} \, \mathrm{d}x \\ &= \sum_{j=0}^{\infty} (-1)^j \frac{(1 - \epsilon)^{2j+1}}{2j + 1}. \end{aligned}$$

The following technical lemma will allow us to conclude the proof.

4 Lemma $\lim_{\epsilon\downarrow 0}\sum_{j=0}^{\infty}(-1)^j\frac{(1-\epsilon)^{2j+1}}{2j+1}=\sum_{j=0}\frac{(-1)^j}{2j+1}$.

Proof By the Alternating Test, the series $\sum_{j=0}^{\infty}(-1)^j\frac{(1-\epsilon)^{2j+1}}{2j+1}$ converges for $\epsilon\in[0,2]$. Define $f\colon[0,2]\to\mathbb{R}$ by

$$f(x)=\sum_{j=0}^{\infty}(-1)^{j+1}\frac{(x-1)^{2j+1}}{2j+1}$$

and define $g\colon[-1,1]\to\mathbb{R}$ by

$$g(x)=\sum_{j=0}^{\infty}(-1)^{j+1}\frac{x^{2j+1}}{2j+1}$$

so that $f(x)=g(x-1)$. Since g is defined by a $\mathbb{R}$-convergent power series, by Corollary 3.5.18 g is continuous. In particular,

$$g(-1)=\lim_{x\downarrow -1}\sum_{j=0}^{\infty}(-1)^{j+1}\frac{x^{2j+1}}{2j+1}.$$

From this it follows that

$$f(0)=\lim_{x\downarrow 0}\sum_{j=0}^{\infty}(-1)^{j+1}\frac{(x-1)^{2j+1}}{2j+1},$$

which is the result. ▼

Combining this with (3.19) we have

$$\frac{p_0}{2}=\lim_{\epsilon\downarrow 0}\int_0^{1-\epsilon}\frac{1}{1+x^2}\,\mathrm{d}x=\lim_{\epsilon\downarrow 0}\sum_{j=0}^{\infty}(-1)^j\frac{(1-\epsilon)^{2j+1}}{2j+1}=\sum_{j=0}\frac{(-1)^j}{2j+1}=\frac{\pi}{4},$$

using the definition of π in Definition 2.4.20. ■

Now that we have on hand a reasonable characterisation of π, we can proceed to state the familiar properties of sin and cos.

3.6.19 Proposition (Properties of sin and cos) *The functions* sin *and* cos *enjoy the following properties:*

(i) sin *and* cos *are infinitely differentiable, and furthermore satisfy* $\sin' = \cos$ *and* $\cos' = -\sin$*;*

(ii) $\sin(-x)=\sin(x)$ *and* $\cos(-x)=\cos(x)$ *for all* $x\in\mathbb{R}$*;*

(iii) $\sin(x)^2+\cos(x)^2=1$ *for all* $x\in\mathbb{R}$*;*

(iv) $\sin(x+2\pi)=\sin(x)$ *and* $\cos(x+2\pi)=\cos(x)$ *for all* $x\in\mathbb{R}$*;*

(v) the map

$$[0,2\pi)\ni x\mapsto(\cos(x),\sin(x))\in\{(x,y)\in\mathbb{R}^2\mid x^2+y^2=1\}$$

is a bijection.

Proof (i) This was proved as Lemma 1 in the proof of Theorem 3.6.18.

(ii) This follows immediately from the $\mathbb{R}$-power series for sin and cos.

(iii) This was proved as (3.18) in the course of the proof of Theorem 3.6.18.

(iv) Since $e^{i\frac{\pi}{2}} = i$ by Theorem 3.6.18, we use Proposition **??** to deduce

$$e^{2\pi i} = (e^{i\frac{\pi}{2}})^4 = i^4 = 1.$$

Again using Proposition **??** we then have

$$e^{z+2\pi i} = e^z e^{2\pi i} = e^z$$

for all $z \in \mathbb{C}$. Therefore, for $x \in \mathbb{R}$, we have

$$\cos(x + 2\pi) + i\sin(x + 2\pi) = e^{i(x+2\pi)} = e^{ix} = \cos(x) + i\sin(x),$$

which gives the result.

(v) Denote $\mathbb{S}^1 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$, and note that, if we make the standard identification of $\mathbb{C}$ with $\mathbb{R}^2$ (as we do), then $\mathbb{S}^1_{\mathbb{C}}$ (see the proof of Theorem 3.6.18) becomes identified with $\mathbb{S}^1$, with the identification explicitly being $x + iy \mapsto (x, y)$. Thus the result we are proving is equivalent to the assertion that the map

$$f : [0, 2\pi) \ni x \mapsto e^{ix} \in \mathbb{S}^1_{\mathbb{C}}$$

is a bijection. This is what we will prove. By part (iii), this map is well-defined in the sense that it actually does take values in $\mathbb{S}^1_{\mathbb{C}}$. Suppose that $e^{ix_1} = e^{ix_2}$ for distinct points $x_1, x_2 \in [0, 2\pi)$, and suppose for concreteness that $x_1 < x_2$. Then $x_2 - x_1 \in (0, 2\pi)$, and $\frac{1}{4}(x_2 - x_1) \in (0, \frac{\pi}{2})$. We then have

$$e^{ix_1} = e^{ix_2} \implies e^{i(x_2-x_1)} = 1 \implies (e^{i\frac{1}{4}(x_2-x_1)})^4 = 1.$$

Let $e^{i\frac{1}{4}(x_2-x_1)} = \xi + i\eta$. Since $\frac{1}{4}(x_2 - x_1) \in (0, \frac{\pi}{2})$, we saw during the course of the proof of Theorem 3.6.18 that $\xi, \eta \in (0, 1)$. We then use the definition of complex multiplication to compute

$$(e^{i\frac{1}{4}(x_2-x_1)})^4 = \xi^4 - 6\xi^2\eta^2 + \eta^4 + 4i\xi\eta(\xi^2 - \eta^2).$$

Since $(e^{i\frac{1}{4}(x_2-x_1)})^4 = 1$ is real, we conclude that $\xi^2 - \eta^2 = 0$. Combining this with $\xi^2 + \eta^2 = 1$ gives $\xi^2 = \eta^2 = \frac{1}{2}$. Since both ξ and η are positive we have $\xi = \eta = \frac{1}{\sqrt{2}}$. Substituting this into the above expression for $(e^{i\frac{1}{4}(x_2-x_1)})^4$ gives $(e^{i\frac{1}{4}(x_2-x_1)})^4 = -1$. Thus we arrive at a contradiction, and it cannot be the case that $e^{ix_1} = e^{ix_2}$ for distinct $x_1, x_2 \in [0, 2\pi)$. Thus f is injective.

To show that f is surjective, we let $z = x + iy \in \mathbb{S}^1_{\mathbb{C}}$, and consider four cases.

1. $x, y \geq 0$: Since cos is monotonically decreasing from 1 to 0 on $[0, \frac{\pi}{2}]$, there exists $\theta \in [0, \frac{\pi}{2}]$ such that $\cos(\theta) = x$. Since $\sin(\theta)^2 = 1 - \cos(\theta)^2 = 1 - x^2 = y^2$, and since $\sin(\theta) \geq 0$ for $\theta \in [0, \frac{\pi}{2}]$, we conclude that $\sin(\theta) = y$. Thus $z = e^{i\theta}$.
2. $x \geq 0$ and $y \leq 0$: Let $\xi = x$ and $\eta = -y$ so that $\xi, \eta \geq 0$. From the preceding case we deduce the existence of $\phi \in [0, \frac{\pi}{2}]$ such that $e^{i\phi} = \xi + i\eta$. Thus $\cos(\phi) = x$ and $\sin(\phi) = -y$. By part (ii) we then have $\cos(-\phi) = x$ and $\sin(-\phi) = y$, and we note that $-\phi \in [-\frac{\pi}{2}, 0]$. Define

$$\theta = \begin{cases} 2\pi - \phi, & \phi \in (0, \frac{\pi}{2}], \\ 0, & \phi = 0. \end{cases}$$

By part (iv) we then have $\cos(\theta) = x$ and $\sin(\theta) = y$, and that $\theta \in [\frac{3\pi}{2}, 2\pi)$ if $\phi \in (0, \frac{\pi}{2}]$.

3. $x \leq 0$ and $y \geq 0$: Let $\xi = -x$ and $\eta = y$ si that $\xi, \eta \geq 0$. As in the first case we have $\phi \in [0, \frac{\pi}{2}]$ such that $\cos(\phi) = \xi$ and $\sin(\phi) = \eta$. We then have $-\cos(\phi) = x$ and $\sin(\phi) = y$. Next define $\theta = \pi - \phi$ and note that

$$e^{i\theta} = e^{i\pi}e^{-i\phi} = -(\cos(\phi) - i\sin(\phi)) = -\cos(\phi) + i\sin(\phi) = x + iy,$$

as desired.

4. $x \leq 0$ and $y \leq 0$: Take $\xi = -x$ and $\eta = -y$ so that $\xi, \eta \geq 0$. As in the first case, we have $\phi \in [0, \frac{\pi}{2}]$ such that $\cos(\phi) = \xi = -x$ and $\sin(\phi) = \eta = -y$. Then, taking $\theta = \pi + \phi$, we have

$$e^{i\theta} = e^{i\pi}e^{i\phi} = -(\cos(\phi) + i\sin(\phi)) = x + iy,$$

as desired. ■

From the basic construction of sin and cos that we give, and the properties that follow directly from this construction, there is of course a great deal that one can proceed to do; the resulting subject is broadly called "trigonometry." Rigorous proofs of many of the facts of basic trigonometry follow easily from our constructions here, particularly since we give the necessary properties, along with a rigorous definition, of π. We do assume that the reader has an acquaintance with trigonometry, as we shall use certain of these facts without much ado.

The reciprocals of sin and cos are sometimes used. Thus we define $\csc\colon (0, 2\pi) \to \mathbb{R}$ and $\sec\colon (-\pi, \pi) \to \mathbb{R}$ by $\csc(x) = \frac{1}{\sin(x)}$ and $\sec(x) = \frac{1}{\cos(x)}$. These are the ***cosecant*** and ***secant*** functions, respectively. One can verify that the restrictions of csc and sec to $(0, \frac{\pi}{2})$ are bijective. In Figure 3.19

One useful and not perfectly standard construction is the following. Define $\tan\colon (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$ by $\tan(x) = \frac{\sin(x)}{\cos(x)}$, noting that the definition makes sense since $\cos(x) > 0$ for $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$. In Figure 3.20 we depict the graph of tan and its inverse $\tan^{-1}$. During the course of the proof of Theorem 3.6.18 we showed that the function tan had the following properties.

3.6.20 Proposition (Properties of tan) *The function* tan *enjoys the following properties:*

(i) tan *is infinitely differentiable;*

(ii) tan *is strictly monotonically increasing;*

(iii) the inverse of tan*, denoted by* $\tan^{-1}\colon \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2})$ *is infinitely differentiable.*

It turns out to be useful to extend the definition of $\tan^{-1}$ to $(-\pi, \pi]$ by defining the function $\operatorname{atan}\colon \mathbb{R}^2 \setminus \{(0,0)\} \to (-\pi, \pi]$ by

$$\operatorname{atan}(x, y) = \begin{cases} \tan^{-1}(\frac{y}{x}), & x > 0, \\ \pi - \tan^{-1}(\frac{y}{x}), & x < 0, \\ \frac{\pi}{2}, & x = 0,\ y > 0, \\ -\frac{\pi}{2}, & x = 0,\ y < 0. \end{cases}$$

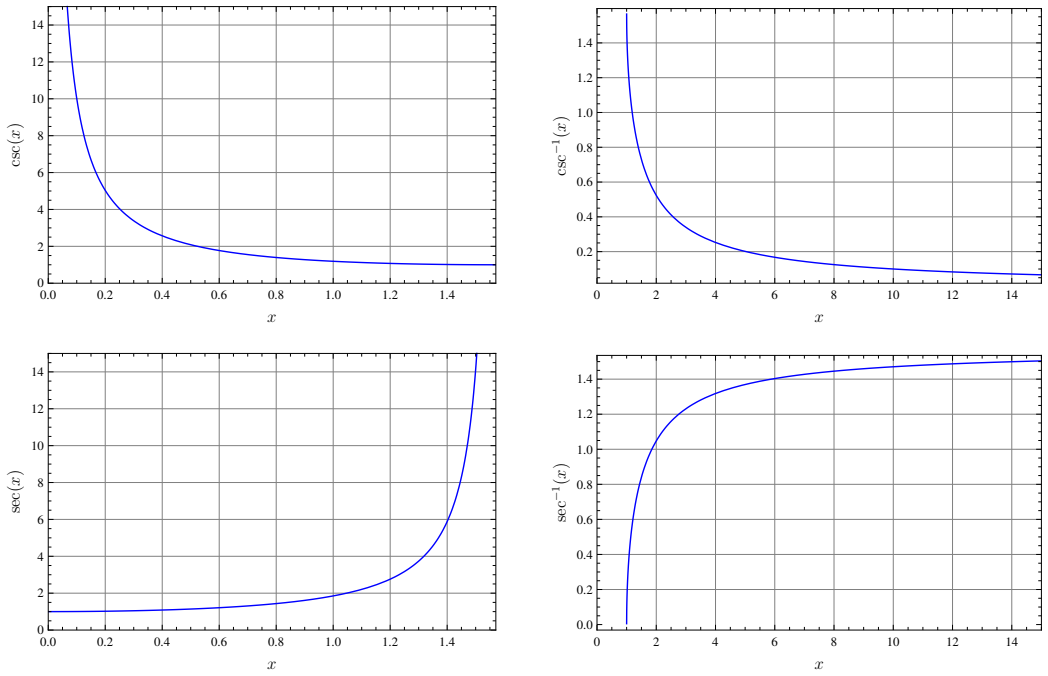


Figure 3.19 Cosecant and its inverse (top) and secant and its inverse (bottom) on $(0, \frac{\pi}{2})$

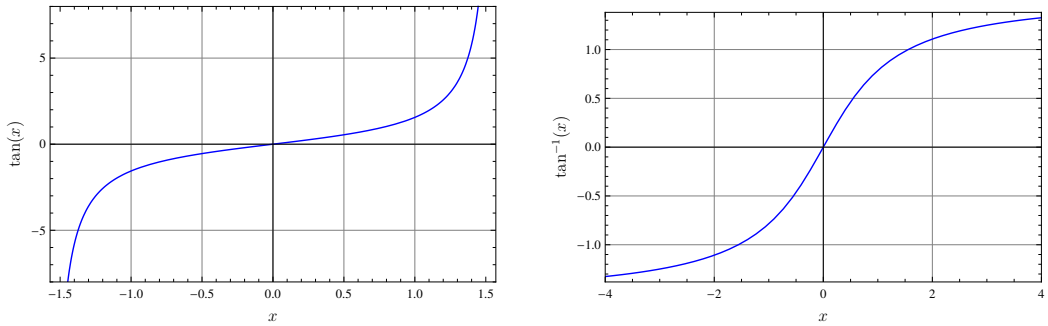


Figure 3.20 The function tan (left) and its inverse $\tan^{-1}$ (right)

As we shall see in ***missing stuff*** when we discuss the geometry of the complex plane, this function returns that angle of a point (x, y) measured from the positive x-axis.

3.6.5 Hyperbolic trigonometric functions

In this section we shall quickly introduce the hyperbolic trigonometric functions. Just why these functions are called "trigonometric" is only best seen in the setting of $\mathbb{C}$-valued functions in ***missing stuff***.

3.6.21 Definition (sinh and cosh) The ***hyperbolic sine function***, denoted by $\sinh\colon \mathbb{R} \to \mathbb{R}$, and the ***hyperbolic cosine function***m denoted by $\cosh\colon \mathbb{R} \to \mathbb{R}$, are defined by

$$\sinh(x) = \sum_{j=1}^{\infty} \frac{x^{2j-1}}{(2j-1)!}, \qquad \cosh(x) = \sum_{j=0}^{\infty} \frac{x^{2j}}{(2j)!},$$

respectively. •

In Figure 3.21 we depict the graphs of sinh and cosh.

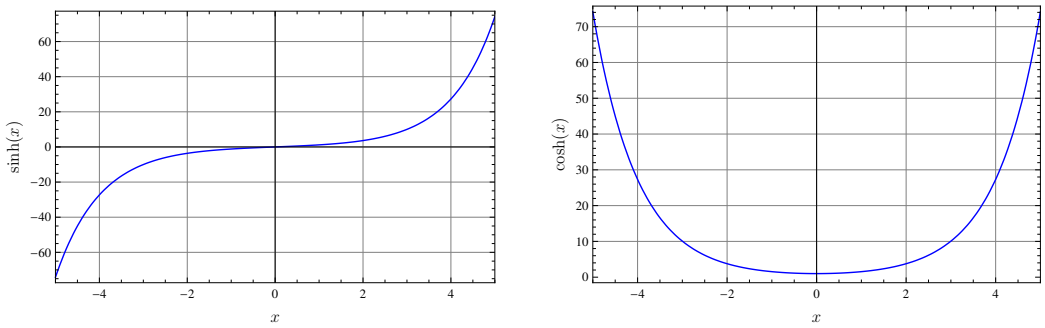


Figure 3.21 The functions sinh (left) and cosh (right)

As with sin and cos, an application of Proposition 2.4.15 and Theorem 3.5.13 shows that the power series expansions for sinh and cosh are convergent for all x.

The following result gives some of the easily determined properties of sinh and cosh.

3.6.22 Proposition (Properties of sinh and cosh) *The functions* sinh *and* cosh *enjoy the following properties:*

(i) $\sinh(x) = \frac{1}{2}(e^x - e^{-x})$ *and* $\cosh(x) = \frac{1}{2}(e^x + e^{-x})$*;*

(ii) sinh *and* cosh *are infinitely differentiable, and furthermore satisfy* $\sinh' = \cosh$ *and* $\cosh' = \sinh$*;*

(iii) $\sinh(-x) = \sinh(x)$ *and* $\cosh(-x) = \cosh(x)$ *for all* $x \in \mathbb{R}$*;*

(iv) $\cosh(x)^2 - \sinh(x)^2 = 1$ *for all* $x \in \mathbb{R}$.

Proof (i) These follows directly from the $\mathbb{R}$-power series definitions for exp, sinh, and cosh.

(ii) This follows from Corollary 3.5.21 and the fact that $\mathbb{R}$-convergent power series can be differentiated term-by-term.

(iii) These follow directly from the $\mathbb{R}$-power series for sinh and cosh.

(iv) This can be proved directly using part (i). ■

Also sometimes useful is the ***hyperbolic tangent function*** $\tanh\colon \mathbb{R} \to \mathbb{R}$ defined by $\tanh(x) = \frac{\sinh(x)}{\cosh(x)}$.

Exercises

3.6.1 For representative values of $a \in \mathbb{R}_{>0}$, give the graph of P_a, showing the features outlined in Proposition 3.6.10.

3.6.2 For representative values of $a \in \mathbb{R}$, give the graph of P^a, showing the features outlined in Proposition 3.6.11.

3.6.3 Prove the following trigonometric identities:

(a) $\cos a \cos b = \frac{1}{2}(\cos(a+b) + \cos(a-b))$;

(b) $\cos a \sin b = \frac{1}{2}(\sin(a+b) - \sin(a-b))$;

(c) $\sin a \sin b = \frac{1}{2}(\cos(a-b) - \cos(a+b))$.

3.6.4 Prove the following trigonometric identities:

(a)

3.6.5 Show that tanh is injective.

Chapter 4

Multiple real variables and functions of multiple real variables

In this chapter we carry on from the preceding chapter and develop the notions of continuity, differentiability, and integrability for functions with multivariable domains and codomains. Much of this development goes in a manner that is strikingly similar to the single-variable case. Therefore, we do not spend as much time with illustrative examples and motivating discussion as we did in Chapter 3. Also some proofs are very similar to their single-variable counterparts, and in these cases we omit detailed proofs. There are, however, some significant differences in the presentation that arise in the extension to multiple variables. For example, the Inverse Function Theorem and the change of variables formula for integrals are far more complicated in the multivariable case. Also, for the multivariable case, one has the important Fubini's Theorem for integrals. Therefore, it is not the case that everything here is simply a trivial extension of what we have already seen in Chapter 3. But it is the case that understanding the material in Chapter 3 will make this chapter far easier to get through.

Do I need to read this chapter? As with the material in Chapter 3, readers who have had a decent sequence of analysis courses can probably skim this chapter on a first reading. This is particularly true if the material in Chapter 3 has been satisfactorily digested. However, there will be occasions where we will use the results in this chapter, so it will have to be come back to at some point if it is not sufficiently well understood. •

Contents

Section 4.1

Norms of Euclidean space and related spaces

In this section we introduce the very most basic structure of Euclidean space: its algebraic structure along with the structure of a norm. Combined, this structure allows us to do analysis in n-dimensional Euclidean space, just as we did in Chapters 2 and 3 for $\mathbb{R}$.

Do I need to read this section? The results in Sections 4.1.1 and 4.1.2 are fundamental to everything in this chapter, and so are required reading. The material in the remaining sections on norms for linear and multilinear maps is required when we define the derivative and higher-order derivatives in Section 4.4. •

4.1.1 The algebraic structure of $\mathbb{R}^n$

We denote by $\mathbb{R}^n$ the n-fold Cartesian product of $\mathbb{R}$ with itself:

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ copies}}.$$

We shall often refer to $\mathbb{R}^n$ as **n-*dimensional Euclidean space***. We shall denote a typical element of $\mathbb{R}^n$ by $\boldsymbol{v} = (v_1, \ldots, v_n)$ when we are talking about the algebraic structure. We call the numbers $v_1, \ldots, _n$ the ***components*** of $\boldsymbol{v}$. We may also use the letters $\boldsymbol{u}$ and $\boldsymbol{w}$. Later in this section, when we discuss properties of $\mathbb{R}^n$ that are not algebraic, we will denote typical points by $\boldsymbol{x} = (x_1, \ldots, x_n)$, and we may also use letters like $\boldsymbol{y}$. Generally speaking, we shall attempt to distinguish between the algebraic and nonalgebraic parts of the structure of $\mathbb{R}^n$.

In $\mathbb{R}$, as we indicated in Section 2.2.1, we can perform familiar algebraic operations like addition, multiplication, and division. Not all of these operations generally carry over to $\mathbb{R}^n$. One *can* add elements of $\mathbb{R}^n$ using the rule

$$\boldsymbol{u} + \boldsymbol{v} = (u_1 + v_1, \ldots, u_n + v_n). \tag{4.1}$$

One can also multiply elements of $\mathbb{R}^n$ by an element of $\mathbb{R}$ using the rule

$$a\boldsymbol{v} = (av_1, \ldots, av_n). \tag{4.2}$$

Let us summarise some of the properties of the algebraic structure of $\mathbb{R}^n$. The following result states that addition (4.1) and multiplication by scalars (4.2) satisfy the axioms for a $\mathbb{R}$-vector space.

4.1.1 Proposition ($\mathbb{R}^n$ is a $\mathbb{R}$-vector space) *The operations (4.1) and (4.2) have the following properties:*

(i) $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1$, $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$ ***(commutativity)***;

(ii) $\mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3) = (\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3$, $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in \mathbb{R}^n$ ***(associativity)***;

(iii) *the element* $\mathbf{0} = (0, \dots, 0) \in \mathbb{R}^n$ *has the property that* $\mathbf{v} + \mathbf{0} = \mathbf{v}$ *for every* $\mathbf{v} \in \mathbb{R}^n$ ***(zero vector)***;

(iv) *for every* $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$ *the element* $-\mathbf{v} = (-v_1, \dots, -v_n) \in \mathbb{R}^n$ *has the property that* $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$ ***(negative vector)***;

(v) $a(b\mathbf{v}) = (ab)\mathbf{v}$, $a, b \in \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$ ***(associativity** again)*;

(vi) $1\mathbf{v} = \mathbf{v}$, $\mathbf{v} \in \mathbb{R}^n$;

(vii) $a(\mathbf{v}_1 + \mathbf{v}_2) = a\mathbf{v}_1 + a\mathbf{v}_2$, $a \in \mathbb{R}$, $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$ ***(distributivity)***;

(viii) $(a_1 + a_2)\mathbf{v} = a_1\mathbf{v} + a_2\mathbf{v}$, $a_1, a_2 \in \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$ ***(distributivity** again)*.

Proof These statements all follow from the properties of algebraic operations on real numbers. ∎

Let us introduce some useful notation for subsets of $\mathbb{R}^n$.***missing stuff***

4.1.2 Definition (Dilation, sum, and difference of sets) Let $A, B \subseteq \mathbb{R}^n$ and let $\lambda \in \mathbb{R}$.

(i) The ***dilation*** of A by λ is the set

$$\lambda A = \{\lambda \boldsymbol{x} \mid \boldsymbol{x} \in A\}.$$

(ii) The ***sum*** of A and B is the set

$$A + B = \{\boldsymbol{x} + \boldsymbol{y} \mid \boldsymbol{x} \in A,\ \boldsymbol{y} \in B\}.$$

(iii) The ***difference*** of A and B is the set

$$A - B = \{\boldsymbol{x} - \boldsymbol{y} \mid \boldsymbol{x} \in A,\ \boldsymbol{y} \in B\}.$$

(iv) If $A = \{\boldsymbol{x}_0\}$ is a singleton, then we denote $A + B = \boldsymbol{x}_0 + B$ and $A - B = \boldsymbol{x}_0 - B$. •

Not all of the algebraic structure of $\mathbb{R}$ carries over to $\mathbb{R}^n$.

1. Generally, one cannot multiply or divide elements of $\mathbb{R}^n$ together in a useful way. However, for $n = 2$ it turns out that multiplication and division *are* also possible, and this is described in Section **??**.[1]
2. Although Zermelo's Well Ordering Theorem tells us that $\mathbb{R}^n$ possesses a well order, apart from $n = 1$ there is no useful (i.e., reacting well with the other structures of $\mathbb{R}^n$) partial order on $\mathbb{R}^n$. Thus any of the results about $\mathbb{R}$ that relate to its natural total order $\leq$ will not generally carry over to $\mathbb{R}^n$.

Let us review some other algebraic concepts and notation associated with $\mathbb{R}^n$. We refer to the general discussions in Sections **??**, **??**, and **??** for more detailed and general discussions.

1. The ***standard basis*** for $\mathbb{R}^n$ is the collection $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ of elements of $\mathbb{R}^n$ given by

$$\boldsymbol{e}_j = (0, \dots, 1, \dots, 0),$$

where the 1 is in the jth position. Obviously we have

$$(v_1, \dots, v_n) = v_1\boldsymbol{e}_1 + \cdots + v_n\boldsymbol{e}_n.$$

[1] There are other values of n for which multiplication and division are possible, but this will not interest us here.

2. The set of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$ is denoted by $\mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^n; \mathbb{R}^m)$ and the set of $m \times n$ matrices with real entries is denoted by $\mathrm{Mat}_{m\times n}(\mathbb{R})$. The sets $\mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^n; \mathbb{R}^m)$ and $\mathrm{Mat}_{m\times n}(\mathbb{R})$ are $\mathbb{R}$-vector spaces and, moreover, are isomorphic in a natural way. Indeed, if $A \in \mathrm{Mat}_{m\times n}(\mathbb{R})$ the corresponding linear map is

$$v \mapsto \Big(\sum_{j=1}^{n} A(1, j)v_j, \ldots, \sum_{j=1}^{n} A(m, j)v_j\Big).$$

4.1.2 The Euclidean inner product and norm, and other norms

There is a generalisation to $\mathbb{R}^n$ of the absolute value function on $\mathbb{R}$. Indeed, this is one of the more valuable features of $\mathbb{R}^n$. In fact, there are many generalisations of the absolute value function which go under the name of "norms;" we shall discuss this idea in detail in Chapter **??**. For now let us just define the norm that is of interest to us. It turns out that the norm we use most in this section is a special sort of norm, derived from an inner product.

4.1.3 Definition (Euclidean inner product) The ***Euclidean inner product*** on $\mathbb{R}^n$ is the map $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$ defined by

$$\langle x, y \rangle_{\mathbb{R}^n} = \sum_{j=1}^{n} x_j y_j. \qquad \bullet$$

This is sometimes called the "dot product" and instead the notation $x \cdot y$ is used. We shall absolutely never use this notation; it is something to be used only by small children.

Let us give some properties of the Euclidean inner product.

4.1.4 Proposition (Properties of the Euclidean inner product) *The Euclidean inner product has the following properties:*

(i) $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^n} = \langle \mathbf{y}, \mathbf{x} \rangle_{\mathbb{R}^n}$ *for* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (***symmetry***);

(ii) $\langle \alpha\mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^n} = \alpha \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^n}$ *for* $\alpha \in \mathbb{R}$ *and* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (***linearity I***);

(iii) $\langle \mathbf{x}_1 + \mathbf{x}_2, \mathbf{y} \rangle_{\mathbb{R}^n} = \langle \mathbf{x}_1, \mathbf{y} \rangle_{\mathbb{R}^n} + \langle \mathbf{x}_2, \mathbf{y} \rangle_{\mathbb{R}^n}$ *for* $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y} \in \mathbb{R}^n$ (***linearity II***);

(iv) $\|\mathbf{x}\|_{\mathbb{R}^n} \mathbf{x} \geq 0$ *for* $\mathbf{x} \in \mathbb{R}^n$ (***positivity***);

(v) $\|\mathbf{x}\|_{\mathbb{R}^n} \mathbf{x} = 0$ *only if* $\mathbf{x} = \mathbf{0}$ (***definiteness***).

Proof These are all elementary deductions using the definition. ■

As we shall see in Definition **??**, a map assigning to a pair of vectors in any $\mathbb{R}$-vector space a number, with the assignment having the five properties above, is called an "inner product." These are studied in some generality in Chapter **??**.

Readers knowing a little Euclidean geometry are familiar with the notion of vectors being "perpendicular." For grownups, the word is "orthogonal."

4.1.5 Definition (Orthogonal, orthogonal complement) Two vectors $x, y \in \mathbb{R}^n$ are ***orthogonal*** if $\langle x, y \rangle_{\mathbb{R}^n} = 0$. If $S \subseteq \mathbb{R}^n$, the ***orthogonal complement*** of S is the set

$$S^{\perp} = \{x \in \mathbb{R}^n \mid \langle x, y \rangle_{\mathbb{R}^n} = 0 \text{ for all } y \in S\}. \qquad \bullet$$

Let us explore the notion of orthogonality with some examples.

4.1.6 Examples (Orthogonality)

1. Consider two vectors $\boldsymbol{x} = (x_1, x_2)$, $\boldsymbol{y} = (y_1, y_2) \in \mathbb{R}^2$. These vectors are orthogonal if and only if $x_1y_1 + x_2y_2 = 0$. Thinking of one of the vectors, say $\boldsymbol{x}$, as being fixed, this is a linear equation in $\boldsymbol{y}$; we refer to Section **??** for a general discussion of such maps. Here we need only note that the subspace of solutions is two-dimensional when $\boldsymbol{x} = \boldsymbol{0}$ and is one-dimensional otherwise. Thus, obviously, every vector is orthogonal to $\boldsymbol{0}$. To describe the one-dimensional subspace of vectors orthogonal to $\boldsymbol{x} \neq \boldsymbol{0}$ we note that one such vector is $\boldsymbol{y} = (-x_2, x_1)$. Thus this is a basis for one-dimensional subspace of vectors orthogonal to $\boldsymbol{x}$. We show the picture in Figure 4.1, noting that, in this case, orthogonality agrees with our

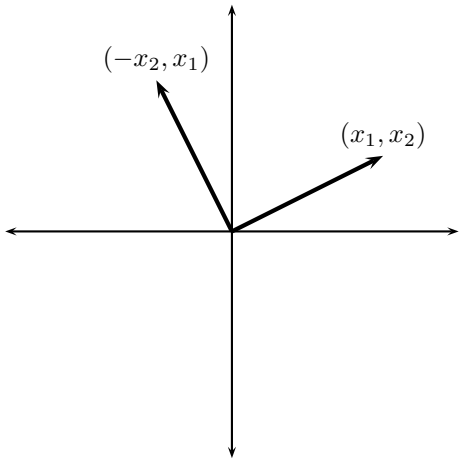


Figure 4.1 Orthogonal vectors in $\mathbb{R}^2$

usual notion of perpendicularity.

2. Let $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ be the standard basis for $\mathbb{R}^n$. Then one readily determines that

$$\langle \boldsymbol{e}_j, \boldsymbol{e}_k \rangle_{\mathbb{R}^n} = \begin{cases} 1, & j = k, \\ 0, & j \neq k. \end{cases}$$

A general basis for $\mathbb{R}^n$ with this property is called ***orthonormal***. Such ideas will be explored in great depth and generality in Chapter **??**. •

We shall not explore the details of what an inner product buys for us, referring the reader to ***missing stuff*** for a general discussion of finite-dimensional vector spaces with inner products. For our purposes the Euclidean inner product is related to the Euclidean norm which is the generalisation of the absolute value function on $\mathbb{R}$ that we shall use to prescribe the structure of Euclidean space.

4.1.7 Definition (Euclidean norm) The ***Euclidean norm*** on $\mathbb{R}^n$ is the function $\|\cdot\|_{\mathbb{R}^n}$ from $\mathbb{R}^n$ to $\mathbb{R}_{\geq 0}$ defined by

$$\|\boldsymbol{x}\|_{\mathbb{R}^n} = \Big(\sum_{j=1}^{n} x_j^2\Big)^{1/2}.$$ •

Note that when $n = 1$ we have $\|\cdot\|_{\mathbb{R}^1} = |\cdot|$. When $n \in \{2, 3\}$, $\|x\|_{\mathbb{R}^n}$ is the usual notion of length in "physical space."

Let us record the properties of the Euclidean norm.

4.1.8 Proposition (Properties of the Euclidean norm) *The Euclidean norm has the following properties:*

(i) $\|\alpha \mathbf{x}\|_{\mathbb{R}^n} = |\alpha| \|\mathbf{x}\|_{\mathbb{R}^n}$ *for* $\alpha \in \mathbb{R}$ *and* $\mathbf{x} \in \mathbb{R}^n$ (***homogeneity***);

(ii) $\|\mathbf{x}\|_{\mathbb{R}^n} \geq 0$ *for all* $\mathbf{x} \in \mathbb{R}^n$ (***positivity***);

(iii) $\|\mathbf{x}\|_{\mathbb{R}^n} = 0$ *only if* $\mathbf{x} = \mathbf{0}$ (***definiteness***);

(iv) $\|\mathbf{x}_1 + \mathbf{x}_2\|_{\mathbb{R}^n} \leq \|\mathbf{x}_1\|_{\mathbb{R}^n} + \|\mathbf{x}_2\|_{\mathbb{R}^n}$ (***triangle inequality***).

Moreover, the Euclidean norm shares the following relationships with the Euclidean inner product:

(v) $\|\mathbf{x}\|_{\mathbb{R}^n} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{R}^n}}$ *for all* $\mathbf{x} \in \mathbb{R}^n$;

(vi) $|\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^n}| \leq \|\mathbf{x}\|_{\mathbb{R}^n} \|\mathbf{y}\|_{\mathbb{R}^n}$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (***Cauchy–Bunyakovsky– Schwarz inequality***).

Proof The only nontrivial properties are the fourth one and the final one. We first prove the Cauchy–Bunyakovsky–Schwarz inequality and then use it to prove the triangle inequality.

The Cauchy–Bunyakovsky–Schwarz inequality is obviously true for $y = \mathbf{0}$, so we shall suppose that $y \neq \mathbf{0}$. We first prove the result for $\|y\|_{\mathbb{R}^n} = 1$. In this case we have

$$\begin{aligned}
0 &\leq \|x - \langle x, y \rangle_{\mathbb{R}^n} y\|_{\mathbb{R}^n}^2 \\
&= \langle x - \langle x, y \rangle_{\mathbb{R}^n} y, x - \langle x, y \rangle_{\mathbb{R}^n} y \rangle_{\mathbb{R}^n} \\
&= \langle x, x \rangle_{\mathbb{R}^n} - \langle x, y \rangle_{\mathbb{R}^n} \langle y, x \rangle_{\mathbb{R}^n} - \langle x, y \rangle_{\mathbb{R}^n} \langle x, y \rangle_{\mathbb{R}^n} + \langle x, y \rangle_{\mathbb{R}^n} \langle x, y \rangle_{\mathbb{R}^n} \langle y, y \rangle_{\mathbb{R}^n} \\
&= \|x\|_{\mathbb{R}^n}^2 - \langle x, y \rangle_{\mathbb{R}^n}^2.
\end{aligned}$$

Thus we have shown that provided $\|y\|_{\mathbb{R}^n} = 1$, $\langle x, y \rangle_{\mathbb{R}^n}^2 \leq \|x\|_{\mathbb{R}^n}^2$. Taking square roots yields the result in this case. For $\|y\|_{\mathbb{R}^n} \neq 1$ we define $z = \frac{y}{\|y\|_{\mathbb{R}^n}}$ so that $\|z\|_{\mathbb{R}^n} = 1$. In this case

$$\big|\langle x, z \rangle_{\mathbb{R}^n}\big| \leq \|x\|_{\mathbb{R}^n} \quad \Longrightarrow \quad \frac{\big|\langle x, y \rangle_{\mathbb{R}^n}\big|}{\|y\|_{\mathbb{R}^n}} \leq \|x\|_{\mathbb{R}^n},$$

and so the inequality follows.

Now, to prove the triangle inequality, we compute

$$\begin{aligned}
\|x + y\|_{\mathbb{R}^n}^2 &= \langle x + y, x + y \rangle_{\mathbb{R}^n} \\
&= \|x\|_{\mathbb{R}^n}^2 + 2\langle x, y \rangle_{\mathbb{R}^n} + \|y\|_{\mathbb{R}^n}^2 \\
&\leq \|x\|_{\mathbb{R}^n}^2 + 2\big|\langle x, y \rangle_{\mathbb{R}^n}\big| + \|y\|_{\mathbb{R}^n}^2 \\
&\leq \|x\|_{\mathbb{R}^n}^2 + 2\|x\|_{\mathbb{R}^n} \|y\|_{\mathbb{R}^n} + \|y\|_{\mathbb{R}^n}^2 \\
&= (\|x\|_{\mathbb{R}^n} + \|y\|_{\mathbb{R}^n})^2,
\end{aligned}$$

where we have used the lemma. The result now follows by taking square roots. ■

As we shall see in Definition **??**, a map assigning to vectors in a $\mathbb{R}$-vector space a number, with the assignment having the three properties above, is a "norm." These are studied in detail in Chapter **??**.

Sometimes we will use other norms for $\mathbb{R}^n$. Two common norms are given in the following definition.

4.1.9 Definition (1- and ∞-norm for Euclidean space) The ***1-norm*** on $\mathbb{R}^n$ is the function $\|\cdot\|_1$ from $\mathbb{R}^n$ to $\mathbb{R}_{\geq 0}$ defined by

$$\|\boldsymbol{x}\|_1 = \sum_{j=1}^{n} |x_j|,$$

and the ***∞-norm*** on $\mathbb{R}^n$ is the function $\|\cdot\|_\infty$ from $\mathbb{R}^n$ to $\mathbb{R}_{\geq 0}$ defined by

$$\|\boldsymbol{x}\|_\infty = \max\{|x_1|, \ldots, |x_n|\}. \qquad \bullet$$

The 1- and ∞-norms enjoy the following properties, as is easily verified (see also Examples Example **??**–**??** and **??** and Section **??**).

4.1.10 Proposition (Properties of the 1- and ∞-norms) *For* $\mathrm{p} \in \{1, \infty\}$*, the* p*-norm has the following properties:*

(i) $\|\alpha \mathbf{x}\|_\mathrm{p} = |\alpha| \|\mathbf{x}\|_\mathrm{p}$ *for* $\alpha \in \mathbb{R}$ *and* $\mathbf{x} \in \mathbb{R}^\mathrm{n}$ ***(homogeneity)****;*

(ii) $\|\mathbf{x}\|_\mathrm{p} \geq 0$ *for all* $\mathbf{x} \in \mathbb{R}^\mathrm{n}$ ***(positivity)****;*

(iii) $\|\mathbf{x}\|_\mathrm{p} = 0$ *only if* $\mathbf{x} = \mathbf{0}$ ***(definiteness)****;*

(iv) $\|\mathbf{x}_1 + \mathbf{x}_2\|_\mathrm{p} \leq \|\mathbf{x}_1\|_\mathrm{p} + \|\mathbf{x}_2\|_\mathrm{p}$ ***(triangle inequality)****.*

When we are simultaneously discussing and contrasting the various norms, we will sometime use $\|\cdot\|_2$ rather than $\|\cdot\|_{\mathbb{R}^n}$ to denote the Euclidean norm, and we may refer to this norm as the ***2-norm***.

The following relationships between the 1-, 2-, and ∞-norms are often useful.

4.1.11 Proposition (Relationships between the 1-, 2-, and ∞-norms) *For* $\mathbf{v} \in \mathbb{R}^\mathrm{n}$ *we have the following inequalities:*

(i) $\|\mathbf{v}\|_1 \leq \sqrt{\mathrm{n}} \|\mathbf{v}\|_2$*;*

(ii) $\|\mathbf{v}\|_1 \leq \mathrm{n} \|\mathbf{v}\|_\infty$*;*

(iii) $\|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1$*;*

(iv) $\|\mathbf{v}\|_2 \leq \sqrt{\mathrm{n}} \|\mathbf{v}\|_\infty$*;*

(v) $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_1$*;*

(vi) $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2$.

Moreover, the above inequalities are the best possible in the sense that, in each case, there exists a vector $\mathbf{v} \in \mathbb{R}^\mathrm{n}$ *such that equality is satisfied.*

Proof (i) Note that the expression

$$\|\boldsymbol{v}\|_1 = \sum_{j=1}^{n} |v_j|$$

means that $n\|\boldsymbol{v}\|_1$ is the average of the positive numbers $|v_1|, \ldots, |v_n|$. Thus we can write each of these numbers as this average divided by n plus the difference: $|v_j| = \frac{\|\boldsymbol{v}\|_1}{n} + \delta_j$. Note that $\sum_{j=1}^n \delta_j = 0$. Now compute

$$\begin{aligned}\|\boldsymbol{v}\|_2 &= \Big(\sum_{j=1}^n |v_j|^2\Big)^{1/2} = \Big(\sum_{j=1}^n \Big(\frac{\|\boldsymbol{v}\|_1}{n} + \delta_j\Big)^2\Big)^{1/2} \\ &= \Big(\sum_{j=1}^n \Big(\frac{\|\boldsymbol{v}\|_1^2}{n^2} + 2\frac{\|\boldsymbol{v}\|_1\delta_j}{n} + \delta_j^2\Big)\Big)^{1/2} \geq \Big(\sum_{j=1}^n \frac{\|\boldsymbol{v}\|_1^2}{n^2}\Big)^{1/2} = \frac{\|\boldsymbol{v}\|_1}{\sqrt{n}},\end{aligned}$$

as desired, using the fact that $\sum_{j=1}^n \delta_j = 0$. The inequality is an equality by taking, for example, $\boldsymbol{v} = (1, \ldots, 1)$.

(ii) We have

$$\|\boldsymbol{v}\|_1 = \sum_{j=1}^n |v_j| \leq \sum_{j=1}^n \max\{|v_j| \mid \ j \in \{1, \ldots, n\}\} = n\|\boldsymbol{v}\|_\infty.$$

The inequality becomes equality, for example, for the vector $(1, \ldots, 1)$.

(iii) We have

$$\|\boldsymbol{v}\|_2 = \Big\|\sum_{j=1}^n v_j \boldsymbol{e}_j\Big\|_2 \leq \sum_{j=1}^n \|v_j \boldsymbol{e}_j\|_2 = \sum_{j=1}^n |v_j| \|e_j\|_2 = \sum_{j=1}^n |v_j| = \|\boldsymbol{v}\|_1.$$

The inequality becomes equality if, for example, $\boldsymbol{v} = (1, 0, \ldots, 0)$.

(iv) First note that the inequality is trivially satisfied when $\boldsymbol{v} = \boldsymbol{0}_{\mathbb{F}^n}$. If $\|\boldsymbol{v}\|_\infty = 1$ we have $|v_j| \leq 1$ whence $|v_j|^2 \leq |v_j|$ for $j \in \{1, \ldots, n\}$. Therefore, in this case we have

$$\|\boldsymbol{v}\|_2^2 = \sum_{j=1}^n \|v_j\|^2 \leq \sum_{j=1}^n |v_j| \leq \sum_{j=1}^n \max\{|v_j| \mid \ j \in \{1, \ldots, n\}\} = n\|\boldsymbol{v}\|_\infty.$$

Therefore, taking square roots, when $\|\boldsymbol{v}\|_\infty = 1$ we have $\|\boldsymbol{v}\|_2 \leq \sqrt{n}\|\boldsymbol{v}\|_\infty$. For general nonzero $\boldsymbol{v}$ we write $\boldsymbol{v} = \lambda \boldsymbol{u}$ where $\|\boldsymbol{u}\|_\infty = 1$ and where $\lambda = \|\boldsymbol{v}\|_\infty$. We then have

$$\|\boldsymbol{v}\|_2 = |\lambda| \|\boldsymbol{u}\|_2 \leq \lambda \sqrt{n} \|\boldsymbol{u}\|_\infty = \sqrt{n}\|\boldsymbol{v}\|_\infty,$$

giving the desired result. The inequality becomes equality by taking, for example, $\boldsymbol{v} = (1, \ldots, 1)$.

(v) Let $j_0 \in \{1, \ldots, n\}$ be such that

$$|v_{j_0}| = \max\{|v_j| \mid \ j \in \{1, \ldots, n\}\}.$$

Then

$$\|\boldsymbol{v}\|_\infty = |v_{j_0}| \leq \sum_{j=1}^n |v_j|.$$

The inequality becomes equality, for example, for the vector $(1, 0, \ldots, 0)$.

(vi) Let $j_0 \in \{1, \ldots, n\}$ be such that

$$|v_{j_0}| = \max\{|v_j| \mid \ j \in \{1, \ldots, n\}\}.$$

Then

$$\|v\|_\infty^2 = |v_{j_0}|^2 \leq \sum_{j=1}^n |v_j|^2 = \|v\|_2^2.$$

Taking square roots gives $\|v\|_\infty \leq \|v\|_2$.

The inequality becomes equality, for example, for the vector $(1, 0, \ldots, 0)$. ■

The ideas of norms and inner products are explored in some detail in Chapters **??** and **??**.

4.1.3 Norms for multilinear maps

One of the places in the development of multivariable differentiation in Section 4.4 departs from the single-variable case is in higher-order derivatives. In the single-variable case, the derivative of a function is again a function, and so higher-order derivatives can be defined inductively as functions. But in the multivariable case, the derivative is a linear map as we shall see, and so to talk about higher-order derivatives one must talk intelligently about functions taking values in the set of linear maps. There are two facets to this. Firstly we must be comfortable with the algebraic aspects of multilinear maps. These are dealt with in Section **??**, and the reader will have to understand some material from this section before proceeding. Secondly, in order to inductively define higher-order derivatives we must have norms on sets of multilinear maps. We implicitly identify the set $\mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^n; \mathbb{R}^m)$ of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$ and the set $\mathrm{Mat}_{m\times n}(\mathbb{R})$ of $m \times n$ matrices with entries in $\mathbb{R}$; see Definition **??**. Thus for linear maps, the norms are sometimes called matrix norms.

First of all, we shall use somewhat more compact notation for multilinear maps than is used in Section **??**. Namely, we denote by $\mathrm{L}(\mathbb{R}^{n_1}, \ldots, \mathbb{R}^{n_k}; \mathbb{R}^m)$ the set of $\mathbb{R}$-multilinear maps from $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ to $\mathbb{R}^m$. (In Section **??** we denoted this set of multilinear maps by $\mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^{n_1}, \ldots, \mathbb{R}^{n_k}; \mathbb{R}^m)$.) In the particular (and in this section usual) case when $n_1 = \cdots = n_k = n$ then we denote the multilinear maps from $(\mathbb{R}^n)^k$ to $\mathbb{R}^m$ by $\mathrm{L}^k(\mathbb{R}^n; \mathbb{R}^m)$. We also recall that a multilinear map $\mathsf{L} \in \mathrm{L}^k(\mathbb{R}^n; \mathbb{R}^m)$ is ***symmetric*** if

$$\mathsf{L}(v_{\sigma(1)}, \ldots, v_{\sigma(k)}) = \mathsf{L}(v_1, \ldots, v_k)$$

for every permutation $\sigma \in \mathfrak{S}_k$. We denote the set of symmetric multilinear maps from $(\mathbb{R}^n)^k$ to $\mathbb{R}^m$ by $S^k(\mathbb{R}^n; \mathbb{R}^m)$.

Our notation for multilinear maps will come back to us in ***missing stuff*** when we talk about continuous linear maps between normed vector spaces and in ***missing stuff*** when we talk about linear maps between topological vector spaces. In finite-dimensions all multilinear maps are continuous and so our notationally identifying $\mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^{n_1}, \ldots, \mathbb{R}^{n_k}; \mathbb{R}^m)$ with the continuous multilinear maps is justified. All that justification aside, all we care about is that $\mathrm{L}(\mathbb{R}^{n_1}, \ldots, \mathbb{R}^{n_k}; \mathbb{R}^m)$ denotes the set of $\mathbb{R}$-multilinear maps from $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ to $\mathbb{R}^m$. Now we need to put norms on sets of linear and multilinear maps. The reader may well wish to refer ahead to Section **??** for a general introduction to norms. Only the elementary definitions and examples from that section are needed here.

We will let $\|\cdot\|$ denote an arbitrary norm on $\mathbb{R}^n$. In practice, we shall most often take $\|\cdot\|$ to be the Euclidean norm, but we stick to a more general setup for simplicity. When talking about maps between $\mathbb{R}^n$ and $\mathbb{R}^m$, we will have norms on both spaces, and we shall denote both of these norms, and any norm induced by them, by $\|\cdot\|$, accepting an abuse of notation that does not cause problems.

With all of this preamble, we can now make the following definition.

4.1.12 Definition (Induced norm on the set of multilinear maps) Let $\|\cdot\|_{\alpha_1}, \dots, \|\cdot\|_{\alpha_k}$ be norms on $\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_k}$, respectively, and let $\|\cdot\|_\beta$ be a norm on $\mathbb{R}^m$. For $\mathsf{L} \in \mathrm{L}(\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_k}; \mathbb{R}^m)$ the ***induced norm*** of L is

$$\|\mathsf{L}\|_{\alpha,\beta} = \inf\{M \in \mathbb{R}_{>0} \mid \|\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq M\|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\}.$$

•

Let us verify that the proposed norm is indeed a norm. The reader may wish to refer to Section **??** for more information in the case of linear maps.

4.1.13 Proposition (The induced norm is a norm) *The induced norm defined in Definition 4.1.12 is a norm on* $\mathrm{L}(\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_k}; \mathbb{R}^m)$. *Moreover, for every* $\mathbf{x}_j \in \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$,

$$\|\mathsf{L}(\mathbf{x}_1, \dots, \mathbf{x}_k)\|_\beta \leq \|\mathsf{L}\|_{\alpha,\beta}\|\mathbf{x}_1\|_{\alpha_1} \cdots \|\mathbf{x}_k\|_{\alpha_k}.$$

Proof Let $\{\boldsymbol{e}_1, \dots, \mathit{vecte}_d\}$ be the standard basis for $\mathbb{R}^d$. For $\mathsf{L} \in \mathrm{L}(\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_k}; \mathbb{R}^m)$ define $\mathsf{L}^l_{j_1 \cdots j_k}$, $j_1 \in \{1, \dots, n_1\}, \dots, j_k \in \{1, \dots, n_k\}$, $l \in \{1, \dots, m\}$, by

$$\mathsf{L}(\boldsymbol{e}_{j_1}, \dots, \boldsymbol{e}_{j_k}) = \sum_{l=1}^m \mathsf{L}^m_{j_1 \cdots j_k} \boldsymbol{e}_l.$$

For $\boldsymbol{x}_j \in \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$, let us write

$$\boldsymbol{x}_j = x_j^1 \boldsymbol{e}_1 + \cdots + x_j^{n_j} \boldsymbol{e}_{n_j}.$$

Then we have, by multilinearity of L,

$$\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k) = \sum_{j_1=1}^{n_1} \cdots \sum_{j_k=1}^{n_k} \sum_{l=1}^m \mathsf{L}^l_{j_1 \cdots j_k} x_1^{j_1} \cdots x_k^{j_k} \boldsymbol{e}_l.$$

This shows that L is continuous since its components are polynomial functions of the components, and such functions are continuous.

Let us denote by $\overline{\mathsf{B}}(r, \boldsymbol{x})$ the closed ball of radius r centred at $\boldsymbol{x}$. We shall use the same notation for balls in any norm. Since L is continuous, by Theorem 4.3.31 it is bounded when restricted to the compact set $\overline{\mathsf{B}}(1, \mathbf{0}) \times \cdots \times \overline{\mathsf{B}}(1, \mathbf{0})$. Let

$$M = \sup\{\|\mathsf{L}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_k)\|_\beta \mid \|\boldsymbol{u}_j\|_{\alpha_j} = 1,\ j \in \{1, \dots, k\}\}.$$

For $\boldsymbol{x}_j \in \mathbb{R}^{n_j} \setminus \{\mathbf{0}\}$, $j \in \{1, \dots, k\}$, we then have

$$\|\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta = \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k} \mathsf{L}\Big(\frac{\boldsymbol{x}_1}{\|\boldsymbol{x}_1\|_{\alpha_1}}, \dots, \frac{\boldsymbol{x}_k}{\|\boldsymbol{x}_k\|_{\alpha_k}}\Big) \leq M\|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k}.$$

This shows that $\|\mathsf{L}\|_{\boldsymbol{\alpha},\beta} < \infty$ and so is well-defined.

Let us next verify the final assertion of the proposition. Suppose that there exists $\boldsymbol{x}_j \in \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$, such that

$$\|\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta > \|\mathsf{L}\|_{\boldsymbol{\alpha},\beta} \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k}.$$

Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that

$$\|\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta > (\|\mathsf{L}\|_{\boldsymbol{\alpha},\beta} - \epsilon) \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},$$

and this contradicts the definition of $\|\mathsf{L}\|_{\boldsymbol{\alpha},\beta}$. Thus we must have

$$\|\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq \|\mathsf{L}\|_{\boldsymbol{\alpha},\beta} \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k}, \tag{4.3}$$

as desired.

Now we show that $\mathsf{L} \mapsto \|\mathsf{L}\|_{\boldsymbol{\alpha},\beta}$ has the properties of a norm. It is clear that $\|\mathsf{L}\|_{\boldsymbol{\alpha},\beta} \geq 0$ and that $\|\mathsf{L}\|_{\boldsymbol{\alpha},\beta} = 0$ when $\mathsf{L} = 0$. Suppose that $\|\mathsf{L}\|_{\boldsymbol{\alpha},\beta} = 0$. Then, by (4.3), for every $\boldsymbol{x}_j \in \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$,

$$\|\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq \|\mathsf{L}\|_{\boldsymbol{\alpha},\beta} \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k} = 0,$$

giving $\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k) = 0$, and so $\mathsf{L} = 0$. Note that $\|0\mathsf{L}\|_{\boldsymbol{\alpha},\beta} = |0| \|\mathsf{L}\|_{\boldsymbol{\alpha},\beta}$. Also, if $a \in \mathbb{R} \setminus \{0\}$, then

$$\begin{aligned}
\|a\mathsf{L}\|_{\boldsymbol{\alpha},\beta} &= \inf\{M \in \mathbb{R}_{>0} \mid \|a\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq M \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\} \\
&= \inf\{M \in \mathbb{R}_{>0} \mid |a| \|\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq M \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\} \\
&= \inf\left\{M \in \mathbb{R}_{>0} \;\middle|\; \|\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq \frac{M}{|a|} \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\right\} \\
&= \inf\{|a|M' \in \mathbb{R}_{>0} \mid \|\mathsf{L}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq M' \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\} \\
&= |a| \|\mathsf{L}\|_{\boldsymbol{\alpha},\beta},
\end{aligned}$$

using Proposition 2.2.28. Finally, if $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{L}(\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_k}; \mathbb{R}^m)$, then

$$\begin{aligned}
\|\mathsf{L}_1 + \mathsf{L}_2\|_{\boldsymbol{\alpha},\beta} &= \inf\{M \in \mathbb{R}_{>0} \mid \|(\mathsf{L}_1 + \mathsf{L}_2)(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \\
&\qquad \leq M \|\boldsymbol{x}\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\} \\
&\leq \inf\{M \in \mathbb{R}_{>0} \mid \|\mathsf{L}_1(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \\
&\qquad + \|\mathsf{L}_2(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq M \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\} \\
&= \inf\{M_1 + M_2 \in \mathbb{R}_{>0} \mid \|\mathsf{L}_1(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq M_1 \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k}, \\
&\qquad \|\mathsf{L}_2(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq M_2 \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\} \\
&= \inf\{M \in \mathbb{R}_{>0} \mid \|\mathsf{L}_1(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \leq \\
&\qquad M \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\} \\
&\qquad + \inf\{M \in \mathbb{R}_{>0} \mid \|\mathsf{L}_2(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_\beta \\
&\qquad \leq M \|\boldsymbol{x}_1\|_{\alpha_1} \cdots \|\boldsymbol{x}_k\|_{\alpha_k},\ \boldsymbol{x}_j \in \mathbb{R}^{n_j},\ j \in \{1, \dots, k\}\} \\
&= \|\mathsf{L}_1\|_{\boldsymbol{\alpha},\beta} + \|\mathsf{L}_2\|_{\boldsymbol{\alpha},\beta},
\end{aligned}$$

using Proposition 2.2.28. ∎

4.1.4 The nine common induced norms for linear maps

Let us consider a collection of special cases for linear maps. We use the three norms

$$\|\boldsymbol{x}\|_1 = \sum_{j=1}^{n} |x_j|, \quad \|\boldsymbol{x}\|_2 = \Big(\sum_{j=1}^{n} x_j^2\Big)^{1/2}, \quad \|\boldsymbol{x}\|_\infty = \max\{|x_1|, \dots, |x_n|\}$$

on $\mathbb{R}^n$, noting that $\|\cdot\|_2$ is the Euclidean norm, which we have also denoted by $\|\cdot\|_{\mathbb{R}^n}$. Let us characterise the nine possible induced norms

$$\|\mathsf{L}\|_{p,q} \triangleq \inf\{M \in \mathbb{R}_{>0} \mid \|\mathsf{L}(\boldsymbol{x})\|_q \le M\|\boldsymbol{x}\|_p,\ \boldsymbol{x} \in \mathbb{R}^n\}, \qquad p, q \in \{1, 2, \infty\},$$

on $\mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$ induced by these three norms. In the statement of the following theorem, recall from Definition **??** that $\boldsymbol{c}(\mathsf{L}, j) \in \mathbb{R}^m$, $j \in \{1, \dots, n\}$, denotes the jth column vector of L and $\boldsymbol{r}(\mathsf{L}, a) \in \mathbb{R}^n$, $a \in \{1, \dots, m\}$, denotes the ath row vector of L, where we recall from Theorem **??** that there is a natural correspondence between finite matrices and linear maps.

4.1.14 Theorem (Induced norms for linear maps) *Let* $\mathrm{p}, \mathrm{q} \in \{1, 2, \infty\}$ *and let* $\mathsf{L} \in \mathrm{L}(\mathbb{R}^{\mathrm{m}}; \mathbb{R}^{\mathrm{m}})$. *The induced norm* $\|\cdot\|_{\mathrm{p},\mathrm{q}}$ *satisfies the following formulae:*

(i) $\|\mathsf{L}\|_{1,1} = \max\{\|\mathbf{c}(\mathsf{L}, \mathrm{j})\|_1 \mid \mathrm{j} \in \{1, \dots, \mathrm{n}\}\}$;

(ii) $\|\mathsf{L}\|_{1,2} = \max\{\|\mathbf{c}(\mathsf{L}, \mathrm{j})\|_2 \mid \mathrm{j} \in \{1, \dots, \mathrm{n}\}\}$;

(iii) $\begin{aligned}\|\mathsf{L}\|_{1,\infty} &= \max\{|\mathsf{L}(\mathrm{a}, \mathrm{j})| \mid \mathrm{a} \in \{1, \dots, \mathrm{m}\},\ \mathrm{j} \in \{1, \dots, \mathrm{n}\}\};\\ &= \max\{\|\mathbf{c}(\mathsf{L}, \mathrm{j})\|_\infty \mid \mathrm{j} \in \{1, \dots, \mathrm{n}\}\}\\ &= \max\{\|\mathbf{r}(\mathsf{L}, \mathrm{a})\|_\infty \mid \mathrm{a} \in \{1, \dots, \mathrm{m}\}\}\end{aligned}$

(iv) $\|\mathsf{L}\|_{2,1} = \max\{\|\mathsf{L}^{\mathrm{T}}(\mathbf{u})\|_2 \mid \mathbf{u} \in \{-1, 1\}^{\mathrm{m}}\}$;

(v) $\|\mathsf{L}\|_{2,2} = \max\{\sqrt{\lambda} \mid \lambda \textit{ is an eigenvalue for } \mathsf{L}^{\mathrm{T}}\mathsf{L}\}$;

(vi) $\|\mathsf{L}\|_{2,\infty} = \max\{\|\mathbf{r}(\mathsf{L}, \mathrm{a})\|_2 \mid \mathrm{a} \in \{1, \dots, \mathrm{m}\}\}$;

(vii) $\|\mathsf{L}\|_{\infty,1} = \max\{\|\mathsf{L}(\mathbf{u})\|_1 \mid \mathbf{u} \in \{-1, 1\}^{\mathrm{n}}\}$;

(viii) $\|\mathsf{L}\|_{\infty,2} = \max\{\|\mathsf{L}(\mathbf{u})\|_2 \mid \mathbf{u} \in \{-1, 1\}^{\mathrm{n}}\}$;

(ix) $\|\mathsf{L}\|_{\infty,\infty} = \max\{\|\mathbf{r}(\mathsf{L}, \mathrm{a})\|_1 \mid \mathrm{a} \in \{1, \dots, \mathrm{m}\}\}$.

Proof In the proof we make free use of results we have not yet proved. We also make frequent use of the obvious formula

$$\mathsf{L}(\boldsymbol{x}) = \Big(\langle \boldsymbol{r}(\mathsf{L}, 1), \boldsymbol{x}\rangle_{\mathbb{R}^n}, \dots, \langle \boldsymbol{r}(\mathsf{L}, m), \boldsymbol{x}\rangle_{\mathbb{R}^n}\Big).$$

Let $\mathsf{L} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$ and note that

$$\begin{aligned}\|\mathsf{L}\| &= \inf\{M \in \mathbb{R}_{>0} \mid \|\mathsf{L}(\boldsymbol{x})\| \le M\|\boldsymbol{x}\|,\ \boldsymbol{x} \in \mathbb{R}^n\}\\ &= \{M \in \mathbb{R}_{>0} \mid \|\mathsf{L}(\boldsymbol{x})\| \le M\|\boldsymbol{x}\|,\ \boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}\}\\ &= \{M \in \mathbb{R}_{>0} \mid \|\mathsf{L}(\tfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|})\| \le M,\ \boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}\}\\ &= \sup\{\|\mathsf{L}(\boldsymbol{x})\| \mid \|\boldsymbol{x}\| = 1\}.\end{aligned}$$

We shall use this characterisation of the norm below.

In the proof, we also let $\{e_1, \dots, e_d\}$ be the standard basis for $\mathbb{R}^d$.
(i) We compute

$$\begin{aligned}
\|\mathsf{L}\|_{1,1} &= \sup\{\|\mathsf{L}(x)\|_1 \mid \|x\|_1 = 1\} \\
&= \sup\Big\{\sum_{a=1}^{m} |\langle r(\mathsf{L}(x)), x\rangle_{\mathbb{R}^n}| \;\Big|\; \|x\|_1 = 1\Big\} \\
&\leq \sup\Big\{\sum_{a=1}^{m}\sum_{j=1}^{n} |\mathsf{L}(a,j)||x_j| \;\Big|\; \|x\|_1 = 1\Big\} \\
&= \sup\Big\{\sum_{j=1}^{n} |x_j|\Big(\sum_{a=1}^{m} |\mathsf{L}(a,j)|\Big) \;\Big|\; \|x\|_1 = 1\Big\} \\
&\leq \max\Big\{\sum_{a=1}^{m} |\mathsf{L}(a,j)| \;\Big|\; j \in \{1,\dots,n\}\Big\} \\
&= \max\{\|c(\mathsf{L},j)\|_1 \mid j \in \{1,\dots,n\}\}.
\end{aligned}$$

To establish the opposite inequality, suppose that $k \in \{1, \dots, n\}$ is such that

$$\|c(\mathsf{L},k)\|_1 = \max\{\|c(\mathsf{L},j)\|_1 \mid j \in \{1,\dots,n\}\}.$$

Then,

$$\|\mathsf{L}(e_k)\|_1 = \sum_{a=1}^{m} \Big|\Big(\sum_{j=1}^{n} \mathsf{L}(a,j)e_k(j)\Big)\Big| = \sum_{a=1}^{m} |\mathsf{L}(a,k)| = \|c(\mathsf{L},k)\|_1.$$

Thus

$$\|\mathsf{L}\|_{1,1} \geq \max\{\|c(\mathsf{L},j)\|_1 \mid j \in \{1,\dots,n\}\},$$

since $\|e_k\|_1 = 1$.
(ii) We compute

$$\begin{aligned}
\|\mathsf{L}\|_{1,2} &= \sup\{\|\mathsf{L}(x)\|_2 \mid \|x\|_1 = 1\} \\
&= \sup\Big\{\Big(\sum_{a=1}^{m} \langle r(\mathsf{L},a), x\rangle_{\mathbb{R}^n}^2\Big)^{1/2} \;\Big|\; \|x\|_1 = 1\Big\} \\
&\leq \sup\Big\{\Big(\sum_{a=1}^{m}\Big(\sum_{j=1}^{n} |\mathsf{L}(a,j)x_j|\Big)^2\Big)^{1/2} \;\Big|\; \|x\|_1 = 1\Big\} \\
&\leq \sup\Big\{\Big(\sum_{a=1}^{m} (\max\{|\mathsf{L}(a,j)| \mid j \in \{1,\dots,n\}\})^2\Big(\sum_{j=1}^{n} |x_j|\Big)^2\Big)^{1/2} \;\Big|\; \|x\|_1 = 1\Big\} \\
&= \Big(\sum_{a=1}^{m} (\max\{|\mathsf{L}(a,j)| \mid j \in \{1,\dots,n\}\})^2\Big)^{1/2} \\
&= \Big(\max\Big\{\sum_{a=1}^{m} \mathsf{L}(a,j)^2 \;\Big|\; j \in \{1,\dots,n\}\Big\}\Big)^{1/2} = \max\{\|c(\mathsf{L},j)\|_2 \mid j \in \{1,\dots,n\}\},
\end{aligned}$$

using Proposition 2.2.27 and the fact that

$$\sup\{\|x\|_2 \mid \|x\|_1 = 1\} = 1.$$

To establish the other inequality, note that if we take $k \in \{1, \dots, n\}$ such that

$$\|\mathbf{c}(\mathsf{L}, k)\|_2 = \max\{\|\mathbf{c}(\mathsf{L}, j)\|_2 \mid j \in \{1, \dots, n\}\},$$

then we have

$$\|\mathsf{L}(\mathbf{e}_k)\|_2 = \Big(\sum_{a=1}^{m}\Big(\sum_{j=1}^{n} \mathsf{L}(a, j)\mathbf{e}_k(j)\Big)^2\Big)^{1/2} = \Big(\sum_{a=1}^{m} \mathsf{L}(a, k)^2\Big)^{1/2} = \|\mathbf{c}(\mathsf{L}, k)\|_2.$$

Thus

$$\|\mathsf{L}\|_{1,2} \geq \max\{\|\mathbf{c}(\mathsf{L}, j)\|_2 \mid j \in \{1, \dots, n\}\},$$

since $\|\mathbf{e}_k\|_1 = 1$.

(iii) Here we compute

$$\begin{aligned}
\|\mathsf{L}\|_{1,\infty} &= \sup\{\|\mathsf{L}(\mathbf{x})\|_\infty \mid \|\mathbf{x}\|_1 = 1\} \\
&= \sup\Big\{\max\Big\{\Big|\sum_{j=1}^{n} \mathsf{L}(a, j)x_j\Big| \;\Big|\; a \in \{1, \dots, m\}\Big\} \;\Big|\; \|\mathbf{x}\|_1 = 1\Big\} \\
&\leq \sup\Big\{\max\Big\{|\mathsf{L}(a, j)| \;\Big|\; j \in \{1, \dots, n\},\ a \in \{1, \dots, m\}\Big\}\Big(\sum_{j=1}^{n} |x_j|\Big) \;\Big|\; \|\mathbf{x}\|_1 = 1\Big\} \\
&= \max\{|\mathsf{L}(a, j)| \mid j \in \{1, \dots, n\},\ a \in \{1, \dots, m\}\}.
\end{aligned}$$

For the converse inequality, let $k \in \{1, \dots, n\}$ be such that

$$\max\{|\mathsf{L}(a, k)| \mid a \in \{1, \dots, m\}\} = \max\{|\mathsf{L}(a, j)| \mid j \in \{1, \dots, n\},\ a \in \{1, \dots, m\}\}.$$

Then

$$\begin{aligned}
\|\mathsf{L}(\mathbf{e}_k)\|_\infty &= \max\Big\{\Big|\sum_{j=1}^{n} \mathsf{L}(a, j)\mathbf{e}_k(j)\Big| \;\Big|\; a \in \{1, \dots, m\}\Big\} \\
&= \max\{|\mathsf{L}(a, k)| \mid a \in \{1, \dots, m\}\}.
\end{aligned}$$

Thus

$$\|\mathsf{L}\|_{1,\infty} \geq \max\{|\mathsf{L}(a, j)| \mid j \in \{1, \dots, n\},\ a \in \{1, \dots, m\}\},$$

since $\|\mathbf{e}_k\|_1 = 1$.

(iv) In this case we maximise the function $\mathbf{x} \mapsto \|\mathsf{L}(\mathbf{x})\|_1$ subject to the constraint that $\|\mathbf{x}\|_2 = 1$, or equivalently, subject to the constraint that $\|\mathbf{x}\|_2^2 = 1$. We shall do this using Theorem 4.4.44 and defining

$$f(\mathbf{x}) = \|\mathsf{L}(\mathbf{x})\|_1, \quad g(\mathbf{x}) = \|\mathbf{x}\|_2^2 - 1.$$

Let us first assume that none of the rows of L are zero. We must exercise some care because f is not differentiable on $\mathbb{R}^n$. Note that

$$\|\mathsf{L}(\mathbf{x})\|_1 = \sum_{a=1}^{m} |\langle \mathbf{r}(\mathsf{L}, a), \mathbf{x}\rangle_{\mathbb{R}^n}|.$$

Thus f is differentiable at points off the set

$$B_{\mathsf{L}} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \text{ there exists } a \in \{1, \dots, m\} \text{ such that } \langle \boldsymbol{r}(\mathsf{L}, a), \boldsymbol{x} \rangle_{\mathbb{R}^n} = 0\}.$$

To facilitate computations, let us define $\boldsymbol{u}_{\mathsf{L}} : \mathbb{R}^n \to \mathbb{R}^m$ by asking that

$$u_{\mathsf{L},a}(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{r}(\mathsf{L}, a), \boldsymbol{x} \rangle_{\mathbb{R}^n}).$$

Note that $B_{\mathsf{L}} = \boldsymbol{u}_{\mathsf{L}}^{-1}(\boldsymbol{0})$. Note that on $\mathbb{R}^n \setminus B_{\mathsf{L}}$ the function $\boldsymbol{u}_{\mathsf{L}}$ is locally constant. That is to say, if $\boldsymbol{x} \in \mathbb{R}^n \setminus B$, then there is a neighbourhood $U \subseteq \mathbb{R}^n \setminus B_{\mathsf{L}}$ of $\boldsymbol{x}$ such that $\boldsymbol{u}_{\mathsf{L}}|U$ is constant (why?). Moreover, it is clear that

$$f(\boldsymbol{x}) = \langle \boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}), \mathsf{L}(\boldsymbol{x}) \rangle_{\mathbb{R}^m}.$$

Now let $\boldsymbol{x}_0 \in \mathbb{R}^n \setminus B_{\mathsf{L}}$ be a maximum of f subject to the constraint that $g(\boldsymbol{x}) = 0$. Note that

$$\boldsymbol{D}g(\boldsymbol{x}) \cdot \boldsymbol{v} = \langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathbb{R}^n} + \langle \boldsymbol{v}, \boldsymbol{x} \rangle_{\mathbb{R}^n} = 2\langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathbb{R}^n},$$

and so, if $\boldsymbol{x} \neq \boldsymbol{0}$, then we can conclude that $\boldsymbol{D}g(\boldsymbol{x})$ has rank 1. Thus, by Theorem 4.4.44, there exists $\lambda \in \mathbb{R}$ such that

$$\boldsymbol{D}(f - \lambda g)(\boldsymbol{x}_0) = \boldsymbol{0}.$$

Since $\boldsymbol{u}_{\mathsf{L}}$ is locally constant,

$$\boldsymbol{D}f(\boldsymbol{x}_0) \cdot \boldsymbol{v} = \langle \boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}_0), \mathsf{L}(\boldsymbol{v}) \rangle_{\mathbb{R}^m}.$$

Moreover, $\boldsymbol{D}g(\boldsymbol{x}) \cdot \boldsymbol{v} = 2\langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathbb{R}^n}$. Thus $\boldsymbol{D}(f - \lambda g)(\boldsymbol{x}_0) = \boldsymbol{0}$ if and only if

$$\mathsf{L}^T(\boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}_0)) = 2\lambda \boldsymbol{x}_0 \quad \Longrightarrow \quad |\lambda| = \frac{1}{2}\|\mathsf{L}^T(\boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}_0))\|_2,$$

since $\|\boldsymbol{x}_0\|_2 = 1$. Thus $\lambda = 0$ if and only if $\mathsf{L}^T(\boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}_0)) = \boldsymbol{0}$. Therefore, if $\lambda = 0$ then

$$f(\boldsymbol{x}_0) = \langle \boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}_0), \mathsf{L}(\boldsymbol{x}_0) \rangle_{\mathbb{R}^m} = \langle \mathsf{L}^T(\boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}_0)), \boldsymbol{x}_0 \rangle_{\mathbb{R}^n} = 0.$$

If $\lambda \neq 0$ then

$$f(\boldsymbol{x}_0) = \langle \mathsf{L}^T(\boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}_0)), \boldsymbol{x}_0 \rangle_{\mathbb{R}^n} = \frac{1}{2\lambda}\|\mathsf{L}^T(\boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}_0))\|_2^2 = \frac{2}{\lambda}\lambda^2 = 2\lambda.$$

Observing that $|\lambda| = \|\mathsf{L}^T(\boldsymbol{u}_{\mathsf{L}}(\boldsymbol{x}_0))\|_2$ and that f is nonnegative-valued, we can conclude that, at solutions of the constrained maximisation problem, we must have

$$f(\boldsymbol{x}_0) = \|\mathsf{L}^T(\boldsymbol{u})\|_2,$$

where $\boldsymbol{u}$ varies over the nonzero points in the image of $\boldsymbol{u}_{\mathsf{L}}$, i.e., over points from $\{-1, 1\}^m$.

This would conclude the proof of this part of the theorem in the case that L has no zero rows, but for the fact that it is possible that f attains its maximum on B_{L}. We now show that this does not happen. Let $\boldsymbol{x}_0 \in B_{\mathsf{L}}$ satisfy $\|\boldsymbol{x}_0\|_2 = 1$ and denote

$$A_0 = \{a \in \{1, \dots, m\} \mid \ u_{\mathsf{L},a}(\boldsymbol{x}_0) = 0\}.$$

Let $A_1 = \{1, \dots, m\} \setminus A_0$. Let $a_0 \in A_0$. For $\epsilon \in \mathbb{R}$ define

$$\boldsymbol{x}_\epsilon = \frac{\boldsymbol{x}_0 + \epsilon \boldsymbol{r}(\mathsf{L}, a_0)}{\sqrt{1 + \epsilon^2 \|\boldsymbol{r}(\mathsf{L}, a_0)\|_2^2}}.$$

Note that

$$\|x_0 + \epsilon r(\mathsf{L}, a_0)\|_2^2 = \|x_0\|_2^2 + \epsilon^2 \|r(\mathsf{L}, a_0)\|_2^2 = 1 + \epsilon^2 \|r(\mathsf{L}, a_0)\|_2^2$$

since $\langle r(\mathsf{L}, a_0), x_0 \rangle_{\mathbb{R}^n} = 0$. Thus x_ϵ satisfies the constraint $\|x_\epsilon\|_2^2 = 1$. Now let $\epsilon_0 \in \mathbb{R}_{>0}$ be sufficiently small that

$$\langle r(\mathsf{L}, a), x_\epsilon \rangle_{\mathbb{R}^n} \neq 0$$

for all $a \in A_1$ and $\epsilon \in [-\epsilon_0, \epsilon_0]$; this is possible since x_ϵ depends continuously on ϵ. Then we compute

$$\begin{aligned} \|\mathsf{L}(x_\epsilon)\|_1 &= \sum_{a=1}^m |\langle r(\mathsf{L}, a), x_\epsilon \rangle_{\mathbb{R}^n}| \\ &= \frac{1}{\sqrt{1 + \epsilon^2 \|r(\mathsf{L}, a_0)\|_2^2}} \sum_{a=1}^m |\langle r(\mathsf{L}, a), x_0 \rangle_{\mathbb{R}^n} + \epsilon \langle r(\mathsf{L}, a), r(\mathsf{L}, a_0) \rangle_{\mathbb{R}^n}|. \end{aligned}$$

Note that, by Taylor Theorem, ***missing stuff***, we can write

$$\frac{1}{\sqrt{1 + \epsilon^2 \|r(\mathsf{L}, a_0)\|_2^2}} = 1 - \epsilon^2 \frac{\|r(\mathsf{L}, a_0)\|_2^2}{2} + O(\epsilon^3),$$

so that, for ϵ sufficiently small,

$$\begin{aligned} \|\mathsf{L}(x_\epsilon)\|_1 &= \sum_{a=1}^m |\langle r(\mathsf{L}, a), x_0 \rangle_{\mathbb{R}^n} + \epsilon \langle r(\mathsf{L}, a), r(\mathsf{L}, a_0) \rangle_{\mathbb{R}^n}| + O(\epsilon^2) \\ &= \sum_{a \in A_0} |\epsilon| |\langle r(\mathsf{L}, a), r(\mathsf{L}, a_0) \rangle_{\mathbb{R}^n}| \\ &\quad + \sum_{a \in A_1} |\langle r(\mathsf{L}, a), x_0 \rangle_{\mathbb{R}^n} + \epsilon \langle r(\mathsf{L}, a), r(\mathsf{L}, a_0) \rangle_{\mathbb{R}^n}| + O(\epsilon^2). \end{aligned} \tag{4.4}$$

Since we are assuming that none of the rows of L are zero,

$$\sum_{a \in A_0} |\epsilon| |\langle r(\mathsf{L}, a), r(\mathsf{L}, a_0) \rangle_{\mathbb{R}^n}| > 0 \tag{4.5}$$

for $\epsilon \in [-\epsilon_0, \epsilon_0]$. Now take $a \in A_1$. If ϵ is sufficiently small we can write

$$|\langle r(\mathsf{L}, a), x_0 \rangle_{\mathbb{R}^n} + \epsilon \langle r(\mathsf{L}, a), r(\mathsf{L}, a_0) \rangle_{\mathbb{R}^n}| = |\langle r(\mathsf{L}, a), x_0 \rangle_{\mathbb{R}^n}| + \epsilon C_a$$

for some $C_a \in \mathbb{R}$. As a result, and using (4.4), we have

$$\|\mathsf{L}(x_\epsilon)\|_1 = \|\mathsf{L}(x_0)\|_1 + \sum_{a \in A_0} (|\epsilon| |\langle r(\mathsf{L}, a), r(\mathsf{L}, a_0) \rangle_{\mathbb{R}^n}| + \epsilon \sum_{a \in A_1} C_a + O(\epsilon^2).$$

It therefore follows, possibly by again choosing ϵ_0 to be sufficiently small, that we have

$$\|\mathsf{L}(x_\epsilon)\|_1 > \|\mathsf{L}(x_0)\|_1$$

either for all $\epsilon \in [-\epsilon_0, 0)$ or for all $\epsilon \in (0, \epsilon_0]$, taking (4.5) into account. Thus if $x_0 \in B_{\mathsf{L}}$ then x_0 is not a local maximum for f subject to the constraint $g^{-1}(0)$.

Finally, suppose that L has some rows that are zero. Let

$$A_0 = \{a \in \{1, \dots, m\} \mid \boldsymbol{r}(\mathsf{L}, a) = \boldsymbol{0}\}$$

and let $A_1 = \{1, \dots, m\} \setminus A_0$. Let $A_1 = \{a_1, \dots, a_k\}$ with $a_1 < \cdots < a_k$, and define $\hat{\mathsf{L}} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^k)$ by

$$\hat{\mathsf{L}}(\boldsymbol{x}) = \sum_{r=1}^{k} \langle \boldsymbol{r}(\mathsf{L}, a_r), \boldsymbol{x} \rangle_{\mathbb{R}^n} \boldsymbol{e}_r,$$

and note that $\|\mathsf{L}(\boldsymbol{x})\|_1 = \|\hat{\mathsf{L}}(\boldsymbol{x})\|_1$ for every $\boldsymbol{x} \in \mathbb{R}^n$. If $\boldsymbol{y} \in \mathbb{R}^m$ define $\hat{\boldsymbol{y}} \in \mathbb{R}^k$ by removing from $\boldsymbol{y}$ the elements corresponding to the zero rows of L:

$$\hat{\boldsymbol{y}} = (y_{a_1}, \dots, y_{a_k}).$$

Then we compute

$$\begin{aligned}
\mathsf{L}^T(\boldsymbol{y}) &= \sum_{j=1}^{n} \langle \boldsymbol{r}(\mathsf{L}^T, j), \boldsymbol{y} \rangle_{\mathbb{R}^n} \boldsymbol{e}_j = \sum_{j=1}^{n} \Big(\sum_{a=1}^{m} \mathsf{L}(a, j) y_a \Big) \boldsymbol{e}_j \\
&= \sum_{j=1}^{n} \Big(\sum_{r=1}^{k} \mathsf{L}(a_r, j) y_{a_r} \Big) \boldsymbol{e}_j = \sum_{j=1}^{n} \langle \boldsymbol{c}(\hat{\mathsf{L}}, r), \hat{\boldsymbol{y}} \rangle_{\mathbb{R}^n} \boldsymbol{e}_j \\
&= \sum_{j=1}^{n} \langle \boldsymbol{t}(\hat{\mathsf{L}}^T, r), \hat{\boldsymbol{y}} \rangle_{\mathbb{R}^n} \boldsymbol{e}_j = \hat{\mathsf{L}}^T(\hat{\boldsymbol{y}}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mathsf{L}\|_{2,1} &= \sup\{\|\mathsf{L}(\boldsymbol{x})\|_1 \mid \|\boldsymbol{x}\|_2 = 1\} \\
&= \sup\{\|\hat{\mathsf{L}}(\boldsymbol{x})\|_1 \mid \|\boldsymbol{x}\|_2 = 1\} = \|\hat{\mathsf{L}}\|_{2,1} \\
&= \max\{\|\hat{\mathsf{L}}^T(\hat{\boldsymbol{u}})\|_2 \mid \hat{\boldsymbol{u}} \in \{-1, 1\}^k\} \\
&= \max\{\|\mathsf{L}^T(\boldsymbol{u})\|_2 \mid \boldsymbol{u} \in \{-1, 1\}^m\},
\end{aligned}$$

and this finally gives the result.

(v) Note that, in this case, we wish to maximise the function $\boldsymbol{x} \mapsto \|\mathsf{L}(\boldsymbol{x})\|_2$ subject to the constraint that $\|\boldsymbol{x}\|_2 = 1$. However, this is equivalent to maximising $\boldsymbol{x} \mapsto \|\mathsf{L}(\boldsymbol{x})\|_2^2$ subject to the constraint that $\|\boldsymbol{x}\|_2^2 = 1$. In this case, the function we are maximising and the function defining the constraint are infinitely differentiable. Therefore, we can use Theorem 4.4.44 below to determine the character of the maxima. Thus we define

$$f(\boldsymbol{x}) = \|\mathsf{L}(\boldsymbol{x})\|_2^2, \quad g(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 - 1.$$

Note that

$$\boldsymbol{D}g(\boldsymbol{x}) \cdot \boldsymbol{v} = \langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathbb{R}^n} + \langle \boldsymbol{v}, \boldsymbol{x} \rangle_{\mathbb{R}^n} = 2\langle \boldsymbol{x}, \boldsymbol{v} \rangle_{\mathbb{R}^n},$$

and so, if $\boldsymbol{x} \neq \boldsymbol{0}$, then we can conclude that $\boldsymbol{D}g(\boldsymbol{x})$ has rank 1. Thus, by Theorem 4.4.44, if a point $\boldsymbol{x}_0 \in \mathbb{R}^n$ solves the constrained maximisation problem, then there exists $\lambda \in \mathbb{R}$ such that

$$\boldsymbol{D}(f - \lambda g)(\boldsymbol{x}_0) = 0.$$

Since

$$f(x) = \langle \mathsf{L}(x), \mathsf{L}(x) \rangle_{\mathbb{R}^n} = \langle \mathsf{L}^T \circ \mathsf{L}(x), x \rangle_{\mathbb{R}^n},$$

we compute

$$\boldsymbol{D}f(x) \cdot v = \langle \mathsf{L}^T \circ \mathsf{L}(x), v \rangle_{\mathbb{R}^n} + \langle \mathsf{L}^T \circ \mathsf{L}(v), x \rangle_{\mathbb{R}^n} = 2\langle \mathsf{L}^T \circ \mathsf{L}(x), v \rangle_{\mathbb{R}^n}.$$

We also have $\boldsymbol{D}g(x) \cdot v = 2\langle x, v \rangle_{\mathbb{R}^n}$. Thus $\boldsymbol{D}(f - \lambda g)(x_0) = 0$ implies that

$$\mathsf{L}^T \circ \mathsf{L}(x_0) = \lambda x_0.$$

Thus it must be the case that λ is an eigenvalue for $\mathsf{L}^T \circ \mathsf{L}$ with eigenvector x_0. Let us record some facts about this eigenvalue/eigenvector combination.

1 Lemma *If $\mathsf{L} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$ then the linear map $\mathsf{L}^T \circ \mathsf{L} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ has the following properties:*

(i) all eigenvalues of $\mathsf{L}^T \circ \mathsf{L}$ are real and nonnegative;

(ii) the exists a basis for $\mathbb{R}^n$, orthonormal with respect to the Euclidean inner product, consisting of eigenvectors of $\mathsf{L}^T \circ \mathsf{L}$.

Proof First of all, note that

$$(\mathsf{L}^T \circ \mathsf{L})^T = \mathsf{L}^T \circ \mathsf{L},$$

and so, by ***missing stuff***, the linear map $\mathsf{L}^T \circ \mathsf{L}$ is symmetric with respect to the Euclidean inner product. Thus the eigenvalues of $\mathsf{L}^T \circ \mathsf{L}$ are real. Also note that

$$\langle \mathsf{L}^T \circ \mathsf{L}(x), x \rangle_{\mathbb{R}^n} = \langle \mathsf{L}(x), \mathsf{L}(x) \rangle_{\mathbb{R}^n} \geq 0$$

by ***missing stuff***, and so the eigenvalues of $\mathsf{L}^T \circ \mathsf{L}$ are nonnegative by ***missing stuff***.

That there is a basis of eigenvectors for $\mathbb{R}^n$, orthonormal with respect to $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$, follows from ***missing stuff***. ▼

Let us proceed with our analysis. The lemma implies that there exist $\lambda_1, \dots, \lambda_n \in \mathbb{R}_{\geq 0}$ and vectors $x_1, \dots, x_n$ such that

$$\lambda_1 \leq \cdots \leq \lambda_n,$$

such that $\mathsf{L}^T \circ \mathsf{L}(x_j) = \lambda_j x_j$, $j \in \{1, \dots, n\}$, and such that a solution to the problem of maximising f with the constraint $g^{-1}(0)$ is obtained by evaluating f at one of the points $x_1, \dots, x_n$. Thus the problem can be solved by evaluating f at this finite collection of points, and determining at which of these f has its largest value. Thus we compute

$$f(x_j) = \|\mathsf{L}(x_j)\|_2^2 = \langle \mathsf{L}(x), \mathsf{L}(x_j) \rangle_{\mathbb{R}^m} = \langle \mathsf{L}^T \circ \mathsf{L}(x_j), x_j \rangle_{\mathbb{R}^n} = \lambda_j \|x_j\|_2^2 = \lambda_j.$$

The maximum value of f subject to the constraint $g^{-1}(0)$ is then attained at x_n and this maximum value is λ_n. Thus the maximum value of the function $x \mapsto \|\mathsf{L}(x)\|_2$ subject to the constraint that $\|x\|_2 = 1$ is $\sqrt{\lambda_n}$, and this gives the desired result.

(vi) First of all, we note that this part of the theorem certainly holds when $\mathsf{L} = 0$. Thus we shall freely assume that L is nonzero when convenient. We maximise the function $x \mapsto \|\mathsf{L}(x)\|_\infty$ subject to the constraint that $\|x\|_2 = 1$, or equivalently subject to the constraint that $\|x\|_2^2 = 1$. We shall use Theorem 4.4.44, defining

$$f(x) = \|\mathsf{L}(x)\|_\infty, \quad g(x) = \|x\|_2^2 - 1.$$

Note that L is not differentiable on $\mathbb{R}^n$, so we first restrict to a subset where f is differentiable. Let us define

$$\begin{aligned} A_{\mathsf{L}} &: \mathbb{R}^n \to \mathbf{2}^{\{1,\dots,m\}} \\ &\quad \boldsymbol{x} \mapsto \{a \in \{1,\dots,m\} \mid \langle \boldsymbol{r}(\mathsf{L},a), \boldsymbol{x}\rangle_{\mathbb{R}^n} = \|\mathsf{L}(\boldsymbol{x})\|_\infty\}. \end{aligned}$$

Then denote

$$B_{\mathsf{L}} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \operatorname{card}(A_{\mathsf{L}}(\boldsymbol{x})) > 1\}.$$

Since

$$\|\mathsf{L}(\boldsymbol{x})\|_\infty = \max\{\langle \boldsymbol{r}(\mathsf{L},1), \boldsymbol{x}\rangle_{\mathbb{R}^n}, \dots, \langle \boldsymbol{r}(\mathsf{L}m), \boldsymbol{x}\rangle_{\mathbb{R}^n}\},$$

we see that f is differentiable at points that are not in the set B_{L}.

Let us first suppose that $\boldsymbol{x}_0 \in \mathbb{R}^n \setminus B_{\mathsf{L}}$ is a maximum of f subject to the constraint that $g(\boldsymbol{x}) = 0$. Then there exists a unique $a_0 \in \{1,\dots,m\}$ such that $f(\boldsymbol{x}_0) = \langle \boldsymbol{r}(\mathsf{L},a_0), \boldsymbol{x}_0\rangle_{\mathbb{R}^n}$. Since we are assuming that L is nonzero, it must be that $\boldsymbol{r}(\mathsf{L},a_0)$ is nonzero. Moreover, there exists a neighbourhood U of $\boldsymbol{x}_0$ such that

$$\operatorname{sign}(\langle \boldsymbol{r}(\mathsf{L},a_0), \boldsymbol{x}\rangle_{\mathbb{R}^n}) = \operatorname{sign}(\langle \boldsymbol{r}(\mathsf{L},a_0), \boldsymbol{x}_0\rangle_{\mathbb{R}^n})$$

and

$$f(\boldsymbol{x}) = \langle \boldsymbol{r}(\mathsf{L},a_0), \boldsymbol{x}\rangle_{\mathbb{R}^n}$$

for each $\boldsymbol{x} \in U$. Abbreviating

$$u_{\mathsf{L},a_0}(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{r}(\mathsf{L},a_0), \boldsymbol{x}\rangle_{\mathbb{R}^n}),$$

we have

$$f(\boldsymbol{x}) = u_{\mathsf{L},j}(\boldsymbol{x}_0)\langle \boldsymbol{r}(\mathsf{L},a_0), \boldsymbol{x}\rangle_{\mathbb{R}^n}$$

for every $\boldsymbol{x} \in U$. Note that, as in the proofs of parts (iv) and (v) above, $\boldsymbol{D}g(\boldsymbol{x})$ has rank 1 for $\boldsymbol{x} \neq 0$. Therefore, by Theorem 4.4.44, there exists $\lambda \in \mathbb{R}$ such that

$$\boldsymbol{D}(f - \lambda g)(\boldsymbol{x}_0) = \mathbf{0}.$$

We compute

$$\boldsymbol{D}(f - \lambda g)(\boldsymbol{x}_0) \cdot \boldsymbol{v} = u_{\mathsf{L},j}(\boldsymbol{x}_0)\langle \boldsymbol{r}(\mathsf{L},a_0), \boldsymbol{v}\rangle_{\mathbb{R}^n} - 2\lambda\langle \boldsymbol{x}_0, \boldsymbol{v}\rangle_{\mathbb{R}^n}$$

for every $\boldsymbol{v} \in \mathbb{R}^n$. Thus we must have

$$2\lambda\boldsymbol{x}_0 = u_{\mathsf{L},a_0}(\boldsymbol{x}_0)\boldsymbol{r}(\mathsf{L},a_0).$$

This implies that $\boldsymbol{x}_0$ and $\boldsymbol{r}(\mathsf{L},a_0)$ are linearly dependent and that

$$|\lambda| = \frac{1}{2}\|\boldsymbol{r}(\mathsf{L},a_0)\|_2$$

since $\|\boldsymbol{x}_0\|_2 = 1$. Therefore,

$$f(\boldsymbol{x}_0) = u_{\mathsf{L},a_0}(\boldsymbol{x}_0)\langle \boldsymbol{r}(\mathsf{L},a_0), \tfrac{1}{2\lambda}u_{\mathsf{L},a_0}(\boldsymbol{x}_0)\boldsymbol{r}(\mathsf{L},a_0)\rangle_{\mathbb{R}^n} = \frac{2}{\lambda}\lambda^2 = 2\lambda.$$

Since $|\lambda| = \frac{1}{2}\|\boldsymbol{r}(\mathsf{L},a_0)\|_2$ it follows that

$$f(\boldsymbol{x}_0) = \|\boldsymbol{r}(\mathsf{L},a_0)\|_2.$$

This completes the proof, but for the fact that maxima of f may occur at points in B_{L}. Thus let $\boldsymbol{x}_0 \in B_{\mathsf{L}}$ be such that $\|\boldsymbol{x}_0\|_2 = 1$. For $a \in A_{\mathsf{L}}(\boldsymbol{x}_0)$ let us write

$$\boldsymbol{r}(\mathsf{L}, a) = \rho_a \boldsymbol{x}_0 + \boldsymbol{y}_a,$$

where $\langle \boldsymbol{x}_0, \boldsymbol{y}_a \rangle_{\mathbb{R}^n} = 0$. Therefore,

$$\langle \boldsymbol{r}(\mathsf{L}, a), \boldsymbol{x}_0 \rangle_{\mathbb{R}^n} = \rho_a.$$

We claim that if there exists $a_0 \in A_{\mathsf{L}}(\boldsymbol{x}_0)$ for which $\boldsymbol{y}_{a_0} \neq \boldsymbol{0}$, then $\boldsymbol{x}_0$ cannot be a maximum of f subject to the constraint $g^{-1}(0)$. Indeed, if $\boldsymbol{y}_{a_0} \neq 0$ then define

$$\boldsymbol{x}_\epsilon = \frac{\boldsymbol{x}_0 + \epsilon \boldsymbol{y}_{a_0}}{\sqrt{1 + \epsilon^2 \|\boldsymbol{y}_{a_0}\|_2^2}}.$$

As in the proof of part (iv) above, one shows that $\|\boldsymbol{x}_\epsilon\|_2 = 1$, and so $\boldsymbol{x}_\epsilon$ satisfies the constraint for every $\epsilon \in \mathbb{R}$. Also as in the proof of part (iv), we have

$$\boldsymbol{x}_\epsilon = \boldsymbol{x}_0 + \epsilon \boldsymbol{y}_0 + O(\epsilon^2).$$

Thus

$$\langle \boldsymbol{r}(\mathsf{L}, a_0), \boldsymbol{x}_\epsilon \rangle_{\mathbb{R}^n} = \rho_a + \epsilon \|\boldsymbol{y}_{a_0}\|_2^2 + O(\epsilon^2)$$

and so, for ϵ sufficiently small,

$$|\langle \boldsymbol{r}(\mathsf{L}, a_0), \boldsymbol{x}_\epsilon \rangle_{\mathbb{R}^n}| = |\langle \boldsymbol{r}(\mathsf{L}, a_0), \boldsymbol{x}_0 \rangle_{\mathbb{R}^n}| + \epsilon C_{a_0} + O(\epsilon^2)$$

where C_{a_0} is nonzero. Therefore, there exists $\epsilon_0 \in \mathbb{R}_{>0}$ such that

$$|\langle \boldsymbol{r}(\mathsf{L}, a_0), \boldsymbol{x}_\epsilon \rangle_{\mathbb{R}^n}| > |\langle \boldsymbol{r}(\mathsf{L}, a_0), \boldsymbol{x}_0 \rangle_{\mathbb{R}^n}|$$

either for all $\epsilon \in [-\epsilon_0, 0)$ or for all $\epsilon \in (0, \epsilon_0]$. In either case, $\boldsymbol{x}_0$ cannot be a maximum for f subject to the constraint $g^{-1}(0)$.

Finally, suppose that $\boldsymbol{x}_0 \in B_{\mathsf{L}}$ is a maximum for f subject to the constraint $g^{-1}(0)$. Then, as we saw in the preceding paragraph, for each $a \in A_{\mathsf{L}}(\boldsymbol{x}_0)$, we must have

$$\boldsymbol{r}(\mathsf{L}, a) = \langle \boldsymbol{r}(\mathsf{L}, a), \boldsymbol{x}_0 \rangle_{\mathbb{R}^n} \boldsymbol{x}_0.$$

It follows that $\|\boldsymbol{r}(\mathsf{L}, a)\|_2^2 = \langle \boldsymbol{r}(\mathsf{L}, a), \boldsymbol{x}_0 \rangle_{\mathbb{R}^n}^2$. Moreover, by definition of $A_{\mathsf{L}}(\boldsymbol{x}_0)$ and since we are supposing that $\boldsymbol{x}_0$ is a maximum for f subject to the constraint $g^{-1}(0)$, we have

$$\begin{aligned} & |\langle \boldsymbol{r}(\mathsf{L}, a), \boldsymbol{x}_0 \rangle_{\mathbb{R}^n}| = \|\mathsf{L}\|_{2,\infty} \\ \implies \quad & \langle \boldsymbol{r}(\mathsf{L}, a), \boldsymbol{x}_0 \rangle_{\mathbb{R}^n}^2 = \|\mathsf{L}\|_{2,\infty}^2 \\ \implies \quad & \|\boldsymbol{r}(\mathsf{L}, a)\|_2 = \|\mathsf{L}\|_{2,\infty}. \end{aligned} \tag{4.6}$$

Now, if $a \in \{1, \dots, m\}$, we claim that

$$\|\boldsymbol{r}(\mathsf{L}, a)\|_2 \leq \|\mathsf{L}\|_{2,\infty}. \tag{4.7}$$

Indeed suppose that $a \in \{1, \dots, m\}$ satisfies

$$\|\boldsymbol{r}(\mathsf{L}, a)\|_2 > \|\mathsf{L}\|_{2,\infty}.$$

Define $\boldsymbol{x} = \frac{\boldsymbol{r}(\mathsf{L},a)}{\|\boldsymbol{r}(\mathsf{L},a)\|_2}$ so that $\boldsymbol{x}$ satisfies the constraint $g(\boldsymbol{x}) = 0$. Moreover,

$$f(\boldsymbol{x}) \geq \langle \boldsymbol{r}(\mathsf{L},a), \boldsymbol{x} \rangle_{\mathbb{R}^n} = \|\boldsymbol{r}(\mathsf{L},a)\|_2 > \|\mathsf{L}\|_{2,\infty},$$

contradicting the assumption that $\boldsymbol{x}_0$ is a maximum for f. Thus, given that (4.6) holds for every $a \in A_{\mathsf{L}}(\boldsymbol{x}_0)$ and (4.7) holds for every $a \in \{1,\ldots,m\}$, we have

$$\|\mathsf{L}\|_{2,\infty} = \max\{\|\boldsymbol{r}(\mathsf{L},a)\|_2 \mid a \in \{1,\ldots,m\}\},$$

as desired.

For the last three parts of the theorem, the following result is useful.

2 Lemma *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and let $|||\cdot|||_\infty$ be the norm induced on $\mathrm{L}(\mathbb{R}^n;\mathbb{R}^m)$ by the norm $\|\cdot\|_\infty$ on $\mathbb{R}^n$ and the norm $\|\cdot\|$ on $\mathbb{R}^m$. Then*

$$|||\mathsf{L}|||_\infty = \max\{\|\mathsf{L}(\mathbf{u})\| \mid \mathbf{u} \in \{-1,1\}^n\}.$$

Proof Note that the set

$$\{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_\infty \leq 1\}$$

is a convex polytope. Therefore, by (**??**) from the proof of Theorem **??**, this set is the convex hull of $\{-1,1\}^n$. Thus, if $\|\boldsymbol{x}\|_\infty = 1$ we can write

$$\boldsymbol{x} = \sum_{\boldsymbol{u} \in \{-1,1\}^n} \lambda_{\boldsymbol{u}} \boldsymbol{u}$$

where $\lambda_{\boldsymbol{u}} \in [0,1]$ for each $\boldsymbol{u} \in \{-1,1\}^n$ and

$$\sum_{\boldsymbol{u} \in \{-1,1\}^n} \lambda_{\boldsymbol{u}} = 1.$$

Therefore,

$$\begin{aligned}
\|\mathsf{L}(\boldsymbol{x})\| &= \Big\| \sum_{\boldsymbol{u} \in \{-1,1\}^n} \lambda_{\boldsymbol{u}} \mathsf{L}(\boldsymbol{u}) \Big\| \leq \sum_{\boldsymbol{u} \in \{-1,1\}^n} \lambda_{\boldsymbol{u}} \|\mathsf{L}(\boldsymbol{u})\| \\
&\leq \Big(\sum_{\boldsymbol{u} \in \{-1,1\}^n} \lambda_{\boldsymbol{u}} \Big) \max\{\|\mathsf{L}(\boldsymbol{u})\| \mid \boldsymbol{u} \in \{-1,1\}^n\} \\
&= \max\{\|\mathsf{L}(\boldsymbol{u})\| \mid \boldsymbol{u} \in \{-1,1\}^n\}.
\end{aligned}$$

Therefore,

$$\sup\{\|\mathsf{L}(\boldsymbol{x})\| \mid \|\boldsymbol{x}\|_\infty = 1\} \leq \max\{\|\mathsf{L}(\boldsymbol{u})\| \mid \boldsymbol{u} \in \{-1,1\}^n\} \leq \sup\{\|\mathsf{L}(\boldsymbol{x})\| \mid \|\boldsymbol{x}\|_\infty = 1\},$$

the last inequality holding since if $\boldsymbol{u} \in \{-1,1\}^n$ then $\|\boldsymbol{u}\|_\infty = 1$. The result follows since the previous inequalities must be equalities. ▼

(vii) This follows immediately from the preceding lemma.
(viii) This too follows immediately from the preceding lemma.
(ix) Note that for $\boldsymbol{u} \in \{-1,1\}^n$ we have

$$|\langle \boldsymbol{r}(\mathsf{L},a), \boldsymbol{u} \rangle_{\mathbb{R}^n}| = \Big| \sum_{j=1}^n \mathsf{L}(a,j) u_j \Big| \leq \sum_{j=1}^n |\mathsf{L}(a,j)| = \|\boldsymbol{r}(\mathsf{L},a)\|_1.$$

Therefore, using the previous lemma,

$$\begin{aligned}\|\mathsf{L}\|_{\infty,\infty} &= \max\{\|\mathsf{L}(\boldsymbol{u})\|_\infty \mid \boldsymbol{u} \in \{-1,1\}^n\}\\ &= \max\{\max\{|\langle \boldsymbol{r}(\mathsf{L},a), \boldsymbol{u}\rangle_{\mathbb{R}^n}| \mid a \in \{1,\dots,m\}\} \mid \boldsymbol{u} \in \{-1,1\}^n\}\\ &\leq \max\{\|\boldsymbol{r}(\mathsf{L},a)\|_1 \mid a \in \{1,\dots,m\}\}.\end{aligned}$$

To establish the other inequality, for $a \in \{1,\dots,m\}$ define $\boldsymbol{u}_a \in \{-1,1\}^n$ by

$$u_{a,j} = \begin{cases} 1, & \mathsf{L}(a,j) \geq 0,\\ -1, & \mathsf{L}(a,j) < 0 \end{cases}$$

and note that a direct computation gives the ath component of $\mathsf{L}(\boldsymbol{u}_a)$ as $\|\boldsymbol{r}(\mathsf{L},a)\|_1$. Therefore,

$$\begin{aligned}\max\{\|\boldsymbol{r}(\mathsf{L},a)\|_1 \mid a \in \{1,\dots,m\}\} &= \max\{|\mathsf{L}(\boldsymbol{u}_a)_a| \mid a \in \{1,\dots,m\}\}\\ &\leq \max\{\|\mathsf{L}(\boldsymbol{u}_a)\|_\infty \mid a \in \{1,\dots,m\}\}\\ &\leq \max\{\|\mathsf{L}(\boldsymbol{u})\|_\infty \mid \boldsymbol{u} \in \{-1,1\}^n\} = \|\mathsf{L}\|_{\infty,\infty},\end{aligned}$$

giving this part of the theorem. ■

Having characterised the nine possible norms on $\mathrm{L}(\mathbb{R}^n;\mathbb{R}^m)$ corresponding to the norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$, we shall always use the norm $\|\cdot\|_{2,2}$, unless explicitly stated to the contrary. And, as we do for the 2-norm for $\mathbb{R}^n$, we will adopt particular notation for the $(2,2)$-norm on $\mathrm{L}(\mathbb{R}^n;\mathbb{R}^m)$, denoting it by $\|\cdot\|_{\mathbb{R}^n,\mathbb{R}^m}$.

4.1.5 The Frobenius norm

Next let us consider a different norm for the set of linear maps. First of all, note that there is an identification of $\mathrm{L}(\mathbb{R}^n;\mathbb{R}^m)$ with $\mathbb{R}^{mn}$. Indeed, there are many such identifications; for example, one could assemble the m rows of $\boldsymbol{A}$, each consisting of n numbers, consecutively to get a vector of length mn. On $\mathbb{R}^{mn}$ one has the Euclidean norm $\|\cdot\|_{\mathbb{R}^{mn}}$, and this then defines a norm on $\mathrm{L}(\mathbb{R}^n;\mathbb{R}^m)$ using whatever identification one chooses. Moreover, since the Euclidean norm is "unbiased" in terms of the ordering of the indices (i.e., the Euclidean norm of a vector is independent on the order of its components), this norm on $\mathrm{L}(\mathbb{R}^n;\mathbb{R}^m)$ will be independent of how one chooses to assemble the components of a matrix into a vector of length mn. Thus, waiting for the dust to settle, we have the following definition.

4.1.15 Definition (Frobenius[2] norm) The *Frobenius norm* of $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathbb{R})$ is

$$\|\boldsymbol{A}\|_{\mathrm{Fr}} = (\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A}))^{1/2}$$ •

Note that, using the definition of transpose, of matrix multiplication, and of trace we have following formula for the Frobenius norm:

$$\|\boldsymbol{A}\|_{\mathrm{Fr}} = \Big(\sum_{a=1}^m \sum_{j=1}^n \boldsymbol{A}(a,j)^2\Big)^{1/2}.$$

[2]Ferdinand Georg Frobenius (1849–1917) was a German mathematician whose primary contributions were to the fields of group theory, operator theory, differential geometry, and other.

Thus the Frobenius norm is indeed just the square root of the sum of the squares of the components of $\boldsymbol{A}$, just as suggested before the definition.

Let us give some properties of the Frobenius norm, including the assertion that it is indeed a norm.

4.1.16 Proposition (Properties of the Frobenius norm) *If* $\mathbf{A}, \mathbf{A}_1, \mathbf{A}_2 \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$, *if* $\mathbf{B} \in \mathrm{L}(\mathbb{R}^k; \mathbb{R}^n)$, *if* $a \in \mathbb{R}$, *and if* $\mathbf{x} \in \mathbb{R}^n$ *then the following statements hold:*

(i) $\|a\mathbf{A}\|_{\mathrm{Fr}} = |a|\|\mathbf{A}\|_{\mathrm{Fr}}$;

(ii) $\|\mathbf{A}\|_{\mathrm{Fr}} \geq 0$;

(iii) $\|\mathbf{A}\|_{\mathrm{Fr}} = 0$ *only if* $\mathbf{A} = \mathbf{0}_{m\times n}$;

(iv) $\|\mathbf{A}_1 + \mathbf{A}_2\|_{\mathrm{Fr}} \leq \|\mathbf{A}_1\|_{\mathrm{Fr}} + \|\mathbf{A}_2\|_{\mathrm{Fr}}$;

(v) $\|\mathbf{A}\mathbf{x}\|_{\mathbb{R}^m} \leq \|\mathbf{A}\|_{\mathrm{Fr}}\|\mathbf{x}\|_{\mathbb{R}^n}$;

(vi) $\|\mathbf{A}\mathbf{B}\|_{\mathrm{Fr}} \leq \|\mathbf{A}\|_{\mathrm{Fr}}\|\mathbf{B}\|_{\mathrm{Fr}}$.

Proof The first four properties of the Frobenius norm follow from the corresponding properties for the Euclidean norm on $\mathbb{R}^{mn}$. Thus we prove only the last two.

For the fifth property we adopt the notation of Proposition 4.3.16 and compute

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{x}\|_{\mathbb{R}^m} &= \Big(\sum_{a=1}^{m} \langle \boldsymbol{r}(\boldsymbol{A}, a), \boldsymbol{x}\rangle_{\mathbb{R}^n}^2\Big)^{1/2} \leq \Big(\sum_{a=1}^{m} \|\boldsymbol{r}(\boldsymbol{A}, a)\|_{\mathbb{R}^n}^2 \|\boldsymbol{x}\|_{\mathbb{R}^n}^2\Big)^{1/2} \\ &= \Big(\sum_{a=1}^{m} \|\boldsymbol{r}(\boldsymbol{A}, a)\|_{\mathbb{R}^n}^2\Big)^{1/2} \|\boldsymbol{x}\|_{\mathbb{R}^n}.\end{aligned}$$

The result follows after we notice, and verify via a direct computation, that

$$\|\boldsymbol{A}\|_{\mathrm{Fr}} = \Big(\sum_{a=1}^{m} \|\boldsymbol{r}(\boldsymbol{A}, a)\|_{\mathbb{R}^n}^2\Big)^{1/2}.$$

For the final assertion we first note that

$$\|\boldsymbol{A}\|_{\mathrm{Fr}} = \Big(\sum_{j=1}^{n} \|\boldsymbol{c}(\boldsymbol{A}, j)\|_{\mathbb{R}^m}^2\Big)^{1/2},$$

where, as in Definition **??**, $\boldsymbol{c}(\boldsymbol{A}, j)$ is the jth column of $\boldsymbol{A}$. Also note that the sth column of $\boldsymbol{AB}$ is given by $\boldsymbol{A}\boldsymbol{c}(\boldsymbol{B}, s)$. Thus we compute

$$\begin{aligned}\|\boldsymbol{AB}\|_{\mathrm{Fr}} &= \Big(\sum_{s=1}^{k} \|\boldsymbol{c}(\boldsymbol{AB}, s)\|_{\mathbb{R}^m}^2\Big)^{1/2} = \Big(\sum_{s=1}^{k} \|\boldsymbol{A}\boldsymbol{c}(\boldsymbol{B}, s)\|_{\mathbb{R}^n}^2\Big)^{1/2} \\ &\leq \Big(\sum_{s=1}^{k} \|\boldsymbol{A}\|_{\mathrm{Fr}}^2 \|\boldsymbol{c}(\boldsymbol{B}, s)\|_{\mathbb{R}^n}^2\Big)^{1/2} \leq \|\boldsymbol{A}\|_{\mathrm{Fr}} \Big(\sum_{s=1}^{k} \|\boldsymbol{c}(\boldsymbol{B}, s)\|_{\mathbb{R}^n}^2\Big)^{1/2} \\ &= \|\boldsymbol{A}\|_{\mathrm{Fr}}\|\boldsymbol{B}\|_{\mathrm{Fr}},\end{aligned}$$

as desired, and where we have used the result from the previous part. ■

It is natural to ask whether the Frobenius norm is the induced norm for some pair of norms, one on $\mathbb{R}^n$ and one on $\mathbb{R}^m$.

4.1.17 Proposition (The Frobenius norm is not often induced) *If* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, then the Frobenius norm on* $\mathrm{L}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}})$ *is the induced norm for any pair of norms, one on* $\mathbb{R}^{\mathrm{n}}$ *and the other on* $\mathbb{R}^{\mathrm{m}}$*, if and only if* m *or* n *are equal to* 1.

Proof If $\|\cdot\|$ is a norm on $\mathbb{R}^n$, then let us define a norm $\|\cdot\|^*$ on $\mathbb{R}^n$ by

$$\|x\|^* = \sup\{|\langle x, v\rangle_{\mathbb{R}^n}| \mid \|v\| = 1\}.$$

It is easy to verify $\|\cdot\|^*$ is indeed a norm. Moreover, it is easy to verify that $\|\cdot\|^{**} = \|\cdot\|$.

Let us give a few lemmata that we will use in the proof. For the following lemma, if $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ then yx^T denotes the linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ defined by

$$yx^T(\xi) = \langle x, \xi\rangle_{\mathbb{R}^n} y.$$

It is evident that $\operatorname{rank}(yx^T) = 1$.

1 Lemma *Let* $\|\cdot\|_\alpha$ *and* $\|\cdot\|_\beta$ *be norms on* $\mathbb{R}^{\mathrm{n}}$ *and* $\mathbb{R}^{\mathrm{m}}$*, respectively, and let* $\|\cdot\|_{\alpha,\beta}$ *be the induced norm on* $\mathrm{L}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}})$*. Then*

$$\|\mathbf{y}\mathbf{x}^{\mathrm{T}}\|_{\alpha,\beta} = \|\mathbf{x}\|_\alpha^* \|\mathbf{y}\|_\beta$$

for every $\mathbf{x} \in \mathbb{R}^{\mathrm{n}}$ *and* $\mathbf{y} \in \mathbb{R}^{\mathrm{m}}$*.*

Proof We compute

$$\|yx^T\|_{\alpha,\beta} = \sup\{\|yx^T(v)\|_\beta \mid \|v\|_\alpha = 1\} = \sup\{|\langle x, v\rangle_{\mathbb{R}^n}| \|y\|_\beta \mid \|v\|_\alpha\} = \|x\|_\alpha^* \|y\|_\beta$$ ▼

For the following lemma, we refer ahead to Definition 4.3.19 for the notion of an orthogonal matrix, or equivalently linear map. We also recall from ***missing stuff*** the notion of singular values for a linear map between inner product spaces.

2 Lemma *If* $\|\cdot\|$ *is a norm on* $\mathrm{L}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}})$ *such that*

$$\|\mathsf{U} \circ \mathsf{L}\mathsf{V}\| = \|\mathsf{L}\|$$

for every $\mathsf{U} \in \mathrm{O}(\mathrm{m})$ *and every* $\mathsf{V} \in \mathrm{O}(\mathrm{n})$*, then there exists* $\mathrm{c} \in \mathbb{R}_{>0}$ *such that, if* $\mathsf{L} \in \mathrm{L}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}})$ *has rank* 1*, it holds that* $\|\mathsf{L}\| = \mathrm{c}\sigma_{\max}(\mathsf{L})$*.*

Proof Let us denote by $\mathsf{L}_{11} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$ the linear map defined by

$$\mathsf{L}(x_1, \ldots, x_n) = (x_1, 0, \ldots, 0).$$

As we show in ***missing stuff***, if L has rank 1, then there exists $\mathsf{U} \in \mathrm{O}(m)$ and $\mathsf{V} \in \mathrm{O}(n)$ such that $\mathsf{L} = \sigma_{\max}(\mathsf{L})\mathsf{U} \circ \mathsf{L}_{11} \circ \mathsf{V}$. It therefore follows that if L has rank 1 then $\|\mathsf{L}\| = \sigma_{\max}\|\mathsf{L}_{11}\|$, giving the result by taking $c = \|\mathsf{L}_{11}\|$. ▼

Now the following lemma is key.

3 Lemma *Let* $\|\cdot\|$ *be a norm on* $\mathrm{L}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}})$ *satisfy*

$$\|\mathsf{U} \circ \mathsf{L}\mathsf{V}\| = \|\mathsf{L}\|$$

for every $\mathsf{U} \in \mathrm{O}(\mathrm{m})$ *and every* $\mathsf{V} \in \mathrm{O}(\mathrm{n})$*. Then the following statements are equivalent:*

(i) there exist norms $\|\cdot\|_\alpha$ *on* $\mathbb{R}^{\mathrm{n}}$ *and* $\|\cdot\|_\beta$ *on* $\mathbb{R}^{\mathrm{m}}$ *such that* $\|\cdot\|$ *is the corresponding induced norm;*

(ii) there exists $c \in \mathbb{R}_{>0}$ *such that* $\|\mathsf{L}\| = c\sigma_{\max}(\mathsf{L})$ *for every* $\mathsf{L} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$.

Proof From Theorem 4.1.14(v), the norm on $\mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$ induced by the norm $\|\cdot\|_2$ on $\mathbb{R}^n$ and $c\|\cdot\|_2$ satisfies $\|\mathsf{L}\| = c\sigma_{\max}(\mathsf{L})$ for every $\mathsf{L} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$. Moreover, since

$$\sigma_{\max}(\mathsf{U} \circ \mathsf{L} \circ \mathsf{V}) = \sigma_{\max}(\mathsf{L})$$

for every $\mathsf{U} \in \mathrm{O}(m)$ and $\mathsf{V} \in \mathrm{O}(n)$, we arrive at the implication (ii) $\implies$ (i).

For the converse implication, suppose that $\|\cdot\|$ is induced by $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. By Lemma 2 there exists $c \in \mathbb{R}_{>0}$ such that $\|\mathsf{L}\| = c\sigma_{\max}(\mathsf{L})$ for every $\mathsf{L} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$ having rank 1. From Lemma 1 and ***missing stuff*** we also have

$$c\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2 = c\sigma_{\max}(\boldsymbol{y}\boldsymbol{x}^T) = \|\boldsymbol{x}\|_\alpha^*\|\boldsymbol{y}\|_\beta$$

for every $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$. By fixing $\boldsymbol{y} \in \mathbb{R}^m$ we see that there exists $c_1 \in \mathbb{R}_{>0}$ such that $\|\boldsymbol{x}\|_\alpha^* = c_1\|\boldsymbol{x}\|_2$ for every $\boldsymbol{x} \in \mathbb{R}^n$. Similarly, by fixing $\boldsymbol{x}$ there exists $c_2 \in \mathbb{R}_{>0}$ such that $\|\boldsymbol{y}\|_\beta = c_2\|\boldsymbol{y}\|_2$ for every $\boldsymbol{y} \in \mathbb{R}^m$. Since $\|\cdot\|_\alpha^{**} = \|\cdot\|\alpha$ and since $\|\cdot\|_2^* = \|\cdot\|_2$ (verify this), we conclude that $\|\cdot\|_\alpha = c_2\|\cdot\|_2$. From Theorem 4.1.14(v) we conclude that $\|\mathsf{L}\| = \frac{c_2}{c_1}\sigma_{\max}(\mathsf{L})$, giving the lemma. ▼

Now we prove the proposition. First of all, note that if $n = 1$ or if $m = 1$, then $\|\cdot\|_{\mathrm{Fr}} = \|\cdot\|_{2,2}$ by Theorem 4.1.14. Conversely, suppose that neither n nor m is equal to 1. For $a \in \mathbb{R}_{>0}$ define $\mathsf{L}_a \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$ by

$$\mathsf{L}_a(x_1, x_2, x_3, \ldots, x_n) = (x_1, ax_2, 0, \ldots, 0).$$

Note that $\sigma_{\max}(\mathsf{L}_a) = \max\{1, a\}$. However, $\|\mathsf{L}_a\|_{\mathrm{Fr}} = \sqrt{1 + a^2}$. Thus we cannot have $\|(\|_{\mathrm{Fr}}\mathsf{L}_a) = c\sigma_{\max}(\mathsf{L}_a)$ for every $a \in \mathbb{R}_{>0}$. By Lemma 3 the theorem follows. ■

4.1.6 Notes

Some parts of the proof we give of Theorem 4.1.14 are new, although much of the result is classically known; see [Horn and Johnson 1990]. The proof of part (iv) of Theorem 4.1.14 comes from [Drakakis and Pearlmutter 2009]. The proof of part (vii) of Theorem 4.1.14 comes from [Rohn 2000]. Note that there is a somewhat different character in certain of the induced norm computations in Theorem 4.1.14. In particular, the induced norms $\|\cdot\|_{2,1}$, $\|\cdot\|_{\infty,1}$, and $\|\cdot\|_{\infty,2}$ involve a search over the 2^m points in $\{-1, 1\}^m$ (in the first two cases) or the 2^n points in $\{-1, 1\}^n$ in the third case. The computations of these norms is correspondingly more involved in terms of the numbers of computations that must be performed. This is discussed by Rohn [2000] for the norm $\|\cdot\|_{\infty,1}$.

The proof we give of Proposition 4.1.17 follows [Chellaboina and Haddad 1995].

Exercises

4.1.1 Show that $S^\perp$ is a subspace of $\mathbb{R}^n$ for every nonempty subset $S \subseteq \mathbb{R}^n$.

4.1.2 Let $r_1, r_2 \in \mathbb{R}_{>0}$ satisfy $r_2 \leq r_1$ and let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$. Show that if $\overline{\mathrm{B}}(r_1, \boldsymbol{x}_1) \cap \overline{\mathrm{B}}(r_2, \boldsymbol{x}_2) \neq \emptyset$ then $\overline{\mathrm{B}}(r_2, \boldsymbol{x}_2) \subseteq \overline{\mathrm{B}}(3r_1, \boldsymbol{x}_1)$. Show that you understand your proof by drawing a picture.

4.1.3 Show that for each $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$, $\big|\|\boldsymbol{x}_1\|_{\mathbb{R}^n} - \|\boldsymbol{x}_2\|_{\mathbb{R}^n}\big| \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\mathbb{R}^n}$.

Section 4.2

The structure of $\mathbb{R}^n$

In this section we summarise the topological (see Chapter **??**) properties of $\mathbb{R}^n$. Many of the properties here are discussed in a more general context in Chapter **??**. Therefore, we limit ourselves here to those features of $\mathbb{R}^n$ that we will make use of without needing the abstract development of Chapter **??**. For example, some of what we do here will be used in Chapter **??**. Because some of what we say here bears a strong resemblance to some of the results of Chapter 2, and because we shall generalise much of this structure in Chapter **??**, we shall omit some of the proofs that resemble their counterparts of Chapters 2 and **??**.

Do I need to read this section? Much of what we say in this section follows in the same vein as does much of Chapter 2. Therefore, perhaps a reader can overlook some of the details of what we say here until specific parts of it are needed. •

4.2.1 Sequences in $\mathbb{R}^n$

Note that for $\mathbb{R}$ the discussion of sequences and their convergence is reliant on the absolute value function. Since this can be generalised to $\mathbb{R}^n$, the ideas of Cauchy sequences and convergent sequences carries over to $\mathbb{R}^n$. Let us give the definitions in this case.

4.2.1 Definition (Cauchy sequence, convergent sequence, bounded sequence) Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}^n$. The sequence:

(i) is a ***Cauchy sequence*** if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $\|x_j - x_k\|_{\mathbb{R}^n} < \epsilon$ for $j, k \geq N$;

(ii) ***converges to*** $\mathbf{x_0}$ if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $\|x_j - x_0\|_{\mathbb{R}^n} < \epsilon$ for $j \geq N$;

(iii) ***diverges*** if it does not converge to any element in $\mathbb{R}^n$;

(iv) is ***bounded*** if there exists $M \in \mathbb{R}_{>0}$ such that $\|x_j\|_{\mathbb{R}^n} < M$ for each $j \in \mathbb{Z}_{>0}$;

(v) is ***constant*** if $x_j = x_1$ for every $j \in \mathbb{Z}_{>0}$;

(vi) is ***eventually constant*** if there exists $N \in \mathbb{Z}_{>0}$ such that $x_j = x_N$ for every $j \geq N$. •

One can show, just as for sequences of real numbers, that convergent sequences are Cauchy and that Cauchy sequences are bounded. Let us state these results here.

4.2.2 Proposition (Convergent sequences are Cauchy) *If a sequence* $(\mathbf{x}_j)_{j\in\mathbb{Z}_{>0}}$ *converges to* $\mathbf{x}_0$ *then it is Cauchy.*

Proof Let $\epsilon \in \mathbb{R}_{>0}$ and let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $\|x_j - x_0\|_{\mathbb{R}^n} < \frac{\epsilon}{2}$ for $j \geq N$. Then, for $j, k \geq N$ we have

$$\|x_j - x_k\|_{\mathbb{R}^n} \leq \|x_j - x_0\|_{\mathbb{R}^n} + \|x_0 - x_k\|_{\mathbb{R}^n} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

as desired. ∎

4.2.3 Proposition (Cauchy sequences are bounded) *If $(\mathbf{x}_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequences then it is bounded.*

Proof Let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $\|\boldsymbol{x}_j - \boldsymbol{x}_k\|_{\mathbb{R}^n} < 1$ for $j,k \geq N$. Let $M_N = \max\{\|\boldsymbol{x}_1\|_{\mathbb{R}^n}, \ldots, \|\boldsymbol{x}_N\|_{\mathbb{R}^n}\}$. For $j \geq N$ we have

$$|\boldsymbol{x}_j| \leq \|\boldsymbol{x}_j - \boldsymbol{x}_N\|_{\mathbb{R}^n} + \|\boldsymbol{x}_N\|_{\mathbb{R}^n} < 1 + M_N,$$

showing that $\|\boldsymbol{x}_j\|_{\mathbb{R}^n} < 1 + M_N$ for each $j \in \mathbb{Z}_{>0}$. ∎

The following result indicates that, to show the convergence of a sequence in $\mathbb{R}^n$, it suffices to show the convergence of the sequence of components.

4.2.4 Proposition (Convergence of a sequence in $\mathbb{R}^n$ equals convergence of each of the components) *Let $(\mathbf{x}_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}^n$ and denote $\mathbf{x}_j = (x_j^1, \ldots, x_j^n)$, $j \in \mathbb{Z}_{>0}$. Then the sequence $(\mathbf{x}_j)_{j\in\mathbb{Z}_{>0}}$ converges to $\mathbf{x}_0 = (x_0^1, \ldots, x_0^n)$ if and only if each of the sequences $(x_j^l)_{j\in\mathbb{Z}_{>0}}$, $l \in \{1, \ldots, n\}$, converges to x_0^l.*

Proof Suppose that $(\boldsymbol{x}_j)_{j\in\mathbb{Z}_{>0}}$ converges to $\boldsymbol{x}_0$. For $\epsilon \in \mathbb{R}_{>0}$ let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $\|\boldsymbol{x}_j - \boldsymbol{x}_0\|_{\mathbb{R}^n} < \epsilon$ for $j \geq N$. Then

$$\left|x_j^l - x_0^l\right| \leq \Big(\sum_{m=1}^n (x_j^m - x_0^m)^2\Big)^{1/2} = \|\boldsymbol{x}_j - \boldsymbol{x}_0\|_{\mathbb{R}^n} < \epsilon,$$

showing that $(x_j^l)_{j\in\mathbb{Z}_{>0}}$ converges to x_0^l.

Now suppose that $(x_j^l)_{j\in\mathbb{Z}_{>0}}$ converges to x_0^l for $l \in \{1, \ldots, n\}$. Let $\epsilon \in \mathbb{R}_{>0}$ and let N be sufficiently large that $\left|x_j^m - x^m\right| < \frac{\epsilon}{\sqrt{n}}$ for $j \geq N$ and for $m \in \{1, \ldots, n\}$. Then

$$\|\boldsymbol{x}_j - \boldsymbol{x}_0\|_{\mathbb{R}^n} = \Big(\sum_{m=1}^n (x_j^m - x^m)^2\Big)^{1/2} < \Big(\sum_{m=1}^n \frac{\epsilon^2}{n}\Big)^{1/2} = \epsilon,$$

as desired. ∎

Thus the convergence tests for sequences in Section 2.3.3 can be used to prove convergence of sequences in $\mathbb{R}^n$ by applying them componentwise.

It is also true that Cauchy sequences converge in $\mathbb{R}^n$. As we see in the proof of the following result, this is reliant on the completeness of $\mathbb{R}$. This notion of completeness is explored in detail in more generality in Section **??**.

4.2.5 Theorem (Cauchy sequences in $\mathbb{R}^n$ converge) *If $(\mathbf{x}_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}^n$ then it converges.*

Proof Let $(\boldsymbol{x}_j)_{j\in\mathbb{Z}_{>0}}$ be a Cauchy sequence in $\mathbb{R}^n$; we write $\boldsymbol{x}_j = (x_j^1, \ldots, x_j^n)$, $j \in \mathbb{Z}_{>0}$. We claim that $(x_j^l)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ for $l \in \{1, \ldots, n\}$. Indeed, for $\epsilon \in \mathbb{R}_{>0}$ let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $\|\boldsymbol{x}_j - \boldsymbol{x}_k\|_{\mathbb{R}^n} < \epsilon$ for $j,k \geq N$. Then

$$\left|x_j^l - x_k^l\right| \leq \Big(\sum_{m=1}^n (x_j^m - x_k^m)^2\Big)^{1/2} = \|\boldsymbol{x}_j - \boldsymbol{x}_k\|_{\mathbb{R}^n} < \epsilon$$

for all $l \in \{1, \dots, n\}$ and $j, k \geq N$. By Theorem 2.3.5 there exists $x^l \in \mathbb{R}$ to which the sequence $(x^l_j)_{j \in \mathbb{Z}_{>0}}$ converges. By Proposition 4.2.4 it follows that $(\boldsymbol{x}_j)_{j \in \mathbb{Z}_{>0}}$ converges to $(x^1, \dots, x^l)$. ■

It is also possible to discuss convergence of multiple sequences in $\mathbb{R}^n$. The definitions and results are just like those in Section 2.3.5 for multiple sequences in $\mathbb{R}$. Multiple sequences are also discussed in Section **??** in a more general context. The reader who wants to use multiple sequences in $\mathbb{R}^n$, and is somehow unable to extrapolate from the results of Section 2.3.5 will find the appropriate definitions in this more general setting.

It is useful to know the relationship between limits and algebraic operations.

4.2.6 Proposition (Algebraic operations on sequences) *Let $(\mathbf{x}_j)_{j \in \mathbb{Z}}$ and $(\mathbf{y}_j)_{j \in \mathbb{Z}_{>0}}$ be sequences in $\mathbb{R}^n$ converging to $\mathbf{x}_0$ and $\mathbf{y}_0$, respectively, let $(a_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}$ converging to a_0, and let $a \in \mathbb{R}$. Then the following statements hold:*

(i) the sequence $(a\mathbf{x}_j)_{j \in \mathbb{Z}_{>0}}$ converges to $a\mathbf{x}_0$;

(ii) the sequence $(\mathbf{x}_j + \mathbf{y}_j)_{j \in \mathbb{Z}_{>0}}$ converges to $\mathbf{x}_0 + \mathbf{y}_0$;

(iii) the sequence $(a_j \mathbf{x}_j)_{j \in \mathbb{Z}_{>0}}$ converges to $a_0 \mathbf{x}_0$.

Proof This proof will be given in a more general context, but with essentially identical notation, for Proposition **??**. The proof is also quite similar to the proof for Proposition 2.3.23. Thus we forgo giving the details here. ■

4.2.2 Series in $\mathbb{R}^n$

The extension of series of real numbers to series in $\mathbb{R}^n$ is fairly easily achieved. One begins by considering a ***series*** in $\mathbb{R}^n$ to bean expression of the form

$$\sum_{j=1}^{\infty} \boldsymbol{x}_j,$$

where $\boldsymbol{x}_j \in \mathbb{R}^n$, $j \in \mathbb{Z}_{>0}$. As we discussed at the beginning of Section 2.4.1, one needs to interpret this expression carefully as it is meaningless as a sum until one says something about its convergence. However, as a formal expression involving the elements of the sequence $(\boldsymbol{x}_j)_{j \in \mathbb{Z}}$ it is sensible, and the summation sign is just a convenience to indicate in what we are interested.

Let us define the sorts of convergence one can consider for series.

4.2.7 Definition (Convergence and absolute convergence of series) Let $(\boldsymbol{x}_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}^n$ and consider the series

$$S = \sum_{j=1}^{\infty} \boldsymbol{x}_j.$$

The corresponding sequence of ***partial sums*** is the sequence $(S_k)_{k \in \mathbb{Z}_{>0}}$ defined by

$$S_k = \sum_{j=1}^{k} \boldsymbol{x}_j.$$

Let $x_0 \in \mathbb{R}^n$. The series:

(i) ***converges to*** $\mathbf{x_0}$, and we write $\sum_{j=1}^{\infty} x_j = x_0$, if the sequence of partial sums converges to x_0;

(ii) has x_0 as a ***limit*** if it converges to x_0;

(iii) is ***convergent*** if it converges to some member of $\mathbb{R}^n$;

(iv) ***converges absolutely***, or is ***absolutely convergent***, if the series

$$\sum_{j=1}^{\infty} \|x_j\|_{\mathbb{R}^n}$$

converges;

(v) ***converges conditionally***, or is ***conditionally convergent***, if it is convergent, but not absolutely convergent;

(vi) ***diverges*** if it does not converge;

(vii) has a limit that ***exists*** if $\lim_{j\to\infty} S_j \in \mathbb{R}^n$. •

We have the following correspondence between convergence and absolute convergence.

4.2.8 Proposition (Absolutely convergent series are convergent) *If a series $\sum_{j=1}^{\infty} \mathbf{x}_j$ is absolutely convergent, then it is convergent.*

Proof Let $\epsilon \in \mathbb{R}_{>0}$ and let $N \in \mathbb{Z}_{>0}$ be such that

$$\sum_{j=N}^{\infty} \|x_j\|_{\mathbb{R}^n} < \epsilon;$$

this is possible by absolute convergence (why?). Let $k, l \geq N$ with $l > k$ and compute

$$\Big\| \sum_{j=k+1}^{l} x_j \Big\| \leq \sum_{j=l+1}^{k} \|x_j\|_{\mathbb{R}^n} \leq \sum_{j=N}^{\infty} \|x_j\|_{\mathbb{R}^n} < \epsilon,$$

showing that the sequence of partial sums is Cauchy. By Theorem 4.2.5 it follows that the sequence is convergent. ■

The importance of the concept of absolute convergence is perhaps not perfectly clear at a first glance. One of the reasons it is important is that absolutely convergent series have the property that if you reorder their terms in an arbitrary way, the resulting series still converges and converges to the same limit. This is shown for real series in Theorem 2.4.5 and is explored in detail in a more general setting in Section **??**.

The following property of absolutely convergent series is often important,

4.2.9 Proposition (Swapping summation and norm) *For a sequence* $(\mathbf{x}_j)_{j\in\mathbb{Z}_{>0}}$*, if the series* $\mathbf{S} = \sum_{j=1}^{\infty} \mathbf{x}_j$ *is absolutely convergent, then*

$$\Big\|\sum_{j=1}^{\infty} \mathbf{x}_j\Big\|_{\mathbb{R}^n} \leq \sum_{j=1}^{\infty} \|\mathbf{x}_j\|_{\mathbb{R}^n}.$$

Proof Define

$$S_m^1 = \Big\|\sum_{j=1}^{m} x_j\Big\|_{\mathbb{R}^n}, \quad S_m^2 = \sum_{j=1}^{m} \|x_j\|_{\mathbb{R}^n}, \qquad m \in \mathbb{Z}_{>0}.$$

By Exercise 4.2.1 we have $S_m^1 \leq S_m^2$ for each $m \in \mathbb{Z}_{>0}$. Moreover, by Proposition 4.2.8 and Theorem 4.2.5 the sequences $(S_m^1)_{m\in\mathbb{Z}_{>0}}$ and $(S_m^2)_{m\in\mathbb{Z}_{>0}}$ are Cauchy sequences in $\mathbb{R}^n$ and so converge. It is then clear that

$$\lim_{m\to\infty} S_m^1 \leq \lim_{m\to\infty} S_m^2,$$

which is the result. ■

One can also talk about multiple series in $\mathbb{R}^n$. The definitions are just like those in Section 2.4.5 for multiple series in $\mathbb{R}$. We shall also give these definitions in a more general setting in Section **??**, so the reader can refer ahead if need be.

We can also give results analogous to those in Section 2.3.6 for series in $\mathbb{R}$. First we give some notation for products of series.

4.2.10 Definition (Scalar multiplication of series) Let $S = \sum_{j=0}^{\infty} x_j$ be a series in $\mathbb{R}^n$ and let $s = \sum_{j=0}^{\infty} a_j$ be series in $\mathbb{R}$.

(i) The ***product*** of s and S is the double series $\sum_{j,k=0}^{\infty} a_j v_k$.

(ii) The ***Cauchy product*** of s and S is the series $\sum_{k=0}^{\infty}\Big(\sum_{j=0}^{k} a_j v_{k-j}\Big)$. •

Now we can state the interaction between convergence of series and the vector space operations.

4.2.11 Proposition (Algebraic operations on series) *Let* $\mathbf{S} = \sum_{j=0}^{\infty} \mathbf{x}_j$ *and* $\mathbf{T} = \sum_{j=0}^{\infty} \mathbf{y}_j$ *be series in* $\mathbb{R}^n$ *converging to* $\mathbf{X}_0$ *and* $\mathbf{Y}_0$*, respectively, let* $s = \sum_{j=0}^{\infty} a_j$ *be a series in* $\mathbb{F}$ *converging to* A_0*, and let* $a \in \mathbb{F}$*. Then the following statements hold:*

(i) the series $\sum_{j=0}^{\infty} a\mathbf{x}_j$ *converges to* $a\mathbf{X}_0$*;*

(ii) the series $\sum_{j=0}^{\infty}(\mathbf{x}_j + \mathbf{y}_j)$ *converges to* $\mathbf{X}_0 + \mathbf{Y}_0$*;*

(iii) if s *and* $\mathbf{S}$ *are absolutely convergent, then the product of* s *and* $\mathbf{S}$ *is absolutely convergent and converges to* $A_0\mathbf{X}_0$*;*

(iv) if s *and* $\mathbf{S}$ *are absolutely convergent, then the Cauchy product of* s *and* $\mathbf{S}$ *is absolutely convergent and converges to* $A_0\mathbf{X}_0$*;*

(v) if s *or* $\mathbf{S}$ *are absolutely convergent, then the Cauchy product of* s *and* $\mathbf{S}$ *is convergent and converges to* $A_0\mathbf{X}_0$*.*

Proof The proof is identical, except for slight notational changes, to that for Proposition **??**. It also bears a resemblance to the proof of Proposition 2.4.30. Thus we do not repeat the proof here. ■

4.2.3 Open and closed balls, rectangles

Note that the definition of open (and therefore closed) sets in $\mathbb{R}$ relies on the absolute value function. Therefore, since the absolute value function has an appropriate generalisation to $\mathbb{R}^n$ as the Euclidean norm, the ideas of open and closed sets carry over to $\mathbb{R}^n$. The key idea is the generalisation of the notion of an open ball as seen in Example **??**–**??**. Here we simply make the following definition.

4.2.12 Definition (Open ball, closed ball) Let $x_0 \in \mathbb{R}^n$ and let $r \in \mathbb{R}_{\geq 0}$.

(i) The ***open ball*** centred at x_0 of radius r is the set

$$\mathsf{B}^n(r, x_0) = \{x \in \mathbb{R}^n \mid \|x - x_0\|_{\mathbb{R}^n} < r\}.$$

(ii) The ***closed ball*** centred at x_0 of radius r is the set

$$\overline{\mathsf{B}}^n(r, x_0) = \{x \in \mathbb{R}^n \mid \|x - x_0\|_{\mathbb{R}^n} \leq r\}.$$ •

For example, in the case when $n = 1$, we have

$$\mathsf{B}^1(r, x_0) = (x_0 - r, x_0 + r), \quad \overline{\mathsf{B}}^1(r, x_0) = [x_0 - r, x_0 + r].$$

Thus open and closed balls can be thought of as generalisations of open and closed intervals. In Figure 4.2 we depict how one should think of open and closed balls.

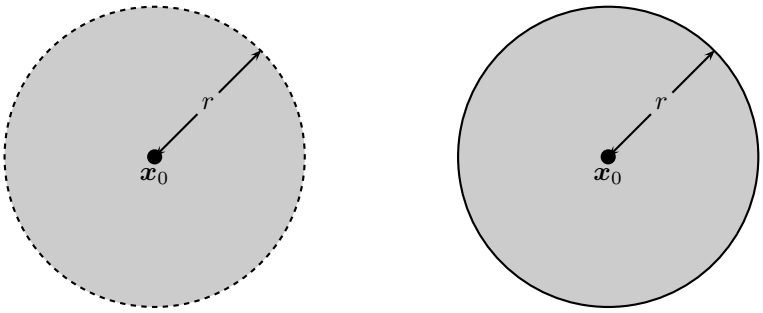


Figure 4.2 Open (left) and closed (right) balls in $\mathbb{R}^n$

4.2.13 Notation ("Balls" versus "spheres") Note that we have defined a ball of radius r as containing all points that are a distance at most r from the centre. It is also interesting to talk about the points that are a distance *exactly* r from the centre. Thus we define

$$\mathbb{S}(r, x_0) = \{x \in \mathbb{R}^n \mid \|x\|_{\mathbb{R}^n} = r\},$$

which is the ***sphere*** of radius r and centre x_0. In common language, "sphere" is often used where we mean "ball." The reader should be aware of our precise convention as we will never violate it, even casually. •

Another natural generalisation of an interval is the following.

4.2.14 Definition (Rectangle, cube) A *rectangle* in $\mathbb{R}^n$ is a subset of the form

$$R = I_1 \times \cdots \times I_n$$

where $I_1, \ldots, I_n \subseteq \mathbb{R}$ are intervals. A rectangle $R = I_1 \times \cdots \times I_n$ is ***fat*** if $\text{int}(I_j) \neq \emptyset$ for each $j \in \mathbb{Z}_{>0}$. If each of the intervals $I_1, \ldots, I_+n$ is bounded and has the same length, the resulting rectangle is called a ***cube***. •

A rectangle is, somehow, a more faithful generalisation of the notion of an interval, it being a product of intervals. Both balls (as we have defined then) and rectangles can serve as the building blocks for what we do in the remainder of this section. This is made precise only after one knows a little about topology and norm topologies; we refer to Section **??** for more details. For now we simply stick to using balls to define many of the useful structural properties of $\mathbb{R}^n$.

However, since we will use rectangles in Section **??** to define the Riemann integral, let us engage in a discussion of some useful constructions involving rectangles. These are direct generalisations of corresponding notions for intervals.

4.2.15 Definition (Partition of a compact rectangle) If

$$R = [a_1, b_1] \times \cdots \times [a_n, b_n],$$

with $a_j < b_j$, $j \in \{1, \ldots, n\}$ is a fat compact rectangle, a ***partition*** of R is an n-tuple $\boldsymbol{P} = (P_1, \ldots, P_n)$ where $P_j = (I_{j1}, \ldots, I_{jk_j})$ is a partition of the interval $[a_j, b_j]$, $j \in \{1, \ldots, n\}$. The rectangles

$$R_{l_1,\ldots,l_n} = I_{1l_1} \times \cdots \times I_{nl_n}, \qquad l_j \in \{1, \ldots, k_j\},\ j \in \{1, \ldots, n\},$$

are the ***subrectangles*** of the partition. •

Thus the partition is applied to each of the coordinate axes of the rectangle R. In Figure 4.3 we depict a partition of a two-dimensional rectangle. Note that

$$R = \mathop{\dot{\bigcup}}_{\substack{l_j \in \{1,\ldots,k_j\} \\ j \in \{1,\ldots,n\}}} R_{l_1,\ldots,l_n}.$$

As with a partition of an interval we can define a "length" of a partition $\boldsymbol{P} = (P_1, \ldots, P_n)$. We suppose that $\text{EP}_j = (x_{j0}, \ldots, x_{jk_j})$ and then define

$$|\boldsymbol{P}| = \min\{|x_{jl} - x_{jm}| \mid\ j \in \{1, \ldots, n\},\ l \in \{1, \ldots, k_j\}\}.$$

Thus $|\boldsymbol{P}|$ is the length of the smallest side of each of the rectangles whose union is R.

It is also possible to say when one partition is contained in another.

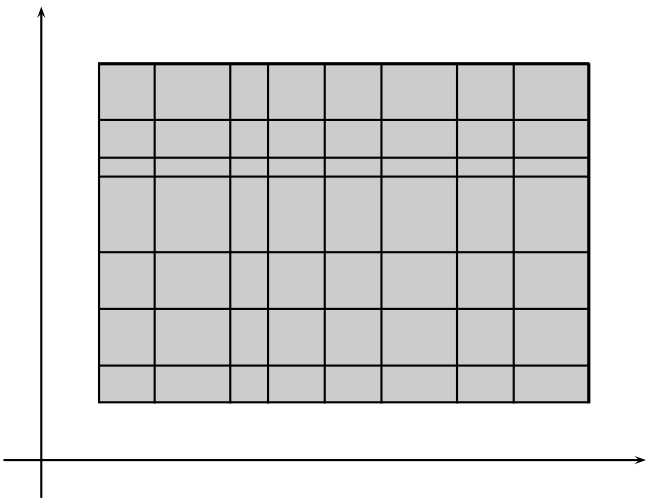


Figure 4.3 A partition of a two-dimensional rectangle

4.2.16 Definition (Refinement of a partition) Let $R \subseteq \mathbb{R}^n$ be a fat rectangle and let $\boldsymbol{P} = (P_1, \dots, P_n)$ and $\boldsymbol{P}' = (P'_1, \dots, P'_n)$ be partitions of R. Then $\boldsymbol{P}'$ is a ***refinement*** of $\boldsymbol{P}$ if P'_j is a refinement of P_j for each $j \in \{1, \dots, n\}$. •

The idea is that each of the rectangles from $\boldsymbol{P}'$ is a subset of a rectangle from $\boldsymbol{P}$.

4.2.4 Open and closed subsets

We now use open balls to define the notion of open and closed subsets of $\mathbb{R}^n$, just as we used intervals in Section 2.5.1 to define open and closed subsets of $\mathbb{R}$.

4.2.17 Definition (Open and closed sets in $\mathbb{R}^n$) A subset $A \subseteq \mathbb{R}^n$

(i) is ***open*** if, for every $\boldsymbol{x} \in A$, there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, \boldsymbol{x}) \subseteq A$ and

(ii) is ***closed*** if $\mathbb{R}^n \setminus A$ is open. •

4.2.18 Remark (Use of the words "topology" and "topological") We shall on occasion, and sometimes more frequently than that, make use of words like "topology" and "topological" in our discussion, although we will not formally introduce such terminology until Chapter **??**. The way to read our use of such words is this: They refer to things broadly related to the use of open subsets of $\mathbb{R}^n$. As we shall see, almost everything we shall say in this chapter depends in some way on open sets, their definition, and their properties. This is exactly what the study of topology consists of. •

The following properties of open and closed sets arise in the general presentation of topological spaces in Chapter **??**.

4.2.19 Proposition (Properties of open and closed sets) *For an arbitrary collection $(\mathrm{U}_a)_{a \in A}$ of open sets and an arbitrary collection $(\mathrm{C}_b)_{b \in B}$ of closed sets the following statements hold:*

(i) $\cup_{a\in A}U_a$ *is open;*

(ii) $\cap_{b\in B}C_b$ *is closed.*

Moreover, for open sets U_1 *and* U_2 *and closed sets* C_1 *and* C_2*, the following statements hold:*

(iii) $U_1 \cap U_2$ *is open;*

(iv) $C_1 \cup C_2$ *is closed.*

Proof This is Exercise 4.2.3. ■

As with open subsets of $\mathbb{R}$ the language "neighbourhood" is often useful.

4.2.20 Definition (Neighbourhood) A ***neighbourhood*** of $x \in \mathbb{R}^n$ is an open set U for which $x \in U$. More generally, a ***neighbourhood*** of a subset $A \subseteq \mathbb{R}^n$ is an open set U for which $A \subseteq U$. •

Many of the properties of open sets in $\mathbb{R}$ also hold for open subsets of $\mathbb{R}^n$.

4.2.21 Proposition (Open subsets of $\mathbb{R}$ are unions of open balls) *If* $U \subseteq \mathbb{R}^n$ *is a nonempty open set then* U *is a countable union of open balls.*

Proof Let $x \in U$ so that there exists $r_x \in \mathbb{R}_{>0}$ for which $B^n(r_x, x) \subseteq U$. By Proposition 2.2.15 there exists $q_x \in \mathbb{Q}_{>0}$ such that $q_x < r_x$. Therefore, $B^n(q_x, x) \subseteq U$. Also by Proposition 2.2.15 there exists $q_x \in \mathbb{R}^n$ with rational components such that $\|x - q_x\|_{\mathbb{R}^n} < \frac{q_x}{s}$. For $y \in B^n(\frac{q_x}{2}, q_x)$ we have

$$\|y - x\|_{\mathbb{R}^n} \leq \|y - q_x\|_{\mathbb{R}^n} + \|q_x - x\|_{\mathbb{R}^n} < \frac{q_x}{2} + \frac{q_x}{2} = q_x,$$

and so $y \in B^n(q_x, x) \subseteq U$. Thus $B^n(\frac{q_x}{2}, q_x)$ is a ball of rational radius centred at a point with rational components, contained in U and containing x. Doing this for each x gives a collection of open balls of rational radius centred at points with rational components that covers U. The result will follow is we can show that the set of balls with rational radius with centres having rational components is countable. For fixed $x \in \mathbb{R}^n$ the set of balls centred at x with rational radius is certainly countable since $\mathbb{Q}_{>0}$ is countable. The subset $\mathbb{Q}^n \subseteq \mathbb{R}^n$ is also countable by since it has cardinality $n \cdot \text{card}(\mathbb{Q})$ which is equal to $\text{card}(\mathbb{Q})$ by Theorem 1.7.17(ii). Thus the set of balls with rational radius centred at points with rational coordinates is a countable union of countable sets. Such sets are countable by Proposition 1.7.16. ■

missing stuff

4.2.5 Interior, closure, boundary, etc.

The definitions and results here are similar to those for $\mathbb{R}$ given in Section 2.5.3. Moreover, they will be discussed in a more general setting in Section **??**. The proofs in the most general setting in Section **??** are virtually identical to the proofs in the least general case in Section 2.5.3. Therefore, we elect to omit the proofs in this section, and merely state the results for reference. Readers unable to translate the results from Section 2.5.3 to this section can refer ahead to Section **??**; the only difference between the proofs in that section and what would appear here are trivial differences in notation. Moreover, examples, discussion, and motivation can be found in Section 2.5.3.

4.2.22 Definition (Accumulation point, cluster point, limit point) For a subset $A \subseteq \mathbb{R}^n$, a point $\boldsymbol{x} \in \mathbb{R}^n$ is:

(i) an ***accumulation point*** for A if, for every neighbourhood U of $\boldsymbol{x}$, the set $A \cap (U \setminus \{\boldsymbol{x}\})$ is nonempty;

(ii) a ***cluster point*** for A if, for every neighbourhood U of $\boldsymbol{x}$, the set $A \cap U$ is infinite;

(iii) a ***limit point*** of A if there exists a sequence $(\boldsymbol{x}_j)_{j \in \mathbb{Z}_{>0}}$ in A converging to $\boldsymbol{x}$.

The set of accumulation points of A is called the ***derived set*** of A, and is denoted by $\operatorname{der}(A)$. •

In Remark 2.5.12 we made some comments about conventions concerning the words "accumulation point," "cluster point," and "limit point." Those remarks apply equally here.

4.2.23 Proposition ("Accumulation point" equals "cluster point") *For a set* $\mathrm{A} \subseteq \mathbb{R}^n$, $\mathbf{x} \in \mathbb{R}^n$ *is an accumulation point for* A *if and only if it is a cluster point for* A.

4.2.24 Proposition (Properties of the derived set) *For* $\mathrm{A}, \mathrm{B} \subseteq \mathbb{R}^n$ *and for a family of subsets* $(\mathrm{A}_i)_{i \in I}$ *of* $\mathbb{R}^n$*, the following statements hold:*

(i) $\operatorname{der}(\emptyset) = \emptyset$*;*

(ii) $\operatorname{der}(\mathbb{R}^n) = \mathbb{R}^n$*;*

(iii) $\operatorname{der}(\operatorname{der}(\mathrm{A})) = \operatorname{der}(\mathrm{A})$*;*

(iv) if $\mathrm{A} \subseteq \mathrm{B}$ *then* $\operatorname{der}(\mathrm{A}) \subseteq \operatorname{der}(\mathrm{B})$*;*

(v) $\operatorname{der}(\mathrm{A} \cup \mathrm{B}) = \operatorname{der}(\mathrm{A}) \cup \operatorname{der}(\mathrm{B})$*;*

(vi) $\operatorname{der}(\mathrm{A} \cap \mathrm{B}) \subseteq \operatorname{der}(\mathrm{A}) \cap \operatorname{der}(\mathrm{B})$.

4.2.25 Definition (Interior, closure, and boundary) Let $A \subseteq \mathbb{R}^n$.

(i) The ***interior*** of A is the set

$$\operatorname{int}(A) = \cup\{U \mid U \subseteq A,\ U \text{ open}\}.$$

(ii) The ***closure*** of A is the set

$$\operatorname{cl}(A) = \cap\{C \mid A \subseteq C,\ C \text{ closed}\}.$$

(iii) The ***boundary*** of A is the set $\operatorname{bd}(A) = \operatorname{cl}(A) \cap \operatorname{cl}(\mathbb{R}^n \setminus A)$. •

4.2.26 Proposition (Characterisation of interior, closure, and boundary) *For* $\mathrm{A} \subseteq \mathbb{R}^n$*, the following statements hold:*

(i) $\mathbf{x} \in \operatorname{int}(\mathrm{A})$ *if and only if there exists a neighbourhood* U *of* $\mathbf{x}$ *such that* $\mathrm{U} \subseteq \mathrm{A}$*;*

(ii) $\mathbf{x} \in \operatorname{cl}(\mathrm{A})$ *if and only if, for each neighbourhood* U *of* $\mathbf{x}$*, the set* $\mathrm{U} \cap \mathrm{A}$ *is nonempty;*

(iii) $\mathbf{x} \in \operatorname{bd}(\mathrm{A})$ *if and only if, for each neighbourhood* U *of* $\mathbf{x}$*, the sets* $\mathrm{U} \cap \mathrm{A}$ *and* $\mathrm{U} \cap (\mathbb{R}^n \setminus \mathrm{A})$ *are nonempty.*

4.2.27 Proposition (Properties of interior) *For* $A, B \subseteq \mathbb{R}^n$ *and for a family of subsets* $(A_i)_{i\in I}$ *of* $\mathbb{R}^n$*, the following statements hold:*

(i) $\operatorname{int}(\emptyset) = \emptyset$;

(ii) $\operatorname{int}(\mathbb{R}^n) = \mathbb{R}^n$;

(iii) $\operatorname{int}(\operatorname{int}(A)) = \operatorname{int}(A)$;

(iv) if $A \subseteq B$ *then* $\operatorname{int}(A) \subseteq \operatorname{int}(B)$;

(v) $\operatorname{int}(A \cup B) \supseteq \operatorname{int}(A) \cup \operatorname{int}(B)$;

(vi) $\operatorname{int}(A \cap B) = \operatorname{int}(A) \cap \operatorname{int}(B)$;

(vii) $\operatorname{int}(\cup_{i\in I} A_i) \supseteq \cup_{i\in I} \operatorname{int}(A_i)$;

(viii) $\operatorname{int}(\cap_{i\in I} A_i) \subseteq \cap_{i\in I} \operatorname{int}(A_i)$.

Moreover, a set $A \subseteq \mathbb{R}^n$ *is open if and only if* $\operatorname{int}(A) = A$.

4.2.28 Proposition (Properties of closure) *For* $A, B \subseteq \mathbb{R}^n$ *and for a family of subsets* $(A_i)_{i\in I}$ *of* $\mathbb{R}^n$*, the following statements hold:*

(i) $\operatorname{cl}(\emptyset) = \emptyset$;

(ii) $\operatorname{cl}(\mathbb{R}^n) = \mathbb{R}^n$;

(iii) $\operatorname{cl}(\operatorname{cl}(A)) = \operatorname{cl}(A)$;

(iv) if $A \subseteq B$ *then* $\operatorname{cl}(A) \subseteq \operatorname{cl}(B)$;

(v) $\operatorname{cl}(A \cup B) = \operatorname{cl}(A) \cup \operatorname{cl}(B)$;

(vi) $\operatorname{cl}(A \cap B) \subseteq \operatorname{cl}(A) \cap \operatorname{cl}(B)$;

(vii) $\operatorname{cl}(\cup_{i\in I} A_i) \supseteq \cup_{i\in I} \operatorname{cl}(A_i)$;

(viii) $\operatorname{cl}(\cap_{i\in I} A_i) \subseteq \cap_{i\in I} \operatorname{cl}(A_i)$.

Moreover, a set $A \subseteq \mathbb{R}^n$ *is closed if and only if* $\operatorname{cl}(A) = A$.

4.2.29 Proposition (Joint properties of interior, closure, boundary, and derived set) *For* $A \subseteq \mathbb{R}^n$*, the following statements hold:*

(i) $\mathbb{R}^n \setminus \operatorname{int}(A) = \operatorname{cl}(\mathbb{R}^n \setminus A)$;

(ii) $\mathbb{R}^n \setminus \operatorname{cl}(A) = \operatorname{int}(\mathbb{R}^n \setminus A)$.

(iii) $\operatorname{cl}(A) = A \cup \operatorname{bd}(A)$;

(iv) $\operatorname{int}(A) = A - \operatorname{bd}(A)$;

(v) $\operatorname{cl}(A) = \operatorname{int}(A) \cup \operatorname{bd}(A)$;

(vi) $\operatorname{cl}(A) = A \cup \operatorname{der}(A)$;

(vii) $\mathbb{R}^n = \operatorname{int}(A) \cup \operatorname{bd}(A) \cup \operatorname{int}(\mathbb{R}^n \setminus A)$.

We close this section by defining a useful notion related to the topics of this section.

4.2.30 Definition (Dense subset) A subset $D \subseteq \mathbb{R}^n$ is ***dense*** if $\operatorname{cl}(D) = \mathbb{R}^n$. •

There is a simple example of a countable dense subset of $\mathbb{R}^n$.

4.2.31 Example (Countable dense subset) The set $\mathbb{Q}^n$ is a dense subset of $\mathbb{R}^n$. To verify this one needs only, for $x \in \mathbb{R}^n$, to construct a sequence $(q_j)_{j\in\mathbb{Z}_{>0}}$ converging to x. That this is possible follows from the fact that $\mathbb{Q} \subseteq \mathbb{R}$ is dense, along with Proposition 4.2.4. Moreover, note that $\mathbb{Q}^n$ is countable by Theorem 1.7.17. •

4.2.6 Compact subsets

The notion of compactness, relying as it does only on the idea of an open set, is transferable from $\mathbb{R}$ to $\mathbb{R}^n$, and indeed to the general setting of Chapter **??** (see Section **??**). That is to say, the idea of an open cover of a subset of $\mathbb{R}^n$ transfers directly from $\mathbb{R}$, and, therefore, the definition of a compact set as being a set for which every open cover possesses a finite subcover also generalises. In this section we explore the details of this for $\mathbb{R}^n$.

We begin with some notions associated to open covers.

4.2.32 Definition (Open cover of a subset of $\mathbb{R}^n$) Let $A \subseteq \mathbb{R}^n$.

(i) An ***open cover*** for A is a family $(U_i)_{i\in I}$ of open subsets of $\mathbb{R}^n$ having the property that $A \subseteq \cup_{i\in I} U_i$.

(ii) A ***subcover*** of an open cover $(U_i)_{i\in I}$ of A is an open cover $(V_j)_{j\in J}$ of A having the property that $(V_j)_{j\in J} \subseteq (U_i)_{i\in I}$. •

The following property of open covers of subsets of $\mathbb{R}^n$ is useful.

4.2.33 Lemma (Lindelöf Lemma for $\mathbb{R}^n$) *If $(U_i)_{i\in I}$ is an open cover of $A \subseteq \mathbb{R}^n$, then there exists a countable subcover of A.*

Proof Let $\mathscr{B} = \{\mathsf{B}^n(r, x) \mid r \in \mathbb{Q}, x \in \mathbb{Q}^n\}$. Note that $\mathscr{B}$ is a countable union of countable sets, and so is countable by Proposition 1.7.16 (also see the last part of the proof of Proposition 4.2.21). Therefore, we can write $\mathscr{B} = (\mathsf{B}^n(r_j, x_j))_{j\in\mathbb{Z}_{>0}}$. Now define

$$\mathscr{B}' = \{\mathsf{B}^n(r_j, x_j) \mid \mathsf{B}^n(r_j, x_j) \subseteq U_i \text{ for some } i \in I\}.$$

Let us write $\mathscr{B}' = (\mathsf{B}^n(r_{j_k}, x_{j_k}))_{k\in\mathbb{Z}_{>0}}$. We claim that $\mathscr{B}'$ covers A. Indeed, if $x \in A$ then $x \in U_i$ for some $i \in I$. Then there exists $k \in \mathbb{Z}_{>0}$ such that $x \in \mathsf{B}^n(r_{j_k}, x_{j_k}) \subseteq U_i$. Now, for each $k \in \mathbb{Z}_{>0}$, let $i_k \in I$ satisfy $\mathsf{B}^n(r_{j_k}, x_{j_k}) \subseteq U_{i_k}$. Then the countable collection of open sets $(U_{i_k})_{k\in\mathbb{Z}_{>0}}$ clearly covers A since $\mathscr{B}'$ covers A. ■

Now we define the important notion of compactness, along with some other related useful concepts.

4.2.34 Definition (Bounded, compact, and totally bounded in $\mathbb{R}^n$) A subset $A \subseteq \mathbb{R}^n$ is:

(i) ***bounded*** if there exists $M \in \mathbb{R}_{>0}$ such that $A \subseteq \overline{\mathsf{B}}^n(M, \mathbf{0})$;

(ii) ***compact*** if every open cover $(U_i)_{i\in I}$ of A possesses a finite subcover;

(iii) ***precompact***[3] if $\mathrm{cl}(A)$ is compact;

(iv) ***totally bounded*** if, for every $\epsilon \in \mathbb{R}_{>0}$ there exists $x_1, \ldots, x_k \in \mathbb{R}^n$ such that $A \subseteq \cup_{j=1}^k \mathsf{B}^n(\epsilon, x_j)$. •

[3]What we call "precompact" is very often called "relatively compact." However, we shall use the term "relatively compact" for something different.

The simplest characterisation of compact subsets of $\mathbb{R}^n$ is the following. We shall freely interchange our use of the word compact between the definition given in Definition 4.2.34 and the conclusions of the following theorem.

4.2.35 Theorem (Heine–Borel Theorem in $\mathbb{R}^n$) *A subset* $\mathrm{K} \subseteq \mathbb{R}^n$ *is compact if and only if* K *is closed and bounded.*

Proof We first prove a couple of lemmata.

1 Lemma *If* $\mathrm{K}_1 \subseteq \mathbb{R}^m$ *is compact and if* $\mathrm{K}_2 \subseteq \mathbb{R}^n$ *is compact then* $\mathrm{K}_1 \times \mathrm{K}_2 \subseteq \mathbb{R}^{m+n}$ *is compact.*

Proof Let us denote points in $\mathbb{R}^{m+n}$ by $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^m \times \mathbb{R}^n$. For $\boldsymbol{x} \in \mathbb{R}^m$ denote

$$K_{2,\boldsymbol{x}} = \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{y} \in K_2\}.$$

Let $(U_a)_{a \in A}$ be an open cover of $K_1 \times K_2$. For $\boldsymbol{x} \in K_1$ denote

$$A_{\boldsymbol{x}} = \{a \in A \mid U_a \cap K_{2,\boldsymbol{x}} \neq \emptyset\}.$$

For $a \in A_{\boldsymbol{x}}$ define

$$V_a = \{\boldsymbol{y} \in U_a \mid (\boldsymbol{x}, \boldsymbol{y}) \in K_{2,\boldsymbol{x}}\}.$$

We claim that V_a is open. Indeed, let $\boldsymbol{y} \in V_a$ so that $(\boldsymbol{x}, \boldsymbol{y}) \in U_a$. Since U_a is open there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^{m+n}(\epsilon, (\boldsymbol{x}, \boldsymbol{y})) \subseteq U_a$. Therefore $\mathsf{B}^n(\epsilon, \boldsymbol{y}) \subseteq V_a$, and so V_a is open as claimed. Therefore, $(V_a)_{a \in A_{\boldsymbol{x}}}$ is an open cover of K_2. Thus there exists $a_{\boldsymbol{x},1}, \ldots, a_{\boldsymbol{x},k_{\boldsymbol{x}}} \in A_{\boldsymbol{x}}$ such that $K_2 \subseteq \cup_{j=1}^{k_{\boldsymbol{x}}} V_{a_{\boldsymbol{x},j}}$.

Now, for $a \in A$ denote

$$W_a = \{\boldsymbol{x} \in \mathbb{R}^m \mid (\boldsymbol{x}, \boldsymbol{y}) \in U_a \text{ for some } \boldsymbol{y} \in \mathbb{R}^n\}.$$

We claim that W_a is open. To see this, let $\boldsymbol{x} \in W_a$ and let $\boldsymbol{y} \in \mathbb{R}^n$ be such that $(\boldsymbol{x}, \boldsymbol{y}) \in U_a$. Since U_a is open there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^{m+n}(\epsilon, (\boldsymbol{x}, \boldsymbol{y})) \subseteq U_a$. Therefore, $\mathsf{B}^m(\epsilon, \boldsymbol{x}) \subseteq W_a$, giving W_a as open, as desired. Now define $W_{\boldsymbol{x}} = \cap_{j=1}^{k_{\boldsymbol{x}}} W_{a_j}$ and note that by Exercise 4.2.3 it follows that $W_{\boldsymbol{x}}$ is open. Thus $(W_{\boldsymbol{x}})_{\boldsymbol{x} \in K_1}$ is an open cover for K_1. By compactness of K_1 there exists $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in K_1$ such that $K_1 = \cup_{l=1}^m W_{\boldsymbol{x}_l}$. Therefore,

$$K_1 \times K_2 = \cup_{\boldsymbol{x} \in K_1} K_{2,\boldsymbol{x}} = \cup_{\boldsymbol{x} \in K_1} \cup_{j=1}^{k_{\boldsymbol{x}}} U_{a_{\boldsymbol{x},j}} = \cup_{l=1}^m \cup_{j=1}^{k_{\boldsymbol{x}_l}} U_{a_{\boldsymbol{x}_l,j}},$$

so giving a finite subcover of $K_1 \times K_2$. ▼

2 Lemma *If* A *is compact and if* $\mathrm{B} \subseteq \mathrm{A}$ *is closed, then* B *is compact.*

Proof Let $(U_i)_{i \in I}$ be an open cover for B and define $V = \mathbb{R}^m \setminus B$. Since B is closed, $(U_i)_{i \in I} \cup (V)$ is an open cover for A. Since A is compact there exists $i_1, \ldots, i_k \in I$ such that $A \subseteq \cup_{j=1}^k U_{i_j} \cup V$. Therefore, $B \subseteq \cup_{j=1}^k U_{i_k}$, giving a finite subcover of B. ▼

Suppose that K is closed and bounded. Let $R \in \mathbb{R}_{>0}$ be sufficiently large that $K \subseteq [-R, R] \times \cdots \times [-R, R]$. By Theorem 2.5.27 it follows that $[-R, R]$ is compact. By induction using Lemma 1 it follows that $[-R, R] \times \cdots \times [-R, R]$ is compact. By Lemma 2 it follows that K is compact.

Next suppose that K is compact. For $\epsilon \in \mathbb{R}_{>0}$ consider the open cover $(\mathsf{B}^n(\epsilon, \boldsymbol{x}))_{\boldsymbol{x} \in K}$ of K. Since K is compact there exists $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in K$ such that $K \subseteq \cup_{j=1}^k \mathsf{B}^n(\epsilon, \boldsymbol{x}_j)$. If

$$M_0 = \max\{\|\boldsymbol{x}_j - \boldsymbol{x}_l\|_{\mathbb{R}^n} \mid j, l \in \{i, \ldots, k\}\} + 2\epsilon,$$

then it is easy to see that $A \subseteq \mathsf{B}^n(M, \mathbf{0})$ for any $M > M_0$; thus K is bounded. Now suppose that K is compact but not closed. Then, by Proposition 4.2.28, there exists $x_0 \in \mathrm{cl}(K) \setminus K$. For each $x \in K$ let $r_x \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^n(\epsilon_x, x) \cap \mathsf{B}^n(\epsilon_x, x_0) = \emptyset$. Then $(\mathsf{B}^n(\epsilon_x, x))_{x \in K}$ is an open cover of K. Therefore, there exists $x_1, \ldots, x_k \in K$ such that $K \subseteq \cup_{j=1}^k \mathsf{B}^n(\epsilon_{x_j}, x_j)$. But this means that K does not intersect the open subset $\cap_{j=1}^k \mathsf{B}^n(\epsilon_{x_j}, x_0)$, so contradicting the existence of $x \in \mathrm{cl}(K) \setminus K$. Thus $K = \mathrm{cl}(K)$, giving the result. ■

The Heine–Borel Theorem has the following useful corollary.

4.2.36 Corollary (Closed subsets of compact sets in $\mathbb{R}^n$ are compact) *If* A $\subseteq \mathbb{R}^n$ *is compact and if* B $\subseteq$ A *is closed, then* B *is compact.*

Proof This was proved as Lemma 2 in the proof of the Heine–Borel Theorem. ■

As we warned the reader in Section 2.5.4, care must be taken when generalising the notion of compactness from $\mathbb{R}^n$ to the more general notion of a topological space as defined in Chapter **??**. A key fact is that compactness and closed and boundedness are not generally equivalent. Perhaps the nicest illustration of this is given in Theorem **??** where it is shown that, for Banach spaces, this equivalence happens only in finite dimensions.

The following result is another equivalent characterisation of compact subsets of $\mathbb{R}^n$, and is often useful.

4.2.37 Theorem (Bolzano–Weierstrass Theorem in $\mathbb{R}^n$) *A subset* K $\subseteq \mathbb{R}^n$ *is compact if and only if every sequence in* K *has a subsequence which converges in* K.

Proof Suppose that there exists a sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ in K having no convergent subsequence. This means that for each $j \in \mathbb{Z}_{>0}$ there exists $\epsilon_j \in \mathbb{R}_{>0}$ such that $x_k \notin \mathsf{B}^n(\epsilon_j, x_j)$ for $k \neq j$. Let $X \triangleq \{x_j \mid j \in \mathbb{Z}_{>0}\}$. The open cover $(\mathsf{B}^n(\epsilon_j, x_j))_{j \in \mathbb{Z}_{>0}}$ of X possesses no finite subcover and so X is not compact. We claim that the set is closed. Indeed, if $x \in \mathrm{cl}(X)$ it follows by Proposition 4.2.26 that x is the limit of a sequence in X. But the only such sequences are those that are eventually constant, and so the claim follows. By Corollary 4.2.36 it now follows that K is not compact since it possesses a closed but not compact subset.

Next suppose that every sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ in K possesses a convergent subsequence. Let $(U_i)_{i \in I}$ be an open cover of K, and by Lemma 4.2.33 choose a countable subcover which we denote by $(U_j)_{j \in \mathbb{Z}_{>0}}$. Now suppose that every finite subcover of $(U_j)_{j \in \mathbb{Z}_{>0}}$ does not cover K. This means that, for every $k \in \mathbb{Z}_{>0}$, the set $C_k = K \setminus \left(\cup_{j=1}^k U_j\right)$ is nonempty. Thus we may define a sequence $(x_k)_{k \in \mathbb{Z}_{>0}}$ in $\mathbb{R}^n$ such that $x_k \in C_k$. Since the sequence $(x_k)_{k \in \mathbb{Z}_{>0}}$ is in K, it possesses a convergent subsequence $(x_{k_m})_{m \in \mathbb{Z}_{>0}}$, by hypotheses. Let x be the limit of this subsequence. Since $x \in K$ and since $K = \cup_{j \in \mathbb{Z}_{>0}} U_j$, $x \in U_l$ for some $l \in \mathbb{Z}_{>0}$. Since the sequence $(x_{k_m})_{m \in \mathbb{Z}_{>0}}$ converges to x, it follows that there exists $N \in \mathbb{Z}_{>0}$ such that $x_{k_m} \in U_l$ for $m \geq N$. But this contradicts the definition of the sequence $(x_k)_{k \in \mathbb{Z}_{>0}}$, forcing us to conclude that our assumption is wrong that there is no finite subcover of K from the collection $(U_j)_{j \in \mathbb{Z}_{>0}}$. ■

The following property of compact subsets of $\mathbb{R}^n$ is useful.

4.2.38 Theorem (Lebesgue number for compact sets) *Let $\mathrm{K} \subseteq \mathbb{R}^n$ be a compact set. Then for any open cover $(\mathrm{U}_\alpha)_{\alpha\in\mathrm{A}}$ of K, there exists $\delta \in \mathbb{R}_{>0}$, called the **Lebesgue number** of K, such that, for each $\mathbf{x} \in \mathrm{K}$, there exists $\alpha \in \mathrm{A}$ such that $\mathsf{B}^n(\delta, \mathbf{x}) \cap \mathrm{K} \subseteq \mathrm{U}_\alpha$.*

Proof Suppose there exists an open cover $(U_\alpha)_{\alpha\in A}$ such that, for all $\delta \in \mathbb{R}_{>0}$, there exists $x \in K$ such that none of the sets U_α, $\alpha \in A$, contains $\mathsf{B}^n(\delta, x) \cap K$. Then there exists a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in K such that

$$\left\{\alpha \in A \;\middle|\; \mathsf{B}^n(\tfrac{1}{j}, x_j) \subseteq U_\alpha\right\} = \emptyset$$

for each $j \in \mathbb{Z}_{>0}$. By the Bolzano–Weierstrass Theorem there exists a subsequence $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$ that converges to a point, say x, in K. Then there exists $\epsilon \in \mathbb{R}_{>0}$ and $\alpha \in A$ such that $\mathsf{B}^n(\epsilon, x) \subseteq U_\alpha$. Now let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $\|x_{j_k} - x\|_{\mathbb{R}^n} < \frac{\epsilon}{2}$ for $k \geq N$ and such that $\frac{1}{j_N} < \frac{\epsilon}{2}$. Now let $k \geq N$. Then, if $y \in \mathsf{B}^n(\frac{1}{j_k}, x_{j_k})$ we have

$$\|y - x\|_{\mathbb{R}^n} = \|y - x_{j_k} + x_{j_k} - x\|_{\mathbb{R}^n} \leq \|y - x_{j_k}\|_{\mathbb{R}^n} + \|x - x_{j_k}\|_{\mathbb{R}^n} < \epsilon.$$

Thus we arrive at the contradiction that $\mathsf{B}^n(\frac{1}{j_k}, x_{j_k}) \subseteq U_\alpha$. ■

The following result is useful and is sometimes known as the ***Cantor Intersection Theorem***.

4.2.39 Proposition (Countable intersections of nested compact sets are nonempty) *Let $(\mathrm{K_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ be a collection of nonempty compact subsets of $\mathbb{R}^n$ satisfying $\mathrm{K_{j+1}} \subseteq \mathrm{K_j}$. Then $\cap_{\mathrm{j}\in\mathbb{Z}_{>0}}\mathrm{K_j}$ is nonempty.*

Proof It is clear that $K = \cap_{j\in\mathbb{Z}_{>0}} K_j$ is bounded, and moreover it is closed by Exercise 4.2.3. Thus K is compact by the Heine–Borel Theorem. Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence for which $x_j \in K_j$ for $j \in \mathbb{Z}_{>0}$. This sequence is thus a sequence in K_1 and so, by the Bolzano–Weierstrass Theorem, has a subsequence $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$ converging to $x \in K_1$. The sequence $(x_{j_{k+1}})_{k\in\mathbb{Z}_{>0}}$ is then a sequence in K_2 which is convergent, so showing that $x \in K_2$. Similarly, one shows that $x \in K_j$ for all $j \in \mathbb{Z}_{>0}$, giving the result. ■

Finally, let us indicate the relationship between the notions of relative compactness and total boundedness. We see that for $\mathbb{R}^n$ these concepts are the same. This may not be true in general.***missing stuff***

4.2.40 Proposition ("Precompact" equals "totally bounded" in $\mathbb{R}^n$) *A subset of $\mathbb{R}^n$ is precompact if and only if it is totally bounded.*

Proof Let $A \subseteq \mathbb{R}^n$.

First suppose that A is precompact. Since $A \subseteq \mathrm{cl}(A)$ and since $\mathrm{cl}(A)$ is bounded by the Heine–Borel Theorem, it follows that A is bounded. We claim that A is then totally bounded. Let $M \in \mathbb{R}_{>0}$ be such that $A \subseteq \overline{\mathsf{B}}^n(M, \mathbf{0})$ so that $\mathrm{cl}(A) \subseteq \overline{\mathsf{B}}^n(M, \mathbf{0})$ by Proposition 4.2.28(iv). Thus $\mathrm{cl}(A)$ is closed and bounded, and so compact by the Heine–Borel Theorem. For $\epsilon \in \mathbb{R}_{>0}$ note that $(\mathsf{B}^n(\epsilon, x))_{x\in\mathrm{cl}(A)}$ is an open cover of $\mathrm{cl}(A)$. Thus there exists a finite collection $x_1, \ldots, x_k \in \mathrm{cl}(A)$ such that $\mathrm{cl}(A) \subseteq \cup_{j=1}^k \mathsf{B}^n(\epsilon, x_j)$. Since $A \subseteq \mathrm{cl}(A)$ this shows that A is totally bounded.

Now suppose that A is totally bounded. For $\epsilon \in \mathbb{R}_{>0}$ let $x_1, \ldots, x_k \in \mathbb{R}^n$ have the property that $A \subseteq \cup_{j=1}^k \mathsf{B}^n(\epsilon, x_j)$. If

$$M_0 = \max\{\|x_j - x_l\|_{\mathbb{R}^n} \mid j, l \in \{i, \ldots, k\}\} + 2\epsilon,$$

then it is easy to see that $A \subseteq \mathsf{B}^n(M, \mathbf{0})$ for any $M > M_0$. Then $\mathrm{cl}(A) \subseteq \overline{\mathsf{B}}^n(M, \mathbf{0})$ by part (iv) of Proposition 4.2.28, and so $\mathrm{cl}(A)$ is bounded. Since $\mathrm{cl}(A)$ is closed, it follows from the Heine–Borel Theorem that A is precompact. ■

We close this section with a discussion of a notion of the size of a set.

4.2.41 Definition (Diameter of a set) The *diameter* of a set $A \subseteq \mathbb{R}^n$ is

$$\mathrm{diam}(A) = \sup\{\|x_1 - x_2\|_{\mathbb{R}^n} \mid x_1, x_2 \in A\}.$$ •

The following properties of the diameter are useful.

4.2.42 Proposition (Properties of diameter) *For* $\mathrm{A} \subseteq \mathbb{R}^n$ *the following statements hold:*

(i) $\mathrm{diam(A)} < \infty$ *if and only if* A *is bounded;*

(ii) $\mathrm{diam(cl(A))} = \mathrm{diam(A)}$.

Proof (i) Suppose that $\mathrm{diam}(A) = D \in \mathbb{R}_{>0}$. Let $x_0 \in A$ and define $M = D + \|x_0\|_{\mathbb{R}^n}$. Then, for $x \in A$ we have

$$\|x\|_{\mathbb{R}^n} = \|x - x_0\|_{\mathbb{R}^n} + \|x_0\|_{\mathbb{R}^n} < M$$

and so $A \subseteq \mathsf{B}^n(M, \mathbf{0})$.

Now suppose that A is bounded and let $M \in \mathbb{R}_{>0}$ be such that $A \subseteq \mathsf{B}^n(M, \mathbf{0})$. Let $x_1, x_2 \in A$ so that

$$\|x_1 - x_2\|_{\mathbb{R}^n} \leq \|x_1\|_{\mathbb{R}^n} + \|x_2\|_{\mathbb{R}^n} < 2M.$$

Therefore,

$$\sup\{\|x_1 - x_2\|_{\mathbb{R}^n} \mid x_1, x_2 \in A\} \leq 2M,$$

and so $\mathrm{diam}(A) \leq 2M$.

(ii) Let $x_1, x_2 \in \mathrm{cl}(A)$ and let $(x_{1,j})_{j \in \mathbb{Z}_{>0}}$ and $(x_{2,j})_{j \in \mathbb{Z}_{>0}}$ be sequences in A converging to x_1 and x_2, respectively. Then, for each $j \in \mathbb{Z}_{>0}$,

$$\|x_{1,j} - x_{2,j}\|_{\mathbb{R}^n} \leq \mathrm{diam}(A),$$

which gives

$$\|x_1 - x_2\|_{\mathbb{R}^n} = \lim_{j \to \infty} \|x_{1,j} - x_{2,j}\|_{\mathbb{R}^n} \leq \mathrm{diam}(A),$$

where we have swapped the limit with the norm using continuity of the norm (***missing stuff***) and Theorem 4.3.2. ■

4.2.7 Connected subsets

It is pretty easy to characterise connectivity in $\mathbb{R}$, as we saw in Section 2.5.5. Here we discuss connectedness in $\mathbb{R}^n$, and as we shall see things are a little more complicated in this case.

One of the reasons why connectedness is more complicated in dimensions higher than one is because there are two natural distinct notions of connectivity. As we shall see, these agree in one dimension, but not in higher dimensions.

The first notion we consider is fairly intuitive. It relies on the notion of paths in Euclidean spaces which are discussed in Section **??**. Readers who cannot imagine what is the definition of a path can refer ahead.

4.2.43 Definition (Path-connected subset of $\mathbb{R}^n$) A subset $A \subseteq \mathbb{R}^n$ is ***path-connected*** if, for every $x_0, x_1 \in \mathbb{R}^n$ there exists a path $\gamma\colon [a,b] \to \mathbb{R}^n$ such that $\gamma(s) \in A$ for every $s \in [a,b]$ and such that $\gamma(a) = x_0$ and $\gamma(b) = x_1$. •

The idea is that the map γ is to be thought of as a curve, or path, from x_1 to x_2. Path-connectedness of A is the property of going from any point in A to any other point in A in a continuous manner while remaining in A. This is depicted in Figure 4.4.

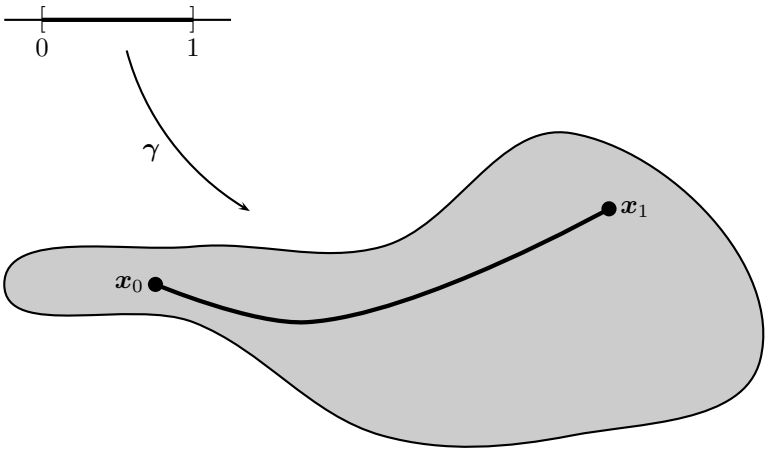


Figure 4.4 A depiction of a path-connected set

Besides this fairly intuitive notion of path-connectedness (which, as we shall see, agrees with our notion of connectedness from Definition 2.5.33) we can duplicate the definition we have already seen for subsets of $\mathbb{R}$.

4.2.44 Definition (Connected subset of $\mathbb{R}^n$) Subsets $A, B \subseteq \mathbb{R}^n$ are ***separated*** if $A \cap \operatorname{cl}(B) = \emptyset$ and $\operatorname{cl}(A) \cap B = \emptyset$. A subset $S \subseteq \mathbb{R}^n$ is ***disconnected*** if $S = A \cup B$ for nonempty separated subsets A and B. A subset $S \subseteq \mathbb{R}^n$ is ***connected*** if it is not disconnected. •

For subsets of $\mathbb{R}$ (i.e., in the case when $n = 1$) we have the simple characterisation of connected sets from Theorem 2.5.34. For subsets of $\mathbb{R}^n$ with $n > 1$ there is no such elementary characterisation. Indeed, as we shall see in Example 4.2.46 below, some connected sets can be pretty complicated, and not "obviously" connected.

But before we get to this, let us give the relationship between connectedness and path-connectedness.

4.2.45 Proposition (Path-connected sets are connected) *If $A \subseteq \mathbb{R}^n$ is path-connected then it is connected.*

Proof Suppose that A is not connected but is path-connected. Let $A = A_1 \cup A_2$ with A_1 and A_2 nonempty separated sets. Let $x_1 \in A_1$ and $x_2 \in A_2$ and let $\gamma\colon [0,1] \to \mathbb{R}^2$ be continuous, A-valued, and have the property that $\gamma(0) = x_1$ and $\gamma(1) = x_2$. Define $B_1 = \gamma^{-1}(A_1)$ and $B_2 = \gamma^{-1}(A_2)$. We claim that B_1 and B_2 are separated. Indeed, suppose that $B_1 \cap \operatorname{cl}(B_2)$ is nonempty and let $s_0 \in B_1 \cap \operatorname{cl}(B_2)$. Since $s_0 \in B_1$ we have $\gamma(s_0) \in A_1$. Note that $\operatorname{cl}(B_2)$ is closed and bounded, and so compact by the Heine–Borel Theorem.

By Proposition 4.3.29 it follows that $\gamma(\text{cl}(B_2))$ is compact, and so in particular closed. Since γ is continuous, since $\text{cl}(B_2)$ is closed, and since $\gamma(\text{cl}(B_2))$ is closed, it follows from Theorem 4.3.2 and Proposition 4.2.26 that $\gamma(s_0) \in \gamma(\text{cl}(B_2))$. But this implies that $\gamma(s_0) \in \text{cl}(A_2)$ and so this contradicts the connectedness of A. Thus A cannot be path-connected. ■

4.2.46 Example (A set that is connected but not path connected) Let us consider the subset S of $\mathbb{R}^2$ defined by

$$S = \{(x, y) \in \mathbb{R}^2 \mid \ y = \sin \tfrac{1}{x}, \ x \neq 0\} \cup \{(0, y) \mid \ y \in [-1, 1]\}.$$

In Figure 4.5 we depict this subset which is sometimes called the ***topologist's sine***

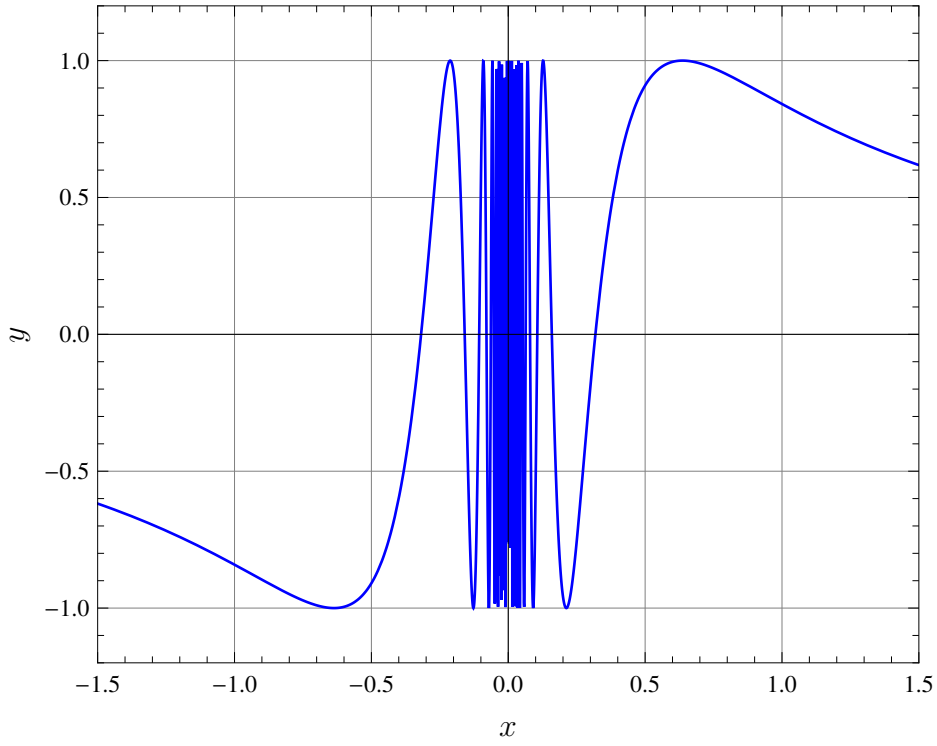


Figure 4.5 The topologist's sine curve

curve. (Actually, usually the first set in the definition of S is what is the topologist's sine curve, and the set S is its closure.)

We first claim that S is connected. Let us write $S = S_1 \cup S_2 \cup S_3$ with

$$\begin{aligned} S_1 &= \{(x, y) \in \mathbb{R}^2 \mid \ y = \sin \tfrac{1}{x}, \ x > 0\}, \\ S_2 &= \{(x, y) \in \mathbb{R}^2 \mid \ y = \sin \tfrac{1}{x}, \ x < 0\}, \\ S_3 &= \{(0, y) \mid \ y \in [-1, 1]\}. \end{aligned}$$

It is evident that S_1, S_2, and S_2 are path-connected since they are images of intervals under continuous maps. Therefore, they are connected by Proposition 4.2.45. Thus

none of S_1, S_2, or S_3 are the union of separated subsets. Moreover, since S is the closure of $S_1 \cup S_2$ (why?) it follows that S is connected by Exercise 4.2.6.

Next we claim that S is not path-connected. To see this, suppose that there exists a continuous map $\gamma\colon [0,1] \to \mathbb{R}^2$ taking values in S and such that $\gamma(0) = (\frac{1}{\pi}, 0)$ and $\gamma(1) = (0,0)$. Let

$$s_* = \inf\{s \in [0,1] \mid \gamma(s) \in \{0\} \times \mathbb{R}\}.$$

Such an s_* exists since $\gamma(1) = (0,0)$ and so $s_* \leq 1$. Therefore, $\gamma([0, s_*])$ intersects the y-axis at exactly one point. However, $S_3 \subseteq \operatorname{cl}(\gamma([0, s_*])$ (why?) which implies that $\gamma([0, s_*])$ is not closed, and so not compact by the Heine–Borel Theorem. But this contradicts the continuity of γ by Proposition 4.3.29. •

An important class of subsets where connectedness and path-connectedness agree are open sets. Here one can connect points with particular paths called polygonal paths. The reader can get the precise definition from Definition **??**, although the intuition is easy: a polygonal path is formed from a finite collection of line segments.

4.2.47 Theorem (Open connected sets are polygonally path connected) *If* $\mathrm{U} \subseteq \mathbb{R}^n$ *is open and connected then, given* $\mathbf{x}_0, \mathbf{x}_1 \in \mathrm{U}$*, there exists a polygonal path lying in* U *connecting* $\mathbf{x}_0$ *and* $\mathbf{x}_1$*.*

Proof Let $x_0 \in U$ and let $A_{x_0} \subseteq U$ be the set of points that can be connected to x_0 with a polygonal path lying in U. We claim that A_{x_0} is a nonempty open set. Since U is open there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, x_0) \subseteq U$. If $v \in \mathbb{R}^n$ such that $\|v\|_{\mathbb{R}^n} = 1$ then

$$x_0 + sv \in \mathsf{B}^n(\epsilon, x_0) \subseteq U, \qquad s \in [0, \epsilon).$$

Thus A_{x_0} is not empty. Now let $x \in A_{x_0}$. Since $x \in U$ there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, x) \subseteq U$. Again, for any vector $v \in \mathbb{R}^n$ such that $\|v\|_{\mathbb{R}^n} = 1$ we have

$$x + sv \in \mathsf{B}^n(\epsilon, x_0) \subseteq U, \qquad s \in [0, \epsilon).$$

Thus $\mathsf{B}^n(\epsilon, x) \subseteq A_{x_0}$ since x_0 can be connected to x by a polygonal path and every point in $\mathsf{B}^n(\epsilon, x)$ can be connected to x by a segment. This shows that A_{x_0} is open.

Next we claim that $\operatorname{bd}(A_{x_0}) \cap U = \emptyset$. Indeed, let $x \in \operatorname{bd}(A_{x_0}) \cap U$. Since $x \in U$ there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, x) \subseteq U$. Since $x \in \operatorname{bd}(A_{x_0})$ and by Proposition 4.2.26, there exists $x' \in \mathsf{B}^n(\epsilon, x)$ such that $x' \in A_{x_0}$. But then x can be connected to x' by a segment (just as in the preceding parts of the proof) and x_0 can be connected to x' by a polygonal path, meaning that x_0 can be connected to x by a polygonal path. Thus $x \in A_{x_0} \cap \operatorname{bd}(A_{x_0})$, contradicting the openness of A_{x_0}.

Let $B_{x_0} = U \setminus A_{x_0}$. We claim that $B_{x_0} = \emptyset$. Suppose otherwise. First we claim that B_{x_0} is open. Let $x \in B_{x_0}$. Since $x \in U$ there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, x) \subseteq U$. As above, x can be connected to any point in $\mathsf{B}^n(\epsilon, x)$ by a segment. This ensures that $\mathsf{B}^n(\epsilon, x) \cap A_{x_0} = \emptyset$ since otherwise this implies the existence of a polygonal path connecting x_0 to x. Thus $\mathsf{B}^n(\epsilon, x) \subseteq B_{x_0}$ and so B_{x_0} is indeed open.

We next claim that $\operatorname{bd}(B_{x_0}) \cap U = \emptyset$. Suppose otherwise and let $x \in \operatorname{bd}(B_{x_0}) \cap U$. Since $x \in U$ there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, x) \subseteq U$. Since $x \in \operatorname{bd}(B_{x_0})$ and by Proposition 4.2.26, there exists $x' \in \mathsf{B}^n(\epsilon, x)$ such that $x' \in B_{x_0}$. This means, as we have seen several times now, that x can be connected to x' by a segment. This means $x \notin A_{x_0}$

since otherwise this would imply the existence of a polygonal path from x_0 to x'. Thus $x \in B_{x_0} \cap \mathrm{bd}(B_{x_0})$, contradicting the openness of B_{x_0}.

Since $\mathrm{cl}(B_{x_0}) = B_{x_0} \cup \mathrm{bd}(B_{x_0})$ and $\mathrm{cl}(A) = A_{x_0} \cup \mathrm{bd}(A_{x_0})$, since $\mathrm{bd}(A_{x_0}) \cap U = \emptyset$, and since $\mathrm{bd}(B_{x_0}) \cap U = \emptyset$, it follows that $\mathrm{cl}(A_{x_0}) \cap B_{x_0} = \emptyset$ and $A_{x_0} \cap \mathrm{cl}(B_{x_0}) = \emptyset$. Thus, by assuming that B_{x_0} we show that U is a disjoint union of separated sets, contradicting the connectedness of U. Thus we must have $B_{x_0} = \emptyset$ and so $U = A_{x_0}$, as desired. ■

The preceding proposition implies the following interesting result.

4.2.48 Corollary (Open connected sets are differentiably path connected) *If* $\mathrm{U} \subseteq \mathbb{R}^n$ *is open and connected then, given* $\mathbf{x}_0, \mathbf{x}_1 \in \mathrm{U}$*, there exists a differentiable path lying in* U *connecting* $\mathbf{x}_0$ *and* $\mathbf{x}_1$*.*

Proof From Theorem 4.2.47 let γ be a polygonal path connecting x_0 and x_1. Let $y_1, \ldots, y_k \in U$ be the points at which γ is not differentiable, i.e., the "corner" points of the polygonal path. Now let $\epsilon \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^n(\epsilon, y_j) \subseteq U$ for each $j \in \{1, \ldots, k\}$. By Theorem **??** (or more precisely, by following the idea of the proof of that theorem as depicted in Figure **??**) there then exists a differentiable path γ_{diff} connecting x_0 with x_1 and that lies in U. ■

missing stuff

4.2.8 Subsets and relative topology

We have thus far been discussing properties of subsets of $\mathbb{R}^n$. However, sometimes it is useful to discuss subsets of subsets, and the properties of the smaller subset relative to the larger subset, not relative to $\mathbb{R}^n$. We shall revisit this idea in a more general (and in some sense, more suitable) setting in Section **??**; one way to think of this section is that it gives s gentle introduction to the more general material to come. We shall in this section make occasional and casual use of the terminology "relative topology," although it will not be defined until Section **??**.

Relatively open and closed sets

The key is the following definition.

4.2.49 Definition (Relatively open and closed subsets) Let $S \subseteq \mathbb{R}^n$ and let $A \subseteq S$.

(i) The subset $A \subseteq S$ is ***relatively open*** in S if, for every $x \in A$ there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, x) \cap S \subseteq A$.

(ii) The subset $A \subseteq S$ is ***relatively closed*** in S if $S \setminus A$ is relatively open in S. •

We shall often omit "in S" in "relatively open in S" when it is understood what set S is being used.

Let us characterise the notion of relatively open and relatively closed sets in a useful way.

4.2.50 Proposition (Characterisation of relatively open and closed subsets) *For* $\mathrm{S} \subseteq \mathbb{R}^n$ *and for* $\mathrm{A} \subseteq \mathrm{S}$ *the following statements hold:*

(i) A *is relatively open in* S *if and only if there exists an open subset* $U \subseteq \mathbb{R}^n$ *such that* $A = S \cap U$;

(ii) A *is relatively closed in* S *if and only if there exists a closed subset* $C \subseteq \mathbb{R}^n$ *such that* $A = S \cap C$.

Proof (i) Suppose that A is relatively open and let $\boldsymbol{x} \in A$. Let $\epsilon_{\boldsymbol{x}} \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^n(\epsilon_{\boldsymbol{x}}, \boldsymbol{x}) \cap S \subseteq A$. Then $U = \cup_{\boldsymbol{x} \in A} \mathsf{B}^n(\epsilon_{\boldsymbol{x}}, \boldsymbol{x})$ is open and has the property that $A = S \cap U$.

Conversely, let $A = S \cap U$ for an open set U. Then, for $\boldsymbol{x} \in A$ there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, \boldsymbol{x}) \subseteq U$. Therefore, $\mathsf{B}^n(\epsilon, \boldsymbol{x}) \cap S \subseteq U \cap S = A$.

(ii) Suppose that A is relatively closed so that $S \setminus A$ is relatively open. By the previous part of the result, $S \setminus A = S \cap U$ for an open subset $U \subseteq \mathbb{R}^n$. Thus

$$A = S \setminus (S \setminus A) = S \setminus (S \cap (U \cap A)) = S \cap (S \setminus (U \cap A)) = S \cap (\mathbb{R}^n \setminus U),$$

using DeMorgan's Laws. Taking $C = \mathbb{R}^n \setminus U$ gives the result.

Conversely, suppose that $A = S \cap C$ for a closed set C. Then $S \setminus A = (\mathbb{R}^n \setminus C) \cap S$ so that $S \setminus A$ is relatively open by the previous part of the result. Thus A is relatively closed. ■

These ideas of relatively open and closed subsets seems simple, but some care must be exercised in using them. Some examples illustrate the possible pitfalls.

4.2.51 Examples (Relatively open and closed subsets)

1. For any subset $S \subseteq \mathbb{R}^n$, the subset $S \subseteq S$ is always both relatively open and relatively closed. It is also true that $\emptyset \subseteq S$ is also both open and closed.
2. Let $S = (0, 1)$. Then, as in the preceding general example, $S \subseteq S$ is closed. Note, however, that S is not a closed subset of $\mathbb{R}$.
3. Let $S = [0, 1]$. Then $S \subseteq S$ is open although S is not an open subset of $\mathbb{R}$.
4. Let us consider $S = \mathbb{Z}$ as a subset of $\mathbb{R}$. We claim every subset of S is open. Indeed, let $A \subseteq \mathbb{Z}$ and let $x \in A$. Then $\mathsf{B}^n(\frac{1}{2}, x) \cap S = \{x\} \subseteq A$, showing that A is indeed open. A subset where every subset is open is called a ***discrete*** subset, and agrees with the usual notion of a discrete subset; see Exercise 4.2.7.
5. Let us examine $\mathbb{Q} \subseteq \mathbb{R}$, and consider some of its open and closed sets.
   (a) We claim that every singleton $\{q\} \subseteq \mathbb{Q}$ is not relatively open but is relatively closed. Since $\{q\} = \{q\} \cap \mathbb{Q}$, $\{q\}$ is relatively closed by Proposition 4.2.50. By Proposition 4.2.50 it follows that a relatively open subset of $\mathbb{Q}$ containing q must be of the form $U \cap \mathbb{Q}$ where U is an open subset of $\mathbb{R}$ containing q. Since U is a disjoint union of open intervals by Proposition 2.5.6, any relatively open subset of $\mathbb{Q}$ containing q will contain $(a, b)\mathbb{Q}$ for an open interval (a, b) containing q. However, *every* subset of $\mathbb{Q}$ of the form $(a, b) \cap \mathbb{Q}$ will contain infinitely many elements. Thus any relatively open subset of $\mathbb{Q}$ containing q will contain infinitely many elements. In particular, $\{q\}$ is not relatively open. Thus $\mathbb{Q}$ is not discrete.
   (b) We claim that for every $q \in \mathbb{Q}$ and for every $\epsilon \in \mathbb{R}_{>0}$ there exists a neighbourhood of q that is both open and closed and is contained in an interval of length at most ϵ. Indeed, let $r_1 \in (q - \frac{\epsilon}{2}, q)$ and $r_2 \in (q, q + \frac{\epsilon}{2})$ be irrational, this

being possible by Proposition 2.2.17. We claim that $(r_1, r_2) \cap \mathbb{Q}$ is both relatively open and relatively closed. It is relatively open by Proposition 4.2.50. Note that

$$\begin{aligned}\mathbb{Q} \setminus ((r_1, r_2) \cap \mathbb{Q}) &= ((-\infty, r_1] \cap \mathbb{Q}) \cup ([r_2, \infty) \cap \mathbb{Q}) \\ &= ((-\infty, r_1) \cap \mathbb{Q}) \cup ((r_2, \infty) \cap \mathbb{Q}),\end{aligned}$$

the latter inequality since r_1 and r_2 are irrational. This shows, by Proposition 4.2.50, that $\mathbb{Q} \setminus ((r_1, r_2) \cap \mathbb{Q})$ is open, and so $(r_1, r_2) \cap \mathbb{Q}$ is closed. •

One can, in the expected way, define the notion of a neighbourhood in this setup.

4.2.52 Definition (Relative neighbourhood) Let $S \subseteq \mathbb{R}^n$. A ***relative neighbourhood*** of $x \in S$ is a relatively open subset $U \subseteq S$ for which $x \in U$. More generally, a ***relative neighbourhood*** of $A \subseteq S$ is a relatively open set $U \subseteq S$ for which $A \subseteq U$. •

Many of the notions we have given above for subsets of $\mathbb{R}^n$ also apply to subsets of subsets of $\mathbb{R}^n$. For example...

4.2.53 Definition (Accumulation point, cluster point, limit point) For $S \subseteq \mathbb{R}^n$ and for $A \subseteq S$, a point $x \in S$ is:

(i) an ***accumulation point*** for A in S if, for every relative neighbourhood U of x, the set $A \cap (U \setminus x)$ is nonempty;

(ii) a ***cluster point*** for A in S if, for every relative neighbourhood U of x, the set $A \cap U$ is infinite;

(iii) a ***limit point*** of A in S if there exists a sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ in A converging to x.

The set of accumulation points of A in S is called the ***derived set*** of A, and is denoted by $\mathrm{der}_S(A)$. •

Relative interior, closure, and boundary

One can also define the notions of interior, closure, and boundary for subsets of subsets.

4.2.54 Definition (Relative interior, closure, and boundary) Let $S \subseteq \mathbb{R}^n$ and let $A \subseteq S$.

(i) The ***relative interior*** of A in S is the set

$$\mathrm{int}_S(A) = \cup\{U \mid U \subseteq A,\ U \text{ relatively open in } S\}.$$

(ii) The ***relative closure*** of A in S is the set

$$\mathrm{cl}_S(A) = \cap\{C \mid A \subseteq C \subseteq S,\ C \text{ relatively closed in } S\}.$$

(iii) The ***relative boundary*** of A in S is the set $\mathrm{bd}_S(A) = \mathrm{cl}_S(A) \cap \mathrm{cl}_S(S \setminus A)$. •

The properties of the interior, closure, and boundary given in Propositions 4.2.26, 4.2.27, 4.2.28, and 4.2.29 are also valid for the relative interior, relative closure, and relative boundary. Indeed, they are valid in the far more general context of topological spaces (see Chapter **??**). Thus we do not present the results here, but we shall occasionally use them.

Let us give some examples to illustrate that these notions should be thought about carefully in examples.

4.2.55 Examples (Relative interior, closure, and boundary)

1. If $S = (0, 1)$ then $\text{int}_S(S) = S$ and $\text{cl}_S(S) = S$ since S is both open and closed. Note, however, that $\text{cl}(S) = [0, 1]$. Also note that $\text{bd}_S(S) = \emptyset$ while $\text{bd}(S) = \{0, 1\}$.
2. If $S = [0, 1]$ then $\text{int}_S(S) = S$ and $\text{cl}_S(S) = S$ since S is both open and closed. Note, however, that $\text{int}(S) = (0, 1)$. Also note that $\text{bd}_S(S) = \emptyset$ while $\text{bd}(S) = \{0, 1\}$. •

For subsets the notion of denseness carries over in an obvious way.

4.2.56 Definition (Dense subset) If $A \subseteq \mathbb{R}^n$ a subset $D \subseteq A$ is ***dense*** in A if $\text{cl}_A(D) = A$. •

Some example illustrate the notion of dense subsets.

4.2.57 Examples (Dense subsets)

1. The set $\mathbb{Q} \cap [0, 1]$ is dense in $[0, 1]$.
2. The set $(0, 1)$ is dense in $[0, 1]$. •

Relatively compact sets

Let us now consider the matter of when a subset of a set is compact. The following definition is the obvious one.

4.2.58 Definition (Relatively compact) Let $A \subseteq \mathbb{R}^n$. A subset $K \subseteq A$ is ***relatively compact***[4] if, for every family $(U_i)_{i \in I}$ of relatively open subsets of A such that $K \subseteq \cup_{i \in I} U_i$, there exists $i_1, \ldots, i_k \in I$ such that $K \subseteq \cup_{j=1}^k U_{i_j}$. •

It turns out that this definition of relative compactness is the same as compactness in the usual sense.

4.2.59 Proposition (Characterisation of relatively compact sets) *Let* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$. *A subset* $\mathrm{K} \subseteq \mathrm{A}$ *is relatively compact if and only if* K *is compact as a subset of* $\mathbb{R}^{\mathrm{n}}$.

Proof First suppose that K is a relatively compact subset of A. Let $(U'_i)_{i \in I}$ be a family of open subsets of $\mathbb{R}^n$ such that $K \subseteq \cup_{i \in I} U'_i$. For each $i \in I$ define $U_i = U'_i \cap A$, noting that U_i is a relatively open subset of A by Proposition 4.2.50. Since $K \subseteq \cup_{i \in I} U_i$ and since K is relatively compact, there exists $i_1, \ldots, i_k \in I$ such that $K \subseteq \cup_{j=1}^k U_{i_j}$. Evidently $K \subseteq \cup_{j=1}^k U'_{i_j}$ and so K is a compact subset of $\mathbb{R}^n$.

[4] This is not the usual meaning given to the words "relatively compact." Most often, "relatively compact" is used to refer to what we call "precompact." However, we think that the meaning we give to "relatively compact" here is far more natural.

Next suppose that K is a compact subset of $\mathbb{R}^n$. Let $(U_i)_{i\in I}$ be a family of relatively open subsets of A such that $K \subseteq \cup_{i\in I} U_i$. By Proposition 4.2.50 let $(U'_i)_{i\in I}$ be a family of open subsets of $\mathbb{R}^n$ such that $U_i = U'_i \cap A$ for every $i \in I$. Clearly $K \subseteq \cup_{i\in I} U'_i$. Since K is compact there exists $i_1, \dots, i_k \in I$ such that $K \subseteq \cup_{j=1}^k U'_{i_j}$. By Proposition 1.1.7 we have

$$K \subseteq (\cup_{j=1}^k U'_{i_j}) \cap A = \cup_{j=1}^k U_{i_j},$$

showing that K is relatively compact. ■

Let us use the preceding result to characterise relatively compact subsets of $\mathbb{Q} \subseteq \mathbb{R}$.

4.2.60 Examples (Relatively compact subsets of $\mathbb{Q}$) Let us examine some properties of relatively compact subsets of $\mathbb{Q}$.

1. A finite subset $K \subseteq \mathbb{Q}$ is easily seen to be compact; see Exercise 4.2.9.
2. We claim that if $K \subseteq \mathbb{Q}$ is compact then K has an isolated point, i.e., there exists a point $q \in K$ and a neighbourhood U of q such that $U \cap K = \{q\}$. Indeed, suppose that K has no isolated points. Since finite subsets of $\mathbb{Q}$ are isolated and compact, we can consider the case when K is countable. Let us enumerate the points in K as $K = \{q_j\}_{j\in\mathbb{Z}_{>0}}$. Let us take $j_1 = 1$ and $p_1 = q_{j_1}$. As we saw in (4.2.51)–5, we can find a sufficiently small relatively closed relative neighbourhood U_1 of p_1 such that $K \not\subset U_1$. The subset $V_1 = K \setminus U_1$ is relatively open and relatively closed since U_1 is relatively open and relatively closed. Moreover, V_1 cannot be finite since K has no isolated points. Denote

$$j_2 = \min\{j \in \mathbb{Z}_{>0} \mid \ j > 1, \ q_j \notin U_1\}$$

and $p_2 = q_{j_2}$. Since $p_2 \in V_1$ and since V_1 is relatively open, by Proposition 4.2.50 we have that

$$\inf\{|p_2 - q| \mid \ q \in U_1\} > 0.$$

Therefore, again using the construction of (4.2.51)–5, there exists a sufficiently small relatively closed relative neighbourhood U_2 of p_2 such that $U_2 \cap U_1 = \emptyset$ and $V_1 \not\subset U_2$. Then define $V_2 = V_1 \setminus U_2$. Again, since K has no isolated points, V_2 is not finite. This process can be carried out to define a sequence $(j_k)_{j\in\mathbb{Z}_{>0}}$ of positive integers, a sequence $(p_k)_{k\in\mathbb{Z}_{>0}}$ of elements of K, and a sequence $(U_k)_{k\in\mathbb{Z}_{>0}}$ of pairwise disjoint subsets of K that are relatively open. We claim that $K \subseteq \cup_{k\in\mathbb{Z}_{>0}} U_k$. Indeed, suppose that $q_m \in K$ but $q_m \notin \cup_{k\in\mathbb{Z}_{>0}} U_k$ for some $m \in \mathbb{Z}_{>0}$. Denote

$$k_m = \min\{k \in \mathbb{Z}_{>0} \mid \ j_k > m\}.$$

Note that $q_m \notin \cup_{k=1}^{k_m-1} U_k$. However, the definition of k_m is that it is the smallest integer such that $q_{k_m} \notin \cup_{k=1}^{k_m-1} U_k$. Since $m < k_m$, we arrive at a contradiction. Thus the relatively open sets $(U_k)_{k\in\mathbb{Z}_{>0}}$ cover K, but clearly admit no finite subcover since they are pairwise disjoint. Thus subsets of $\mathbb{Q}$ with no isolated points cannot be compact.

3. The question raised by the previous two points is: "Are all relatively compact subsets of $\mathbb{Q}$ comprised only of isolated points, or, equivalently, are all relatively compact subsets of $\mathbb{Q}$ finite?" The answer is, "No." For example, the set

$$K = \{0\} \cup \{\tfrac{1}{k} \mid k \in \mathbb{Z}_{>0}\}$$

is relatively compact. To see this, by Proposition 4.2.59 and the Heine–Borel Theorem we need only show that it is closed and bounded as a subset of $\mathbb{R}$. It is clearly bounded. By Proposition 4.2.26 we can easily that $\mathrm{cl}(K) = K$ and so K is closed. Thus this is an example of a relatively compact subset of $\mathbb{Q}$ with a nonisolated point, since 0 is not isolated.

4. Finally, let us show that there are relatively compact subsets of $\mathbb{Q}$ having infinitely many nonisolated points. Let us define

$$K = \{0\} \cup \{\tfrac{1}{k} \mid k \in \mathbb{Z}_{>0}\} \cup \{\tfrac{1}{j} + \tfrac{1}{k} \mid j, k \in \mathbb{Z}_{>0}\}.$$

Let us first identity the accumulation points and limit of K.

1 Lemma *The set of accumulation points of* K *is* $\{0\} \cup \{\frac{1}{k}\}_{k \in \mathbb{Z}_{>0}}$ *and the set of limit points of* K *is* K.

Proof Let the sequence $(\frac{1}{j_l} + \frac{1}{k_l})_{l \in \mathbb{Z}_{>0}}$ converge to $r \in \mathbb{R}$. The sequence $(\frac{1}{j_l})_{l \in \mathbb{Z}_{>0}}$ has a convergent subsequence $(\frac{1}{j_{l_m}})_{m \in \mathbb{Z}_{>0}}$ since it is bounded (see Proposition 2.3.4). Since

$$\lim_{m \to \infty} \frac{1}{k_{l_m}} = r - \lim_{m \to \infty} \frac{1}{j_{l_m}},$$

the subsequence $(\frac{1}{k_{l_m}})_{m \in \mathbb{Z}_{>0}}$ also converges. By Proposition 2.3.23 we have

$$\lim_{m \to \infty} \frac{1}{j_{l_m}} = \frac{1}{\lim_{m \to \infty} j_{l_m}}.$$

There are two possible cases.

1. $\lim_{m \to \infty} j_{l_m} = \infty$: In this case $\lim_{m \to \infty} \frac{1}{j_{l_m}} = 0$.

2. $\lim_{m \to \infty} j_{l_m} \neq \infty$: In this case there must be a positive integer j_0 such that $\lim_{m \to \infty} j_{l_m} = j_0$. Thus $\lim_{m \to \infty} \frac{1}{j_{l_m}} = \frac{1}{j_0}$.

Similarly, either $\lim_{m \to \infty} \frac{1}{k_{l_m}} = 0$ or there exists $k_0 \in \mathbb{Z}_{>0}$ such that $\lim_{m \to \infty} \frac{1}{k_{l_m}} = \frac{1}{k_0}$. Thus, in all cases,

$$\lim_{m \to \infty} \Big(\frac{1}{j_{l_m}} + \frac{1}{k_{l_m}} \Big) \in K$$

and we conclude that the set of limit points of K is K, as claimed. The accumulation points of K arise as limits of sequences that are not eventually constant. From the various cases presented above, the converging subsequences that are not eventually constant arise when one or both of the cases

$$\lim_{m \to \infty} j_{l_m} = \infty, \quad \lim_{m \to \infty} k_{l_m} = \infty$$

occur. In this case,

$$\lim_{m\to\infty}\Big(\frac{1}{j_{l_m}}+\frac{1}{k_{l_m}}\Big)\in\{0\}\cup\{\tfrac{1}{k}\}_{k\in\mathbb{Z}_{>0}},$$

as desired. ▼

The lemma allows us to conclude that the set of nonisolated points of K is exactly $\{0\}\cup\{\frac{1}{k}\}_{k\in\mathbb{Z}_{>0}}$. Thus the set of nonisolated points is infinite. Moreover, K is closed (because every point is a limit point) and bounded, and hence compact.

3. We claim that relatively compact subsets of $\mathbb{Q}$ have empty relative interior. If a subset of $\mathbb{Q}$ has an nonempty interior, it must contain a nonempty relatively open subset. This means that it must contain a subset of the form $I\cap\mathbb{Q}$ where I is an open interval.
We claim that if $I\subseteq\mathbb{R}$ is an interval with a nonempty interior, then $I\cap\mathbb{Q}$ is not relatively compact in $\mathbb{Q}$. By the Bolzano–Weierstrass Theorem it suffices to show that there are sequences in $I\cap\mathbb{Q}$ that contain no subsequences converging in $I\cap\mathbb{Q}$. To exhibit such a sequence, let $r\in\operatorname{int}(I)$ be irrational and let $(q_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $I\cap\mathbb{Q}$ converging to r (by Proposition 2.2.15). Any subsequence of this sequence also converges to $r\notin I\cap\mathbb{Q}$. •

While we are talking about compactness, let us characterise compact subsets of $\mathbb{R}^n$ using relatively open and closed sets.

4.2.61 Proposition (Characterisation of compactness in terms of relatively open sets) *A subset $K\subseteq\mathbb{R}^n$ is compact if and only if, for every collection $(U_a)_{a\in A}$ of relatively open subsets of K for which $K=\cup_{a\in A}U_a$, there exists $a_1,\dots,a_k\in A$ such that $K=\cup_{j=1}^k U_{a_j}$.*

Proof First suppose that K is compact. For a collection $(U_a)_{a\in A}$ of relatively open subsets of K that covers K, let $V_a\subseteq\mathbb{R}^n$ be open and such that $U_a=K\cap V_a$, $a\in A$, using Proposition 4.2.50. Thus $(V_a)_{a\in A}$ is an open cover of K. Since K is compact there exists $a_1,\dots,a_k\in A$ such that $K\subseteq\cup_{j=1}^k V_{a_j}$. Thus

$$K=\cup_{j=1}^k(V_{a_j}\cap K)=\cup_{j=1}^k U_{a_j},$$

as desired.

For the converse, let $(V_a)_{a\in A}$ be an open cover of K so that $(U_a=V_a\cap K)_{a\in A}$ is a cover of K by relatively open sets by Proposition 4.2.50. Thus there exists $a_1,\dots,a_k\in A$ such that $K=\cup_{j=1}^k U_{a_j}$ and so $K\subseteq\cup_{j=1}^k V_{a_j}$. That is, K is compact. ■

It is also possible to characterise compactness deftly in terms of relatively closed sets.

4.2.62 Definition (Finite intersection property) Let $A\subseteq\mathbb{R}^n$ and let $(B_j)_{j\in J}$ be a family of subset of A. The family has the ***finite intersection property*** if, for any finite subset $\{j_1,\dots,j_k\}\subseteq J$, the set $\cap_{m=1}^k B_{j_m}\neq\emptyset$. •

We then have the following characterisation of compact sets.

4.2.63 Proposition (Compactness and the finite intersection property) *A subset* $\mathrm{K} \subseteq \mathbb{R}^{\mathrm{n}}$ *is compact if and only if every family* $(\mathrm{C_j})_{\mathrm{j}\in\mathrm{J}}$ *of relatively closed subsets of* K *with the finite intersection property has the property that* $\cap_{\mathrm{j}\in\mathrm{J}}\mathrm{C_j} \neq \emptyset$.

Proof Suppose that K is compact. Let $(C_j)_{j\in J}$ be a family of closed sets with the finite intersection property and suppose that $\cap_{j\in J}C_j = \emptyset$. Then we have

$$K = K \setminus (\cap_{j\in J}C_j) = \cup_{j\in J}(K \setminus C_j)$$

by DeMorgan's Laws. Then, since K is compact, there exists $j_1, \dots, j_k \in J$ such that $K = \cup_{m=1}^{k}(A \setminus C_{j_m})$. But this gives $K = K \setminus (\cap_{m=1}^{k}C_{j_m})$, again by DeMorgan's Laws. This means that $\cap_{m=1}^{k}C_{j_m} = \emptyset$, contradicting the finite intersection property of $(C_j)_{j\in J}$.

Conversely, suppose that $(U_j)_{j\in J}$ is an open cover of K and suppose that there is no finite subcover of this open cover. We claim that the family $(K \setminus U_j)_{j\in J}$ has the finite intersection property. Indeed, let $\{j_1, \dots, j_k\} \subseteq J$ so that

$$\cap_{m=1}^{k}(K \setminus U_{j_m}) = K \setminus (\cup_{m=1}^{k}U_{j_m}) \neq \emptyset$$

since $(U_j)_{j\in J}$ possesses no finite subcover. Now, for any finite subset $\{j_1, \dots, j_k\} \subseteq J$ we have

$$\emptyset \neq \cap_{j\in J}(K \setminus U_j) = K \setminus (\cup_{j\in J}U_j)$$

since $(K \setminus U_j)_{j\in J}$ has the finite intersection property. But this contradicts the fact that $(U_j)_{j\in J}$ covers K. ■

Connectedness using relative constructions

The use of relatively open and closed sets provides an elegant characterisation of connectedness. This characterisation will generalise to the notion of connectedness for general topological spaces in Section **??**.

4.2.64 Theorem (Characterisation of connectedness in terms of relative topology) *A subset* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *is connected if and only if the only subsets of* A *that are both relatively open and relatively closed in* A *are* $\emptyset$ *and* A.

Proof First suppose that A is disconnected so that $A = S \cup T$ for nonempty sets S and T with $\mathrm{cl}(S) \cap T = \emptyset$ and $S \cap \mathrm{cl}(T) = \emptyset$. Note that $S = A \cap \mathrm{cl}(S)$ since $S \subseteq \mathrm{cl}(S)$ and $\mathrm{cl}(S) \cap T = \emptyset$. By Proposition 4.2.50 this means that S is relative closed. In like manner T is relatively closed. Thus both S and T are also relatively open.

Now suppose that $S \subseteq A$ is relatively open and relatively closed, and that $S \neq A$ and $S \neq \emptyset$. Then $A = S \cup (A \setminus S)$ where S and $T \triangleq A \setminus S$ are both relatively open and relatively closed. We claim that $\mathrm{cl}(S) \cap T = \emptyset$. Indeed, if $\boldsymbol{x} \in T$ there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, \boldsymbol{x}) \cap A \subseteq T$ since T is relatively open. Since $S \cap T = \emptyset$ this implies that $\mathsf{B}^n(\epsilon, \boldsymbol{x}) \cap S = \emptyset$. By the analogue of Proposition 4.2.26 for the relative closure this implies that $\boldsymbol{x} \notin \mathrm{cl}(S)$. Thus we indeed have $\mathrm{cl}(S) \cap T = \emptyset$. The same argument gives $S \cap \mathrm{cl}(T) = \emptyset$ and so A is disconnected. ■

4.2.9 Local compactness

In this section we introduce the important idea of local compactness. This property turns out to be exactly what is needed for certain constructions. Our investigation here will be rather elementary. In Section **??** we give a deeper treatment of local compactness.

We begin with the definition.

4.2.65 Definition (Locally compact) A subset $A \subseteq \mathbb{R}^n$ is ***locally compact*** if, for every $x \in A$, there exists a relative neighbourhood $U \subseteq A$ of x such that $\mathrm{cl}_A(U)$ is a relatively compact subset of A. •

Let us give some examples and counterexamples.

4.2.66 Examples (Locally compact subsets)

1. We claim that every open subset U of $\mathbb{R}^n$ is locally compact. Indeed, let $x \in U$ and, since U is open, let $\epsilon \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^n(\epsilon, x) \subseteq U$. Then $\mathsf{B}(\frac{\epsilon}{2}, x) \subseteq U$ is a relative neighbourhood of x whose closure is a relatively compact subset of U.
2. We claim that every closed subset A of $\mathbb{R}^n$ is locally compact. Indeed, let $x \in A$, let $\epsilon \in \mathbb{R}_{>0}$, and denote $U = \mathsf{B}^n(\epsilon, x) \cap A$, noting that U is a relative neighbourhood of x by Proposition 4.2.50. We claim that
$$\mathrm{cl}_A(U) = \overline{\mathsf{B}}^n(\epsilon, x) \cap A. \tag{4.8}$$
Note that
$$U \subseteq \overline{\mathsf{B}}^n(\epsilon, x) \cap A \subseteq A,$$
the latter inclusion holding since A is closed. Thus
$$\overline{\mathsf{B}}^n(\epsilon, x) \cap A \subseteq \mathrm{cl}_A(U)$$
by definition of $\mathrm{cl}_A(U)$. The opposite inclusion holds by Proposition 4.2.28. Thus we have (4.8). By Proposition 4.2.59 we have that $\mathrm{cl}_A(U)$ is relatively compact in A. This shows that U is a relative neighbourhood of x possessing a relatively compact closure.
3. We claim that the subset $\mathbb{Q} \subseteq \mathbb{R}$ is not locally compact. Indeed, we showed in Example 4.2.60–3 that all relatively compact subsets of $\mathbb{Q}$ have empty relative interior.
4. Let
$$A = \{(0,0)\} \cup \{(x,y) \in \mathbb{R}^2 \mid x \in \mathbb{R}_{>0}\} \subseteq \mathbb{R}^2$$
(see Figure 4.6). We claim that A is not locally compact. Indeed, let $\epsilon \in \mathbb{R}_{>0}$ and let $U = U' \cap A$ (with $U' \subseteq \mathbb{R}^2$ a neighbourhood of $(0,0)$) be a relative neighbourhood of $(0,0)$. We claim that $\mathrm{cl}_A(U)$ is not compact.
Let $\epsilon \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^2(\epsilon, (0,0)) \subseteq U'$. For $j \in \mathbb{Z}_{>0}$ define an open subset $U'_j \subseteq \mathbb{R}^2$ by
$$U'_j = (\{(x,y) \in \mathbb{R}^2 \mid x > j^{-1}y,\ y \geq 0\} \cup \{(x,y) \in \mathbb{R}^2 \mid x > -j^{-1}y,\ y \leq 0\}) \cap \mathsf{B}^2(3\epsilon, (0,0))$$
(see Figure 4.7 for a depiction). Also let
$$U'_0 = \mathsf{B}^2(\frac{\epsilon}{2}, (0,0)), \quad V' = \mathbb{R}^2 \setminus \overline{\mathsf{B}}^2(2\epsilon, (0,0)).$$

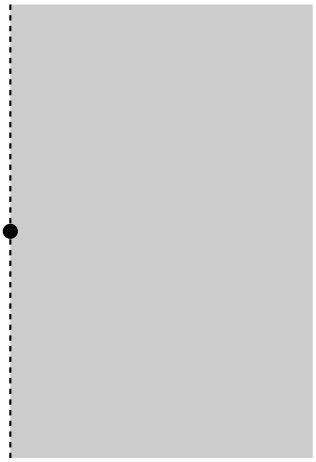


Figure 4.6 A subset of $\mathbb{R}^2$ that is not locally compact

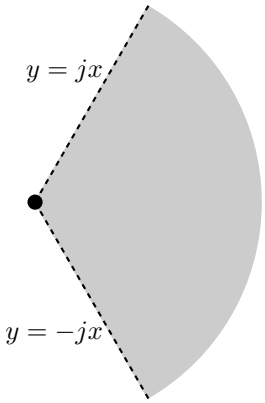


Figure 4.7 The open set U'_j

Note that

$$A \subseteq V' \cup U'_0 \cup_{j \in \mathbb{Z}_{>0}} U'_j.$$

Thus, if we define $U_j = U'_j \cap A$, $j \in \mathbb{Z}_{>0}$, $U_0 = U'_0 \cap A$, and $V = V' \cap A$, then

$$\mathrm{cl}_A(U) \subseteq V \cup U_0 \cup_{j \in \mathbb{Z}_{>0}} U_j.$$

We claim that there is no finite subset of the relatively open cover $\mathcal{O} = \{V\} \cup \{U_0\} \cup \{U_j\}_{j \in \mathbb{Z}_{>0}}$ that covers $\mathrm{cl}_A(U)$. Indeed, note that $\overline{\mathsf{B}}^2(\epsilon, (0,0)) \cap A \subseteq \mathrm{cl}_A(U)$. Therefore, any subset of $\mathcal{O}$ covering $\mathrm{cl}_A(U)$ must also cover $\overline{\mathsf{B}}^2(\epsilon, (0,0)) \cap A$. This, however, implies that all of the subsets U_j, $j \in \mathbb{Z}_{>0}$, must be contained in any subcover covering $\mathrm{cl}_A(U)$, and this ensures that $\mathrm{cl}_A(U)$ is not compact. •

4.2.10 Products of subsets

Next we consider subsets of Cartesian products of Euclidean spaces. Specifically, we consider sets of the form $A_1 \times \cdots \times A_k$ where $A_j \subseteq \mathbb{R}^{n_j}$, $j \in \{1, \ldots, k\}$.

For such subsets we shall give their properties in terms of properties of the subsets $A_1, \dots, A_k$. In studying these sets we make the natural identification of $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ with $\mathbb{R}^{n_1+\cdots+n_k}$ given by

$$\begin{aligned}\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \ni ((x_{1,1}, \dots, x_{1,n_1}), \dots, (x_{k,1}, \dots, x_{k,n_k})) \\ \mapsto (x_{1,1}, \dots, x_{1,n_1}, x_{2,1}, \dots, x_{k-1,n_{k-1}}, x_{k,1}, \dots, x_{k,n_k}) \in \mathbb{R}^{n_1+\cdots+n_k}.\end{aligned}$$

Thus, on $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ we shall use the Euclidean norm $\|\cdot\|_{\mathbb{R}^{n_1+\cdots+n_k}}$, and notions of openness, closedness, etc., will be derived from this. It is useful to relate this norm to the separate norms for $\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_k}$.

4.2.67 Lemma *For $\mathbf{x}_j \in \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$, we have*

$$\begin{aligned}\|\mathbf{x}_1\|_{\mathbb{R}^{n_1}} + \cdots + \|\mathbf{x}_k\|_{\mathbb{R}^{n_k}} \leq \sqrt{k}\|(\mathbf{x}_1, \dots, \mathbf{x}_k)\|_{\mathbb{R}^{n_1+\cdots+n_k}}, \\ \|(\mathbf{x}_1, \dots, \mathbf{x}_k)\|_{\mathbb{R}^{n_1+\cdots+n_k}} \leq \|\mathbf{x}_1\|_{\mathbb{R}^{n_1}} + \cdots + \|\mathbf{x}_k\|_{\mathbb{R}^{n_k}}.\end{aligned}$$

Proof Define

$$\delta_j = \|\boldsymbol{x}_j\|_{\mathbb{R}^{n_j}} - \tfrac{1}{k}(\|\boldsymbol{x}_1\|_{\mathbb{R}^{n_1}} + \cdots + \|\boldsymbol{x}_k\|_{\mathbb{R}^{n_k}}), \qquad j \in \{1, \dots, k\},$$

noting that $\delta_1 + \cdots + \delta_k = 0$ and that

$$\|\boldsymbol{x}_j\|_{\mathbb{R}^{n_j}} = \tfrac{1}{k}(\|\boldsymbol{x}_1\|_{\mathbb{R}^{n_1}} + \cdots + \|\boldsymbol{x}_k\|_{\mathbb{R}^{n_k}}) + \delta_j, \qquad j \in \{1, \dots, k\}.$$

A computation then gives

$$\begin{aligned}\|(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_{\mathbb{R}^{n_1+\cdots+n_k}} &= (\|\boldsymbol{x}_1\|^2_{\mathbb{R}^{n_1}} + \cdots + \|\boldsymbol{x}_k\|^2_{\mathbb{R}^{n_k}})^{1/2} \\ &= \left(\tfrac{1}{k}(\|\boldsymbol{x}_1\|_{\mathbb{R}^{n_1}} + \cdots + \|\boldsymbol{x}_k\|_{\mathbb{R}^{n_k}})^2 + \delta_1^2 + \cdots + \delta_k^2\right)^{1/2}\end{aligned}$$

which gives

$$\|(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_{\mathbb{R}^{n_1+\cdots+n_k}} \geq \tfrac{1}{\sqrt{k}}(\|\boldsymbol{x}_1\|_{\mathbb{R}^{n_1}} + \cdots + \|\boldsymbol{x}_k\|_{\mathbb{R}^{n_k}}),$$

as desired.

For the second inequality we have

$$\begin{aligned}\|(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)\|_{\mathbb{R}^{n_1+\cdots+n_k}} &= \|(\boldsymbol{x}_1, \boldsymbol{0}, \dots, \boldsymbol{0}) + (\boldsymbol{0}, , \dots, \boldsymbol{0}, \boldsymbol{x}_k)\|_{\mathbb{R}^{n_1+\cdots+n_k}} \\ &\leq \|(\boldsymbol{x}_1, \boldsymbol{0}, \dots, \boldsymbol{0})\|_{\mathbb{R}^{n_1+\cdots+n_k}} + \|(\boldsymbol{0}, \dots, \boldsymbol{0}, \boldsymbol{x}_k)\|_{\mathbb{R}^{n_1+\cdots+n_k}} \\ &= \|\boldsymbol{x}_1\|_{\mathbb{R}^{n_1}} + \cdots + \|\boldsymbol{x}_k\|_{\mathbb{R}^{n_k}},\end{aligned}$$

as desired. ■

Using these inequalities one can directly check that

$$\begin{aligned}\mathsf{B}^{n_1}(\epsilon, \boldsymbol{x}_1) \times \cdots \times \mathsf{B}^{n_k}(\epsilon, \boldsymbol{x}_k) \subseteq \mathsf{B}^{n_1+\cdots+n_k}(k\epsilon, (\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)), \\ \mathsf{B}^{n_1+\cdots+n_k}(\epsilon, (\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)) \subseteq \mathsf{B}^{n_1}(\sqrt{k}\epsilon, \boldsymbol{x}_1) \times \cdots \times \mathsf{B}^{n_k}(\sqrt{k}\epsilon, \boldsymbol{x}_k).\end{aligned} \tag{4.9}$$

The following theorem states the results in which we are interested.

4.2.68 Theorem (Properties of products derived from properties of components) *If* $A_j \subseteq \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$, *then the following statements hold:*

(i) $A_1 \times \cdots \times A_k$ *is open if and only if each of the sets* A_j, $j \in \{1, \dots, k\}$, *is open;*

(ii) $A_1 \times \cdots \times A_k$ *is closed if and only if each of the sets* A_j, $j \in \{1, \dots, k\}$, *is closed;*

(iii) $A_1 \times \cdots \times A_k$ *is compact if and only if each of the sets* A_j, $j \in \{1, \dots, k\}$, *is compact;*

(iv) $A_1 \times \cdots \times A_k$ *is connected if and only if each of the sets* A_j, $j \in \{1, \dots, k\}$, *is connected.*

Proof By an elementary induction argument in each case it suffices to prove the theorem in the case when $k = 2$. In this case, for simplicity of notation, we denote $n_1 = m$ and $n_2 = n$, and write a typical point in $\mathbb{R}^m \times \mathbb{R}^n$ as (x, y).

(i) Suppose that $A \times B$ is open and let $x_0 \in A$ and $y_0 \in B$. Since $A \times B$ is open there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^{m+n}(2\epsilon, (x_0, y_0)) \subseteq A \times B$. By (4.9) it follows that $\mathsf{B}^m(\epsilon, x_0) \times \mathsf{B}^n(\epsilon, y_0) \subseteq A \times B$ and so $\mathsf{B}^m(\epsilon, x_0) \subseteq A$ and $\mathsf{B}^n(\epsilon, y_0) \subseteq B$. Thus both A and B are open.

Now suppose that A and B are open and let $(x_0, y_0) \in A \times B$. Let $\epsilon \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^m(\sqrt{2}\epsilon, x_0) \subseteq A$ and $\mathsf{B}^n(\sqrt{2}\epsilon, y_0) \subseteq B$. Then $\mathsf{B}^m(\sqrt{2}\epsilon, x_0) \times \mathsf{B}^n(\sqrt{2}\epsilon, y_0) \subseteq A \times B$. By (4.9) it follows that $\mathsf{B}^{m+n}(\epsilon, (x_0, y_0)) \subseteq A \times B$ and so $A \times B$ is open.

(ii) Suppose that $A \times B$ is closed and let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in A that converges to some $x_0 \in \mathbb{R}^m$. We will show that $x_0 \in A$ which will show that A is closed by Proposition 4.2.26. Note that for $y_0 \in B$ the sequence $((x_j, y_0))_{j \in \mathbb{Z}}$ is in $A \times B$. Moreover, since

$$\|(x_j, y_0) - (x_0, y_0)\|_{\mathbb{R}^{m+n}} = \|x_j - x_0\|_{\mathbb{R}^m},$$

the sequence converges to (x_0, y_0). Since $A \times B$ is closed it follows that $(x_0, y_0) \in A \times B$ and so $x_0 \in B$, as desired.

Conversely, suppose that both A and B are closed. Then, by part (i), $(\mathbb{R}^m \setminus A) \times \mathbb{R}^n$ and $\mathbb{R}^m \times (\mathbb{R}^n \setminus B)$ are open and so too is their union. However,

$$(\mathbb{R}^m \times \mathbb{R}^n) \setminus (A \times B) = ((\mathbb{R}^m \setminus A) \times \mathbb{R}^n) \cup (\mathbb{R}^m \times (\mathbb{R}^n \setminus B))$$

and so $(\mathbb{R}^m \times \mathbb{R}^n) \setminus (A \times B)$ is open. Thus $A \times B$ is closed.

(iii) Suppose that $A \times B$ is compact, i.e., is closed and bounded by the Heine–Borel Theorem. Then A and B are closed by part (ii). Moreover, A and B are also bounded. Indeed, suppose that, say, A were unbounded and let $M \in \mathbb{R}_{>0}$. Then there exists $x_1, x_2 \in A$ such that $\|x_1 + x_2\|_{\mathbb{R}^m} \geq M$. Therefore, for $y \in B$ we have

$$\|(x_1, y) - (x_2, y)\|_{\mathbb{R}^{m+n}} \|x_1 - x_2\|_{\mathbb{R}^m} \geq M,$$

giving $A \times B$ as unbounded since $M \in \mathbb{R}_{>0}$ is arbitrary. Thus both A and B are closed and bounded, and so compact by the Heine–Borel Theorem.

Conversely, suppose that A and B are compact, i.e., closed and bounded by the Heine–Borel Theorem. Then $A \times B$ is closed by part (ii). To see that $A \times B$ is bounded, let $M \in \mathbb{R}_{>0}$ be such that

$$\|x_1 - x_2\|_{\mathbb{R}^m} < \tfrac{M}{2}, \quad \|y_1 - y_2\|_{\mathbb{R}^n} < \tfrac{M}{2}$$

for all $x_1, x_2 \in A$ and $y_1, y_2 \in B$. Then

$$\|(x_1, y_1) - (x_2, y_2)\|_{\mathbb{R}^{m+n}} \leq \|x_1 - x_2\|_{\mathbb{R}^m} + \|y_1 - y_2\|_{\mathbb{R}^n} < M,$$

using Lemma 4.2.67.

(iv) Suppose that A is not connected. Then $A = S \cup T$ where S and T are nonempty sets satisfying $\text{cl}(S) \cap T = \emptyset$ and $S \cap \text{cl}(T) = \emptyset$. Then $A \times B = (S \times B) \cup (T \times B)$. We claim that $\text{cl}(S \times B) \cap (T \times B) = \emptyset$. Let $((x_j, y_j))_{j \in \mathbb{Z}_{>0}}$ be a sequence in $S \times B$ converging to $(x_0, y_0) \in \text{cl}(S \times B)$. It is evident that $(x_j)_{j \in \mathbb{Z}_{>0}} \subseteq S$ converges to x_0 and so $x_0 \in \text{cl}(S)$. Therefore, $x_0 \notin T$ and so $(x_0, y_0) \notin T \times B$. Thus $\text{cl}(S \times B) \cap (T \times B) = \emptyset$, as claimed. We similarly show that $(S \times B) \cap \text{cl}(T \times B) = \emptyset$. This shows that $A \times B$ is disconnected if A is disconnected. Similarly one shows that $A \times B$ is disconnected if B is disconnected.

Now suppose that A and B are connected but that $A \times B$ are disconnected. Thus we suppose that $A \times B = S \cup T$ for nonempty sets S and T such that $\text{cl}(S) \cap T = \emptyset$ and $S \cap \text{cl}(T) = \emptyset$. Let $(x_1, y_1) \in S$ and $(x_2, y_2) \in T$. We claim that $\{x_1\} \times B$ and $A \times \{y_2\}$ are connected. This is clear since if, for example, $\{x_1\} \times B$ is disconnected then B is disconnected. Now note that $(\{x_1\} \times B) \cap (A \times \{y_2\}) \neq \emptyset$ since it contains the point (x_2, y_1). By Exercise 4.2.5 it follows that $X = (\{x_1\} \times B) \cup (A \times \{y_2\})$ is connected. However, this is a contradiction since the disconnectedness of $A \times B$ implies that

$$X = (X \cap S) \cup (X \cap T)$$

where $\text{cl}(X \cap S) \cap (X \cap T) = \emptyset$ and $(X \cap S) \cap \text{cl}(X \cap T) = \emptyset$. Thus it must be that $A \times B$ is connected. ■

These characterisations of products allows us to prove the following result.

4.2.69 Proposition (Interior, closure, and boundary of products) *If* $A \subseteq \mathbb{R}^m$ *and* $B \subseteq \mathbb{R}^n$ *then*

(i) $\text{int}(A \times B) = \text{int}(A) \times \text{int}(B)$,

(ii) $\text{cl}(A \times B) = \text{cl}(A) \times \text{cl}(B)$, *and*

(iii) $\text{bd}(A \times B) = (\text{bd}(A) \times \text{cl}(B)) \cup (\text{cl}(A) \times \text{bd}(B))$.

Proof (i) Since $\text{int}(A) \times \text{int}(B) \subseteq A \times B$ we have $\text{int}(A) \times \text{int}(B) \subseteq \text{int}(A \times B)$ by the definition of interior. Now let $(x, y) \in \text{int}(A \times B)$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^{m+n}(2\epsilon, (x, y)) \subseteq A \times B$. By (4.9) it then follows that

$$\mathsf{B}^m(\epsilon, x) \times \mathsf{B}^n(\epsilon, y) \subseteq A \times B,$$

and so $\mathsf{B}^m(\epsilon, x) \subseteq A$ and $\mathsf{B}^m(\epsilon, y) \subseteq B$. Thus $x \in \text{int}(A)$ and $y \in \text{int}(B)$.

(ii) Since $A \times B \subseteq \text{cl}(A) \times \text{cl}(B)$ and since $\text{cl}(A) \times \text{cl}(B)$ is closed by Theorem 4.2.68, it follows that $\text{cl}(A \times B) \subseteq \text{cl}(A) \times \text{cl}(B)$. Now let $(x, y) \in \text{cl}(A) \times \text{cl}(B)$. Then, for every $\epsilon \in \mathbb{R}_{>0}$ we have

$$\mathsf{B}^m(\tfrac{\epsilon}{2}, x) \cap A \neq \emptyset, \quad \mathsf{B}^n(\tfrac{\epsilon}{2}, y) \cap B \neq \emptyset \quad \implies \quad (\mathsf{B}^m(\tfrac{\epsilon}{2}, x) \times \mathsf{B}^n(\tfrac{\epsilon}{2}, y)) \cap (A \times B) \neq \emptyset.$$

Therefore, by (4.9) we have

$$\mathsf{B}^{m+n}(\epsilon, (x, y)) \cap (A \times B) \neq \emptyset.$$

Thus $(x, y) \in \text{cl}(A \times B)$ since this holds for every $\epsilon \in \mathbb{R}_{>0}$.

(iii) Let $(x, y) \in \text{bd}(A \times B)$. By Proposition 4.2.26 this means that for every $\epsilon \in \mathbb{R}_{>0}$

$$\mathsf{B}^{m+n}(\tfrac{\epsilon}{\sqrt{2}}, (x, y)) \cap (A \times B) \neq \emptyset, \quad \mathsf{B}^{m+n}(\tfrac{\epsilon}{\sqrt{2}}, (x, y)) \cap ((\mathbb{R}^n \times \mathbb{R}^m) \setminus (A \times B)) \neq \emptyset.$$

Therefore, by (4.9),

$$(\mathsf{B}^m(\epsilon, x) \times \mathsf{B}^n(\epsilon, y)) \cap (A \times B) \neq \emptyset, \quad (\mathsf{B}^m(\epsilon, x) \times \mathsf{B}^n(\epsilon, y)) \cap ((\mathbb{R}^n \times \mathbb{R}^m) \setminus (A \times B)) \neq \emptyset$$

for every $\epsilon \in \mathbb{R}_{>0}$. The condition

$$(\mathsf{B}^m(\epsilon, x) \times \mathsf{B}^n(\epsilon, y)) \cap (A \times B) \neq \emptyset$$

means that $x \in \mathrm{cl}(A)$ and $y \in \mathrm{cl}(B)$. Let us now these conditions along with the condition

$$(\mathsf{B}^m(\epsilon, x) \times \mathsf{B}^n(\epsilon, y)) \cap ((\mathbb{R}^n \times \mathbb{R}^m) \setminus (A \times B)) \neq \emptyset, \qquad \epsilon \in \mathbb{R}_{>0}.$$

This condition is exactly the condition that $(x, y) \in \mathrm{cl}((\mathbb{R}^n \times \mathbb{R}^m) \setminus (A \times B))$. We thus have the following possibilities.

1. $x \in \mathrm{cl}(A)$, $y \in \mathrm{cl}(B)$, $x \in A$, and $y \notin B$: In this case we must have $y \in \mathrm{bd}(\mathbb{R}^m \setminus B)$.
2. $x \in \mathrm{cl}(A)$, $y \in \mathrm{cl}(B)$, $x \in A$, and $y \in B$: In this case we cannot have $x \in \mathrm{int}(A)$ *and* $y \in \mathrm{int}(B)$ and so we must have either (a) $x \in \mathrm{bd}(A)$ and $y \in B$ or (b) $x \in B$ and $y \in \mathrm{bd}(B)$.
3. $x \in \mathrm{cl}(A)$, $y \in \mathrm{cl}(B)$, $x \notin A$, and $y \in A$: In this case we must have $x \in \mathrm{bd}(A)$.
4. $x \in \mathrm{cl}(A)$, $y \in \mathrm{cl}(B)$, $x \notin A$ and $y \notin B$: In this case we must have $x \in \mathrm{bd}(A)$ and $y \in \mathrm{bd}(B)$.

This means that we have either (1) $(x, y) \in \mathrm{bd}(A) \times \mathrm{cl}(B)$ or (2) $(x, y) \in \mathrm{cl}(A) \times \mathrm{bd}(B)$. Thus gives

$$\mathrm{bd}(A \times B) \subseteq (\mathrm{bd}(A) \times \mathrm{cl}(B)) \cup (\mathrm{cl}(A) \times \mathrm{bd}(B)).$$

Next suppose that $(x, y) \in \mathrm{bd}(A) \times \mathrm{cl}(B)$. This means that for every $\epsilon \in \mathbb{R}_{>0}$ the following sets are nonempty:

$$\mathsf{B}^m(\sqrt{2}\epsilon, x) \cap A, \quad \mathsf{B}^m(\sqrt{2}\epsilon, x) \cap (\mathbb{R}^n \setminus A), \quad \mathsf{B}^n(\sqrt{2}\epsilon, y) \cap B.$$

Thus take

$$x' \in \mathsf{B}^m(\sqrt{2}\epsilon, x) \cap A, \quad x'' \in \mathsf{B}^m(\sqrt{2}\epsilon, x) \cap (\mathbb{R}^n \setminus A), \quad y' \in \mathsf{B}^n(\sqrt{2}\epsilon, y) \cap B.$$

Then

$$\begin{aligned} & (x', y') \in (\mathsf{B}^m(\sqrt{2}\epsilon, x) \times \mathsf{B}^n(\sqrt{2}\epsilon, y)) \cap (A \times B) \\ \implies \quad & (x', y') \in \mathsf{B}^{m+n}(\epsilon, (x, y)) \cap (A \times B). \end{aligned}$$

Also

$$\begin{aligned} & (x'', y') \in (\mathsf{B}^m(\sqrt{2}\epsilon, x) \times \mathsf{B}^n(\sqrt{2}\epsilon, y)) \cap ((\mathbb{R}^m \setminus A) \times B) \\ \implies \quad & (x'', y') \in \mathsf{B}^{m+n}(\epsilon, (x, y)) \cap ((\mathbb{R}^m \setminus A) \times B) \\ \implies \quad & (x'', y') \in \mathsf{B}^{m+n}(\epsilon, (x, y)) \cap ((\mathbb{R}^m \times \mathbb{R}^m) \setminus (A \times B)). \end{aligned}$$

In like manner one shows that if $(x, y) \in \mathrm{cl}(A) \times \mathrm{bd}(B)$ then

$$\mathsf{B}^{m+n}(\epsilon, (x, y)) \cap (A \times B), \quad \mathsf{B}^{m+n}(\epsilon, (x, y)) \cap ((\mathbb{R}^m \times \mathbb{R}^m) \setminus (A \times B))$$

are nonempty. That is, for every $\epsilon \in \mathbb{R}_{>0}$ the sets

$$\mathsf{B}^{m+n}(\epsilon, (x, y)) \cap (A \times B), \quad \mathsf{B}^{m+n}(\epsilon, (x, y)) \cap ((\mathbb{R}^n \times \mathbb{R}^m) \setminus (A \times B))$$

are nonempty. Thus $(\mathrm{bd}(A) \times \mathrm{cl}(B)) \cup (\mathrm{cl}(A) \times \mathrm{bd}(B)) \subseteq \mathrm{bd}(A \times B)$. ∎

4.2.70 Remark (Finite Cartesian products) By an elementary induction argument, the first two statements in the preceding result carry over to finite Cartesian products of sets $A_1 \times \cdots \times A_k \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$. The generalisation of the third statement is tedious, but straightforward, and left to the reader. •

4.2.11 Sets of measure zero

One can also talk about subsets of $\mathbb{R}^n$ which have measure zero. This is done in the obvious way, using balls instead of intervals to cover sets. While the "volume" (i.e., length) of an interval is obviously defined, the volume of a ball in $\mathbb{R}^n$ is not so easily deduced. Let us here just define this volume, saving for ***missing stuff*** the calculations needed to verify the formula. Thus we denote by

$$\operatorname{vol}(\mathsf{B}^n(r, \mathbf{0})) = \frac{\pi^{n/2} r^n}{\Gamma(\frac{n}{2} + 1)} \tag{4.10}$$

volume of the ball of radius r, and we (reasonably) declare that the volume of a ball is independent of its centre. In the above formula, the function Γ (called, unsurprisingly, the **Γ-*function***) is defined by

$$\Gamma(x) = \int_0^\infty \mathrm{e}^{-y} y^{x-1} \, \mathrm{d}y.$$

This expression can be made more familiar by using property of the Γ-function that

$$\Gamma(\tfrac{k}{2} + 1) = \begin{cases} (\frac{k}{2})!, & k \text{ an even nonnegative integer,} \\ \frac{k! \pi^{1/2}}{2^k (\frac{k-1}{2})!}, & k \text{ an odd nonnegative integer.} \end{cases}$$

The reader is asked to explore some properties of the Γ-function in Exercise 4.2.16. In any case, we suppose that we know the volume of an n-dimensional ball.

With this we can make the following definition.

4.2.71 Definition (Set of measure zero) A subset $A \subseteq \mathbb{R}^n$ has ***measure zero*** if

$$\inf \Big\{ \sum_{j=1}^\infty \operatorname{vol}(\mathsf{B}^n(r_j, \boldsymbol{x}_j)) \;\Big|\; A \subseteq \bigcup_{j \in \mathbb{Z}_{>0}} \mathsf{B}^n(r_j, \boldsymbol{x}_j) \Big\} = 0. \quad \bullet$$

We refer the reader to Section 2.5.6 for examples of sets of zero measure, some interesting and some not. Ideas concerning the generalisation to $\mathbb{R}^n$ of sets of measure zero are discussed in Section **??**.

4.2.12 Convergence in $\mathbb{R}^n$-nets and a second glimpse of Landau symbols

In Section 2.3.7 we discussed convergence for generalisations of sequences where the index set is a subset of $\mathbb{R}$. In 2.3.8 we used this general notion of convergence to define Landau symbols. In this section we make a further generalisation to the case of generalised sequences where the index set is a subset of $\mathbb{R}^n$.

We begin by defining the sorts of directed sets we consider. The definition we give is a generalisation of that given for $\mathbb{R}$ in Section 2.3.7, but now we use the topology of $\mathbb{R}^n$ in a more fancy way.

4.2.72 Definition ($\mathbb{R}^n$-directed set) Let $A \subseteq \mathbb{R}^n$ and let $x_0 \in \mathbb{R}^n$.

(i) The ***$\mathbb{R}^n$-directed set in $\mathbf{A}$ at $\mathbf{x_0}$*** is the family of subsets

$$D(A, x_0) = \{U \cap A \mid U \subseteq \mathbb{R}^n \text{ open}, \ x_0 \in U\}$$

with the partial order $\supseteq$.

(ii) The ***$\mathbb{R}$-directed set in $\mathbf{A}$ at ∞*** is the family of subsets

$$D(A, \infty) = \{U \cap A \mid U \subseteq \mathbb{R}^n \text{ open}, \ \mathbb{R}^n \setminus \overline{\mathrm{B}}^n(R, \mathbf{0}) \subseteq U \text{ for some } R \in \mathbb{R}_{>0}\}$$

with the partial order $\supseteq$. •

Let us verify that $\mathbb{R}^n$-directed sets are indeed directed sets.

4.2.73 Proposition ($\mathbb{R}^n$-directed sets are directed sets) *If* $\mathrm{A} \subseteq \mathbb{R}^n$ *and if* $\mathbf{x}_0 \in \mathbb{R}^n$, *then* $(\mathrm{D}(\mathrm{A}, \mathbf{x}_0), \supseteq)$ *and* $(\mathrm{D}(\mathrm{A}, \infty), \supseteq)$ *are directed sets.*

Proof In the first case, let $U_1 \cap A, U_2 \cap A \in D(A, x_0)$ and note that, since $x_0 \in U_1 \cap A$ and $x_0 \in U_2 \cap A$, we have $U_1 \cap U_2$ is open and $x_0 \in (U_1 \cap U_2) \cap A$. Thus, $(U_1 \cap U_2) \cap A \in D(A, x_0)$ and

$$U_1 \cap A, U_2 \cap Q \supseteq (U_1 \cap U_2) \cap A.$$

In the second case, let $U_1 \cap A, U_2 \cap A \in D(A, \infty)$. Let $R_1, R_2 \in \mathbb{R}_{>0}$ be such that $\mathbb{R}^n \setminus \overline{\mathrm{B}}^n(R_1, \mathbf{0}) \subseteq U_1$ and $\mathbb{R}^n \setminus \overline{\mathrm{B}}^n(R_2, \mathbf{0}) \subseteq U_2$ and define $R = \max\{R_1, R_2\}$. Then $U_1 \cap U_2$ is open and $\mathbb{R}^n \setminus \overline{\mathrm{B}}^n(R, \mathbf{0}) \subseteq (U_1 \cap U_2) \cap A$. Thus $(U_1 \cap U_2) \cap A \in D(A, \infty)$ and

$$U_1 \cap A, U_2 \cap A \supseteq (U_1 \cap U_2) \cap A.$$

■

Now we define the sort of nets we consider in this case.

4.2.74 Definition ($\mathbb{R}^n$-net, convergence in $\mathbb{R}^n$-nets) Let $A \subseteq \mathbb{R}^n$, let $x_0 \in \mathbb{R}^n$, and let $D \in \{D(A, x_0), D(A, \infty)\}$. A ***$\mathbb{R}^n$-net*** in D is a map $\phi \colon A \to \mathbb{R}^m$ for some $m \in \mathbb{Z}_{>0}$. A $\mathbb{R}^n$-net $\phi \colon A \to \mathbb{R}^m$ in the $\mathbb{R}^n$-directed set D

(i) ***converges to $\mathbf{s_0} \in \mathbb{R}^m$*** if, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $U \cap A \in D$ such that, for any $V \cap A \in D$ for which $U \cap A \supseteq V \cap A$, $\|\phi(x) - s_0\|_{\mathbb{R}^m} < \epsilon$ for every $x \in V \cap A$;

(ii) has $s_0 \in \mathbb{R}^m$ as a ***limit*** if it converges to s_0, and we write $s_0 = \lim_D \phi$;

(iii) ***diverges*** if it does not converge,

(iv) has a limit that ***exists*** if $\lim_D \phi \in \mathbb{R}^m$, and

(v) is ***oscillatory*** if the limit of the $\mathbb{R}^n$-net does not exist, does not diverge to ∞, and does not diverge to $-\infty$. •

As with $\mathbb{R}$-nets, it is convenient to have some notation for $\mathbb{R}^n$-nets that allows us to understand more easily the sort of convergence that is taking place.

4.2.75 Notation (Limits of $\mathbb{R}^n$-nets) Let $A \subseteq \mathbb{R}^n$, let $x_0 \in \mathbb{R}^n$, let $D \in \{D(A, x_0), D(A, \infty)\}$, and let $\phi \colon A \to \mathbb{R}^m$ be a $\mathbb{R}^n$-net in D. Let us look at the two cases and give notation for each.

(i) $D = D(A, x_0)$: In this case we write $\lim_D \phi = \lim_{x \to_A x_0} \phi(x)$.

(ii) $D = D(A, \infty)$: In this case we write $\lim_D \phi = \lim_{x \to_A \infty} \phi(x)$. •

As with $\mathbb{R}$-nets, convergence in $\mathbb{R}^n$-nets can be characterised in terms of sequences in the case when x_0 is a limit of points in A.

4.2.76 Proposition (Convergence in $\mathbb{R}^n$-nets in terms of sequences) *Let* $A \subseteq \mathbb{R}^n$, *let* $x_0 \in \mathbb{R}^n$, *let* $D \in \{D(A, x_0), D(A, \infty)\}$, *and let* $\phi\colon A \to \mathbb{R}^m$ *be a* $\mathbb{R}^n$*-net in* D. *Then, corresponding to the two cases in Notation 4.2.75, we have the following statements:*

(i) if $x_0 \in \mathrm{cl}(A)$, *then the following statements are equivalent:*

(a) $\lim_{x \to_A x_0} \phi(x) = s_0$;

(b) $\lim_{j\to\infty} \phi(x_j) = s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *converging to* x_0;

(ii) if $\sup\{\|x\|_{\mathbb{R}^n} \mid x \in A\} = \infty$, *then the following statements are equivalent:*

(a) $\lim_{x \to_A \infty} \phi(x) = s_0$;

(b) $\lim_{j\to\infty} \|\phi(x_j)\|_{\mathbb{R}^m} = s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *such that* $\lim_{j\to\infty} \|x_j\|_{\mathbb{R}^n} = \infty$.

Proof For the first equivalence, suppose that $\lim_{x\to_A x_0} \phi(x) = s_0$ and let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in A converging to x_0. Let $\epsilon \in \mathbb{R}_{>0}$ and let $U \cap A \in D(A, x_0)$ be such that, for any $V \cap A \in D(A, x_0)$ for which $U \cap A \supseteq V \cap A$, we have $\|\phi(x) - s_0\|_{\mathbb{R}^m} < \epsilon$ for any $x \in V \cap A$. Let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $x_j \in U \cap A$ for every $j \geq N$, this being possible since $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to x_0. Now note that $\|\phi(x_j) - s_0\|_{\mathbb{R}^m} < \epsilon$ for every $j \geq N$ since $x_j \in U \cap A$ for every $j \geq N$. This gives $\lim_{j\to\infty} \phi(x_j) = s_0$, as desired.

For the converse, suppose that $\lim_{x\to_A x_0} \phi(x) \neq s_0$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that, for any $U \cap A \in D(A, x_0)$, we have a $V \cap A \in D(A, x_0)$ with $U \cap A \supseteq V \cap A$ for which $\|\phi(x) - s_0\|_{\mathbb{R}^m} \geq \epsilon$ for some $x \in V \cap A$. Since $x_0 \in \mathrm{cl}(A)$ it follows that, for any $j \in \mathbb{Z}_{>0}$, there exists $x_j \in \mathsf{B}^n(\frac{1}{j}, x_0) \cap A$ such that $\|\phi(x_j) - s_0\|_{\mathbb{R}^m} \geq \epsilon$. Thus the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in A converging to x_0 has the property that $(\phi(x_j))_{j\in\mathbb{Z}_{>0}}$ does not converge to s_0.

For the second equivalence, suppose that $\lim_{x\to_A \infty} \phi(x) = s_0$ and let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in A such that $\lim_{j\to\infty} \|x_j\|_{\mathbb{R}^n} = \infty$. Let $M \in \mathbb{R}_{>0}$ and let $U \cap A \in D(A, \infty)$ be such that, for any $V \cap A \in D(A, \infty)$ for which $U \cap A \supseteq V \cap A$, we have $\|\phi(x) - s_0\|_{\mathbb{R}^m} < \epsilon$ for every $x \in V \cap A$. Let $R \in \mathbb{R}_{>0}$ be such that $\mathbb{R}^n \setminus \overline{\mathsf{B}}^n(R, \mathbf{0}) \subseteq U$ and let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $\|x_j\|_{\mathbb{R}^n} > R$ for $j \geq N$. It then follows that $\|\phi(x_j) - s_0\|_{\mathbb{R}^m} < \epsilon$ for every $j \geq N$ since $x_j \in U \cap A$. Thus $\lim_{j\to\infty} \phi(x_j) = s_0$.

For the converse, suppose that $\lim_{x\to_A \infty} \phi(x)$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that, for any $U \cap A \in D(A, \infty)$, we have a $V \cap A \in D(A, \infty)$ with $U \cap A \supseteq V \cap A$ for which $\|\phi(x) - s_0\|_{\mathbb{R}^m} \geq \epsilon$ for some $x \in V \cap A$. By our assumption that A is unbounded, it follows that, for any $j \in \mathbb{Z}_{>0}$, there exists $x_j \in (\mathbb{R}^n \setminus \overline{\mathsf{B}}^n(j, x_0)) \cap A$ such that $\|\phi(x_j) - s_0\|_{\mathbb{R}^m} \geq \epsilon$. Thus the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in A for which $(\|x_j\|_{\mathbb{R}^n})_{j\in\mathbb{Z}_{>0}}$ diverges to ∞ has the property that $(\phi(x_j))_{j\in\mathbb{Z}_{>0}}$ does not converge to s_0. ■

From the preceding result, we can easily establish the equivalence of convergence of $\mathbb{R}^n$-nets with $n = 1$ with convergence of $\mathbb{R}$-nets from Section 2.3.7.

Now let us give some examples to make the preceding construction concrete.

4.2.77 Examples (Convergence in $\mathbb{R}^n$-nets) In the examples below we will simply give "the answer," leaving to the reader the mundane details of verification.

1. Define $\phi\colon \mathbb{R}^n \to \mathbb{R}$ by
$$\phi(x) = \frac{1}{1 + \|x\|^2_{\mathbb{R}^n}}.$$
If we think of ϕ as a $\mathbb{R}^n$-net in $D = D(\mathbb{R}^n, \mathbf{0})$ then $\lim_D \phi = 1$. If we think of ϕ as a $\mathbb{R}^n$-net in $D = D(\mathbb{R}^n, \infty)$ then $\lim_D \phi = 0$.

2. Define $\phi\colon \mathbb{R}^n \to \mathbb{R}$ by $\phi(x) = \sin(\|x\|_{\mathbb{R}^n})$. If we think of ϕ as a $\mathbb{R}^n$-net in $D = D(\mathbb{R}^n, \mathbf{0})$ then $\lim_D \phi = 1$. If we think of ϕ as a $\mathbb{R}^n$-net in $D = D(\mathbb{R}^n, \infty)$ then $\lim_D \phi$ does not exist. •

There are also generalisations of lim sup and lim inf to $\mathbb{R}^n$-nets. We let $A \subseteq \mathbb{R}^n$, $x_0 \in \mathbb{R}^n$, $D =\in \{D(A, x_0), D(A, \infty)\}$, and $\phi\colon A \to \mathbb{R}$. We denote by $\sup_D \phi, \inf_D \phi\colon A \to \mathbb{R}$ the $\mathbb{R}$-nets in D given by

$$\sup_D \phi(x) = \sup\{\phi(y) \mid y \in U \cap A \text{ for all } U \cap A \in D \text{ for which } x \in U \cap A\},$$
$$\inf_D \phi(x) = \inf\{\phi(y) \mid y \in U \cap A \text{ for all } U \cap A \in D \text{ for which } x \in U \cap A\}.$$

Then we define

$$\limsup_D \phi = \lim_D \sup_D \phi, \quad \liminf_D \phi = \lim_D \inf_D \phi.$$

Let us now adapt our notion of Landau symbols from Section 2.3.8 to $\mathbb{R}^n$-nets.

4.2.78 Definition (Landau symbols "O" and "o") Let $A \subseteq \mathbb{R}^n$, let $x_0 \in \mathbb{R}^n$, let $D \in \{D(A, x_0), D(A, \infty)\}$ be a $\mathbb{R}^n$-directed set, and let $\phi\colon A \to \mathbb{R}$.

(i) Denote by $O_D(\phi)$ the functions $\psi\colon A \to \mathbb{R}^m$ for which there exists $U \cap A \in D$ and $M \in \mathbb{R}_{>0}$ such that, for every $V \cap A \in D$ for which $U \cap A \supseteq V \cap A$, $\|\psi(x)\|_{\mathbb{R}^m} \leq M|\phi(x)|$ for every $x \in V \cap A$.

(ii) Denote by $o_D(\phi)$ the functions $\psi\colon A \to \mathbb{R}$ such that, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $U \cap A \in D$ such that $\|\psi(x)\|_{\mathbb{R}^m} < \epsilon|\phi(x)|$ for $x \in V \cap A$.

If $\psi \in O_D(\phi)$ (resp. $\psi \in o_D(\phi)$) then we say that ψ is ***big oh of ϕ*** (resp. ***little oh of ϕ***). •

It is often the case that the comparison function ϕ is positive on A. In such cases, one can give a somewhat more concrete characterisation of O_D and o_D.

4.2.79 Proposition (Alternative characterisation of Landau symbols) *Let* $A \subseteq \mathbb{R}^n$, *let* $x_0 \in \mathbb{R}^n$, *let* $D \in \{D(A, x_0), D(A, \infty)\}$ *be a* $\mathbb{R}^n$*-directed set, and let* $\phi\colon A \to \mathbb{R}_{>0}$ *and* $\psi\colon A \to \mathbb{R}$. *Then*

(i) $\psi \in O_D(\phi)$ *if and only if* $\limsup_D \frac{\|\psi\|_{\mathbb{R}^m}}{\phi} < \infty$ *and*

(ii) $\psi \in o_D(\phi)$ *if and only if* $\lim_D \frac{\|\psi\|_{\mathbb{R}^m}}{\phi} = 0$.

Proof We leave this as Exercise 4.2.15. ■

4.2.80 Examples (Landau symbols)

1. Generalising what we saw in Example 2.3.34 for differentiability of $\mathbb{R}$-valued functions defined on intervals, let $U \subseteq \mathbb{R}^n$ be open, let $x_0 \in U$, and let $f\colon U \to \mathbb{R}^m$. Let $k \in \mathbb{Z}_{\geq 0}$ and for $A_j \in S^j(\mathbb{R}^n; \mathbb{R}^m)$, $j \in \{0, 1, \ldots, k\}$, define $g_{f,x_0,A}\colon U \to \mathbb{R}^m$ by

$$g_{f,x_0,A}(x) = \frac{A_0}{0!} + \frac{A_1(x)}{1!} + \frac{A_2(x,x)}{2!} + \cdots + \frac{A_k(x,\ldots,x)}{k!}.$$

Define a $\mathbb{R}^n$-net in $D = D(U, x_0)$ by $\phi_k(x) = \|x - x_0\|^k_{\mathbb{R}^n}$. Then one can verify (this is Taylor's Theorem) that f is k-times continuously differentiable at x_0 with $D^j f(x_0) = A_j$, $j \in \{0, 1, \ldots, k\}$, if and only if $\|f - g_{f,x_0,A}\|_{\mathbb{R}^m} \in o_D(\phi_m)$.

Exercises

4.2.1 Show that

$$\left\|\sum_{j=1}^{m} \boldsymbol{x}_j\right\|_{\mathbb{R}^n} \leq \sum_{j=1}^{m} \|\boldsymbol{x}_j\|_{\mathbb{R}^n}$$

for any finite family $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ in $\mathbb{R}^n$.

4.2.2 Let $A \subseteq \mathbb{R}^n$ be closed and let $(\boldsymbol{x}_j)_{j \in \mathbb{Z}_{>0}}$ be a Cauchy sequence. Show that the sequence converges to a point in A.

4.2.3 Prove Proposition 4.2.19.

4.2.4 Show that a subset $C \subseteq \mathbb{R}^n$ is closed if and only if $C \cap K$ is closed for every compact subset K of $\mathbb{R}^n$.

4.2.5 Let $(A_i)_{i \in I}$ be a family of connected subsets of $\mathbb{R}^n$ and suppose that $\cap_{i \in I} A_i \neq \emptyset$. Show that $\cup_{i \in I} A_i$ is connected.

4.2.6 Show that the closure of a connected set is connected.

4.2.7 Show that for a subset $D \subseteq \mathbb{R}^n$ the following two statements are equivalent:

1. D is discrete, i.e., every subset of D is relatively open in D;
2. there exists $\epsilon \in \mathbb{R}_{>0}$ such that, for every $\boldsymbol{x} \in D$, $\mathsf{B}^n(\epsilon, \boldsymbol{x}) \cap D = \{\boldsymbol{x}\}$.

4.2.8 Let $U \subseteq \mathbb{R}^n$ be open and $C \subseteq \mathbb{R}^n$ be closed.

(a) If $A \subseteq U$ show that $\text{int}_U(A) = \text{int}(A)$.

(b) If $A \subseteq C$ show that $\text{cl}_C(A) = \text{int}(A)$.

4.2.9 Show that finite subsets of $\mathbb{Q}$ are relatively compact.

4.2.10 Show that if $r, s \in \mathbb{R}_{>0}$, and $\boldsymbol{x}_0 \in \mathbb{R}^m$ and $\boldsymbol{y}_0 \in \mathbb{R}^n$, then $\mathsf{B}^m(r, \boldsymbol{x}_0) \times \mathsf{B}^n(s, \boldsymbol{y}_0)$ is an open subset of $\mathbb{R}^n \times \mathbb{R}^m$.

4.2.11 Let $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^n$. Show that a sequence $((\boldsymbol{x}_j, \boldsymbol{y}_j))_{j \in \mathbb{Z}_{>0}}$ converges to $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ if and only if $(\boldsymbol{x}_j)_{j \in \mathbb{Z}_{>0}}$ and $(\boldsymbol{y}_j)_{j \in \mathbb{Z}_{>0}}$ converge to $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$, respectively.

4.2.12 Let $(Z_j)_{j \in \mathbb{Z}_{>0}}$ be a family of subsets of $\mathbb{R}^n$ that each have measure zero. Show that $\cup_{j \in \mathbb{Z}_{>0}} Z_j$ also has measure zero.

4.2.13 If $\mathsf{V} \subseteq \mathbb{R}^n$ is a subspace of dimension at most $n-1$ show that V has measure zero.

4.2.14 Let $D \in \{D(A, \boldsymbol{x}_0), D(A, \infty)\}$ be a $\mathbb{R}^n$-directed set and let $\boldsymbol{\phi} \colon A \to \mathbb{R}^m$ be a $\mathbb{R}^n$-net in D. For $\boldsymbol{s}_0 \in \mathbb{R}^m$ define the corresponding $\mathbb{R}^n$ net $\boldsymbol{\phi}_{\boldsymbol{x}_0, \boldsymbol{s}_0} \colon A \to \mathbb{R}_{\geq 0}$ by $\boldsymbol{\phi}_{\boldsymbol{x}_0, \boldsymbol{s}_0}(\boldsymbol{x}) = \|\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{s}_0\|_{\mathbb{R}^m}$. Show that $\lim_D \boldsymbol{\phi} = \boldsymbol{s}_0$ if and only if $\lim_D \boldsymbol{\phi}_{\boldsymbol{x}_0, \boldsymbol{s}_0} = 0$.

4.2.15 Prove Proposition 4.2.79.

4.2.16

Section 4.3

Continuous functions of multiple variables

With the structure of $\mathbb{R}^n$ as given in Section 4.2 it is fairly easy to generalise the notion of continuity from the single-variable case to the multivariable case. Thus much of what we say in this section bears a strong resemblance to the material in Section 3.1. We do, however, add more depth and detail in this section than we did in Section 3.1. For example, we discuss the structure of linear maps, affine maps, isometries of $\mathbb{R}^n$, and homeomorphisms. Reading this section will be excellent preparation for understanding the general notion of a continuous map and its properties as presented in Section **??**.

Since this section does repeat some of the material from Section 3.1, we omit reproducing the illustrative examples that we have already given, and only give examples that reveal something interesting about the multivariable case.

Do I need to read this section? If one is reading this chapter then one should read this section. Certain of the sections can be skipped, and these are clearly labelled. •

4.3.1 Definition and properties of continuous multivariable maps

First let us establish our notation for multivariable functions. If $A \subseteq \mathbb{R}^n$ we use a bold font, $\boldsymbol{f} \colon A \to \mathbb{R}^m$ to represent a multivariable function on A, reflecting the fact that we use a similar bold font to denote points in $\mathbb{R}^n$ for $n > 1$. In keeping with this convention, we will denote by $f \colon A \to \mathbb{R}$ a typical function taking values in $\mathbb{R}$, even though the domain is multi-dimensional. Note that, since $\boldsymbol{f} \colon A \to \mathbb{R}^m$ takes values in $\mathbb{R}^m$ we can write

$$\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \dots, f_m(\boldsymbol{x})),$$

where the functions $f_j \colon A \to \mathbb{R}$, $j \in \{1, \dots, m\}$, are the ***components*** of $\boldsymbol{f}$.

If a function $\boldsymbol{f} \colon A \to \mathbb{R}^m$ takes values in $B \subseteq \mathbb{R}^m$ we may write $\boldsymbol{f} \colon A \to B$.

The definition of continuity for $\mathbb{R}$-values functions on $\mathbb{R}$ is made using the absolute value function $|\cdot|$ on $\mathbb{R}$ in an essential way. Since the Euclidean norm $\|\cdot\|_{\mathbb{R}^n}$ provides a generalisation of the absolute value function, we shall use this to extend to multiple dimensions our definitions of continuity.

4.3.1 Definition (Continuous map) Let $n, m \in \mathbb{Z}_{>0}$ and let $A \subseteq \mathbb{R}^n$ be a subset. A map $\boldsymbol{f} \colon A \to \mathbb{R}^m$ is:

(i) ***continuous at*** $\mathbf{x_0} \in \mathbf{A}$ if, for every $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that $\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_0)\|_{\mathbb{R}^n} < \epsilon$ whenever $\boldsymbol{x} \in A$ satisfies $\|\boldsymbol{x} - \boldsymbol{x}_0\|_{\mathbb{R}^m} < \delta$;

(ii) ***continuous*** if it is continuous at each $\boldsymbol{x}_0 \in A$;

(iii) ***discontinuous at*** $\mathbf{x_0} \in \mathbf{A}$ if it is not continuous at $\boldsymbol{x}_0$;

(iv) ***discontinuous*** if it is not continuous.

Note that if f takes values in $B \subseteq \mathbb{R}^m$ we shall say that $f\colon A \to B$ is continuous if it is continuous as a map into $\mathbb{R}^m$, i.e., if the map $i_B \circ f$ is continuous, where i_B is the inclusion of B into $\mathbb{R}^m$. •

Note that we define continuity for multivariable maps defined on *arbitrary* subsets of $\mathbb{R}^n$, whereas for the single-variable case we only considered functions defined on intervals. We do this principally because there is no really useful generalisation to higher-dimensions of the notion of an interval. We will mostly only use fairly well-behaved subsets of $\mathbb{R}^n$, e.g., open sets, or closures of open sets, although our definition allows rather degenerate domains for maps.

The following equivalent characterisations of continuity, except for the last, are just as they are in the case when $m = n = 1$, and, indeed, the proof also generalises the one-dimensional proof only by replacing open intervals by open balls. Here, for simplicity, we only consider maps whose domain is an open set (see Example **??–??** for the definition of an open set in this case).

4.3.2 Theorem (Alternative characterisations of continuity) *For a map* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}$ *defined on a subset* $\mathrm{A} \subseteq \mathbb{R}^n$ *and for* $\mathbf{x}_0 \in \mathrm{A}$*, the following statements are equivalent:*

(i) $\mathbf{f}$ *is continuous at* $\mathbf{x}_0$*;*

(ii) for every neighbourhood V *of* $\mathbf{f}(\mathbf{x}_0)$ *there exists a neighbourhood* U *of* $\mathbf{x}_0$ *such that* $\mathbf{f}(\mathrm{U} \cap \mathrm{A}) \subseteq \mathrm{V}$*;*

(iii) $\lim_{\mathbf{x} \to_{\mathrm{A}} \mathbf{x}_0} \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0)$*;*

(iv) the components of $\mathbf{f}$ *are continuous at* $\mathbf{x}_0$*.*

Proof We shall show the equivalence of the first three statements, leaving the last as Exercise 4.3.2.

(i) $\implies$ (ii) Let $V \subseteq \mathbb{R}^m$ be a neighbourhood of $f(x_0)$ and let $\epsilon \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^m(\epsilon, f(x_0)) \subseteq V$. Then, by continuity of f, let $\delta \in \mathbb{R}_{>0}$ be such that $\|f(x) - f(x_0)\|_{\mathbb{R}^m} < \epsilon$ if $x \in A$ satisfies $\|x - x_0\|_{\mathbb{R}^n} < \delta$. That is, if $U = \mathsf{B}^n(\delta, x_0)$ then $f(U \cap A) \subseteq \mathsf{B}^m(\epsilon, f(x_0)) \subseteq V$.

(ii) $\implies$ (iii) Let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in A converging to x_0. For $\epsilon \in \mathbb{R}_{>0}$ let U be a neighbourhood of x_0 such that $f(U \cap A) \subseteq \mathsf{B}^m(\epsilon, f(x_0))$. Now let $\delta \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^n(\delta, x_0) \subseteq U$ and let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $\|x_j - x_0\|_{\mathbb{R}^n} < \delta$ for $j \geq N$. Then, for $j \geq N$, $f(x_j) \subseteq \mathsf{B}^m(\epsilon, f(x_0))$, i.e., $\|f(x_j) - f(x_0)\|_{\mathbb{R}^m} < \epsilon$ for $j \geq N$. Thus $(f(x_0))_{j \in \mathbb{Z}_{>0}}$ converges to $f(x_0)$.

(iii) $\implies$ (i) Suppose that f is not continuous at x_0. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that, for any $\delta \in \mathbb{R}_{>0}$, $f(\mathsf{B}^n(\delta, x_0) \cap A) \not\subset \mathsf{B}^m(\epsilon, f(x_0))$. For each $j \in \mathbb{Z}_{>0}$, therefore, let $x_j \in A$ satisfy $\|x_j - x_0\|_{\mathbb{R}^n} < \frac{1}{j}$ and $f(x_j) \notin \mathsf{B}^m(\epsilon, f(x_0))$. Then the sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ converges to x_0 but the sequence $(f(x_j))_{j \in \mathbb{Z}_{>0}}$ does not converge to $f(x_0)$. ■

Note that the last part of the preceding theorem says that "f is continuous if and only if its components are continuous." This is not to be confused with the incorrect statement that "f is continuous if and only if it is a continuous function of each component." The following example illustrates the distinction.

4.3.3 Example (A discontinuous function that is continuous in each of its vari-

ables) Consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{x_1^2 + x_2^2}, & (x_1, x_2) \neq (0, 0), \\ 0, & (0, 0). \end{cases}$$

We first claim that this function is discontinuous at $(0, 0)$. Indeed, consider points in $\mathbb{R}^2$ of the form (a, a) for $a \in \mathbb{R}^*$. At such points we have $f(a, a) = \frac{1}{2}$. Since $f(0, 0) = 0$ and since every neighbourhood of $(0, 0)$ contains a point of the form (a, a) for some $a \in \mathbb{R}^*$, it follows that f cannot be continuous at $(0, 0)$.

We also claim that for fixed $x_{10} \in \mathbb{R}$ (resp. $x_{20} \in \mathbb{R}$) the function $x_2 \mapsto f(x_{10}, x_2)$ (resp. $x_1 \mapsto f(x_1, x_{20})$) is continuous. First fix $x_{10} \in \mathbb{R}^*$. Then the function $x_2 \mapsto \frac{x_{10} x_2}{x_{10}^2 + x_2^2}$ is clearly continuous (since the denominator is nonzero and since sums, products, and quotients by nonzero functions preserve continuity). If $x_{10} = 0$ then we have $f(x_{10}, x_2) = 0$ for all $x_2 \in \mathbb{R}$, and this is obviously a continuous function. This shows that $x_2 \mapsto f(x_{10}, x_2)$ is continuous for every $x_{10} \in \mathbb{R}$. An entirely similar argument shows that $x_1 \mapsto f(x_1, x_{20})$ is continuous for all $x_{20} \in \mathbb{R}$. •

The previous theorem also has the following useful restatement which employs the relative topology discussed in Section 4.2.8.

4.3.4 Corollary (Characterisation of continuous maps) *For* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *and for* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}}$ *the following statements are equivalent:*

(i) $\mathbf{f}$ *is continuous;*

(ii) $\mathbf{f}^{-1}(\mathrm{V})$ *is relatively open in* A *for every open subset* V *of* $\mathbb{R}^{\mathrm{m}}$.

Proof First suppose that f is continuous and let $V \subseteq \mathbb{R}^m$ be open. Let $\boldsymbol{x}_0 \in f^{-1}(V)$ so that $f(\boldsymbol{x}_0) \in V$. Since V is open and so a neighbourhood of $f(\boldsymbol{x}_0)$, by Theorem 4.3.2 there exists a neighbourhood U of $\boldsymbol{x}_0$ such that $f(U \cap A) \subseteq V$. Thus $U \cap A$ is a relative neighbourhood of $\boldsymbol{x}_0$ in $f^{-1}(V)$ and so $f^{-1}(V)$ is open.

Now suppose that $f^{-1}(V)$ is relatively open in A for every open subset V of $\mathbb{R}^m$. Let $\boldsymbol{x}_0 \in A$ and let V be a neighbourhood of $f(\boldsymbol{x}_0)$. Then $f^{-1}(V)$ is a relative neighbourhood of $\boldsymbol{x}_0$ in A. By Proposition 4.2.50 there exists an open set U in $\mathbb{R}^n$ such that $f^{-1}(V) = U \cap A$. Therefore, since $f(f^{-1}(V)) \subseteq V$ by Proposition 1.3.5, it follows that f is continuous at $\boldsymbol{x}_0$ using Theorem 4.3.2. ■

The notion of uniform continuity can be extended to multivariable functions.

4.3.5 Definition (Uniform continuity) Let $A \subseteq \mathbb{R}^n$. A map $f\colon A \to \mathbb{R}^m$ is ***uniformly continuous*** if, for every $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that $\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|_{\mathbb{R}^n} < \epsilon$ whenever $\boldsymbol{x}_1, \boldsymbol{x}_2 \in A$ satisfy $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\mathbb{R}^n} < \delta$. •

Obviously all uniformly continuous functions are continuous. We refer the reader to Example 3.1.7 for an example of a continuous but not uniformly continuous function.

We close this section by initiating a discussion of the relationship between continuity, interior, closure, and boundary.

4.3.6 Proposition (Continuity and interior, closure, and boundary) *If* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$, *if* $\mathrm{S} \subseteq \mathrm{A}$, *if* $\mathrm{B} \subseteq \mathbb{R}^{\mathrm{m}}$, *and if* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}}$ *is continuous then the following statements hold:*

(i) $\mathrm{int}_{\mathrm{B}}(\mathbf{f}(\mathrm{S}))) \subseteq \mathbf{f}(\mathrm{int}_{\mathrm{A}}(\mathrm{S}))$;

(ii) $\mathbf{f}(\mathrm{cl}_{\mathrm{S}}(\mathrm{A})) \subseteq \mathrm{cl}_{\mathrm{B}}(\mathbf{f}(\mathrm{S}))$;

(iii) $\mathbf{f}(\mathrm{bd}_{\mathrm{S}}(\mathrm{A})) \subseteq \mathrm{bd}_{\mathrm{B}}(\mathbf{f}(\mathrm{S}))$.

Proof Let $\boldsymbol{y} \in \mathrm{int}_B(\boldsymbol{f}(S))$ then there exists a relative neighbourhood U of $\boldsymbol{y}$ in $\boldsymbol{f}(S)$ in B such that $U \subseteq \boldsymbol{f}(S)$. Then $\boldsymbol{f}^{-1}(U)$ is relatively open in A. That is, if $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ for $\boldsymbol{x} \in S$ then $\boldsymbol{x} \in \mathrm{int}_A(S)$. Thus $\boldsymbol{y} \in \boldsymbol{f}(\mathrm{int}_A(S))$.

Let $\boldsymbol{y} \in \boldsymbol{f}(\mathrm{cl}_A(S))$ with $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathrm{cl}_A(S)$. Then there exists a sequence $(\boldsymbol{x}_j)_{j\in\mathbb{Z}_{>0}}$ in S converging to $\boldsymbol{x}$. By Theorem 4.3.2 it follows that $(\boldsymbol{f}(\boldsymbol{x}_j))_{j\in\mathbb{Z}_{>0}}$ converges to $\boldsymbol{y}$. Since $\boldsymbol{f}(\boldsymbol{x}_j) \in \boldsymbol{f}(S)$ it follows that $\boldsymbol{y} \in \mathrm{cl}_B(\boldsymbol{f}(S))$.

Let $\boldsymbol{y} \in \boldsymbol{f}(\mathrm{bd}_A(S))$ with $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathrm{bd}_A(S)$. Then there exist sequences $(\boldsymbol{x}_j)_{j\in\mathbb{Z}_{>0}}$ in S and $(\boldsymbol{x}'_j)_{j\in\mathbb{Z}_{>0}}$ in $A \setminus S$, both converging to $\boldsymbol{x}$. By continuity of $\boldsymbol{f}$ the sequences $(\boldsymbol{f}(\boldsymbol{x}_j))_{j\in\mathbb{Z}_{>0}}$ in $\boldsymbol{f}(S)$ and $(\boldsymbol{f}(\boldsymbol{x}'_j))_{j\in\mathbb{Z}_{>0}}$ in $\boldsymbol{f}(A \setminus S) = \boldsymbol{f}(A) \setminus \boldsymbol{f}(S)$ both converge to $\boldsymbol{y}$. Thus $\boldsymbol{y} \in \mathrm{bd}_{\boldsymbol{f}(S)}(\boldsymbol{f}(B))$. ■

In general, the converse inclusions of the preceding result are not true.

4.3.7 Examples (Continuity and interior, closure, and boundary)

1. Consider $A = S = [0, \pi] \subseteq \mathbb{R}$, $B = [0, 1] \subseteq \mathbb{R}$, and take $f\colon A \to B$ given by $f(x) = \sin(x)$. Note that $f(S) = [0, 1]$. Then $f(\frac{\pi}{2}) = 1$ and so $1 \in f(\mathrm{int}_A(S))$. However, $1 \notin \mathrm{int}_B(f(S))$.
2. Take $A = S = \mathbb{R} \subseteq \mathbb{R}$, $B = [-\frac{\pi}{2}, \frac{\pi}{2}]$, and let $f\colon A \to B$ be given by $f(x) = \tan^{-1}(x)$. Note that $f(S) = (-\frac{\pi}{2}, \frac{\pi}{2})$. Thus $\frac{\pi}{2} \in \mathrm{cl}_B(f(S))$ but $\frac{\pi}{2} \notin f(\mathrm{cl}_A(S))$ since S is closed.
3. The same example as the preceding works here since $\frac{\pi}{2} \in \mathrm{bd}_B(f(S))$ but $\frac{\pi}{2} \notin f(\mathrm{bd}_A(S))$ since $\mathrm{bd}_A(S) = \emptyset$. •

4.3.2 Discontinuous maps

This section is rather specialised and technical and so can be omitted until needed. However, the material is needed at certain points in the text.

Next we consider the discontinuities of multivariable functions. The discussion here is not much different from that in a single variable, so we keep things brief.

4.3.8 Definition (Types of discontinuity) Let $A \subseteq \mathbb{R}^n$ and suppose that $f\colon A \to \mathbb{R}^m$ is discontinuous at $\boldsymbol{x}_0 \in A$. The point $\boldsymbol{x}_0$ is:

(i) a ***removable discontinuity*** if $\lim_{x\to_A x_0} \boldsymbol{f}(\boldsymbol{x})$ exists;

(ii) an ***essential discontinuity*** if the limit $\lim_{x\to_A x_0} \boldsymbol{f}(\boldsymbol{x})$ exists.

The set of all discontinuities of $\boldsymbol{f}$ is denoted by D_f. •

Note that we are not quite able to give as refined a characterisation of a point of discontinuity as we did in the single-variable case. This is because the discontinuities of multiple-variable functions can be rather more general that those for single-variable functions. Let us explore this in the context of an example.

4.3.9 Example (Strangeness of discontinuities for multivariable functions) We again consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ considered in Example 4.3.3:

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{x_1^2 + x_2^2}, & (x_1, x_2) \neq (0, 0), \\ 0, & (0, 0). \end{cases}$$

In Example 4.3.3 we showed that this function was continuous when thought of separately as a function of x_1 and of x_2, but was actually discontinuous at $(0, 0)$. Here we shall further explore the nature of the discontinuity at $(0, 0)$. First let us consider how the function behaves as we approach the origin along lines. Thus consider the line

$$s \mapsto (0, 0) + s(u_1, u_2), \qquad s \in \mathbb{R}$$

through $(0, 0)$ in the direction (u_1, u_2). We easily compute

$$f((0, 0) + s(u_1, u_2)) = \frac{u_1 u_2}{u_1^2 + u_2^2}.$$

If $u_1 = 0$ or $u_2 = 0$ then we have

$$\lim_{s \to 0} f((0, 0) + s(u_1, u_2)) = 0.$$

For $u_1 \neq 0$ let us take $u_2 = a u_1$, i.e., the line has slope $a \in \mathbb{R}$. In this case we have

$$\lim_{s \to 0} f((0, 0) + s(u_1, u_2)) = \frac{a}{1 + a^2}.$$

Similarly, if $u_2 \neq 0$ and $u_1 = b u_2$ then we have

$$\lim_{s \to 0} f((0, 0) + s(u_1, u_2)) = \frac{b}{1 + b^2}.$$

Thus all of these limits are finite, but the value of the limit depends on the direction in which one approaches $(0, 0)$. •

As in the single-variable case, we can use the oscillation to measure the discontinuity of a function.

4.3.10 Definition (Oscillation) Let $A \subseteq \mathbb{R}^n$ and let $\boldsymbol{f}\colon A \to \mathbb{R}^m$ be a map. The ***oscillation*** of $\boldsymbol{f}$ is the map $\omega_{\boldsymbol{f}}\colon A \to \mathbb{R}$ defined by

$$\omega_{\boldsymbol{f}}(\boldsymbol{x}) = \inf\{\sup\{\|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)\|_{\mathbb{R}^m} \mid \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathsf{B}^n(\delta, \boldsymbol{x}) \cap A\} \mid \delta \in \mathbb{R}_{>0}\}.$$ •

Note that the definition makes sense since the function

$$\delta \mapsto \sup\{\|\boldsymbol{f}(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\|_{\mathbb{R}^m} \mid \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathsf{B}^n(\delta, \boldsymbol{x}) \cap A\}$$

is monotonically increasing. In particular, if $\boldsymbol{f}$ is bounded (see Definition 4.3.30 below) then $\omega_{\boldsymbol{f}}$ is also bounded. The following result indicates in what way $\omega_{\boldsymbol{f}}$ measures the continuity of $\boldsymbol{f}$.

4.3.11 Proposition (Oscillation measures discontinuity) *For a subset* $A \subseteq \mathbb{R}$ *and a map* $\mathbf{f}\colon A \to \mathbb{R}$, $\mathbf{f}$ *is continuous at* $\mathbf{x} \in A$ *if and only if* $\omega_{\mathbf{f}}(\mathbf{x}) = 0$.

Proof Suppose that f is continuous at x and let $\epsilon \in \mathbb{R}_{>0}$. Choose $\delta \in \mathbb{R}_{>0}$ such that if $y \in \mathsf{B}^n(\delta, x) \cap A$ then $\|f(y) - f(x)\|_{\mathbb{R}^m} < \frac{\epsilon}{2}$. Then, for $x_1, x_2 \in \mathsf{B}^n(\delta, x)$ we have

$$\|f(x_1) - f(x_2)\|_{\mathbb{R}^m} \leq \|f(x_1) - f(x)\|_{\mathbb{R}^m} + \|f(x) - f(x_2)\|_{\mathbb{R}^m} < \epsilon.$$

Therefore,

$$\sup\{\|f(x_1) - f(x_2)\|_{\mathbb{R}^n} \mid x_1, x_2 \in \mathsf{B}^n(\delta, x) \cap A\} < \epsilon.$$

Since ϵ is arbitrary this gives

$$\inf\{\sup\{\|f(x_1) - f(x_2)\|_{\mathbb{R}^m} \mid x_1, x_2 \in \mathsf{B}^n(\delta, x) \cap A\} \mid \delta \in \mathbb{R}_{>0}\} = 0,$$

meaning that $\omega_f(x) = 0$.

Now suppose that $\omega_f(x) = 0$. For $\epsilon \in \mathbb{R}_{>0}$ let $\delta \in \mathbb{R}_{>0}$ be chosen such that

$$\sup\{\|f(x_1) - f(x_2)\|_{\mathbb{R}^m} \mid x_1, x_2 \in \mathsf{B}^n(\delta, x) \cap A\} < \epsilon.$$

In particular, $\|f(y) - f(x)\|_{\mathbb{R}^m} < \epsilon$ for all $y \in \mathsf{B}^n(\delta, x) \cap A$, giving continuity of f at x. ■

Let us consider an example where we can compute the oscillation.

4.3.12 Example (Oscillation for a discontinuous function) We again consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}$

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{x_1^2 + x_2^2}, & (x_1, x_2) \neq (0, 0), \\ 0, & (0, 0) \end{cases}$$

that is discontinuous at $(0, 0)$. Let us determine $\omega_f(0, 0)$. As we saw in Example 4.3.9, the function is constant on lines through $(0, 0)$. Therefore, all values of the function in any neighbourhood of $(0, 0)$ are attained by considering the values of the function along lines through $(0, 0)$. Moreover, in Example 4.3.9 we did this computation and we recall that the results were as follows.

1. On the line $s \mapsto (s, 0)$, $f(s, 0) = 0$.
2. On the line $s \mapsto (0, s)$, $f(0, s) = 0$.
3. On the line $s \mapsto (s, as)$, $f(s, as) = \frac{a}{1+a^2}$.
4. On the line $s \mapsto (bs, s)$, $f(bs, s) = \frac{b}{1+a^2}$.

The bottom line is that the values of f in any neighbourhood of $(0, 0)$ are in 1–1 correspondence with the elements of the set $\{\frac{a}{1+a^2} \mid a \in \mathbb{R}\}$. Thus one should look at the graph of the function $g\colon a \mapsto \frac{a}{1+a^2}$ to determine its maxima and minima. Since g is differentiable and $\lim_{a \to \pm\infty} g(a) = 0$, by Theorem 3.2.16 the maxima and minima occur where g' vanishes. We compute $g'(a) = \frac{1-a^2}{(a+1^2)^2}$ which means that maxima and minima must occur at $a \in \{-1, 1\}$. Also by Theorem 3.2.16, minima occur when $g''(a) > 0$ and maxima occur when $g''(a) < 0$. We compute $g''(1) = -\frac{1}{2}$ and $g''(-1) = \frac{1}{2}$. That $a = 1$ is a maximum for g and $a = -1$ is a minimum. We compute $g(1) = \frac{1}{2}$ and $g(-1) = -\frac{1}{2}$. This then gives $\omega_f(0, 0) = 1$.

Normally it will be quite difficult to explicitly compute the oscillation of a function. •

Let us now describe the possible set of discontinuities of an arbitrary multivariable function. The key to this, just as in the single-variable case, is the following result.

4.3.13 Proposition (Closed preimages of the oscillation of a function) *Let* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *and let* $\mathbf{f}\colon \mathrm{I} \to \mathbb{R}$ *be a function. Then, for every* $\alpha \geq 0$*, the set*

$$\mathrm{A}_\alpha = \{\mathbf{x} \in \mathrm{A} \mid \omega_{\mathbf{f}}(\mathbf{x}) \geq \alpha\}$$

is relatively closed in A.

Proof The result where $\alpha = 0$ is clear, so we assume that $\alpha \in \mathbb{R}_{>0}$. For $\delta \in \mathbb{R}_{>0}$ define

$$\omega_f(\boldsymbol{x}, \delta) = \sup\{\|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)\|_{\mathbb{R}^m} \mid \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathsf{B}^n(\delta, \boldsymbol{x}) \cap A\}$$

so that $\omega_f(\boldsymbol{x}) = \lim_{\delta \to 0} \omega_f(\boldsymbol{x}, \delta)$. Let $(\boldsymbol{x}_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in A_α converging to $\boldsymbol{x} \in \mathbb{R}^n$ and let $(\epsilon_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $(0, \alpha)$ converging to zero. Let $j \in \mathbb{Z}_{>0}$. We claim that there exists points $\boldsymbol{y}_j, \boldsymbol{z}_j \in \mathsf{B}^n(\epsilon_j, \boldsymbol{x}_j) \cap A$ such that $\|\boldsymbol{f}(\boldsymbol{y}_j) - \boldsymbol{f}(\boldsymbol{z}_j)\|_{\mathbb{R}^m} \geq \alpha - \epsilon_j$. Suppose otherwise so that for every $\boldsymbol{y}, \boldsymbol{z} \in \mathsf{B}^n(\epsilon_j, \boldsymbol{x}_j) \cap A$ we have $\|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{z})\|_{\mathbb{R}^m} < \alpha - \epsilon_j$. It then follows that $\lim_{\delta \to 0} \omega_f(\boldsymbol{x}_j, \delta) \leq \alpha - \epsilon_j < \alpha$, contradicting the fact that $\boldsymbol{x}_j \in A_\alpha$. We claim that $(\boldsymbol{y}_j)_{j \in \mathbb{Z}_{>0}}$ and $(\boldsymbol{z}_j)_{j \in \mathbb{Z}_{>0}}$ converge to $\boldsymbol{x}$. Indeed, let $\epsilon \in \mathbb{R}_{>0}$ and choose $N_1 \in \mathbb{Z}_{>0}$ sufficiently large that $\epsilon_j < \frac{\epsilon}{2}$ for $j \geq N_1$ and choose $N_2 \in \mathbb{Z}_{>0}$ such that $\|\boldsymbol{x}_j - \boldsymbol{x}\|_{\mathbb{R}^n} < \frac{\epsilon}{2}$ for $j \geq N_2$. Then, for $j \geq \max\{N_1, N_2\}$ we have

$$\|\boldsymbol{y}_j - \boldsymbol{x}\|_{\mathbb{R}^n} \leq \|\boldsymbol{y}_j - \boldsymbol{x}_j\|_{\mathbb{R}^n} + \|\boldsymbol{x}_j - \boldsymbol{x}\|_{\mathbb{R}^n} < \epsilon.$$

Thus $(\boldsymbol{y}_j)_{j \in \mathbb{Z}_{>0}}$ converges to $\boldsymbol{x}$, and the same argument, and therefore the same conclusion, also applies to $(\boldsymbol{z}_j)_{j \in \mathbb{Z}_{>0}}$.

Thus we have sequences of points $(\boldsymbol{y}_j)_{j \in \mathbb{Z}_{>0}}$ and $(\boldsymbol{z}_j)_{j \in \mathbb{Z}_{>0}}$ in A converging to $\boldsymbol{x}$ and a sequence $(\epsilon_j)_{j \in \mathbb{Z}_{>0}}$ in $(0, \alpha)$ converging to zero for which $\|\boldsymbol{f}(\boldsymbol{y}_j) - \boldsymbol{f}(\boldsymbol{z}_j)\|_{\mathbb{R}^m} \geq \alpha - \epsilon_j$. We claim that this implies that $\omega_f(\boldsymbol{x}) \geq \alpha$. Indeed, suppose that $\omega_f(\boldsymbol{x}) < \alpha$. There exists $N \in \mathbb{Z}_{>0}$ such that $\alpha - \epsilon_j > \alpha - \omega_f(\boldsymbol{x})$ for every $j \geq N$. Therefore,

$$\|\boldsymbol{f}(\boldsymbol{y}_j) - \boldsymbol{f}(\boldsymbol{z}_j)\|_{\mathbb{R}^m} \geq \alpha - \epsilon_j > \alpha - \omega_f(\boldsymbol{x})$$

for every $j \geq N$. This contradicts the definition of $\omega_f(\boldsymbol{x})$ since the sequences $(\boldsymbol{y}_j)_{j \in \mathbb{Z}_{>0}}$ and $(\boldsymbol{z}_j)_{j \in \mathbb{Z}_{>0}}$ converge to $\boldsymbol{x}$.

Now we claim that the sequence $(\boldsymbol{x}_j)_{j \in \mathbb{Z}_{>0}}$ converges to $\boldsymbol{x}$. Let $\epsilon \in \mathbb{R}_{>0}$ and let $N_1 \in \mathbb{Z}_{>0}$ be large enough that $\|\boldsymbol{x} - \boldsymbol{y}_j\|_{\mathbb{R}^n} < \frac{\epsilon}{2}$ for $j \geq N_1$ and let $N_2 \in \mathbb{Z}_{>0}$ be large enough that $\epsilon_j < \frac{\epsilon}{2}$ for $j \geq N_2$. Then, for $j \geq \max\{N_1, N_2\}$ we have

$$\|\boldsymbol{x} - \boldsymbol{x}_j\|_{\mathbb{R}^n} \leq \|\boldsymbol{x} - \boldsymbol{y}_j\|_{\mathbb{R}^n} + \|\boldsymbol{y}_j - \boldsymbol{x}_j\|_{\mathbb{R}^n} < \epsilon,$$

as desired.

This shows that every sequence in A_α converges to a point in A_α. It follows from Exercise 2.5.2 that A_α is closed. ∎

For readers who like the fancy language, we comment that the preceding result means exactly that ω_f is upper semicontinuous, cf. Proposition **??**.

The following corollary is somewhat remarkable, in that it shows that the set of discontinuities of a function cannot be arbitrary.

4.3.14 Corollary (Discontinuities are the countable union of closed sets) *Let* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *and let* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}}$ *be a function. Then the set*

$$\mathrm{D}_{\mathbf{f}} = \{\mathbf{x} \in \mathrm{A} \mid \mathbf{f} \textit{ is not continuous at } \mathbf{x}\}$$

is the countable union of closed sets.

Proof This follows immediately from Proposition 4.3.13 after we note that

$$D_f = \cup_{k \in \mathbb{Z}_{>0}} \{x \in A \mid \omega_f(x) \geq \tfrac{1}{k}\}. \qquad \blacksquare$$

4.3.3 Linear and affine maps

In this section we study a particularly simple, but as it turns out, very interesting class of continuous maps. While we studied linear maps in detail in Chapter **??**, let us redefine them here for fun, along with another, closely related type of map. The reader will recall that if $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathbb{R})$ is an $m \times n$-matrix with real entries (see Definition **??**) then the product of $\boldsymbol{A}$ with $\boldsymbol{x} \in \mathbb{R}^n$ is the element $\boldsymbol{Ax} \in \mathbb{R}^m$ defined by

$$(\boldsymbol{Ax})_a = \sum_{j=1}^{n} \boldsymbol{A}(a, j)x_j.$$

With this recollection we then make the following definition.

4.3.15 Definition (Linear map, affine map) A map $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is

(i) ***linear*** if there exists $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathbb{R})$ such that $f(\boldsymbol{x}) = \boldsymbol{Ax}$ for every $\boldsymbol{x} \in \mathbb{R}^n$ and is

(ii) ***affine*** if there exists $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathbb{R})$ and $\boldsymbol{b} \in \mathbb{R}^m$ such that $f(\boldsymbol{x}) = \boldsymbol{Ax} + \boldsymbol{b}$ for every $\boldsymbol{x} \in \mathbb{R}^n$. •

Recall from Theorem **??** that in the above definition we are establishing the natural identification of $\mathrm{Mat}_{m\times n}(\mathbb{R})$ with $\mathrm{Hom}_{\mathbb{R}}(\mathbb{R}^n; \mathbb{R}^m)$. Moreover, according to Proposition **??** this identification is of a matrix with the matrix representative of the linear map with respect to the standard basis. In this chapter we shall unblinkingly use this identification, and use the words "matrix" and "linear map" interchangeably, keeping in mind the natural identifications we are making.

Let us give some of the elementary properties of linear and affine maps. Since linear maps are special cases of affine maps, we sometimes need only consider them.

4.3.16 Proposition (Affine maps are uniformly continuous) *For* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathbb{R})$ *and* $\mathbf{b} \in \mathbb{R}^{\mathrm{m}}$*, the affine map* $\mathbf{f}\colon \mathbf{x} \mapsto \mathbf{Ax} + \mathbf{b}$ *is uniformly continuous.*

Proof Note that the ath component of $\boldsymbol{Ax}$ is exactly $\langle \boldsymbol{r}(\boldsymbol{A}, a), \boldsymbol{x}\rangle_{\mathbb{R}^n}$, where we recall from Definition **??** that $\boldsymbol{r}(\boldsymbol{A}, a)$ denotes the ath row of $\boldsymbol{A}$. Let

$$M = \max\{|\boldsymbol{r}(\boldsymbol{A}, a)| \mid a \in \{1, \ldots, m\}\}.$$

For $\epsilon \in \mathbb{R}_{>0}$ let $\delta = \frac{\epsilon}{\sqrt{m}M}$ and compute

$$\begin{aligned}\|f(x) - f(y)\|_{\mathbb{R}^m} &= \Big(\sum_{a=1}^{m} \langle r(A,a), x - y\rangle_{\mathbb{R}^n}\Big)^{1/2} \\ &\leq \Big(\sum_{a=1}^{m} \|r(A,a)\|_{\mathbb{R}^n}^2 \|x - y\|_{\mathbb{R}^n}^2\Big)^{1/2} \\ &\leq \sqrt{m}M\|x - y\|_{\mathbb{R}^n}.\end{aligned}$$

Thus, if $\|x - y\|_{\mathbb{R}^n} < \delta$ then $\|f(x) - f(y)\|_{\mathbb{R}^m} < \epsilon$, giving uniform continuity as desired. ■

4.3.4 Isometries

There is a special class of maps on $\mathbb{R}^n$ which (as we shall see) are affine. Let us first define the desired property of such maps.

4.3.17 Definition (Isometry of $\mathbb{R}^n$) A map $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is an ***isometry*** if

$$\|f(x_1) - f(x_2)\|_{\mathbb{R}^n} = \|x_1 - x_2\|_{\mathbb{R}^n}$$

for every $x_1, x_2 \in \mathbb{R}^n$. •

The idea of an isometry, then, is that it preserves the distance between points. It is not immediately obvious, but the set of isometries has a very simple structure. To get at this, we begin by considering linear isometries.

4.3.18 Theorem (Characterisation of linear isometries of $\mathbb{R}^n$) *For a matrix* $\mathbf{R} \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ *the following statements are equivalent:*

(i) $\mathbf{R}$ *is a linear isometry;*

(ii) $\|\mathbf{Rx}\|_{\mathbb{R}^n} = \|\mathbf{x}\|_{\mathbb{R}^n}$ *for all* $\mathbf{x} \in \mathbb{R}^n$;

(iii) $\langle \mathbf{Rx}, \mathbf{Ry}\rangle_{\mathbb{R}^n} = \langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{R}^n}$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$;

(iv) $\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{I}_n$;

(v) $\mathbf{R}$ *is invertible and* $\mathbf{R}^{-1} = \mathbf{R}^{\mathrm{T}}$.

Proof (i) $\implies$ (ii) If $\boldsymbol{R}$ is a linear isometry then

$$\|\boldsymbol{R}x - \boldsymbol{R}\mathbf{0}\|_{\mathbb{R}^n} = \|x - \mathbf{0}\|_{\mathbb{R}^n}$$

or $\|\boldsymbol{R}x\|_{\mathbb{R}^n} = \|x\|_{\mathbb{R}^n}$, as desired.

(ii) $\implies$ (iii) We are assuming that $\|\boldsymbol{R}x\|_{\mathbb{R}^n} = \|x\|_{\mathbb{R}^n}$ which implies that

$$\|\boldsymbol{R}x\|_{\mathbb{R}^n}^2 = \|x\|_{\mathbb{R}^n}^2 \quad \implies \quad \langle \boldsymbol{R}x, \boldsymbol{R}x\rangle_{\mathbb{R}^n} = \langle x, x\rangle_{\mathbb{R}^n},$$

this holding for all $x \in \mathbb{R}^n$. Thus, for every $x, y \in \mathbb{R}^n$,

$$\begin{aligned}&\langle \boldsymbol{R}(x+y), \boldsymbol{R}(x+y)\rangle_{\mathbb{R}^n} = \langle x+y, x+y\rangle_{\mathbb{R}^n} \\ \implies \quad &\langle \boldsymbol{R}x, \boldsymbol{R}x\rangle_{\mathbb{R}^n} + \langle \boldsymbol{R}y, \boldsymbol{R}y\rangle_{\mathbb{R}^n} + 2\langle \boldsymbol{R}x, \boldsymbol{R}y\rangle_{\mathbb{R}^n} = \langle x, x\rangle_{\mathbb{R}^n} + \langle y, y\rangle_{\mathbb{R}^n} + 2\langle x, y\rangle_{\mathbb{R}^n} \\ \implies \quad &\langle \boldsymbol{R}x, \boldsymbol{R}y\rangle_{\mathbb{R}^n} = \langle x, y\rangle_{\mathbb{R}^n},\end{aligned}$$

as desired.

(iii) $\implies$ (iv) Letting $\{e_1, \dots, e_n\}$ be the standard basis for $\mathbb{R}^n$ we have

$$\langle \boldsymbol{R}e_j, \boldsymbol{R}e_k \rangle_{\mathbb{R}^n} = \langle e_j, e_k \rangle_{\mathbb{R}^n}, \qquad j, k \in \{1, \dots, n\}.$$

We have

$$\langle e_j, e_k \rangle_{\mathbb{R}^n} = \boldsymbol{I}_n(j,k) = \begin{cases} 1, & j = k, \\ 0, & j \neq k \end{cases}$$

and a direct calculation shows that

$$\langle \boldsymbol{R}e_j, \boldsymbol{R}e_k \rangle_{\mathbb{R}^n} = \sum_{i=1}^{n} \boldsymbol{R}(i,j)\boldsymbol{R}(i,k) = (\boldsymbol{R}^T\boldsymbol{R})(j,k).$$

Thus $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}_n$. From Theorem **??** this means that $\boldsymbol{R}$ is invertible with inverse $\boldsymbol{R}^T$. This means that we also have $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}_n$.

(iv) $\implies$ (v) This was proved in the preceding part of the proof.

(v) $\implies$ (i) We first note that a direct computation shows that

$$\langle \boldsymbol{A}x, y \rangle_{\mathbb{R}^n} = \langle x, \boldsymbol{A}^T y \rangle_{\mathbb{R}^n} \tag{4.11}$$

for all $x, y \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathbb{R})$; this idea will be revealed in a more general setting in ***missing stuff***. If $\boldsymbol{R}$ is invertible with inverse $\boldsymbol{R}^T$ we have

$$\begin{aligned} & \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}_n \\ \implies \quad & \boldsymbol{R}^T\boldsymbol{R}x = x, \qquad x \in \mathbb{R}^n \\ \implies \quad & \langle \boldsymbol{R}^T\boldsymbol{R}x, x \rangle_{\mathbb{R}^n} = \langle x, x \rangle_{\mathbb{R}^n}, \qquad x \in \mathbb{R}^n \\ \implies \quad & \langle \boldsymbol{R}x, \boldsymbol{R}x \rangle_{\mathbb{R}^n} = \langle x, x \rangle_{\mathbb{R}^n}, \qquad x \in \mathbb{R}^n, \end{aligned}$$

using (4.11). Thus $\|\boldsymbol{R}x\|_{\mathbb{R}^n} = \|x\|_{\mathbb{R}^n}$ for every $x \in \mathbb{R}^n$. Therefore,

$$\|\boldsymbol{R}x_1 - \boldsymbol{R}x_2\|_{\mathbb{R}^n} = \|\boldsymbol{R}(x_1 - x_2)\|_{\mathbb{R}^n} = \|x_1 - x_2\|_{\mathbb{R}^n}$$

for all $x_1, x_2 \in \mathbb{R}^n$, meaning that $\boldsymbol{R}$ is an isometry. ■

Clearly linear isometries are very special. They are also very important, although we will not engage in a general investigation of these until ***missing stuff***. For now we just make a definition.

4.3.19 Definition (Orthogonal matrix) A matrix $\boldsymbol{R} \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ is ***orthogonal*** if it is a linear isometry. The set of orthogonal $n \times n$ matrices is denoted by $\mathsf{O}(n)$ and is called the ***orthogonal group*** in n-dimensions. •

Since we call $\mathsf{O}(n)$ the orthogonal group, it ought to be a group. The reader can verify that this is the case in Exercise 4.3.8.

With an understanding of linear isometries, it is possible to understand the structure of a general isometry. The following result gives the characterisation.

4.3.20 Theorem (Characterisation of isometries of $\mathbb{R}^n$) *A map $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^n$ is an isometry if and only if there exists $\mathbf{R} \in \mathsf{O}(n)$ and $\mathbf{r} \in \mathbb{R}^n$ such that*

$$\mathbf{f}(\mathbf{x}) = \mathbf{R}\mathbf{x} + \mathbf{r}, \qquad \mathbf{x} \in \mathbb{R}^n.$$

Proof First let us verify that the map $x \mapsto Rx + r$ is an isometry. We compute

$$\|(Rx_1 + r) - (Rx_2 + r)\|_{\mathbb{R}^n} = \|R(x_1 - x_2)\|_{\mathbb{R}^n} = \|x_1 - x_2\|_{\mathbb{R}^n},$$

using Theorem 4.3.18. Thus maps of the form given in the theorem statement are isometries.

Now suppose that f is an isometry. First suppose that f fixes $\mathbf{0} \in \mathbb{R}^n$: $f(\mathbf{0}) = \mathbf{0}$. We shall use the fact (see Exercise 4.3.1) that the Euclidean norm space satisfies the parallelogram law:

$$\|x + y\|^2_{\mathbb{R}^n} + \|x - y\|^2_{\mathbb{R}^n} = 2\Big(\|x\|^2_{\mathbb{R}^n} + \|y\|^2_{\mathbb{R}^n}\Big).$$

Using this equality, and the fact that f is an isometry fixing $\mathbf{0}$, we compute

$$\begin{aligned}\|f(x) + f(y)\|^2_{\mathbb{R}^n} &= 2\|f(x)\|^2_{\mathbb{R}^n} + 2\|f(y)\|^2_{\mathbb{R}^n} - \|f(x) - f(y)\|^2_{\mathbb{R}^n}\\ &= 2\|f(x) - f(\mathbf{0})\|^2_{\mathbb{R}^n} + 2\|f(y) - f(\mathbf{0})\|^2_{\mathbb{R}^n} - \|f(x) - f(y)\|^2_{\mathbb{R}^n}\\ &= 2\|x\|^2_{\mathbb{R}^n} + 2\|y\|^2_{\mathbb{R}^n} - \|x - y\|^2_{\mathbb{R}^n} = \|x + y\|^2_{\mathbb{R}^n}. \end{aligned} \tag{4.12}$$

By the polarization identity, see Exercise 4.3.1, we obtain

$$\langle x, y\rangle_{\mathbb{R}^n} = \tfrac{1}{2}\Big(\|x + y\|^2_{\mathbb{R}^n} - \|x\|^2_{\mathbb{R}^n} - \|y\|^2_{\mathbb{R}^n}\Big)$$

for every $x, y \in \mathbb{R}^n$. In particular, using (4.12) and the fact that f is an isometry fixing $\mathbf{0}$, we compute

$$\begin{aligned}\langle f(x), f(y)\rangle_{\mathbb{R}^n} &= \tfrac{1}{2}\Big(\|f(x) + f(y)\|^2_{\mathbb{R}^n} - \|f(x)\|^2_{\mathbb{R}^n} - \|f(y)\|^2_{\mathbb{R}^n}\Big)\\ &= \tfrac{1}{2}\Big(\|f(x) + f(y)\|^2_{\mathbb{R}^n} - \|f(x) - f(\mathbf{0})\|^2_{\mathbb{R}^n} - \|f(y) - f(\mathbf{0})\|^2_{\mathbb{R}^n}\Big)\\ &= \tfrac{1}{2}\Big(\|x + y\|^2_{\mathbb{R}^n} - \|x\|^2_{\mathbb{R}^n} - \|y\|^2_{\mathbb{R}^n}\Big) = \langle x, y\rangle_{\mathbb{R}^n}. \end{aligned} \tag{4.13}$$

We now claim that this implies that f is a linear map. Indeed, let $\{e_1, \dots, e_n\}$ be the standard basis for $\mathbb{R}^n$ and let $(x_1, \dots, x_n)$ be the components of $x \in \mathbb{R}^n$ in this basis (thus $x_i = \langle x, e_i\rangle_{\mathbb{R}^n}$, $i \in \{1, \dots, n\}$). Since

$$\langle f(e_i), f(e_j)\rangle_{\mathbb{R}^n} = \langle e_i, e_j\rangle_{\mathbb{R}^n}, \qquad i, j \in \{1, \dots, n\},$$

the vectors $\{f(e_1), \dots, f(e_n)\}$ form an orthonormal basis for $\mathbb{R}^n$ (see ***missing stuff*** for the notion of an orthonormal basis). The components of $f(x)$ in this basis are given by $\langle f(x, f(e_i)\rangle_{\mathbb{R}^n}$, $i \in \{1, \dots, n\}$. By (4.13) this means that the components of $f(x)$ are precisely $(x_1, \dots, x_n)$. That is,

$$f\Big(\sum_{i=1}^n x_i e_i\Big) = \sum_{i=1}^n x_i f(e_i).$$

Therefore, if f fixes $\mathbf{0} \in \mathbb{R}^n$ then f is linear and so, by Theorem 4.3.18, there exists $R \in \mathsf{O}(n)$ such that $f(x) = Rx$. Thus the theorem holds when f fixes $\mathbf{0}$.

Now, suppose that f fixes not $\mathbf{0}$, but some other point $x_0 \in \mathbb{R}^n$: $f(x_0) = x_0$. Then define $f_{x_0} \colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$f_{x_0}(x) = f(x + x_0) - x_0,$$

and note that $f_{x_0}(\mathbf{0}) = \mathbf{0}$. Thus $f_{x_0}(x) = \boldsymbol{R}(x)$ for some $\boldsymbol{R} \in \mathsf{O}(n)$. Therefore,

$$f(x) = f_{x_0}(x - x_0) + x_0 = \boldsymbol{R}x + x_0 - \boldsymbol{R}x_0.$$

Thus the theorem holds when f fixes a general point in $\mathbb{R}^n$.

Finally, suppose that f maps $x_1 \in \mathbb{R}^n$ to $x_2 \in \mathbb{R}^n$. In this most general case define $f_{x_1,x_2} \colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$f_{x_1,x_2}(x) = f(x) - (x_2 - x_1),$$

noting that $f_{x_1,x_2}(x_1) = x_1$. Therefore, by the previous part of the proof,

$$f_{x_1,x_2}(x) = \boldsymbol{R}x + r'$$

for some $\boldsymbol{R} \in \mathsf{O}(n)$ and some $r' \in \mathbb{R}^n$. Thus we get the theorem by taking $r = r' + (x_2 - x_1)$. ■

Now that we have described the set of isometries, let us name them.

4.3.21 Definition (Euclidean group) The ***Euclidean group*** in n-dimensions is the set of isometries of $\mathbb{R}^n$ and is denoted by $\mathsf{E}(n)$. •

Of course, the Euclidean group is a group, as the reader may verify in Exercise 4.3.11.

Note that there are two fundamental sorts of isometries. The first are ***translations*** which are of the form $x \mapsto x + r$ for some $r \in \mathbb{R}^n$. The second fundamental sort of isometry are those that are linear: $x \mapsto \boldsymbol{R}x$ for $\boldsymbol{R} \in \mathsf{O}(n)$. These are called ***rotations***. Theorem 4.3.20 tells us that a general isometry is a rotation followed by a translation.

4.3.5 Continuity and operations on functions

In this section we prove the hoped for properties of continuous functions with respect to the algebraic and topological properties of Euclidean space. First of all let us note that if $A \subseteq \mathbb{R}^n$ then the set of $\mathbb{R}^m$-valued maps on A is a $\mathbb{R}$-vector space. Indeed, the operations of vector addition and scalar multiplication are defined by

$$(f + g)(x) = f(x) + g(x), \qquad (af)(x) = a(f(x)),$$

where $f, g \colon A \to \mathbb{R}^m$ and where $a \in \mathbb{R}$. These operations respect continuous functions.

4.3.22 Proposition (Continuity, and addition and scalar multiplication) *If* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$, *if* $\mathbf{f}, \mathbf{g} \colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}}$ *are continuous, and if* $\mathrm{a} \in \mathbb{R}$ *then* $\mathbf{f} + \mathbf{g}$ *and* $\mathrm{a}\mathbf{f}$ *are continuous.*

Proof The proof differs from the relevant parts of the proof of Proposition 3.1.15 only by change of notation so we omit it here. ■

4.3.23 Proposition (Continuity and composition) *Let* $A \subseteq \mathbb{R}^m$, $B \subseteq \mathbb{R}^m$ *and let* $\mathbf{f}\colon A \to \mathbb{R}^m$ *and* $\mathbf{g}\colon B \to \mathbb{R}^k$ *have the properties that* $\text{image}(A) \subseteq B$ *and that* $\mathbf{f}$ *is continuous at* $\mathbf{x}_0 \in A$ *and* $\mathbf{g}$ *is continuous at* $\mathbf{f}(\mathbf{x}_0)$. *Then* $\mathbf{g} \circ \mathbf{f}$ *is continuous at* $\mathbf{x}_0$.

Proof This is proved in the same manner as Proposition 3.1.16. ■

4.3.24 Proposition (Continuity and restriction) *If* $A \subseteq \mathbb{R}^n$, *if* $B \subseteq A$, *and if* $\mathbf{f}\colon A \to \mathbb{R}^m$ *be continuous at* $\mathbf{x}_0 \in B$, *then* $\mathbf{f}|B$ *is continuous at* $\mathbf{x}_0$.

Proof The manner of proof here is like that in Proposition 3.1.17. ■

Note that the converse of the previous result is not generally true.

4.3.25 Example (Continuity and restriction) Define $f\colon \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1, & x \in \mathbb{Z}, \\ 0, & x \notin \mathbb{Z}. \end{cases}$$

Then $f|\mathbb{Z}$ is continuous (it is constant), but f is not continuous at points in $\mathbb{Z}$. •

Let us also indicate how continuity interacts with products.

4.3.26 Proposition (Continuity and products) *The following statements hold:*

(i) *if* $A_j \subseteq \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$, *and if* $\mathbf{f}\colon A_1 \times \cdots \times A_k \to \mathbb{R}^k$ *is continuous at* $(\mathbf{x}_{10}, \dots, \mathbf{x}_{k0})$, *then the maps*

$$\mathbf{x}_j \mapsto \mathbf{f}(\mathbf{x}_{10}, \dots, \mathbf{x}_j, \dots, \mathbf{x}_{k0}), \qquad j \in \{1, \dots, k\},$$

are continuous at $\mathbf{x}_{j0}$;

(ii) *if* $C \subseteq \mathbb{R}^k$ *and if* $\mathbf{g}\colon C \to \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ *is given by* $\mathbf{g}(\mathbf{z}) = (\mathbf{g}_1(\mathbf{z}), \dots, \mathbf{g}_k(\mathbf{z}))$ *for* $\mathbf{g}_j\colon C \to \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$, *then* $\mathbf{g}$ *is continuous at* $\mathbf{z}_0 \in C$ *if and only if each of the maps* $\mathbf{g}_j$, $j \in \{1, \dots, k\}$, *are continuous at* $\mathbf{z}_0$.

Proof By induction it suffices to prove the result for $k = 2$. We denote $n_1 = m$, $n_2 = n$, and write a typical point in $\mathbb{R}^m \times \mathbb{R}^n$ as (x, y).

(i) Suppose that f is continuous at (x_0, y_0) and let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence converging to x_0. Then the sequence $((x_j, y_0))_{j \in \mathbb{Z}_{>0}}$ is easily verified to converge to (x, y_0). Continuity of f and Theorem 4.3.2 ensures that

$$\lim_{j \to \infty} f(x_j, y_0) = f(x_0, y_0),$$

which in turn gives continuity of $x \mapsto f(x, y_0)$ at x_0 by Theorem 4.3.2. An entirely similar argument gives continuity of $y \mapsto f(x_0, y)$ at y_0.

(ii) First suppose that g is continuous at z_0. Then, for a sequence $(z_j)_{j \in \mathbb{Z}_{>0}}$ in C converging to z_0, the sequence $((g_1(z_j), g_2(z_j)))_{j \in \mathbb{Z}_{>0}}$ converges to $(g_1(z_0), g_2(z_0))$ by Theorem 4.3.2. From Exercise 4.2.11 we know that the sequences $(g_1(z_j))_{j \in \mathbb{Z}_{>0}}$ and $(g_2(z_j))_{j \in \mathbb{Z}_{>0}}$ converge to $g_1(z_0)$ and $g_2(z_0)$, respectively. By Theorem 4.3.2 it follows that g_1 and g_2, respectively.

The argument can be reversed, using Exercise 4.2.11 and Theorem 4.3.2, to show that g is continuous at (x_0, y_0) if g_1 is continuous at x_0 and g_2 is continuous at y_0. ■

The reader will notice that an implication is missing from the preceding result. This is not an oversight.

4.3.27 Example (Discontinuous function continuous in both variables) Define $f\colon \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x_1, x_2) = \begin{cases} \frac{x_1^2 x_2}{x_1^4 + x_2^2}, & (x_1, x_2) \neq (0, 0), \\ 0, & (x_1, x_2) = (0, 0). \end{cases}$$

We claim that f is not continuous at $(0, 0)$. Consider a point in $\mathbb{R}^2$ of the form (a, a^2) for $a \in \mathbb{R}$. At such points we have $f(a, a^2) = \frac{1}{2}$. Since $f(0, 0) = 0$ and since any neighbourhood of $(0, 0)$ contains a point of the form (a, a^2) for some $a \in \mathbb{R}^*$, it follows that f cannot be continuous at $(0, 0)$.

However, the two functions

$$x_1 \mapsto f(x_1, 0) = 0, \quad x_2 \mapsto f(0, x_2) = 0$$

are obviously continuous. •

Let us finally consider the behaviour of continuity with respect to the operations of selection of maximums and minimums.

4.3.28 Proposition (Continuity and min and max) *If* $\mathrm{A} \subseteq \mathbb{R}^n$ *and if* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *are continuous functions, then the functions*

$$\mathrm{A} \ni \mathbf{x} \mapsto \min\{\mathrm{f}(\mathbf{x}), \mathrm{g}(\mathbf{x})\} \in \mathbb{R}, \qquad \mathrm{A} \ni \mathbf{x} \mapsto \max\{\mathrm{f}(\mathbf{x}), \mathrm{g}(\mathbf{x})\} \in \mathbb{R}$$

are continuous.

Proof Let $x_0 \in A$ and let $\epsilon \in \mathbb{R}_{>0}$. Let us first assume that $f(x_0) > g(x_0)$. That is to say, assume that $(f - g)(x_0) \in \mathbb{R}_{>0}$. Continuity of f and g ensures that there exists $\delta_1 \in \mathbb{R}_{>0}$ such that if $x \in \mathsf{B}^n(\delta_1, x_0) \cap A$ then $(f - g)(x) \in \mathbb{R}_{>0}$. That is, if $x \in \mathsf{B}^n(\delta_1, x_0) \cap A$ then

$$\min\{f(x), g(x)\} = g(x), \quad \max\{f(x), g(x)\} = f(x).$$

Continuity of f ensures that there exists $\delta_2 \in \mathbb{R}_{>0}$ such that if $x \in \mathsf{B}^n(\delta_2, x_0) \cap A$ then $|f(x) - f(x_0)| < \epsilon$. Similarly, continuity of f ensures that there exists $\delta_3 \in \mathbb{R}_{>0}$ such that if $x \in \mathsf{B}^n(\delta_3, x_0) \cap A$ then $|g(x) - g(x_0)| < \epsilon$. Let $\delta_4 = \min\{\delta_1, \delta_2\}$. If $x \in \mathsf{B}(\delta_4, x_0) \cap A$ then

$$|\min\{f(x), g(x)\} - \min\{f(x_0), g(x_0)\}| = |g(x) - g(x_0)| < \epsilon$$

and

$$|\max\{f(x), g(x)\} - \max\{f(x_0), g(x_0)\}| = |f(x) - f(x_0)| < \epsilon.$$

This gives continuity of the two functions in this case. Similarly, swapping the rôle of f and g, if $f(x_0) < g(x_0)$ one can arrive at the same conclusion. Thus we need only consider the case when $f(x_0) = g(x_0)$. In this case, by continuity of f and g, choose $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(x_0)| < \epsilon$ and $|g(x) - g(x_0)| < \epsilon$ for $x \in \mathsf{B}(\delta, x_0) \cap A$. Then let $x \in \mathsf{B}(\delta, x_0) \cap A$. If $f(x) \geq g(x)$ then we have

$$|\min\{f(x), g(x)\} - \min\{f(x_0), g(x_0)\}| = |g(x) - g(x_0)| < \epsilon$$

and

$$|\max\{f(x), g(x)\} - \max\{f(x_0), g(x_0)\}| = |f(x) - f(x_0)| < \epsilon.$$

This gives the result in this case, and one similarly gets the result when $f(x) < g(x)$. ■

4.3.6 Continuity, and compactness and connectedness

As we saw in Section 3.1.4 for single-variable functions, continuity acts nicely with respect to certain topological notions including compactness and connectedness. We give these results here in the multivariable case, noting that there is a great deal in common with the single-variable case. Thus we will go through this fairly quickly.

4.3.29 Proposition (The continuous image of a compact set is compact) *If* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *is compact and if* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}}$ *is continuous, then* image($\mathbf{f}$) *is compact.*

Proof Let $(U_i)_{i\in I}$ be an open cover of image(f). Then $(f^{-1}(U_i))_{i\in I}$ is an open cover of A, and so there exists a finite subset $(i_1, \dots, i_k) \subseteq I$ such that $\cup_{j=1}^{k} f^{-1}(U_{i_k}) = A$. It is then clear that $(f(f^{-1}(U_{i_1})), \dots, f(f^{-1}(U_{i_k})))$ covers image(f). Moreover, by Proposition 1.3.5, $f(f^{-1}(U_{i_j})) \subseteq U_{i_j}$, $j \in \{1, \dots, k\}$. Thus $(U_{i_1}, \dots, U_{i_k})$ is a finite subcover of $(U_i)_{i\in I}$. ■

The following properties of functions interact well with compactness.

4.3.30 Definition (Bounded map) For an subset $A \subseteq \mathbb{R}^n$, a map $f\colon A \to \mathbb{R}^m$ is:

(i) ***bounded*** if there exists $M \in \mathbb{R}_{>0}$ such that image(f) $\subseteq \overline{\mathrm{B}}^n(M, \mathbf{0})$;

(ii) ***locally bounded*** if $f|K$ is bounded for every compact subset $K \subseteq A$;

(iii) ***unbounded*** if it is not bounded. •

4.3.31 Theorem (Continuous functions on compact sets are bounded) *If* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *is compact, then a continuous function* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}}$ *is bounded.*

Proof Let $x \in A$. As f is continuous, there exists $\delta \in \mathbb{R}_{>0}$ so that $\|f(y) - f(x)\|_{\mathbb{R}^m} < 1$ provided that $\|y - x\|_{\mathbb{R}^n} < \delta$. In particular, if $x \in A$, there is a neighbourhood U_x of x such that $\|f(y)\|_{\mathbb{R}^n} \leq \|f(x)\|_{\mathbb{R}^m} + 1$ for all $x \in U_x \cap A$. Thus f is bounded on $U_x \cap A$. This can be done for each $x \in A$, so defining a family of open sets $(U_x)_{x\in A}$. Clearly $A \subseteq \cup_{x\in A} U_x$, and so, by Theorem 4.2.35, there exists a finite collection of points $x_1, \dots, x_k \in A$ such that $A \subseteq \cup_{j=1}^{k} U_{x_j}$. Obviously for any $x \in A$,

$$\|f(x)\|_{\mathbb{R}^m} \leq 1 + \max\{f(x_1), \dots, f(x_k)\},$$

thus showing that f is bounded. ■

4.3.32 Theorem (Continuous functions on compact sets achieve their extreme values) *If* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *is a compact interval and if* $\mathrm{f}\colon \mathrm{A} \to \mathbb{R}$ *is continuous, then there exist points* $\mathbf{x}_{\min}, \mathbf{x}_{\max} \in \mathrm{A}$ *such that*

$$\mathrm{f}(\mathbf{x}_{\min}) = \inf\{\mathrm{f}(\mathbf{x}) \mid \mathbf{x} \in \mathrm{A}\}, \quad \mathrm{f}(\mathbf{x}_{\max}) = \sup\{\mathrm{f}(\mathbf{x}) \mid \mathbf{x} \in \mathrm{A}\}.$$

Proof It suffices to show that f achieves its maximum on A since if f achieves its maximum, then $-f$ will achieve its minimum. So let $M = \sup\{f(x) \mid x \in A\}$, and suppose that there is no point $x_{\max} \in A$ for which $f(x_{\max}) = M$. Then $f(x) < M$ for each $x \in A$. For a given $x \in A$ we have

$$f(x) = \tfrac{1}{2}(f(x) + f(x)) < \tfrac{1}{2}(f(x) + M).$$

Continuity of f ensures that there is an open set U_x containing x such that, for each $y \in U_x \cap A$, $f(y) < \frac{1}{2}(f(x) + M)$. Since $A \subseteq \cup_{x \in A} U_x$, by the Heine–Borel theorem, there exists a finite number of points $x_1, \ldots, x_k$ such that $A \subseteq \cup_{j=1}^{k} U_{x_j}$. Let $m = \max\{f(x_1), \ldots, f(x_k)\}$ so that, for each $y \in I_{x_j}$, and for each $j \in \{1, \ldots, k\}$, we have

$$f(y) < \tfrac{1}{2}(f(x_j) + M) < \tfrac{1}{2}(m + M),$$

which shows that $\frac{1}{2}(m+M)$ is an upper bound for f. However, since f attains the value m on A, we have $m < M$ and so $\frac{1}{2}(m + M) < M$, contradicting the fact that M is the least upper bound. Thus our assumption that f cannot attain the value M on A is false. ■

As in the single-variable case we saw that continuity and compactness conspire to give uniform continuity. This is true in the multivariable case as well, and serves to further establish the connection between "compactness" and "uniformly."

4.3.33 Theorem (Heine–Cantor Theorem) *Let* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *be compact. If* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}}$ *is continuous, then it is uniformly continuous.*

Proof Let $x \in A$ and let $\epsilon \in \mathbb{R}_{>0}$. Since f is continuous, then there exists $\delta_x \in \mathbb{R}_{>0}$ such that, if $y \in \mathsf{B}^n(\delta_x, x) \cap A$ then $f(y) \in \mathsf{B}^m(\frac{\epsilon}{2}, f(x))$. Note that $A \subseteq \cup_{x \in A} \mathsf{B}^n(\frac{\delta_x}{2}, x)$, so that the open sets $(\mathsf{B}^n(\frac{\delta_x}{2}, x))_{x \in A}$ cover A. By definition of compactness, there then exists a finite number of these open sets that cover A. Denote this finite family by $(\mathsf{B}^n(\frac{\delta_{x_1}}{2}, x_1), \ldots, \mathsf{B}^n(\frac{\delta_{x_k}}{2}, x_k))$ for some $x_1, \ldots, x_k \in A$. Take $\delta = \frac{1}{2}\min\{\delta_{x_1}, \ldots, \delta_{x_k}\}$. Now let $x, y \in A$ satisfy $\|x - y\|_{\mathbb{R}^n} < \delta$. Then there exists $j \in \{1, \ldots, k\}$ such that $x \in \mathsf{B}^n(\frac{\delta_{x_k}}{2}, x_j)$. We also have

$$\|y - x_j\|_{\mathbb{R}^n} \leq \|y - x\|_{\mathbb{R}^n} + \|x - x_j\|_{\mathbb{R}^n} < \delta_{x_j},$$

using the triangle inequality. Therefore,

$$\|f(y) - f(x)\|_{\mathbb{R}^m} \leq \|f(y) - f(x_j)\|_{\mathbb{R}^m} + \|f(x_j) - f(x)\|_{\mathbb{R}^m} < \epsilon,$$

again using the triangle inequality. Since this holds for *any* $x \in A$, it follows that f is uniformly continuous. ■

Now let us turn to connectedness and its relation to continuity.

4.3.34 Proposition (The continuous image of a (path) connected set is (path) connected) *If* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *is (path) connected and if* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}}$ *is continuous, then* $\mathbf{f}(\mathrm{A})$ *is (path) connected.*

Proof Suppose that $f(A)$ is not connected. Then there exist nonempty separated sets S and T such that $f(A) = S \cup T$. Let $S' = f^{-1}(S)$ and $T' = f^{-1}(T)$ so that $A = S' \cup T'$. By Propositions 4.2.28 and 1.3.5, and since $f^{-1}(\mathrm{cl}(S))$ is closed, we have

$$\mathrm{cl}(S') = \mathrm{cl}(f^{-1}(S)) \subseteq \mathrm{cl}(f^{-1}(\mathrm{cl}(S)) = f^{-1}(\mathrm{cl}(S)).$$

Therefore, by Proposition 1.3.5,

$$\mathrm{cl}(S') \cap T' \subseteq f^{-1}(\mathrm{cl}(S)) \cap f^{-1}(T) = f^{-1}(\mathrm{cl}(S) \cap T) = \emptyset.$$

We also similarly have $S' \cap \mathrm{cl}(T') = \emptyset$. Thus A is not connected, which gives the result for connectedness.

Now suppose that A is path connected and let $y_1, y_2 \in \text{image}(f)$. Thus $y_1 = f(x_1)$ and $y_2 = f(x_2)$. Since A is path connected there exists a continuous path $\gamma\colon [a,b] \to A$ such that $\gamma(a) = x_1$ and $x_2 = \gamma(b)$. The path $f \circ \gamma$ in $\text{image}(f)$ is continuous by Proposition 4.3.23 and has the property that $f \circ \gamma(a) = y_1$ and $f \circ \gamma(b) = y_2$. Thus $\text{image}(f)$ is path connected. ■

In multiple variables, the Intermediate Value Theorem is actually significantly more revealing than it is in the single-variable case. Indeed, it illustrates that it is connectivity that is the crucial ingredient in the theorem.

4.3.35 Theorem (Intermediate Value Theorem) *Let* $A \subseteq \mathbb{R}^n$ *be connected and let* $f\colon A \to \mathbb{R}$ *be continuous. If* $\mathbf{x}_1, \mathbf{x}_2 \in A$ *then, for any* $y \in [f(\mathbf{x}_1), f(\mathbf{x}_2)]$*, there exists* $\mathbf{x} \in A$ *such that* $f(\mathbf{x}) = y$.

Proof From Proposition 4.3.34 we know that $\text{image}(f)$ is connected and so is an interval by virtue of Theorem 2.5.34. The points $f(x_1)$ and $f(x_2)$ lie in this interval, and so too, therefore, does every point between $f(x_1)$ and $f(x_1)$. ■

4.3.7 Homeomorphisms

As we become more mature, we become more able to digest advanced concepts. In this section introduce the idea of a homeomorphism. The idea of a homeomorphism is an important one; it plays the rôle played by isomorphism for algebraic objects. That is, a homeomorphism gives the backdrop for understanding those things that are "continuous invariants," meaning that they are invariant under continuous maps. Obviously, not just any continuous map will do. Upon reflection, the following sort of continuous map is the reasonable one to generate the notion of "continuous invariants."

4.3.36 Definition (Homeomorphism, homeomorphic) If $A \subseteq \mathbb{R}^n$ and $B \subseteq \mathbb{R}^m$, a ***homeomorphism*** from A to B is a continuous bijection $f\colon A \to B$ whose inverse $f^{-1}\colon B \to A$ is also continuous. If $A \subseteq \mathbb{R}^n$ and $B \subseteq \mathbb{R}^m$ have the property that there exists a homeomorphism $f\colon A \to B$, then A and B are ***homeomorphic***. •

The following result is obvious, but is worth recording so it is out in the open.

4.3.37 Proposition ("Homeomorphic" is an equivalence relation) *If* $A \subseteq \mathbb{R}^n$, $B \subseteq \mathbb{R}^m$, *and* $C \subseteq \mathbb{R}^k$ *then the following statements hold:*

(i) A *is homeomorphic to* A*;*

(ii) if A *is homeomorphic to* B *then* B *is homeomorphic to* A*;*

(iii) if A *and* B *are homeomorphic and if* B *and* C *are homeomorphic, then* A *and* C *are homeomorphic.*

In other words, the relation "$A \sim B$ *if* A *and* B *are homeomorphic" between subsets of Euclidean spaces is an equivalence relation.*

Let us give some examples so that we develop some feeling for what a homeomorphism is and is not.

4.3.38 Examples (Homeomorphisms)

1. For any subset $A \subseteq \mathbb{R}^n$ the identity map $\mathrm{id}_A \colon A \to A$ is a homeomorphism. This is easy to check.
2. Let $\mathsf{V} \subseteq \mathbb{R}^n$ be a subspace and let $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}$ be a basis for V. We claim that the map $\mathsf{L} \colon \mathbb{R}^k \to \mathsf{V}$ defined by

$$\mathsf{L}(x_1, \dots, x_k) = x_1\boldsymbol{v}_1 + \cdots + x_k\boldsymbol{v}_k$$

is a homeomorphism. Certainly it is bijective (if you do not immediately see this, this means you need to read up on linear independence in Section **??**).
To see that it is continuous, denote

$$M = \max\{\|\boldsymbol{v}_1\|_{\mathbb{R}^k}, \dots, \|\boldsymbol{v}_k\|_{\mathbb{R}^k}\}$$

and, for $\epsilon \in \mathbb{R}_{>0}$, choose $\delta = \frac{\epsilon}{kM}$. If $\|\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^k} < \delta$ then $|x_j - y_j| < \delta$ for every $j \in \{1, \dots, k\}$. Thus we have, for $\|\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^k} < \delta$,

$$\begin{aligned} \|\mathsf{L}(\boldsymbol{x}) - \mathsf{L}(\boldsymbol{y})\|_{\mathbb{R}^n} &= \|(x_1 - y_1)\boldsymbol{v}_1 + \cdots + (x_k - y_k)\boldsymbol{v}_k\|_{\mathbb{R}^m} \\ &\leq |x_1 - y_1|\|\boldsymbol{v}_1\|_{\mathbb{R}^m} + \cdots + |x_k - y_k|\|\boldsymbol{v}_k\|_{\mathbb{R}^m} \\ &< kM\delta = \epsilon. \end{aligned}$$

This shows that L is continuous, indeed uniformly continuous, consistent with Proposition 4.3.16.
Now let us show that L^{-1} is continuous. By Theorem **??** we take vectors $\boldsymbol{v}_{k+1}, \dots, \boldsymbol{v}_n \in \mathbb{R}^n$ such that $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ is a basis for $\mathbb{R}^n$. Then define a linear map $\hat{\mathsf{L}} \colon \mathbb{R}^n \to \mathbb{R}^k$ by asking that

$$\hat{\mathsf{L}}(\boldsymbol{v}_j) = \begin{cases} \boldsymbol{e}_j, & j \in \{1, \dots, k\}, \\ \boldsymbol{0}, & j \in \{k+1, \dots, n\}, \end{cases}$$

cf. Theorem **??**. By Proposition 4.3.16 we know that $\hat{\mathsf{L}}$ is continuous and by Proposition 4.3.24 we know that, as a result, $\mathsf{L} = \hat{\mathsf{L}}|\mathsf{V}$ is continuous.
3. Let $A = (0, \infty)$ and let $B = \mathbb{R}$. Define $f \colon A \to B$ by $f(x) = \log(x)$. By Proposition 3.6.6 f is a homeomorphism. Since every open unbounded interval that is a strict subset of $\mathbb{R}$ is of the form (a, ∞) or $(-\infty, b)$, one can easily modify our construction to show that all such intervals homeomorphic to $\mathbb{R}$; see Exercise 4.3.12.
4. Let $A = (0, 1) \subseteq \mathbb{R}$ and let $B = \mathbb{R}$. The map $f \colon A \to B$ given by $f(x) = \tan^{-1}(\pi(x - \frac{1}{2}))$ is a homeomorphism, this following from Proposition 3.6.20. It is possible to modify this example to show that every bounded open interval is homeomorphic to $\mathbb{R}$; see Exercise 4.3.12.
5. Let $A = (-\pi, \pi] \subseteq \mathbb{R}$ and let

$$B = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 = 1\}.$$

Thus B is the unit circle in $\mathbb{R}^2$. Any point in $(x_1, x_2) \in B$ is expressed in the form $(x_1, x_2) = (\cos(x), \sin(x))$ for some $x \in \mathbb{R}$; see Proposition 3.6.19(iii). Moreover, if we ask that $x \in (-\pi, \pi]$ then there exists a unique such point such that $(x_1, x_2) = (\cos(x), \sin(x))$. That is, the map $f\colon A \to B$ defined by $f(x) = (\cos(x), \sin(x))$ is a bijection. We claim that f is continuous. This follows directly from the continuity of cos and sin; see Proposition 3.6.19(i). We also claim that f^{-1} is discontinuous at $(-1, 0)$. To see why this is so, note that $f^{-1}(-1, 0) = \pi$. Now let $(x_1, x_2) \in B$ satisfy $x_1, x_2 < 0$. Then $f^{-1}(x_1, x_2) \in (-\pi, -\frac{\pi}{2})$. Thus, for all such points we have

$$|f^{-1}(x_1, x_2) - f^{-1}(-1, 0)| > \tfrac{\pi}{2}.$$

However, for any $\delta \in \mathbb{R}_{>0}$ there exists a point $(x_1, x_2) \in B$ with $(x_1, x_2) < 0$ such that $\|(x_1, x_2) - (-1, 0)\|_{\mathbb{R}^2} < \delta$. Thus $f^{-1}(\mathsf{B}^2(\delta, (-1, 0))) \not\subset \mathsf{B}^1(1, \pi)$, giving discontinuity of f^{-1} at $(-1, 0)$.

The point is that a continuous bijection need not be a homeomorphism. •

The second of the preceding examples is worth expounding on a little.

4.3.39 Remark (The topology of a subspace) If one has two bases $\{v_1, \dots, v_k\}$ and $\{v'_1, \dots, v'_k\}$ for a subspace $\mathsf{V} \subseteq \mathbb{R}^n$, these induce as in Example 2 two homeomorphisms $\mathsf{L}, \mathsf{L}' \colon \mathbb{R}^k \to \mathsf{V}$. Thus, by Proposition 4.3.37, the subspace V is homeomorphic to $\mathbb{R}^k$ in a manner not depending in the use of a basis to establish the homeomorphism. In other words, a k-dimensional subspace inherits in a natural way the topological structure of $\mathbb{R}^k$. We shall use this fact in the sequel to, without loss of generality, work with all of $\mathbb{R}^n$ rather than a subspace of $\mathbb{R}^n$. This is a special case of the general principle that it is sometimes convenient to work with a set homeomorphic to the one in a given problem. •

As mentioned in the preparatory comments of this section, the notion of a homeomorphism has the intent of allowing us to consider properties that are "continuous invariants." The reader may understand this idea by comparing it to a statement from linear algebra; Proposition **??** says that the dimension of a vector space is an isomorphism invariant (indeed, it is actually the only isomorphism invariant). We are interested in properties of subsets of Euclidean space that are homeomorphism invariant. Let us make an actual definition so we know what we are talking about.

4.3.40 Definition (Topological invariant) A property P is a ***topological invariant*** if, whenever $A \subseteq \mathbb{R}^n$ has property P then every subset $B \subseteq \mathbb{R}^m$ that is homeomorphic to A also has property P. •

Unlike the comparatively simple situation in linear algebra where the only isomorphism invariant is dimension, an exhaustive list of topological invariants (okay, well, "simple" topological invariants) seems not to be practical. However, let us list some topological invariants that we have already encountered, as well as some concepts that are not topological invariants.

4.3.41 Theorem (Some topological invariants) *The following properties are topological invariants:*

(i) *compactness;*
(ii) *connectedness;*
(iii) *path-connectedness;*
(iv) *existence of a continuous map into given subset* $S \subseteq \mathbb{R}^n$*;*
(v) *existence of a continuous map from a given subset* $S \subseteq \mathbb{R}^n$*.*

The following properties are not topological invariants:

(vi) *openness;*
(vii) *closedness;*
(viii) *boundedness;*
(ix) *total boundedness.*

Proof Suppose that $A \subseteq \mathbb{R}^n$ is a compact (resp. connected, path connected) and let $f\colon A \to B \subseteq \mathbb{R}^m$ be a homeomorphism. Then B is compact (resp. connected, path connected) by Proposition 4.3.29 (resp. Proposition 4.3.34). This gives the first three properties as being topological invariants.

That the last two properties asserted as being topological invariants are, in fact, topological invariants is a consequence of the composition of continuous maps being continuous, i.e., of Proposition 4.3.23. For example, if A is homeomorphic to B with a homeomorphism $\boldsymbol{h}\colon A \to B$ and if $\boldsymbol{f}\colon S \to A$ is continuous, then $\boldsymbol{h} \circ \boldsymbol{f}$ is a continuous map of S into B.

To show that a property is not a topological invariant it suffices to give an example, and this is what we do for the last four parts of the theorem.

Note that $A = \mathbb{R}$ is open and is homeomorphic to the set

$$B = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 = 0\}$$

which is not open. Also, B is closed and homeomorphic to $(0, 1)$ (cf. Example 4.3.38–4) which is not closed.

The same example will suffice in each of the last two statements. Indeed, let $A = (0, 1)$ which is both bounded and totally bounded. However, B is homeomorphic to $\mathbb{R}$ by Example 4.3.38–4, and $\mathbb{R}$ is neither bounded nor totally bounded. ■

4.3.42 Remark ("Intrinsic" versus "extrinsic" properties) It is interesting to note that the three topological invariants we give in the preceding theorem differ in a fundamental way from the four properties that are not topological invariants. Indeed, note that the four properties that are not topological invariants have to do, not with the set itself, but with its properties as a subset of the Euclidean space in which it resides. The three properties that *are* topological invariants, however, have to do with the set itself, not how it sits in Euclidean space. There is something in this observation. •

Note that Example 4.3.38–5 shows that for a map to be a homeomorphism it is not sufficient for it to be a continuous bijection. Let us now turn to cases where it is possible to make this inference.

4.3.43 Theorem (Continuous bijections on compact sets are homeomorphisms) *If* $\mathrm{A} \subseteq \mathbb{R}^{\mathrm{n}}$ *is compact and if* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}}$ *is a continuous injection then* $\mathbf{f}$ *is a homeomorphism of* A *with* image($\mathbf{f}$).

Proof Let us denote $B = \text{image}(f)$ and $f^{-1}\colon B \to A$ the inverse. By Proposition 4.3.29 it follows that B is compact. We claim that the image of a relatively closed subset of A is relatively closed in B. Thus let $C \subseteq A$ be relatively closed so that, by Corollary 4.2.36, C is compact. Then $f(C)$ is a compact subset of B and so relatively closed, again by Corollary 4.2.36. Therefore, f maps relatively closed sets to relatively closed sets, and so also maps relatively open sets to relatively open sets by virtue of f being a bijection. Thus f^{-1} is continuous. ■

In our proof of the topological invariance of the property of openness in Proposition 4.3.41 we showed that the open subset $\mathbb{R} \subseteq \mathbb{R}$ is homeomorphic to the non-open subset of $\mathbb{R}^2$ consisting of the x_1-axis. The reader might protest that this is unfair, and that to make the statement interesting we should produce an open subset of $\mathbb{R}^n$ that is homeomorphic to a subset of $\mathbb{R}^n$ (the same "n," note) that is not open. It turns out, however, that such an example does not exist. This is nontrivial, but we will give the proof here anyway. The following theorem which gives the desired conclusion is an extremely important one, and is difficult to prove by "elementary" methods; the result is most naturally viewed from the point of view of either dimension theory or algebraic topology (see Section **??** for references). Our long but elementary proof relies crucially on Theorem **??**, which itself relies on the Weierstrass Approximation Theorem (Theorem 4.5.4), the Tietze Extension Theorem (Theorem **??**), and the Brouwer Fixed Point Theorem (Theorem **??**).

4.3.44 Theorem (Domain Invariance Theorem) *If* U *is an open subset of* $\mathbb{R}^{\mathrm{n}}$ *and if* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^{\mathrm{n}}$ *is an injective continuous map, then* image($\mathbf{f}$) *is open and* $\mathbf{f}$ *is a homeomorphism between* U *and* image($\mathbf{f}$).

Proof We begin with a couple of lemmata that contain the crux of the proof. We note that

$$\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_{\mathbb{R}^n} = 1\}$$

denotes the unit sphere in $\mathbb{R}^n$.

1 Lemma *If* $\mathrm{C} \subseteq \mathbb{R}^{\mathrm{n}}$ *is closed then the following two statements regarding* $\mathbf{x} \in \mathrm{C}$ *are equivalent:*

(i) $\mathbf{x} \in \mathrm{bd}(\mathrm{C})$*;*

(ii) for any relative neighbourhood V *of* $\mathbf{x}$ *in* C *there exists a relative neighbourhood* U *of* $\mathbf{x}$ *in* C *having the properties that*

  (a) $\mathrm{U} \subseteq \mathrm{V}$ *and*

  (b) if $\mathbf{g}\colon \mathrm{C} \setminus \mathrm{U} \to \mathbb{S}^{\mathrm{n}-1}$ *is continuous then there exists a continuous map* $\hat{\mathbf{g}}\colon \mathrm{C} \to \mathbb{S}^{\mathrm{n}-1}$ *such that* $\mathbf{g} = \hat{\mathbf{g}}|(\mathrm{C} \setminus \mathrm{U})$.

Proof (i) $\implies$ (ii) Suppose that $x_0 \in \mathrm{bd}(C)$ and let V be a relative neighbourhood of x_0 in C. By Proposition 4.2.50 there exists an open subset V' in $\mathbb{R}^n$ such that $V = C \cap V'$. Then let $\epsilon \in \mathbb{R}_{>0}$ be sufficiently small that $\mathsf{B}^n(\epsilon, x_0) \subseteq V'$ and take $U = C \cap \mathsf{B}^n(\epsilon, x_0)$. Let

$$\mathbb{S}^{n-1}(\epsilon, x_0) = \{x \in \mathbb{R}^n \mid \|x - x_0\|_{\mathbb{R}^n} = \epsilon\}$$

be the sphere of radius ϵ centred at x_0, i.e., $\mathbb{S}^{n-1}(\epsilon, x_0) = \mathrm{bd}(\mathsf{B}^n(\epsilon, x_0))$. Define

$$C_0 = C \cap \overline{\mathsf{B}}^n(\epsilon, x_0), \quad C_1 = C \setminus \mathsf{B}^n(\epsilon, x_0),$$

noting that $C = C_0 \cup C_1$, that $C_0 \cap C_1 \subseteq \mathbb{S}^{n-1}(\epsilon, x_0)$ and that

$$C_0 \cap \mathbb{S}^{n-1}(\epsilon, x_0) = C_1 \cap \mathbb{S}^{n-1}(\epsilon, x_0).$$

Now let $g\colon C_1 \to \mathbb{S}^{n-1}$ be continuous. We shall define the extension $\hat{g}\colon C \to \mathbb{S}^{n-1}$ by defining it on C_0 and then showing that the resulting map is consistently defined on $C_0 \cap C_1$.

The first observation to make is that $\mathbb{S}^{n-1}$ is homeomorphic to $\mathbb{S}^{n-1}(\epsilon, x_0)$ (see Exercise 4.3.15) and so any homeomorphism $\iota\colon \mathbb{S}^{n-1} \to \mathbb{S}^{n-1}(\epsilon, x_0)$ of these two sets will give a continuous map $h = \iota \circ g\colon C_1 \to \mathbb{S}^{n-1}(\epsilon, x_0)$. We shall define a map $\hat{h}\colon C \to \mathbb{S}^{n-1}(\epsilon, x_0)$ which extends h, and the desired map $\hat{g}$ is then given by $\hat{g} = \iota^{-1} \circ \hat{h}$.

Next note that by Corollary **??** there exists a continuous map $h'\colon \mathbb{S}^{n-1}(\epsilon, x_0) \to \mathbb{S}^{n-1}(\epsilon, x_0)$ that agrees with h on $C_1 \cap \mathbb{S}^{n-1}(\epsilon, x_0)$.

To define $\hat{h}$ on C_0 we note that, since $x_0 \in \mathrm{bd}(C)$, there exists a point $x_1 \in \mathsf{B}^n(\epsilon, x) - C$. If $x \in C_0 \subseteq \overline{\mathsf{B}}^n(\epsilon, x_0)$ define

$$y_x = x_1 + \frac{\|x - x_1\|^2_{\mathbb{R}^n} - \|x - x_0\|^2_{\mathbb{R}^n} + \epsilon\|x - x_1\|_{\mathbb{R}^n}}{\|x - x_1\|^2_{\mathbb{R}^n}}(x - x_1). \tag{4.14}$$

Note that y_x is the point on the sphere $\mathbb{S}^{n-1}(\epsilon, x_0)$ obtained as the intersection of the sphere with the ray from x_1 passing through x. The essential feature of y_x is that it is a continuous function of x. We take $\hat{h}(x) = h'(y_x)$. Since $y_x = x$ for $x \in C_0 \cap \mathbb{S}^{n-1}(\epsilon, x)$ we have $\hat{h}(x) = h(x)$ for $x \in C_0 \cap C_1$.

Thus we can take $\hat{h}(x) = h(x)$ for $x \in C_1$ and the result will be a consistently defined continuous $\mathbb{S}^{n-1}(\epsilon, x_0)$-valued map on C.

(ii) $\implies$ (i) Now suppose that $x_0 \in \mathrm{int}(C)$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}^n(\epsilon, x_0) \subseteq C$. Now let U be a relatively open neighbourhood of x_0 in C with the property that $U \subseteq \mathsf{B}^n(\epsilon, x_0)$. Now define $h\colon C \setminus U \to \mathbb{S}^{n-1}(\epsilon, x_0)$ by

$$h(x) = x_0 + \epsilon\frac{x - x_0}{\|x - x_0\|_{\mathbb{R}^n}}.$$

Note that $h(x)$ is the point on the sphere $\mathbb{S}^{n-1}(\epsilon, x_0)$ which is the intersection of the sphere with the ray from x_0 passing through x. Now suppose that there exists $\hat{h}\colon C \to \mathbb{S}^{n-1}(\epsilon, x_0)$ which extends h. Since $\mathbb{S}^{n-1}(\epsilon, x_0) \subseteq C \setminus U$ and since $h(x) = x$ for $x \in \mathbb{S}^{n-1}(\epsilon, x_0)$, it follows that $\hat{h}|\overline{\mathsf{B}}^n(\epsilon, x_0)$ is a retraction of $\overline{\mathsf{B}}^n(\epsilon, x_0)$ onto $\mathbb{S}^{n-1}(\epsilon, x_0)$. This is not possible by Proposition **??**, after recalling, as above, that $\overline{\mathsf{B}}^n(\epsilon, x_0)$ is homeomorphic to $\mathbb{D}^n$ (see Exercise 4.3.14). ▼

2 Lemma *If* $\mathrm{A} \subseteq \mathbb{R}^n$ *and* $\mathrm{B} \subseteq \mathbb{R}^m$ *are closed and if* $\mathbf{f}\colon \mathrm{A} \to \mathrm{B}$ *is a homeomorphism then* $\mathbf{f}(\mathrm{bd}(\mathrm{A})) = \mathrm{bd}(\mathrm{B})$.

Proof By Proposition 4.3.6 we have $f(\mathrm{bd}(A)) \subseteq \mathrm{bd}(B)$. Let $y \in \mathrm{bd}(B)$ so that $y = f(x)$ for some $x \in A$. Let V be a relative neighbourhood of x in A. Then continuity of f^{-1} gives $V' = f(V)$ as a relative neighbourhood of y in B. By Lemma 1 there exists a relative neighbourhood U' of y in B such that

1. $U' \subseteq V'$ and
2. if $g' : B \setminus U' \to \mathbb{S}^{n-1}$ is continuous then there exists a continuous map $\hat{g}' : B \to \mathbb{S}^{n-1}$ such that $g = \hat{g}|(B \setminus U)$.

Then define $U = f^{-1}(U')$ which, by continuity of f, is a relative neighbourhood of x. Moreover, $U \subseteq V$. Now let $g : A \setminus U \to \mathbb{S}^{n-1}$ be continuous. Then $g' \triangleq g \circ f^{-1}$ is a continuous map from $B \setminus U'$ to $\mathbb{S}^{n-1}$. There that exists $\hat{g}' : B \to \mathbb{S}^{n-1}$ extending g'. Now define $\hat{g} : A \to \mathbb{S}^{n-1}$ by $\hat{g} = \hat{g}' \circ f$. The continuity of $\hat{g}$ allows us to conclude that $x \in \text{bd}(A)$ and so $y \in f(\text{bd}(A))$. ▼

Proceeding with the proof, if $U' \subseteq U$ is open we claim that $f(U')$ is open. Let us denote $V' = f(U')$ and let $y \in V'$. Thus $y = f(x)$ for some $x \in U'$. Let $r \in \mathbb{R}_{>0}$ be such that $\overline{\mathsf{B}}^n(r, x) \subseteq U'$. Then $f|\mathsf{B}^n(r, x)$ is a homeomorphism onto its image by Theorem 4.3.43. Therefore, $f(x) \in \text{int}(f(U'))$ by Lemma 2. This shows that every point in V' is an interior point and so V' is open. In other words, if $V = f(U)$ then $f^{-1} : V \to U$ is continuous, as desired. ■

As we have said, the Domain Invariance Theorem is very important. Let us explore interpretations of it and some important consequences of it. First of all, the following result follows directly, and gives a useful topological invariance property.

4.3.45 Corollary (Openness in $\mathbb{R}^n$ is a topological invariant) *Let* $n \in \mathbb{Z}_{>0}$. *Then the property "*A *is an open subset of* $\mathbb{R}^n$*" is a topological invariant.*

Proof Suppose that $A \subseteq \mathbb{R}^n$ is open and that $B \subseteq \mathbb{R}^n$ is homeomorphic to A. Then there exists a homeomorphism $f : A \to B$. The map $f \circ i_B : A \to \mathbb{R}^n$ is then injective and continuous. Thus, by Theorem 4.3.44, its image is open. But its image is B. ■

Now let us attempt to understand the Domain Invariance Theorem by trying to gain some appreciation for why it is nontrivial. Let us see if we can do this for $n = 1$. Thus we consider an open subset $U \subseteq \mathbb{R}$ and a continuous injective map $f : U \to \mathbb{R}$. Since U is open, it is a union of intervals by Proposition 2.5.6. Thus we may as well restrict our attention to the case when U is an interval. In this case a continuous function will be strictly monotonically increasing or strictly monotonically decreasing; this is Exercise **??**. In the case when f is differentiable with positive or negative derivative the Domain Invariance Theorem is more or less obvious since, in this case, f is approximately linear with a positive or negative slope. So the real content of the Domain Invariance Theorem in this case occurs at points where f is either not differentiable, or has derivative zero. Let us then give an example which illustrates some facets of the Domain Invariance Theorem.

4.3.46 Example (A continuous, strictly monotonically increasing function that is not differentiable on a dense set) We give another peculiar sort of function to illustrate a rather subtle point. We define a sequence of functions $(f_k)_{k \in \mathbb{Z}_{\geq 0}}$ on $[0, 1]$ as follows. We take $f_0(x) = x$. To define f_1 take

$$f_1(0) = f_0(0) = 0, \ f_1(1) = f_0(1) = 1,$$
$$f_1(\tfrac{1}{2}) = (1 - \alpha) f_0(0) + \alpha f_0(1) = \alpha,$$

where $\alpha \in (0,1)$. We then define f_1 on $(0, \frac{1}{2})$ and $(\frac{1}{2}, 1)$ by asking that it be continuous and linear on these intervals. Now suppose that we have defined $f_0, f_1, \ldots, f_k$ and define f_{k+1} as follows. We require that

$$f_{k+1}(\tfrac{j}{2^k}) = f_k(\tfrac{j}{2^k}), \qquad j \in \{0, 1, \ldots, 2^k\},$$
$$f_{k+1}(\tfrac{2j+1}{2^{k+1}}) = (1-\alpha) f_k(\tfrac{j}{2^k}) + \alpha f_k(\tfrac{j+1}{2^k}), \qquad j \in \{0, 1, \ldots, 2^k - 1\}.$$

We then define f_{k+1} on all of $[0,1]$ by asking that it be linear on each of the subintervals $[\frac{j}{2^{k+1}}, \frac{j+1}{2^{k+1}}]$, $j \in \{0, 1, \ldots, 2^{k+1} - 1\}$. We then define $f_\alpha \colon [0,1] \to \mathbb{R}$ by

$$f_\alpha(x) = \lim_{k \to \infty} f_k(x), \qquad x \in [0,1].$$

In Figure 4.8 we show the first step in this construction for various α. The idea is

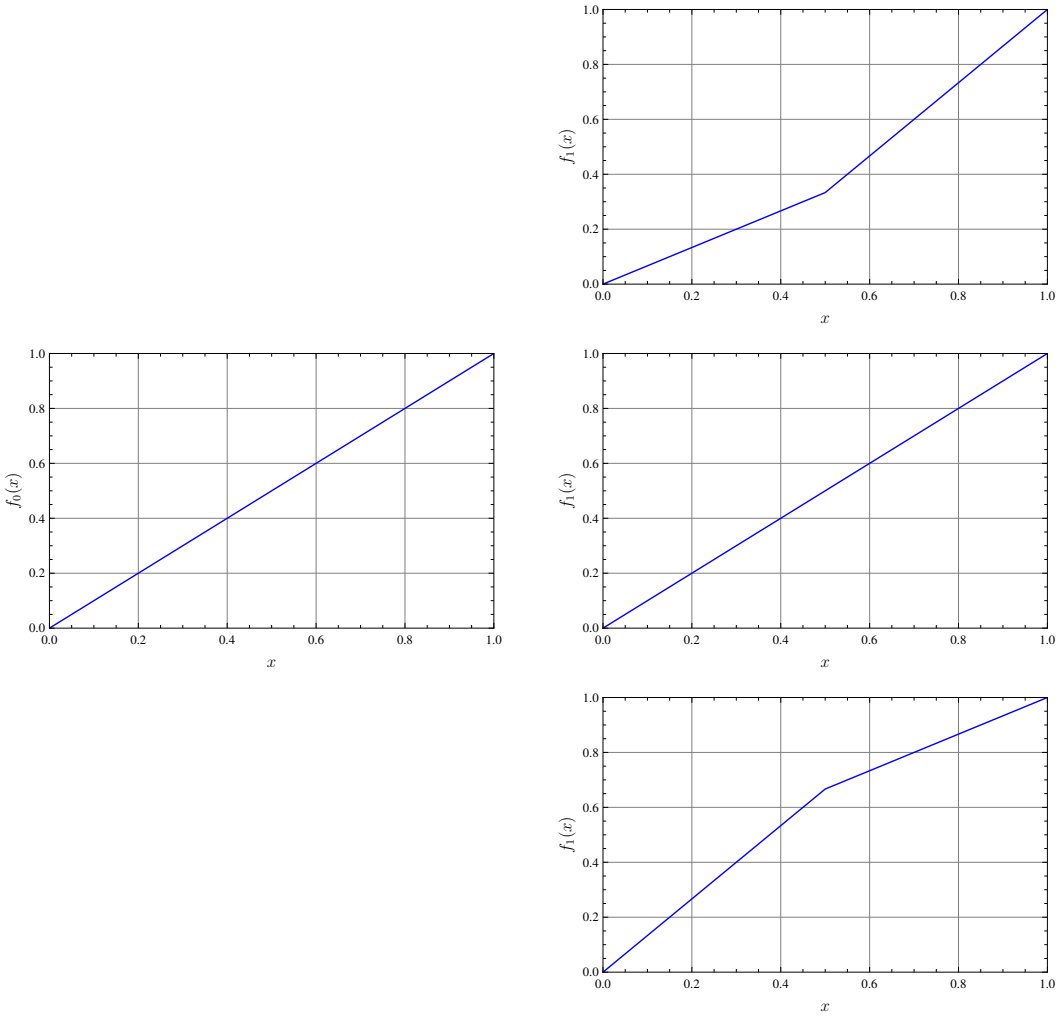


Figure 4.8 The first step in constructing the function f_α for $\alpha < \frac{1}{2}$ (top), $\alpha = \frac{1}{2}$ (middle), and $\alpha > \frac{1}{2}$ (bottom)

that this construction is applied recursively to each on the subintervals on which the function is linear.

Now we record some of the features of this function by proving a series of lemmata. First let us show that the definition of f_α makes sense.

1 Lemma *For each* $x \in [0,1]$ *and* $\alpha \in (0,1)$ *the limit* $\lim_{k\to\infty} f_k(x)$ *exists.*

Proof Using the linearity of f_k between the endpoints of the intervals used to define it, we compute

$$\begin{aligned} f_{k+1}(\tfrac{2j+1}{2^{k+1}}) - f_k(\tfrac{2j+1}{2^{k+1}}) &= (1-\alpha) f_k(\tfrac{j}{2^k}) + \alpha f_k(\tfrac{j+1}{2^k}) - \tfrac{1}{2}(f_k(\tfrac{j}{2^k}) + f_k(\tfrac{j+1}{2^k})) \\ &= (\alpha - \tfrac{1}{2})(f_k(\tfrac{j+1}{2^k}) - f_k(\tfrac{j}{2^k})), \end{aligned}$$

for $k \in \mathbb{Z}_{\geq 0}$ and $j \in \{0, 1, \dots, 2^k - 1\}$. Thus we have three cases.

1. When $\alpha = \frac{1}{2}$ we have $f_{k+1}(\frac{2j+1}{2^{k+1}}) = f_k(\frac{2j+1}{2^{k+1}})$, giving $f_{k+1} = f_k$.
2. When $\alpha < \frac{1}{2}$ then the sequence $(f_k(\frac{2j+1}{2^{k+1}}))_{k\in\mathbb{Z}_{\geq 0}}$ is strictly monotonically decreasing and bounded below by zero. Thus it converges.
3. When $\alpha > \frac{1}{2}$ then the sequence $(f_k(\frac{2j+1}{2^{k+1}}))_{k\in\mathbb{Z}_{\geq 0}}$ is strictly monotonically increasing and bounded above by zero. Thus it converges. ▼

2 Lemma *The function* f_α *is strictly monotonically increasing for* $\alpha \in (0,1)$.

Proof We shall first show that each of the functions f_k, $k \in \mathbb{Z}_{\geq 0}$, are strictly monotonically increasing. We show this by induction. It is clear that f_0 is strictly monotonically increasing. Now suppose that f_k is strictly monotonically increasing. We have

$$\begin{aligned} f_{k+1}(\tfrac{j}{2^k}) - f_{k+1}(\tfrac{2j+1}{2^{k+1}}) &= f_k(\tfrac{j}{2^k}) - f_{k+1}(\tfrac{2j+1}{2^{k+1}}) \\ &= f_k(\tfrac{j}{2^k}) - (1-\alpha) f_k(\tfrac{j}{2^k}) - \alpha f_k(\tfrac{j+1}{2^k}) \\ &= \alpha(f_k(\tfrac{j}{2^k}) - f_k(\tfrac{j+1}{2^k})) < 0 \end{aligned}$$

and

$$\begin{aligned} f_{k+1}(\tfrac{j+1}{2^k}) - f_{k+1}(\tfrac{2j+1}{2^{k+1}}) &= f_k(\tfrac{j+1}{2^k}) - f_{k+1}(\tfrac{2j+1}{2^{k+1}}) \\ &= f_k(\tfrac{j+1}{2^k}) - (1-\alpha) f_k(\tfrac{j}{2^k}) - \alpha f_k(\tfrac{j+1}{2^k}) \\ &= (1-\alpha)(f_k(\tfrac{j+1}{2^k}) - f_k(\tfrac{j}{2^k})) > 0. \end{aligned}$$

Thus we have

$$f_{k+1}(\tfrac{j}{2^k}) < f_{k+1}(\tfrac{2j+1}{2^{k+1}}) < f_{k+1}(\tfrac{j+1}{2^k}), \qquad j \in \{0, 1, \dots, 2^k - 1\}.$$

Since f_{k+1} is defined to be linear on the subintervals $[\frac{j}{2^{k+1}}, \frac{j+1}{2^{k+1}}]$, $j \in \{0, 1, \dots, 2^{k+1} - 1\}$, it follows that f_{k+1} is strictly monotonically increasing. It therefore follows that f_α is nondecreasing. To show that f_α is, in fact, strictly monotonically increasing, let $x_1, x_2 \in [0,1]$ satisfy $x_1 < x_2$. By Exercise 2.1.5 let $j, k \in \mathbb{Z}_{>0}$ satisfy $\frac{j}{2^k} \in (x_1, x_2)$. We consider three cases.

1. In the case when $\alpha = \frac{1}{2}$ it follows easily that f_α is strictly monotonically increasing since, as we showed in Lemma 1, $f_{1/2}(x) = x$.
2. If $\alpha > \frac{1}{2}$ we have
$$f_\alpha(x_1) \leq f_\alpha(\tfrac{j}{2^k}) = f_k(\tfrac{j}{2^k}) \leq f_k(x_2) \leq f_\alpha(x_2).$$
3. When $\alpha < \frac{1}{2}$ we have
$$f_\alpha(x_1) \leq f_k(x_1) < f_k(\tfrac{j}{2^k}) = f_\alpha(\tfrac{j}{2^k}) \leq f_\alpha(x_2).$$ ▼

3 Lemma *The function* f_α *is continuous for* $\alpha \in (0,1)$.

Proof Let us first make a preliminary construction. We call a sequence $([a_k, b_k])_{k\in\mathbb{Z}_{\geq 0}}$ of subintervals of $[0,1]$ ***binary*** if $a_0 = 1$ and $b_0 = 1$, and if, for each $k \in \mathbb{Z}$, either

1. $a_{k+1} = a_k$ and $b_{k+1} = b_k - \frac{1}{2^{k+1}}$ or
2. $a_{k+1} = a_k + \frac{1}{2^{k+1}}$ and $b_{k+1} = b_k$.

Thus, for example, either $[a_1, b_1] = [0, \frac{1}{2}]$ or $[a_1, b_2] = [\frac{1}{2}, 1]$. If $([a_k, b_k])_{k\in\mathbb{Z}_{\geq 0}}$ is a binary sequence, if $k \in \mathbb{Z}_{\geq 0}$, and if $a_{k+1} = a_k$, then we compute

$$\begin{aligned} f_\alpha(b_{k+1}) - f_\alpha(a_{k+1}) &= f_{k+1}(b_{k+1}) - f_{k+1}(a_{k+1}) \\ &= (1-\alpha)f_k(a_j) + \alpha f_k(b_j) - f_k(a_k) \\ &= \alpha(f_k(b_j) - f_k(a_k)). \end{aligned}$$

In the case when $b_{k+1} = b_k$ we similarly compute

$$\begin{aligned} f_\alpha(b_{k+1}) - f_\alpha(a_{k+1}) &= f_{k+1}(b_{k+1}) - f_{k+1}(a_{k+1}) \\ &= f_k(b_k) - ((1-\alpha)f_k(a_k) + \alpha f_k(b_k)) \\ &= (1-\alpha)(f_k(b_k) - f_k(a_k)). \end{aligned}$$

Therefore, using $f_0(b_0) - f_0(a_0) = 1$, a trivial inductive argument gives

$$f_\alpha(b_k) - f_\alpha(a_k) = \prod_{j=1}^{k} \sigma_j,$$

where $\sigma_j \in \{\alpha, 1-\alpha\}$, depending on whether $a_j = a_{j-1}$ or $b_j = b_{j=1}$. In any case, the above computations show that

$$f_\alpha(b_k) - f_\alpha(a_k) \leq \begin{cases} (1-\alpha)^k, & \alpha \leq \frac{1}{2}, \\ \alpha^k, & \alpha > \frac{1}{2}. \end{cases}$$

Now we show the continuity of f_α. Let $\epsilon \in \mathbb{R}_{>0}$ and let $N \in \mathbb{Z}_{>0}$ be sufficiently large that $(1-\alpha)^N < \frac{\epsilon}{2}$ if $\alpha \leq \frac{1}{2}$ or $\alpha^N < \frac{\epsilon}{2}$ if $\alpha > \frac{1}{2}$. Let $x_0 \in (0,1)$ and let $([a_k, b_k])_{k\in\mathbb{Z}_{\geq 0}}$ and $([a'_k, b'_k])_{k\in\mathbb{Z}_{\geq 0}}$ be binary intervals such that $a_N < x_0$, $x_0 < b'_N$, and $b_N = a'_N$. (By choosing N large enough we can ensure that $a_N > 0$ and $b'_N < 1$.) Then let $\delta \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^1(\delta, x_0) \subseteq [a_N, b'_N]$. Then we have

$$f_\alpha(b'_N) - f_\alpha(a_N) < \epsilon \quad \Longrightarrow \quad |f_\alpha(x) - f_\alpha(x_0)| < \epsilon,\ x \in \mathsf{B}^1(\delta, x_0),$$

by monotonicity of f_α. Continuity of f_α at 0 and 1 is shown in a similar manner, so we forgo the routine details. ▼

4 Lemma *Suppose that* $\mathrm{x} \in [0,1]$ *has a binary expansion* $\mathrm{x} = \sum_{\mathrm{j}=1}^{\infty} \frac{\mathrm{x}_\mathrm{j}}{2^\mathrm{j}}$ *with* $\mathrm{x}_\mathrm{j} \in \{0,1\}$, $\mathrm{j} \in \mathbb{Z}_{>0}$, *and suppose that the sets*

$$\{\mathrm{j} \in \mathbb{Z}_{>0} \mid \mathrm{x}_\mathrm{j} = 0\}, \quad \{\mathrm{j} \in \mathbb{Z}_{>0} \mid \mathrm{x}_\mathrm{j} = 1\}$$

are infinite, i.e., suppose that x *is irrational in base* 2. *Then* $\mathrm{f}'_\alpha(\mathrm{x}) = 0$. *In particular,* f_α *is differentiable with zero derivative on a subset of* $[0,1]$ *that has full measure.*

Proof Since x is irrational in base 2 it follows that for each $k \in \mathbb{Z}$ there exists a unique $j \in \mathbb{Z}_{\geq 0}$ such that $x \in (\frac{j}{2^k}, \frac{j+1}{2^k})$ (the binary irrationality of x ensures that the endpoints are not included in the interval $(\frac{j}{2^k}, \frac{j+1}{2^k})$). Moreover, if we write

$$\frac{j}{2^k} = \frac{y_1}{2} + \cdots + \frac{y_n}{2^n}$$

as the binary decimal expansion, then we have

$$a_k \triangleq \frac{l}{2^k} = \frac{y_1}{2} + \cdots + \frac{y_k}{2^k} < x < \frac{y_1}{2} + \cdots + \frac{y_k}{2^k} + \frac{1}{2^k} = \frac{l+1}{2^k} \triangleq b_k,$$

which implies that $y_j = x_j$, $j \in \{1, \dots, k\}$. Therefore, if $x_k = 0$ then

$$a_k = a_{k-1},\ b_k = a_{k-1} + \tfrac{1}{2^k} = \tfrac{a_{k-1}+b_{k-1}}{2},$$

and if $x_k = 1$ then

$$a_k = a_{k-1} + \tfrac{1}{2^k} = \tfrac{a_{k-1}+b_{k-1}}{2},\ b_k = b_{k-1}.$$

Therefore, if $x_k = 0$ then

$$\begin{aligned}\frac{f_\alpha(b_k) - f_\alpha(a_k)}{\frac{1}{2^k}} &= 2^k\Big((1-\alpha)f_{k-1}(a_{k-1}) + \alpha f_{k-1}(b_{k-1}) - f_{k-1}(a_{k-1})\Big)\\ &= 2^k\alpha(f_{k-1}(b_{k-1}) - f_{k-1}(a_{k-1}))\end{aligned}$$

and if $x_k = 1$ then

$$\begin{aligned}\frac{f_\alpha(b_k) - f_\alpha(a_k)}{\frac{1}{2^k}} &= 2^k\Big(f_{k-1}(b_{k-1}) - (1-\alpha)f_{k-1}(a_{k-1}) - \alpha f_{k-1}(b_{k-1})\Big)\\ &= 2^k(1-\alpha)(f_{k-1}(b_k) - f_{k-1}(a_k)).\end{aligned}$$

In either case, we have

$$\frac{f_\alpha(b_k) - f_\alpha(a_k)}{\frac{1}{2^k}} = 2(x_k + (-1)^{x_k}\alpha)2^{k-1}(f_{k-1}(b_k) - f_{k-1}(a_k)),$$

and so a simple induction gives

$$\frac{f_\alpha(b_k) - f_\alpha(a_k)}{\frac{1}{2^k}} = \prod_{j=1}^{k} 2(x_j + (-1)^{x_j}\alpha).$$

Thus

$$f'(x) = \lim_{k\to\infty} \frac{f_\alpha(b_k) - f_\alpha(a_k)}{\frac{1}{2^k}} = 0$$

since $\alpha, (1-\alpha) < 1$.

The final assertion follows since the irrational numbers in base 2 have measure 1. This can be proved in exactly the same way as it is proved in base 10; see Exercise 2.1.4. ▼

In Figure 4.9 we show the graph of f_α for a few α's. Since this function is contin-

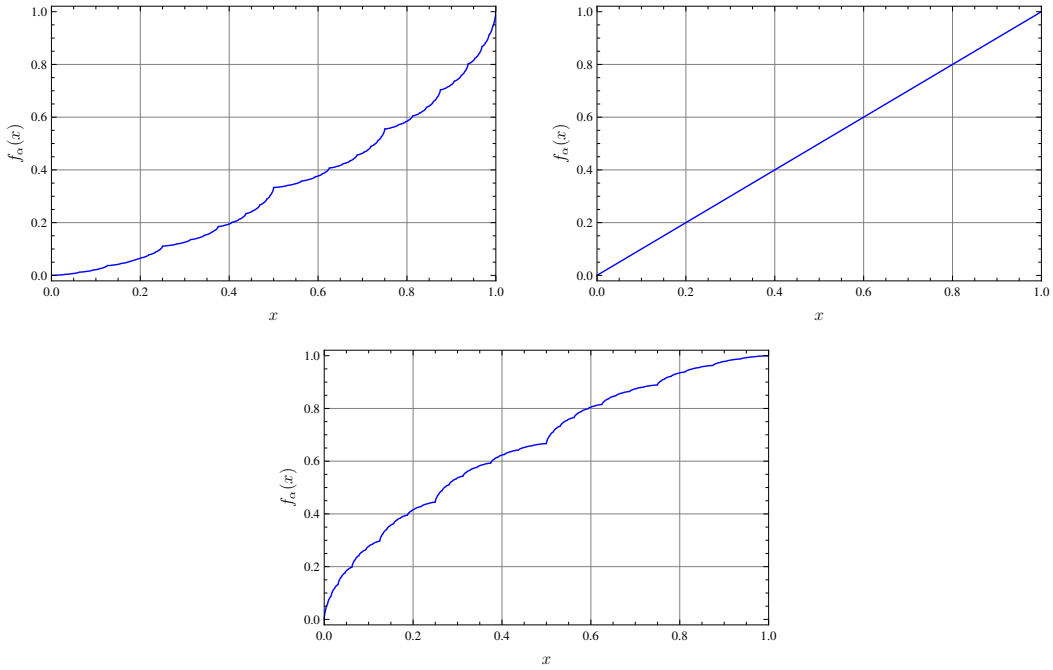


Figure 4.9 The function f_α for $\alpha = \frac{1}{3}$ (top left), $\alpha = \frac{1}{2}$ (top left), and $\alpha = \frac{2}{3}$ (bottom)

uous and monotonically increasing it is injective by Exercise **??**. Therefore, by the Domain Invariance Theorem, $f|(0,1)$ is a homeomorphism onto $(0,1)$. In particular, the Domain Invariance Theorem allows us to conclude that f^{-1} is continuous. This may not be perfectly clear from the construction.

Interestingly, there are a number of places where the function f_α comes up in applications. The most common of these is in the "bold play" strategy in probability. The situation is this. A gambler possesses a fraction $x \in [0,1]$ of what she wants, and wishes to play a game at even money (i.e., the same amount is either paid out on a loss or collected on a win) until the desired goal is achieved or the gambler is bankrupt. The probability of winning a game is the quantity $\alpha \in (0,1)$. It then turns out that the probability of eventual success is $f_\alpha(x)$. Note that if $\alpha < \frac{1}{2}$ (i.e., the game is biased against the gambler) then the gambler must start with a fraction

$x > \frac{1}{2}$ of the desired goal in order to have a greater that 50% chance of winning. This makes sense, I guess. •

Let us give another consequence of the Domain Invariance Theorem. One expects, and it is true, that two Euclidean spaces are homeomorphic if and only if they have the same dimension. Perhaps this seems "obvious," but it becomes less so the more one gets to know about the possible complex behaviour of continuous maps between Euclidean spaces and their subsets. Indeed, the following theorem is intimately and essentially connected to the Domain Invariance Theorem.

4.3.47 Theorem (Dimension Invariance Theorem) *The sets* $\mathbb{R}^n$ *and* $\mathbb{R}^m$ *are homeomorphic if and only if* $m = n$.

Proof Since "homeomorphic" is an equivalence relation, we suppose without loss of generality that $m \leq n$. Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is a homeomorphism. Consider the m-dimensional subspace V of $\mathbb{R}^n$ defined by

$$\mathsf{V} = \{x \in \mathbb{R}^n \mid x_{m+1} = \cdots = x_n = 0\}.$$

By Example 4.3.38–2 we know that V is homeomorphic to $\mathbb{R}^m$. That there exists a homeomorphism $g\colon \mathbb{R}^m \to \mathsf{V}$. Therefore, the composition of homeomorphisms being a homeomorphism, $g \circ f\colon \mathbb{R}^n \to \mathsf{V}$ is a homeomorphism. By the Domain Invariance Theorem this means that V is open in $\mathbb{R}^n$, and this is the case if and only if $m = n$. ■

4.3.8 Notes

Theorem 4.3.44 on "invariance of domain" is due to Brouwer [1912]. For a "basic" result, it is rather difficult to prove, and its proof properly belongs to the domains of dimension theory ([Hurewicz and Wallman 1941] is the classical reference here) and algebraic topology (Munkres [1984] has a good treatment).

Exercises

4.3.1 Answer the following questions:

(a) Verify that the Euclidean inner product satisfies the ***parallelogram law***:

$$\|x_1 + x_2\|^2_{\mathbb{R}^n} + \|x_1 - x_2\|^2_{\mathbb{R}^n} = 2\Big(\|x_1\|^2_{\mathbb{R}^n} + \|x_2\|^2_{\mathbb{R}^n}\Big).$$

(b) Give an interpretation of the parallelogram law in $\mathbb{R}^2$.

(c) Verify that the Euclidean inner product satisfies the ***polarisation identity***:

$$4\langle x_1, x_2\rangle_{\mathbb{R}^n} = \langle x_1 + x_2, x_1 + x_2\rangle_{\mathbb{R}^n} - \langle x_1 - x_2, x_1 - x_2\rangle_{\mathbb{R}^n}.$$

4.3.2 Let $A \subseteq \mathbb{R}^n$ and let $f\colon A \to \mathbb{R}^m$ be a map. Show that f is continuous at $x_0 \in A$ if and only if the components of f are continuous at x_0.

4.3.3 For $A \subseteq \mathbb{R}^n$, show that $f\colon A \to \mathbb{R}^m$ is continuous if and only if $f^{-1}(B)$ is relatively closed in A for every closed subset B of $\mathbb{R}^m$.

4.3.4 Is the preimage of a (path) connected set under a continuous map (path) connected?

4.3.5 Consider the subset

$$S = \{(x_1, 0) \in \mathbb{R}^2 \mid x_1 \in \mathbb{R}\} \cup \{(0, x_2) \in \mathbb{R}^2 \mid x_2 > 0\}$$

of $\mathbb{R}^2$ and the subset $A = \{(x_1, 0) \mid x_1 \in \mathbb{R}\}$ of S.

(a) Is A relatively open in S?

(b) Is A relatively closed in S?

(c) Determine $\text{int}_S(A)$, $\text{cl}_S(A)$, and $\text{bd}_S(A)$.

4.3.6 Show that the image of an affine map is an affine subspace.

4.3.7 Let $\boldsymbol{R} \in \mathsf{O}(3)$.

(a) Show that $\boldsymbol{R}$ has at least one real eigenvalue and that its magnitude must 1.

Let $\boldsymbol{v}$ be an eigenvector for the real eigenvalue ± 1 and let $\boldsymbol{v}^\perp$ be the subspace orthogonal to $\boldsymbol{v}$.

(a) Show that $\boldsymbol{R}(\boldsymbol{v}^\perp) \subseteq \boldsymbol{v}^\perp$.

(b) Argue that if $\boldsymbol{R} \neq \boldsymbol{I}_3$ then $\boldsymbol{R}$ has no eigenvectors that are not collinear with $\boldsymbol{v}$?
Hint: *Use the fact that* $\mathbf{v}^\perp$ *is two-dimensional.*

(c) Which of the preceding parts of the exercise fail if $\boldsymbol{R} \in \mathsf{O}(n)$ for $n \neq 3$?

4.3.8 Answer the following questions.

(a) Show that $\mathsf{O}(n)$ is a group with the group operation given by matrix multiplication.

(b) Is $\mathsf{O}(n)$ a subspace of the $\mathbb{R}$-vector space $\text{Mat}_{n\times n}(\mathbb{R})$?

4.3.9 Show that if $\boldsymbol{R} \in \mathsf{O}(n)$ then $\det \boldsymbol{R} \in \{-1, 1\}$.

4.3.10 Show that if $\boldsymbol{R} \in \mathsf{O}(n)$ and if $\lambda \in \mathbb{C}$ is an eigenvalue for the complexification $\boldsymbol{R}_{\mathbb{C}}$, then $|\lambda| = 1$.

4.3.11 Show that $\mathsf{E}(n)$ is a group with the group operation of map composition. Be sure to explicitly given the formulae for the product of two elements and the inverse of an element.

4.3.12 Let $I \subseteq \mathbb{R}$ be an open interval. Explicitly construct a homeomorphism from I to $\mathbb{R}$.

4.3.13 Show that $\mathsf{B}^n(1, \mathbf{0})$, the open ball of radius 1 centred at the origin in $\mathbb{R}^n$, is homeomorphic to $\mathbb{R}^n$.

4.3.14 Show that the following sets are homeomorphic:

1. $\mathbb{D}^n = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_{\mathbb{R}^n} \leq 1\}$;
2. $\mathbb{D}^n(r, \boldsymbol{x}_0) = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{x}_0\|_{\mathbb{R}^n} \leq \epsilon\}$ where $r \in \mathbb{R}_{>0}$ and $\boldsymbol{x}_0 \in \mathbb{R}^n$;
3. a fat compact rectangle R;
4. $\mathbb{S}^n_+ = \{\boldsymbol{x} \in \mathbb{S}^n \subseteq \mathbb{R}^{n+1} \mid x_{n+1} \geq 0\}$.

4.3.15 Show that the following sets are homeomorphic:

1. $\mathbb{S}^n = \{\boldsymbol{x} \in \mathbb{R}^{n+1} \mid \|\boldsymbol{x}\|_{\mathbb{R}^{n+1}} = 1\}$;
2. $\mathbb{S}^n(r, \boldsymbol{x}_0) = \{\boldsymbol{x} \in \mathbb{R}^{n+1} \mid \|\boldsymbol{x} - \boldsymbol{x}_0\|_{\mathbb{R}^{n+1}} = \epsilon\}$ where $r \in \mathbb{R}_{>0}$ and $\boldsymbol{x}_0 \in \mathbb{R}^{n+1}$;
3. $\text{bd}(R)$ where $R \subseteq \mathbb{R}^{n+1}$ is a fat compact rectangle.

Section 4.4

Differentiable multivariable functions

Unlike our discussion of continuity, the notion of differentiability for maps involving multiple variables is not so much a straightforward generalisation of the single-variable case. For example, we shall see that the appropriate way to think about the derivative in the multivariable case (and therefore, by specialisation, the single-variable case) is as a linear map. This turns out to be an important conceptual idea in understanding just what the derivative "is."

Some of the ideas in this section can be illustrated using single-variable examples, and we refer to Section 3.2 for these. However, there are phenomenon in the multivariable case that do not arise in the single-variable case, and we give particular examples to exhibit these phenomenon.

Do I need to read this section? If you want to understand differentiability of multivariable functions, and you do not already, then you need to read this section. It is true that we do not make a great deal of use of the material in this section, but it does come up on occasion. •

missing stuff

4.4.1 Definition and basic properties of the derivative

The definition of what it means for a map to be differentiable immediately emphasises the linear algebraic character that is essential to the picture in higher-dimensions. The definition we give for the derivative in this case should be thought of as the generalisation of Proposition 3.2.4; let us therefore present a result along these lines that will ensure that our definition of derivative makes sense.

4.4.1 Proposition (Uniqueness of linear approximation) *Let $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}}$ be an open set and let $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^{\mathrm{m}}$ be a map. For $\mathbf{x}_0 \in \mathrm{U}$, there exists at most one $\mathsf{L} \in \mathrm{L}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}})$ such that*

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \frac{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) - \mathsf{L}(\mathbf{x} - \mathbf{x}_0))\|_{\mathbb{R}^{\mathrm{m}}}}{\|\mathbf{x} - \mathbf{x}_0\|_{\mathbb{R}^{\mathrm{n}}}} = 0. \tag{4.15}$$

Proof Suppose there are two such maps L_1 and L_2. For any $x \in U$, we may write $x = x_0 + av$ for some $a \in \mathbb{R}_{>0}$ and $v \in \mathbb{R}^n$ such that $\|v\|_{\mathbb{R}^n} = 1$. We compute

$$\begin{aligned}
\|\mathsf{L}_1(v) - \mathsf{L}_2(v)\|_{\mathbb{R}^m} &= \frac{\|\mathsf{L}_1(x - x_0) - \mathsf{L}_2(x - x_0)\|_{\mathbb{R}^m}}{\|x - x_0\|_{\mathbb{R}^n}} \\
&= \frac{\|-f(x) + f(x_0) + \mathsf{L}_1(x - x_0) + f(x) - f(x_0) - \mathsf{L}_2(x - x_0)\|_{\mathbb{R}^m}}{\|x - x_0\|_{\mathbb{R}^n}} \\
&\leq \frac{\|f(x) - f(x_0) - \mathsf{L}_1(x - x_0)\|_{\mathbb{R}^m}}{\|x - x_0\|_{\mathbb{R}^n}} + \frac{\|f(x) - f(x_0) - \mathsf{L}_2(x - x_0)\|_{\mathbb{R}^m}}{\|x - x_0\|_{\mathbb{R}^n}}.
\end{aligned}$$

Since L_1 and L_2 both satisfy (4.15), as we let $x \to x_0$ the right-hand side goes to zero showing that $\|\mathsf{L}_1(v) - \mathsf{L}_2(v)\|_{\mathbb{R}^m} = \|(\mathsf{L}_1 - \mathsf{L}_2)(v)\|_{\mathbb{R}^m} = 0$ for every v with $\|v\|_{\mathbb{R}^n} = 1$. Thus $\mathsf{L}_1 - \mathsf{L}_2$ is the trivial map sending any vector to zero, or equivalently $\mathsf{L}_1 = \mathsf{L}_2$. ■

We can now state the definition of the derivative for multivariable maps.

4.4.2 Definition (Derivative and differentiable map) Let $U \subseteq \mathbb{R}^n$ be an open subset and let $f\colon U \to \mathbb{R}^m$ be a map.

(i) The map f is ***differentiable at*** $\mathbf{x_0} \in \mathbf{U}$ if there exists a linear map $\mathsf{L}_{f,x_0}\colon \mathbb{R}^n \to \mathbb{R}^m$ such that
$$\lim_{x\to x_0} \frac{\|f(x) - f(x_0) - \mathsf{L}_{f,x_0}(x - x_0)\|_{\mathbb{R}^m}}{\|x - x_0\|_{\mathbb{R}^n}} = 0.$$

(ii) If f is differentiable at x_0, then the linear map L_{f,x_0} is denoted by $Df(x_0)$ and is called the ***derivative*** of f at x_0.

(iii) If f is differentiable at each point $x \in U$, then f is ***differentiable***.

(iv) If f is differentiable and if the map $x \mapsto Df(x)$ is continuous (using any norm one wishes on $\mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$) then f is ***continuously differentiable***, or of ***class*** $\mathbf{C^1}$. •

Sometimes the derivative is called the ***total derivative*** or the ***Fréchet derivative***. Similarly, differentiability in the sense of the preceding definition is sometimes called ***Fréchet differentiability***. The reason for this is that the existence of this derivative implies the existence of other derivatives, such as the directional derivative which we discuss in Section 4.4.3.

4.4.3 Notation (Evaluation of the derivative) Since $Df(x_0) \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$, we can write $Df(x_0)(v)$ as the image of $v \in \mathbb{R}^n$ under the derivative thought of as a linear map. To avoid the somewhat cumbersome looking double parentheses, we shall often write $Df(x_0) \cdot v$ instead of $Df(x_0)(v)$. •

With the derivative defined, it is now possible to talk about higher-order derivatives in a systematic way. We let $U \subseteq \mathbb{R}^n$ be open and let $f\colon U \to \mathbb{R}^m$ be continuously differentiable. The derivative is then a map $U \ni x \mapsto Df(x) \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$. Given that from Section 4.1.3 we have a norm on $\mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$, this map is a candidate for having its derivative defined. The derivative of Df at $x_0 \in U$, if it exists, is the linear map $D^2f \in \mathrm{L}(\mathbb{R}^n; \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m))(x_0)$ satisfying
$$\lim_{x\to x_0} \frac{\|Df(x) - Df(x_0) - D^2f(x_0) \cdot (x - x_0)\|_{\mathbb{R}^n,\mathbb{R}^m}}{\|x - x_0\|_{\mathbb{R}^n}} = 0.$$
By Proposition **??** we implicitly think of D^2f as being an element of $\mathrm{L}^2(\mathbb{R}^n; \mathbb{R}^m)$. Now we can carry on this process recursively to define derivatives of arbitrary order.

4.4.4 Definition (Higher-order derivatives) Let $U \subseteq \mathbb{R}^n$ be open, let $f\colon U \to \mathbb{R}^m$ be a function, let $r \in \mathbb{Z}_{>0}$, and suppose that f is $(r-1)$ times differentiable with $G\colon U \to \mathrm{L}^{r-1}(\mathbb{R}^n; \mathbb{R}^m)$ denoting the $(r-1)$st derivative.

(i) The map f is **r** ***times continuously differentiable at*** $\mathbf{x_0} \in \mathbf{U}$ if there exists $DG(x_0) \in \mathrm{L}(\mathbb{R}^n; \mathrm{L}^{r-1}(\mathbb{R}^n; \mathbb{R}^m))$ such that
$$\lim_{x\to x_0} \frac{\|G(x) - G(x_0) - DG(x_0) \cdot (x - x_0)\|_{\mathbb{R}^n,\mathrm{L}^{r-1}(\mathbb{R}^n;\mathbb{R}^m)}}{\|x - x_0\|_{\mathbb{R}^n}} = 0. \tag{4.16}$$

(ii) If (4.16) holds then the map $\boldsymbol{DG}(\boldsymbol{x}_0)$ is identified, using Proposition **??**, with the multilinear map $\boldsymbol{D}^r\boldsymbol{f}(\boldsymbol{x}_0) \in \mathrm{L}^r(\mathbb{R}^n;\mathbb{R}^m)$ and called the ***r*th *derivative*** of $\boldsymbol{f}$ at $\boldsymbol{x}_0$.

(iii) If $\boldsymbol{f}$ is r times differentiable at each point $\boldsymbol{x} \in U$, then $\boldsymbol{f}$ is **r *times differentiable***.

(iv) If $\boldsymbol{f}$ is r times differentiable and if the function $\boldsymbol{x} \mapsto \boldsymbol{D}^r\boldsymbol{f}(\boldsymbol{x})$ is continuous, then $\boldsymbol{f}$ is **r *times continuously differentiable***, or of ***class* $\mathbf{C}^\mathbf{r}$**.

If $\boldsymbol{f}$ is of class C^r for each $r \in \mathbb{Z}_{>0}$, then $\boldsymbol{f}$ is ***infinitely differentiable***, or of ***class* $\mathbf{C}^\infty$**. •

The following result gives an important property of higher-order derivatives. Parts of the proof rely on properties of the derivative we have yet to prove. Specifically, the proof properly belongs after the proof of Theorem 4.4.33, but we give it here since this is where it fits best in terms of the flow of ideas.

4.4.5 Theorem (The derivative is symmetric) *If* $\mathrm{U} \subseteq \mathbb{R}^\mathrm{n}$ *is open and if* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^\mathrm{m}$ *is of class* C^r*, then* $\mathbf{D}^\mathrm{r}\mathbf{f} \in \mathrm{S}^\mathrm{r}(\mathbb{R}^\mathrm{n};\mathbb{R}^\mathrm{m})$.

Proof By Proposition 4.4.17 we can assume, without loss of generality, that $m = 1$. We thus take $m = 1$ and write our function as f. When $r = 1$ we have $S^1(\mathbb{R}^n;\mathbb{R}) = \mathrm{L}(\mathbb{R}^n;\mathbb{R})$ so the result is vacuous in this case. We next consider the case when $r = 2$. Let $\boldsymbol{x}_0 \in U$ and let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$. Let $a \in \mathbb{R}_{>0}$ be sufficiently small that $\boldsymbol{x}_0 + s\boldsymbol{u} + t\boldsymbol{v} \in U$ for all $(s,t) \in \mathsf{B}^2(a,(0,0))$, this being possible since U is open and since the map $(s,t) \mapsto \boldsymbol{x}_0 + s\boldsymbol{u} + t\boldsymbol{v}$ is linear, and so infinitely differentiable by Corollary 4.4.9. Then define $g\colon \mathsf{B}^2(a,(0,0)) \to \mathbb{R}$ by

$$g(s,t) = f(\boldsymbol{x}_0 + s\boldsymbol{u} + t\boldsymbol{v}).$$

The Chain Rule (Theorem 4.4.49) implies that g is of class C^2. We then compute the following iterated partial derivatives using the Chain Rule and Proposition 4.4.7:

$$\begin{aligned}
\boldsymbol{D}_1 g(s,t) \cdot 1 &= \boldsymbol{D}f(\boldsymbol{x}_0 + s\boldsymbol{u} + t\boldsymbol{v}) \cdot \boldsymbol{u},\\
\boldsymbol{D}_2 g(s,t) \cdot 1 &= \boldsymbol{D}f(\boldsymbol{x}_0 + s\boldsymbol{u} + t\boldsymbol{v}) \cdot \boldsymbol{v},\\
\boldsymbol{D}_2\boldsymbol{D}_1 g(s,t) \cdot (1,1) &= \boldsymbol{D}^2 f(\boldsymbol{x}_0 + s\boldsymbol{u} + t\boldsymbol{v}) \cdot (\boldsymbol{v},\boldsymbol{u}),\\
\boldsymbol{D}_1\boldsymbol{D}_2 g(s,t) \cdot (1,1) &= \boldsymbol{D}^2 f(\boldsymbol{x}_0 + s\boldsymbol{u} + t\boldsymbol{v}) \cdot (\boldsymbol{u},\boldsymbol{v}).
\end{aligned}$$

Thus the result for $r = 2$ will follow if $\boldsymbol{D}_1\boldsymbol{D}_2 g(0,0) = \boldsymbol{D}_2\boldsymbol{D}_1 g(0,0)$. This, however, is a special case of Theorem 4.4.33.

For $r > 2$ we proceed by induction, assuming the result true for $r = s - 1$ and then supposing that f is of class C^r. For $\boldsymbol{x} \in U$ and $\boldsymbol{v}_1, \dots, \boldsymbol{v}_s \in \mathbb{R}^n$ we compute

$$\begin{aligned}
\boldsymbol{D}^s f(\boldsymbol{x}) \cdot (\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_s) &= (\boldsymbol{D}^2(\boldsymbol{D}^{s-2} f)(\boldsymbol{x}) \cdot (\boldsymbol{v}_1, \boldsymbol{v}_2)) \cdot (\boldsymbol{v}_3, \dots, \boldsymbol{v}_s)\\
&= (\boldsymbol{D}^2(\boldsymbol{D}^{s-2} f)(\boldsymbol{x}) \cdot (\boldsymbol{v}_2, \boldsymbol{v}_1)) \cdot (\boldsymbol{v}_3, \dots, \boldsymbol{v}_s)\\
&= \boldsymbol{D}^s f(\boldsymbol{x}) \cdot (\boldsymbol{v}_2, \boldsymbol{v}_1, \dots, \boldsymbol{v}_s),
\end{aligned}$$

showing that

$$\boldsymbol{D}^s f(\boldsymbol{x}) \cdot (\boldsymbol{v}_{\sigma(1)}, \boldsymbol{v}_{\sigma(2)}, \dots, \boldsymbol{v}_{\sigma(s)}) = \boldsymbol{D}^s f(\boldsymbol{x}) \cdot (\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_s)$$

for $\sigma = (1\ 2)$. Now let $\sigma \in \mathfrak{S}_{s-1}$ and by the induction hypothesis note that

$$\boldsymbol{D}^{s-1} f(\boldsymbol{x}) \cdot (\boldsymbol{v}_{\sigma(1)}, \dots, \boldsymbol{v}_{\sigma(s-1)}) = \boldsymbol{D}^{s-1} f(\boldsymbol{x}) \cdot (\boldsymbol{v}_1, \dots, \boldsymbol{v}_{s-1})$$

for all $x \in U$ and $v_1, \dots, v_{s-1}$. Then, by Proposition 4.4.7, we have, for any $v_0 \in \mathbb{R}^n$,

$$\begin{aligned} \boldsymbol{D}^s f(x) \cdot (v_0, v_{\sigma(1)}, \dots, v_{\sigma(s-1)}) &= (\boldsymbol{D}(\boldsymbol{D}^{s-1} f)(x) \cdot v_0) \cdot (v_{\sigma(1)}, \dots, v_{\sigma(s-1)}) \\ &= (\boldsymbol{D}(\boldsymbol{D}^{s-1} f)(x) \cdot v_0) \cdot (v_1, \dots, v_{s-1}) \\ &= \boldsymbol{D}^s f(x) \cdot (v_0, v_1, \dots, v_{s-1}), \end{aligned}$$

giving

$$\boldsymbol{D}^s f(x) \cdot (v_{\sigma(1)}, v_{\sigma(2)}, \dots, v_{\sigma(s)}) = \boldsymbol{D}^s f(x) \cdot (v_1, v_2, \dots, v_s)$$

when σ leaves 1 fixed. Now, by Exercise **??** any permutation $\sigma \in \mathfrak{S}_s$ can be written as a finite product of (1 2) and permutations leaving 1 fixed. From this the result follows. ■

We now deal with the problem of having potentially competing definitions of the derivative for a $\mathbb{R}$-valued function of a single real variable. Let us resolve this.

4.4.6 Theorem (Consistency of differentiability definitions for $\mathbb{R}$-valued functions of a single variable) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an open interval, let* $\mathrm{f} \colon \mathrm{I} \to \mathbb{R}$, *let* $\mathrm{x}_0 \in \mathrm{I}$, *and let* $\mathrm{r} \in \mathbb{Z}_{\geq 0}$. *Then* f *is* r *times differentiable at* x_0 *in the sense of Definition 3.2.5 if and only if* f *is* r *times differentiable at* x_0 *in the sense of Definition 4.4.4. Moreover, if* f *is* r *times continuously differentiable at* x_0 *then*

$$\mathbf{D}^{\mathrm{r}} \mathrm{f}(\mathrm{x}_0)(\mathrm{v}_1, \dots, \mathrm{v}_{\mathrm{r}}) = \mathrm{f}^{(\mathrm{r})}(\mathrm{x}_0)\mathrm{v}_1 \cdots \mathrm{v}_{\mathrm{r}}$$

for every $\mathrm{v}_1, \dots, \mathrm{v}_{\mathrm{r}} \in \mathbb{R}$.

Proof We first observe that there is a natural isomorphism from $\mathbb{R}$ to $S^r(\mathbb{R}; \mathbb{R})$ assigning to $a \in \mathbb{R}$ the symmetric multilinear map

$$(v_1, \dots, v_r) \mapsto a\, v_1 \cdots v_r.$$

This isomorphism is easily verified to preserve the standard norms on $\mathbb{R}$ and $S^r(\mathbb{R}; \mathbb{R})$. We shall implicitly use this isomorphism is the proof.

For $r = 0$ the result is clearly true since 0 times differentiable means continuous in the case of each definition. Assume the result is true for $r \in \{0, 1, \dots, k-1\}$. Thus assume that existence of $\boldsymbol{D}^{k-1} f(x_0)$ is equivalent to existence of $f^{(k-1)}(x_0)$ and that

$$\boldsymbol{D}^{k-1} f(x_0)(v_1, \dots, v_{k-1}) = f^{(k-1)}(x_0) v_1 \cdots v_{k-1}$$

for all $v_1, \dots, v_{k-1} \in \mathbb{R}$.

First let us suppose that f is k times differentiable at x_0 in the sense of Definition 4.4.4. Then $\boldsymbol{D}^{k-1} f$ is continuous at x_0. Let $g \colon I \to \mathbb{R}$ be defined by asking that $g(x)$ be the image of $\boldsymbol{D}^{k-1} f(x)$ under the isomorphism of $S^{k-1}(\mathbb{R}; \mathbb{R})$ with $\mathbb{R}$. It then holds that g is differentiable at x_0 in the sense of Definition 4.4.4 since $\boldsymbol{D}^{k-1} f$ is differentiable at x_0 in the sense of Definition 4.4.4. By the induction hypothesis it then follows from Proposition 3.2.4 that $f^{(k-1)}$ is differentiable in the sense of Definition 3.2.5. This means that f is k times differentiable at x_0 in the sense of Definition 3.2.5.

Next suppose that f is k times differentiable at x_0 in the sense of Definition 4.4.4. Let $\mathsf{L} \colon I \to S^{k-1}(\mathbb{R}; \mathbb{R})$ be defined by asking that $\mathsf{L}(x)$ be the image of $f^{(k-1)}(x)$ under isomorphism of $\mathbb{R}$ with $S^{k-1}(\mathbb{R}; \mathbb{R})$. Since $f^{(k-1)}$ is differentiable at x_0 in the sense of Definition 3.2.5 it follows that L is differentiable at x_0 in the sense of Definition 3.2.5. By the induction hypothesis and Proposition 3.2.4 it follows that $\boldsymbol{D}^{k-1} f$ is differentiable

at x_0 in the sense of Definition 4.4.4. This means that f is k times differentiable at x_0 in the sense of Definition 4.4.4.

For the final assertion of the proof, for fixed $v_1, \dots, v_{k-1} \in \mathbb{R}$ consider the function $h\colon I \to \mathbb{R}$ defined by

$$h(x) = f^{(k-1)}(x)v_1 \cdots v_{k-1}.$$

We claim that h is differentiable at x_0 if f is k times differentiable at x_0. We use the derivative of Definition 3.2.5 to verify this assertion. We have

$$\begin{aligned}\lim_{x\to x_0} \frac{h(x) - h(x_0)}{x - x_0} &= \lim_{x\to x_0} \frac{f^{(k-1)}(x)v_1 \cdots v_{k-1} - f^{(k-1)}(x_0)v_1 \cdots v_{k-1}}{x - x_0}\\ &= \lim_{x\to x_0} \frac{f^{(k-1)}(x) - f^{(k-1)}(x_0)}{x - x_0} v_1 \cdots v_{k-1}\\ &= f^{(k)}(x_0)v_1 \cdots v_{k-1},\end{aligned}$$

where we have used Proposition 2.3.23 and Proposition 2.3.29. This gives the differentiability of h at x_0 as well as an explicit formula for the derivative. Using Proposition 3.2.4 we have

$$\boldsymbol{D}h(x_0) \cdot v_0 = f^{(k)}(x_0)v_0v_1 \cdots v_{k-1},$$

which gives the theorem. ∎

The reader will have noticed that we give no examples to illustrate the multi-dimensional derivative. There is a reason for this. Based on the definition it is not that easy to actually compute the derivative in multiple-dimensions. However, it is actually easy to compute this derivative in practice only knowing how to differentiate $\mathbb{R}$-valued functions of a single variable. But the development of this connection is actually a little involved, and we postpone it until Theorem 4.4.22, at which time we will also provide some examples.

We close this section with a useful characterisation of differentiability that can simplify how one handles computations with derivatives.

4.4.7 Proposition (Swapping of differentiation and evaluation) *For $\mathrm{U} \subseteq \mathbb{R}^n$ open, for $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^m$, and for $\mathbf{x}_0 \in \mathrm{U}$, the following statements are equivalent:*

(i) *$\mathbf{f}$ is r times differentiable at $\mathbf{x}_0$;*

(ii) *$\mathbf{f}$ is $r-1$ times continuously differentiable in a neighbourhood of $\mathbf{x}_0$ and, for each $\mathbf{v}_1, \dots, \mathbf{v}_{r-1} \in \mathbb{R}^n$, the map $\boldsymbol{\delta}_{\mathbf{f};\mathbf{v}_1,\dots,\mathbf{v}_{r-1}}\colon \mathrm{U} \to \mathbb{R}^m$ defined by*

$$\boldsymbol{\delta}_{\mathbf{f};\mathbf{v}_1,\dots,\mathbf{v}_{r-1}}(\mathbf{x}) = \mathbf{D}^{r-1}\mathbf{f}(\mathbf{x}) \cdot (\mathbf{v}_1, \dots, \mathbf{v}_{r-1})$$

is differentiable at $\mathbf{x}_0$.

Moreover, if $\mathbf{f}$ is r times differentiable at $\mathbf{x}_0 \in \mathrm{U}$ then

$$\mathbf{D}^r\mathbf{f}(\mathbf{x}_0) \cdot (\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{r-1}) = \mathbf{D}\boldsymbol{\delta}_{\mathbf{f};\mathbf{v}_1,\dots,\mathbf{v}_{r-1}}(\mathbf{x}_0) \cdot \mathbf{v}_0 \tag{4.17}$$

for every $\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{r-1} \in \mathbb{R}^n$.

Proof First suppose that f is r times differentiable at x_0. From Proposition 4.4.35 it follows that f is $r-1$ times continuously differentiable in a neighbourhood of x_0. For $v_1, \dots, v_{r-1} \in \mathbb{R}^n$ let us define $\mathrm{Ev}_{v_1,\dots,v_{r-1}} : \mathrm{L}^{r-1}(\mathbb{R}^n; \mathbb{R}^m; \to)\mathbb{R}^m$ by

$$\mathrm{Ev}_{v_1,\dots,v_{r-1}}(\mathsf{L}) = \mathsf{L}(v_1, \dots, v_{r-1}).$$

Then we have $\delta_{f;v_1,\dots,v_{r-1}} = \mathrm{Ev}_{v_1,\dots,v_{r-1}} \circ D^{r-1}f$. Since $\mathrm{Ev}_{v_1,\dots,v_{r-1}}$ is linear (this is a simple verification), it follows from Corollary 4.4.9 that it is infinitely differentiable. Thus $\delta_{f;v_1,\dots,v_{r-1}}$ is differentiable by the Chain Rule, Theorem 4.4.49. Moreover, also by the Chain Rule and Corollary 4.4.9, it follows that

$$\begin{aligned} D\delta_{f;v_1,\dots,v_{r-1}}(x_0) \cdot v_0 &= \mathrm{Ev}_{v_1,\dots,v_{r-1}}(D(D^{r-1}f)(x_0) \cdot v_0) \\ &= D^r f(x_0) \cdot (v_0, v_1, \dots, v_{r-1}), \end{aligned}$$

using Proposition **??**. This gives (4.17).

Next suppose that f is $r-1$ times continuously differentiable in a neighbourhood of x_0 and that $\delta_{f;v_1,\dots,v_{r-1}}$ is differentiable at x_0 for every $v_1, \dots, v_{r-1} \in \mathbb{R}^n$. To show that f is r times differentiable at x_0 we claim that it suffices to show that the components of $D^{r-1}f$ are differentiable at x_0 (see Definition **??** for definition of the components of a multilinear map). That this is so essentially follows from Proposition 4.4.17 below. However, the "essentially" warrants a little explanation.

In Proposition 4.4.17 we show that a map taking values in $\mathbb{R}^m$ is differentiable if and only if each of its components is differentiable. But here we are not talking about a map taking values in $\mathbb{R}^m$, but taking values in $\mathrm{L}^{r-1}(\mathbb{R}^n; \mathbb{R}^m)$. But, the assignment taking a multilinear map in $\mathrm{L}^{r-1}(\mathbb{R}^n; \mathbb{R}^m)$ to its components is a linear isomorphism taking values in $\mathbb{R}^{mn^{r-1}}$. Moreover, the Frobenius norm on $\mathrm{L}^{r-1}(\mathbb{R}^n; \mathbb{R}^m)$ is "the same as" the Euclidean norm on $\mathbb{R}^{mn^{r-1}}$ under this isomorphism; in the language of ***missing stuff***, the isomorphism is norm-preserving. Therefore, Proposition 4.4.17 can essentially be applied to assert that $D^{r-1}f$ is differentiable at x_0 if its components are differentiable at x_0.

The matter of showing that the components of $D^{r-1}f$ are differentiable at x_0 is straightforward. Indeed, the components of $D^{r-1}f$ are simply given by the $\mathbb{R}$-valued functions

$$x \mapsto (D^{r-1}f(x) \cdot (e_{j_1}, \dots, e_{j_{r-1}}))_a = (\delta_{f;e_{j_1},\dots,e_{j_{r-1}}}(x))_a, \qquad j_1, \dots, j_{r-1} \in \{1, \dots, n\},\ a \in \{1, \dots, m\},$$

defined in a neighbourhood of x_0. By assumption and by Proposition 4.4.17 these functions are, indeed, differentiable at x_0. ■

While in these volumes we do not adhere to presentation dictated solely by logical implications always flowing forwards, we do feel compelled to warn the reader that in this section we make an abuse of logical ordering so dire as to merit comment. We shall in the next several sections (and already in the proofs of Theorem 4.4.5 and Proposition 4.4.7 above) make repeated and crucial use of the multivariable Chain Rule which we do not prove until Theorem 4.4.49. A reader who might be bothered by this can go ahead and read the Chain Rule and its proof right now since the proof relies only on ideas that are presently at our disposal.

4.4.2 Derivatives of multilinear maps

In this section we consider a special class of maps, and show that they are infinitely differentiable and compute their derivatives of all orders. The maps we consider are multilinear maps $L\colon \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}^m$. It will turn out that these maps come up many times for various reasons, and for this reason it is useful to determine their derivatives. Moreover, it is a good exercise in using the definition of the derivatives to compute the derivatives of multilinear maps.

Since derivatives are themselves multilinear maps, it will be useful to discriminate notationally between points in the domain of the map and points in the domain of the derivative of the map. Thus we shall write a point in $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ as $(x_1, \dots, x_k)$ when we mean it to be in the domain of the map L and we shall write a point in $\mathbb{R}^{n_1} \oplus \cdots \oplus \mathbb{R}^{n_k}$ as $(v_1, \dots, v_k)$ when we mean it to be an argument of the derivative. The argument of the rth derivative is an element of $(\mathbb{R}^{n_1} \oplus \cdots \oplus \mathbb{R}^{n_k})^r$ and will be written as

$$((v_{11}, \dots, v_{1k}), \dots, (v_{r1}, \dots, v_{rk})).$$

For $r \in \{1, \dots, k\}$ define

$$D_{r,k} = \{\{j_1, \dots, j_r\} \mid \ j_1, \dots, j_r \in \{1, \dots, k\} \text{ distinct}\}.$$

For $\{j_1, \dots, j_r\} \in D_{r,k}$ let us denote by $\{j'_1, \dots, j'_{k-r}\}$ the complement of $\{j_1, \dots, j_r\}$ in $\{1, \dots, k\}$. Now, for $\{j_1, \dots, j_r\} \in D_{r,k}$ define

$$\lambda_{j_1,\dots,j_r} \in \mathrm{L}((\mathbb{R}^{n_{j'_1}} \oplus \cdots \oplus \mathbb{R}^{n_{j'_{k-r}}}) \oplus (\mathbb{R}^{n_{j_1}} \oplus \cdots \oplus \mathbb{R}^{n_{j_r}}); \mathbb{R}^{n_1} \oplus \cdots \oplus \mathbb{R}^{n_k})$$

by asking that

$$\lambda_{j_1,\dots,j_k}((x_1, \dots, x_{k-r}), (v_1, \dots, v_r))$$

be obtained by placing x_l in slot j'_l for $l \in \{1, \dots, k-r\}$ and by placing v_l in slot j_l for $l \in \{1, \dots, r\}$.

With the above notation we have the following description of the derivative of a multilinear map.

4.4.8 Theorem (Derivatives of multilinear maps) *If* $L \in \mathrm{L}(\mathbb{R}^{n_1} \oplus \cdots \oplus \mathbb{R}^{n_k}; \mathbb{R}^m)$ *then* L *is infinitely differentiable. Moreover, for* $r \in \{1, \dots, k\}$ *we have*

$$\mathbf{D}^r L(\mathbf{x}_1, \dots, \mathbf{x}_k) \cdot ((\mathbf{v}_{11}, \dots, \mathbf{v}_{1k}), \dots, (\mathbf{v}_{r1}, \dots, \mathbf{v}_{rk})) = \sum_{\sigma \in \mathfrak{S}_r} \sum_{\{j_1,\dots,j_r\} \in D_{r,k}} L \circ \lambda_{j_1,\dots,j_r}((\mathbf{x}_{j'_1}, \dots, \mathbf{x}_{j'_{k-r}}), (\mathbf{v}_{\sigma(1)j_1}, \dots, \mathbf{v}_{\sigma(r)j_r}))$$

and for $r > k$ *we have* $\mathbf{D}^r L(\mathbf{x}_1, \dots, \mathbf{x}_k) = \mathbf{0}$.

Proof We prove the result by induction on r. For $r = 1$ the theorem asserts that

$$DL(x_{01}, \dots, x_{0k}) \cdot (v_1, \dots, v_k) = L(v_1, x_{02}, \dots, x_{0k}) + L(x_{01}, v_2, \dots, x_{0k}) + \cdots + L(x_{01}, x_{02}, \dots, v_k).$$

To verify this we must show that

$$\lim_{\substack{(x_1,\dots,x_k)\\ \to(x_{01},\dots,x_{0k})}} \Big\| \mathsf{L}(x_1,\dots,x_k) - \mathsf{L}(x_{01},\dots,x_{0k}) - \mathsf{L}(x_1-x_{01},\dots,x_{0k}) - \mathsf{L}(x_{01},\dots,x_k-x_{0k}) \Big\|_{\mathbb{R}^m} / \|(x_1-x_{01},\dots,x_k-x_{0k})\|_{\mathbb{R}^{n_1+\cdots+n_k}} = 0. \quad (4.18)$$

We do this by induction on k. For $k = 1$ we have

$$\mathsf{L}(x_1) - \mathsf{L}(x_{01}) - \mathsf{L}(x_1 - x_{01}) = \mathbf{0},$$

and so (4.18) holds trivially. Now suppose that (4.18) holds for $k = s \geq 2$ and let $\mathsf{L} \in \mathrm{L}(\mathbb{R}^{n_1},\dots,\mathbb{R}^{n_{s+1}};\mathbb{R}^m)$. We first note that the numerator in the limit in (4.18) can be written as

$$\begin{aligned}\mathsf{L}(x_1,\dots,x_s,x_{0(s+1)}) - \mathsf{L}(x_{01},\dots,x_{0s},x_{0(s+1)}) + \mathsf{L}(x_1,\dots,x_s,x_s-x_{0(s+1)})\\ - \mathsf{L}(x_1-x_{01},\dots,x_{0s},x_{0(s+1)}) - \cdots - \mathsf{L}(x_{01},\dots,x_s-x_{0s},x_{0(s+1)})\\ - \mathsf{L}(x_{01},\dots,x_{0s},x_{s+1}-x_{0(s+1)}).\end{aligned}$$

By the induction hypothesis we have

$$\lim_{\substack{(x_1,\dots,x_s)\\ \to(x_{01},\dots,x_{0s})}} \Big\| \mathsf{L}(x_1,\dots,x_s,x_{0(s+1)}) - \mathsf{L}(x_{01},\dots,x_{0s},x_{0(s+1)}) - \mathsf{L}(x_1-x_{01},\dots,x_{0s},x_{0(s+1)}) - \mathsf{L}(x_{01},\dots,x_s-x_{0s},x_{0(s+1)}) \Big\|_{\mathbb{R}^m} / \|(x_1-x_{01},\dots,x_s-x_{0s})\|_{\mathbb{R}^{n_1+\cdots+n_s}} = 0.$$

Since

$$\|(x_1-x_{01},\dots,x_s-x_{0s})\|_{\mathbb{R}^{n_1+\cdots+n_s}} \leq \|(x_1-x_{01},\dots,x_s-x_{0s},x_{s+1}-x_{0(s+1)})\|_{\mathbb{R}^{n_1+\cdots+n_s+n_{s+1}}}$$

this implies that

$$\lim_{\substack{(x_1,\dots,x_s,x_{s+1})\\ \to(x_{01},\dots,x_{0s},x_{0(s+1)})}} \Big\| \mathsf{L}(x_1,\dots,x_s,x_{0(s+1)}) - \mathsf{L}(x_{01},\dots,x_{0s},x_{0(s+1)}) - \mathsf{L}(x_1-x_{01},\dots,x_{0s},x_{0(s+1)}) - \mathsf{L}(x_{01},\dots,x_s-x_{0s},x_{0(s+1)}) \Big\|_{\mathbb{R}^m} / \|(x_1-x_{01},\dots,x_s-x_{0s},x_{s+1}-x_{0(s+1)})\|_{\mathbb{R}^{n_1+\cdots+n_s+n_{s+1}}} = 0. \quad (4.19)$$

We also have

$$\lim_{\substack{(x_1,\dots,x_s,x_{s+1})\\ \to(x_{01},\dots,x_{0s},x_{0(s+1)})}} \Big\| \mathsf{L}\Big(x_1,\dots,x_s,\tfrac{x_{s+1}-x_{0(s+1)}}{\|x_{s+1}-x_{0(s+1)}\|_{\mathbb{R}^{n_{s+1}}}}\Big) - \mathsf{L}\Big(x_{01},\dots,x_{0s},\tfrac{x_{s+1}-x_{0(s+1)}}{\|x_{s+1}-x_{0(s+1)}\|_{\mathbb{R}^{n_{s+1}}}}\Big) \Big\|_{\mathbb{R}^m} = 0$$

by continuity of L. Since

$$\|x_{s+1}-x_{0(s+1)}\|_{\mathbb{R}^{n_{s+1}}} \leq \|(x_1-x_{01},\dots,x_s-x_{0s},x_{s+1}-x_{0(s+1)})\|_{\mathbb{R}^{n_1+\cdots+n_s+n_{s+1}}}$$

this gives

$$\lim_{\substack{(x_1,\dots,x_s,x_{s+1})\\ \to(x_{01},\dots,x_{0s},x_{0(s+1)})}} \Big\| \mathsf{L}(x_1,\dots,x_s,x_{s+1}-x_{0(s+1)}) - \mathsf{L}(x_{01},\dots,x_{0s},x_{s+1}-x_{0(s+1)}) \Big\|_{\mathbb{R}^m} / \|(x_1-x_{01},\dots,x_s-x_{0s},x_{s+1}-x_{0(s+1)})\|_{\mathbb{R}^{n_1+\cdots+n_s+n_{s+1}}} = 0. \quad (4.20)$$

Combining (4.19) and (4.20) gives (4.18) for the case when $k = s + 1$ and so gives the conclusion of the theorem in the case when $r = 1$.

Now suppose that the theorem holds for $r \in \{1, \dots, s\}$ with $s < k$ and let $\mathsf{L} \in \mathrm{L}(\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_k}; \mathbb{R}^m)$. Let us fix $\{j_1, \dots, j_s\} \in D_{s,k}$ and denote the complement of $\{j_1, \dots, j_s\}$ in $\{1, \dots, k\}$ by $\{j'_1, \dots, j'_{k-s}\}$, just as in our definitions before the theorem statement. Let us also fix $v_{j_l} \in \mathbb{R}^{n_{j_l}}$ for $l \in \{1, \dots, s\}$. Then define

$$\begin{aligned}\mathsf{P}_{v_{j_1},\dots,v_{j_s}} : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} &\to (\mathbb{R}^{n_{j'_1}} \times \cdots \times \mathbb{R}^{n'_{j_{k-s}}}) \times (\mathbb{R}^{n_{j_1}} \times \cdots \times \mathbb{R}^{n_{j_s}}) \\ (x_1, \dots, x_k) &\mapsto ((x_{j'_1}, \dots, x_{j'_{k-s}}), (v_{j_1}, \dots, v_{j_s})).\end{aligned}$$

Now define $g_{v_{j_1},\dots,v_{j_s}} : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}^m$ by $g_{v_{j_1},\dots,v_{j_s}} = \mathsf{L} \circ \lambda_{j_1,\dots,j_s} \circ \mathsf{P}_{v_{j_1},\dots,v_{j_s}}$ and note that

$$g_{v_{j_1},\dots,v_{j_s}}(x_1, \dots, x_k) = \mathsf{L} \circ \lambda_{j_1,\dots,j_s}((x_{j'_1}, \dots, x_{j'_{k-s}}), (v_{j_1}, \dots, v_{j_s})).$$

By the Chain Rule, Theorem 4.4.49 below, we have

$$\begin{aligned}&Dg_{v_{j_1},\dots,v_{j_s}}(x_1, \dots, x_k) \cdot (u_1, \dots, u_k) \\ &\qquad = D(\mathsf{L} \circ \lambda_{j_1,\dots,j_r})(P(x_1, \dots, x_k)) \circ D\mathsf{P}_{v_{j_1},\dots,v_{j_s}}(x_1, \dots, x_k) \cdot (u_1, \dots, u_k).\end{aligned}$$

Note that since $\mathsf{P}_{v_{j_1},\dots,v_{j_s}}$ is essentially a linear map (precisely, it is affine, meaning linear plus constant) we have

$$D\mathsf{P}_{v_{j_1},\dots,v_{j_s}}(x_1, \dots, x_k) \cdot (u_1, \dots, u_k) = ((u_{j'_1}, \dots, u_{j'_{k-s}}), (0, \dots, 0)).$$

Note that since $\mathsf{L} \circ \lambda_{j_1,\dots,j_s} \in \mathrm{L}(\mathbb{R}^{n_{j'_1}}, \dots, \mathbb{R}^{n_{j'_{k-s}}}, \mathbb{R}^{n_{j_1}}, \dots, \mathbb{R}^{n_{j_s}}; \mathbb{R}^m)$ (as is readily verified), by the induction hypothesis,

$$\begin{aligned}&D(\mathsf{L} \circ \lambda_{j_1,\dots,j_s})(x_{j'_1}, \dots, x_{j'_{k-s}}, x_{j_1}, \dots, x_{j_s}) \cdot ((u_{j'_1}, \dots, u_{j'_{k-s}}), (u_{j_1}, \dots, u_{j_s})) \\ &\qquad = \mathsf{L} \circ \lambda_{j_1,\dots,j_s}((u_{j'_1}, \dots, x_{j'_{k-s}}), (x_{j_1}, \dots, x_{j_s})) + \dots \\ &\qquad\qquad + \mathsf{L} \circ \lambda_{j_1,\dots,j_s}((x_{j'_1}, \dots, x_{j'_{k-s}}), (x_{j_1}, \dots, u_{j_s})).\end{aligned}$$

Therefore,

$$\begin{aligned}&Dg_{v_{j_1},\dots,v_{j_s}}(x_1, \dots, x_k) \cdot (u_1, \dots, u_k) \\ &\qquad = \mathsf{L} \circ \lambda_{j_1,\dots,j_s}((u_{j'_1}, \dots, x_{j'_{k-s}}), (v_{j_1}, \dots, v_{j_s})) + \dots \\ &\qquad\qquad + \mathsf{L} \circ \lambda_{j_1,\dots,j_s}((x_{j'_1}, \dots, u_{j'_{k-s}}), (v_{j_1}, \dots, v_{j_s})).\end{aligned}$$

Thus, for $v_j \in \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$, we have

$$\begin{aligned}&Dg_{v_{j_1},\dots,v_{j_s}}(x_1, \dots, x_k) \cdot (v_1, \dots, v_k) \\ &\qquad = \sum_{j_{s+1} \notin \{j_1,\dots,j_s\}} \mathsf{L} \circ \lambda_{j_1,\dots,j_s,j_{s+1}}((x_{j'_1}, \dots, x_{j'_{k-(s+1)}}), (v_{j_1}, \dots, v_{j_{s+1}})).\end{aligned}$$

Thus, using this relation along with Proposition 4.4.7, linearity of the derivative (see Proposition 4.4.47), the Chain Rule (see Theorem 4.4.49), and the induction hypothesis,

we compute

$$\begin{aligned}
D^{s+1}f(x_1,\dots,x_k)\cdot((v_{11},\dots,v_{1k}),(v_{21},\dots,v_{2k}),\dots,\dots,(v_{(s+1)1},\dots,v_{(s+1)k}))\\
&= \sum_{\sigma\in\mathfrak{S}_s}\sum_{\{j_2,\dots,j_{s+1}\}\in D_{s,k}} Dg_{v_{\sigma(2)j_2},\dots,v_{\sigma(s+1)j_{s+1}}}(x_1,\dots,x_k)\cdot(v_{11},\dots,v_{1k})\\
&= \sum_{\sigma\in\mathfrak{S}_s}\sum_{\{j_2,\dots,j_{s+1}\}\in D_{s,k}}\sum_{j_1\notin\{j_2,\dots,j_{s+1}\}} L\circ\lambda_{j_1,\dots,j_s,j_{s+1}}((x_{j'_1},\dots,x_{j'_{k-(s+1)}}),\\
&\quad (v_{j_1},v_{\sigma(2)j_2},\dots,v_{\sigma(s+1)j_{s+1}}))\\
&= \sum_{\sigma\in\mathfrak{S}_{s+1}}\sum_{\{j_1,\dots,j_{s+1}\}\in D_{s+1,k}} L\circ\lambda_{\{j_1,\dots,j_{s+1}\}}((x_{j'_1},\dots,x_{j'_{k-(s+1)}}),\\
&\quad (v_{\sigma(1)j_1},\dots,v_{\sigma(s+1)j_{s+1}})),
\end{aligned}$$

where, in the second and third line, we define $\sigma\in\mathfrak{S}_s$ to be a bijection of $\{1,\dots,s+1\}$ by permutation of the last s elements.

The preceding argument gives the result when $r\in\{1,\dots,k\}$. For $r>k$ we argue as follows. We first note that

$$D^kL((v_{11},\dots,v_{1k}),\dots,(v_{k1},\dots,v_{kk}))=\sum_{\sigma\in\mathfrak{S}_k}L(v_{\sigma(1)1},\dots,v_{\sigma(k)k}). \tag{4.21}$$

By Proposition 4.4.7 it follows that $D^rL=\mathbf{0}$ for $r>k$. ∎

The proof of the preceding theorem, and indeed the statement, is marred by notational baggage needed to state the result in full generality. However, the result is actually simple to use, and to illustrate this we explicitly write the result when $k=3$. In this case we have the following formulae:

$$\begin{aligned}
DL(x_1,x_2,x_3)\cdot(v_{11},v_{12},v_{13}) &= L(v_1,x_2,x_3)+L(x_1,v_2,x_3)+L(x_1,x_2,v_3),\\
D^2L(x_1,x_2,x_3)\cdot((v_{11},v_{12},v_{13}),(v_{21},v_{22},v_{23}))&\\
&= L(v_{21},v_{12},x_3)+L(v_{11},v_{22},x_3)+L(v_{21},x_2,v_{13})\\
&\quad +L(v_{11},x_2,v_{23})+L(x_1,v_{22},v_{13})+L(x_1,v_{12},v_{23}),\\
D^3L(x_1,x_2,x_3)\cdot((v_{11},v_{12},v_{13}),(v_{21},v_{22},v_{23}),(v_{31},v_{32},v_{33}))&\\
&= L(v_{11},v_{22},v_{33})+L(v_{11},v_{32},v_{23})+L(v_{21},v_{12},v_{33})\\
&\quad +L(v_{21},v_{32},v_{13})+L(v_{31},v_{12},v_{23})+L(v_{31},v_{22},v_{13}).
\end{aligned}$$

For readers who understand the product rule of differentiation well, cf. Theorem 4.4.48, the preceding formulae are easy to derive. For readers for whom the formulae look mysterious, it is well to develop some facility in using them and like formulae since they come up often.

A case of particular importance occurs when $n_1=\cdots=n_k=n$ and when all arguments of L are the same.

4.4.9 Corollary (Derivatives of multilinear maps II) *Let $L\in L^k(\mathbb{R}^n;\mathbb{R}^m)$ and define $\mathbf{f}_L:\mathbb{R}^n\to\mathbb{R}^m$ by $\mathbf{f}_L(\mathbf{x})=L(\mathbf{x},\dots,\mathbf{x})$. Then $\mathbf{f}_L$ is infinitely differentiable and, moreover, for $r\in\{1,\dots,k\}$ we have*

$$\mathbf{D}^r\mathbf{f}_L(\mathbf{x})\cdot(\mathbf{v}_1,\dots,\mathbf{v}_r)=\sum_{\sigma\in\mathfrak{S}_r}\sum_{\{j_1,\dots,j_r\}\in D_{r,k}} L\circ\lambda_{j_1,\dots,j_r}((\mathbf{x},\dots,\mathbf{x}),(\mathbf{v}_{\sigma(1)},\dots,\mathbf{v}_{\sigma(r)})).$$

Proof Define $\mathsf{D} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n \oplus \cdots \oplus \mathbb{R}^n)$ by $\mathsf{D}(x) = (x, \ldots, x)$. Then $f_{\mathsf{L}} = \mathsf{L} \circ \mathsf{D}$. Let us also define, for any $r \in \mathbb{Z}_{>0}$, $\mathsf{D}^*_r \colon \mathrm{L}^r((\mathbb{R}^n)^k; \mathbb{R}^m) \to \mathrm{L}^r(\mathbb{R}^n; \mathbb{R}^m)$ by

$$\mathsf{D}^*_r(\mathsf{A}) \cdot (v_1, \ldots, v_r) = \mathsf{A}(\mathsf{D}(v_1), \ldots, \mathsf{D}(v_r)).$$

Let us record the derivative of f_{L} in this case.

1 Lemma $\mathbf{D}^{\mathrm{r}} \mathbf{f}_{\mathsf{L}} = \mathsf{D}^*_{\mathrm{r}} \circ \mathbf{D}^{\mathrm{r}} \mathsf{L} \circ \mathsf{D}$.

Proof We prove the lemma by induction on r. For $r = 1$ we have

$$\boldsymbol{D} f_{\mathsf{L}}(x) \cdot v_1 = \boldsymbol{D}\mathsf{L}(\mathsf{D}(x)) \circ \mathsf{D}(v_1),$$

using the Chain Rule below and the fact that the derivative of D is D since D is linear. This gives the result when $r = 1$, using the definition of D^*_1. So suppose the result holds for $r \in \{1, \ldots, s\}$. Thus

$$\boldsymbol{D}^s f_{\mathsf{L}}(x) \cdot (v_1, \ldots, v_s) = \boldsymbol{D}^s \mathsf{L}(\mathsf{D}(x)) \cdot (\mathsf{D}(v_1), \ldots, \mathsf{D}(v_s)).$$

Using Proposition 4.4.7 and the Chain Rule we then have

$$\begin{aligned}(\boldsymbol{D}^{s+1} f_{\mathsf{L}}(x) \cdot (v_0)) \cdot (v_1, \ldots, v_s) &= (\boldsymbol{D}(\boldsymbol{D}^s \mathsf{L})(\mathsf{D}(x)) \circ \mathsf{D}(v_0)) \cdot (\mathsf{D}(v_1), \ldots, \mathsf{D}(v_s)) \\ &= \boldsymbol{D}^{s+1} \mathsf{L}(\mathsf{D}(x)) \cdot (\mathsf{D}(v_0), \mathsf{D}(v_1), \ldots, \mathsf{D}(v_s)),\end{aligned}$$

where we use the isomorphism of Proposition **??**. This gives the lemma. ▼

The result now follows directly from Theorem 4.4.8. ■

The following trivial corollary is also worth recording separately.

4.4.10 Corollary (The derivative of a linear map) *If* $\mathsf{L} \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^m)$ *then* $\mathbf{D}\mathsf{L}(\mathbf{x}) = \mathsf{L}$ *for each* $\mathbf{x} \in \mathbb{R}^n$.

4.4.3 The directional derivative

In this section we describe another way of differentiating a function. As we shall see, this type of derivative is weaker than the derivative in the preceding section. However, it is perhaps a more intuitive notion of derivative, so we discuss it here to assist in understanding how one might interpret the derivative.

4.4.11 Definition (Directional derivative) Let $U \subseteq \mathbb{R}^n$ be open, let $f \colon U \to \mathbb{R}^m$, let $x_0 \in U$, and let $v \in \mathbb{R}^n$. The map f is ***differentiable in the direction*** $\mathbf{v}$ at x_0 if the map $s \mapsto f(x_0 + sv)$ is differentiable at $s = 0$. If f has a directional derivative at x_0 in the direction v then we denote by

$$\boldsymbol{D} f(x_0; v) = \frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} f(x_0 + sv)$$

the ***directional derivative***. If, for all $v \in \mathbb{R}^n$, f is differentiable in the direction v at x_0 then f is ***Gâteaux differentiable*** at x_0. •

We advise the reader to carefully note the distinction in the notation between the derivative at x_0 evaluated at v and the directional derivative at x_0 in the direction v. The former is denoted by $Df(x_0) \cdot v$ while the latter is denoted by $Df(x_0; v)$.

It is probably the case that the directional derivative is a more easily understood concept that the derivative. The idea of the directional derivative of f at x_0 in the direction of v is that one measures what is happening to the values of f as one steps away from x_0 in a specific direction. One might imagine that the existence of the derivative is equivalent to the existence of all partial derivatives. This, however, is false! Let us explore, therefore, the relationship between the derivative and the directional derivative.

4.4.12 Proposition (Differentiable maps are directionally differentiable) *Let $\mathrm{U} \subseteq \mathbb{R}^n$ be open and let $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^m$ be differentiable at $\mathbf{x}_0$. Then, for any $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{f}$ has a directional derivative at $\mathbf{x}_0$ in the direction of $\mathbf{v}$ and, moreover,*

$$\mathbf{Df}(\mathbf{x}_0; \mathbf{v}) = \mathbf{Df}(\mathbf{x}_0) \cdot \mathbf{v}.$$

Proof Let $\epsilon \in \mathbb{R}_{>0}$ be such that $x_0 + sv \in U$ for each $s \in (-\epsilon, \epsilon)$; this is possible since U is open. Then let $g\colon (-\epsilon, \epsilon) \to U$ be given by $g(s) = x_0 + sv$. The existence of the directional derivative of f at x_0 in the direction of v is then exactly the differentiability of $s \mapsto f \circ g(s)$ at $s = 0$. However, by the Chain Rule (Theorem 4.4.49), this function is indeed differentiable at $s = 0$ and, moreover,

$$Df(x_0; v) = Df(x_0) \circ Dg(0).$$

Note that $Dg(0) \in \mathrm{L}(\mathbb{R}; \mathbb{R}^n)$ is simply the linear map $\alpha \mapsto \alpha v$ and so

$$Df(x_0) \circ Dg(0) \in \mathrm{L}(\mathbb{R}; \mathbb{R}^m)$$

is the linear map

$$\alpha \mapsto \alpha(Df(x_0) \cdot v).$$

Upon making the natural identification of $\mathbb{R}^m$ with $\mathrm{L}(\mathbb{R}; \mathbb{R}^m)$ (i.e., the identification which assigns to $u \in \mathbb{R}^m$ the linear map $\alpha \mapsto \alpha u$) we see that we have the equality of derivatives asserted in the proposition. ■

In some sense the preceding result is reassuring since it tells us that the directional derivative interpretation can be made for the derivative when the latter exists. The following example shows, however, that the converse of the preceding result does not hold in general. Thus it is not the case that differentiability in all directions is equivalent to differentiability.

4.4.13 Example (Discontinuous function possessing all directional derivatives) We consider the function of Example 4.3.27:

$$f(x_1, x_2) = \begin{cases} \frac{x_1^2 x_2}{x_1^4 + x_2^2}, & (x_1, x_2) \neq (0, 0), \\ 0, & (x_1, x_2) = (0, 0). \end{cases}$$

In Example 4.3.27 we show that f is discontinuous at $(0, 0)$.

We further claim that f possesses all directional derivatives at $(0,0)$. Indeed, let $(u_1, u_2) \in \mathbb{R}^2$ and consider the line

$$s \mapsto (0,0) + s(u_1, u_2), \qquad s \in \mathbb{R},$$

through $(0,0)$ in the direction of (u_1, u_2). Along this line we have

$$f((0,0) + s(u_1, u_2)) = \frac{su_1^2u_2}{s^2u_1^4 + u_2^2}.$$

A direct computation gives

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} f((0,0) + s(u_1, u_2)) = \begin{cases} \frac{u_1^2}{u_2}, & u_2 \neq 0, \\ 0, & u_2 = 0, \end{cases}$$

which shows that f possesses all directional derivatives at $(0,0)$. •

Having settled the relationship between the derivative and the directional derivative, let us give some of the properties of the directional derivative.

4.4.14 Proposition (Properties of the directional derivative) *Let* $\mathrm{U} \subseteq \mathbb{R}^n$, *let* $\mathbf{f}, \mathbf{g} \colon \mathrm{U} \to \mathbb{R}^m$, *let* $\mathbf{x}_0 \in \mathrm{U}$, *let* $\mathbf{v} \in \mathbb{R}^n$, *and let* $\mathrm{a} \in \mathbb{R}$. *If* $\mathbf{f}$ *and* $\mathbf{g}$ *are differentiable in the direction* $\mathbf{v}$ *at* $\mathbf{x}_0$ *then the following statements hold:*

(i) $\mathbf{f}$ *is differentiable in the direction* $\alpha\mathbf{v}$ *at* $\mathbf{x}_0$ *for each* $\alpha \in \mathbb{R}$ *and the map* $\alpha \mapsto \mathbf{D}\mathbf{f}(\mathbf{x}_0; \alpha\mathbf{v})$ *is linear;*

(ii) $\mathbf{f} + \mathbf{g}$ *is differentiable in the direction* $\mathbf{v}$ *at* $\mathbf{x}_0$ *and*

$$\mathbf{D}(\mathbf{f} + \mathbf{g})(\mathbf{x}_0; \mathbf{v}) = \mathbf{D}\mathbf{f}(\mathbf{x}_0; \mathbf{v}) + \mathbf{D}\mathbf{g}(\mathbf{x}_0; \mathbf{v});$$

(iii) $\mathrm{a}\mathbf{f}$ *is differentiable in the direction* $\mathbf{v}$ *at* $\mathbf{x}_0$ *and*

$$\mathbf{D}(\mathrm{a}\mathbf{f})(\mathbf{x}_0; \mathbf{v}) = \mathrm{a}(\mathbf{D}\mathbf{f}(\mathbf{x}_0; \mathbf{v}).$$

Moreover, if $\mathrm{m} = 1$ *and we denote* $\mathbf{f}$ *and* $\mathbf{g}$ *by* f *and* g, *respectively, then under the same hypotheses as above we additionally have the following statements:*

(iv) fg *is differentiable in the direction* $\mathbf{v}$ *at* $\mathbf{x}_0$ *and*

$$\mathbf{D}(\mathrm{fg})(\mathbf{x}_0; \mathbf{v}) = \mathrm{g}(\mathbf{x}_0)\mathbf{D}\mathrm{f}(\mathbf{x}_0; \mathbf{v}) + \mathrm{f}(\mathbf{x}_0)\mathbf{D}\mathrm{g}(\mathbf{x}_0; \mathbf{v});$$

(v) if $\mathrm{g}(\mathbf{x}_0) \neq 0$ *then* $\frac{\mathrm{f}}{\mathrm{g}}$ *is differentiable in the direction* $\mathbf{v}$ *at* $\mathbf{x}_0$ *and*

$$\mathbf{D}(\tfrac{\mathrm{f}}{\mathrm{g}})(\mathbf{x}_0; \mathbf{v}) = \frac{\mathrm{g}(\mathbf{x}_0)\mathbf{D}\mathrm{f}(\mathbf{x}_0; \mathbf{v}) - \mathrm{f}(\mathbf{x}_0)\mathbf{D}\mathrm{g}(\mathbf{x}_0; \mathbf{v})}{\mathrm{g}(\mathbf{x}_0)^2}.$$

Proof (i) For $\alpha = 0$ we clearly have $Df(x_0; \alpha v) = \mathbf{0}$. So suppose that $\alpha \neq 0$. Then, letting $\sigma = \alpha s$ and using the Chain Rule, Theorem 4.4.49,

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} f(x_0 + s\alpha v) = \frac{\mathrm{d}\sigma}{\mathrm{d}s}\,\frac{\mathrm{d}}{\mathrm{d}\sigma}\Big|_{\sigma=0} f(x_0 + \sigma v) = \alpha Df(x_0; v),$$

giving this part of the result.

(ii) We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0}(f+g)(x_0+sv) &= \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0}f(x_0+sv)+\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0}g(x_0+sv)\\ &= Df(x_0;v)+Dg(x_0;v),\end{aligned}$$

as desired, where we have used Proposition 3.2.10.

(iii) This part of the result also follows from Proposition 3.2.10.

(iv) We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0}(fg)(x_0+sv) &= \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0}f(x_0+sv)g(x_0+sv)\\ &= Df(x_0;v)+Dg(x_0;v),\end{aligned}$$

where we have used Proposition 3.2.10.

(v) This also follows from Proposition 3.2.10. ∎

It is also possible to define higher-order directional derivatives. We let $U \subseteq \mathbb{R}^n$ be open, let $f\colon U \to \mathbb{R}^m$, let $x \in U$, and let $v_1, v_2 \in \mathbb{R}^n$. We suppose that the directional derivative $Df(x_0 + sv_2; v_1)$ exists for each s sufficiently close to zero for some $x_0 \in U$. This allows the possibility of defining the directional derivative of the directional derivative:

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} Df(x_0 + sv_2; v_1).$$

This procedure can be continued inductively.

4.4.15 Definition (Higher-order directional derivatives) Let $U \subseteq \mathbb{R}^n$ be open, let $f\colon U \to \mathbb{R}^m$, let $x_0 \in U$ and $v_0, v_1, \dots, v_{r-1} \in \mathbb{R}^n$, and suppose that f is differentiable in the directions $v_1, \dots, v_{r-1}$ at $x_0 + sv_0$ for $s \in (-\epsilon, \epsilon)$ with $\epsilon \in \mathbb{R}_{>0}$, with $D^{r-1}f(x_0 + sx_0; v_1, \dots, v_{r-1})$ be the directional derivative. The vector

$$D^r f(x_0; v_0, v_1, \dots, v_{r-1}) \triangleq \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} D^{r-1}f(x_0 + sx_0; v_1, \dots, v_{r-1})$$

in $\mathbb{R}^m$ is the ***directional derivative of*** $\mathbf{f}$ at x_0 in the directions $v_0, v_1, \dots, v_{r-1}$, when the derivative exists. •

We now have the following generalisation of Proposition 4.4.12.

4.4.16 Proposition (Higher-order derivative and directional derivatives) *Let* $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}}$ *and let* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^{\mathrm{m}}$ *be* r *times differentiable at* $\mathbf{x}_0 \in \mathrm{U}$. *Then, for any* $\mathbf{v}_1, \dots, \mathbf{v}_{\mathrm{r}} \in \mathbb{R}^{\mathrm{n}}$, *the directional derivative of* $\mathbf{f}$ *at* $\mathbf{x}_0$ *and in the directions* $\mathbf{v}_1, \dots, \mathbf{v}_{\mathrm{r}}$ *exists and, moreover,*

$$\mathbf{D}^{\mathrm{r}}\mathbf{f}(\mathbf{x}_0; \mathbf{v}_1, \dots, \mathbf{v}_{\mathrm{r}}) = \mathbf{D}^{\mathrm{r}}\mathbf{f}(\mathbf{x}_0) \cdot (\mathbf{v}_1, \dots, \mathbf{v}_{\mathrm{r}}).$$

Proof We prove the result by induction on r, the case of $r = 1$ being Proposition 4.4.12. Suppose the result holds for $r = s$ and let f be $s+1$ times differentiable at x_0. By Proposition 4.4.35 and by the induction hypothesis the directional derivatives $D^s f(x; v_1, \dots, v_s)$ exist for x in a neighbourhood of x_0 and for all $v_1, \dots, v_s$. Since

$$D^s f(x; v_1, \dots, v_s) = D^s f(x) \cdot (v_1, \dots, v_s)$$

by the induction hypothesis, it follows from Proposition 4.4.7 that

$$x \mapsto D^s f(x; v_1, \dots, v_s)$$

is differentiable at x_0. By Proposition 4.4.12 it then holds that this map has a directional derivative at x_0 in the direction $v_0 \in \mathbb{R}^n$. Also by Proposition 4.4.12 it follows that

$$\begin{aligned} D^{s+1} f(x_0; v_0, v_1, \dots, v_s) &= (D(D^s f)(x) \cdot v_0) \cdot (v_1, \dots, v_s) \\ &= D^{s+1} f(x_0) \cdot (v_0, v_1, \dots, v_s), \end{aligned}$$

giving the result. ■

4.4.4 Derivatives and products, partial derivatives

The notion of a partial derivatives is one that is easy to understand in practice. That is to say, if one can compute derivatives, the matter of computing partial derivatives poses no problems in principle. However, this simplicity of computation can serve to obscure the rather important contribution of the *concept* of partial derivative to the theory of the derivative, and particularly higher-order derivatives. Therefore, in this section we present the partial derivative in a slightly general setting in order to give the partial derivative a little context. The appropriate general setting is that of functions defined on and taking values in products.

We first consider the case when we have a map $f \colon A \to \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_k}$ from a subset $A \subseteq \mathbb{R}^n$ into a product of Euclidean spaces. In this case, following Example 1.3.3–9, we write $f = f_1 \times \cdots \times f_k$ for maps $f_j \colon A \to \mathbb{R}^{m_j}$, $j \in \{1, \dots, k\}$; that is,

$$f(x) = (f_1(x), \dots, f_k(x)), \qquad x \in A.$$

We note that if f is differentiable at $x_0 \in A$ then $Df(x_0) \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^{m_1} \oplus \cdots \oplus \mathbb{R}^{m_k})$. As in Exercise **??** we note that a linear map L from $\mathbb{R}^n$ into $\mathbb{R}^{m_1} \oplus \cdots \oplus \mathbb{R}^{m_k}$ can be written as

$$\mathsf{L}(v) = \mathsf{L}_1(v) + \cdots + \mathsf{L}_k(v)$$

for linear maps $\mathsf{L}_j \colon \mathbb{R}^n \to \mathbb{R}^{m_j}$, $j \in \{1, \dots, k\}$. Let us use the notation $\mathsf{L} = \mathsf{L}_1 \oplus \cdots \oplus \mathsf{L}_k$ to represent this fact. This notation can be extended to multilinear maps as well. Thus if $\mathsf{L} \in \mathrm{L}^k(\mathbb{R}^n; \mathbb{R}^{m_1} \oplus \cdots \oplus \mathbb{R}^{m_k})$ then we can write

$$\mathsf{L}(v_1, \dots, v_k) = \mathsf{L}_1(v_1, \dots, v_k) + \cdots + \mathsf{L}_k(v_1, \dots, v_k)$$

for $\mathsf{L}_j \in \mathrm{L}^k(\mathbb{R}^n; \mathbb{R}^{m_j})$, $j \in \{1, \dots, k\}$. We also write $\mathsf{L} = \mathsf{L}_1 \oplus \cdots \oplus \mathsf{L}_k$ in this case.

With all this notation we have the following result.

4.4.17 Proposition (Derivatives of maps taking values in products) *Let* $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}}$ *be open and let* $\mathbf{f}\colon \mathrm{A} \to \mathbb{R}^{\mathrm{m}_1} \times \cdots \times \mathbb{R}^{\mathrm{m}_k}$ *be a map which we write as* $\mathbf{f} = \mathbf{f}_1 \times \cdots \times \mathbf{f}_{\mathrm{k}}$. *Then* $\mathbf{f}$ *is* r *times differentiable at* $\mathbf{x}_0 \in \mathrm{U}$ *if and only if* $\mathbf{f}_{\mathrm{j}}$ *is* r *times differentiable at* $\mathbf{x}_0$ *for each* $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$. *Moreover, if* $\mathbf{f}$ *is* r *times differentiable at* $\mathbf{x}_0$ *then*

$$\mathbf{D}^{\mathrm{r}}\mathbf{f}(\mathbf{x}_0) = \mathbf{D}^{\mathrm{r}}\mathbf{f}_1(\mathbf{x}_0) \oplus \cdots \oplus \mathbf{D}^{\mathrm{r}}\mathbf{f}_{\mathrm{k}}(\mathbf{x}_0).$$

Proof Via an elementary inductive argument it suffices to prove the result in the case of $r = 1$, and so we restrict ourselves to this case.

Suppose that f is differentiable at x_0 with derivative written as $Df(x_0) = \mathsf{L}_1 \oplus \cdots \oplus \mathsf{L}_k$. Then, using the triangle inequality,

$$\frac{\|f_j(x) - f_j(x_0) - \mathsf{L}_j(x - x_0)\|_{\mathbb{R}^{m_j}}}{\|x - x_0\|_{\mathbb{R}^n}} \leq \frac{\|f(x) - f(x_0) - Df(x_0) \cdot (x - x_0)\|_{\mathbb{R}^{m_1 + \cdots + m_k}}}{\|x - x_0\|_{\mathbb{R}^n}}, \qquad j \in \{1, \dots, k\}.$$

Therefore,

$$\lim_{x \to x_0} \frac{\|f_j(x) - f_j(x_0) - \mathsf{L}_j(x - x_0)\|_{\mathbb{R}^{m_j}}}{\|x - x_0\|_{\mathbb{R}^n}} = 0, \qquad j \in \{1, \dots, k\},$$

giving differentiability of f_j at x_0 with derivative L_j for each $j \in \{1, \dots, k\}$.

For the converse, suppose that $f_1, \dots, f_k$ are differentiable at x_0 and let

$$\mathsf{L} = Df_1(x_0) \oplus \cdots \oplus Df_k(x_0).$$

Then, using the triangle inequality,

$$\frac{\|f(x) - f(x_0) - \mathsf{L}(x - x_0)\|_{\mathbb{R}^{m_1 + \cdots + m_k}}}{\|x - x_0\|_{\mathbb{R}^n}} \leq \sum_{j=1}^{k} \frac{\|f_j(x) - f_j(x_0) - Df_j(x_0) \cdot (x - x_0)\|_{\mathbb{R}^{m_j}}}{\|x - x_0\|_{\mathbb{R}^n}}.$$

Thus

$$\lim_{x \to x_0} \frac{\|f(x) - f(x_0) - \mathsf{L}(x - x_0)\|_{\mathbb{R}^{m_1 + \cdots + m_k}}}{\|x - x_0\|_{\mathbb{R}^n}} = 0,$$

giving differentiability of f at x_0. Uniqueness of the derivative now also ensures that the final assertion of the result holds. ■

Now we turn to the case of primary interest, that when the domain of the function is a product.

4.4.18 Definition (Partial derivative) Let $U \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ be open, let $f\colon U \to \mathbb{R}^m$, let $x_0 = (x_{01}, \dots, x_{0k}) \in U$, and let $j \in \{1, \dots, k\}$.

(i) The map f is ***differentiable at*** $\mathbf{x_0}$ ***with respect to the*** $\mathbf{j}$***th component*** if the map

$$U \cap (\{x_{01}\} \times \cdots \times \mathbb{R}^{n_j} \times \cdots \times \{x_{0k}\} \ni x_j) \mapsto f(x_{01}, \dots, x_j, \dots, x_{0k}) \in \mathbb{R}^m \tag{4.22}$$

is differentiable at x_{0j}.

(ii) If f is differentiable at x_0 with respect to the jth component, then the derivative at x_{j0} of the map (4.22) is denoted by $D_j f(x_0)$ and is called the ***j*th *partial derivative*** of f at x_0. •

For the reader who cannot quite imagine what is the connection with the usual notion of partial derivative, we ask that they hang on for just a moment as this will be made clear soon enough. First let us record the relationship between the derivative and the partial derivatives.

4.4.19 Theorem (Partial derivatives and derivatives) *If* $\mathrm{U} \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ *is an open set and if* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^m$ *is a map differentiable at* $\mathbf{x}_0 \in \mathrm{U}$*, then* $\mathbf{f}$ *is differentiable at* $\mathbf{x}_0$ *with respect to the* j*th component for each* $\mathrm{j} \in \{1, \ldots, \mathrm{k}\}$*. Moreover, if* $\mathbf{f}$ *is differentiable at* $\mathbf{x}_0$ *then we have the following relationships between the derivative and the partial derivatives:*

$$\mathbf{D}_\mathrm{j}\mathbf{f}(\mathbf{x}_0) \cdot \mathbf{v}_\mathrm{j} = \mathbf{Df}(\mathbf{x}_0) \cdot (\mathbf{0}, \ldots, \mathbf{v}_\mathrm{j}, \ldots, \mathbf{0})$$

$$\mathbf{Df}(\mathbf{x}_0) \cdot (\mathbf{v}_1, \ldots, \mathbf{v}_\mathrm{k}) = \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathbf{D}_\mathrm{j}\mathbf{f}(\mathbf{x}_0) \cdot \mathbf{v}_\mathrm{j}.$$

Proof Let us denote $x_0 = (x_{01}, \ldots, x_{0k})$. Differentiability of f at x_0 implies, in particular, that

$$\lim_{x_j \to x_{0j}} \Big(\| f(x_{01}, \ldots, x_j, \ldots, x_{k0}) - f(x_{01}, \ldots, x_{0j}, \ldots, x_{0k}) - Df(x_0) \cdot (\mathbf{0}, \ldots, x_j - x_{0j}, \ldots, \mathbf{0}) \|_{\mathbb{R}^m} \Big) / \Big(\|x_j - x_{0j}\|_{\mathbb{R}^{n_j}} \Big) = 0.$$

This precisely means that f is differentiable at x_0 with respect to the jth component.

Now let $v_j \in \mathbb{R}^{n_j}$ and denote $v = (\mathbf{0}, \ldots, v_j, \ldots, \mathbf{0})$. By twice applying Proposition 4.4.12 we have

$$Df(x_0) \cdot v = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} f(x_0 + sv) = f(x_{01}, \ldots, x_{0j} + sv_j, \ldots, x_{0k}) = D_j f(x_0) \cdot v_j.$$

By linearity of the derivative we then have

$$Df(x_0) \cdot (v_1, \ldots, v_k) = \sum_{j=1}^{k} Df(x_0) \cdot (\mathbf{0}, \ldots, v_j, \ldots, \mathbf{0}) = \sum_{j=1}^{k} D_j f(x_0) \cdot v_j,$$

which completes the proof. ■

If we combine Proposition 4.4.17 and Theorem 4.4.19 then we get the following general result concerning derivatives and products.

4.4.20 Corollary (Derivatives and products) *Let* $\mathrm{U} \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_r}$ *be an open set and let* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_s}$ *be a map that we write as* $\mathbf{f} = \mathbf{f}_1 \times \cdots \times \mathbf{f}_\mathrm{s}$*. If* $\mathbf{f}$ *is differentiable at* $\mathbf{x}_0 \in \mathrm{U}$ *then, for each* $\mathrm{j} \in \{1, \ldots, \mathrm{r}\}$ *and* $\mathrm{k} \in \{1, \ldots, \mathrm{s}\}$*,* $\mathbf{f}_\mathrm{k}$ *is differentiable at* $\mathbf{x}_0$ *with respect to the* j*th component. Moreover, if* $\mathbf{f}$ *is differentiable at* $\mathbf{x}_0 \in \mathrm{U}$ *then*

$$\mathbf{Df}(\mathbf{x}_0) \cdot (\mathbf{v}_1, \ldots, \mathbf{v}_\mathrm{r}) = \Big(\sum_{\mathrm{j}_1=1}^{\mathrm{r}} \mathbf{D}_{\mathrm{j}_1}\mathbf{f}_1(\mathbf{x}_0) \cdot \mathbf{v}_{\mathrm{j}_1}, \ldots, \sum_{\mathrm{j}_\mathrm{s}=1}^{\mathrm{r}} \mathbf{D}_{\mathrm{j}_\mathrm{s}}\mathbf{f}_\mathrm{s}(\mathbf{x}_0) \cdot \mathbf{v}_{\mathrm{j}_\mathrm{s}} \Big),$$

While the above presentation makes it look like the product structure is special, of course this is not the case. Every Euclidean space is a product of copies of $\mathbb{R}^1$, by definition. Therefore, the above presentation can always be applied to this natural product structure of every Euclidean space. Moreover, using this product structure sheds some light on the derivative and how to compute it. We see this as follows.

4.4.21 Definition (Jacobian matrix) Let $U \subseteq \mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ be differentiable, let $f\colon U \to \mathbb{R}^m = \mathbb{R} \times \cdots \times \mathbb{R}$ be differentiable at $x_0 \in U$, and write $f = f_1 \times \cdots \times f_m$ for $f_1, \dots, f_m\colon U \to \mathbb{R}$.

(i) The ***j*th partial derivative** of f at x_0 is $D_j f(x_0) \in \mathbb{R}^m$ (noting that $\mathrm{L}(\mathbb{R}; \mathbb{R}^m)$ is isomorphic to $\mathbb{R}^m$ by Exercise **??**).

(ii) The ***j*th partial derivative** of the kth component of f at x_0 is $D_j f_k(x_0) \in \mathbb{R}$ (noting that $\mathrm{L}(\mathbb{R}; \mathbb{R})$ is isomorphic to $\mathbb{R}$ by Exercise **??**).

(iii) The ***Jacobian matrix*** of f at x_0 is the $m \times n$ matrix

$$\begin{bmatrix} D_1 f_1(x_0) & \cdots & D_n f_1(x_0) \\ \vdots & \ddots & \vdots \\ D_1 f_m(x_0) & \cdots & D_n f_m(x_0) \end{bmatrix}. \qquad \bullet$$

Note that we use the same terminology "jth partial derivative" for the specific case of the preceding definition as we used in the more general case of Definition 4.4.18. This is a legitimate source of possible confusion, but is also standard practice.

The next result follows immediately from Corollary 4.4.20, and is quite important since it tells us how one computes the derivative in practice.

4.4.22 Theorem (Explicit formula for the derivative) *If $\mathrm{U} \subseteq \mathbb{R}^n$ is an open set and if $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^m$ is a map differentiable at $\mathbf{x}_0 \in \mathrm{U}$ written as $\mathbf{f} = \mathrm{f}_1 \times \cdots \times \mathrm{f}_m$, then the components $\mathrm{f}_1, \dots, \mathrm{f}_m\colon \mathrm{U} \to \mathbb{R}$ of $\mathbf{f}$ are differentiable at $\mathbf{x}_0$ with respect to the jth coordinate for each $\mathrm{j} \in \{1, \dots, n\}$. Furthermore, the matrix representative of $\mathbf{Df}(\mathbf{x}_0)$ with respect to the standard bases $\mathscr{B}_n$ and $\mathscr{B}_m$ for $\mathbb{R}^n$ and $\mathbb{R}^m$ is the Jacobian matrix of $\mathbf{f}$ at $\mathbf{x}_0$.*

We shall frequently think of the derivative as being *equal* to its Jacobian matrix with the understanding that we are using the standard basis to represent the components of the derivative as a linear map. This is convenient to do, and is only a mild abuse at worst.

4.4.23 Notation (Alternative notation for the partial derivative) As with the notation for the derivative as discussed in Notation 3.2.2, there is notation for the partial derivative that sees more common use that the notation we give. Specifically, it is frequent to see the symbol $\frac{\partial f}{\partial x_j}$ used for what we denote by $D_j f$. This more common notation suffers from the same drawbacks as the notation $\frac{\mathrm{d}f}{\mathrm{d}x}$ for the ordinary derivative. Namely, it introduces the independent variable x_j in a potentially confusing way. Much of the time, this does not cause problems, and indeed we will use this notation when it is not imprudent to do so. $\bullet$

In Exercise 4.4.3 the reader can provide a rule that is often helpful in computing partial derivatives with respect to coordinates. Let us give a couple of examples to illustrate the notion of partial derivative and its connection with the derivative.

4.4.24 Examples (Partial derivative)

1. Let $U = \mathbb{R}^2 \setminus \{(0,0)\}$ and define $f\colon U \to \mathbb{R}^2$ by $f(x_1, x_2) = \left(\frac{x_1}{\sqrt{x_1^2+x_2^2}}, \frac{x_2}{\sqrt{x_1^2+x_2^2}}\right)$. We claim that f possesses both partial derivatives at all points in U. Indeed, we compute

$$\begin{aligned}\lim_{h\to 0} &\frac{\left(\frac{x_1+h}{\sqrt{(x_1+h)^2+x_2^2}}, \frac{x_2}{\sqrt{(x_1+h)^2+x_2^2}}\right) - \left(\frac{x_1}{\sqrt{x_1^2+x_2^2}}, \frac{x_2}{\sqrt{x_1^2+x_2^2}}\right)}{h} \\ &= \left(\lim_{h\to 0} \frac{\frac{x_1+h}{\sqrt{(x_1+h)^2+x_2^2}} - \frac{x_1}{\sqrt{x_1^2+x_2^2}}}{h}, \lim_{h\to 0} \frac{\frac{x_2}{\sqrt{(x_1+h)^2+x_2^2}} - \frac{x_2}{\sqrt{x_1^2+x_2^2}}}{h}\right) \\ &= \left(\frac{x_2^2}{(x_1^2+x_2^2)^{3/2}}, -\frac{x_1x_2}{(x_1^2+x_2^2)^{3/2}}\right),\end{aligned}$$

where, in the last step, we have simply computed the usual derivative, using the rules given in Section 3.2. In like manner we have

$$\lim_{h\to 0} \frac{\left(\frac{x_1}{\sqrt{x_1^2+x_2^2}}, \frac{x_2+h}{\sqrt{x_1^2+(x_2+h)^2}}\right) - \left(\frac{x_1}{\sqrt{x_1^2+(x_2+h)^2}}, \frac{x_2}{\sqrt{x_1^2+x_2^2}}\right)}{h} = \left(-\frac{x_1x_2}{(x_1^2+x_2^2)^{3/2}}, \frac{x_1^2}{(x_1^2+x_2^2)^{3/2}}\right)$$

Thus both partial derivatives indeed exist, and we moreover have

$$\begin{aligned}\boldsymbol{D}_1 f(x_1, x_2) &= \left(\frac{x_2^2}{(x_1^2+x_2^2)^{3/2}}, -\frac{x_1x_2}{(x_1^2+x_2^2)^{3/2}}\right), \\ \boldsymbol{D}_2 f(x_1, x_2) &= \left(-\frac{x_1x_2}{(x_1^2+x_2^2)^{3/2}}, \frac{x_1^2}{(x_1^2+x_2^2)^{3/2}}\right),\end{aligned}$$

and so the partial derivatives are also continuous functions on U.
Therefore, *if* f is differentiable at some point $(x_{01}, x_{02}) \in \mathbb{R}^2 \setminus \{(0,0)\}$ then it must hold that

$$\boldsymbol{D}f(x_{01}, x_{02}) = \begin{bmatrix} \frac{x_{02}^2}{(x_{01}^2+x_{02}^2)^{3/2}} & -\frac{x_{01}x_{02}}{(x_{01}^2+x_{02}^2)^{3/2}} \\ -\frac{x_{01}x_{02}}{(x_{01}^2+x_{02}^2)^{3/2}} & \frac{x_{01}^2}{(x_{01}^2+x_{02}^2)^{3/2}} \end{bmatrix},$$

where we identify the derivative with its matrix representative in the standard basis. We should, at this point since we know no better, actually verify that f is differentiable with this derivative. This can be done directly using the definition of derivative. Thus one can check directly, using rules for limits as in Proposition 2.3.23, that

$$\begin{aligned}\lim_{(x_1,x_2)\to(x_{01},x_{02})} &\Big(\|f(x_1, x_2) - f(x_{01}, x_{02}) \\ &\quad - \boldsymbol{D}f(x_{01}, x_{02}) \cdot (x_1 - x_{01}, x_2 - x_{02})\|_{\mathbb{R}^2}\Big) / \|(x_1, x_2) - (x_{01}, x_{02})\|_{\mathbb{R}^2} = 0.\end{aligned}$$

We leave the tedious verification of this to the reader, particularly as we shall see in Theorem 4.4.25 below that in this example there is an easy way to verify that this function is, in fact, of class C^1.

2. Define $f\colon \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x_1, x_2) = \begin{cases} \frac{2x_1^2 x_2}{x_1^4 + x_2^2}, & (x_1, x_2) \neq (0, 0), \\ 0, & (x_1, x_2) = (0, 0). \end{cases}$$

We claim that f possesses both partial derivatives at $(0, 0)$, but is not differentiable at $(0, 0)$. Let us first show that f possesses both partial derivative at $(0, 0)$. By definition, this amounts to checking the differentiability (in the sense of Definition 3.2.1) of the function $x_1 \mapsto f(x_1, 0) = 0$. This function, being constant, is obviously differentiable at $(0, 0)$ with derivative zero. In like manner one can show that f possesses the second partial derivative at $(0, 0)$ and that this second partial derivative is also zero. Now let us show that f is discontinuous, and therefore not differentiable, at $(0, 0)$. Consider the sequence $((\frac{1}{j}, \frac{1}{j^2}))_{j \in \mathbb{Z}_{>0}}$ in $\mathbb{R}^2$. This sequence converges to $(0, 0)$. We directly compute that $f(\frac{1}{j}, \frac{1}{j^2}) = 1$ for all $j \in \mathbb{Z}_{>0}$. Therefore

$$\lim_{j \to \infty} f(\tfrac{1}{j}, \tfrac{1}{j^2}) = 1 \neq f(0, 0).$$

Therefore, f is indeed discontinuous, and so not differentiable, at $(0, 0)$ by Proposition 4.4.35 below.

Note that the function of Example 4.4.13 also has the property that its partial derivatives exist, but the function is not differentiable. •

The preceding examples illustrate one of the problems that one has with the derivative: it is often not so easy to verify its existence since the mere existence of all partial derivatives is not sufficient. There is an important case, however, where one can infer differentiability from the properties of the partial derivatives. Here we return to the general setup for the partial derivative in terms of products.

4.4.25 Theorem (Equivalence of continuous differentiability and continuity of partial derivatives) *For an open set* $\mathrm{U} \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$, *a map* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^m$, *and for* $\mathrm{r} \in \mathbb{Z}_{>0}$, *the following statements are equivalent:*

(i) $\mathbf{f}$ *is of class* C^r;

(ii) the partial derivatives $\mathbf{D}_\mathrm{j}\mathbf{f}(\mathbf{x})$ *exist for each* $\mathrm{j} \in \{1, \ldots, \mathrm{k}\}$ *and* $\mathbf{x} \in \mathrm{U}$, *and, moreover, the maps* $\mathbf{x} \mapsto \mathbf{D}_\mathrm{j}\mathbf{f}(\mathbf{x})$ *are of class* $\mathrm{C}^{\mathrm{r}-1}$.

Proof By induction we can assume without loss of generality that $k = 2$. Moreover, by Propositions 4.3.26 and 4.4.17 we can take $m = 1$ without loss of generality. Thus we prove the theorem for $k = 2$ and $m = 1$. Consistent with our standing conventions we write "$\boldsymbol{f}$" as "f."

(i) $\implies$ (ii) From Theorem 4.4.19 we know that the partial derivatives $\boldsymbol{D}_1 f(\boldsymbol{x})$ and $\boldsymbol{D}_2 f(\boldsymbol{x})$ exist at all points $\boldsymbol{x} \in U$. To prove continuity of the partial derivatives, define maps

$$\phi_1\colon \mathrm{L}(\mathbb{R}^{n_1} \oplus \mathbb{R}^{n_2}; \mathbb{R}) \to \mathrm{L}(\mathbb{R}^{n_1}; \mathbb{R}), \quad \phi_2\colon \mathrm{L}(\mathbb{R}^{n_1} \oplus \mathbb{R}^{n_2}; \mathbb{R}) \to \mathrm{L}(\mathbb{R}^{n_2}; \mathbb{R})$$

by

$$\phi_1(\mathsf{L}_1)(v_1) = \mathsf{L}_1(v_1, \mathbf{0}), \quad \phi_2(\mathsf{L}_2)(v_2) = \mathsf{L}_1(\mathbf{0}, v_2)$$

for $v_1 \in \mathbb{R}^{n_1}$ and $v_2 \in \mathbb{R}^{n_2}$. These maps are easily verified to be linear and so in particular are infinitely differentiable, cf. Corollary 4.4.10. Moreover, we easily see that

$$\boldsymbol{D}_1 f = \phi_1 \circ \boldsymbol{D} f, \quad \boldsymbol{D}_2 f = \phi_2 \circ \boldsymbol{D} f.$$

Therefore, if $\boldsymbol{D}f$ is of class C^{r-1} (as it is by Proposition 4.4.35) then the partial derivatives are also of class C^{r-1}.

(ii) $\implies$ (i) First we show that $\boldsymbol{D}f(x)$ exists for all $x \in U$ if all partial derivatives exist at each point, and are continuous. Let us fix $x = (x_1, x_2) \in U$ and let $(h_1, h_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be such that $(x_1 + s_1 h_1, x_2 + s_2 h_2) \in U$ for $s_1, s_2 \in [0, 1]$, this being possible since U is open. Consider the map

$$s \mapsto f(x_1, x_2 + s h_2).$$

By the Chain Rule (Theorem 4.4.49), it being applicable since the partial derivative of f with respect to the second component exists, we have

$$\frac{\mathrm{d}}{\mathrm{d}s} f(x_1, x_2 + s h_2) = \boldsymbol{D}_2 f(x_1, x_2 + s h_2) \cdot h_2.$$

By the multivariable Fundamental Theorem of Calculus (this is obtained in this case by applying the single-variable Fundamental Theorem componentwise, but the reader can also refer ahead to ***missing stuff***) we have

$$f(x_1, x_2 + h_2) - f(x_1, x_2) = \int_0^1 \boldsymbol{D}_2 f(x_1, x_2 + s h_2) \cdot h_2 \, \mathrm{d}s. \tag{4.23}$$

The same argument can be applied to the map

$$s \mapsto f(x_1 + s h_1, x_2 + h_2)$$

to give

$$f(x_1 + h_1, x_2 + h_2) - f(x_1, x_2 + h_2) = \int_0^1 \boldsymbol{D}_1 f(x_1 + s h_1, x_2 + h_2) \cdot h_1 \, \mathrm{d}s. \tag{4.24}$$

Combining (4.23) and (4.24) we get

$$\begin{aligned}
&f(x_1 + h_1, x_2 + h_2) - f(x_1, h_2) - \boldsymbol{D}_1 f(x_1, x_2) \cdot h_1 - \boldsymbol{D}_2 f(x_1, x_2) \cdot h_2 \\
&\quad = \int_0^1 \boldsymbol{D}_1 f(x_1 + s h_1, x_2 + h_2) \cdot h_1 \, \mathrm{d}s + \int_0^1 \boldsymbol{D}_2 f(x_1, x_2 + s h_2) \cdot h_2 \, \mathrm{d}s \\
&\qquad - \boldsymbol{D}_1 f(x_1, x_2) \cdot h_1 - \boldsymbol{D}_2 f(x_1, x_2) \cdot h_2 \\
&\quad = \Big(\int_0^1 \big(\boldsymbol{D}_2 f(x_1 + s h_1, x_2 + h_2) - \boldsymbol{D}_1 f(x_1, x_2) \big) \mathrm{d}s \Big) \cdot h_1 + \\
&\qquad \Big(\int_0^1 \big(\boldsymbol{D}_2 f(x_1, x_2 + s h_2) - \boldsymbol{D}_2 f(x_1, x_2) \big), \mathrm{d}s \Big) \cdot h_2
\end{aligned}$$

Now let $\epsilon \in \mathbb{R}_{>0}$ and by continuity of the partial derivatives choose $(\boldsymbol{h}_1, \boldsymbol{h}_2)$ such that

$$\sup\{\|\boldsymbol{D}_2 f(\boldsymbol{x}_1 + s\boldsymbol{h}_1, \boldsymbol{x}_2 + \boldsymbol{h}_2) - \boldsymbol{D}_1 f(\boldsymbol{x}_1, \boldsymbol{x}_2)\|_{\mathbb{R}^n, \mathbb{R}^m} \mid s \in [0,1]\} < \frac{\epsilon}{2\sqrt{2}}$$
$$\sup\{\|\boldsymbol{D}_2 f(\boldsymbol{x}_1, \boldsymbol{x}_2 + s\boldsymbol{h}_2) - \boldsymbol{D}_2 f(\boldsymbol{x}_1, \boldsymbol{x}_2)\|_{\mathbb{R}^n, \mathbb{R}^m} \mid s \in [0,1]\} < \frac{\epsilon}{2\sqrt{2}}.$$

With $(\boldsymbol{h}_1, \boldsymbol{h}_2)$ so chosen we have

$$\begin{aligned}\Big|\Big(\boldsymbol{D}_2 f(\boldsymbol{x}_1 + s\boldsymbol{h}_1, \boldsymbol{x}_2 + \boldsymbol{h}_2) - \boldsymbol{D}_1 f(\boldsymbol{x}_1, \boldsymbol{x}_2)\Big) \cdot \boldsymbol{h}_1 \\ + \Big(\boldsymbol{D}_2 f(\boldsymbol{x}_1, \boldsymbol{x}_2 + s\boldsymbol{h}_2) - \boldsymbol{D}_2 f(\boldsymbol{x}_1, \boldsymbol{x}_2)\Big) \cdot \boldsymbol{h}_2\Big| &\leq \tfrac{\epsilon}{2\sqrt{2}}\|\boldsymbol{h}_1\|_{\mathbb{R}^{n_1}} + \tfrac{\epsilon}{2\sqrt{2}}\|\boldsymbol{h}_2\|_{\mathbb{R}^{n_2}} \\ &\leq \tfrac{\epsilon}{\sqrt{2}}(\|\boldsymbol{h}_1\|_{\mathbb{R}^{n_1}} + \|\boldsymbol{h}_2\|_{\mathbb{R}^{n_2}}) \leq \epsilon\|(\boldsymbol{h}_1, \boldsymbol{h}_2)\|_{\mathbb{R}^{n_1+n_2}},\end{aligned}$$

using Lemma 4.2.67. Therefore,

$$\begin{aligned}\Big|f(\boldsymbol{x}_1 + \boldsymbol{h}_1, \boldsymbol{x}_2 + \boldsymbol{h}_2) - f(\boldsymbol{x}_1, \boldsymbol{h}_2) - \boldsymbol{D}_1 f(\boldsymbol{x}_1, \boldsymbol{x}_2) \cdot \boldsymbol{h}_1 - \\ \boldsymbol{D}_2 f(\boldsymbol{x}_1, \boldsymbol{x}_2) \cdot \boldsymbol{h}_2\Big| / \|(\boldsymbol{h}_1, \boldsymbol{h}_2)\|_{\mathbb{R}^{n_1+n_2}} < \epsilon,\end{aligned}$$

and so we conclude that f is differentiable at $(\boldsymbol{x}_1, \boldsymbol{x}_2)$.

Finally, we show that $\boldsymbol{D}f$ is of class C^{r-1} if both $\boldsymbol{D}_1 f$ and $\boldsymbol{D}_2 f$ are of class C^{r-1}. Define maps

$$\psi_1 \colon \mathrm{L}(\mathbb{R}^{n_1}; \mathbb{R}) \to \mathrm{L}(\mathbb{R}^{n_1} \oplus \mathbb{R}^{n_2}; \mathbb{R}), \quad \psi_2 \colon \mathrm{L}(\mathbb{R}^{n_2}; \mathbb{R}) \to \mathrm{L}(\mathbb{R}^{n_1} \oplus \mathbb{R}^{n_2}; \mathbb{R})$$

by

$$\psi_1(\mathsf{L}_1)(\boldsymbol{v}_1, \boldsymbol{v}_2) = \mathsf{L}_1(\boldsymbol{v}_1), \quad \psi_2(\mathsf{L}_2)(\boldsymbol{v}_1, \boldsymbol{v}_2) = \mathsf{L}_2(\boldsymbol{v}_2).$$

These maps are linear and so infinitely differentiable. Moreover, since

$$\boldsymbol{D}f(\boldsymbol{x}) = \psi_1 \circ \boldsymbol{D}_1 f + \psi_2 \circ \boldsymbol{D}_2 f$$

it follows that $\boldsymbol{D}f$ is of class C^{r-1} if $\boldsymbol{D}_1 f$ and $\boldsymbol{D}_2 f$ are of class C^{r-1} by virtue of Proposition 4.4.47. ■

Let us consider the theorem in view of the examples we introduced above.

4.4.26 Examples (Partial derivatives (cont'd))

1. We take $U = \mathbb{R}^2 \setminus \{(0,0)\}$ and take $\boldsymbol{f} \colon U \to \mathbb{R}^2$ given by $\boldsymbol{f}(x_1, x_2) = \Big(\frac{x_1}{\sqrt{x_1^2 + x_2^2}}, \frac{x_2}{\sqrt{x_1^2 + x_2^2}}\Big)$. In Example 4.4.24–1 we computed

$$\boldsymbol{D}\boldsymbol{f}(x_1, x_2) = \begin{bmatrix} \frac{x_2^2}{(x_1^2 + x_2^2)^{3/2}} & -\frac{x_1 x_2}{(x_1^2 + x_2^2)^{3/2}} \\ -\frac{x_1 x_2}{(x_1^2 + x_2^2)^{3/2}} & \frac{x_1^2}{(x_1^2 + x_2^2)^{3/2}} \end{bmatrix}.$$

Since the components of this matrix are continuous functions on U, it follows from Theorem 4.4.25 that $\boldsymbol{f}$ is of class C^1 on U.

2. Here we take $f\colon \mathbb{R}^2 \to \mathbb{R}$ to be defined by

$$f(x_1, x_2) = \begin{cases} \frac{2x_1^2 x_2}{x_1^4 + x_2^2}, & (x_1, x_2) \neq (0,0), \\ 0, & (x_1, x_2) = (0,0). \end{cases}$$

In Example 4.4.24–2 we showed that both partial derivatives of f exist at $(0,0)$ and are zero. For $(x_1, x_2) \neq (0,0)$ we can compute, using Theorem 4.4.22,

$$\boldsymbol{D}_1 f(x_1, x_2) = 2x_1 x_2 \frac{x_2^2 - x_1^4}{(x_1^4 + x_2^2)^2}, \quad \boldsymbol{D}_2 f(x_1, x_2) = x_1^2 \frac{x_1^4 - x_2^2}{(x_1^4 + x_2^2)^2}.$$

These partial derivatives are continuous on $\mathbb{R}^2 \setminus \{(0,0)\}$, and so it follows from Theorem 4.4.25 that f is of class C^1 on $\mathbb{R}^2 \setminus \{(0,0)\}$. However, the partial derivatives are readily verified to be discontinuous at $(0,0)$, cf. Example 4.4.24–2, and so it follows from Theorem 4.4.25 that f is not of class C^1 in any neighbourhood of $(0,0)$. Of course, we knew this already since f is actually discontinuous at $(0,0)$. •

4.4.5 Iterated partial derivatives

Now that we have used the notion of partial derivative to get better handle on how to compute the derivative of a multivariable map, let us see if we can similarly compute higher-order derivatives of multivariable maps using partial derivatives. In addressing this matter we will also shed some light on an important property of higher-order derivatives in the usual sense. In particular, we shall illuminate clearly the significance of the classical statement that "partial derivatives commute" by showing that this statement is not true in general.

Suppose we have an open set $U \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ and a map $\boldsymbol{f}\colon U \to \mathbb{R}^m$. Suppose that for $j_1 \in \{1, \dots, k\}$, $\boldsymbol{f}$ is continuously differentiable with respect to the j_1st component. That is, the map

$$U \ni \boldsymbol{x} \mapsto \boldsymbol{D}_{j_1} \boldsymbol{f}(\boldsymbol{x}) \in \mathrm{L}(\mathbb{R}^{n_1}; \mathbb{R}^m)$$

is defined and continuous. While there are weaker conditions that will guarantee this, to keep things simple let us suppose that $\boldsymbol{f}$ is of class C^1 so the existence and continuity of the partial derivative is ensured by Theorem 4.4.19. Now let $j_2 \in \{1, \dots, k\}$. We can then talk about the differentiability of the map $U \ni \boldsymbol{x} \mapsto \boldsymbol{D}_{j_1} \boldsymbol{f}(\boldsymbol{x})$ with respect to the j_2nd component. Indeed, while again weaker hypotheses are possible, if we assume that $\boldsymbol{f}$ is of class C^2 then the map

$$U \ni \boldsymbol{x} \mapsto \boldsymbol{D}_{j_2} \boldsymbol{D}_{j_1} \boldsymbol{f}(\boldsymbol{x}) \in \mathrm{L}(\mathbb{R}^{n_{j_2}}, \mathbb{R}^{n_{j_1}}; \mathbb{R}^m)$$

is defined and continuous by virtue of Theorem 4.4.19. (We use Proposition **??** to describe the codomain of this map.) Clearly, if $\boldsymbol{f}$ is of class C^r and if $j_1, \dots, j_r \in \{1, \dots, k\}$ then we can inductively define

$$U \ni \boldsymbol{x} \mapsto \boldsymbol{D}_{j_r} \cdots \boldsymbol{D}_{j_1} \boldsymbol{f}(\boldsymbol{x}) \in \mathrm{L}(\mathbb{R}^{n_{j_r}}, \dots, \mathbb{R}^{n_{j_1}}; \mathbb{R}^m),$$

again using Proposition **??**.

Let us organise the preceding discussion by naming the objects.

4.4.27 Definition (Iterated partial derivative) Let $U \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ be open, let $f\colon U \to \mathbb{R}^m$, let $x_0 \in U$, and let $j_1, \dots, j_r \in \{1, \dots, k\}$. The multilinear map

$$D_{j_r} \cdots D_{j_1} f(x_0) \in \mathrm{L}(\mathbb{R}^{n_{j_r}}, \dots, \mathbb{R}^{n_{j_1}}; \mathbb{R}^m),$$

when it is defined, is an ***iterated partial derivative*** of f at x_0. The number $r \in \mathbb{Z}_{>0}$ is the ***degree*** of the iterated partial derivative. •

Let us relate the rth derivative of f to the iterated partial derivatives of degree r. To do so we generalise the relationship in the case of $r = 1$ given in Theorem 4.4.19. This requires that we represent elements of $(\mathbb{R}^{n_1} \oplus \cdots \oplus \mathbb{R}^{n_k})^r$ is an appropriate way. A vector in $\mathbb{R}^{n_1} \oplus \cdots \oplus \mathbb{R}^{n_k}$ we write as $(v_1, \dots, v_k)$ for $v_j \in \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$. Thus we write an element of $(\mathbb{R}^{n_1} \oplus \cdots \oplus \mathbb{R}^{n_k})^r$ as

$$((v_{r1}, \dots, v_{rk}), \dots, (v_{11}, \dots, v_{1k}))$$

for $v_{aj} \in \mathbb{R}^{n_j}$, $a \in \{1, \dots, r\}$, $j \in \{1, \dots, k\}$. Note the ordering with respect to the first index: we list the vectors from r to 1, not from 1 to r. This is to be consistent with our ordering of indices for iterated partial derivatives from r to 1 as we go from left to right.

We now have the following generalisation of Theorem 4.4.19.

4.4.28 Theorem (Iterated partial derivatives and higher-order derivatives) *If $\mathrm{U} \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ is an open set and if $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^m$ is a map that is r times differentiable at $\mathbf{x}_0 \in \mathrm{U}$, then all iterated partial derivatives of $\mathbf{f}$ degree r are defined $\mathbf{x}_0$. Moreover, if $\mathbf{f}$ is r times differentiable at $\mathbf{x}_0$ then we have the following relationships between the derivative and the partial derivatives:*

(i) for $((\mathbf{v}_{r1}, \dots, \mathbf{v}_{rk}), \dots, (\mathbf{v}_{11}, \dots, \mathbf{v}_{1k})) \in (\mathbb{R}^{n_1} \oplus \cdots \oplus \mathbb{R}^{n_k})^r$ we have

$$\mathbf{D}^r \mathbf{f}(\mathbf{x}) \cdot ((\mathbf{v}_{r1}, \dots, \mathbf{v}_{1k}), \dots, (\mathbf{v}_{11}, \dots, \mathbf{v}_{1k})) = \sum_{j_1, \dots, j_r = 1}^{k} \mathbf{D}_{j_r} \cdots \mathbf{D}_{j_1} \mathbf{f}(\mathbf{x}) \cdot (\mathbf{v}_{rj_r}, \dots, \mathbf{v}_{1j_1}); \quad (4.25)$$

(ii) for $j_1, \dots, j_r \in \{1, \dots, k\}$ and $(\mathbf{v}_r, \dots, \mathbf{v}_1) \in \mathbb{R}^{n_{j_r}} \oplus \cdots \oplus \mathbb{R}^{n_{j_1}}$ we have

$$\mathbf{D}_{j_r} \cdots \mathbf{D}_{j_1} \mathbf{f}(\mathbf{x}) \cdot (\mathbf{v}_r, \dots, \mathbf{v}_1) = \mathbf{D}^r \mathbf{f}(\mathbf{x}) \cdot (\underbrace{(\mathbf{0}, \dots, \mathbf{v}_r, \dots, \mathbf{0})}_{\mathbf{v}_r \text{ in } j_r \text{th slot}}, \dots, \underbrace{(\mathbf{0}, \dots, \mathbf{v}_1, \dots, \mathbf{0})}_{\mathbf{v}_1 \text{ in } j_1 \text{st slot}}). \quad (4.26)$$

.

Proof We prove the first implication of the theorem by induction on r. We do this by simultaneously proving (4.26) by in the induction argument. For $r = 1$ the assertion and (4.26) is simply Theorem 4.4.19. So suppose the result true for $r \in \{1, \dots, s\}$ and suppose that f is $s + 1$ times differentiable at x_0. By the induction hypothesis we have that all iterated partial derivatives of degree s exist and satisfy

$$D_{j_s} \cdots D_{j_1} f(x) \cdot (v_s, \dots, v_1) = D^s f(x) \cdot (\underbrace{(\mathbf{0}, \dots, v_s, \dots, \mathbf{0})}_{v_s \text{ in } j_s \text{th slot}}, \dots, \underbrace{(\mathbf{0}, \dots, v_1, \dots, \mathbf{0})}_{v_1 \text{ in } j_1 \text{st slot}}).$$

By Proposition 4.4.7 and Theorem 4.4.19, differentiability of $D^s f$ at x_0 implies that all iterated partial derivatives of degree $s+1$ exist at x_0. To prove that (4.26) holds for $r = s+1$ we compute

$$\begin{aligned}
&D_{j_{s+1}} D_{j_s} \cdots D_{j_1} f(x) \cdot (v_{s+1}, v_s, \dots, v_1) \\
&\qquad = \Big(D_{j_{s+1}} (D^s f(x) \cdot (\underbrace{(0, \dots, v_s, \dots, 0)}_{v_s \text{ in } j_s\text{th slot}}, \dots, \underbrace{(0, \dots, v_1, \dots, 0)}_{v_1 \text{ in } j_1\text{st slot}})) \Big) \cdot v_{s+1} \\
&\qquad = D^{s+1} f(x) \cdot (\underbrace{(0, \dots, v_{s+1}, \dots, 0)}_{v_{s+1} \text{ in } j_{s+1}\text{st slot}}, \underbrace{(0, \dots, v_s, \dots, 0)}_{v_s \text{ in } j_s\text{th slot}}, \dots, \\
&\qquad\quad \underbrace{(0, \dots, v_1, \dots, 0)}_{v_1 \text{ in } j_1\text{st slot}})
\end{aligned}$$

using the induction hypotheses and Theorem 4.4.19. This gives (4.26) for $r = s+1$.

Finally we need to show that (4.25) holds. We prove this also by induction on r. For $r = 1$ the formula holds by Theorem 4.4.19. Suppose, then, that (4.25) holds for $r = s$ and that f is $s+1$ times differentiable at x_0. Using the fact that the formula holds for $r = s$, we compute

$$\begin{aligned}
&D^{s+1} f(x) \cdot ((v_{(s+1)1}, \dots, v_{(s+1)k}), (v_{s1}, \dots, v_{sk}), \dots, (v_{11}, \dots, v_{1k})) \\
&\qquad = \sum_{j_{s+1}=1}^{k} \Big(D_{j_{s+1}} (D^s f \cdot ((v_{s1}, \dots, v_{sk}), \dots, (v_{11}, \dots, v_{1k}))) \Big) \cdot v_{(s+1)j_{s+1}} \\
&\qquad = \sum_{j_{s+1}=1}^{k} \Big(D_{j_{s+1}} \Big(\sum_{j_1, \dots, j_s = 1}^{k} D_{j_s} \cdots D_{j_1} f(x) \cdot (v_{sj_s}, \dots, v_{1j_1}) \Big) \Big) \cdot v_{(s+1)j_{s+1}} \\
&\qquad = \sum_{j_1, \dots, j_s, j_{s+1} = 1}^{k} D_{j_{s+1}} D_{j_s} \cdots D_{j_1} f(x) \cdot (v_{(s+1)j_{s+1}}, v_{sj_s}, \dots, v_{1j_1}),
\end{aligned}$$

giving (4.25) for $r = s+1$. ∎

Since the preceding theorem contains Theorem 4.4.19 as a special case, it follows that the converse does not hold. That is to say, the existence of iterated partial derivatives of degree r does not imply that f is r times differentiable. We refer to the discussion surrounding Theorem 4.4.19 for more details.

Just as Theorem 4.4.19 allowed us to give an explicit formula for the derivative in Theorem 4.4.22, we can use apply Theorem 4.4.28 to give an explicit formula for higher-order derivatives.

4.4.29 Definition (Iterated partial derivative) Let $U \subseteq \mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ be open, let $f \colon U \to \mathbb{R}^m$, let $x_0 \in U$, and let $j_1, \dots, j_r \in \{1, \dots, n\}$. The multilinear map

$$D_{j_r} \cdots D_{j_1} f(x_0) \in \mathbb{R}^m$$

(noting that $\mathrm{L}^r(\mathbb{R}; \mathbb{R}^m)$ is isomorphic to $\mathbb{R}^m$ by Exercise **??**) when it is defined, is an ***iterated partial derivative*** of f at x_0. The number $r \in \mathbb{Z}_{>0}$ is the ***degree*** of the iterated partial derivative. •

Now, an application of Proposition 4.4.17 and Theorem 4.4.28 gives the following result.

4.4.30 Theorem (Explicit formula for higher-order derivatives) *If* $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}}$ *is open and if* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^{\mathrm{m}}$ *is a map that is* r *times differentiable at* $\mathbf{x}_0$ *and is written as* $\mathbf{f} = \mathrm{f}_1 \times \cdots \times \mathrm{f}_{\mathrm{m}}$, *then all iterated partial derivatives of degree* r *of components* $\mathrm{f}_1, \ldots, \mathrm{f}_{\mathrm{m}}\colon \mathrm{U} \to \mathbb{R}$ *exist at* $\mathbf{x}_0$. *Furthermore, the components of* $\mathbf{D}^{\mathrm{r}}\mathbf{f}(\mathbf{x}_0) \in \mathrm{L}^{\mathrm{r}}(\mathbb{R}^{\mathrm{n}}; \mathbb{R}^{\mathrm{m}})$ *are defined by*

$$(\mathbf{D}^{\mathrm{r}}\mathbf{f}(\mathbf{x}_0)(\mathbf{e}_{\mathrm{j}_{\mathrm{r}}}, \ldots, \mathbf{e}_{\mathrm{j}_1}))_{\mathrm{a}} = \mathbf{D}_{\mathrm{j}_{\mathrm{r}}} \cdots \mathbf{D}_{\mathrm{j}_1} \mathrm{f}_{\mathrm{a}}(\mathbf{x}_0),$$

for $\mathrm{j}_1, \ldots, \mathrm{j}_{\mathrm{r}} \in \{1, \ldots, \mathrm{n}\}$ *and* $\mathrm{a} \in \{1, \ldots, \mathrm{m}\}$.

In terms of more commonly used notation, the components of $\boldsymbol{D}^r \boldsymbol{f}(\boldsymbol{x}_0)$ are written as

$$\frac{\partial^r f_a}{\partial x_{j_r} \cdots \partial x_{j_1}}(\boldsymbol{x}_0), \qquad j_1, \ldots, j_r \in \{1, \ldots, n\},\ a \in \{1, \ldots, m\}.$$

The following theorem generalises Theorem 4.4.25 and shows that, as long as the iterated partial derivatives are continuous, one can assert higher-order continuous differentiability.

4.4.31 Theorem (Higher-order continuous differentiability and continuity of iterated partial derivatives) *Let* $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}_1} \times \cdots \times \mathbb{R}^{\mathrm{n}_{\mathrm{k}}}$ *be open, let* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^{\mathrm{m}}$, *and let* $\mathrm{r} \in \mathbb{Z}_{>0}$. *Then the following statements are equivalent:*

(i) $\mathbf{f}$ *is of class* C^{r};

(ii) all iterated partial derivatives of $\mathbf{f}$ *of degree* r *exist and are continuous.*

Proof (i) $\implies$ (ii) By Theorem 4.4.25, if f is of class C^r then $\boldsymbol{D}_{j_1} f$ is of class C^{r-1} for every $j_1 \in \{1, \ldots, k\}$. Inductively using Theorem 4.4.25, it then follows that $\boldsymbol{D}_{j_r} \cdots \boldsymbol{D}_{j_1} f$ is defined and continuous for every $j_1, \ldots, j_r \in \{1, \ldots, k\}$.

(ii) $\implies$ (i) We prove this implication by induction on r. As part of the proof we shall prove, included in the induction, that (4.25) holds under the assumption that iterated partial derivatives of degree r exist. By Theorem 4.4.25 it holds that if all iterated partial derivatives of degree 1 (i.e., all partial derivatives) exist and are continuous then f is of class C^1. Moreover, we showed in the proof of Theorem 4.4.25 that (4.25) holds for $r = 1$. Suppose the implication and (4.25) are true for $r \in \{1, \ldots, s\}$ and suppose that all iterated partial derivatives of degree $s+1$ exist and are continuous. By the induction hypothesis the map $\boldsymbol{x} \mapsto \boldsymbol{D}^s f(\boldsymbol{x})$ is defined and continuous. Moreover, the assumption that all iterated derivatives of degree $s+1$ exist and are continuous implies, by Proposition 4.4.7 and (4.25) with $r = s$, that all partial derivatives of $\boldsymbol{D}^s f$ exist and are continuous. Thus, by Theorem 4.4.25, $\boldsymbol{D}^s f$ is continuously differentiable and so f is of class C^{s+1}. The proof that (4.25) holds for $r = s+1$ is then carried out just as in the proof of Theorem 4.4.28. ∎

Next we discuss an important idea, that of commutativity of iterated partial derivatives. That is, we consider an open subset $U \subseteq \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$ and a map $f\colon U \to \mathbb{R}^m$ for which the iterated partial derivatives $\boldsymbol{D}_{j_1}\boldsymbol{D}_{j_2} f$ and $\boldsymbol{D}_{j_2}\boldsymbol{D}_{j_1} f$ exist at $\boldsymbol{x}_0 \in U$ for some $j_1, j_2 \in \{1, \ldots, k\}$. The question is, "When are these iterated partial derivatives equal?" Clearly they cannot be equal when $n_{j_1} \neq n_{j_2}$ since $\boldsymbol{D}_{j_1}\boldsymbol{D}_{j_2} f \in \mathrm{L}(\mathbb{R}^{n_{j_1}}, \mathbb{R}^{n_{j_2}}; \mathbb{R}^m)$ and $\boldsymbol{D}_{j_2}\boldsymbol{D}_{j_1} f \in \mathrm{L}(\mathbb{R}^{n_{j_2}}, \mathbb{R}^{n_{j_1}}; \mathbb{R}^m)$. Even when $n_{j_1} = n_{j_2}$ they are not generally equal.

4.4.32 Example (Partial derivatives do not generally commute) Let $\mathsf{B} \in \bigwedge^2(\mathbb{R}^n; \mathbb{R}^m)$; that is, B is a skew-symmetric bilinear map from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^m$. Let us define $f_{\mathsf{B}} \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ by $f_{\mathsf{B}}(x_1, x_2) = \mathsf{B}(x_1, x_2)$. By Theorem 4.4.8 we have

$$\begin{aligned} D_1 f_{\mathsf{B}}(x_1, x_2) \cdot v &= \mathsf{B}(v, x_2) \\ D_2 f_{\mathsf{B}}(x_1, x_2) \cdot v &= \mathsf{B}(x_1, v) \\ D_1 D_1 f_{\mathsf{B}}(x_1, x_2) \cdot (v_1, v_2) &= 0 \\ D_1 D_2 f_{\mathsf{B}}(x_1, x_2) \cdot (v_1, v_2) &= \mathsf{B}(v_1, v_2) \\ D_2 D_1 f_{\mathsf{B}}(x_1, x_2) \cdot (v_1, v_2) &= \mathsf{B}(v_2, v_1) \\ D_2 D_2 f_{\mathsf{B}}(x_1, x_2) \cdot (v_1, v_2) &= 0 \end{aligned}$$

for all $x_1, x_2, v, v_1, v_2 \in \mathbb{R}^n$. Since B is skew-symmetric, we have

$$D_1 D_2 f_{\mathsf{B}}(x_1, x_2) = D_2 D_1 f_{\mathsf{B}}(x_1, x_2)$$

if and only if $\mathsf{B} = 0$ (why?). Since the only case when B *must* be zero is when $n = 1$ (why?), we conclude that there are lots of possible choices for B when $n \geq 2$ for which the partial derivatives do not commute. •

The preceding example showing that partial derivative do not generally commute is not deep. However, it does help to provide a context as to why, when $n_{j_1} = n_{j_2} = 1$ it follows that partial derivatives do, indeed, commute. In particular, we hope that this suggests that the commuting of partial derivatives in this case is somewhat deep.

4.4.33 Theorem (One-dimensional partial derivatives commute) *Let* $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}_1} \times \cdots \times \mathbb{R}^{\mathrm{n}_\mathrm{k}}$ *be open, let* $\mathbf{f} \colon \mathrm{U} \to \mathbb{R}^{\mathrm{m}}$ *be of class* C^2*, and let* $\mathrm{j}_1, \mathrm{j}_2 \in \{1, \dots, \mathrm{k}\}$ *have the property that* $\mathrm{n}_{\mathrm{j}_1} = \mathrm{n}_{\mathrm{j}_2} = 1$*. Then*

$$\mathbf{D}_{\mathrm{j}_1} \mathbf{D}_{\mathrm{j}_2} \mathbf{f}(\mathbf{x}) = \mathbf{D}_{\mathrm{j}_2} \mathbf{D}_{\mathrm{j}_1} \mathbf{f}(\mathbf{x})$$

for all $\mathbf{x} \in \mathrm{U}$*.*

Proof By Proposition 4.4.17 we can assume that $m = 1$ without loss of generality. We thus denote "f" by "f." Let us write a point in U as

$$(x_1, \dots, x_{j_1}, \dots, x_{j_2}, \dots, x_k) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R} \times \cdots \times \mathbb{R} \times \cdots \times \mathbb{R}^{n_k}.$$

For each $j \in \{1, \dots, k\}$ choose $x_{0j} \in \mathbb{R}^{n_j}$ so that

$$x_0 \triangleq (x_{01}, \dots, x_{0k}) \in U.$$

If f is of class C^2 then the map

$$g \colon (s_1, s_2) \mapsto f(x_{01}, \dots, x_{j_1 0} + s_1, \dots, x_{j_2 0} + s_2, \dots, x_{0k})$$

is of class C^2 in a neighbourhood of $(0, 0) \in \mathbb{R}^2$. Moreover, by definition of the partial derivatives,

$$D_1 D_2 g(0, 0) = D_{j_1} D_{j_2} f(x_0), \quad D_2 D_1 g(0, 0) = D_{j_2} D_{j_1} f(x_0).$$

Thus is suffices to show that $\boldsymbol{D}_1\boldsymbol{D}_2 g(0,0) = \boldsymbol{D}_2\boldsymbol{D}_1 g(0,0)$.

For (s_1, s_2) in a neighbourhood of $(0,0)$ define

$$D(s_1, s_2) = g(s_1, s_2) - g(s_1, 0) - g(0, s_2) + g(0,0).$$

For fixed s_2 define $g_{s_2}(s_1) = g(s_1, s_2) - g(s_1, 0)$ so that $D(s_1, s_2) = g_{s_2}(s_1) - g_{s_2}(0)$. By the Mean Value Theorem, Theorem 3.2.19, we have

$$D(s_1, s_2) = g_{s_2}(s_1) - g_{s_2}(0) = s_1 g'_{s_2}(\tilde{s}_1) = s_1(\boldsymbol{D}_1 g(\tilde{s}_1, s_2) - \boldsymbol{D}_1 g(\tilde{s}_1, 0))$$

for some $\tilde{s}_1 \in [0, s_1]$. Now we apply the Mean Value Theorem again to the function $s_2 \mapsto \boldsymbol{D}_1 g(\tilde{s}_1, s_2)$ to get

$$\boldsymbol{D}_1 g(\tilde{s}_1, s_2) - \boldsymbol{D}_1 g(\tilde{s}_1, 0) = s_2 \boldsymbol{D}_2\boldsymbol{D}_1 g(\tilde{s}_1, \tilde{s}_2).$$

Putting the preceding two formulae together we get

$$\boldsymbol{D}_2\boldsymbol{D}_1 g(\tilde{s}_1, \tilde{s}_2) = \frac{D(s_1, s_2)}{s_1 s_2}.$$

Continuity of the iterated partial derivatives of length two gives

$$\boldsymbol{D}_2\boldsymbol{D}_1 g(0,0) = \lim_{(s_1,s_2)\to(0,0)} \frac{D(s_1, s_2)}{s_1 s_2}$$

The above construction can be repeated, swapping the rôles of s_1 and s_2, to give

$$\boldsymbol{D}_1\boldsymbol{D}_2 g(0,0) = \lim_{(s_1,s_2)\to(0,0)} \frac{D(s_1, s_2)}{s_1 s_2},$$

giving the result. ∎

Let us give a few examples to illuminate this important theorem.

4.4.34 Examples (Commutativity of one-dimensional partial derivatives)

1.

4.4.6 The derivative and function behaviour

Why is the derivative and differentiability important? Of course, this is an important question, and in this section we give some simple results that indicate why one might study the derivative of a map. Somewhat more profound illustrations of this are given in Section **??**.

As in the single-variable case, differentiability implies continuity.

4.4.35 Proposition (Differentiable maps are continuous) *If $\mathrm{U} \subseteq \mathbb{R}^n$ is an open set and if $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^m$ is differentiable at $\mathbf{x}_0 \in \mathrm{U}$, then there exists $\mathrm{M} \in \mathbb{R}_{>0}$ and a neighbourhood $\mathrm{V} \subseteq \mathrm{U}$ of $\mathbf{x}_0$ such that*

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\|_{\mathbb{R}^m} \leq \mathrm{M}\|\mathbf{x} - \mathbf{x}_0\|_{\mathbb{R}^n}, \qquad \mathbf{x} \in \mathrm{V}.$$

In particular, $\mathbf{f}$ is continuous at $\mathbf{x}_0$.

Proof By definition of "differentiable at x_0" there exists a neighbourhood V of x_0 such that

$$\begin{aligned} &\frac{\|f(x) - f(x_0) - Df(x_0) \cdot (x - x_0)\|_{\mathbb{R}^m}}{\|x - x_0\|_{\mathbb{R}^n}} < 1 \\ \implies \quad &\|f(x) - f(x_0) - Df(x_0) \cdot (x - x_0)\|_{\mathbb{R}^m} < \|x - x_0\|_{\mathbb{R}^n} \end{aligned}$$

for $x \in V$. By Proposition 4.1.13 we have

$$\|Df(x_0) \cdot v\|_{\mathbb{R}^m} \leq \|Df(x_0)\|_{\mathbb{R}^n,\mathbb{R}^m} \|v\|_{\mathbb{R}^n}$$

for all $v \in \mathbb{R}^n$. Thus the triangle inequality gives

$$\begin{aligned} \|f(x) - f(x_0)\|_{\mathbb{R}^m} &\leq \|f(x) - f(x_0) - Df(x_0) \cdot (x - x_0)\|_{\mathbb{R}^m} \\ &\quad + \|Df(x_0)\|_{\mathbb{R}^n,\mathbb{R}^m} \|x - x_0\|_{\mathbb{R}^n} \\ &\leq \|x - x_0\|_{\mathbb{R}^n} + \|Df(x_0)\|_{\mathbb{R}^n,\mathbb{R}^m} \|x - x_0\|_{\mathbb{R}^n} \end{aligned}$$

for all $x \in V$, giving the first assertion of the result if we take $M = 1 + \|Df(x_0)\|_{,\mathbb{R}^n} \mathbb{R}^m$.

For the final assertion, let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta' \in \mathbb{R}_{>0}$ be such that $\mathsf{B}(\delta', x_0) \subseteq V$. Taking $\delta = \min\{\delta', \frac{\epsilon}{M}\}$ and letting $x \in \mathsf{B}(\delta, x_0)$ gives

$$\|f(x) - f(x_0)\|_{\mathbb{R}^m} \leq M \|x - x_0\|_{\mathbb{R}^n} < \epsilon,$$

giving continuity of f at x_0. ∎

If the derivative of the function is bounded, then one can infer uniform continuity.

4.4.36 Proposition (Functions with bounded derivatives are sometimes uniformly continuous) *If* $\mathrm{U} \subseteq \mathbb{R}^n$ *is open and if* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^m$ *is continuously differentiable, then the following two statements hold:*

(i) if U *is convex (see the comments before the statement of Theorem 3.2.19 below) and if* $\mathbf{Df}$ *is bounded, then* $\mathbf{f}$ *is uniformly continuous;*

(ii) if $\mathrm{K} \subseteq \mathrm{U}$ *is compact, then* $\mathbf{f}|\mathrm{K}$ *is uniformly continuous.*

Proof (i) From the Mean Value Theorem, Theorem 3.2.19 below. there exists $M \in \mathbb{R}_{>0}$ such that

$$\|f(x) - f(y)\|_{\mathbb{R}^m} \leq M \|x - y\|_{\mathbb{R}^n}$$

for every $x, y \in U$. Now let $\epsilon \in \mathbb{R}_{>0}$ and let $x \in U$. Define $\delta = \frac{\epsilon}{M}$ and note that if $y \in U$ satisfies $\|x - y\|_{\mathbb{R}^n} < \delta$ then we have

$$\|f(x) - f(y)\|_{\mathbb{R}^m} < \epsilon,$$

giving the desired uniform continuity.

(ii) Let

$$\begin{aligned} A &= \sup\{\|f\|_{\mathbb{R}^m}(x) \mid x \in K\}, \\ B &= \sup\{\|Df(x)\|_{,|} \mathbb{R}^n\} \mathbb{R}^m x \in K, \end{aligned}$$

noting that $A, B < \infty$ by Theorem 4.3.31. Let $x \in K$ and let $r_x \in \mathbb{R}_{>0}$ be such that $\mathsf{B}^n(2r_x, x) \subseteq U$. For $y_1, y_2 \in \mathsf{B}^n(r_x, x)$, the Mean Value Theorem gives

$$\|f(y_1) - f(y_2)\|_{\mathbb{R}^m} \leq B\|y_1 - y_2\|_{\mathbb{R}^n}.$$

Since $(\mathsf{B}^n(r_x, x))_{x \in K}$ covers K, there exists $x_1, \ldots, x_k \in K$ such that $K \subseteq \cup_{j=1}^k \mathsf{B}^n(r_{x_j}, x_j)$. Let us abbreviate $N_j = \mathsf{B}^n(r_{x_j}, x_j)$ for $j \in \{1, \ldots, k\}$. By Theorem 4.2.38 there exists $r \in \mathbb{R}_{>0}$ such that if $x, y \in K$ satisfy $\|x - y\|_{\mathbb{R}^n} < r$ then $x, y \in N_j$ for some $j \in \{1, \ldots, k\}$.

We let $x, y \in K$. If $\|x - y\|_{\mathbb{R}^n} < r$ then $x, y \in N_j$ for some $j \in \{1, \ldots, k\}$ and so

$$\|f(x) - f(y)\|_{\mathbb{R}^m} \leq B\|x - y\|_{\mathbb{R}^n}.$$

If $\|x - y\|_{\mathbb{R}^n} \geq r$ then

$$\|f(x) - f(y)\|_{\mathbb{R}^m} \leq \|f(x)\|_{\mathbb{R}^m} + \|f(y)\|_{\mathbb{R}^m} \leq 2A = \frac{2Ar}{r} \leq 2r^{-1}A\|x - y\|_{\mathbb{R}^n}.$$

Taking $M = \max\{B, 2r^{-1}A\}$, we then have

$$\|f(x) - f(y)\|_{\mathbb{R}^m} \leq M\|x - y\|_{\mathbb{R}^n}$$

for all $x, y \in K$. Uniform continuity of f follows as in the proof of the first part of the result. ■

The two conditions in the preceding result are generally necessary, as the following example shows.

4.4.37 Example (A function with a bounded derivative that is not uniformly continuous) Consider the curve $\gamma\colon (1, \infty) \to \mathbb{R}^2$ defined by

$$\gamma(t) = (1 + \tanh(t - 1))(\cos(2\pi t), \sin(2\pi t)).$$

In Figure 4.10 we depict the traces of this curve, which is a spiral whose radius grows from a radius of 1 to a limiting radius of 2. Define

$$\begin{aligned} \phi\colon (-\tfrac{1}{8}, \tfrac{1}{8}) \times (1, \infty) &\to \mathbb{R}^2 \\ (s, t) &\mapsto (1 + \tanh(t + s - 1))(\cos(2\pi t), \sin(2\pi t)). \end{aligned}$$

Let us verify some of the elementary properties of this map.

1 Lemma *The map ϕ has the following properties:*

(i) it is injective;

(ii) it is continuously differentiable and there exists $c \in \mathbb{R}_{>0}$ such that $\|\mathbf{D}\phi(s, t)\|_{\mathbb{R}^2, \mathbb{R}^2} \geq c$ for all $(s, t) \in (-\frac{1}{8}, \frac{1}{8}) \times (1, \infty)$;

(iii) ϕ^{-1} is continuously differentiable with bounded derivative.

Proof (i) Suppose that $\phi(s_1, t_1) = \phi(s_2, t_2)$, and without loss of generality take $t_2 \geq t_1$. Since the two image points must lie on the same ray through the origin we must have

$$(\cos(2\pi t_1), \sin(2\pi t_1)) = (\cos(2\pi t_2), \sin(2\pi t_2)),$$

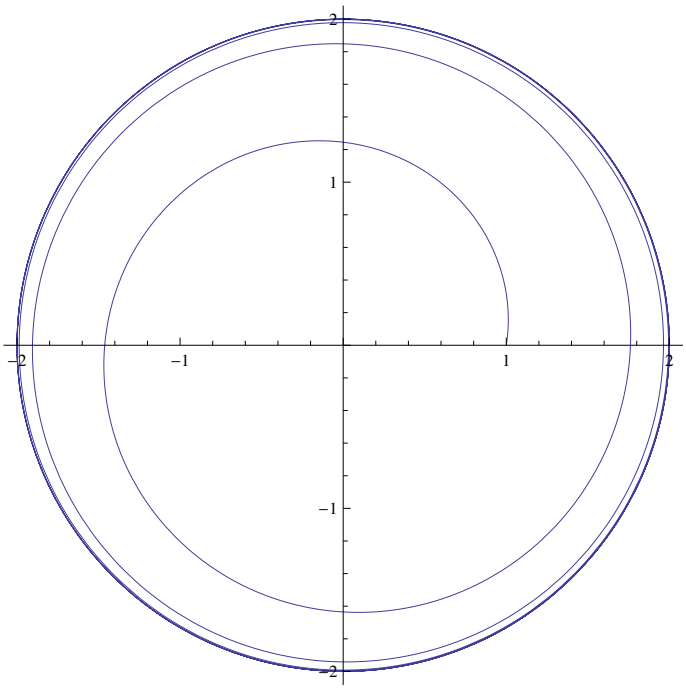


Figure 4.10 A spiral curve

implying that $t_2 - t_1 \in \mathbb{Z}_{\geq 0}$. If $t_1 = t_2$ then we must immediately have $1 + \tanh(t_1 + s_1 - 1) = 1 + \tanh(t_1 + s_2 - 1)$ giving $s_1 = s_2$ since tanh is injective (see Exercise 3.6.5). So suppose that $t_2 - t_1 = k \in \mathbb{Z}_{>0}$. Then we must have

$$\begin{aligned} & 1 + \tanh(t_1 + s_1 - 1) = 1 + \tanh(t_1 + k + s_2 - 1) \\ \implies \quad & t_1 + s_1 - 1 = s_1 + k + s_2 - 1 \\ \implies \quad & s_1 - s_2 = k. \end{aligned}$$

again using injectivity of tanh. However, since $s_1, s_2 \in (\frac{1}{8}, \frac{1}{8})$ we have

$$|s_1 - s_2| < \tfrac{1}{4} \neq k,$$

and so we conclude that we must have $t_2 = t_1$ and so $s_2 = s_1$. This gives the desired injectivity of $\boldsymbol{\phi}$.

(ii) We directly compute

$$\|\boldsymbol{D}\boldsymbol{\phi}(s,t)\|_{\mathbb{R}^2,\mathbb{R}^2} = 2(\cosh(t+s-1)^(- 4) + 2\pi^2(1 - \tanh(t+s-1))^2).$$

Note that, for (s,t) in the domain of $\boldsymbol{\phi}$, we have

$$\tanh(t+s-1) \geq \tanh(-\tfrac{1}{8}) = -\tanh(\tfrac{1}{8}) \in (-1,0).$$

Thus

$$\|\boldsymbol{D}\boldsymbol{\phi}(s,t)\|_{\mathbb{R}^2,\mathbb{R}^2} \geq 4\pi^2(1 + \tanh(\tfrac{1}{8})^2 > 0,$$

giving this part of the lemma.

(iii) This follows from the Inverse Function Theorem, Theorem **??** below. ▼

Now we let $U = \text{image}(\phi)$, noting that U is a "thickening" of the trace from Figure 4.10. By ***missing stuff*** U is open. Next define

$$\begin{aligned} g\colon (-\tfrac{1}{8}, \tfrac{1}{2}) \times (1, \infty) &\to \mathbb{R} \\ (s, t) &\mapsto t \end{aligned}$$

and we note that clearly g is continuously differentiable with a bounded derivative. If we define $f\colon U \to \mathbb{R}$ by $f = g \circ \phi^{-1}$, then, by the Chain Rule and Proposition 4.1.16(vi), f is continuously differentiable with a bounded derivative. It remains to show that f is not uniformly continuous.

For $k \in \mathbb{Z}$ with $k \geq 2$ note that $x_k \triangleq (1 + \tanh(k-1)(1, 0) \in U$ and that $f(x_k) = k$. Let $\delta \in \mathbb{R}_{>0}$. Since $\lim_{k\to\infty}(1 + \tanh(k-1)) = 2$ let $N \in \mathbb{Z}_{>0}$ be sufficiently large that

$$|(1 + \tanh(j-1)) - (1 + \tanh(k-1))| < \delta, \qquad j, k \geq N.$$

Then let $k \in \mathbb{Z}_{>0}$ and note that

$$|f(x_{N+k}) - f(x_N)| = k.$$

Note that

$$\|x_{N+k} - x_N\|_{\mathbb{R}^2} = |(1 + \tanh(N + k - 1)) - (1 + \tanh(N - 1))| < \delta.$$

Therefore, for any $\delta \in \mathbb{R}_{>0}$, there are points $x, y \in U$ such that $\|x - y\|_{\mathbb{R}^2} < \delta$ but $|f(x) - f(y)| \geq 1$. This prohibits uniform continuity of f. •

As we showed to dramatic effect in Example 3.2.9, it is very much not the case that a continuous function is differentiable.

Next we consider the multivariable version of the Mean Value Theorem that we stated in the single-variable case as Theorem 3.2.19. The fact that the natural domain for functions in the single-variable case is an interval needs to be appropriately generalised to the multivariable case. A natural way to do this is with the notion of a convex set. We shall investigate convexity in some detail in Section **??**, so let us just recall the basic definition here. For $x_1, x_2 \in \mathbb{R}^n$ denote by $\{(1 - s)x_1 + sx_2 \mid s \in [0, 1]\}$ the line segment between x_1 and x_2. A subset $C \subseteq \mathbb{R}^n$ is ***convex*** if the line segment between any two points in C is a subset of C.

4.4.38 Theorem (Mean Value Theorem) *Let $\mathrm{C} \subseteq \mathbb{R}^n$ be an open convex set and let $\mathbf{f}\colon \mathrm{C} \to \mathbb{R}^m$ be of class C^1. If $\mathbf{x}_1, \mathbf{x}_2 \in \mathrm{C}$ then*

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\|_{\mathbb{R}^m} \leq \sup\{\|\mathbf{Df}((1 - s)\mathbf{x}_1 + s\mathbf{x}_2)\|_{\mathbb{R}^n, \mathbb{R}^m} \mid s \in [0, 1]\}\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbb{R}^n}.$$

Moreover, if $\mathbf{Df}$ is uniformly bounded, i.e., if there exists $\mathrm{M} \in \mathbb{R}_{>0}$ such that $\|\mathbf{Df}(\mathbf{x})\|_{\mathbb{R}^n, \mathbb{R}^m} \leq \mathrm{M}$ for every $\mathbf{x} \in \mathrm{C}$, then

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\|_{\mathbb{R}^m} \leq \mathrm{M}\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbb{R}^n}.$$

Proof Let $\gamma\colon [0,1] \to \mathbb{R}^n$ be defined by $\gamma(s) = (1-s)x_1 + sx_2$. Then $\operatorname{image}(\gamma) \subseteq C$ since C is convex. By the Chain Rule, Theorem 4.4.49, we have

$$D(f \circ \gamma)(s) = Df(\gamma(s)) \circ D\gamma(s).$$

Using the Fundamental Theorem of Calculus applied to the components of the map $g = f \circ \gamma\colon C \to \mathbb{R}^m$ we have

$$g(1) - g(0) = \int_0^1 Dg(s)\,\mathrm{d}s,$$

which gives

$$f(x_1) - f(x_2) = \int_0^1 Df((1-s)x_1 + sx_2) \cdot (x_2 - x_1)\,\mathrm{d}s.$$

Thus, using Proposition 4.1.16(v) and ***missing stuff***,

$$\begin{aligned}
\|f(x_1) - f(x_2)\|_{\mathbb{R}^m} &= \Big\| \int_0^1 Df((1-s)x_1 + sx_2) \cdot (x_2 - x_1)\,\mathrm{d}s \Big\|_{\mathbb{R}^m} \\
&\leq \int_0^1 \|Df((1-s)x_1 + sx_2) \cdot (x_2 - x_1)\|_{\mathbb{R}^m}\,\mathrm{d}s \\
&\leq \Big(\int_0^1 \|Df((1-s)x_1 + sx_2)\|_{\mathbb{R}^n,\mathbb{R}^m}\,\mathrm{d}s \Big) \cdot \|x_1 - x_2\|_{\mathbb{R}^m} \\
&\leq \sup\{\|Df((1-s)x_1 + sx_2)\|_{\mathbb{R}^n,\mathbb{R}^m} \mid s \in [0,1]\} \|x_1 - x_2\|_{\mathbb{R}^m},
\end{aligned}$$

as desired.

The final assertion of the theorem follows immediately from the first. ■

4.4.7 Derivatives and maxima and minima

Next we generalise to multiple-dimensions the relationships between derivatives and maxima and minima of functions. First let us define the relevant function properties.

4.4.39 Definition (Local maximum and local minimum) Let $A \subseteq \mathbb{R}^n$ and let $A\colon I \to \mathbb{R}$ be a function. A point $x_0 \in A$ is a:

(i) ***local maximum*** if there exists a neighbourhood U of x_0 such that $f(x) \leq f(x_0)$ for every $x \in U \cap A$;

(ii) ***strict local maximum*** if there exists a neighbourhood U of x_0 such that $f(x) < f(x_0)$ for every $x \in U \cap (A \setminus \{x_0\})$;

(iii) ***local minimum*** if there exists a neighbourhood U of x_0 such that $f(x) \geq f(x_0)$ for every $x \in U \cap A$;

(iv) ***strict local minimum*** if there exists a neighbourhood U of x_0 such that $f(x) > f(x_0)$ for every $x \in U \cap (A \setminus \{x_0\})$. •

To generalise the single-variable characterisation of maxima and minima given in Theorem 3.2.16 the reader will want to recall properties of symmetric bilinear maps from Section **??**.

4.4.40 Theorem (Derivatives, and maxima and minima) *If $U \subseteq \mathbb{R}^n$ is open, if $f\colon U \to \mathbb{R}$ is a function, and if $\mathbf{x}_0 \in U$ then the following statements hold:*

(i) if f is differentiable at $\mathbf{x}_0$ and if $\mathbf{x}_0$ is a local maximum or a local minimum for f, then $\mathbf{D}f(\mathbf{x}_0) = \mathbf{0}$;

(ii) if f is twice differentiable at $\mathbf{x}_0$, and if $\mathbf{x}_0$ is a local maximum (resp. local minimum) for f, then $\mathbf{D}^2f(\mathbf{x}_0)$ is negative-semidefinite (resp. positive-semidefinite);

(iii) if f is twice differentiable at $\mathbf{x}_0$, and if $\mathbf{D}f(\mathbf{x}_0) = \mathbf{0}$ and $\mathbf{D}^2f(\mathbf{x}_0)$ is negative definite (resp. positive-definite), then $\mathbf{x}_0$ is a strict local maximum (resp. strict local minimum) for f;

(iv) if f is twice differentiable at $\mathbf{x}_0$, if $\mathbf{D}f(\mathbf{x}_0) = \mathbf{0}$ and if $\mathbf{D}^2f(\mathbf{x}_0)$ is neither positive- nor negative-semidefinite, then $\mathbf{x}_0$ is neither a local minimum nor a local maximum for f.

Proof (i) We shall give the proof for the case when x_0 is a local minimum; the case of a local maximum is similar. Let $v \in \mathbb{R}^n$. Since x_0 a local minimum we have

$$f(x_0 + sv) - f(x_0) \geq 0$$

for all s sufficiently near 0. Thus

$$\tfrac{1}{s}(f(x_0 + sv) - f(x_0)) \geq 0$$

for $s \in \mathbb{R}_{\geq 0}$ and so, by Proposition 4.4.12,

$$Df(x_0) \cdot v = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} f(x_0 + sv) = \lim_{s \downarrow 0} \tfrac{1}{s}(f(x_0 + sv) - f(x_0)) \geq 0.$$

Similarly, since

$$\tfrac{1}{s}(f(x_0 + sv) - f(x_0)) \leq 0$$

for $s \in \mathbb{R}_{\leq 0}$ we have $Df(x_0) \cdot v \leq 0$ and so we conclude that $Df(x_0) \cdot v = 0$. Since this holds for any $v \in \mathbb{R}^n$ we must have $Df(x_0) = 0$.

(ii) We prove the result for the case when x_0 is a local minimum; the case of a local maximum is proved similarly. By the multivariable Taylor Theorem, ***missing stuff***, and noting the definition of the Landau symbol from ***missing stuff***, we have

$$0 \leq f(x_0 + sv) - f(x_0) = \tfrac{1}{2}s^2 D^2 f(x_0) \cdot (v, v) + o((sv)^2)$$

for every $v \in \mathbb{R}^n$ and for s sufficiently near 0. Therefore,

$$\begin{aligned} & D^2 f(x_0) \cdot (v, v) + \tfrac{2}{s^2} o((sv)^2) \geq 0 \\ \implies \quad & D^2 f(x_0) \cdot (v, v) + \lim_{s \to 0} \tfrac{2}{s^2} o((sv)^2) \geq 0 \\ \implies \quad & D^2 f(x_0) \cdot (v, v) \geq 0, \end{aligned}$$

giving $D^2 f(x_0)$ as positive-semidefinite, as desired.

(iii) We first prove a lemma.

1 Lemma *If* $\mathsf{B} \in \mathrm{S}^2(\mathbb{R}^n; \mathbb{R})$ *is positive-definite then there exists* $\mathrm{m}, \mathrm{M} \in \mathbb{R}_{>0}$ *such that*

$$\mathrm{m}\|\mathbf{v}\|^2_{\mathbb{R}^n} \leq \mathsf{B}(\mathbf{v}, \mathbf{v}) \leq \mathrm{M}\|\mathbf{v}\|^2_{\mathbb{R}^n}$$

for every $\mathbf{v} \in \mathbb{R}^n$.

Proof Define $\boldsymbol{B} \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ by $\boldsymbol{B}(i, j) = \mathsf{B}(\boldsymbol{e}_i, \boldsymbol{e}_j)$ so that

$$\mathsf{B}(v, v) = \sum_{i,j=1}^{n} \boldsymbol{B}(i, j)v(i)v(j).$$

Then $\boldsymbol{B}^T = \boldsymbol{B}$ and

$$\sum_{i,j=1}^{n} \boldsymbol{B}(i, j)v(i)v(j) > 0$$

for every $v \in \mathbb{R}^n$, cf. the proof of Theorem **??**. By ***missing stuff*** there exists an orthogonal matrix $\boldsymbol{R} \in \mathsf{O}(n)$ such that

$$\boldsymbol{B} = \boldsymbol{R}^T \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix} \boldsymbol{R}$$

for $d_1, \ldots, d_n \in \mathbb{R}_{>0}$. Therefore, for any $v \in \mathbb{R}^n$, we have

$$\sum_{i,j=1}^{n} \boldsymbol{B}(i, j)v(i)v(j) = \sum_{j=1}^{n} d_j(\boldsymbol{R}v)(j)^2 = \sum_{j=1}^{n} d_j v(j)^2.$$

Therefore, we directly have

$$\min\{d_1, \ldots, d_n\}\|v\|^2_{\mathbb{R}^n} \leq \mathsf{B}(v, v) \leq \max\{d_1, \ldots, d_n\}\|v\|^2_{\mathbb{R}^n}$$

for every $v \in \mathbb{R}^n$, giving the result. ▼

We now prove this part of the theorem for the case when $\boldsymbol{D}^2 f(\boldsymbol{x}_0)$ is positive-definite; the case when it is negative-definite follows in the same manner with a suitable trivial modification to the signs of m and M in the lemma above.

From the lemma there exists $m \in \mathbb{R}_{>0}$ such that $\boldsymbol{D}^2 f(\boldsymbol{x}_0) \cdot (v, v) \geq m\|v\|^2_{\mathbb{R}^n}$ for every $v \in \mathbb{R}^n$. Therefore, by the multivariable Taylor Theorem, ***missing stuff***, we have

$$f(\boldsymbol{x}_0 + v) - f(\boldsymbol{x}_0) = \tfrac{1}{2}\boldsymbol{D}^2 f(\boldsymbol{x}_0) \cdot (v, v) + o(v^2) \geq \tfrac{1}{2}m\|v\|^2_{\mathbb{R}^n} + o(v^2),$$

for v sufficiently small in norm that $\boldsymbol{x}_0 + v \in U$. Now choose $\epsilon \in \mathbb{R}_{>0}$ sufficiently small that $|o(v^2)| \leq \frac{1}{4}m\|v\|^2_{\mathbb{R}^n}$ for $v \in \overline{\mathsf{B}}^n(\epsilon, \mathbf{0})$. Then

$$f(\boldsymbol{x}_0 + v) - f(\boldsymbol{x}_0) \geq \tfrac{1}{4}m\|v\|^2_{\mathbb{R}^n},$$

for all $v \in \overline{\mathsf{B}}^n(\epsilon, \mathbf{0})$, giving $\boldsymbol{x}_0$ as a strict local minimum for f.

(iv) Since $\boldsymbol{D}^2 f(\boldsymbol{x}_0)$ is neither positive- nor negative-semidefinite, there exists $v_-, v_+ \in \mathbb{R}^n$ such that

$$\boldsymbol{D}^2 f(\boldsymbol{x}_0) \cdot (v_-, v_-) \in \mathbb{R}_{<0}, \quad \boldsymbol{D}^2 f(\boldsymbol{x}_0) \cdot (v+, v_+) \in \mathbb{R}_{>0}.$$

As above, write

$$f(x_0 + sv_-) - f(x_0) = \tfrac{1}{2}s^2 D^2 f(x_0) \cdot (v_-, v_-) + o((sv_-)^2).$$

for $s \in \mathbb{R}_{>0}$ be sufficiently small that $x_0 + sv_- \in U$. Further choosing s_0 sufficiently small that

$$\left|\frac{o((sv_-)^2)}{s^2}\right| < \frac{1}{4} Df(x_0) \cdot (v_-, v_-)$$

for $s \in (0, s_0]$, we have

$$\begin{aligned}\tfrac{1}{2}s^2 D^2 f(x_0) \cdot (v_-, v_-) + o((sv_-)^2) &= s^2\Big(\frac{1}{2} D^2 f(x_0) \cdot (v_-, v_-) + \frac{o((sv_-)^2)}{s^2}\Big) \\ &< \frac{s^2}{4} Df(x_0) \cdot (v_-, v_-) < 0,\end{aligned}$$

giving $f(x_0 + sv_-) < f(x_0)$ for $s \in (0, s_0]$. In a similar manner, one shows that $f(x_0 + sv_+) > f(x_0)$ for s sufficiently small. Thus x_0 is neither a local minimum nor a local minimum. ■

We refer to Example 3.2.17 for illustrations of the above theorem in the single-variable case. The same conclusions concerning the lack of converses to the theorem hold as were drawn from Example 3.2.17. It is, however, slightly insightful to give a few additional examples in multiple-variables.

4.4.41 Examples (Derivatives, and maxima and minima)

1. We define $f_\alpha \colon \mathbb{R}^2 \to \mathbb{R}$ by $f_\alpha(x_1, x_2) = x_1^2 + \alpha x_2^2$ for $\alpha \in \mathbb{R}$. We see that $(0, 0)$ is a local minimum (resp. strict local minimum) when $\alpha \in \mathbb{R}_{\geq 0}$ (resp. $\alpha \in \mathbb{R}_{>0}$). When $\alpha \in \mathbb{R}_{<0}$ we have that $(0, 0)$ is neither a local minimum nor a local maximum. We compute

$$Df_\alpha(0, 0) = 0, \quad D^2 f_\alpha(0, 0) \cdot ((v_1, v_2), (v_1, v_2)) = 2v_1^2 + 2\alpha v_2^2.$$

Thus $D^2 f_\alpha$ is positive-semidefinite when $\alpha = 0$, positive-definite when $\alpha \in \mathbb{R}_{>0}$, and indefinite when $\alpha \in \mathbb{R}_{<0}$. From Theorem 4.4.40 we see that $(0, 0)$ is a strict local minimum for f_α when $\alpha \in \mathbb{R}_{>0}$. When $\alpha \in \mathbb{R}_{\leq 0}$ we can only conclude that $(0, 0)$ is not a local minimum for f_α.

2. We take $f_\alpha \colon \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x_1, x_2) = x_1^2 + \alpha x_2^2$ for $\alpha \in \mathbb{R}$. When $\alpha \in \mathbb{R}_{>0}$ we see that $(0, 0)$ is a strict local minimum for f_α and that when $\alpha \in \mathbb{R}_{\geq 0}$ we have $(0, 0)$ as a (not strict) local minimum. When $\alpha \in \mathbb{R}_{<0}$, $(0, 0)$ is neither a local minimum nor a local maximum. We compute

$$Df_\alpha(0, 0) = 0, \quad D^2 f_\alpha(0, 0) \cdot ((v_1, v_2), (v_1, v_2)) = 2v_1^2.$$

Thus $D^2 f_\alpha(0, 0)$ is positive-semidefinite for every α. By Theorem 4.4.40 we can conclude that $(0, 0)$ cannot be a local minimum for f_α for every α, and this is indeed the case. However, the conclusion that $(0, 0)$ is a strict local minimum of f_α for $\alpha \in \mathbb{R}_{>0}$ cannot be deduced from Theorem 4.4.40.

3. Finally, we take $f_\alpha\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $f_\alpha(x_1, x_2) = x_1^4 + \alpha x_2^4$. We see that when $\alpha \in \mathbb{R}_{>0}$ (resp. $\alpha \in \mathbb{R}_{\geq 0}$), $(0,0)$ is a strict local minimum (resp. local minimum) for f_α. For $\alpha \in \mathbb{R}_{<0}$ we have that $(0,0)$ is neither a local minimum nor a local maximum. Moreover, we compute $D^2 f(0,0) \cdot ((v_1, v_2), (v_1, v_2)) = 0$ and so $D^2 f(0,0)$ is both positive- and negative-semidefinite. No conclusions can be drawn using Theorem 4.4.40 to determine whether $(0,0)$ is a local maximum or minimum. •

4.4.8 Derivatives and constrained extrema

Let us next consider an important modification of the problem of finding minima and maxima, that where constraints are added to the mix. We wish to allow equality and inequality constraints, so let us set this up properly. Given $x, y \in \mathbb{R}^n$, let us write $x \leq y$ when $x_j \leq y_j$ for each $j \in \{1, \dots, n\}$. With this convention, we make the following definition.

4.4.42 Definition (Equality and inequality constraints) Let $A \subseteq \mathbb{R}^n$ and let $g\colon A \to \mathbb{R}^m$. A point $x \in A$ satisfies the ***equality constraint*** defined by g if $g(x) = 0$ and satisfies the ***inequality constraint*** defined by g if $g(x) \leq 0$. •

Thus, with the notation of the definition, the set of points in A satisfying the equality constraint is $g^{-1}(0)$ and the set of points satisfying the inequality constraint defined by g is $g^{-1}(\mathbb{R}^m_{\leq 0})$. We can now define the sorts of minima and maxima in which we are interested.

4.4.43 Definition (Constrained local maximum and minimum) Let $A \subseteq \mathbb{R}^n$ and consider maps $f\colon A \to \mathbb{R}$ and $g\colon A \to \mathbb{R}^m$ and $h\colon A \to \mathbb{R}^k$. A point $x_0 \in g^{-1}(0)$ is

(i) ***local maximum*** of the triple (f, g, h) if there exists a relative neighbourhood U of x_0 in A such that $f(x) \leq f(x_0)$ for every $x \in g^{-1}(0) \cap h^{-1}(\mathbb{R}^k_{\leq 0}) \cap U$;

(ii) ***strict local maximum*** of the triple (f, g, h) if there exists a relative neighbourhood U of x_0 in A such that $f(x) < f(x_0)$ for every $x \in g^{-1}(0) \cap h^{-1}(\mathbb{R}^k_{\leq 0}) \cap (U \setminus \{x_0\})$;

(iii) ***local minimum*** of the triple (f, g, h) if there exists a relative neighbourhood U of x_0 in A such that $f(x) \geq f(x_0)$ for every $x \in g^{-1}(0) \cap h^{-1}(\mathbb{R}^k_{\leq 0}) \cap U$;

(iv) ***strict local minimum*** of the triple (f, g, h) if there exists a relative neighbourhood U of x_0 in A such that $f(x) > f(x_0)$ for every $x \in g^{-1}(0) \cap h^{-1}(\mathbb{R}^k_{\leq 0}) \cap (U \setminus \{x_0\})$.

If there are no inequality constraints, we shall say that x_0 is a local maximum (etc.) of (f, g) with equality constraints. If there are no inequality constraints, we shall say that x_0 is a local maximum (etc.) of (f, h) with inequality constraints. •

The following theorem gives conditions for minimising (f, g, h) under hypotheses of differentiability.

4.4.44 Theorem (Lagrange Multiplier Rule) *Let* $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}}$ *be open, and let* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}$, $\mathbf{g}\colon \mathrm{U} \to \mathbb{R}^{\mathrm{m}}$, *and* $\mathbf{h}\colon \mathrm{U} \to \mathbb{R}^{\mathrm{k}}$ *be continuously differentiable. For* $\lambda_0 \in \mathbb{R}$, $\boldsymbol{\lambda} \in \mathbb{R}^{\mathrm{m}}$, *and* $\boldsymbol{\mu} \in \mathbb{R}^{\mathrm{k}}$, *define*

$$\begin{aligned} \mathrm{f}_{\lambda_0, \boldsymbol{\lambda}, \boldsymbol{\mu}}\colon \mathrm{U} &\to \mathbb{R} \\ \mathbf{x} &\mapsto \lambda_0 \mathrm{f}(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{g}(\mathbf{x}) \rangle_{\mathbb{R}^{\mathrm{m}}} + \langle \boldsymbol{\mu}, \mathbf{h}(\mathbf{x}) \rangle_{\mathbb{R}^{\mathrm{k}}}. \end{aligned}$$

If $\mathbf{x}_0$ is a local minimum of $(\mathrm{f}, \mathbf{g}, \mathbf{h})$, then there exist $\lambda_0 \in \mathbb{R}$, $\boldsymbol{\lambda} \in (\mathbb{R}^{\mathrm{m}})^*$, and $\boldsymbol{\mu} \in \mathbb{R}^{\mathrm{k}}$, not simultaneously zero, such that $\mathbf{D}\mathrm{f}_{\lambda_0,\boldsymbol{\lambda},\boldsymbol{\mu}}(\mathbf{x}_0) = \mathbf{0}$. Furthermore, the following statements hold:

(i) $\lambda_0 \in \mathbb{R}_{\geq 0}$ *and* $\boldsymbol{\mu} \geq \mathbf{0}$*;*

(ii) if, for $\mathrm{r} \in \{1, \dots, \mathrm{k}\}$, $\mathrm{h}_\mathrm{r}(\mathbf{x}_0) < 0$, *then* $\mu_\mathrm{r} = 0$*;*

(iii) if the vectors satisfy the ***Kuhn–Tucker condition****, namely that*

$$\{\mathbf{D}\mathrm{g}_1(\mathbf{x}_0), \dots, \mathbf{D}\mathrm{g}_\mathrm{m}(\mathbf{x}_0)\} \cup \{\mathbf{D}\mathrm{h}_\mathrm{r}(\mathbf{x}_0) \mid \mathrm{h}_\mathrm{r}(\mathbf{x}_0)\}$$

are linearly independent, then λ_0 can be taken to be 1*.*

Proof We assume, without loss of generality, that $\boldsymbol{x}_0 = \mathbf{0}$, that $f(\mathbf{0}) = 0$, and that

$$h_1(\mathbf{0}) = \cdots = h_s(\mathbf{0}) = 0,\ h_{s+1}(\mathbf{0}), \dots, h_k(\mathbf{0}) \in \mathbb{R}_{<0}.$$

It will be convenient to denote $a_+ = \max\{0, a\}$ for $a \in \mathbb{R}$.

Suppose that $\bar{\epsilon} \in \mathbb{R}_{>0}$ is such that $\overline{\mathsf{B}}^n(\bar{\epsilon}, \mathbf{0}) \subseteq U$ and such that $h_r(\boldsymbol{x}) < 0$ for every $\boldsymbol{x} \in \overline{\mathsf{B}}(\bar{\epsilon}, \mathbf{0})$, $r \in \{s+1, \dots, k\}$, the latter being possible since $\boldsymbol{h}$ is continuous. We prove a lemma.

1 Lemma *If $\epsilon \in (0, \bar{\epsilon}]$, then there exists $\mathrm{M} \in \mathbb{R}_{>0}$ such that*

$$\mathrm{f}(\mathbf{x}) + \|\mathbf{x}\|^2_{\mathbb{R}^\mathrm{n}} + \mathrm{M}\Big(\sum_{\mathrm{a}=1}^{\mathrm{m}} \mathrm{g}_\mathrm{a}(\mathbf{x})^2 + \sum_{\mathrm{r}=1}^{\mathrm{s}} (\mathrm{h}_\mathrm{r}(\mathbf{x})_+)^2\Big) \in \mathbb{R}_{>0}$$

for all $\mathbf{x}$ such that $\|\mathbf{x}\|_{\mathbb{R}^\mathrm{n}} = \epsilon$.

Proof Suppose the conclusions of the lemma do not hold. Then there exists a sequence $(M_j)_{j \in \mathbb{Z}_{>0}}$ in $\mathbb{R}_{>0}$ and a sequence $(\boldsymbol{x}_j)_{j \in \mathbb{Z}_{>0}}$ such that (1) $\lim_{j\to\infty} M_j = \infty$, (2) $\|\boldsymbol{x}_j\|_{\mathbb{R}^n} = \epsilon$ for each $j \in \mathbb{Z}_{>0}$, and (3)

$$f(\boldsymbol{x}_j) + \|\boldsymbol{x}_j\|^2_{\mathbb{R}^n} \leq -M_j\Big(\sum_{a=1}^{m} g_a(\boldsymbol{x}_j)^2 + \sum_{r=1}^{s} (h_r(\boldsymbol{x}_j)_+)^2\Big) \tag{4.27}$$

for each $j \in \mathbb{Z}_{>0}$. Note that the set of points

$$\{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_{\mathbb{R}^n} = \epsilon\}$$

is closed and bounded, and so compact. By the Bolzano–Weierstrass Theorem, we can assume that the sequence $(\boldsymbol{x}_j)_{j \in \mathbb{Z}_{>0}}$ converges to $\bar{\boldsymbol{x}}$ such that $\|\bar{\boldsymbol{x}}\|_{\mathbb{R}^n} = \epsilon$. Since $\boldsymbol{g}$ is continuous and since the function $\boldsymbol{x} \mapsto h_r(\boldsymbol{x})_+$ is continuous, we have

$$\sum_{a=1}^{m} g_a(\bar{\boldsymbol{x}})^2 + \sum_{r=1}^{s} (h_r(\bar{\boldsymbol{x}})_+)^2 = \lim_{j\to\infty}\Big(\sum_{a=1}^{m} g_a(\boldsymbol{x}_j)^2 + \sum_{r=1}^{s} (h_r(\boldsymbol{x}_j)_+)^2\Big) = 0.$$

Thus $\boldsymbol{g}(\bar{\boldsymbol{x}}) = \mathbf{0}$ and $h_r(\bar{\boldsymbol{x}}) = 0$, $r \in \{1, \dots, s\}$. Then $\bar{\boldsymbol{x}}$ satisfies the equality constraints defined by $\boldsymbol{g}$ and the inequality constraints defined by $\boldsymbol{h}$. As such, since $\mathbf{0}$ is a local minimum of $(f, \boldsymbol{g}, \boldsymbol{h})$, $f(\bar{\boldsymbol{x}}) \geq f(\mathbf{0}) = 0$. However, by (4.27), $f(\boldsymbol{x}_j) \leq -\epsilon^0$ for each $j \in \mathbb{Z}_{>0}$, and so, by continuity of f,

$$f(\bar{\boldsymbol{x}}) = \lim_{j\to\infty} f(\boldsymbol{x}_j) \leq -\epsilon^2,$$

giving a contradiction. ▼

Now another lemma.

2 Lemma *If $\epsilon \in (0, \bar{\epsilon}]$, then there exists $\bar{\mathbf{x}} \in \mathsf{B}(\epsilon, \mathbf{0})$, $\lambda_0 \in \mathbb{R}$, $\boldsymbol{\lambda} \in \mathbb{R}^{\mathrm{m}}$, and $\boldsymbol{\mu} \in \mathbb{R}^{\mathrm{k}}$ such that*

(i) $\lambda_0, \mu_1, \ldots, \mu_{\mathrm{s}} \in \mathbb{R}_{\geq 0}$,

(ii) $\mu_{\mathrm{s}+1} = \cdots = \mu_{\mathrm{k}} = 0$,

(iii) $\|(\lambda_0, \lambda_1, \ldots, \lambda_{\mathrm{m}}, \mu_1, \ldots, \mu_{\mathrm{k}})\|_{\mathbb{R}^{\mathrm{m}+\mathrm{k}+1}} = 1$, *and*

(iv) for each $\mathrm{j} \in \{1, \ldots, \mathrm{n}\}$,

$$\lambda_0(\mathbf{D}_2\mathrm{f}(\bar{\mathbf{x}}) + 2\bar{\mathbf{x}}) + \sum_{\mathrm{a}=1}^{\mathrm{m}} \lambda_{\mathrm{a}} \mathbf{D}_{\mathrm{j}} \mathrm{g}_{\mathrm{a}}(\bar{\mathbf{x}}) + \sum_{\mathrm{r}=1}^{\mathrm{s}} \mu_{\mathrm{r}} \mathbf{D}_{\mathrm{j}} \mathrm{h}_{\mathrm{r}}(\bar{\mathbf{x}}) = 0.$$

Proof Let M be as in Lemma 1. Define

$$F(\boldsymbol{x}) = f(\boldsymbol{x}) + \|\boldsymbol{x}\|^2_{\mathbb{R}^n} + M\Big(\sum_{a=1}^{m} g_a(\boldsymbol{x})^2 + \sum_{r=1}^{s} (h_r(\boldsymbol{x})_+)^2\Big)$$

for $\boldsymbol{x} \in U$. Since $\overline{\mathsf{B}}^n(\epsilon, \mathbf{0})$ is compact and F is continuous, by Theorem 4.3.32 there exists $\bar{\boldsymbol{x}} \in \overline{\mathsf{B}}^n(\epsilon, \mathbf{0})$ such that

$$F(\bar{\boldsymbol{x}}) = \inf\{F(\boldsymbol{x}) \mid \boldsymbol{x} \in \overline{\mathsf{B}}^n(\epsilon, \mathbf{0})\}.$$

In particular, $F(\bar{\boldsymbol{x}}) \leq F(\mathbf{0}) = 0$. Thus, by the definition of M from Lemma 1, $\|\bar{\boldsymbol{x}}\|_{\mathbb{R}^n} \neq \epsilon$. By Theorem 4.4.40 it follows that $\boldsymbol{D}F(\bar{\boldsymbol{x}}) = \mathbf{0}$ since $\bar{\boldsymbol{x}}$ is a local minimum for $F|\mathsf{B}^n(\epsilon, \mathbf{0})$. Note that the function $x \mapsto (x_+)^2$ is continuously differentiable. Therefore, by the Chain Rule, the function $\boldsymbol{x} \mapsto (h_r(\boldsymbol{x})_+)^2$ is continuously differentiable for each $r \in \{1, \ldots, s\}$. Moreover, also by the Chain Rule, its j partial derivative is given by

$$\boldsymbol{x} \mapsto 2h_s(\boldsymbol{x})_+ \boldsymbol{D}_j h_r(\boldsymbol{x}).$$

Thus an elementary computation gives

$$0 = \boldsymbol{D}_j F(\bar{\boldsymbol{x}}) = \boldsymbol{D}_j f(\bar{\boldsymbol{x}}) + 2x_j + \sum_{a=1}^{m} 2Mg_a(\bar{\boldsymbol{x}}) \boldsymbol{D}_j g_a(\bar{\boldsymbol{x}}) + \sum_{r=1}^{s} 2Mh_s(\boldsymbol{x})_+ \boldsymbol{D}_j h_r(\boldsymbol{x}).$$

Now define

$$\lambda_0' = 1, \quad \lambda_a' = 2Mg_a(\bar{\boldsymbol{x}}),\ a \in \{1, \ldots, m\}, \quad \mu_r' = 2Mh_r(\bar{\boldsymbol{x}})_+,\ r \in \{1, \ldots, s\}, \quad \mu_{s+1}' = \cdots = \mu_k' = 0.$$

Then let $\ell = \|(\lambda_0', \lambda_1', \ldots, \lambda_m', \mu_1', \ldots, \mu_k')\|_{\mathbb{R}^{m+k+1}}$ and define $\lambda_a = \ell^{-1}\lambda_a'$, $a \in \{0, 1, \ldots, m\}$, and $\mu_r = \ell^{-1}\mu_r'$, $r \in \{1, \ldots, k\}$. One easily sees that these definitions satisfy the conclusions of the lemma. ▼

Now let $(\epsilon_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $(0, \bar{\epsilon}]$ converging to 0. For each $j \in \mathbb{Z}_{>0}$, let $\bar{\boldsymbol{x}}_j \in \mathsf{B}^n(\epsilon_j, \mathbf{0})$, $\lambda_{0,j} \in \mathbb{R}_{\geq 0}$, $\boldsymbol{\lambda}_j \in \mathbb{R}^m$, and $\boldsymbol{\mu}_j \in \mathbb{R}^k$ satisfy the conclusions of Lemma 2 for ϵ_j. Then, since $\lim_{j\to\infty} \bar{\boldsymbol{x}}_j = \mathbf{0}$,

$$0 = \lim_{j\to\infty} \Big(\lambda_0(\boldsymbol{D}_2 f(\bar{\boldsymbol{x}}_j) + 2\bar{\boldsymbol{x}}_j) + \sum_{a=1}^{m} \lambda_a \boldsymbol{D}_j g_a(\bar{\boldsymbol{x}}_j) + \sum_{r=1}^{s} \mu_r \boldsymbol{D}_j h_r(\bar{\boldsymbol{x}}_j)\Big)$$
$$= \lambda_0 \boldsymbol{D}_2 f(\mathbf{0}) + \sum_{a=1}^{m} \lambda_a \boldsymbol{D}_j g_a(\mathbf{0}) + \sum_{r=1}^{s} \mu_r \boldsymbol{D}_j h_r(\mathbf{0}).$$

This gives the conclusions of the theorem, with the exception of the final assertion.

For the final assertion, if $\lambda_0 = 0$ then the condition $\boldsymbol{D}f_{\lambda_0,\boldsymbol{\lambda},\boldsymbol{\mu}}(\mathbf{0}) = 0$ with $\lambda = 0$ ensures that the set

$$\{\boldsymbol{D}g_1(\mathbf{0}), \dots, \boldsymbol{D}g_m(\mathbf{0}), \boldsymbol{D}h_1(\mathbf{0}), \dots, \boldsymbol{D}h_s(\mathbf{0})\}$$

is linearly dependent. As $\lambda_0 \in \mathbb{R}_{>0}$ we can define $\lambda_0' = 1$, $\lambda_a' = \lambda_0^{-1}\lambda_a$, $a \in \{1, \dots, m\}$, and $\mu_r' = \lambda_0^{-1}\mu_r$, $r \in \{1, \dots, k\}$, and the resulting λ_0', $\boldsymbol{\lambda}'$, and $\boldsymbol{\mu}'$ will satisfy the conclusions of the theorem with $\lambda_0 = 1$. ■

Many presentations of the Lagrange Multiplier Rule will omit the rôle of the constant λ_0, assuming it to be equal to 1. However, this is only valid when the condition (iii) of the theorem is satisfied, as the following example shows.

4.4.45 Example (Constrained extrema when the constraints are not linearly independent) We take $n = 2$, and define $f\colon \mathbb{R}^2 \to \mathbb{R}$ by $f(x_1, x_2) = x_1$ $g\colon \mathbb{R}^2 \to \mathbb{R}$ by $g(x_1, x_2) = x_1^2 + x_2^2$. We do not consider inequality constraints. Note that the only point satisfying the equality constraint defined by g is $(0, 0)$. Thus there is only one choice for a local minimum of (f, g) and so the solution of the problem is trivial. However, it is not possible to satisfy the conclusions of Theorem 4.4.44 for this solution unless $\lambda_0 = 0$. Indeed, if $\lambda_0 = 0$ then Theorem 4.4.44 tells us that there exists $\lambda_1, \lambda_2 \in \mathbb{R}$, not both zero, such that

$$\boldsymbol{D}_j f(0,0) + \lambda_1 \boldsymbol{D}_j g_1(0,0) + \lambda_2 \boldsymbol{D}_j g_2(0,0) = 0, \qquad j \in \{1, 2\}.$$

Thus gives $1 = 0$, which is rather absurd. However, the conclusions of Theorem 4.4.44 *are* satisfied for arbitrary $\lambda_1, \lambda_2 \in \mathbb{R}$ if we take $\lambda_0 = 0$. •

The preceding result gives necessary conditions for a point $\boldsymbol{x}_0$ to be a local minimum for $(f, \boldsymbol{g}, \boldsymbol{h})$. Let us now consider sufficient conditions involving the second derivative.

To conveniently state the theorem, we introduce some notation. If $\boldsymbol{\lambda} \in \mathbb{R}^m$ then we denote $\boldsymbol{f}_\lambda\colon U \to \mathbb{R}$ the function given by

$$f_\lambda(\boldsymbol{x}) = f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{g}(\boldsymbol{x}) \rangle_{\mathbb{R}^m}.$$

Let $Q_\lambda(\boldsymbol{x})$ denote the restriction of the symmetric bilinear map $\boldsymbol{D}^2 f_\lambda(\boldsymbol{x})$ to the subspace $\ker(\boldsymbol{D}\boldsymbol{g}(\boldsymbol{x}))$. With this notation, we have the following theorem.

4.4.46 Theorem (Second-derivative tests for constrained minima) *Let* $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}}$ *be open, and let* $\mathrm{f}\colon \mathrm{U} \to \mathbb{R}$ *and* $\mathbf{g}\colon \mathrm{U} \to \mathbb{R}^{\mathrm{m}}$ *be twice continuously differentiable. For* $\mathbf{x}_0 \in \mathrm{U}$*, assume that* $\mathbf{Dg}(\mathbf{x}_0)$ *has rank* m *and that there exists* $\boldsymbol{\lambda} \in \mathbb{R}^{\mathrm{m}}$ *such that* $\mathbf{D}\mathrm{f}_\lambda(\mathbf{x}_0) = \mathbf{0}$*. Then the following statements hold:*

(i) if $\mathbf{x}_0$ *is a local maximum (resp. local minimum) for* $(\mathrm{f}, \mathbf{g})$*, then* $\mathrm{Q}_\lambda(\mathbf{x}_0)$ *is negative-semidefinite (resp. positive-semidefinite);*

(ii) if $\mathrm{Q}_\lambda(\mathbf{x}_0)$ *is negative definite (resp. positive-definite), then* $\mathbf{x}_0$ *is a strict local maximum (resp. strict local minimum) for* f*;*

(iii) if $Q_\lambda(\mathbf{x}_0)$ *is neither positive- nor negative-semidefinite, then* $\mathbf{x}_0$ *is neither a local minimum nor a local maximum.*

Proof Let

$$S = \{v \in \ker(\boldsymbol{D}\boldsymbol{g}(\boldsymbol{x}_0)) \mid \|v\|_{\mathbb{R}^n} = 1\}.$$

The following lemma, relying on the Implicit Function Theorem stated below as Theorem **??**, is key to our proof.

1 Lemma *If* $\mathbf{v} \in \mathrm{S}$ *there exists* $\delta \in \mathbb{R}_{>0}$ *and a continuously differentiable curve* $\gamma\colon [-\delta, \delta] \to \mathbf{g}^{-1}(\mathbf{0})$ *such that* $\gamma(0) = \mathbf{x}_0$ *and* $\mathbf{D}\gamma(0) = \mathbf{v}$.

Proof For $\sigma \in \mathfrak{S}_n$ let $\mathsf{L}_\sigma\colon \mathbb{R}^n \to \mathbb{R}^n$ be defined by

$$\mathsf{L}_\sigma(x_1, \dots, x_n) = (x_{\sigma(1)}, \dots, x_{\sigma(n)}).$$

Note that

$$\boldsymbol{D}(\boldsymbol{g} \circ \mathsf{L}_\sigma)(\mathsf{L}_\sigma^{-1}(\boldsymbol{x}_0)) = \boldsymbol{D}\boldsymbol{g}(\boldsymbol{x}_0) \circ \mathsf{L}_\sigma.$$

Let $\sigma \in \mathfrak{S}_n$ be such that the matrix

$$\begin{bmatrix} \boldsymbol{D}_{\sigma(1)}g_1(\boldsymbol{x}_0) & \cdots & \boldsymbol{D}_{\sigma(m)}g_1(\boldsymbol{x}_0) \\ \vdots & \ddots & \vdots \\ \boldsymbol{D}_{\sigma(1)}g_m(\boldsymbol{x}_0) & \cdots & \boldsymbol{D}_{\sigma(m)}g_m(\boldsymbol{x}_0) \end{bmatrix}$$

is invertible. Such a σ exists since $\boldsymbol{D}\boldsymbol{g}(\boldsymbol{x}_0)$ has rank m, and so has m linearly independent columns. The permutation σ is chosen to shift these columns to be leftmost. Let $U' \subseteq \mathbb{R}^m$ and $V' \subseteq \mathbb{R}^{n-m}$ be open sets such that $\boldsymbol{x}_0 \in \mathsf{L}_\sigma(U' \times V') \subseteq U$, making the obvious identification of $\mathbb{R}^n$ with $\mathbb{R}^m \times \mathbb{R}^{n-m}$. Now note that the map $\boldsymbol{g} \circ \mathsf{L}_\sigma\colon U' \times V' \to \mathbb{R}^m$ satisfies the hypotheses of the Implicit Function Theorem at $\mathsf{L}_\sigma^{-1}(\boldsymbol{x}_0)$, and so, after shrinking V' if necessary, there exists a continuously differentiable map $\boldsymbol{h}\colon V' \to U'$ such that

$$(\boldsymbol{h}(\boldsymbol{y}), \boldsymbol{y}) = (\boldsymbol{g} \circ \mathsf{L}_\sigma)^{-1}(\mathbf{0}) \cap U' \times V'.$$

Moreover, also by the Implicit Function Theorem,***missing stuff***

$$\ker(\boldsymbol{D}(\boldsymbol{g} \circ \mathsf{L}_\sigma)(\mathsf{L}_\sigma^{-1}(\boldsymbol{x}_0))) = \{(\boldsymbol{D}\boldsymbol{h}(\mathbf{0}) \cdot \boldsymbol{u}, \boldsymbol{u}) \mid \boldsymbol{u} \in \mathbb{R}^{n-m}\}.$$

Let $\boldsymbol{y}_0 \in V'$ be such that $\boldsymbol{x}_0 = \mathsf{L}_\sigma(\boldsymbol{h}(\boldsymbol{y}_0), \boldsymbol{y}_0)$. Note that

$$\mathsf{L}_\sigma^{-1}(v) \in \ker(\boldsymbol{D}(\boldsymbol{g} \circ \mathsf{L}_\sigma)(\mathsf{L}_\sigma^{-1}(\boldsymbol{x}_0)))$$

and thus there exists $\boldsymbol{u} \in \mathbb{R}^{n-m}$ such that $(\boldsymbol{D}\boldsymbol{h}(\boldsymbol{y}_0) \cdot \boldsymbol{u}, \boldsymbol{u}) = \mathsf{L}_\sigma^{-1}(v)$. The curve

$$\gamma'(s) = (\boldsymbol{h}(\boldsymbol{y}_0 + s\boldsymbol{u}), \boldsymbol{y}_0 + s\boldsymbol{u}),$$

defined for s sufficiently small, satisfies $\boldsymbol{D}\gamma'(0) = \mathsf{L}_\sigma^{-1}(v)$. Therefore, the curve

$$\gamma(s) = \mathsf{L}_\sigma \circ \gamma'(s)$$

satisfies $\boldsymbol{D}\gamma(0) = v$. Moreover,

$$\boldsymbol{g}(\gamma(s)) = \boldsymbol{g} \circ \mathsf{L}_\sigma(\gamma'(s)) = \mathbf{0}$$

by definition of γ', and so we get the lemma. ▼

With the lemma at hand, the remainder of the proof is more or less straightforward, following the proofs of parts (ii), (ii), and (iv) of Theorem 4.4.40. Moreover, we shall only prove the statements corresponding to local maxima, as the statements for local minima follow using the same ideas.

(i) Suppose that $Q_\lambda(x_0)$ is not positive-semidefinite. Then there exists $v \in S$ such that $Q_\lambda(x_0) \cdot (v, v) < 0$. By the lemma, let γ be a curve in $g^{-1}(\mathbf{0})$ such that $\gamma(0) = x_0$ and $D\gamma(0) = v$. Following the ideas in Theorem 4.4.40, write

$$f_\lambda(\gamma(s)) - f_\lambda(x_0) = \tfrac{1}{2} D^2 f_\lambda(x_0) + o(s^2).$$

Let $s_0 \in \mathbb{R}_{>0}$ be sufficiently small that

$$|o(s^2)| < \frac{1}{4} D^2 f_\lambda(x_0) \cdot (v, v)$$

for every $s \in (0, s_0]$. Then

$$\tfrac{1}{2} D^2 f_\lambda(x_0) + o(s^2) < 0$$

and so $f_\lambda(\gamma(s)) < f_\lambda(x_0)$ for every $s \in (0, s_0]$, showing that x_0 is not a local minimum for f_λ. Since $f_\lambda | g^{-1}(\mathbf{0}) = f_\lambda | g^{-1}(\mathbf{0})$, this part of the result follows.

(ii) Suppose that $D^2 f_\lambda(x_0)$ is positive-definite. Let

$$m = \inf\{\tfrac{1}{2} D^2 f_\lambda(x_0) \cdot (v, v) \mid v \in S\},$$

noting that $m \in \mathbb{R}_{>0}$. Let

$$M = \{v \in \mathbb{R}^n \mid x_0 + v \in g^{-1}(\mathbf{0})\}.$$

Note that

$$\begin{aligned} \lim_{\substack{v \to \mathbf{0} \\ v \in M}} Dg(x_0) \cdot (\tfrac{v}{\|v\|_{\mathbb{R}^n}}) &= \lim_{\substack{v \to \mathbf{0} \\ v \in M}} \Big(Dg(x_0) \cdot (\tfrac{v}{\|v\|_{\mathbb{R}^n}}) - \frac{g(x_0 + v) - g(x_0)}{\|v\|_{\mathbb{R}^n}} \Big) \\ &= \lim_{\substack{v \to \mathbf{0} \\ v \in M}} \Big(\frac{Dg(x_0) \cdot v - g(x_0 + v) - g(x_0)}{\|v\|_{\mathbb{R}^n}} \Big) = 0. \end{aligned}$$

Thus

$$\lim_{\substack{v \to \mathbf{0} \\ v \in M}} \frac{v}{\|v\|_{\mathbb{R}^n}} \in S.$$

Thus, given $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $\|v\|_{\mathbb{R}^n} < \delta$ then there exists $u \in S$ such that $\|\frac{v}{\|v\|_{\mathbb{R}^n}} - u\|_{\mathbb{R}^n} < \epsilon$. Because the function

$$v \mapsto \tfrac{1}{2} D^2 f_\lambda(x_0) \cdot (v, v)$$

is continuous, it follows from Theorem 4.3.33 that it is uniformly continuous on the compact set

$$\{u + v \in \mathbb{R}^n \mid u \in S,\ \|v\|_{\mathbb{R}^n} \leq \epsilon\}.$$

Therefore, by choosing δ (and thus ϵ) sufficiently small, we can ensure that $\|\frac{v}{\|v\|_{\mathbb{R}^n}}\|_{\mathbb{R}^n} > \frac{1}{2} m$ for $v \in M$ such that $\|v\|_{\mathbb{R}^n} < \delta$. As in the proof of Theorem 4.4.40, write

$$f_\lambda(x_0 + v) - f_\lambda(x_0) = \tfrac{1}{2} D^2 f_\lambda(x_0) \cdot (v, v) + o(v^2).$$

By making δ smaller if necessary, we can ensure that

$$\frac{o(v^2)}{\|v\|_{\mathbb{R}^n}^2} < \frac{1}{4}m.$$

In this case, for $v \in M$,

$$\tfrac{1}{2}D^2 f_\lambda(x_0)\cdot(v,v) + o(v^2) = \|v\|_{\mathbb{R}^n}^2\Big(\tfrac{1}{2}D^2 f_\lambda(x_0)\cdot(\tfrac{v}{\|v\|_{\mathbb{R}^n}},\tfrac{v}{\|v\|_{\mathbb{R}^n}^2}) + \frac{o(v^2)}{\|v\|_{\mathbb{R}^n}^2}\Big) \geq \frac{1}{4}m\|v\|_{\mathbb{R}^n}^2 > 0.$$

This shows that $f_\lambda(x) > f_\lambda(x_0)$ for $x \in g^{-1}(\mathbf{0})$ is a neighbourhood of x_0. Since $f_\lambda|g^{-1}(\mathbf{0}) = f|g^{-1}(\mathbf{0})$, this part of the theorem follows.

(iii) The proof here follows the proof of part (iii). ∎

4.4.9 The derivative and operations on functions

In this section we give the usual results concerning how differentiation interacts with the usual function operations.

Our first result deals with algebraic operations on functions, and for this we note that if $A \subseteq \mathbb{R}^n$, if $f, g\colon A \to \mathbb{R}^m$, and if $\alpha \in \mathbb{R}$ then we define $f+g, \alpha f\colon U \to \mathbb{R}^m$ by

$$(f+g)(x) = f(x) + g(x), \quad (\alpha f)(x) = \alpha(f(x)), \qquad x \in A.$$

If, moreover, $m = 1$ and we denote the maps by $f, g\colon A \to \mathbb{R}$, then we define $fg, \frac{f}{g}\colon A \to \mathbb{R}$ by

$$(fg)(x) = f(x)g(x), \quad \Big(\frac{f}{g}\Big)(x) = \frac{f(x)}{g(x)}, \qquad x \in A.$$

With this notation we have the following result.

4.4.47 Proposition (The derivative, and addition and multiplication) *Let* $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}}$ *be open, let* $\mathbf{f}, \mathbf{g}\colon \mathrm{U} \to \mathbb{R}^{\mathrm{m}}$ *be* r *times differentiable at* $\mathbf{x}_0 \in \mathrm{U}$*, and let* $\alpha \in \mathbb{R}$*. Then the following statements hold:*

(i) $\mathbf{f}+\mathbf{g}$ *is* r *times differentiable at* $\mathbf{x}_0$ *and* $\mathbf{D}^{\mathrm{r}}(\mathbf{f}+\mathbf{g})(\mathbf{x}_0) = \mathbf{D}^{\mathrm{r}}\mathbf{f}(\mathbf{x}_0) + \mathbf{D}^{\mathrm{r}}\mathbf{g}(\mathbf{x}_0)$*;*

(ii) $\alpha\mathbf{f}$ *is* r *times differentiable at* $\mathbf{x}_0 \in \mathrm{U}$ *and* $\mathbf{D}(\alpha\mathbf{f})(\mathbf{x}_0) = \alpha\mathbf{D}\mathbf{f}(\mathbf{x}_0)$*.*

Moreover, if $\mathrm{m} = 1$ *and if* $\mathrm{f}, \mathrm{g}\colon \mathrm{U} \to \mathbb{R}$ *are differentiable at* $\mathbf{x}_0$ *then the following statements hold:*

(iii) fg *is differentiable at* $\mathbf{x}_0$ *and*

$$\mathbf{D}(\mathrm{fg})(\mathbf{x}_0) = \mathrm{g}(\mathbf{x}_0)\mathbf{D}\mathrm{f}(\mathbf{x}_0) + \mathrm{g}(\mathbf{x}_0)\mathbf{D}\mathrm{f}(\mathbf{x}_0);$$

(iv) if $\mathrm{g}(\mathbf{x}_0) \neq 0$ *then* $\frac{\mathrm{f}}{\mathrm{g}}$ *is differentiable at* $\mathbf{x}_0$ *and*

$$\mathbf{D}\Big(\frac{\mathrm{f}}{\mathrm{g}}\Big)(\mathbf{x}_0) = \frac{\mathrm{g}(\mathbf{x}_0)\mathbf{D}\mathrm{f}(\mathbf{x}_0) - \mathrm{f}(\mathbf{x}_0)\mathbf{D}\mathrm{g}(\mathbf{x}_0)}{\mathrm{g}(\mathbf{x}_0)^2}.$$

Proof (i) We shall prove the assertion for $r = 1$, the general assertion following from this case by a simple induction. We compute

$$\lim_{x\to x_0} \frac{\|(f+g)(x)-(f+g)(x_0)-(Df(x_0)+Dg(x_0))\cdot(x-x_0)\|_{\mathbb{R}^m}}{\|x-x_0\|_{\mathbb{R}^n}} = \lim_{x\to x_0} \frac{\|f(x)-f(x_0)-Df(x_0)\cdot(x-x_0)\|_{\mathbb{R}^m}}{\|x-x_0\|_{\mathbb{R}^n}} + \lim_{x\to x_0} \frac{\|g(x)-g(x_0)-Dg(x_0)\cdot(x-x_0)\|_{\mathbb{R}^m}}{\|x-x_0\|_{\mathbb{R}^n}} = 0,$$

using Proposition 4.2.6.

(ii) Again we only prove the result for $r = 1$, the general case following by induction. We again use Proposition 4.2.6 to get

$$\lim_{x\to x_0} \frac{\|(\alpha f)(x)-(\alpha f)(x_0)-(\alpha Df(x_0))\cdot(x-x_0)\|_{\mathbb{R}^m}}{\|x-x_0\|_{\mathbb{R}^n}} = \alpha\Big(\lim_{x\to x_0} \frac{\|f(x)-f(x_0)-Df(x_0)\cdot(x-x_0)\|_{\mathbb{R}^m}}{\|x-x_0\|_{\mathbb{R}^n}}\Big) = 0.$$

(iii) We shall simply show how this part of the result follows from Theorem 4.4.48. Define $\mathsf{B} \in \mathrm{L}^2(\mathbb{R};\mathbb{R})$ by $\mathsf{B}(a_1,a_2) = a_1a_2$ so that $(fg)(x) = \mathsf{B}(f(x),g(x))$, and this then immediately gives this part of the result.

(iv) Since $g(x_0) \neq 0$ and since g is continuous at x_0 by Proposition 4.4.35 there exists a neighbourhood $V \subseteq U$ of x_0 such that $g(x)$ has the same sign as $g(x_0)$ for all $x \in V$. Thus the function $\iota\colon y \mapsto \frac{1}{y}$ is differentiable on $g(V)$. If we define $h\colon V \to \mathbb{R}$ by $h(x) = \frac{1}{g(x)}$ then h is differentiable at x_0 by the Chain Rule and, moreover,

$$Dh(x_0) = D\iota(g(x_0)) \circ Dg(x_0) = -\frac{Dg(x_0)}{g(x_0)^2}.$$

The result now follows from part (iii) noting that $\frac{f}{g} = hf$. ∎

Part (iii) of the preceding result is the ***product rule***. Sometimes a more sophisticated version of this is useful, and so we state this here.

4.4.48 Theorem (Leibniz Rule) *Let* $\mathrm{U} \subseteq \mathbb{R}^n$ *be open, let* $\mathbf{f}\colon \mathrm{U} \to \mathbb{R}^r$ *and* $\mathbf{g}\colon \mathrm{U} \to \mathbb{R}^s$ *be differentiable at* $\mathbf{x}_0 \in \mathrm{U}$*, and let* $\mathsf{B} \in \mathrm{L}(\mathbb{R}^r,\mathbb{R}^s;\mathbb{R}^m)$*. If* $\mathbf{h}\colon \mathrm{U} \to \mathbb{R}^m$ *is defined by* $\mathbf{h}(\mathbf{x}) = \mathsf{B}(\mathbf{f}(\mathbf{x}),\mathbf{g}(\mathbf{x}))$ *then* $\mathbf{h}$ *is differentiable at* $\mathbf{x}_0$ *and, moreover,*

$$\mathbf{Dh}(\mathbf{x}_0)\cdot\mathbf{v} = \mathsf{B}(\mathbf{Df}(\mathbf{x}_0)\cdot\mathbf{v},\mathbf{g}(\mathbf{x}_0)) + \mathsf{B}(\mathbf{f}(\mathbf{x}_0),\mathbf{Dg}(\mathbf{x}_0)\cdot\mathbf{v})$$

for every $\mathbf{v} \in \mathbb{R}^n$.

Proof By Theorem 4.4.8 the map $\mathsf{B}\colon \mathbb{R}^r \times \mathbb{R}^s \to \mathbb{R}^m$ is differentiable and

$$D\mathsf{B}(p_0,q_0)\cdot(u,w) = \mathsf{B}(u,q_0) + \mathsf{B}(p_0,w) \tag{4.28}$$

for every $(u,w) \in \mathbb{R}^r \oplus \mathbb{R}^s$. Since $h = \mathsf{B}\circ(f\times g)$ it follows from the Chain Rule below that

$$Dh(x_0)\cdot v = D\mathsf{B}((f\times g)(x_0))\circ D(f\times g)(x_0)\cdot v.$$

By Proposition 4.4.17 we have

$$D(f \times g)(x_0) \cdot v = (Df(x_0) \cdot v, Dg(x_0) \cdot v),$$

and the result then follows from (4.28). ■

We next state the multivariable Chain Rule, this being one of the most important theorems concerning the derivative. Indeed, we have already used this result many times in this section.

4.4.49 Theorem (Chain Rule) *Let* $\mathrm{U} \subseteq \mathbb{R}^{\mathrm{n}}$ *and* $\mathrm{V} \subseteq \mathbb{R}^{\mathrm{m}}$ *be open, consider maps* $\mathbf{f}\colon \mathrm{U} \to \mathrm{V}$ *and* $\mathbf{g}\colon \mathrm{V} \to \mathbb{R}^{\mathrm{k}}$*, and let* $\mathbf{x}_0 \in \mathrm{U}$*. If* $\mathbf{f}$ *is differentiable at* $\mathbf{x}_0$ *and if* $\mathbf{g}$ *is differentiable at* $\mathbf{f}(\mathbf{x}_0)$*, then* $\mathbf{g} \circ \mathbf{f}$ *is differentiable at* $\mathbf{x}_0$ *and, moreover,*

$$\mathbf{D}(\mathbf{g} \circ \mathbf{f})(\mathbf{x}_0) = \mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{x}_0)) \circ \mathbf{D}\mathbf{f}(\mathbf{x}_0).$$

Proof Let $\epsilon \in \mathbb{R}_{>0}$.

By Proposition 4.4.35 let $\delta_1, M \in \mathbb{R}_{>0}$ be such that

$$\|f(x) - f(x_0)\|_{\mathbb{R}^m} \leq M\|x - x_0\|_{\mathbb{R}^n}$$

for $x \in \mathsf{B}(\delta_1, x_0)$. Since g is differentiable at $f(x_0)$ there exists $\eta \in \mathbb{R}_{>0}$ such that

$$\|g(y) - g \circ f(x_0) - Dg(f(x_0)) \cdot (y - f(x_0))\|_{\mathbb{R}^k} \leq \tfrac{\epsilon}{2M}\|y - f(x_0)\|_{\mathbb{R}^m}$$

for $y \in \mathsf{B}(\eta, f(x_0))$. Since f is continuous at x_0 there exists $\delta_1 \in \mathbb{R}_{>0}$ such that

$$\|f(x) - f(x_0)\|_{\mathbb{R}^m} \leq \eta$$

for $x \in \mathsf{B}(\delta_1, x_0)$. Then, letting $\delta_3 = \min\{\delta_1, \delta_2\}$, if $x \in \mathsf{B}(\delta_3, x_0)$ we have

$$\|g \circ f(x) - g \circ f(x_0) - Dg(f(x_0)) \cdot (f(x) - f(x_0))\|_{\mathbb{R}^k} \leq \tfrac{\epsilon}{2M}\|f(x) - f(x_0)\|_{\mathbb{R}^m} \leq \tfrac{\epsilon}{2}\|x - x_0\|_{\mathbb{R}^n}.$$

By differentiability of f at x_0 let $\delta_4 \in \mathbb{R}_{>0}$ be such that

$$\|f(x) - f(x_0) - Df(x_0) \cdot (x - x_0)\|_{\mathbb{R}^m} \leq \frac{\epsilon}{2\|Dg(f(x_0))\|_{\mathbb{R}^n, \mathbb{R}^m}}\|x - x_0\|_{\mathbb{R}^n}$$

for $x \in \mathsf{B}(\delta_4, x_0)$. By Proposition 4.1.16(v) we then have

$$\|Dg(f(x_0)) \cdot (f(x) - f(x_0) - Df(x_0) \cdot (x - x_0))\|_{\mathbb{R}^k} \leq \|Dg(f(x_0))\|_{\mathbb{R}^n, \mathbb{R}^m}\|f(x) - f(x_0) - Df(x_0) \cdot (x - x_0)\|_{\mathbb{R}^m} \leq \tfrac{\epsilon}{2}\|x - x_0\|_{\mathbb{R}^n}$$

for $x \in \mathsf{B}(\delta_4, x_0)$.

Now let $\delta \in \min\{\delta_3, \delta_4\}$ and note that if $x \in \mathsf{B}(\delta, x_0)$ then we have, using the triangle inequality,

$$\begin{aligned}
&\|g \circ f(x) - g \circ f(x_0) - Dg(f(x_0)) \circ Df(x_0) \cdot (x - x_0)\|_{\mathbb{R}^k} \\
&\quad \leq \|g \circ f(x) - g \circ f(x_0) - Dg(f(x_0)) \cdot (f(x) - f(x_0))\|_{\mathbb{R}^k} \\
&\qquad + \|Dg(f(x_0)) \cdot (f(x) - f(x_0) - Df(x_0) \cdot (x - x_0))\|_{\mathbb{R}^k} \\
&\quad \leq \tfrac{\epsilon}{2}\|x - x_0\|_{\mathbb{R}^n} + \tfrac{\epsilon}{2}\|x - x_0\|_{\mathbb{R}^n} = \epsilon\|x - x_0\|_{\mathbb{R}^n}.
\end{aligned}$$

This gives

$$\frac{\|g \circ f(x) - g \circ f(x_0) - Dg(f(x_0)) \circ Df(x_0) \cdot (x - x_0)\|_{\mathbb{R}^k}}{\|x - x_0\|_{\mathbb{R}^n}} < \epsilon,$$

for $x \in \mathsf{B}(\delta, x_0)$, giving differentiability of $g \circ f$ at x_0 with derivative as asserted in the theorem. ■

For completeness let us also give the higher-order versions of the Leibniz and Chain Rules. To state these results in a compact way it is convenient to borrow some of our notation concerning the symmetric group that was given preceding Proposition **??**. Let $r, r_1, \dots, r_k \in \mathbb{Z}_{\geq 0}$ have the property that $r_1 + \cdots + r_k = r$. Then we recall the subgroup $\mathfrak{S}_{r_1|\cdots|r_k}$ of $\mathfrak{S}_r$ that leaves the "slots" of length $r_1, \dots, r_k$ in $\{1, \dots, r\}$ invariant. The situation here is slightly different than that preceding the statement of Proposition **??** in that we allow some of the numbers $r_1, \dots, r_k$ to be zero. However, this amounts to the same thing since the "slots" of length zero do not contribute materially. We also denote by $\mathfrak{S}_{r_1,\dots,r_k}$ the subset of $\mathfrak{S}_r$ having the property that $\sigma \in \mathfrak{S}_{r_1,\dots,r_k}$ satisfies

$$\sigma(r_1 + \cdots + r_j + 1) < \cdots < \sigma(r_1 + \cdots + r_j + r_{j+1}), \qquad j \in \{0, 1, \dots, k-1\}.$$

Again, this notation is in slight conflict with that preceding Proposition **??** in that some of the numbers $r_1, \dots, r_k$ are allowed to be zero. With this notation we may state the following version of Leibniz' Rule, generalising to arbitrary derivatives and arbitrary multilinear maps.

4.4.50 Theorem (General Leibniz Rule) *Let* $\mathrm{U} \subseteq \mathbb{R}^n$ *be open, let* $\mathbf{f}_j \colon \mathrm{U} \to \mathbb{R}^{n_j}$, $j \in \{1, \dots, k\}$, *be* r *times differentiable at* $\mathbf{x}_0 \in \mathrm{U}$, *and let* $\mathsf{L} \in \mathrm{L}(\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_k}; \mathbb{R}^m)$. *If* $\mathbf{f} \colon \mathrm{U} \to \mathbb{R}^m$ *is defined by*

$$\mathbf{f}(\mathbf{x}) = \mathsf{L}(\mathbf{f}_1(\mathbf{x}), \dots, \mathbf{f}_k(\mathbf{x}))$$

then $\mathbf{f}$ *is* r *times differentiable at* $\mathbf{x}_0$ *and, moreover,*

$$\mathbf{D}^r \mathbf{f}(\mathbf{x}_0) \cdot (\mathbf{v}_1, \dots, \mathbf{v}_r) = \sum_{\substack{r_1,\dots,r_k \in \mathbb{Z}_{\geq 0} \\ r_1 + \cdots + r_k = r}} \sum_{\sigma \in \mathfrak{S}_{r_1,\cdots,r_k}} \mathsf{L}(\mathbf{D}^{r_1} \mathbf{f}_1(\mathbf{x}_0) \cdot (\mathbf{v}_{\sigma(1)}, \dots, \mathbf{v}_{\sigma(r_1)}), \dots, \mathbf{D}^{r_k} \mathbf{f}_k(\mathbf{x}_0) \cdot (\mathbf{v}_{\sigma(r_1 + \cdots + r_{k-1} + 1)}, \dots, \mathbf{v}_{\sigma(r)}))$$

for $\mathbf{v}_1, \dots, \mathbf{v}_r \in \mathbb{R}^n$.

Proof We prove the theorem by induction on r, noting that the case of $r = 1$ follows from the Chain Rule, Theorem 4.4.8, and Proposition 4.4.17, using the fact that $f = \mathsf{L} \circ (f_1 \times \cdots \times f_k)$.

Assume the result is true for $r \in \{1, \dots, s\}$ and suppose that $f_1, \dots, f_k$ are of class C^{s+1}. Thus, by Proposition 4.4.7, for fixed $v_1, \dots, v_s \in \mathbb{R}^n$ the function

$$x \mapsto D^s f(x) \cdot (v_2, \dots, v_{s+1}) = \sum_{\substack{s_1,\dots,s_k \in \mathbb{Z}_{\geq 0} \\ s_1 + \cdots + s_k = s}} \sum_{\sigma \in \mathfrak{S}_{s_1,\cdots,s_k}} \mathsf{L}(D^{s_1} f_1(x) \cdot (v_{\sigma(2)}, \dots, v_{\sigma(s_1+1)}), \dots, D^{s_k} f_k(x) \cdot (v_{\sigma(s_1 + \cdots + s_{k-1} + 2)}, \dots, v_{\sigma(s+1)})),$$

is differentiable at $\boldsymbol{x}_0$, where we think of $\sigma \in \mathfrak{S}_s$ as a permutation of the set $\{2, \dots, s+1\}$ in the obvious way.

Let us now make an observation about permutations. Let $s'_1, \dots, s'_k \in \mathbb{Z}_{>0}$ have the property that $s'_1 + \cdots + s'_k = s+1$ and let $\sigma' \in \mathfrak{S}_{s'_1,\dots,s'_k}$. For brevity denote $t'_j = s'_1 + \cdots + s'_j$ for $j \in \{1, \dots, k\}$. Then there exist unique $s_1, \dots, s_k \in \mathbb{Z}_{\geq 0}$ (denote $t_j = s_1 + \cdots + s_j$, $j \in \{1, \dots, k\}$), $\sigma \in \mathfrak{S}_{s_1,\dots,s_k}$, and $j_0 \in \{1, \dots, k\}$ such that

$$s_j = \begin{cases} s'_j, & j \neq j_0, \\ s'_j - 1, & j = j_0 \end{cases}$$

and

$$\begin{aligned} ((\sigma'(t'_1 - s'_1 + 1), \dots, \sigma'(t'_1)), \dots, (\sigma'(t'_{j_0} - s'_{j_0} + 1), \dots, \sigma'(t'_{j_0})), \dots, \\ (\sigma'(t'_k - s'_k + 1) + \cdots + \sigma'(t'_k))) = ((\sigma(t_1 - s_1 + 1), \dots, \sigma(t_1)), \dots, \\ (1, \sigma(t_{j_0} - s_{j_0}), \dots, \sigma(t_{j_0} + 1)), \dots, \\ (\sigma(t_k - s_k), \dots, \sigma(t_k + 1))), \end{aligned} \tag{4.29}$$

with the convention that σ permutes the set $\{1, \dots, t'_{j_0} - s'_{j_0}, t'_{j_0} - s'_{j_0} + 2, \dots, s+1\}$ in the obvious way. The point is that $\sigma'(t'_{j_0} - s'_{j_0} + 1) = 1$, and by definition of $\mathfrak{S}_{s'_1,\dots,s'_k}$ this means that $\sigma'(t'_{j_0} - s'_{j_0} + 1)$ must appear at the beginning of one of the "slots" of length $s'_1, \dots, s'_k$. Conversely, let $s_1, \dots, s_k \in \mathbb{Z}_{\geq 0}$ be such that $s_1 + \cdots + s_k = s \geq 2$ and let $\sigma \in \mathfrak{S}_{s_1,\dots,s_k}$. Denote $t_j = s_1 + \cdots + s_j$ for $j \in \{1, \dots, k\}$. Then, for each $j_0 \in \{1, \dots, k\}$ there exist unique $s'_1, \dots, s'_k \in \mathbb{Z}_{\geq 0}$ (denote $t'_j = s'_1 + \cdots + s'_j$, $j \in \{1, \dots, k\}$) such that

$$s'_j = \begin{cases} s_j, & j \neq j_0, \\ s_j + 1, & j = j_0 \end{cases}$$

and $\sigma' \in \mathfrak{S}_{s'_1,\dots,s'_k}$ such that (4.29) holds.

Using this observation, and since the result holds for $r = 1$ and $r = s$, we can apply Proposition 4.4.7 to get

$$\begin{aligned} \boldsymbol{D}^{s+1} f(\boldsymbol{x}_0) \cdot (\boldsymbol{v}_1, \dots, \boldsymbol{v}_{s+1}) &= (\boldsymbol{D}(\boldsymbol{D}^s f)(\boldsymbol{x}_0) \cdot (\boldsymbol{v}_2, \dots, \boldsymbol{v}_{s+1})) \cdot \boldsymbol{v}_1 \\ &= \Big(\sum_{\substack{s_1,\dots,s_k \in \mathbb{Z}_{\geq 0} \\ s_1 + \cdots + s_k = s}} \sum_{\sigma \in \mathfrak{S}_{s_1,\cdots,s_k}} \mathsf{L}(\boldsymbol{D}^{s_1+1} f_1(\boldsymbol{x}_0) \cdot (\boldsymbol{v}_1, \boldsymbol{v}_{\sigma(2)}, \dots, \boldsymbol{v}_{\sigma(s_1+1)}), \dots, \\ &\quad \boldsymbol{D}^{s_k} f_k(\boldsymbol{x}_0) \cdot (\boldsymbol{v}_{\sigma(s_1+\cdots+s_{k-1}+2)}, \dots, \boldsymbol{v}_{\sigma(s+1)})) \Big) + \dots \\ &\quad + \Big(\sum_{\substack{s_1,\dots,s_k \in \mathbb{Z}_{\geq 0} \\ s_1 + \cdots + s_k = s}} \sum_{\sigma \in \mathfrak{S}_{s_1,\cdots,s_k}} \mathsf{L}(\boldsymbol{D}^{s_1} f_1(\boldsymbol{x}_0) \cdot (\boldsymbol{v}_{\sigma(2)}, \dots, \boldsymbol{v}_{\sigma(s_1+1)}), \dots, \\ &\quad \boldsymbol{D}^{s_k+1} f_k(\boldsymbol{x}_0) \cdot (\boldsymbol{v}_1, \boldsymbol{v}_{\sigma(s_1+\cdots+s_{k-1}+2)}, \dots, \boldsymbol{v}_{\sigma(s+1)})) \Big) \\ &= \sum_{\substack{s'_1,\dots,s'_k \in \mathbb{Z}_{\geq 0} \\ s'_1 + \cdots + s'_k = s+1}} \sum_{\sigma \in \mathfrak{S}_{s'_1,\cdots,s'_k}} \mathsf{L}(\boldsymbol{D}^{s'_1} f_1(\boldsymbol{x}_0) \cdot (\boldsymbol{v}_{\sigma(1)}, \dots, \boldsymbol{v}_{\sigma(s'_1+1)}), \dots, \\ &\quad \boldsymbol{D}^{s'_k} f_k(\boldsymbol{x}_0) \cdot (\boldsymbol{v}_{\sigma(s'_1+\cdots+s'_{k-1}+1)}, \dots, \boldsymbol{v}_{\sigma(s+1)})), \end{aligned}$$

as desired. ∎

In Exercise 4.4.4 we ask the reader to come to grips with the formula in the theorem by writing it down explicitly in some simple cases.

Now let us consider the Chain Rule for higher-order derivatives. To conveniently state the result we introduce the following notation. Let $r \in \mathbb{Z}_{>0}$ and let $r_1, \dots, r_j \in \mathbb{Z}_{\geq 0}$ have the property that $r_1 + \cdots + r_j = r$. Let us denote by $\mathfrak{S}^<_{r_1,\dots,r_j}$ the subset of $\mathfrak{S}_{r_1,\dots,r_j}$ given by

$$\mathfrak{S}^<_{r_1,\dots,r_j} = \{\sigma \in \mathfrak{S}_{r_1,\dots,r_j} \mid \sigma(1) < \sigma(r_1 + 1) < \cdots < \sigma(r_{j-1} + 1)\}.$$

Note, for example, that if $\sigma \in \mathfrak{S}^<_{r_1,\dots,r_j}$ then $\sigma(1) = 1$.

With this notation we have the following statement of the Chain Rule.

4.4.51 Theorem (General Chain Rule) *Let $\mathrm{U} \subseteq \mathbb{R}^n$ and $\mathrm{V} \subseteq \mathbb{R}^m$ be open, consider maps $\mathbf{f}\colon \mathrm{U} \to \mathrm{V}$ and $\mathbf{g}\colon \mathrm{V} \to \mathbb{R}^k$, and let $\mathbf{x}_0 \in \mathrm{U}$. If $\mathbf{f}$ is r times differentiable at $\mathbf{x}_0$ and if $\mathbf{g}$ is r times differentiable at $\mathbf{f}(\mathbf{x}_0)$, then $\mathbf{g} \circ \mathbf{f}$ is r times differentiable at $\mathbf{x}_0$ and, moreover,*

$$\begin{aligned}\mathbf{D}^r(\mathbf{g} \circ \mathbf{f})(\mathbf{x}_0) \cdot (\mathbf{v}_1, \dots, \mathbf{v}_r) &= \sum_{j=1}^{r} \sum_{\substack{r_1,\dots,r_j \in \mathbb{Z}_{>0} \\ r_1+\cdots+r_j=r}} \sum_{\sigma \in \mathfrak{S}^<_{r_1,\dots,r_j}} \mathbf{D}^j\mathbf{g}(\mathbf{f}(\mathbf{x}_0)) \cdot (\mathbf{D}^{r_1}\mathbf{f}(\mathbf{x}_0) \cdot (\mathbf{v}_{\sigma(1)}, \dots, \mathbf{v}_{\sigma(r_1)}), \dots, \\ &\qquad \mathbf{D}^{r_j}\mathbf{f}(\mathbf{x}_0) \cdot (\mathbf{v}_{\sigma(r_1+\cdots+r_{j-1}+1)}, \dots, \mathbf{v}_{\sigma(r)}))\end{aligned}$$

for $\mathbf{v}_1, \dots, \mathbf{v}_r \in \mathbb{R}^n$.

Proof The proof is by induction on r. For $r = 1$ the result is simply Theorem 4.4.49. Assume the result is true for $r \in \{1, \dots, s\}$ and let f and g be $s + 1$ times differentiable at x_0. We thus have

$$\begin{aligned}D^s(g \circ f)(x_0) \cdot (v_2, \dots, v_{s+1}) &= \sum_{j=1}^{s} \sum_{\substack{s_1,\dots,s_j \in \mathbb{Z}_{>0} \\ s_1+\cdots+s_j=s}} \sum_{\sigma \in \mathfrak{S}^<_{s_1,\dots,s_j}} D^j g(f(x_0)) \cdot (D^{s_1} f(x_0) \cdot (v_{\sigma(2)}, \dots, v_{\sigma(s_1+1)}), \dots, \\ &\qquad D^{s_j} f(x_0) \cdot (v_{\sigma(s_1+\cdots+s_{j-1}+2)}, \dots, v_{\sigma(s+1)}))\end{aligned}$$

for every $v_2, \dots, v_{s+1} \in \mathbb{R}^n$, and where $\sigma \in \mathfrak{S}^<_{s_1,\dots,s_j} \subseteq \mathfrak{S}_s$ permutes the set $\{2, \dots, s+1\}$ in the obvious way.

Let us now make an observation about permutations. Let $j' \in \{1, \dots, s+1\}$, let $s'_1, \dots, s'_{j'} \in \mathbb{Z}_{>0}$ satisfy $s'_1 + \cdots + s'_{j'} = s + 1$, and let $\sigma' \in \mathfrak{S}^<_{s'_1,\dots,s'_{j'}}$. For brevity denote $t'_l = s'_1 + \cdots + s'_l$ for $l \in \{1, \dots, j'\}$. We have two cases.

1. $s'_1 = 1$: In this case let $j = j' - 1$, define $s_l = s'_{l+1}$ for $l \in \{1, \dots, j' - 1\}$, and let $t_l = s_1 + \cdots + s_l$ for $l \in \{1, \dots, j\}$. We then have

$$\begin{aligned}&((1), (\sigma'(t'_2 - s'_2 + 1), \dots, \sigma'(t'_2)), \dots, (\sigma'(t'_{j'} - s'_{j'} + 1), \dots, \sigma'(t'_{j'}))) \\ &\qquad = ((1), (\sigma(t_2 - s_2 + 1), \dots, \sigma(t_2)), \dots, (\sigma(t_{j'} - s_{j'} + 1), \dots, \sigma(t_{j'}))), \quad (4.30)\end{aligned}$$

where $\sigma \in \mathfrak{S}^<_{s'_1,\dots,s'_j} \subseteq \mathfrak{S}_s$ permutes $\{2, \dots, s+1\}$ in the obvious way. Note that this uniquely specifies $s_1, \dots, s_j$ and σ.

2. $s'_1 \neq 1$: Here we take $j = j'$, $s_1 = s'_1 - 1$, $s_l = s'_l$ for $l \in \{2, \dots, j\}$. Let us denote $t_l = s_1 + \cdots + s_l$ for $l \in \{1, \dots, j\}$. Then there exist $l_0 \in \{1, \dots, j\}$ giving the corresponding cycle $\tau \in \mathfrak{S}_j$ given by $\tau = (1 \ \cdots \ l_0)$ and $\sigma \in \mathfrak{S}_{s_{\tau(1)}, s_{\tau(2)}, \dots, s_{\tau(j)}}$ such that

$$\begin{aligned}&((\sigma'(t'_1 - s'_1 + 1), \dots, \sigma'(t'_1)), \dots, (\sigma'(t'_{j'} - s'_{j'} + 1), \dots, \sigma'(t'_{j'})))\\ &\quad = ((1, \sigma(t_{\tau(1)} - s_{\tau(1)} + 1), \dots, \sigma(t_{\tau(1)})), \dots, (\sigma(t_{\tau(j)} - s_{\tau(j)} + 1), \dots, \sigma(t_{\tau(j)}))),\end{aligned} \tag{4.31}$$

where σ permutes $\{2, \dots, s+1\}$ in the obvious way. Note that this uniquely specifies $s_1, \dots, s_j$, τ, and σ. Note that the cycle τ is necessary to ensure that $\sigma'(1) = 1$, a necessary condition that $\sigma' \in \mathfrak{S}^<_{s'_1, \dots, s'_{j'}}$. The cycle serves to place the slot into which the "1" is inserted at the beginning of the slot list.

Conversely, let $j \in \{1, \dots, s\}$, let $s_1, \dots, s_j \in \mathbb{Z}_{>0}$ have the property that $s_1 + \cdots + s_j = s$, and let $\sigma \in \mathfrak{S}^<_{s_1, \dots, s_k}$. Denote $t_l = s_1 + \cdots + s_l$ for $l \in \{1, \dots, j\}$. Then we have two scenarios.

1. We take $j' = j + 1$, let $s'_1 = 1$ and $s'_l = s_{l-1}$ for $l \in \{2, \dots, s+1\}$. Define $t_l = s_1 + \cdots + s_l$. Then there exists $\sigma' \in \mathfrak{S}^<_{s'_1, \dots, s'_j}$ such that (4.30) holds. Moreover, this uniquely determines $s'_1, \dots, s'_{j'}$ and σ'.
2. We take $j = j'$ and let $l_0 \in \{1, \dots, j\}$. Then take $\tau \in \mathfrak{S}_j$ to be the cycle $(1 \ \cdots \ l_0)$. We then define $s'_1 = s_{\tau(1)} + 1$ and $s'_l = s_{\tau(l)}$ for $l \in \{2, \dots, j\}$. Then there exists $\sigma' \in \mathfrak{S}^<_{s'_1, \dots, s'_{j'}}$ such that (4.31) holds. Note that this uniquely specifies $s'_1, \dots, s'_{j'}$ and σ'.

Using this observation, along with Proposition 4.4.7, Theorems 4.4.49 and 4.4.50, and the symmetry of the derivatives of g of order up to s, we then compute

$$\begin{aligned}&D^{s+1}(g \circ f)(x_0) \cdot (v_1, \dots, v_{s+1})\\ &\quad = \sum_{j=1}^{s} \sum_{\substack{s_1, \dots, s_j \in \mathbb{Z}_{>0}\\ s_1 + \cdots + s_j = s}} \sum_{\sigma \in \mathfrak{S}^<_{s_1, \dots, s_j}} D^{j+1} g(f(x_0)) \cdot (Df(x_0) \cdot v_1,\\ &\qquad D^{s_1} f(x_0) \cdot (v_{\sigma(2)}, \dots, v_{\sigma(s_1+1)}), \dots,\\ &\qquad D^{s_j} f(x_0) \cdot (v_{\sigma(s_1 + \cdots + s_{j-1} + 2)}, \dots, v_{\sigma(s+1)}))\\ &\qquad + D^j g(f(x_0)) \cdot (D^{s_1+1} f(x_0) \cdot (v_1, v_{\sigma(2)}, \dots, v_{\sigma(s_1+1)}), \dots,\\ &\qquad D^{s_j} f(x_0) \cdot (v_{\sigma(s_1 + \cdots + s_{j-1} + 2)}, \dots, v_{\sigma(s+1)})) + \dots\\ &\qquad + D^j g(f(x_0)) \cdot (D^{s_1} f(x_0) \cdot (v_{\sigma(2)}, \dots, v_{\sigma(s_1+1)}), \dots,\\ &\qquad D^{s_j} f(x_0) \cdot (v_1, v_{\sigma(s_1 + \cdots + s_{j-1} + 2)}, \dots, v_{\sigma(s+1)}))\\ &\quad = \sum_{j'=1}^{s+1} \sum_{\substack{s'_1, \dots, s'_{j'} \in \mathbb{Z}_{>0}\\ s'_1 + \cdots + s'_{j'} = s+1}} \sum_{\sigma' \in \mathfrak{S}^<_{s'_1, \dots, s'_{j'}}} D^{j'} g(f(x_0)) \cdot (D^{s'_1} f(x_0) \cdot (v_{\sigma'(1)}, \dots, v_{\sigma'(s'_1)}),\\ &\qquad \dots, D^{s'_{j'}} f(x_0) \cdot (v_{\sigma'(s'_1 + \cdots + s'_{j'-1} + 1)}, \dots, v_{\sigma'(s+1)})),\end{aligned}$$

as desired. ∎

Let us parse the formula of the preceding result in the case where $r = 2$. We denote the components of f by $f_1, \dots, f_m$ and the components of g by $g_1, \dots, g_k$. The

components of $D^2(g \circ f)(x)$ are

$$\sum_{a,b=1}^{m} \frac{\partial^2 g_\alpha(f(x))}{\partial y_a \partial y_b} \frac{\partial f_a(x)}{\partial x^i} \frac{\partial f_b(x)}{\partial x_j} + \sum_{a=1}^{m} \frac{\partial g_\alpha(f(x))}{\partial y_a} \frac{\partial^2 f_a(x)}{\partial x_i \partial x_j}, \qquad \alpha \in \{1, \dots, k\},\ i, j \in \{1, \dots, n\}.$$

Of course, if you are familiar with how the Chain Rule and the product rule work, this is exactly the formula you would produce. In Exercise 4.4.5 we ask the reader to directly parse the formula in the theorem in the case when $r = 3$.

4.4.10 Notes

We refer to [Abraham, Marsden, and Ratiu 1988, Chapter 2] for a thorough presentation of multivariable calculus, including definitions of higher-order derivatives, the general version of Taylor's Theorem, the Inverse Function Theorem in the multivariable case, the multivariable Chain Rule, and much more. We comment that the approach in [Abraham, Marsden, and Ratiu 1988] also extends the presentation from the multivariable case to the infinite-dimensional case, and that this is important in some applications.

The proof we give of Theorem 4.4.44 follows the excellent presentation of McShane [1973]. The companion second-derivative test, Theorem 4.4.46, has hypotheses the checking of which has caused many papers to be written. The most common technique is that of "bordered Hessians" introduced by Mann [1943] and reiterated, for example, by Spring [1985].

Exercises

4.4.1 Let $\mathsf{L} \in S^k(\mathbb{R}^n; \mathbb{R}^m)$ and define $f_{\mathsf{L}} : \mathbb{R}^n \to \mathbb{R}^m$ by $f_{\mathsf{L}}(x) = \mathsf{L}(x, \dots, x)$. Show that for $r \in \{1, \dots, k\}$ we have

$$D^r f_{\mathsf{L}}(x) \cdot (v_1, \dots, v_r) = \frac{k!}{(k-r)!} \mathsf{L}(x, \dots, x, v_1, \dots, v_r).$$

4.4.2 Consider the map $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) = \|x\|^2_{\mathbb{R}^n}$.

(a) Give explicit and attractive formulae for $Df(x) \cdot v$ and $D^2 f(x) \cdot (v_1, v_2)$ for $x, v, v_1, v_2 \in \mathbb{R}^n$.

(b) Show that $D^j f(x) = \mathbf{0}$ for $x \in \mathbb{R}^n$ and $j \geq 3$.

4.4.3 Let $U \subseteq \mathbb{R}^n$ be an open set and let $f : U \to \mathbb{R}^m$ be differentiable at $x_0 \in U$. Show that $D_j f(x_0) = Df(x_0; e_j)$ for each $j \in \{1, \dots, n\}$.

4.4.4 Expand the formula of Theorem 4.4.50 in the case of $r = k = 3$.

4.4.5 Expand the formula of Theorem 4.4.51 in the case of $r = 3$.

Section 4.5

Sequences and series of functions

In this section we generalise the results of Section 3.4 to functions defined on subsets of $\mathbb{R}^n$. Much of the discussion will take a similar form to our discussion of functions whose domain is $\mathbb{R}$. However, because of the more general context, we will give some results that are of a more advanced nature.

Do I need to read this section? If a reader is acquainted with the results in Section 3.4 then this section can be bypassed on a first reading. However, when we come to use the greater generality of functions of multiple variables, the reader will want to refer back to this section to be sure that all of the extensions from the single variable case work as expected. •

4.5.1 Uniform convergence

4.5.1 Theorem (Weierstrass M-test)

4.5.2 The Weierstrass Approximation Theorem

We now give the multivariable version of the Weierstrass Approximation Theorem presented in Section 3.4.6 for the single-variable case. As we shall see, there are no substantial difficulties with adapting our single-variable proof to the multivariable case. Thus we limit the discussion, and get right to the point.

4.5.2 Definition (Polynomial functions) A function $P\colon \mathbb{R}^n \to \mathbb{R}$ is a polynomial function if

$$P(x_1, \dots, x_n) = \sum_{(k_1,\dots,k_n)\in\mathbb{Z}^n_{\geq 0}} a_{k_1\cdots k_n} x_1^{k_1} \cdots x_n^{k_n},$$

where the set of numbers $a_{k_1\cdots k_n} \in \mathbb{R}$, $(k_1, \dots, k_n) \in \mathbb{Z}^n_{\geq 0}$ have the property that the set

$$\left\{(k_1, \dots, k_n) \in \mathbb{Z}^n_{\geq 0} \;\middle|\; a_{k_1\cdots k_n} \neq 0\right\}$$

is finite. •

In Section **??** we discuss multivariable polynomials in a little detail, so the reader may be interested in reading about this material there. However, we shall be interested in only the most pedestrian aspects of such objects. Indeed our interest is in the following polynomials, recalling from Definition 3.4.19 the notation P_k^m for the single-variable Bernstein polynomials.

4.5.3 Definition (Multivariate Bernstein polynomial, multivariate Bernstein approximation) For $m_1, \dots, m_n \in \mathbb{Z}_{\geq 0}$ and for $k_j \in \{0, 1, \dots, m_j\}$, $j \in \{1, \dots, n\}$, the polynomial

function

$$P^{m_1\cdots m_n}_{k_1\cdots k_n}(x_1,\dots,x_n) = P^{m_1}_{k_1}(x_1)\cdots P^{m_n}_{k_n}(x_n)$$
$$= \binom{m_1}{k_1}\cdots\binom{m_n}{k_n}x_1^{k_1}(1-x_1)^{m_1-k_1}\cdots x_n^{k_n}(1-x_n)^{m_n-k_n}$$

is a ***Bernstein polynomial*** in n-variables. For a continuous function $f\colon R\to\mathbb{R}$ defined on a fact compact rectangle

$$R = [a_1,b_1]\times\cdots\times[a_n,b_n],$$

the $\mathbf{(m_1,\dots,m_n)}$***th Bernstein approximation*** of f is the function $B^R_{m_1\cdots m_n}f\colon R\to\mathbb{R}$ defined by

$$B^R_{m_1\cdots m_n}f(x_1,\dots,x_n) = \sum_{k_1=0}^{m_1}\cdots\sum_{k_n=0}^{m_n} f\Big(\tfrac{k_1}{m_1},\dots,\tfrac{k_n}{m_n}\Big)P^{m_1\cdots m_n}_{k_1\cdots k_n}(x_1,\dots,x_n). \qquad \bullet$$

We may now state the multivariable Weierstrass Approximation Theorem.

4.5.4 Theorem (Multivariable Weierstrass Approximation Theorem) *Let* $\mathrm{K}\subseteq\mathbb{R}^{\mathrm{n}}$ *be a compact set and let* $\mathrm{f}\colon \mathrm{K}\to\mathbb{R}$ *be continuous. Then there exists a sequence* $(\mathrm{P_m})_{\mathrm{m}\in\mathbb{Z}_{>0}}$ *of polynomial functions on* $\mathbb{R}^{\mathrm{n}}$ *such that the sequence* $(\mathrm{P_m}|\mathrm{K})_{\mathrm{m}\in\mathbb{Z}_{>0}}$ *converges uniformly to* f.

Proof First let us consider the case when $K = R$ is a fact compact rectangle. We can without loss of generality take the case when $R = [0,1]^n$, and for brevity denote $B_{m_1\cdots m_n}f = B^R_{m_1\cdots m_n}f$. We will show that the sequence $(B_{m_1\cdots m_n}f)_{(m_1,\dots,m_n)\in\mathbb{Z}_{\geq 0}}$ converges uniformly to f on R. That is to say, given $\epsilon\in\mathbb{R}_{>0}$ there exists $N\in\mathbb{Z}_{>0}$ such that, whenever $m_1,\dots,m_n\geq N$,

$$\big|B_{m_1\cdots m_n}f(\boldsymbol{x}) - f(\boldsymbol{x})\big| < \epsilon, \qquad \boldsymbol{x}\in R.$$

Let $\epsilon\in\mathbb{R}_{>0}$. Since a continuous function on the compact set R is uniformly continuous (Theorem 4.3.33) it follows that there exists $\delta\in\mathbb{R}_{>0}$ such that

$$\|\boldsymbol{x}-\boldsymbol{y}\|_{\mathbb{R}^n}\leq\delta \quad\Longrightarrow\quad |f(\boldsymbol{x})-f(\boldsymbol{y})|\leq\tfrac{\epsilon}{2}.$$

Also define

$$M = \sup\{|f(\boldsymbol{x})| \mid \boldsymbol{x}\in R\},$$

noting that this is finite by Theorem 4.3.31. Now, it $\|\boldsymbol{x}-\boldsymbol{y}\|_{\mathbb{R}^n}\leq\delta$ then

$$|f(\boldsymbol{x})-f(\boldsymbol{y})|\leq\tfrac{\epsilon}{2}\leq\tfrac{\epsilon}{2}+\tfrac{2M}{n\delta^2}(x_j-y_j)^2$$

for every $j\in\{1,\dots,n\}$. If $\|\boldsymbol{x}-\boldsymbol{y}\|_{\mathbb{R}^n}>\delta$ then

$$(x_1-y_1)^2+\cdots+(x_n-y_n)^2>\delta^2.$$

This means that, for some $j_0\in\{1,\dots,n\}$, $(x_{j_0}-y_{j_0})^2>\frac{\delta^2}{n}$. Therefore,

$$|f(\boldsymbol{x})-f(\boldsymbol{y})|\leq 2M\leq 2M\Big(\tfrac{x_{j_0}-y_{j_0}}{\sqrt{n}\delta}\Big)^2\leq\tfrac{\epsilon}{2}+\tfrac{2M}{n\delta^2}(x_{j_0}-y_{j_0})^2.$$

Thus, for every $x, y \in R$ we have

$$|f(x) - f(y)| \leq \tfrac{\epsilon}{2} + \tfrac{2M}{n\delta^2}(x_{j_0} - y_{j_0})^2$$

for some $j_0 \in \{1, \dots, n\}$.

Define $f_0 \colon R \to \mathbb{R}$ by $f_0(x) = 1$ and, for $j_0 \in \{1, \dots, n\}$, define $f_{1,j}, f_{2,j} \colon R \to \mathbb{R}$ by

$$f_{1,j}(x) = x_j, \quad f_{2,j}(x) = x_j^2.$$

Using the lemma from the proof of Theorem 3.4.21 and the Binomial Theorem, one can easily verify the following identities:

$$\begin{aligned} B_{m_1 \cdots m_n} f_0(x) &= 1; \\ B_{m_1 \cdots m_n} f_{1,j}(x) &= x_j; \\ B_{m_1 \cdots m_n} f_{2,j}(x) &= x_j^2 + \tfrac{1}{m_j}(x_j - x_j^2). \end{aligned}$$

In like manner one can also use the lemma of Theorem 3.4.21 to verify that

$$|B_{m_1 \cdots m_n} f(x)| \leq B_{m_1 \cdots m_n} g(x), \qquad x \in R$$

if $|f(x)| \leq g(x)$ for every $x \in R$.

Now fix $x_0 = (x_{0,1}, \dots, x_{0,n}) \in R$. For $x \in R$ let $j(x) \in \{1, \dots, n\}$ be such that

$$|f(x) - f(x_0)| \leq \tfrac{\epsilon}{2} + \tfrac{2M}{n\delta^2}(x_{j(x)} - x_{0,j(x)})^2$$

For every $m_1, \dots, m_n \in \mathbb{Z}_{\geq 0}$ we have

$$\begin{aligned} |B_{m_1 \cdots m_n} f(x) - f(x_0)| &= |B_{m_1 \cdots m_n}(f - f(x_0) f_0)(x)| \\ &\leq B_{m_1 \cdots m_n}\Big(\tfrac{\epsilon}{2} f_0 + \tfrac{2M}{n\delta^2}(f_{1,j(x)} - x_{0,j(x)} f_0)^2\Big)(x) \\ &= \tfrac{\epsilon}{2} + \tfrac{2M}{n\delta^2}(x_{j(x)}^2 + \tfrac{1}{m_{j(x)}}(x_{j(x)} - x_{j(x)}^2) - 2x_{0,j(x)} x_{j(x)} + x_{0,j(x)}^2) \\ &= \tfrac{\epsilon}{2} + \tfrac{2M}{n\delta^2}(x_{j(x)} - x_{0,j(x)})^2 + \tfrac{2M}{nm_{j(x)}\delta^2}(x_{j(x)} - x_{j(x)}^2). \end{aligned}$$

Now take $x = x_0$, note that $j(x_0)$ can be arbitrary, and then get, for any $j \in \{1, \dots, n\}$,

$$|B_{m_1 \cdots m_n} f(x_0) - f(x_0)| \leq \tfrac{\epsilon}{2} + \tfrac{2M}{nm_j\delta^2}(x_{0,j} - x_{0,j}^2) \leq \tfrac{\epsilon}{2} + \tfrac{M}{2nm_j\delta^2},$$

using the fact that $x - x^2 \leq \frac{1}{4}$ for $x \in [0, 1]$. Therefore, if $N \in \mathbb{Z}_{>0}$ is sufficiently large that $\frac{M}{2nm\delta^2} < \frac{\epsilon}{2}$ for $m \geq N$ we have

$$|B_{m_1 \cdots m_n} f(x_0) - f(x_0)| < \epsilon,$$

and this holds for every $x_0 \in R$, giving us the desired uniform convergence in the case where K is a rectangle.

Now consider the case where K is a general compact set and let R be a fat compact rectangle such that $K \subseteq R$. By the Tietze Extension Theorem extend f to a continuous function $\hat{f} \colon R \to \mathbb{R}$ such that $\hat{f}|K = f$. Our computations above ensure that, for $\epsilon \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that, whenever $m_1, \dots, m_n \geq N$,

$$|B_{m_1 \cdots m_n} \hat{f}(x) - \hat{f}(x)| < \epsilon, \qquad x \in R.$$

If we $P_m = B_{m \cdots m} \hat{f}$, $m \in \mathbb{Z}_{>0}$, this gives the sequence of polynomial functions converging uniformly to f on K. ∎

This can then easily be extended to maps taking values in Euclidean space by applying the preceding theorem to each component. Let us say that a map $\boldsymbol{P}\colon \mathbb{R}^n \to \mathbb{R}^m$ is a ***polynomial map*** if

$$\boldsymbol{P}(x_1, \dots, x_n) = (P_1(x_1, \dots, x_n), \dots, P_m(x_1, \dots, x_n))$$

for polynomial functions $P_j\colon \mathbb{R}^n \to \mathbb{R}$.

4.5.5 Corollary (Weierstrass Approximation Theorem for vector-valued maps) *Let* $\mathrm{K} \subseteq \mathbb{R}^n$ *be a compact set and let* $\mathbf{f}\colon \mathrm{K} \to \mathbb{R}^m$ *be continuous. Then there exists a sequence* $(\mathbf{P}_m)_{m\in\mathbb{Z}_{>0}}$ *of polynomial maps on* $\mathbb{R}^n$*, taking values in* $\mathbb{R}^m$*, such that the sequence* $(\mathbf{P}_m|\mathrm{K})_{m\in\mathbb{Z}_{>0}}$ *converges uniformly to* $\mathbf{f}$*.*

4.5.3 Swapping limits with other operations

In this section we prove some of the same results as in Section 3.4.7 concerning the swapping of limits and other operations, like integration and differentiation. One significant extension we give in this section concerns limit theorems for Riemann integration. In Section 3.4.7 we showed that for uniformly convergent sequences one can swap limit and integral. However, this is true, even for the Riemann integral in a more general setting. Here we state these results. These results are really best suited to the domain of Lebesgue integration which we discuss in Chapter **??**. However, since some version of these results are valid for the more easily understood Riemann integral, it is interesting to record them. Moreover, by comparing what is true for the Riemann integral with what is true for the more general Lebesgue integral, one can get a better appreciation of the value of the Lebesgue integral.

First we record the commutativity of the Riemann integral with increasing sequences of functions.

4.5.6 Theorem (The Monotone Convergence Theorem for the Riemann integral) *Let* $\mathrm{R} \subseteq \mathbb{R}^n$ *be a fat compact rectangle and let* $(\mathrm{f}_j)_{j\in\mathbb{Z}_{>0}}$ *be a sequence of* $\mathbb{R}$*-valued functions on* R *satisfying the following conditions:*

(i) $\mathrm{f}_j(\mathbf{x}) \geq 0$ *for each* $\mathbf{x} \in \mathrm{R}$ *and* $j \in \mathbb{Z}_{>0}$*;*

(ii) $\mathrm{f}_{j+1}(\mathbf{x}) \geq \mathrm{f}_j(\mathbf{x})$ *for each* $\mathbf{x} \in \mathrm{R}$ *and* $j \in \mathbb{Z}_{>0}$*;*

(iii) f_j *is Riemann integrable (in the sense of Definition* **??***) for each* $j \in \mathbb{Z}_{>0}$*;*

(iv) the map $\mathrm{f}\colon \mathrm{R} \to \mathbb{R}_{\geq 0}$ *defined by* $\mathrm{f}(\mathbf{x}) = \lim_{j\to\infty} \mathrm{f}_j(\mathbf{x})$ *exists and is Riemann integrable (in the sense of Definition* **??***).*

Then

$$\lim_{j\to\infty} \int_{\mathrm{R}} \mathrm{f}_j(\mathbf{x})\, d\mathbf{x} = \int_{\mathrm{R}} \mathrm{f}(\mathbf{x})\, d\mathbf{x}.$$

Proof We first prove a couple of lemmata.

1 Lemma *Let* $\mathrm{R} \subseteq \mathbb{R}^{\mathrm{n}}$ *be a fat compact rectangle, let* $\mathrm{f}\colon \mathrm{R} \to \mathbb{R}$ *be bounded and Riemann integrable with*

$$\mathrm{M} = \sup\{|\mathrm{f}(\mathbf{x})| \mid \mathbf{x} \in \mathrm{R}\},$$

and suppose that $\int_{\mathrm{R}} \mathrm{f}(\mathbf{x})\,\mathrm{d}\mathbf{x} \geq \mathrm{m}\,\mathrm{vol}(\mathrm{R})$ *for some* $\mathrm{m} \in \mathbb{R}_{>0}$. *Then the set*

$$\left\{\mathbf{x} \in \mathrm{R} \;\middle|\; \mathrm{f}(\mathbf{x}) \geq \tfrac{\mathrm{m}}{2}\mathrm{vol}(\mathrm{R})\right\}$$

contains a finite union of rectangles whose total volume is bounded below by $\frac{\mathrm{m}}{4\mathrm{M}}\mathrm{vol}(\mathrm{R})$.

Proof Let $\boldsymbol{P}$ be a partition of R for which

$$0 \leq \int_R f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - A_-(f, \boldsymbol{P}) \leq \frac{m}{4}\mathrm{vol}(R).$$

Therefore $A_-(f, \boldsymbol{P}) \geq \frac{3m}{4}\mathrm{vol}(R)$. Let us write $\boldsymbol{P} = (P_1, \ldots, P_n)$ with $P_j = (I_{j1}, \ldots, I_{jk_j})$, $j \in \{1, \ldots, n\}$. Let

$$E = \left\{\boldsymbol{x} \in R \;\middle|\; f(\boldsymbol{x}) \geq \tfrac{m}{2}\right\}$$

and denote

$$L_1 = \{(l_1, \ldots, l_n) \in \{1, \ldots, k_1\} \times \cdots \times \{1, \ldots, k_n\} \mid R_{l_1,\ldots,l_m} \subseteq E\}$$

and

$$L_2 = (\{1, \ldots, k_1\} \times \cdots \times \{1, \ldots, k_n\}) \setminus L_1.$$

We then have

$$\begin{aligned}
\frac{3m}{4}\mathrm{vol}(R) \leq A_-(f, \boldsymbol{P}) &= \sum_{(l_1,\ldots,l_n)\in L_1} \inf\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{cl}(R_{l_1,\ldots,l_n})\}\mathrm{vol}(R_{l_1,\ldots,l_n}) \\
&\quad + \sum_{(l_1,\ldots,l_n)\in L_2} \inf\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{cl}(R_{l_1,\ldots,l_n})\}\mathrm{vol}(R_{l_1,\ldots,l_n}) \\
&\leq \sum_{(l_1,\ldots,l_n)\in L_1} M\mathrm{vol}(R_{l_1,\ldots,l_n}) + \sum_{(l_1,\ldots,l_n)\in L_1} \frac{m}{2}\mathrm{vol}(R_{l_1,\ldots,l_n}) \\
&\leq \sum_{(l_1,\ldots,l_n)\in L_1} M\mathrm{vol}(R_{l_1,\ldots,l_n}) + \frac{m}{2}\mathrm{vol}(R).
\end{aligned}$$

Therefore,

$$\sum_{(l_1,\ldots,l_n)\in L_1} \mathrm{vol}(R_{l_1,\ldots,l_n}) \geq \frac{m}{4M}\mathrm{vol}(R),$$

giving the lemma. ▼

Using the preceding lemma we prove the following result.

2 Lemma *Let* $\mathrm{R} \subseteq \mathbb{R}^{\mathrm{n}}$ *be a fat compact rectangle, let* $(\mathrm{g_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence of* $\mathbb{R}$*-valued functions on* R *satisfying the following conditions:*

(i) $\mathrm{g_j}(\mathbf{x}) \geq 0$ *for each* $\mathbf{x} \in \mathrm{R}$ *and* $\mathrm{j} \in \mathbb{Z}_{>0}$;

(ii) $\mathrm{g_{j+1}}(\mathbf{x}) \leq \mathrm{g_j}(\mathbf{x})$ *for each* $\mathbf{x} \in \mathrm{R}$ *and* $\mathrm{j} \in \mathbb{Z}_{>0}$;

(iii) $\mathrm{g_j}$ *is Riemann integrable (in the sense of Definition* **??***) for each* $\mathrm{j} \in \mathbb{Z}_{>0}$;

(iv) $\lim_{\mathrm{j}\to\infty} \mathrm{g_j}(\mathbf{x}) = 0$ *for all* $\mathbf{x} \in \mathrm{R}$.

Then

$$\lim_{j\to\infty} \int_R g_j(x)\,dx = 0.$$

Proof The hypotheses ensure that the sequence whose jth term is $\int_R g_j(x)\,dx$ is monotonically decreasing and positive. Therefore, it converges by Theorem 2.3.8. Let us denote the limit by $\tilde{L} \geq 0$ and suppose, in fact, that $\tilde{L} > 0$. Let us denote $L = \frac{\tilde{L}}{\text{vol}(R)}$. For $j \in \mathbb{Z}_{>0}$, let $g_{j,M} \colon R \to \mathbb{R}_{\geq 0}$ be defined by $g_{j,M}(x) = \min\{g_j(x), M\}$. Since g_1 is Riemann integrable in the sense of Definition **??**, let $M_0 \in \mathbb{R}_{>0}$ be such that $M_0 > \frac{2L}{5}\text{vol}(R)$ and such that

$$\int_R g_1(x)\,dx - \int_R g_{1,M_0}(x)\,dx \leq \frac{L}{5}\text{vol}(R).$$

For each $j \in \mathbb{Z}_{>0}$ we have

$$\{x \in R \mid g_j(x) \geq M_0\} \subseteq \{x \in R \mid g_1(x) \geq M_0\}$$

since $g_j(x) \leq g_1(x)$ for all $x \in R$. This gives

$$0 \leq \int_R (g_j(x) - g_{j,M_0}(x))\,dx \leq \int_R (g_1(x) - g_{1,M_0}(x))\,dx \leq \frac{L}{5}\text{vol}(R).$$

Since $\int_R g_j(x)\,dx \geq L\text{vol}(R)$ (by definition of L) it follows that $\int_R g_{j,M_0}(x)\,dx \geq \frac{4L}{5}\text{vol}(R)$. Now, for $j \in \mathbb{Z}_{>0}$, define

$$E_j = \left\{x \in R \;\middle|\; g_j(x) \geq \tfrac{2L}{5}\text{vol}(R)\right\}.$$

Since $M_0 \geq \frac{2L}{5}\text{vol}(R)$ we also have

$$E_j = \left\{x \in R \;\middle|\; g_{j,M_0}(x) \geq \tfrac{2L}{5}\text{vol}(R)\right\}.$$

By Lemma 1 the set E_j contains a finite number of rectangles whose total volume is bounded below by $\frac{L}{5M_0}\text{vol}(R)$. By Theorem **??** and Exercise 4.2.12 it follows that the set

$$D = \cup_{j\in\mathbb{Z}_{>0}}\{x \in R \mid g_j(x) \text{ is discontinuous at } x\}$$

has measure zero. Therefore, there is a countable collection of open rectangles covering D and having total volume bounded above by $\frac{L}{10M_0}\text{vol}(R)$. Denote by U the union of these rectangles. We claim that $E_j \not\subset U$ for each $j \in \mathbb{Z}_{>0}$. Indeed, if $E_j \subseteq U$ then $\text{vol}(E_j) \leq \text{vol}(U)$, but this cannot be since $\text{vol}(E_j) \geq \frac{L}{5M_0}\text{vol}(R)$ and $\text{vol}(U) \leq \frac{L}{10M_0}\text{vol}(R)$. Let $x \in \text{cl}(E_j)\backslash E_j$. Thus $g_j(x) < \frac{2L}{5}\text{vol}(R)$ by definition of E_j. There then exists a sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$ in E_j converging to x_0. Since $g_j(x_k) \geq \frac{2L}{5}\text{vol}(R)$ for each $k \in \mathbb{Z}_{>0}$ by definition of E_j it follows that $\lim_{k\to\infty} g_j(x_k) \neq g_j(x)$, and so g_j is discontinuous at x. Thus $x \in D \subseteq U$. This shows that $\text{cl}(E_j) \subseteq E_j \cup U$. Now, for $j \in \mathbb{Z}_{>0}$, define $F_j = \text{cl}(E_j) - U$ so that $F_j \subseteq E_j$. Thus F_j is bounded since E_j is bounded. We claim that it is also closed. To see this, let $(x_k)_{k\in\mathbb{Z}_{>0}}$ be a sequence in F_j converging to x. Since $F_j \subseteq \text{cl}(E_j)$ it follows that $x \in \text{cl}(E_j)$. We also claim that $x \notin U$. Indeed, since U is open, if $x \in U$ it must follow that $x_k \in U$ for sufficiently large k, contradicting the fact that $(x_k)_{k\in\mathbb{Z}_{>0}}$ is a sequence in F_j. Thus $x \in \text{cl}(E_j) - U = F_j$. Thus F_j is closed and so compact by the Heine–Borel Theorem. Since $E_{j+1} \subseteq E_j$ it follows that $F_{j+1} \subseteq F_j$. By Proposition 4.2.39 it follows that $\cap_{j\in\mathbb{Z}_{>0}} F_j$ is nonempty. Thus, $\cap_{j\in\mathbb{Z}_{>0}} E_j$ is nonempty. Thus there exists $x \in R$ such that $g_j(x) \geq \frac{2L}{5}\text{vol}(R)$, contradicting the fact that the sequence $(g_j)_{j\in\mathbb{Z}_{>0}}$ converges pointwise to zero. ▼

Now we proceed with the proof of the theorem. With $(f_j)_{j\in\mathbb{Z}_{>0}}$ and f as in the statement of the theorem, let $g_j = f - f_j$ for $j \in \mathbb{Z}_{>0}$. One can easily verify that the sequence $(g_j)_{j\in\mathbb{Z}_{>0}}$ satisfies the hypotheses of Lemma 2. Thus, by the lemma,

$$0 = \lim_{j\to\infty} \int_R g_j(x)\,\mathrm{d}x = \lim_{j\to\infty} \int_R (f(x) - f_j(x))\,\mathrm{d}x = \int_R f(x)\,\mathrm{d}x - \lim_{j\to\infty} \int_R f_j(x)\,\mathrm{d}x,$$

where we have used Proposition **??**. This gives the result. ∎

Let us give some examples which show the value and limitations of the Monotone Convergence Theorem for the Riemann integral.

4.5.7 Examples (The Monotone Convergence Theorem for the Riemann integral)

1. Let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be an enumeration of the rational numbers in the interval $[0,1]$; such an enumeration is possible by Exercise 2.1.3. Define a sequence of functions $(g_j)_{j\in\mathbb{Z}_{>0}}$ from $[0,1]$ to $\mathbb{R}_{\geq 0}$ by

$$g_j(x) = \begin{cases} 1, & x = q_j, \\ 0, & \text{otherwise.} \end{cases}$$

Then define $f_k = \sum_{j=1}^{k} g_j$. One easily verifies that the sequence of functions $(f_k)_{k\in\mathbb{Z}_{>0}}$ satisfies the first three hypotheses of the Monotone Convergence Theorem. Moreover, since the Riemann integral of each of the functions g_j, $j \in \mathbb{Z}_{>0}$, is zero (why?) it follows by Proposition **??** that each of functions f_k, $k \in \mathbb{Z}_{>0}$, has Riemann integral zero. Thus

$$\lim_{k\to\infty} f_k(x)\,\mathrm{d}x = 0.$$

However, the pointwise limit of the sequence $(f_k)_{k\in\mathbb{Z}_{>0}}$ is the function $f\colon [0,1] \to \mathbb{R}_{\geq 0}$ defined by

$$f(x) = \begin{cases} 1, & x \in \mathbb{Q}, \\ 0, & \text{otherwise,} \end{cases}$$

i.e., the characteristic function of $\mathbb{Q} \cap [0,1]$. However, we have already seen in Example 3.3.10 that this function is not Riemann integrable. Thus the Monotone Convergence Theorem for the Riemann integral does not hold in this case. *Punchline:* The condition that the pointwise limit function f is Riemann integrable appears in the *hypotheses* of the Monotone Convergence Theorem, not in its *conclusions*. This is a significant defect of the Riemann integral. As we shall see with the various versions of the Monotone Convergence Theorem in Chapter **??**, for more general notions of the integral the integrability of the pointwise limit function follows as a conclusion.

2. On $[0,1]$ consider the sequence of functions $(f_j)_{j\in\mathbb{Z}_{>0}}$ given by

$$f_j(x) = \begin{cases} \frac{1}{(jx)^{1/2}}, & x \in (0,1], \\ 0, & x = 0. \end{cases}$$

One can readily verify (cf. Example **??**) that each of the functions f_j is Riemann integrable. Moreover, for each $x \in [0,1]$ it follows that $\lim_{j\to\infty} f_j(x) = 0$. Thus the pointwise limit of the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ is the zero function. Therefore, the limit function is Riemann integrable. Note that this sequence does not quite satisfy the hypotheses of the Monotone Convergence Theorem since the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is monotonically decreasing, not increasing, for each $x \in [0,1]$. However, the Monotone Convergence Theorem more or less obviously applies to this case as well (also see Lemma 2 in the proof of the Monotone Convergence Theorem). Indeed, the Monotone Convergence Theorem gives

$$\lim_{j\to\infty} \int_0^1 \frac{1}{(jx)^{1/2}}\,\mathrm{d}x = 0.$$

This can also be checked directly.
Punchline: The Monotone Convergence Theorem applies to sequences of possibly unbounded functions. •

The following result gives conditions, in the absence of positivity of the functions in the sequence, under which we can swap limit and integral.

4.5.8 Theorem (Dominated Convergence Theorem for the Riemann integral) *Let* $\mathrm{R} \subseteq \mathbb{R}^n$ *be a fat compact rectangle and let* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence of* $\mathbb{R}$*-valued functions on* R *satisfying the following conditions:*

(i) *there exists* $\mathrm{M} \in \mathbb{R}_{>0}$ *such that* $\mathrm{f_j}(\mathbf{x}) \le \mathrm{M}$ *for each* $\mathbf{x} \in \mathrm{R}$ *and* $\mathrm{j} \in \mathbb{Z}_{>0}$*;*

(ii) $\mathrm{f_j}$ *is Riemann integrable for each* $\mathrm{j} \in \mathbb{Z}_{>0}$*;*

(iii) *the map* $\mathrm{f}\colon \mathrm{R} \to \mathbb{R}$ *defined by* $\mathrm{f}(\mathbf{x}) = \lim_{\mathrm{j}\to\infty} \mathrm{f_j}(\mathbf{x})$ *exists and is Riemann integrable.*

Then

$$\lim_{\mathrm{j}\to\infty} \int_{\mathrm{R}} \mathrm{f_j}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathrm{R}} \mathrm{f}(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Proof We first prove a lemma.

1 Lemma *Let* $(\mathrm{A_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence of subsets of* $\mathbb{R}^n$ *having the properties*

(i) $\mathrm{A_{j+1}} \subseteq \mathrm{A_j}$, $\mathrm{j} \in \mathbb{Z}_{>0}$*, and*

(ii) $\cap_{\mathrm{j}\in\mathbb{Z}_{>0}} \mathrm{A_j} = \emptyset$*.*

For $\mathrm{j} \in \mathbb{Z}_{>0}$ *define*

$$\nu_{\mathrm{j}} = \inf\{\mathrm{vol(B)} \mid \mathrm{B} \subseteq \mathrm{A_j} \text{ is a finite union of rectangles}\}.$$

Then $\lim_{\mathrm{j}\to\infty} \nu_{\mathrm{j}} = 0$*.*

Proof If there exists $N \in \mathbb{Z}_{>0}$ such that A_N contains no set which is a finite union of fat rectangles then it follows that the sets A_j, $j \ge N$, contain no set which is a finite union of fat rectangles. In this case, the lemma holds vacuously. Thus we can suppose, without loss of generality, that each set A_j contains a set which is a finite union of fat rectangles. Since the sequence of subsets $(A_j)_{j\in\mathbb{Z}_{>0}}$ is decreasing with respect to the partial order of inclusion it follows that the sequence $(\nu_j)_{j\in\mathbb{Z}_{>0}}$ is a decreasing sequence of strictly positive numbers. This sequence converges by Theorem 2.3.8. Suppose that

it converges to $L \in \mathbb{R}_{>0}$. For each $j \in \mathbb{Z}_{>0}$ let $B_j \subseteq A_j$ be a finite union of closed fat rectangles having the property that

$$\operatorname{vol}(B_j) = \nu_j - \tfrac{L}{2^j}. \tag{4.32}$$

For $m \in \mathbb{Z}_{>0}$ let us define $K_m = \cap_{j=1}^m B_j$. Since K_m is an intersection of closed sets it is closed by Exercise 4.2.3. Since the sets K_m, $m \in \mathbb{Z}_{>0}$, are obviously bounded it follows from the Heine–Borel Theorem that they are compact. We next claim that K_m is nonempty for each $m \in \mathbb{Z}_{>0}$. Let $j \in \mathbb{Z}_{>0}$. If $B \subseteq A_j \setminus B_j$ is a finite union of rectangles then we have

$$\operatorname{vol}(B) + \operatorname{vol}(B_j) = \operatorname{vol}(B \cup B_j) \leq \nu_j$$

since B and B_j are disjoint. By (4.32) it then follows that

$$\operatorname{vol}(B) \leq \frac{L}{2^j}. \tag{4.33}$$

Now, for $m \in \mathbb{Z}_{>0}$, let $B \subseteq A_m \setminus K_m$. By Proposition 1.1.5 we have

$$B = (B \setminus B_1) \cup \cdots \cup (B \setminus B_m). \tag{4.34}$$

Since B and B_j are each finite unions of rectangles, $B \setminus B_1$ is a finite union of rectangles for each $j \in \{1, \ldots, m\}$ (why?). Therefore, for each $j \in \{1, \ldots, m\}$, $B \setminus B_j$ is a subset of $A_j \setminus E_j$ that is a finite union of rectangles. By (4.33) this means that $\operatorname{vol}(B \setminus B_j) < \frac{L}{2^n}$, $j \in \{1, \ldots, m\}$. By (4.34) it follows that

$$\operatorname{vol}(B) \leq \sum_{j=1}^m \operatorname{vol}(B \setminus B_j) \leq L \sum_{j=1}^m \frac{1}{2^j} < L.$$

Now, since A_m must contain a set which is a finite union of rectangles with the union having volume at least L, and since any subset of $A_m \setminus K_m$ that is a finite union of rectangles has volume at most L, it follows that $K_m \neq \emptyset$. Now, by Proposition 4.2.39 it follows that $\cap_{m=1}^\infty K_m \neq \emptyset$. Since $K_j \subseteq A_j$ for each $j \in \mathbb{Z}_{>0}$, it then follows that $\cap_{j=1}^\infty A_j \neq \emptyset$, so violating the hypotheses of the lemma. Thus the assumption that the sequence $(\nu_j)_{j \in \mathbb{Z}_{>0}}$ converges to a positive number is invalid. ▼

Next we prove the theorem for the case when the functions f_j, $j \in \mathbb{Z}_{>0}$, take values in $\mathbb{R}_{\geq 0}$ and when the limit function f is the zero function. In this case, let $\epsilon \in \mathbb{R}_{>0}$ and for $j \in \mathbb{Z}_{>0}$ define

$$A_j = \left\{ \boldsymbol{x} \in R \;\middle|\; f_k(\boldsymbol{x}) \geq \tfrac{\epsilon}{4\operatorname{vol}(R)} \text{ for some } k \geq j \right\}.$$

Clearly $A_{j+1} \subseteq A_j$ for all $j \in \mathbb{Z}_{>0}$. Moreover, since the sequence $(f_j)_{j \in \mathbb{Z}_{>0}}$ converges pointwise to zero, $\cap_{j \in \mathbb{Z}_{>0}} A_j = \emptyset$. By the lemma above let $N \in \mathbb{Z}_{>0}$ be sufficiently large that, for $j \geq N$, if $B \subseteq A_j$ is a finite union of rectangles then $\operatorname{vol}(B) < \frac{\epsilon}{4M}$. Let $j \geq N$ and let $\boldsymbol{P}$ be a partition such that

$$\int_R f_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \int_R A_-(f_j, \boldsymbol{P})(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} < \frac{\epsilon}{2}.$$

Define

$$B = \left\{ \boldsymbol{x} \in R \;\middle|\; A_-(f_j, \boldsymbol{P})(\boldsymbol{x}) \geq \tfrac{\epsilon}{4\operatorname{vol}(R)} \right\}$$

and $B' = R \setminus B$. Since $A_-(f, \boldsymbol{P})$ is a step function, B, and therefore B', is a finite union of rectangles. We then have

$$\begin{aligned}\int_R f_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \int_R f_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_R A_-(f,\boldsymbol{P})(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int_R A_-(f,\boldsymbol{P})(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &\leq \frac{\epsilon}{2} + \int_B A_-(f,\boldsymbol{P})(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int_{B'} A_-(f,\boldsymbol{P})(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &\leq \frac{\epsilon}{2} + M\mathrm{vol}(B) + \frac{\epsilon}{4\mathrm{vol}(R)}\mathrm{vol}(B') \leq \frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \epsilon.\end{aligned}$$

Thus $\lim_{j\to\infty} \int_R f_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$ giving the theorem in this case.

Finally to prove the theorem, given the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ and f as in the statement of the theorem, define $g_j(\boldsymbol{x}) = |f(\boldsymbol{x}) - f_j(\boldsymbol{x})|$, $\boldsymbol{x} \in R$, $j \in \mathbb{Z}_{>0}$. By Propositions **??** and **??** it follows that the functions g_j, $j \in \mathbb{Z}_{>0}$, are Riemann integrable. Moreover, they take values in $\mathbb{R}_{\geq 0}$ and converge pointwise to zero. Therefore, by the special case we considered above we have

$$\lim_{j\to\infty}\left|\int_R f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_R f_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right| \leq \lim_{j\to\infty}\int_R |f(\boldsymbol{x}) - f_j(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x} = 0,$$

using Proposition **??**. Thus the theorem follows. ■

4.5.4 Notes

The Dominated Convergence Theorem for the Riemann integral is due to Arzelà [1885] and Arzelà [1900], and the proof we give is an adaptation of the proof of Lewin [1986]. See also [Gordon 2000].

Bibliography

Abraham, R., Marsden, J. E., and Ratiu, T. S. [1988] *Manifolds, Tensor Analysis, and Applications*, number 75 in Applied Mathematical Sciences, Springer-Verlag: New York/Heidelberg/Berlin, ISBN: 978-0-387-96790-5.

Arzelà, C. [1885] *Sulla integrazione per serie*, Atti della Accademia Nazionale dei Lincei. Memorie. Classe di Scienze Fisiche, Matematiche e Naturali. Sezione Ia. Matematica, Meccanica, Astronomia, Geodesia e, **4**(1), pages 532–537, ISSN: 0391-8149.

— [1900] *Sulle serie di funzioni*, Atti della Accademia delle Scienze dell'Istituto di Bologna. Classe di Scienze Fisiche. Rendiconti. Serie XIII, **5**(8), pages 701–704, ISSN: 1122-4142.

Bernstein, S. N. [1912] *Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités*, Communication de la Société Mathématique de Kharkov, **13**, pages 1–2.

Bridges, D. S. and Richman, F. [1987] *Varieties of Constructive Mathematics*, number 97 in London Mathematical Society Lecture Note Series, Cambridge University Press: New York/Port Chester/Melbourne/Sydney, ISBN: 978-0-521-31802-0.

Brouwer, L. E. J. [1912] *Beweis zur Invarianz des n-dimensionalen Gebiets*, Mathematische Annalen, **72**(1), pages 55–56, ISSN: 0025-5831, DOI: 10.1007/BF01456846.

Chellaboina, V.-S. and Haddad, W. M. [1995] *Is the Frobenius matrix norm induced?*, Institute of Electrical and Electronics Engineers. Transactions on Automatic Control, **40**(12), pages 2137–2139, ISSN: 0018-9286, DOI: 10.1109/9.478340.

Cohen, P. J. [1963] *A minimal model for set theory*, American Mathematical Society. Bulletin. New Series, **69**, pages 537–540, ISSN: 0273-0979, DOI: 10.1090/S0002-9904-1963-10989-1.

Dirichlet, J. P. G. L. [1842] *Verallgemeinerung eines Satzes aus der Lehre von den Kettenbrüchen nebst einigen Anwendungen auf die Theorie der Zahlen*, Bericht über die Verhandlungen der Königlich Preussischen Akademie der Wissenschaften, pages 93–95.

Drakakis, K. and Pearlmutter, B. A. [2009] *On the calculation of the $\ell_2 \to \ell_1$ induced matrix norm*, International Journal of Algebra, **3**(5), pages 231–240, ISSN: 1312-8868, URL: http://www.m-hikari.com/ija/ija-password-2009/ija-password5-8-2009/drakakisIJA5-8-2009.pdf.

Gödel, K. [1931] *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme*, Monatshefte für Mathematik, **38**(1), pages 173–189, ISSN: 0026-9255, DOI: 10.1007/s00605-006-0423-7.

Gordon, R. A. [1998] *The use of tagged partitions in elementary real analysis*, The American Mathematical Monthly, **105**(2), pages 107–147, ISSN: 0002-9890, DOI: 10.2307/2589642.

Gordon, R. A. [2000] *A convergence theorem for the Riemann integral*, Mathematics Magazine, **73**(2), pages 141–147, ISSN: 0025-570X, DOI: 10.2307/2691086.

Hörmander, L. [1966] *An Introduction to Complex Analysis in Several Variables*, Van Nostrand Reinhold Co.: London, Reprint: [Hörmander 1990].

— [1990] *An Introduction to Complex Analysis in Several Variables*, 3rd edition, number 7 in North Holland Mathematical Library, North-Holland: Amsterdam/New York, ISBN: 978-0-444-88446-6, Original: [Hörmander 1966].

Horn, R. A. and Johnson, C. R. [1990] *Matrix Analysis*, Cambridge University Press: New York/Port Chester/Melbourne/Sydney, ISBN: 978-0-521-38632-6.

Hurewicz, W. and Wallman, H. [1941] *Dimension Theory*, number 4 in Princeton Mathematical Series, Princeton University Press: Princeton, NJ, ISBN: 978-0-691-07947-9.

Krantz, S. G. and Parks, H. R. [2002] *A Primer of Real Analytic Functions*, 2nd edition, Birkhäuser Advanced Texts, Birkhäuser: Boston/Basel/Stuttgart, ISBN: 978-0-8176-4264-8.

Kronecker, L. [1899] *Werke*, volume 3, Teubner: Leipzig.

Kueh, K.-L. [1986] *A note on Kronecker's approximation theorem*, The American Mathematical Monthly, **93**(7), pages 555–556, ISSN: 0002-9890, DOI: 10.2307/2323034.

Lewin, J. [1986] *A truly elementary approach to the bounded convergence theorem*, The American Mathematical Monthly, **93**(5), pages 395–397, ISSN: 0002-9890, DOI: 10.2307/2323608.

— [1991] *A simple proof of Zorn's lemma*, The American Mathematical Monthly, **98**(4), pages 353–354, ISSN: 0002-9890, DOI: 10.2307/2323807.

Mann, H. B. [1943] *Quadratic forms with linear constraints*, The American Mathematical Monthly, **50**(7), pages 430–433, ISSN: 0002-9890, DOI: 10.2307/2303666.

McCarthy, J. [1953] *An everywhere continuous nowhere differentiable function*, The American Mathematical Monthly, **60**(10), page 709, ISSN: 0002-9890, DOI: 10.2307/2307157.

McShane, E. J. [1973] *The Lagrange multiplier rule*, The American Mathematical Monthly, **80**(8), pages 922–925, ISSN: 0002-9890, DOI: 10.2307/2319406.

Moore, G. H. [1982] *Zermelo's Axiom of Choice: Its Origins, Development, and Influence*, Springer-Verlag: New York/Heidelberg/Berlin, ISBN: 0-387-90670-3, Reprint: [Moore 2013].

— [2013] *Zermelo's Axiom of Choice: Its Origins, Development, and Influence*, Dover Publications, Inc.: New York, NY, ISBN: 978-0-486-48841-7, Original: [Moore 1982].

Munkres, J. R. [1984] *Elements of Algebraic Topology*, Addison Wesley: Reading, MA, ISBN: 978-0-201-04586-4.

Niven, I. [1947] *A simple proof that π is irrational*, American Mathematical Society. Bulletin. New Series, **53**, page 509, ISSN: 0273-0979, DOI: 10.1090/S0002-9904-1947-08821-2.

Robinson, A. [1974] *Non-Standard Analysis*, Princeton Mathematical Series, Princeton University Press: Princeton, NJ, Reprint: [Robinson 1996].

— [1996] *Non-Standard Analysis*, Princeton Landmarks in Mathematics, Princeton University Press: Princeton, NJ, ISBN: 978-0-691-04490-3, Original: [Robinson 1974].

Rohn, J. [2000] *Computing the norm $\|A\|_{\infty,1}$ is NP-hard*, Linear and Multilinear Algebra, **47**(3), ISSN: 0308-1087, DOI: 10.1080/03081080008818644.

Siksek, S. and El-Sedy, E. [2004] *Points of non-differentiability of convex functions*, Applied Mathematics and Computation, **148**(3), pages 725–728, ISSN: 0096-3003, DOI: 10.1016/S0096-3003(02)00932-3.

Spring, D. [1985] *On the second derivative test for constrained local extrema*, The American Mathematical Monthly, **92**(9), pages 631–643, ISSN: 0002-9890, DOI: 10.2307/2323709.

Suppes, P. [1960] *Axiomatic Set Theory*, The University Series in Undergraduate Mathematics, Van Nostrand Reinhold Co.: London, Reprint: [Suppes 1972].

— [1972] *Axiomatic Set Theory*, Dover Publications, Inc.: New York, NY, ISBN: 978-0-486-61630-8, Original: [Suppes 1960].