Mathematics for AI

Oliver Roche-Newton

August 4, 2023

Abstract

Lecture notes for the courses Mathematics for AI 1 and 2 at JKU (Winter Semester 2022/23 and Summer Semester 2023). The notes are based on Mario Ullrich's combined notes for Mathematics for AI 1-3. I am grateful to Jan-Michael Holzinger, Severin Bergsmann, Sara Plavsic and Tereza Votýpková for their help in constructing the notes.

Contents

1 Sets, Numbers and Functions

In this section we will introduce some of the most fundamental objects of mathematics. In particular, we will focus on numbers, sets (usually sets of numbers), and relations between them. We will see some simple proofs and learn about some important proof techniques, particularly proof by induction and proof by contradiction. Finally, we treat complex numbers, which are necessary to give solutions to arbitrary polynomial equations.

The content of this section will form the basis for the mathematics we learn throughout this course and also the upcoming courses Mathematics for AI 2 and 3, and so it is essential to have a solid understanding of the concepts we introduce here.

1.1 Sets

A set M is a collection of different 'objects' which we call elements of M. We use the following notation:

$$x \text{ belongs to } M, \qquad \text{we write } x \in M$$

or

$$x \text{ does not belong to } M, \qquad \text{we write } x \notin M.$$

Some particularly important sets are assigned names and symbols.

- $\mathbb{N} := \{1, 2, 3, \dots\}$ is the set of **natural numbers**.
- $\mathbb{N}_0 := \{0, 1, 2, 3, \dots\}$ is the set of **non-negative integers**.
- $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ is the set of **integers**.
- $\mathbb{Q}$ is the set of rational numbers.
- $\mathbb{R}$ is the set of real numbers.
- $\mathbb{C}$ is the set of complex numbers.
- $\mathbb{P}$ is the set of prime numbers.

All these sets will be precisely defined and discussed later in this chapter. First, let us see that there are multiple ways to define sets. The easiest way would be to list all its elements, for instance:

$$A = \{0, 1, 2\}, \quad B = \{\text{Austria, Germany, Italy, Liechtenstein}\}.$$

However, if we have sets containing an infinite amount of elements, we cannot name all of them. In this case we use dots if it is clear what is contained in the set. We did this already for the sets $\mathbb{N}$ and $\mathbb{Z}$. Some more examples are the even and odd natural numbers:

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

However, this may lead to difficulties of interpretation as this is not guaranteed to give a unique description. For example, we may define the set of even numbers, as above, via

$$E = \{2, 4, \dots\},$$

and the set of all powers of 2 as

$$G = \{2, 4, \dots\}.$$

Since these sets do not differ until the third element, they appear identical in this notation. For a good definition of an infinite set, it is therefore formally necessary to precisely specify the properties of its elements. For instance, the set of even numbers can be written as

$$E = \{2k : k \in \mathbb{N}\},$$

and the set of all powers of 2 as

$$G = \{2^j : j \in \mathbb{N}\}.$$

A special but important set is the empty set, denoted $\emptyset$, which does not contain any element.

Two sets M and N can also be related to each other. If for all $m \in M$ we also have $m \in N$ then we say M is a **subset** of N, and we write $M \subset N$ In this case, we also call N a **superset** of M. If we have a look at the sets defined above, we have e.g. $E \subset \mathbb{N}$ and $G \subset E$. Note that, for any set M we have the relations $M \subset M$ and $\emptyset \subset M$.

Sets M and N are called **equal** if they contain the same elements, i.e. $M \subset N$ and $N \subset M$. For example we have

$$\{0, 1, 2\} = \{2, 0, 1\}$$

and

$$\{2, 3, 5, 7\} = \{p \in \mathbb{P} \text{ such that } p \leq 9\}.$$

To verify that two sets X and Y are equal, we need to check that both $X \subset Y$ and $Y \subset X$.

Exercise - Show that

$$\{2, 3, 5, 7\} = \{p \in \mathbb{P} \text{ such that } p \leq 9\}.$$

If we have $M \subset N$ and $M \neq N$, then we say that M is a **proper** or **strict subset** of N and write $M \subsetneq N$. For example,

$$\mathbb{N} \subsetneq \mathbb{N}_0 \subsetneq \mathbb{Z}. \tag{1}$$

Some authors prefer to use “$\subseteq$” for a generic subset instead of “$\subset$‘”, in order to make it more explicit that equality is not excluded. The same authors typically use the notation “$\subset$” instead of “$\subsetneq$” for proper subsets. So, one should be mindful of this variation in notation when using different literature.

The elements of sets can also be sets! An important example is the **power set** $\mathcal{P}(M)$ for a given set M, which consists of all subsets of M. That is,

$$\mathcal{P}(M) := \{A : A \subset M\}.$$

For example, consider once more the set $A = \{0, 1, 2\}$. The power set of A is

$$\mathcal{P}(A) = \{\emptyset, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, A\}.$$

Note that we always have $M \in \mathcal{P}(M)$ and $\emptyset \in \mathcal{P}(M)$.

We can also create new sets from given sets, say M and N, by using **set operations**.

The **union** $M \cup N$ contains all elements which belong to the set M and all elements which belong to the set N.

The **intersection** $M \cap N$ consists of all elements which are in both M and N.

The **set difference** of M and N, written as $M \setminus N$, is the set of all elements of M which are not contained in N.

If we only work with subsets $M \subset \Omega$ for a fixed set Ω, then we call Ω the **underlying set** or the **ground set**. In this case, the complement of M (in Ω), denoted M^c, is the set $M^c = \Omega \setminus M$.

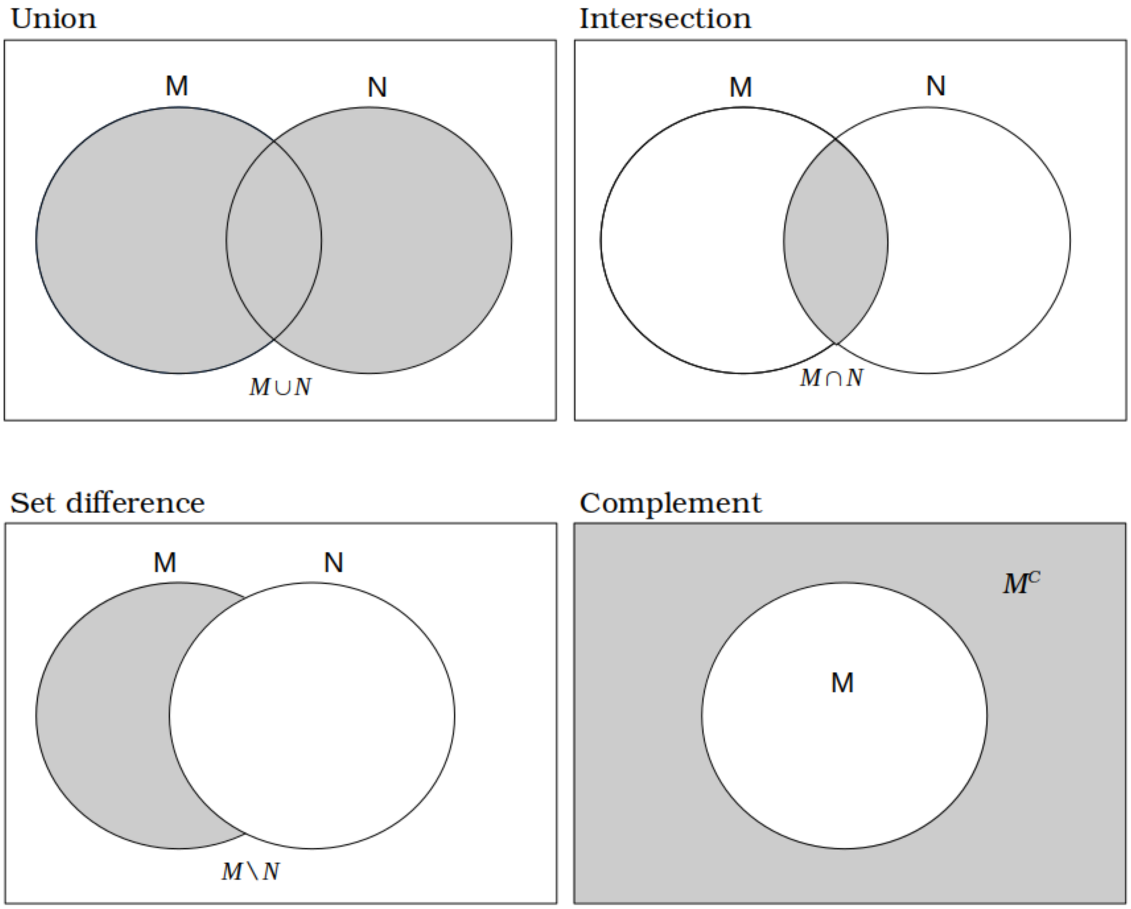


Figure 1: Venn-diagrams

The illustrations above are called **Venn-diagrams**. Such pictures can be a helpful tool for thinking about sets and how they interact with each other.

Example. Let $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and let $A, B \subset \Omega$ be the sets

$$A = \{2, 3, 4\}, \quad B = \{3, 5, 7, 9\}.$$

Then

$$\begin{aligned} A \cup B &= \{2, 3, 4, 5, 7, 9\} \\ A \cap B &= \{3\} \\ B \setminus A &= \{5, 7, 9\} \\ A^c &= \{1, 5, 6, 7, 8, 9, 10\} \\ B^c &= \{1, 2, 4, 6, 8, 10\}. \end{aligned}$$

All of the sets we have seen so far have elements that can be listed, even if in some cases the list is infinite. However, this is not the case with all sets. Another important family of sets to consider are **intervals**.

Definition 1.1. *Let $a, b \in \mathbb{R}$ such that $a \leq b$. Then we define the* ***closed interval*** *between a and b to be the set*

$$[a, b] := \{x \in \mathbb{R} \text{ such that } a \leq x \leq b\}.$$

The ***half open intervals*** *between a and b are the sets*

$$[a, b) := \{x \in \mathbb{R} \text{ such that } a \leq x < b\}$$

and

$$(a, b] := \{x \in \mathbb{R} \text{ such that } a < x \leq b\}.$$

The ***open interval*** *between a and b is the set*

$$(a, b) := \{x \in \mathbb{R} \text{ such that } a < x < b\},$$

Moreover, we write

$$[a, \infty) := \{x \in \mathbb{R} \text{ such that } x \geq a\}, \ (a, \infty) := \{x \in \mathbb{R} \text{ such that } x > a\}$$

and

$$(-\infty, a] := \{x \in \mathbb{R} \text{ such that } x \leq a\}, \ (-\infty, a) := \{x \in \mathbb{R} \text{ such that } x < a\}.$$

Elements of sets are not ordered, and so the sets $\{a, b\}$ and $\{b, a\}$ are the same. Nevertheless, it is often important to order the objects under consideration, and for this purpose we have the Cartesian product.

Definition 1.2. *Let A and B be sets, and let $a \in A$ and $b \in B$ be arbitrary elements.*

- *The expression (a, b), which is sensitive to order, is called an* ***ordered pair****.*
- *Two tuples (a, b) and (a', b') are equal if and only if $a = a'$ and $b = b'$.*
- *The set of all ordered pairs*

  $$A \times B := \{(a, b) : a \in A, b \in B\}$$

  is called the ***Cartesian product*** *of the sets A and B.*

Some remarks about Cartesian products are given below.

- In general $(a, b) \neq (b, a)$. In fact, the second part of the definition implies that $(a, b) = (b, a)$ if and only if $a = b$.
- An ordered pair (a, b) and a set $\{a, b\}$ are completely different objects.
- If we consider more than two sets, say $A_1, A_2, \ldots, A_d$ for some $d \in \mathbb{N}$, then we can also define the (d-fold) Cartesian product

  $$A_1 \times A_2 \times \cdots \times A_d := \{(a_1, \ldots, a_d) : a_i \in A_i \text{ for all } i = 1, \ldots, d\},$$

  whose elements $(a_1, \ldots, a_d)$ are called d-tuples.

- For the Cartesian product
$$\underbrace{A \times A \times \cdots \times A}_{d \text{ times}},$$
we use the shorthand A^d.

Example - Let $A = \{\pi, \sqrt{2}\}$ and $B = \{1, 2, 3\}$. Then

$$A \times B = \{(\pi, 1), (\pi, 2), (\pi, 3), (\sqrt{2}, 1), (\sqrt{2}, 2), (\sqrt{2}, 3)\}$$

and

$$B \times A = \{(1, \pi), (1, \sqrt{2}), (2, \pi), (2, \sqrt{2}), (3, \pi), (3, \sqrt{2})\}.$$

1.2 Propositional logic

Next, we will introduce some notation from logic that can be helpful for making precise mathematical statements. We begin with the **universal and existential quantifiers**. These are essentially just abbreviations.

- The notation $\forall$ means "for all".
- The notation $\exists$ means "there exists".

We also sometimes use the colon symbol ":" to denote "such that". With these notational conventions, we can also give a more concise description of some of the sets and concepts we discussed in the previous subsection. Here are some examples:

- The statement
$$\exists x \in \mathbb{N} \, : \, x \text{ is even}$$
simply says that there is at least one even natural number. This is a true statement, obviously!
- We can also use this notation to make false statements. For instance,
$$\forall x \in \mathbb{N}, \; x \text{ is even.}$$
- We have that $M \subset N$ if and only if
$$\forall x \in M, \, x \in N.$$
- We have that $M \subsetneq N$ if and only if $M \subset N$ and
$$\exists x \in N : x \notin M.$$

As you might have already noticed, we will often need the terms "if" or "if and only if", and therefore we define a mathematical symbol for them. In order to do this properly, we need to learn some basic formal logic.

Definition 1.3. *A **proposition** is a statement which is either true or false.*

We can use **connectives** to build more complicated propositions depending on two propositions A and B.

Definition 1.4. *Let A and B be propositions, then*

- *$\neg A$ (**not** A) is the **negation** of A,*
- *$A \wedge B$ (A **and** B) is the **conjunction** of A and B,*
- *$A \vee B$ (A **or** B) is the **disjunction** of A and B,*

Examples - Let A be the proposition "3 is a prime number" and let B be the proposition "$2^2 = 5$". So, A is true and B is false.

- $\neg A$ is the statement "3 is not a prime". This is false.
- $\neg B$ is the statement "$2^2 \neq 5$". This is true.
- $A \vee B$ is the statement "3 is a prime or $2^2 = 5$". This is true.
- $A \wedge B$ is the statement "3 is a prime and $2^2 = 5$". This is false.
- $A \wedge \neg B$ is the statement "3 is a prime and $2^2 \neq 5$". This is true.
- $A \vee \neg B$ is the statement "3 is a prime or $2^2 \neq 5$". This is true.

The last of these propositions highlights a difference between the meaning of "or" in mathematics and in common language. The proposition $A \vee B$ is true if both A and B are true also. In common language, when we say something like "please buy me some apples or oranges", we are expecting only one of the kinds of fruit to arrive (this is the exclusive or). The common language equivalent of the symbol $\vee$ is "and/or".

We sometimes use **truth tables** to see relations between truthfulness of propositions and their component parts. Here is an example of a truth table for the negation, conjunction, and disjunction.

A	B	$\neg A$	$A \wedge B$	$A \vee B$
$\mathbb{T}$	$\mathbb{T}$	$\mathbb{F}$	$\mathbb{T}$	$\mathbb{T}$
$\mathbb{T}$	$\mathbb{F}$	$\mathbb{F}$	$\mathbb{F}$	$\mathbb{T}$
$\mathbb{F}$	$\mathbb{T}$	$\mathbb{T}$	$\mathbb{F}$	$\mathbb{T}$
$\mathbb{F}$	$\mathbb{F}$	$\mathbb{T}$	$\mathbb{F}$	$\mathbb{F}$

The **truth value** of a proposition is either $\mathbb{T}$ or $\mathbb{F}$. We sometimes write $|A| = \mathbb{T}$ or $|A| = \mathbb{F}$ to indicate the truth value of A.

Definition 1.5. *Let A and B be two propositions. Then,*

- *"$A \implies B$" means "A **implies** B". In other words, if A is true then so is B.*
- *"$A \iff B$" means that both $A \implies B$ and $B \implies A$. In other words, A is true if and only if B is true.*

Using the notation for the conjunction, we see that the proposition $A \iff B$ is the same as $(A \implies B) \wedge (B \implies A)$. We also sometimes use "iff" as an abbreviation for "if and only if".

Examples - Here are some examples of true mathematical statements which involve implications.

- $A \subset \mathbb{N} \implies A \subset \mathbb{Z}$.
- $(x \in \mathbb{P}$ and $x > 2) \implies x$ is odd.
- $b^2 - 4ac > 0 \iff (ax^2 + bx + c = 0$ has two distinct real solutions).

For the three statements above, it seems to be intuitively clear that they are true. We said above that the implication $A \Rightarrow B$ means that "if A is true then B is true". However, what if A is not true? This situation can be rather confusing. Consider the following two statements:

$$\begin{aligned} (1 = 2) &\implies (\text{Every prime number is odd}), \\ (\pi \in \emptyset) &\implies (2 \text{ is even }). \end{aligned}$$

Although it is not so intuitively obvious, both of these statements are in fact also true. In fact, if A is false, then the statement $A \Rightarrow B$ is true, regardless of what B says! This can be confusing, since it can be used to build mathematical statements which are logically true but appear to be nonsensical, like these two above.

Truth tables may be helpful for clearing up potential confusion surrounding this issue. The truth table showing both directions of implication is given below.

A	B	$A \Rightarrow B$	$B \Rightarrow A$	$A \Leftrightarrow B$
$\mathbb{T}$	$\mathbb{T}$	$\mathbb{T}$	$\mathbb{T}$	$\mathbb{T}$
$\mathbb{T}$	$\mathbb{F}$	$\mathbb{F}$	$\mathbb{T}$	$\mathbb{F}$
$\mathbb{F}$	$\mathbb{T}$	$\mathbb{T}$	$\mathbb{F}$	$\mathbb{F}$
$\mathbb{F}$	$\mathbb{F}$	$\mathbb{T}$	$\mathbb{T}$	$\mathbb{T}$

Here are some remarks about what this truth table shows us.

- We can see from the truth table above that the only way that the statement $A \implies B$ is false is if A is true and B is false.
- The final column shows that the statement $A \iff B$ is true when A and B have the same truth value. If two propositions A and B have the same truth value then we say that they are **logically equivalent**, and write $A \equiv B$.

We can build longer truth tables and use them to compare other logical statements. Let us compare the statement $A \implies B$ with $\neg B \implies \neg A$.

A	B	$\neg A$	$\neg B$	$A \Rightarrow B$	$\neg B \Rightarrow \neg A$
$\mathbb{T}$	$\mathbb{T}$	$\mathbb{F}$	$\mathbb{F}$	$\mathbb{T}$	$\mathbb{T}$
$\mathbb{T}$	$\mathbb{F}$	$\mathbb{F}$	$\mathbb{T}$	$\mathbb{F}$	$\mathbb{F}$
$\mathbb{F}$	$\mathbb{T}$	$\mathbb{T}$	$\mathbb{F}$	$\mathbb{T}$	$\mathbb{T}$
$\mathbb{F}$	$\mathbb{F}$	$\mathbb{T}$	$\mathbb{T}$	$\mathbb{T}$	$\mathbb{T}$

We see that the truth values of the statements $A \implies B$ and $\neg B \implies \neg A$ are in fact identical. This means that the statements are logically equivalent; if one is true then the other is true, and if one is false then the other is false. Another way of writing this is

$$(A \implies B) \iff (\neg B \implies \neg A).$$

We say that the statement $\neg B \implies \neg A$ is the **contrapositive** of the statement $A \implies B$. There are times when it is very helpful to know that these two statements are equivalent to one another, particularly as it can help us to transfer a statement that is difficult to work with into an equivalent statement that may be easier to understand.

We use implications in mathematics to build a chain of logic, using existing statements to prove new ones. When we do this, we are using the *transitivity of implication.* This is the statement that, for propositions A, B and C,

$$[(A \implies B) \wedge (B \implies C)] \implies (A \implies C). \tag{2}$$

Exercise - Use a truth table to verify (2).

We often consider propositions which depend on one or more variables, such as the statements "x is even" or "$a \leq b$". The truth value of these statements depends on the choice of the variables x, a and b. A proposition which depends on a variable is called a **predicate**. We may use the notation $P(x)$ or $Q(a, b)$ for such predicates.

Example - Let $P(A)$ be the statement "$A \subset \{1, 2, 3, 4\}$". Then, for example, $P(\{1, 3\})$ and $P(\emptyset)$ are true, while $P(\mathbb{Z})$ is false. Let $Q(A, B)$ be the statement "$B \subset A$". Then $Q(\{1\}, \{1, 2\})$ is false and $Q(\mathbb{Z}, \mathbb{N})$ is true. Observe that

$$(P(A) \wedge (Q(A, B))) \implies (P(B)).$$

Finally, with all of this notation, we may write certain definitions or statements without using any words, but rather by exclusively using mathematical symbols. For example,

$$M \subset N \iff (\forall x \in M, x \in N) \iff (x \in M \Rightarrow x \in N).$$

The "sentence" above shows three different symbolic descriptions for the statement that M is a subset of N.

However, just because we can use these symbolic shorthand descriptions, it does not mean that we always should! Sometimes people forget that words themselves are a very valuable tool for describing mathematical ideas.

1.3 Relations and functions

Roughly speaking, relations shall describe connections between two objects. Here, we give a formal description and important properties. We then introduce functions; a special kind of relation which describe connections *from* one set *to* another. We also discuss special relations that are used to compare, group or order elements of a given set.

Definition 1.6. *A relation R between two sets M and N is a subset of the Cartesian product of M and N, i.e. $R \subset M \times N$.*

To make things clearer, see the upcoming illustration, which depicts every element of R as a "connection" between an element of M and an element of N. As you can see it is possible that $x \in M$ is connected to some $y \in N$, which we denote by $(x, y) \in R$. However, this does not have to be the case for every $x \in M$, and different elements of M may be connected to the same $y \in N$. Moreover, $x \in M$ can be connected to more than one element in N, or can even be connected to none of the elements of N.

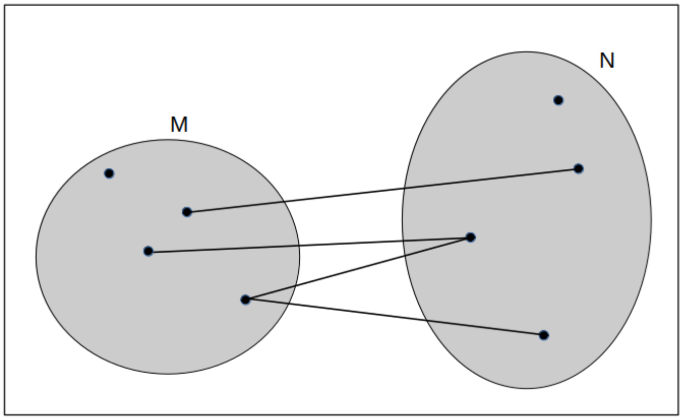


Figure 2: An illustration of a relation

Example - Let $M = \{2, 3, 4, 5\}$ and $N = \{4, 5, 6, 7\}$ and define a relation $R \subset M \times N$ whereby

$$(x, y) \in R \text{ iff } x \text{ or } y \text{ are prime.}$$

Then

$$R = \{(2,4),(2,5),(2,6),(2,7),(3,4),(3,5),(3,6),(3,7),(5,4),(5,5),(5,6),(5,7),(4,5),(4,7)\}.$$

Now we head to a very important type of relation, namely functions, which assign to each element of M exactly one element of N.

Definition 1.7. *Let M and N be non-empty sets. We call $f : M \to N$ a **function** from M to N, if each $x \in M$ is assigned exactly one element $f(x) \in N$.*

*M is called the **domain** of f and N is the **codomain** of f.*

Examples - The function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = x^2$$

is a function. It satisfies the key property that every element x in the domain is assigned a value $f(x)$.

Similarly, the function $g : \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = x + 1$$

is a function.

Definition 1.8. *Let $M, N \neq \emptyset$ and let $f : M \to N$ be a function from M to N.*

*For $S \subset M$, we define the **image** of S under f as*

$$f(S) := \{f(x) : x \in S\} \subset N,$$

*and the **range** of f as*

$$f(M) := \{f(x) : x \in M\} \subset N.$$

In other words, the range of f is the image of the whole domain under f.

*For $T \subset N$, we define the **preimage** of T under f by*

$$f^{-1}(T) := \{x : f(x) \in T\} \subset M.$$

Examples - Consider again the functions $f(x) = x^2$ and $g(x) = x + 1$ introduced above. Let $S \subset \mathbb{R}$ be the closed interval $S = [1, 3]$. Then

$$f(S) = [1, 9], \text{ and } g(S) = [2, 4].$$

The range of f is $[0, \infty)$, while the range of g is $\mathbb{R}$.

Since S is also a subset of the codomain of both f and g, we can also consider its preimage in each case. We obtain

$$f^{-1}(S) = [-\sqrt{3}, -1] \cup [1, \sqrt{3}]$$

and

$$g^{-1}(S) = [0, 2].$$

The next definition shows the connection between relations and functions.

Definition 1.9. *Let $f : M \to N$ be a function. We define the **graph** of f as*

$$G_f := \{(x, f(x)) : x \in M\} \subset M \times N.$$

Note that the graph of a function is a relation. In this sense, all functions induce a relation, but not vice versa.

We can visualize a function with domain and codomain in $\mathbb{R}$ by plotting its graph in $\mathbb{R}^2$. For $f(x) = x^2$ and $g(x) = x + 1$ this is demonstrated in the next illustration (Figure 3).

In what follows, we will define several important properties of relations. We will always demonstrate afterwards what this means for functions.

Definition 1.10. *Let $R \subset M \times N$ be a relation.*

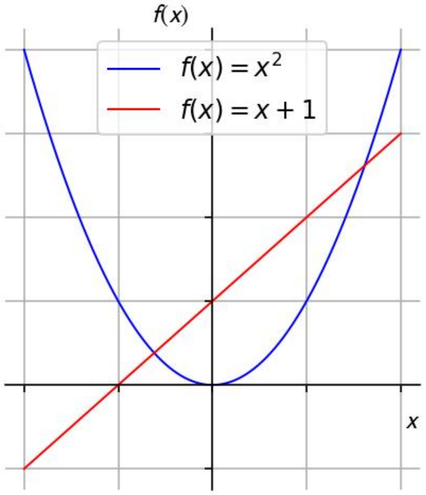


Figure 3: The graph of x^2 and $x + 1$

- *R is* ***injective*** *if and only if*

$$\forall\,(x_1, y_1), (x_2, y_2) \in R,\ x_1 \neq x_2 \Rightarrow y_1 \neq y_2.$$

This is equivalent to

$$\forall (x_1, y_1), (x_2, y_2) \in R,\ y_1 = y_2 \Rightarrow x_1 = x_2.$$

- *R is* ***surjective*** *if and only if*

$$\forall y \in N,\ \exists x \in M\ : (x, y) \in R.$$

- *R is* ***bijective*** *if and only if it is injective and surjective.*

- *R is* ***functional*** *if and only if*

$$\forall x \in M,\ \exists\, y \in N\ : (x, y) \in R$$

and

$$\forall x \in M, y_1, y_2 \in N,\ ((x, y_1), (x, y_2) \in R) \Rightarrow y_1 = y_2.$$

The two requirements for a relation to be functional can be written more succinctly as follows:

$$\forall x \in M,\ \exists!\, y \in N\ : (x, y) \in R.$$

Note that the graph of a function is a functional relation, and vice versa. We can therefore rephrase the above definitions for functions.

Definition 1.11. *Let $f : M \to N$ be a function.*

- *We say f is* ***injective*** *if and only if*

$$\forall x_1, x_2 \in M,\, x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2).$$

- *We say f is* ***surjective*** *if and only if*

$$\forall y \in N,\, \exists x \in M\ : f(x) = y.$$

- *We say f is* ***bijective*** *if and only if*

$$\forall y \in N,\, \exists! x \in M\ : f(x) = y.$$

Let us see some illustrations for better understanding.

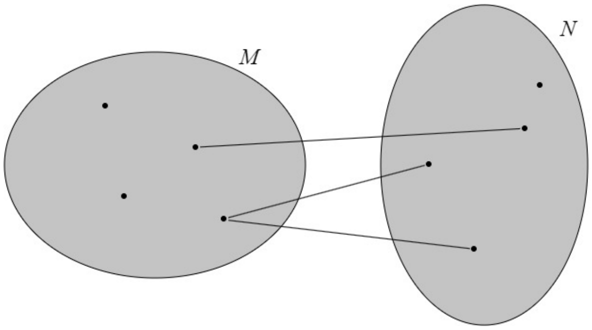


Figure 4: injective relation

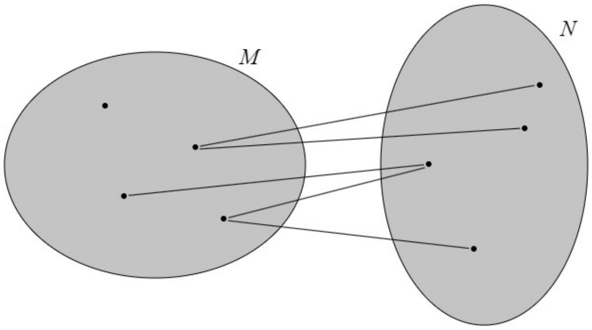


Figure 5: surjective relation

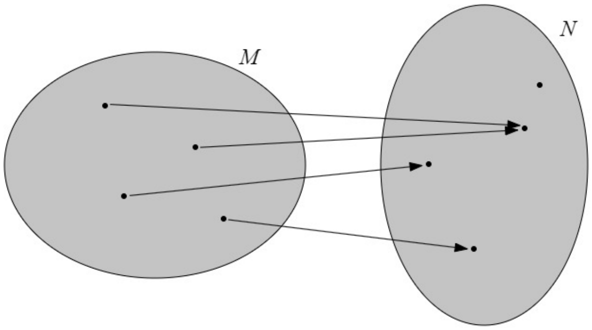


Figure 6: a function (not injective and not surjective)

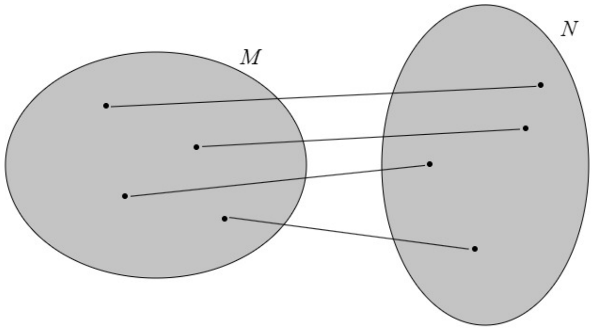


Figure 7: bijective function

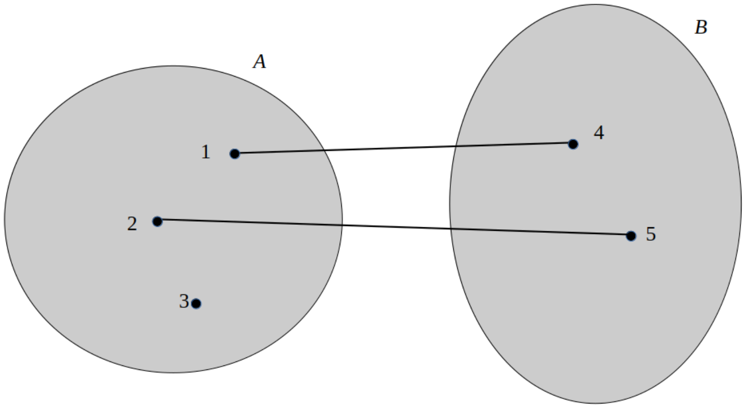


Figure 8: bijective relation but not a function

We will now define two simple functions that can be defined on an arbitrary set M. First, we define the **identity function** $Id_M : M \to M$ which maps each element to itself, i.e., $Id_M(x) = x$.

A constant function is a function which takes the same value for every element. Let M and N be arbitrary non-empty sets and let $c \in N$ be fixed. The function $f : M \to N$, which is defined by

$$f(x) = c,$$

for all $x \in M$, is called a **constant function**.

We now ask ourselves the following question: given a function $f : M \to N$, can we find another function which reverses the effect of f. This leads us to the concept of an inverse function.

Definition 1.12. *Let $f : M \to N$ and $g : N \to M$ be functions with the properties*

$$\forall x \in M, \, g(f(x)) = x$$

and

$$\forall y \in N, \, f(g(y)) = y.$$

*Then f is the **inverse** of g and g is the **inverse** of f. In this case we write $f^{-1} := g$ and $g^{-1} := f$ and call f and g **invertible**.*

Note that we already used the notation f^{-1} in the context of the preimage of a set under a function f. This similarity in notation is intentional, as the following exercise indicates.

Exercise - Let $f : M \to N$ be an invertible function with inverse $f^{-1} : N \to M$ and $S \subset N$. Prove that

$$\{f^{-1}(y) : y \in S\} = \{x \in M : f(x) \in S\}. \tag{3}$$

Let us revisit Definition 1.8. The left hand side of (3) is the image of S under f^{-1}, which is denoted $f^{-1}(S)$. The right hand side is the preimage of S under f, which is also denoted $f^{-1}(S)$. So, the exercise shows that these two notational conventions do not clash with one another.

Example Let $\mathbb{R}_+$ denote the set of all non-negative real numbers (i.e. $\mathbb{R}_+ = [0, \infty)$) and let $f : \mathbb{R}_+ \to \mathbb{R}_+$ and $g : \mathbb{R}_+ \to \mathbb{R}_+$ be defined by

$$f(x) = x^2, \quad \text{and} \quad g(x) = \sqrt{x}.$$

For fixed $x > 0$, we have $g(f(x)) = \sqrt{x^2} = x$. On the other hand, for all $y > 0$ we have $f(g(y)) = (\sqrt{y})^2 = y$. Therefore, f and g are inverses of each other.

The following theorem provides us with a tool to check if a function has an inverse or not.

Theorem 1.13. *Let $f : M \to N$ be a function. Then,*

$$f \text{ is invertible} \iff f \text{ is bijective.}$$

This is the first theorem we have seen in this course, and it will be the first of many. We are trying to build a rigorous mathematical foundation in this course, and so we will (with a few exceptions) give a *proof* of every result we use during the course. The skill of understanding and writing mathematical proofs is rather specialised, and is likely to be unfamiliar to many students of this course. We will devote special attention to proof techniques in a section that will be shortly forthcoming, and will include the proof of Theorem 1.13.

Invertible (or bijective) functions may be used to formally define the **cardinality** of a set.

Definition 1.14. *A finite set M containing n elements has* ***cardinality*** *n. We write $|M| = n$.*

In other words, the cardinality of a finite set is simply its size.

Note that the existence of a bijective function $f : M \to N$ means that there is a **one-to-one correspondence between M and N**. In particular, both sets must have the same cardinality. Moreover, the next exercise shows how we can compare the cardinality of sets by looking at the properties of functions that exist between them.

Exercise - Show that, for two finite sets A and B, the following statements are true.

- $|A| = |B|$ if and only if there is a bijection $f : A \to B$.
- $|A| \leq |B|$ if and only if there is an injection $f : A \to B$.
- $|A| \geq |B|$ if and only if there is a surjection $f : A \to B$.

As well as finite sets, bijections also allow us (to some extent) to characterise the cardinality of an infinite set.

Definition 1.15. *Let A be a set.*

- *If there exists $n \in \mathbb{N}$ such that $|A| = n$, then we call A **finite**, or a **finite set**.*
- *If A is not finite, then we call A **infinite**, or an **infinite set**.*
- *If there exists a bijection $f : \mathbb{N} \to A$, then we call A **countably infinite**.*
- *If A is either finite or countably infinite then we call A **countable**.*
- *If A is not countable, then we call A **uncountable**.*

Note that countability is the precise definition of the "simple" property that the elements of A can be enumerated by the natural numbers $\{1, 2, 3, \dots\}$, which is a fancy way to say, the elements of A can be counted.

Examples - Define a function $f : \mathbb{N} \to \mathbb{Z}$ such that, for all $n \in \mathbb{N}$,

$$f(2n) := n, \text{ and } f(2n-1) := -n+1.$$

This is a bijection. Indeed, $f(1) = 0$, $f(2) = 1$, $f(3) = -1$, $f(4) = 2$, $f(5) = -2$, and so on. We can see from this pattern that every element of $\mathbb{Z}$ is mapped to by exactly one element of $\mathbb{N}$. By Definition 1.15, it follows that $\mathbb{Z}$ is a countable set.

Note the following strange feature of this example. Intuitively, it seems that the set $\mathbb{Z}$ is "bigger" than $\mathbb{N}$. Indeed, $\mathbb{N}$ is a proper subset of $\mathbb{Z}$, and appears to contain half as many elements. Still, with the notion of size (i.e. cardinality) given by Definition 1.15, the two sets have the same size; they are both countably infinite.

Exercise - Show that the set $E = \{2, 4, 6, \dots\}$ is countably infinite.

Exercise - Let M and N be finite sets. Show that $|M \times N| = |M||N|$.

We can also consider the **composition** of functions. Let X, Y, Z be non-empty sets and let $f : X \to Y$ and $g : Y \to Z$ be functions. We then define a function $(g \circ f) : X \to Z$ by first applying f and then applying g. That is, we define $(g \circ f)(x) = g(f(x))$.

Exercise - Check that $g \circ f$ is indeed a function.

Note that it is important that the codomain of f matches the domain of g in order for the definition of $g \circ f$ to make sense.

Example - If $f : M \to N$ is an invertible function and $f^{-1} : N \to M$ is its inverse, then, for all $x \in M$ and $y \in N$, we have

$$f^{-1}(f(x)) = x, \text{ and } f(f^{-1}(y)) = y.$$

In particular, $f \circ f^{-1} = Id_N$ and $f^{-1} \circ f = Id_M$.

Example - Consider the functions $f, g, h : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = x^2, \quad g(x) = \sin(x) \quad \text{and} \quad h(x) = \sin(x^2).$$

We get

$$(g \circ f)(x) = g(f(x)) = \sin(x^2) = h(x).$$

However, if we reverse the order of composition and consider the function $f \circ g$, we see that

$$f(g(x)) = (\sin x)^2.$$

In particular the functions $f \circ g$ and $g \circ f$ are not the same. Indeed, in a typical case we have $f \circ g \neq g \circ f$, and it can even be the case that only one of the compositions is defined.

The next theorem shows how to calculate the inverse of a composition of two functions.

Theorem 1.16. *Let $f : X \to Y$ and $g : Y \to Z$ be invertible functions. Then $g \circ f$ is invertible and*

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

We will prove Theorem 1.16 in the forthcoming introductory section on proof.

Finally, let us discuss some relations that are used to **compare, group or order** elements of a set M . For this we consider relations $R \subset M^2$. Such relations usually have nothing to do with functions, but are still essential. Again, relations with some particularly important characteristics are given special names.

Definition 1.17. *Let $R \subset M \times M$ be a relation for an arbitrary $M \neq \emptyset$. We call R*

- ***reflexive*** *if and only if*
$$\forall\, x \in M,\ (x, x) \in R,$$
- ***symmetric*** *if and only if*
$$(x, y) \in R \Rightarrow (y, x) \in R,$$
- ***antisymmetric*** *if and only if*
$$(x, y), (y, x) \in R \Rightarrow x = y,$$
- ***transitive*** *if and only if*
$$(x, y), (y, z) \in R \Rightarrow (x, z) \in R,$$
- ***total*** *if and only if*
$$\forall x, y \in M,\ (x, y) \in R \text{ or } (y, x) \in R.$$

Some relations which have certain combinations of these properties are especially important.

Definition 1.18.

- *An **equivalence relation** is a relation that is reflexive, symmetric and transitive.*
- *A **partial order** is a relation that is reflexive, antisymmetric and transitive.*
- *A **linear order** is a relation that is a partial order and total.*

Example - Consider the relation $L \subset \mathbb{R}^2$, where we define

$$(x, y) \in L \Leftrightarrow x \leq y.$$

This is the well-known smaller or equal relation in $\mathbb{R}$ and we just write $x \leq y$ for $(x, y) \in L$ in the following.

- This relation is reflexive, since $x \leq x$ holds for all $x \in \mathbb{R}$.
- This relation is antisymmetric, since $x \leq y$ and $y \leq x$ both holding implies that $x = y$.
- This relation is transitive, since $x \leq y$ and $y \leq z$ implies that $x \leq z$.
- This relation is total: for all $x, y \in \mathbb{R}$, at least one of $x \leq y$ or $y \leq x$ holds.

Therefore, L is a linear order.

Example - Consider the usual equality relation in $\mathbb{R}$. That is, define

$$L = \{(x, x) : x \in \mathbb{R}\} \subset \mathbb{R}^2.$$

Then L is reflexive, symmetric, antisymmetric and transitive, but not total. It is therefore an equivalence relation and a partial order.

Exercise - Show that the relation L from the previous example is the only relation that is symmetric and antisymmetric.

Example - Define the "strictly less" relation "<" by

$$L = \{(a, b) \in \mathbb{R}^2 : a < b\}.$$

We seek to determine which of the characteristics defined in Definition 1.17 L has.

- L is not reflexive, since, for example, $1 < 1$ does not hold. (Of course, we know that $x < x$ is never true, but in order to show that reflexivity does not hold, we only need to find a single instance of a counterexample.)
- L is not symmetric. For instance $1 < 2$ holds, but $2 < 1$ is false.
- L *is* antisymmetric. We will discuss this further below.
- L is transitive, since $x < y$ and $y < z$ implies that $x < z$.
- L is not total. Once again, this is because, for example, $1 < 1$ does not hold.

Let us now verify that L is antisymmetric, as claimed above. This is an instance of a potentially confusing feature of implication that we discussed earlier. We want to verify that the implication

$$(x < y \text{ and } y < x) \Rightarrow x = y$$

is a valid statement. However, the left hand side of this implication is a false statement, and false statements imply everything!

This is an instance of a more general situation that occurs in several mathematical proofs. Suppose that $P(x)$ is a statement that depends on some real number x, and we want to verify that

$$x \in A \Rightarrow P(x) \text{ is true.} \tag{4}$$

In the particular case when $A = \emptyset$, the implication (4) is certainly true, because the left hand side is certainly false.

Exercise - Let m and n be integers. We say that m **divides** n, and write $m|n$ if there exists $k \in \mathbb{Z}$ such that

$$mk = n.$$

Show that the divisibility relation

$$R := \{(a, b) \in \mathbb{N} \times \mathbb{N} : a|b\}$$

is a partial order.

Example - One can define a partial order on a set of sets by using the subset relation $\subset$. For example, for

$$M := \{\emptyset, \{1\}, \{2\}, \{1, 2\}\},$$

we define the relation

$$L = \{(A, B) \in M \times M : A \subset B\}.$$

It follows from the basic properties of inclusion that the relation L is reflexive, antisymmetric and transitive. Hence it is a partial order on M . However, since $\{1\} \not\subset \{2\}$ and $\{2\} \not\subset \{1\}$, it is not total, and so not a linear order on M.

For the remainder of this subsection, we will pay special attention to equivalence relations. Equivalence relations have the particularly useful and interesting property that they can be used to partition the elements of a set into related chunks. These partitioning sets are called equivalence classes.

Definition 1.19. *Let $R \subset X \times X$ be an equivalence relation. For $x \in X$, we define the **equivalence class of** x by*

$$[x]_R := \{y \in X : (x, y) \in R\}.$$

When it is clear which relation we are referring to (which it usually is in practice) we abbreviate $[x]_R$ to $[x]$.

Each element $y \in [x]$ is called a **representative** of the equivalence class $[x]$. Note that, by reflexivity $(x, x) \in R$, and hence $x \in [x]$. So an equivalence class is never the empty set and x is a representative of $[x]$.

Example - Consider again the equivalence relation

$$L = \{(x, x) : x \in \mathbb{R}\} \subset \mathbb{R}^2.$$

Then, for each $x \in \mathbb{R}$, we have $[x] = \{x\}$, and so the equivalence classes for this relation are all singleton sets (i.e. sets with exactly one element).

Example - Fix an arbitrary $m \in \mathbb{N}$. We say that two integers $a, b \in \mathbb{Z}$ are **congruent modulo** m, if

$$m|(a - b)$$

and we write

$$a \equiv b \mod m.$$

Define

$$R = \{(a, b) \in \mathbb{Z} : a \text{ and } b \text{ are congruent modulo } m\}.$$

Then R is an equivalence relation. Indeed, let us check that the necessary three properties hold.

- Reflexivity: Let $a \in \mathbb{Z}$. Then m divides $a - a = 0$, and so $(a, a) \in R$.
- Symmetry: Let $a, b \in \mathbb{N}$ such that $(a, b) \in R$. That is, $m|a - b$. In other words, there is some $k \in \mathbb{Z}$ such that

  $$a - b = mk.$$

  But then

  $$b - a = m(-k),$$

  which means that m divides $b - a$ and thus $(b, a) \in R$.
- Transitivity: Let $a, b, c \in \mathbb{Z}$ such that $(a, b), (b, c) \in R$. By the definition of divisibility, there exist integers k_1, k_2 such that

  $$a - b = mk_1, \quad \text{and} \quad b - c = mk_2.$$

  Therefore,

  $$a - c = (a - b) + (b - c) = mk_1 + mk_2 = m(k_1 + k_2).$$

  This means that m divides $a - c$, and thus $(a, c) \in R$.

Note that, in both of the examples above, the equivalence classes do not overlap with one another, and every element of the underlying set belongs to exactly one equivalence class. This is not a coincidence; the following theorem shows that equivalence classes always partition the underlying set in such a way.

Theorem 1.20. *Let $R \subset X \times X$ be an equivalence relation and $x, y \in X$. Then,*

$$(x, y) \in R \iff [y] = [x] \iff [x] \cap [y] \neq \emptyset.$$

We will prove Theorem 1.20 in the forthcoming subsection "Introduction to Proof".

In particular, the last equivalence

$$[y] = [x] \iff [x] \cap [y] \neq \emptyset$$

can be rewritten as

$$[y] \neq [x] \iff [x] \cap [y] = \emptyset.$$

Therefore, we indeed have that two distinct equivalence classes must be disjoint.

1.4 Real numbers

The natural starting point when thinking about numbers is the set $\mathbb{N} = \{1, 2, 3, \dots\}$. These are the first numbers we learn about as small children, and it feels right that they are called the natural numbers. However, this alone is not enough to solve all of the equations that we encounter using simple mathematical operations like addition, subtraction, multiplication, and division. And so, we gradually extend our universe of numbers in order to be able to deal with more of these simple mathematical questions. Continuing in this way, we soon reach the set of *real numbers*.

Let's give a little more detail about this vague idea described in the paragraph above. The set $\mathbb{N}$ is closed under addition and multiplication. That is, for all $n, m \in N$ we have $m + n \in \mathbb{N}$ and $m \cdot n \in \mathbb{N}$. However, we already get in trouble when we try to work with subtraction, since, for example $21 - 42 = -21 \notin \mathbb{N}$.

The set $\mathbb{Z}$ of **integers** is closed under addition and multiplication, but is additionally closed under subtraction. That is, for any $a, b \in \mathbb{Z}$ we have $a + b, ab, a - b \in \mathbb{Z}$. However, division is still a problem if we use integer numbers only. For instance 1 and -3 are both in $\mathbb{Z}$, but the ratio $\frac{1}{-3}$ is not.

We therefore define the set of all **rational numbers**

$$\mathbb{Q} := \left\{ \frac{a}{b} : a, b \in \mathbb{Z}, b \neq 0 \right\},$$

where we call a the **numerator** and b the **denominator**. It is possible that the same rational number can have two different representations. For instance, $\frac{1}{2}$ and $\frac{2}{4}$ are the same number. We call two rational numbers $\frac{k}{n}$ and $\frac{l}{m}$ equal if and only if $km = ln$ with $m, n \neq 0$.

It is sometimes convenient to exclude 0 from $\mathbb{Q}$ in order to avoid the problem of potentially dividing by zero, and we use the notation $\mathbb{Q}^* := \mathbb{Q} \setminus \{0\}$. We now have a set which is closed under division. Indeed, if we take two elements $\frac{a}{b}$ and $\frac{c}{d}$ in $\mathbb{Q}^*$ (with $a, b, c, d \in \mathbb{Z} \setminus \{0\}$), then the ratio

$$\frac{a/b}{c/d} = \frac{ad}{bc}$$

is an element of $\mathbb{Q}^*$.

Note also that any integer $k \in \mathbb{Z}$ can be represented as an element of $\mathbb{Q}$ by writing $\frac{k}{1} = k$. Consequently, we can expand the earlier chain of inclusions (1) to

$$\mathbb{N} \subsetneq \mathbb{N}_0 \subsetneq \mathbb{Z} \subsetneq \mathbb{Q}. \tag{5}$$

The real reason that we have introduced "new" sets of numbers above is that we wanted to solve certain equations and, in particular, we want to know if a solution exists in a given set. For example, if we want to solve the equation $a \cdot x + b = 0$, where $a, b \in \mathbb{Q}$ are fixed constants and $a \neq 0$, we see that

$$x = \frac{-b}{a} \in \mathbb{Q}$$

and so we can find a solution to the equation in $\mathbb{Q}$.

However, what about the simple equation $x^2 - 2 = 0$? Is there some $x \in \mathbb{Q}$ such that $x^2 = 2$? The following theorem will show that this is not possible.

Theorem 1.21. *There is no $x \in \mathbb{Q}$ such that $x^2 = 2$.*

We will come back and prove this theorem in the next subsection, using the method of proof by contradiction.

Theorem 1.21 shows that the equation $x^2 - 2 = 0$ is not solvable in $\mathbb{Q}$. But we would like there to be a solution somewhere! Furthermore, we all know from school that there is a number $\sqrt{2}$ such that $(\sqrt{2})^2 = 2$.

Making the number line complete, we finally get to the set of **real numbers**. Using the familiar decimal expansion, we can define the set of real numbers to be

$$\mathbb{R} := \left\{ b + \frac{a_1}{10} + \frac{a_2}{100} + \cdots : b \in \mathbb{Z}\,, a_1, a_2 \cdots \in \{0,1,2,3,4,5,6,7,8,9\} \right\}. \tag{6}$$

One may think of $\mathbb{R}$ as the set of all points on the number line, i.e., the number line without any holes. Note that there are infinitely many terms in the sum in (6) defining an element of $\mathbb{R}$. It is possible that, after a certain point, all of the a_i are equal to zero. For instance, an integer n can be expressed in this form, with *all* of the a_i equal to zero. More generally, the rational numbers, written using the decimal expansion as in (6), either have a finite number of digits or the sequence of digits is periodic. This means that, if we consider only the rational numbers, some points on the line are missing. These correspond to numbers which have a non-periodic infinite number of decimals, and cannot be written as fractions. As well as $\sqrt{2}$, some other prominent examples of such numbers are π and e. These elements of $\mathbb{R} \setminus \mathbb{Q}$ are called **irrational numbers**.

We can extend the hierarchy described in (5) to include $\mathbb{R}$ as follows:

$$\mathbb{N} \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{R}.$$

It turns out that the set of rational numbers is countable, whereas the set of real numbers is uncountable. So, there are "more" irrational numbers than there are rationals. The task of proving this statement will be considered in a future exercise sheet.

The set of real numbers satisfies several convenient arithmetic and algebraic properties. These properties turn out to be useful and interesting in a much more general setting, which is why we give the following definition.

Definition 1.22. *Let $\mathbb{F}$ be a set with operations of addition (formally, addition is a function $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$) and multiplication (formally, multiplication is a function $\cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$) defined over $\mathbb{F}$ which satisfy the following properties.*

- ***Commutativity**: for all $x, y \in \mathbb{F}$, $x + y = y + x$ and $x \cdot y = y \cdot x$.*
- ***Associativity**: for all $x, y, z \in \mathbb{F}$, $(x + y) + z = x + (y + z)$ and $(x \cdot y) \cdot z = x \cdot (y \cdot z)$.*
- ***Distributivity**: for all $x, y, z \in \mathbb{F}$, $x \cdot (y + z) = (x \cdot y) + (x \cdot z)$.*
- ***Identity elements**: there exists two distinct elements $0, 1 \in \mathbb{F}$ such that, for all $x \in \mathbb{F}$*
    1. *$x + 0 = 0 + x = x$ and*
    2. *$x \cdot 1 = 1 \cdot x = x$.*

- ***Inverse elements***:

  1. *for all* $x \in \mathbb{F}$, *there exists* $y \in \mathbb{F}$ *such that* $x + y = y + x = 0$, *and*
  2. *for all* $x \in \mathbb{F} \setminus \{0\}$, *there exists* $y \in \mathbb{F}$ *such that* $x \cdot y = y \cdot x = 1$.

Then we call $\mathbb{F}$ *a* ***field***.

All of these properties hold for the set of reals, as we can see by using elementary rules for manipulating algebraic equations, and so $\mathbb{R}$ is a field.

Exercise - With the usual operations of addition and multiplication, is $\mathbb{Q}$ a field? How about $\mathbb{Z}$?

Exercise - Suppose that $\mathbb{F}$ is a field. Show that, for all $x, y, z \in \mathbb{F}$,

$$(x + y) \cdot z = (x \cdot z) + (y \cdot z).$$

1.5 An introduction to proof

The concept of rigorous proof is absolutely vital in mathematics. Unlike other branches of the sciences, we cannot be satisfied with evidence that something is very likely to be true, or true beyond reasonable doubt, as one might accept in a court of law. Rather, we need to know that something is true (or false) with absolutely certainty.

For instance, the Goldbach Conjecture states that every even integer greater than 2 can be written as a sum of two prime numbers. We can get some intuition for this conjecture by writing down some examples:

$$\begin{aligned} 4 &= 2+2, \\ 6 &= 3+3, \\ 8 &= 3+5 \\ 10 &= 5+5 \\ &\vdots \\ 100 &= 3+97. \end{aligned}$$

This list provides some initial evidence that the conjecture is correct. In fact, a whole load more evidence exists; the conjecture has been verified, with the help of computers, for all even integers n up to $4 \cdot 10^{18}$. That is a lot of cases checked!

Still, mathematicians continue to seek a proof of the Goldbach conjecture that guarantees that the statement is valid for *all* even n, with absolute certainty. A great deal of effort has been invested in this pursuit, including efforts from many great mathematicians spanning several generations, and a million dollar prize was even offered at one point.

In this subsection, we will discuss some basic techniques that allow us to prove some of the statements that were given earlier in Section 1, and in doing so develop some tools that will allow us to prove many interesting statements during the remainder of the course. We will focus on three particular techniques: proof by definition, proof by contradiction, and proof by induction.

We begin with **proof by definition**. These are often the easiest kinds of proofs, and the technique essentially consists of stringing together some basic properties that follow from definitions. This is another advertisement for the importance of knowing all of the definitions in this course!

We begin with a fairly simple proof that uses only the definitions involved. The following lemma (a lemma is like a small theorem, and usually is used to prove a more important result) will be helpful in proving that $\sqrt{2}$ is irrational. We make use of the obvious definition of an even integer; $n \in \mathbb{Z}$ is even iff there exists $k \in \mathbb{Z}$ such that $n = 2k$.

Lemma 1.23. *Let $n \in \mathbb{N}$. Then*

$$n \text{ is even} \iff n^2 \text{ is even.}$$

Proof. This is a two-way implication, and so we need to prove both directions. We begin by proving the "⇒" direction. Suppose that n is even. Then, by the definition of evenness, we can write $n = 2k$ for some $k \in \mathbb{N}$. Therefore,

$$n^2 = (2k)^2 = 4k^2 = 2(2k^2).$$

Now we prove the "⇐" direction. In fact we will prove the equivalent contrapositive statement

$$n \text{ is odd} \implies n^2 \text{ is odd}.$$

Suppose that n is odd, and so $n = 2k + 1$ for some $k \in \mathbb{N}_0$. But then

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

Thus n^2 is odd, by definition, and the proof is complete. □

For the next instance of these methods, let us restate and then prove Theorem 1.13.

Theorem 1.24. *Let $f : M \to N$ be a function. Then,*

$$f \text{ is invertible} \iff f \text{ is bijective.}$$

Proof. We are trying to prove that an " $\iff$ " holds, and so we need to prove two implications. We begin with the "⇒" direction. We assume that f is invertible, and we need to show that it is a bijection.

We begin by using the definition of invertibility. Since f is invertible, there exists $f^{-1} : N \to M$ such that

$$f^{-1}(f(x)) = x, \ \ \forall\, x \in M, \tag{7}$$

and

$$f(f^{-1}(y)) = y, \ \ \forall\, y \in N. \tag{8}$$

To show that f is bijective, we need to show that it is surjective and injective. We can take care to write down the definitions that tell us exactly what we need to show. We begin with surjectivity. We need to show that,

$$\forall\, y \in N, \, \exists\, x \in M \ : f(x) = y.$$

By (8), we can take $x = f^{-1}(y) \in M$ and see that $f(x) = y$.

Secondly we verify that f is injective. Let us again recall the definition of the property we are trying to prove. f is injective if we have

$$\forall\, x_1, x_2 \in M, \, x_1 \neq x_2 \implies f(x_1) \neq f(x_2).$$

It is easier to work with the logically equivalent contrapositive of this statement. That is, we will show that

$$\forall\, x_1, x_2 \in M, \, f(x_1) = f(x_2) \implies x_1 = x_2.$$

Let $f(x_1) = f(x_2)$ for some $x_1, x_2 \in M$. Since $f(x_1) \in N$ and $f(x_2) \in N$ we may apply the function f^{-1} to $f(x_1) = f(x_2)$ and get

$$x_1 = f^{-1}(f(x_1)) = f^{-1}(f(x_2)) = x_2,$$

as required.

We now move to the "⇐" direction. We assume that f is bijective and need to prove that f is invertible. Since f is bijective we have

$$\forall\, y \in N, \, \exists!\, x \in M \ : \ f(x) = y. \tag{9}$$

In words, for each $y \in N$ we can find a unique $x \in M$ such that $f(x) = y$. Now we define $g : N \to M$ such that it maps each $y \in N$ to this unique $x \in M$, i.e. $g(y) = x$, where x is given by (9). This shows that, for each $y \in N$, we have $f(g(y)) = y$. Indeed,

$$f(g(y)) = f(x) = y.$$

We also need to check that, for all $x \in M$

$$g(f(x)) = x.$$

But this is an immediate consequence of the way that the function g has been defined. Therefore, f is invertible with $f^{-1} = g$. □

Exercises

- Let $f : M \to N$ be a function. Prove that
$$\forall x \in M,\ (f \circ Id_M)(x) = f(x) = (Id_N \circ f)(x). \tag{10}$$
- Prove that composition of functions is associative. That is, for any functions
$$f : M \to N,\ g : N \to O, \text{ and } h : O \to P,$$
prove that, $\forall\, x \in M$,
$$(h \circ (g \circ f))(x) = ((h \circ g) \circ f)(x).$$
- Show that the inverse function is unique. That is, let $f : M \to N$ be an invertible function and suppose that $g_1, g_2 : N \to M$ are inverses of f. Prove that $g_1 = g_2$.

Next, we prove Theorem 1.16, which relates inverses and compositions. We restate and prove the result now.

Theorem 1.25. *Let $f : X \to Y$ and $g : Y \to Z$ be invertible functions. Then $g \circ f$ is invertible and*

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

Proof. We will directly check that $f^{-1} \circ g^{-1}$ is the inverse of $g \circ f$. Recalling the definition of the inverse of a function, there are two things we need to check:

$$(f^{-1} \circ g^{-1})((g \circ f)(x)) = x \quad \forall x \in X \tag{11}$$

and

$$(g \circ f)((f^{-1} \circ g^{-1})(z)) = z \quad \forall z \in Z. \tag{12}$$

Let us first focus on the goal of checking that (11) holds. This can be rewritten as

$$((f^{-1} \circ g^{-1}) \circ (g \circ f))(x) = x. \tag{13}$$

By the associativity of function composition (see the previous exercise) applied twice,

$$\begin{aligned}((f^{-1} \circ g^{-1}) \circ (g \circ f))(x) &= (f^{-1} \circ (g^{-1} \circ (g \circ f)))(x) \\ &= (f^{-1} \circ ((g^{-1} \circ g) \circ f))(x) \\ &= (f^{-1} \circ (Id_Y \circ f))(x).\end{aligned}$$

By (10), it follows that

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

which proves (13) (and thus also (11).

The proof of (12) is similar and is left as an exercise.

□

We continue to collect some proofs of statements that were given earlier in the course. The proof of the next statement (stated earlier as Theorem 1.20) also uses only the relevant definitions.

Theorem 1.26. *Let $R \subset X \times X$ be an equivalence relation and $x, y \in X$. Then,*

$$(x, y) \in R \iff [x] = [y] \iff [x] \cap [y] \neq \emptyset.$$

Proof. We will prove the following chain of implications:

$$(x, y) \in R \implies [x] = [y] \implies [x] \cap [y] \neq \emptyset \implies (x, y) \in R.$$

Once we have completed this cycle of implication, we can see that all three of the statements imply one another by moving around the cycle.

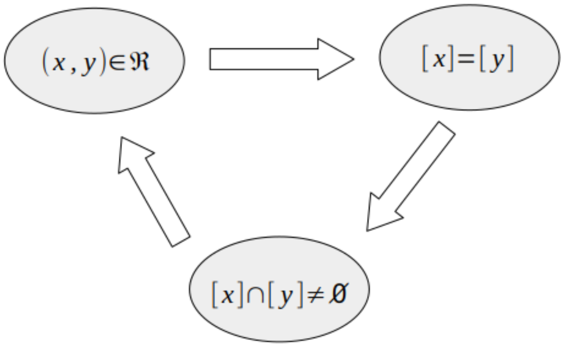


Figure 9: Chain of implications

Proof that $(x, y) \in R \implies [x] = [y]$. Suppose that $(x, y) \in R$. We will first show that $[x] \subset [y]$. Let $z \in [x]$. By definition $(x, z) \in R$. By the symmetry of R, we also know that $(y, x) \in R$. Then, by the transitivity of R, we have

$$(y, x) \in R \text{ and } (x, z) \in R \implies (y, z) \in R.$$

That is $z \in [y]$, and we have shown that $[x] \subset [y]$.

To prove the reverse inclusion $[y] \subset [x]$, let $z \in [y]$ be arbitrary. So $(y, z) \in R$, and then by transitivity

$$(x, y), (y, z) \in R \implies (x, z) \in R.$$

This shows that $z \in [x]$, and hence $[y] \subset [x]$. Thus

$$[y] = [x].$$

Proof that $[x] = [y] \implies [x] \cap [y] \neq \emptyset$. This is clear, we know that $[x]$ is not the empty set (since $x \in [x]$) and so $[x] \cap [y]$ is not the empty set (for instance $x \in [x] \cap [y]$.)

Proof that $[x] \cap [y] \neq \emptyset \implies (x, y) \in R$. If $[x] \cap [y] \neq \emptyset$ then there is some $z \in [x] \cap [y]$. By definition $(x, z) \in R$ and $(y, z) \in R$. By symmetry, $(z, y) \in R$. Then, by transitivity,

$$(x, z), (z, y) \in R \implies (x, y) \in R.$$

□

We now move on to **proof by contradiction**. The basic idea is as follows. Suppose that we want to prove that a given statement A is true. We assume the opposite, which is that A is not true, and look to see what consequences of this assumption can be logically derived. If we can derive a statement B which we know to be false, then there must be something wrong with our initial assumption that A is false. We therefore conclude that A is true. Actually, what we are doing with this approach is proving the equivalent *contrapositive* statement (recall the definition from Section 1.2), and some students may find this to be an easier way to think about this approach.

We first give one of the most famous proofs by contradiction in mathematics; the proof that $\sqrt{2}$ is irrational.

Theorem 1.27. *There is no* $x \in \mathbb{Q}$ *such that* $x^2 = 2$.

Proof. We assume for a contradiction that there is some $x \in \mathbb{Q}$ such that $x^2 = 2$. We write

$$x = \frac{m}{n} \tag{14}$$

for some integers m and n. Moreover, we may assume that at least one of m and n is odd. Indeed, if we have $m = 2k$ and $n = 2l$ then we can rewrite x as $x = \frac{k}{l}$ with $k < m$. We repeat this process until one of the parts of the fraction is odd. The process must terminate eventually, as the value of the integer k gets strictly smaller each time, and this cannot happen infinitely often.

Since $x^2 = 2$, (14) gives

$$2 = x^2 = \frac{m^2}{n^2}.$$

This rearranges to give

$$m^2 = 2n^2. \tag{15}$$

It therefore follows from Lemma 1.23 that m is even, and so we can write $m = 2k$ for some $k \in \mathbb{Z}$. We can plug this back into (15) to get $2n^2 = m^2 = (2k)^2 = 4k^2$, and thus

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

Therefore, n^2 is even, and it again follows from Lemma 1.23 that n is even.

We have reached a contradiction, since we stated earlier that at least one of m and n must be odd, but we have also shown that they must both be even. Therefore, our original assumption that $x^2 = 2$ for some $x \in \mathbb{Q}$ must be false. The proof is complete. □

We will see another famous example of a proof by contradiction a little later in this section, but now we move on to **proof by induction**. Proof by induction is a very structured proof technique that is particularly useful for proving that some proposition $P(n)$ is true for all integers n. The basic principle of mathematical induction is explained in the following statement.

Theorem 1.28. *A predicate $P(n)$ is true for all $n \in \mathbb{N}$ if the following two steps hold:*

- ***Induction basis****: $P(1)$ is true.*
- ***Induction step****: for all $n \in \mathbb{N}$*

$$P(n) \text{ is true} \implies P(n+1) \text{ is true.} \tag{16}$$

Proof. We know that $P(1)$ is true, by the first assumption. We can then iteratively apply (16) to obtain the chain of implications

$$P(1) \text{ is true} \implies P(2) \text{ is true} \implies P(3) \text{ is true} \implies \dots,$$

which completes the proof. □

We call the statement that $P(1)$ is true the **induction basis** or **base case**. The step $P(n) \implies P(n+1)$ is called the **induction step**, and the assumption that $P(n)$ is true is called the **induction hypothesis**.

Let us discuss two examples to demonstrate this type of proof. The first one is (at least without proof) known to many from school. The young Carl Friedrich Gauss (1777-1855) knew it already by the age of nine.

Theorem 1.29. *For all $n \in \mathbb{N}$,*

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}.$$

Proof. We first need to check that the statement is valid in the base case when $n = 1$. This is true, since

$$\sum_{k=1}^{1} k = 1 = \frac{1(1+1)}{2}.$$

Now, we need to show that, if the statement is true for n, it is also true for $n+1$. That is, we assume the induction hypothesis

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2},$$

and we need to use this to show that

$$\sum_{k=1}^{n+1} k = \frac{(n+1)(n+2)}{2}.$$

Indeed,

$$\sum_{k=1}^{n+1} k = \left(\sum_{k=1}^{n} k\right) + n + 1 = \frac{n(n+1)}{2} + n + 1 = \frac{n(n+1)}{2} + \frac{2n+2}{2} = \frac{(n+1)(n+2)}{2}.$$

We used the induction hypothesis in the second equality. □

Another important instance of a proof by induction is **Bernoulli's inequality**, which will be essential later.

Theorem 1.30. *Fix a real number $x \geq -1$ and let $n \in \mathbb{N}$. Then*

$$(1+x)^n \geq 1+nx. \tag{17}$$

Proof. We regard $x \geq -1$ as fixed, and seek to prove that (17) holds for all $n \in \mathbb{N}$ by induction. We first check the base case with $n=1$. Then (17) becomes

$$(1+x)^1 \geq 1+1x.$$

The two sides of this inequality are the same, and so this is true. Thus the base case is valid. We now assume the induction hypothesis

$$(1+x)^n \geq 1+nx \tag{18}$$

and seek to use this to prove that

$$(1+x)^{n+1} \geq 1+(n+1)x.$$

Indeed,

$$(1+x)^{n+1} = (1+x)^n(1+x) \overset{(18)}{\geq} (1+nx)(1+x) = 1+(n+1)x+nx^2 \geq 1+(n+1)x. \tag{19}$$

The last inequality simply uses the fact that $nx^2 \geq 0$. □

Where have we used the assumption that $x \geq -1$ in this proof? It is not so easy to see that we have used it if we do not pay proper attention. We have in fact used this assumption in the first inequality of (19), since we need $1+x \geq 0$ for the logic of this step to make sense.

It is sometimes helpful to use the following interpretation of proof by induction, where we strengthen our induction hypothesis to assume that the statement we want to prove is valid for *all* smaller values. This is called the method of **proof by strong induction**.

Theorem 1.31. *Let $k \geq 1$ be an integer. A predicate $P(n)$ is true for all $n \in \mathbb{N}$ such that $n \geq k$ if the following two steps hold*

- ***Induction basis**: $P(k)$ is true.*
- ***Induction step**: If $P(i)$ is true for all $k \leq i \leq n$ then $P(n+1)$ is true.*

Proof. We know that $P(k)$ is true, by the first assumption. Similarly to the proof of Theorem 1.29, we can then repeatedly apply the second assumption to prove that $P(n)$ holds for all $n \geq k$. Indeed,

$$\begin{aligned} P(k) \text{ is true} &\implies P(k+1) \text{ is true },\\ P(k) \text{ is true and } P(k+1) \text{ is true} &\implies P(k+2) \text{ is true },\\ P(k) \text{ is true and } P(k+1) \text{ is true and } P(k+2) \text{ is true} &\implies P(k+3) \text{ is true },\\ &\vdots \end{aligned}$$

which completes the proof. □

Theorem 1.31 looks slightly more complicated than Theorem 1.29, and one of the reasons for this is that we have made a generalisation that allows us to consider a base case which is not $k = 1$. However, the two methods are essentially the same, and we can also take a larger base case for the usual proof by induction.

We will use strong induction to prove that every integer can be written as a product of primes. This is one part of the Fundamental Theorem of Arithmetic. As the name suggests, this is an extremely important fact in number theory, which tells us that prime numbers form the building blocks for our number system.

Theorem 1.32. *Let $n \geq 2$ be an integer. Then there is some $i \in \mathbb{N}$ and prime numbers $p_1, \dots, p_i$ such that*

$$n = p_1 \cdot p_2 \cdots p_i. \tag{20}$$

Note that this statement does not require that the primes $p_1, \dots, p_i$ are distinct. For instance, to verify that the theorem holds for $n = 4$, we can observe that $4 = 2 \cdot 2$, and this decomposition satisfies the required form (20).

Proof. We will use proof by strong induction. For the base case $n = 2$, we know that n is a prime, and so it is already in the required form (20) (with $i = 1$).

Now we assume the induction hypothesis that every $2 \leq i \leq n$ can be written as a product of primes, and we seek to show that $n + 1$ can be written as a product of primes. If $n + 1$ is a prime number then it automatically has the required form of (20). If $n + 1$ is not prime then there exist two integers $2 \leq a, b < n + 1$ such that

$$n + 1 = a \cdot b. \tag{21}$$

By the induction hypothesis, we can write

$$a = p_1 \cdots p_i, \quad \text{and} \quad b = q_1 \cdots q_j,$$

for some primes $p_1, \dots, p_i$ and $q_1, \dots, q_j$. It then follows from (21) that

$$n + 1 = a \cdot b = p_1 \cdots p_i \cdot q_i \cdots q_j.$$

After relabelling the primes q_i suitably, this gives the required form (20), and the proof is complete. □

We conclude this section by using Theorem 1.32 as part of another ancient and famous proof by contradiction.

Theorem 1.33. *The set $\mathbb{P}$ of all prime numbers is infinite.*

Proof. Suppose for a contradiction that $\mathbb{P}$ is finite. Then we can write

$$\mathbb{P} = \{p_1, p_2, \dots, p_n\}$$

for some integer n. In this notation, p_i denotes the ith prime number. So, $p_1 = 2$, $p_2 = 3$, $p_5 = 11$, and p_n is the largest prime. Then consider the number

$$m := p_1 \cdot p_2 \cdots p_n + 1.$$

Since m is larger than all of the elements of $\mathbb{P}$, it is not a prime. But we know from Theorem 1.32 that m can be written as a product of primes. In particular, it follows that, for some $1 \leq i \leq n$, p_i divides m. But $p_i \geq 2$, and so

$$\frac{m}{p_i} = p_1 \cdots p_{i-1} \cdot p_{i+1} \cdots p_n + \frac{1}{p_i},$$

which is not an integer. This is a contradiction (since p_i both divides and does not divide m) and so the proof is complete. □

1.6 Bounded sets, infimum and supremum

In this subsection, we want to understand the "smallest" and the "largest" element of a set, where possible. To define this formally, we first need the definition of a bounded set (in $\mathbb{R}$).

Definition 1.34. *Let $A \subset \mathbb{R}$.*

- *We say A is* ***bounded from above*** *if and only if*

$$\exists\, c \in \mathbb{R} \,:\, \forall\, a \in A,\, a \leq c.$$

  We call c an ***upper bound*** *for A, and write $c \geq A$ or $A \leq c$.*

- *We say A is* ***bounded from below*** *if and only if*

$$\exists\, c \in \mathbb{R} \,:\, \forall\, a \in A,\, c \leq a.$$

  We call c a ***lower bound*** *for A, and write $c \leq A$ or $A \geq c$.*

- *We say A is* ***bounded*** *if and only if A is bounded from above and bounded from below.*

Example - Let $a, b \in \mathbb{R}$ such that $a < b$. Then for the closed interval $I = [a, b]$ we have that $a, a - 1$ and $a - 42$ are lower bounds for I and b and $b + 42$ are upper bounds, and the same is true for the corresponding open interval (a, b). So, upper and lower bounds are not unique in these cases (and in general).

To fix a specific upper/lower bound, let us first define the minimum and maximum of a set.

Definition 1.35. *Let $A \subset \mathbb{R}$ be a non-empty set and $t \in \mathbb{R}$. Then, t is called a* ***minimal element*** *or* ***minimum*** *of A, denoted by* $\min A := t$ *, if and only if*

- *$t \leq A$ (i.e., t is a lower bound for A), and*
- *$t \in A$.*

t is called a ***maximal element*** *or* ***maximum*** *of A, denoted by* $\max A := t$ *, if and only if*

- *$A \leq t$ (i.e., t is an upper bound for A), and*
- *$t \in A$.*

If the maximum or minimum exists, it is unique. (Why?)

Example - Let $a, b \in \mathbb{R}$ with $a < b$. Then, $\min[a, b] = a$ and $\max[a, b] = b$.

Example - The set of the natural numbers $\mathbb{N} \subset \mathbb{R}$ is bounded from below, with $\min \mathbb{N} = 1$.

However, maxima and minima do not have to exist. For instance, while b is the least upper bound of both $[a, b]$ and (a, b), we have that $b \in [a, b]$, but $b \notin (a, b)$ Hence, b is not the maximum of (a, b), and in fact the set (a, b) does not have a maximum (or minimum). But still, we would like to work with the "best possible" upper and lower bounds for such a set, which are clearly a and b in this example. For this we define the infimum and the supremum as the **greatest lower bound** and the **least upper bound**, respectively. These objects will be very important in the upcoming analysis.

Definition 1.36. *Let $A \subset \mathbb{R}$ be non-empty and $t \in \mathbb{R}$. Then, t is called the* ***greatest lower bound*** *or* ***infimum*** *of A, denoted by $\inf A := t$, if and only if*

- *$t \le A$ (i.e. t is a lower bound), and*
- *$x \le A \implies x \le t$ (i.e. there is no greater lower bound).*

t is called the ***least upper bound*** *or* ***supremum*** *of A, denoted by $\sup A := t$, if and only if*

- *$t \ge A$ (i.e. t is an upper bound), and*
- *$x \ge A \implies x \ge t$ (i.e. there is no smaller upper bound).*

If A is not bounded from above we set $\sup A := \infty$. If A is not bounded from below then we put $\inf A = -\infty$.

For the empty set, we define

$$\sup \emptyset = -\infty, \quad \textit{and} \quad \inf \emptyset = \infty.$$

If $\inf A \in A$, then $\inf A = \min A$, and if $\sup A \in A$, then $\sup A = \max A$. In words, if the infimum (or supremum) of a set A is contained in A, then A has a minimum (or maximum) which has the same value.

Moreover, the infimum and supremum are uniquely determined. To see this for the supremum, assume that there are two suprema t_1 and t_2 for A. Since $\sup A = t_1$, we have $A \le t_1$. In addition, since $\sup A = t_2$, we obtain by the second defining property above that $x \ge A \implies x \ge t_2$. Setting $x = t_1$, we have $t_1 \ge t_2$. But we can also make this argument in reverse to show that $t_1 \le t_2$. Hence $t_1 = t_2$.

Example - Let $a, b \in \mathbb{R}$ with $a < b$. Then,

$$\min[a,b] = \min[a,b) = \inf[a,b] = \inf[a,b) = \inf(a,b] = \inf(a,b) = a$$

and

$$\max[a,b] = \max(a,b] = \sup[a,b] = \sup[a,b) = \sup(a,b] = \sup(a,b) = b.$$

However, $\min(a,b)$, $\min(a,b]$, $\max(a,b)$ and $\max[a,b)$ do not exist.

Example - Let $A = \{x^2 : x \in (-1,1)\}$. Then $A = [0,1)$. Hence

$$\inf A = \min A = 0, \text{ and } \sup A = 1.$$

Let us state an equivalent definition of the supremum and infimum for bounded sets in $\mathbb{R}$. Although it looks more complicated at first sight, this formulation is sometimes very helpful. Moreover, thinking in this way will be a helpful preparation for some of the real analysis we consider later, and particularly for understanding important definitions of convergence, continuity, and much more.

Definition 1.37. *Let $A \subset \mathbb{R}$ be bounded from below. Then, $t = \inf A$ if and only if*

- *$t \le A$ (i.e., t is a lower bound for A), and*

- *$\forall\, \epsilon > 0$, $\exists\, a \in A\,:\, a < t + \epsilon$ (i.e. t comes arbitrarily close to A).*

Analogously, let $A \subset \mathbb{R}$ be bounded from above. Then, $t = \sup A$ if and only if

- *$A \leq t$ (i.e., t is an upper bound for A), and*
- *$\forall\, \epsilon > 0$, $\exists\, a \in A\,:\, a > t - \epsilon$ (i.e. t comes arbitrarily close to A).*

Exercise - Let $A \subset \mathbb{R}$. Define

$$-A := \{-a : a \in A\}$$

and suppose that $\sup A = k$ for some $k \in \mathbb{R}$. Prove that $\inf(-A)$ exists and $\inf(-A) = -k$.

The next result reflects the intuitive notion that the real number line is complete, i.e. there are no gaps in $\mathbb{R}$. This is an important assumption (i.e. an axiom) that underlies much of the work we will do in the Mathematics for AI curriculum, although it is mostly working behind the scenes of the mathematics we study.

Axiom 1.38 (Completeness Axiom)**.** *Every non-empty set $A \subset \mathbb{R}$ that is bounded from above has a least upper bound. That is, there always exists $t \in \mathbb{R}$ such that $t = \sup A$.*

Exercise - Prove that every non-empty set $A \subset \mathbb{R}$ that is bounded from below has a greatest lower bound. (Hint: Use the completeness axiom and a previous exercise.)

Note that the completeness axiom is false if we consider $\mathbb{Q}$ instead of $\mathbb{R}$. Consider for instance the set

$$A = \{x \in \mathbb{Q} : x \leq \sqrt{2}\}.$$

This set is bounded, and we can find a number $c \in \mathbb{Q}$ such that $c \geq A$ (for instance, $c = \frac{3}{2}$ works). However, we cannot find a least upper bound among all such rational c. Suppose that we try to do this. We are looking for a rational number c which is larger than $\sqrt{2}$ but also as close to $\sqrt{2}$ as possible. But, no matter which C we choose, we can always find a smaller one which is still larger than $\sqrt{2}$.

This idea is considered more formally in the next theorem, where we state the Archimedean property. This is based on the fact that the set of natural numbers is unbounded. Even though this property seems rather obvious and unimpressive, it was of significant importance for real analysis, and will be implicitly used repeatedly throughout the Mathematics for AI study.

Theorem 1.39. *The following assertions hold:*

1. *The **Archimedean property:***

   $$\forall x \in \mathbb{R},\, \exists\, n \in \mathbb{N}\,:\, n > x.$$

   In other words, $\mathbb{N}$ has no upper bound in $\mathbb{R}$.

2. *$\forall \epsilon > 0$, $\exists\, n \in \mathbb{N}\,:\, \frac{1}{n} < \epsilon$.*

3. *For any $x, y \in \mathbb{R}$ with $x < y$, there exists a rational number $\frac{m}{n} \in \mathbb{Q}$ such that*

   $$x < \frac{m}{n} < y.$$

The last of these three points is interesting; it tells us that, no matter how close the two real numbers x and y we consider are, we can always find a rational number in between them. Another way to think of this is as follows: no matter where you are on the real number line, you are always arbitrarily close to a rational number.

Proof. We first prove the first point using proof by contradiction. For this, we assume that $\mathbb{N}$ is bounded. Thus, the supremum $x = \sup \mathbb{N}$ exists by the completeness axiom. As x is the least upper bound of $\mathbb{N}$, $x - 1$ is not an upper bound for $\mathbb{N}$, so there exists some $n \in \mathbb{N}$ with

$$x - 1 < n.$$

Adding 1 to both sides of this inequality gives

$$x < n + 1.$$

But $n + 1 \in \mathbb{N}$ (since $n \in \mathbb{N}$) and therefore x is not an upper bound for $\mathbb{N}$. This contradicts the fact that $x = \sup \mathbb{N}$. So, $\mathbb{N}$ must be unbounded.

Now we prove the second point of the theorem. Let $\epsilon > 0$ be arbitrary. We can apply the Archimedean property with $x = \frac{1}{\epsilon}$, which implies that there exists $n \in \mathbb{N}$ with

$$n > \frac{1}{\epsilon}.$$

Rearranging this inequality, we obtain

$$\frac{1}{n} < \epsilon,$$

as required.

We now prove the third bullet-point. We will only check the case $x, y > 0$, since the other possible cases can be checked similarly. Then we are looking for $m, n \in \mathbb{N}$ such that $nx < m < ny$. By the Archimedean property, there exists $n \in \mathbb{N}$ with $n > \frac{1}{y-x}$, which rearranges to give

$$ny > 1 + nx. \tag{22}$$

Now take the smallest natural number m such that $m > nx$ (such an m certainly exists). Then $m > nx$ and $m - 1 \leq nx$. The latter inequality combined with (22) implies that

$$m \leq nx + 1 < ny.$$

We have therefore proved that

$$nx < m < ny.$$

□

Based on this, we can easily calculate certain infima and suprema of (discrete) sets.

Example Let $A = \{\frac{1}{n} : n \in \mathbb{N}\}$. Then

$$\inf A = 0, \quad \text{and} \quad \max A = 1 \tag{23}$$

To prove (23), first note that $0 < \frac{1}{n} \leq 1$ for all $n \in \mathbb{N}$. So, 0 is a lower bound for A, and 1 is an upper bound. Since $1 \in A$ (for $n = 1$), we therefore obtain that $\max A = 1$.

For the infimum we need to show the there is no larger lower bound. For this, let $\epsilon > 0$ be arbitrary, and suppose for a contradiction that ϵ is a lower bound for A. By the Archimedean property there exists an $n \in \mathbb{N}$ with $\frac{1}{n} < \epsilon$. Therefore, we have found some $a = \frac{1}{n} \in A$ such that $a < \epsilon$. This contradicts the assumption that ϵ is a lower bound, and completes the proof.

Exercise - Let

$$A = \left\{ \frac{1}{n^2 - n - 3} : n \in \mathbb{N} \right\}.$$

Calculate $\inf A$, $\sup A$, $\min A$ and $\max A$, if these quantities exist.

1.7 Some basic combinatorial objects, identities and inequalities

We begin this section by defining some important combinatorial quantities. These quantities are heavily used when it comes to discrete mathematics or elementary probability theory, as they represent the number of permutations or subsets of certain size.

Definition 1.40. *The **factorial** $n!$ of a natural number $n \in \mathbb{N}$ is the product*

$$n! = 1 \cdot 2 \cdots n = \prod_{i=1}^{n} i.$$

In addition, we define $0! = 1$.

*For any $n, k \in \mathbb{N}_0$ with $n \geq k$, the **binomial coefficient** $\binom{n}{k}$ (we say "n choose k") is defined by the formula*

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k \cdot (k-1) \cdots 1}.$$

Observe that

$$\binom{n}{0} = 1, \quad \binom{n}{1} = n, \quad \text{and} \quad \binom{n}{k} = \binom{n}{n-k}.$$

The factorial $n!$ is the number of permutations (orderings) of a set of n objects. If you pick your first arbitrary element, you have n possibilities for your choice. When it comes to your second decision you have just $n-1$ options left, and so on.

The binomial coefficient $\binom{n}{k}$ represents the number of ways to choose k (unordered) outcomes from a set of n possibilities. For example, $\binom{n}{3}$ is the number of three-element subsets of $\{1, \ldots, n\}$ and $\binom{4}{2}$ is the number of football matches that are required in a world cup group (4 teams, with each pair of teams playing each other once).

The following is a useful formula for binomial coefficients.

Lemma 1.41. *Let $n, k \in \mathbb{N}$ with $k \leq n-1$. Then we have*

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}.$$

The "simple" explanation of this equality is, that subsets of the subsets $\{1, \ldots, n+1\}$ with $k+1$ elements can be split into those that contain the number $n+1$ and those that don't contain it. To find the number of sets that contain $n+1$, we need to count all k-element subsets of $\{1, \ldots, n\}$, and there are $\binom{n}{k}$ of them. The number of all sets that don't contain $n+1$, is the same as the number of all $(k+1)$-element subsets of $\{1, \ldots, n\}$, which is $\binom{n}{k+1}$. So the total number is their sum.

Here is a more formal and formulaic proof of Lemma 1.41.

Proof.

$$\begin{aligned}\binom{n}{k}+\binom{n}{k+1} &= \frac{n!}{k!(n-k)!}+\frac{n!}{(k+1)!(n-(k+1))!}\\ &= \frac{n!}{k!(n-k)!}+\frac{n!}{(k+1)!(n-k-1)!}\\ &= \frac{n!(k+1)}{(k+1)!(n-k)!}+\frac{n!(n-k)}{(k+1)!(n-k)!}\\ &= \frac{n!((k+1)+(n-k))}{(k+1)!(n-k)!}\\ &= \frac{n!(1+n)}{(k+1)!(n-k)!}\\ &= \frac{(n+1)!}{(k+1)!(n-k)!}\\ &= \frac{(n+1)!}{(k+1)!((n+1)-(k+1))!}=\binom{n+1}{k+1}.\end{aligned}$$

□

Based on this, we can prove another famous and highly applicable theorem; the **Binomial Theorem**.

Theorem 1.42. *Let $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$. Then we have*

$$(x+y)^n = \sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}.$$

Proof. Consider expanding the product $(x+y)^n$ (i.e. multiplying out all of the brackets). Every term we obtain must take the form $x^k y^{n-k}$ for some $0 \leq k \leq n$. This is simply because the powers must add up to n (each product consists of exactly n parts).

It remains to consider how many times the term $x^k y^{n-k}$ appears. This will indeed be $\binom{n}{k}$, since a term $x^k y^{n-k}$ appears if we choose exactly k x from the total of n choices.

□

Alternatively, we can prove the Binomial Theorem by induction, making use of Lemma 1.41. The proof by induction is a bit longer, but some may prefer this more formal approach.

Example - If we set $x = y = 1$ in Theorem 1.42 we get

$$2^n = \sum_{k=0}^{n}\binom{n}{k}.$$

We can also justify this identity by thinking in terms of sets. Both sides of this identity count the number of subsets of $\{1, 2, \ldots, n\}$ (or, moreover, any set of n elements). To see that the number of such sets is 2^n, simply observe that for each of the n elements, we have 2 choices in our construction of a set $A \subset \{1, \ldots, n\}$; we either include the element in A, or we do not.

On the other hand, we can also count the number of subsets of $\{1, \dots, n\}$ by counting the number of subsets with a fixed size k, which gives $\binom{n}{k}$, and then summing over all possible values of k.

Example - Setting $x = -1$ and $y = 1$ in Theorem 1.42 gives the identity

$$0 = \sum_{k=0}^{n} \binom{n}{k} (-1)^k.$$

The binomial theorem can also be used to (easily) improve upon Bernoulli's inequality (see Theorem 1.30.

Corollary 1.43. *Let $x \geq 0$. Then, for any $m \in \{1, \dots, n\}$, we have*

$$(1+x)^n \geq 1 + \binom{n}{m} x^m.$$

Proof. Apply Theorem 1.42 with $y = 1$. We obtain

$$(x+1)^n = \sum_{k=0}^{n} \binom{n}{k} x^k \geq 1 + \binom{n}{m} x^m.$$

For the inequality above, we simply take only the $k = 0$ and $k = m$ terms from the sum. The other terms are all non-negative from the assumption that $x \geq 0$. □

If we apply this theorem with $m = 1$, we recover the lower bound from Bernoulli's inequality. Note however that this Corollary 1.43 is only valid for $x \geq 0$, and it is only guaranteed to give a proper improvement when $x \geq 1$. On the other hand, Bernoulli's inequality is valid for all $x \geq -1$.

1.8 Some important functions

In this subsection we want to briefly discuss some particularly important functions. We start with the **absolute value** of a number, since this function and its generalisations are heavily used throughout the Mathematics for AI courses.

The absolute value of $x \in \mathbb{R}$ is defined by:

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}. \tag{24}$$

The absolute value is a function from $\mathbb{R}$ to $\mathbb{R}_+$ (recall that $\mathbb{R}_+$ denotes the set of all non-negative real numbers, i.e. $\mathbb{R}_+ = [0, \infty)$). We write $|\cdot| : \mathbb{R} \to \mathbb{R}_+$. The graph of the function is shown below.

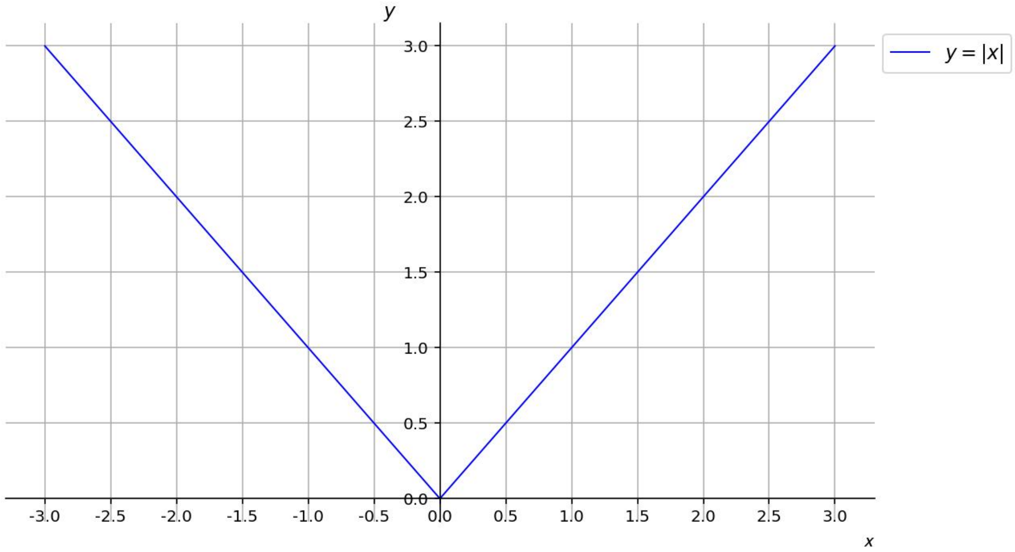


Figure 10: The graph of the absolute value function $f(x) = |x|$

We list some important properties of the absolute value function in the following lemma.

Lemma 1.44. *The absolute value function satisfies the following properties.*

1. *For all* $x \in \mathbb{R}$, $|x| \geq 0$.
2. $|x| = 0 \iff x = 0$.
3. *For all* $x \in \mathbb{R}$, $|x| \geq x$ *and* $|x| \geq -x$.
4. *For all* $x, y \in \mathbb{R}$, $|xy| = |x||y|$.
5. *For all* $x \in \mathbb{R}$, $|-x| = |x|$.
6. *For all* $x \in \mathbb{R}$ *and* $z \in [0, \infty)$,

$$|x| \leq z \iff -z \leq x \leq z.$$

and

$$|x| < z \iff -z < x < z.$$

Proof. The first 3 points follow immediately from the definition of the absolute value.

To prove point 4, we split into 4 cases.

Case 1 - Suppose that $x, y \geq 0$. Then $xy \geq 0$, and so $|xy| = xy$. Also, $|x| = x$ and $|y| = y$. It follows that $|x||y| = xy = |xy|$.

Case 2 - Suppose that $x, y < 0$. Then $xy > 0$, and so $|xy| = xy$. Also, $|x| = -x$ and $|y| = -y$. It follows that $|x||y| = (-x)(-y) = xy = |xy|$.

Case 3 - Suppose that $x \geq 0$ and $y < 0$. Then $xy \leq 0$, and so $|xy| = -xy$. Also, $|x| = x$ and $|y| = -y$. It follows that $|x||y| = x(-y) = -xy = |xy|$.

Case 4 - Suppose that $x < 0$ and $y \geq 0$. Then $xy \leq 0$, and so $|xy| = -xy$. Also, $|x| = -x$ and $|y| = y$. It follows that $|x||y| = (-x)y = -xy = |xy|$.

The proof of points 5 and 6 is left as an exercise. □

The case distinction that we have made in the proof of point 4 in Lemma 1.44 gives a good illustration of how to deal with absolute values in general. We will use a similar approach again now, as we try to understand inequalities involving absolute values.

Example - Consider the inequality $|x - 1| < 2$. We would like to determine all values of x which satisfy this inequality. Define the set

$$L := \{x \in \mathbb{R} : |x - 1| < 2\}.$$

There are two situations that need to be considered separately, according to whether or not $x - 1$ is positive. We split L into two parts. Write $L = L_1 \cup L_2$ where

$$L_1 := \{x \in L : x < 1\}, \quad \text{and} \quad L_2 := \{x \in L : x \geq 1\}.$$

The sets L_1 and L_2 are disjoint (i.e. the intersection $L_1 \cap L_2$ is equal to the empty set). It remains to find a simplified expression for the set L_1 and L_2, and thus to determine precisely what L looks like.

Case 1 - Consider $x < 1$. In this range $|x - 1| = -(x - 1) = 1 - x$, and so the inequality becomes

$$1 - x < 2,$$

which rearranges to $x > -1$. So,

$$L_1 = \{x \in \mathbb{R} : x < 1 \text{ and } x > -1\} = (-1, 1).$$

Case 2 - Consider the range $x \geq 1$. Then $|x-1| = x-1$ and the inequality under consideration simplifies to the form

$$x - 1 < 2,$$

which is equivalent to $x < 3$. Therefore,

$$L_2 = \{x \in \mathbb{R} : x \geq 1 \text{ and } x < 3\} = [1, 3).$$

Putting things together, we have

$$L = L_1 \cup L_2 = (-1, 1) \cup [1, 3) = (-1, 3).$$

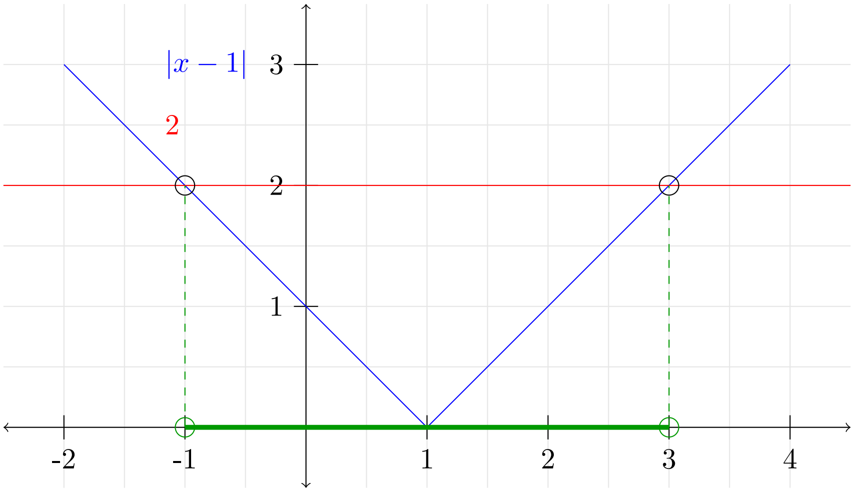


Figure 11: A graphical representation of the inequality and its solution set

We remark here that we also could have reached the conclusion that $L = (-1, 3)$ by applying point 6 from Lemma 1.44. However, for more complicated inequalities involving absolute values Lemma 1.44 is not sufficient, and we really do need to use such a case distinction. The following is such an example.

Example - Consider the inequality

$$2|x+3| - 4|x-1| \geq 8x - 2. \tag{25}$$

We want to determine the nature of the set $L := \{x \in \mathbb{R} : 2|x+3| - 4|x-1| \geq 8x - 2\}$. In this case, we have to distinguish three regions ($x \leq -3$, $-3 < x \leq 1$ and $x > 1$), since the expressions in the absolute values change their signs at these points. Write $L = L_1 \cup L_2 \cup L_3$ where

$$L_1 := \{x \in L : x \leq -3\}, \quad L_2 := \{x \in L : -3 < x \leq 1\}, \quad \text{and} \quad L_3 := \{x \in L : x > 1\}.$$

Case 1 - Suppose that $x \leq -3$. Then the inequality (25) becomes

$$2(-x-3) - 4(-x+1) \geq 8x - 2.$$

After some basic algebra, this is equivalent to $x \leq -\frac{4}{3}$. It then follows that

$$L_1 = \left\{ x \in \mathbb{R} : x \leq -3 \text{ and } x \leq -\frac{4}{3} \right\} = (-\infty, -3].$$

Case 2 Suppose that $-3 < x \leq 1$. Then the inequality (25) becomes

$$2(x+3) - 4(-x+1) \geq 8x - 2.$$

After some basic algebra, this is equivalent to $x \leq 2$. It then follows that

$$L_2 = \{x \in \mathbb{R} : -3 < x \leq 1 \text{ and } x \leq 2\} = (-3, 1].$$

Case 3 - Suppose that $x > 1$. Then the inequality (25) becomes

$$2(x+3) - 4(x-1) \geq 8x - 2.$$

After some basic algebra, it is equivalent to $x \leq \frac{6}{5}$. It then follows that

$$L_3 = \left\{ x \in \mathbb{R} : x > 1 \text{ and } x \leq \frac{6}{5} \right\} = \left(1, \frac{6}{5}\right].$$

Putting everything together, we conclude that

$$L = L_1 \cup L_2 \cup L_3 = \left(-\infty, \frac{6}{5}\right].$$

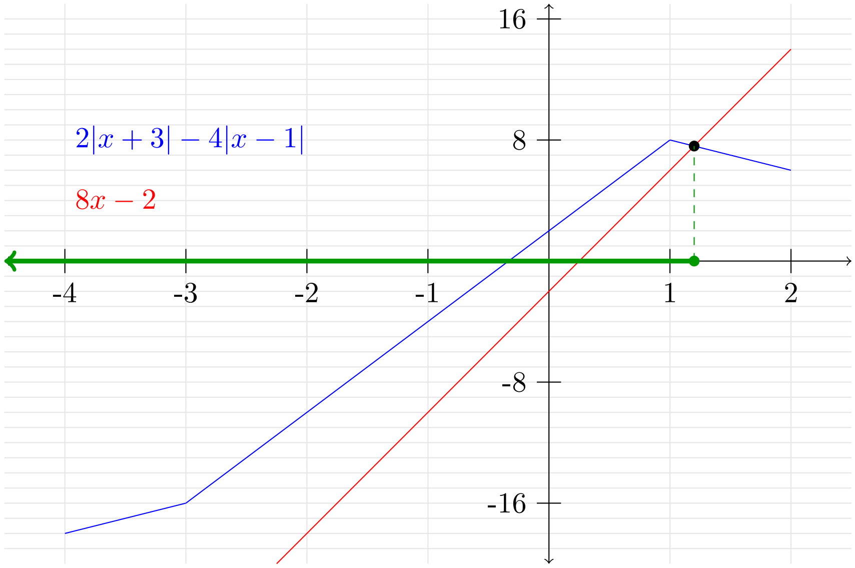


Figure 12: A graphical representation of the inequality (25) and its solution set

The following is probably the most well known inequality involving absolute values. We will see this inequality reappearing in more general scenarios throughout the Mathematics for AI material. This is called the **triangle inequality** for the absolute value function.

Theorem 1.45. *Let $x, y \in \mathbb{R}$. Then*

$$|x+y| \leq |x| + |y|.$$

Proof. We already know from Lemma 1.44 that $|z| \geq z$ and $|z| \geq -z$ for all $z \in \mathbb{R}$. We therefore have

$$|x| + |y| \geq x + y$$

and

$$|x| + |y| \geq -x - y = -(x+y).$$

Since $|x+y|$ is equal to either $x+y$ or $-(x+y)$, we always have $|x+y| \leq |x| + |y|$, as required. □

The triangle inequality (Theorem 1.45) also implies that

$$|x-z| = |x-y+y-z| \leq |x-y| + |y-z|, \tag{26}$$

for all $x, y, z \in \mathbb{R}$. Hence, we can consider the absolute value as a "distance" between numbers. The triangle inequality in the form (26) can be interpreted as saying that "the shortest route from x to z is to go directly, rather than stopping at another number y".

Moreover, we obtain the following (less obvious) inequality, which will also be useful later.

Corollary 1.46. *Let $x, y \in \mathbb{R}$. Then*

$$\big||x| - |y|\big| \leq |x-y|.$$

Proof. Theorem 1.45 yields

$$|a+b| - |b| \leq |a| \tag{27}$$

for all $a, b \in \mathbb{R}$. Set $a = x - y$ and $b = y$, so that (27) gives

$$|x| - |y| \leq |x-y|.$$

On the other hand, if we set $a = x - y$ and $b = -x$, (27) gives

$$|-y| - |-x| \leq |x-y|,$$

which is equivalent to

$$-(|x| - |y|) \leq |x-y|.$$

Since $\big||x| - |y|\big|$ is equal to either $|x| - |y|$ or $-(|x| - |y|)$, it follows that we always have the intended inequality

$$\big||x| - |y|\big| \leq |x-y|.$$

□

We now turn to some other elementary functions. An **affine linear function** is a function $f : \mathbb{R} \to \mathbb{R}$ of the form

$$f(x) = ax + b$$

with $a, b \in \mathbb{R}$. If $a = 0$, then f is called a **constant function**. If $a \neq 0$ and $b = 0$ we call f a **linear function**. Note that linear functions satisfy $f(x+y) = f(x) + f(y)$. (Is the same true for affine linear functions?) Moreover, they are special cases of **polynomial functions**, which are defined to be functions $p : \mathbb{R} \to \mathbb{R}$ of the form

$$p(x) = \sum_{i=0}^{n} a_i x^i,$$

with $a_0, \dots a_n \in \mathbb{R}$. For a non-zero polynomial, i.e. a polynomial where at least one of the coeffecients a_i is not equal to zero, we may assume that $a_n \neq 0$. The value n is the **degree** of the polynomial p. The degree of the zero polynomial (i.e. the function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = 0$ for all $x \in \mathbb{R}$) is not defined.

Examples - Consider the functions

$$f(x) = 2x - 3, \quad g(x) = 4x^3 + 5x, \quad h(x) = \pi, \quad j(x) = \sqrt{2}x.$$

All 4 functions are polynomials. f, h and j are also affine linear functions. j is also a linear function. h is also a constant function.

Closely related to polynomials are **power functions**. These are functions $f : \mathbb{R}_+ \to \mathbb{R}$ of the form $f(x) = x^a$, for some (fixed) $a \in \mathbb{R}$. In the case when a is a natural number, such a power function is a special kind of polynomial called a **monomial**.

The restriction of the domain to the positive reals is necessary in order for these functions to be well defined in general. For instance, if we consider the power function $f(x) = x^{1/2}$, this is the function which takes the positive square root of x. In order for this to be a real number, we may only consider $x \geq 0$.

The next image illustrates the behaviour of the power function for some different choices of a. Notice that there are some crucial points where the behaviour changes. The function grows faster as a increases, and we see a different shape of curve for $a > 1$, $a = 1$ and $0 < a < 1$. Moreover, the function looks very different when a is negative; instead of growing with x, the function is now decreasing.

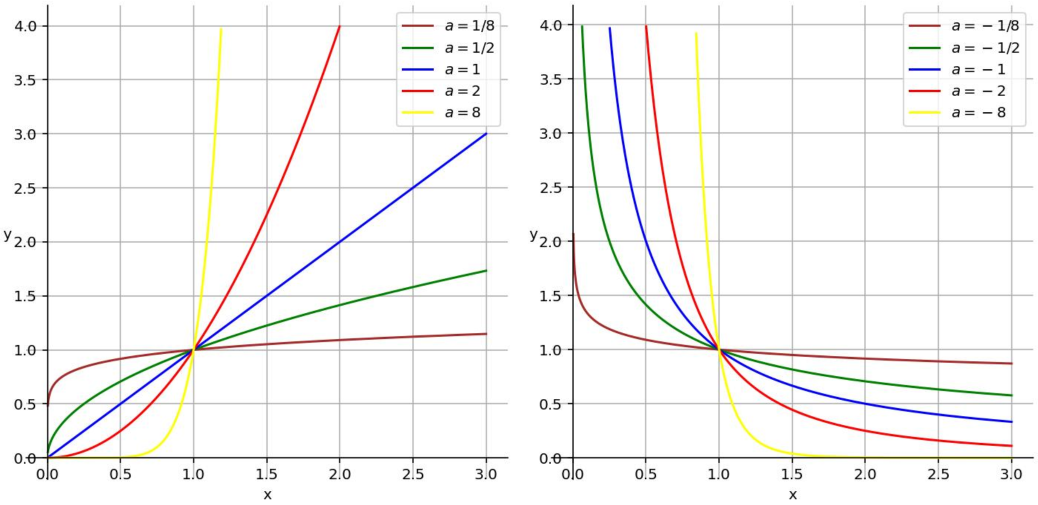


Figure 13: The graph of the function $f(x) = x^a$ for different a.

The most natural power functions to consider are those for which $a = m/n$ is a rational number, in which case we can build a power function $f(x) = x^a$ as a composition of familiar functions $g(x) = x^m$ and $h(x) = \sqrt[n]{x}$. One can also consider power functions of the form $f(x) = x^a$ where a is an irrational number. We avoid giving a full formal definition, since it is rather technical. Interested students may wish to consult Mario Ullrich's notes from the previous Mathematics for AI 1 course (see page 37 of the notes). Roughly speaking, we can approximate $f(x) = x^a$ by finding a rational number b which is *very* close to a and calculating x^b.

Another well-known class of functions are the **exponential functions**, which characterise rapid growth or decay. Exponential functions are defined to be functions $f : \mathbb{R} \to \mathbb{R}_+$ of the form

$$f(x) = a^x,$$

where the base $a > 0$ is a fixed real number. That is, in contrast to power functions, the variable $x \in \mathbb{R}$ is in the exponent.

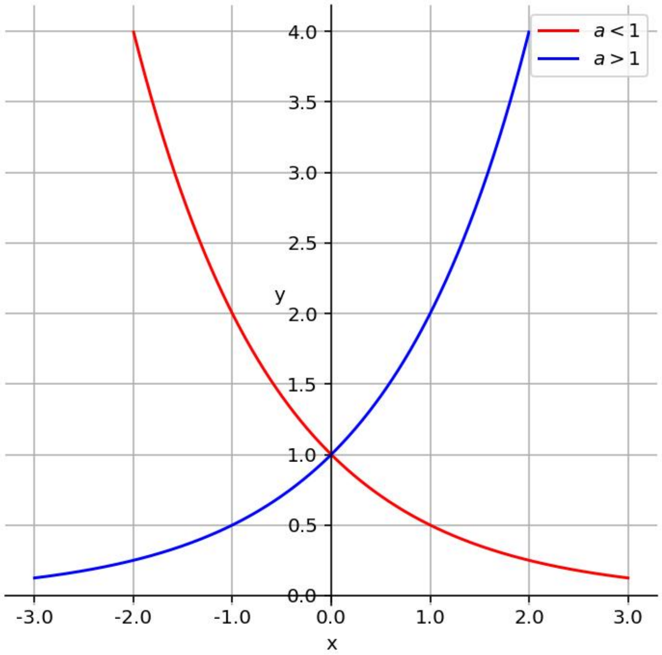


Figure 14: The graph of the function $f(x) = a^x$ for different a.

Among all exponential functions, one is particularly important. This is when we choose the basis $a = e \approx 2.7182818284$, where e is **Euler's number**. This special (irrational) number e is one of the most important constants in mathematics. It connects many different concepts and appears in some beautiful formulae. It may be defined by different means. Indeed,

$$e = \sup\left\{\left(1+\frac{1}{n}\right)^n : n \in \mathbb{N}\right\} = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = \sum_{k=0}^{\infty}\frac{1}{k!}.$$

Don't worry too much about these complicated looking expressions for now. We will come back to them much later.

The exponential functions often appear together with the **logarithmic functions**, which are defined as their inverses. Let $b > 0$ be a real number (we usually restrict our attention to the case $b > 1$. The function $\log_b : \mathbb{R}_+ \to \mathbb{R}$ is defined to be the inverse of the exponential function $g(x) = b^x$. Recalling the properties of inverse functions, this means that

$$\log_b(b^x) = x, \quad \forall\, x \in \mathbb{R}$$

and

$$b^{\log_b(x)} = x, \quad \forall\, x \in \mathbb{R}_+. \tag{28}$$

We can use (28) to give a wordy definition of the function $\log_b$. "The value of $\log_b(x)$ is defined to be the number y which satisfies $b^y = x$."

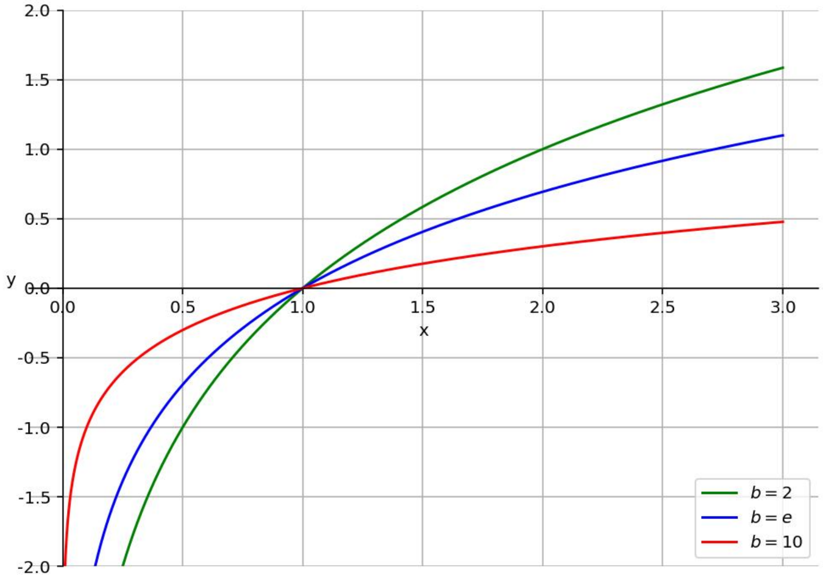


Figure 15: The graph of the function $f(x) = \log_b(x)$ for different b.

Once again, an important special case occurs when $b = e$ (i.e. Euler's number). We use the shorthand $\ln(x) = \log_e(x)$.

Using familiar calculation rules for exponentiation we obtain the following calculation rules for logarithms. For all $a, b, x, y > 0$:

$$\begin{aligned}\log_b x^y &= y \log_b x\\ \log_b(xy) &= \log_b x + \log_b y\\ \log_b\left(\frac{x}{y}\right) &= \log_b x - \log_b y\\ \log_b x &= \frac{\log_a x}{\log_a b}.\end{aligned}$$

Trigonometric functions play a crucial role in analysis and geometry. The most important are $\sin x$, $\cos x$ and $\tan x$. We can give a geometric definition of these function using the unit circle, as the following image illustrates.

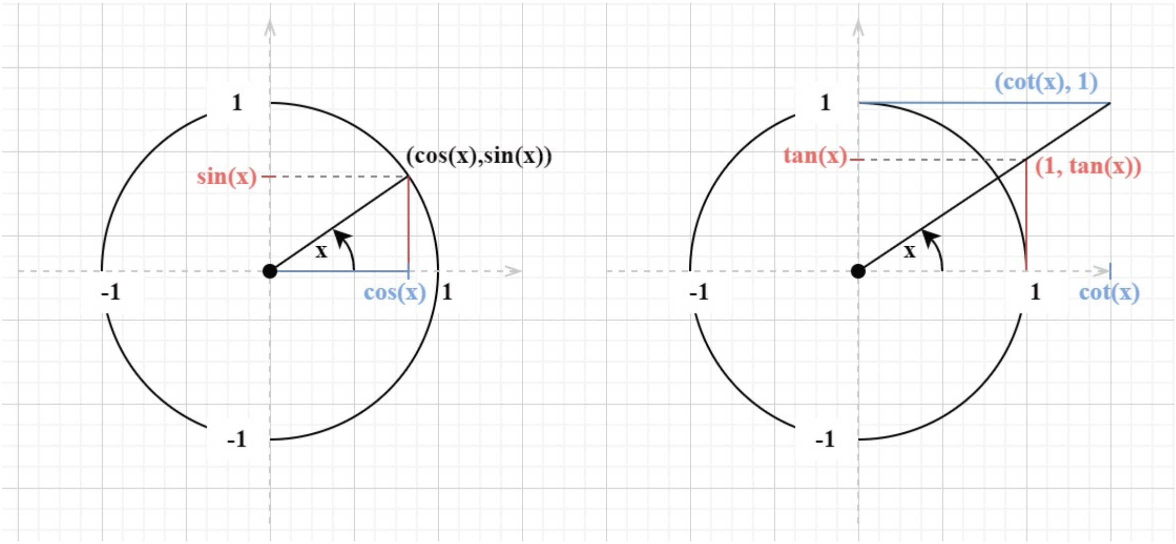


Figure 16: Illustration of trigonometric functions.

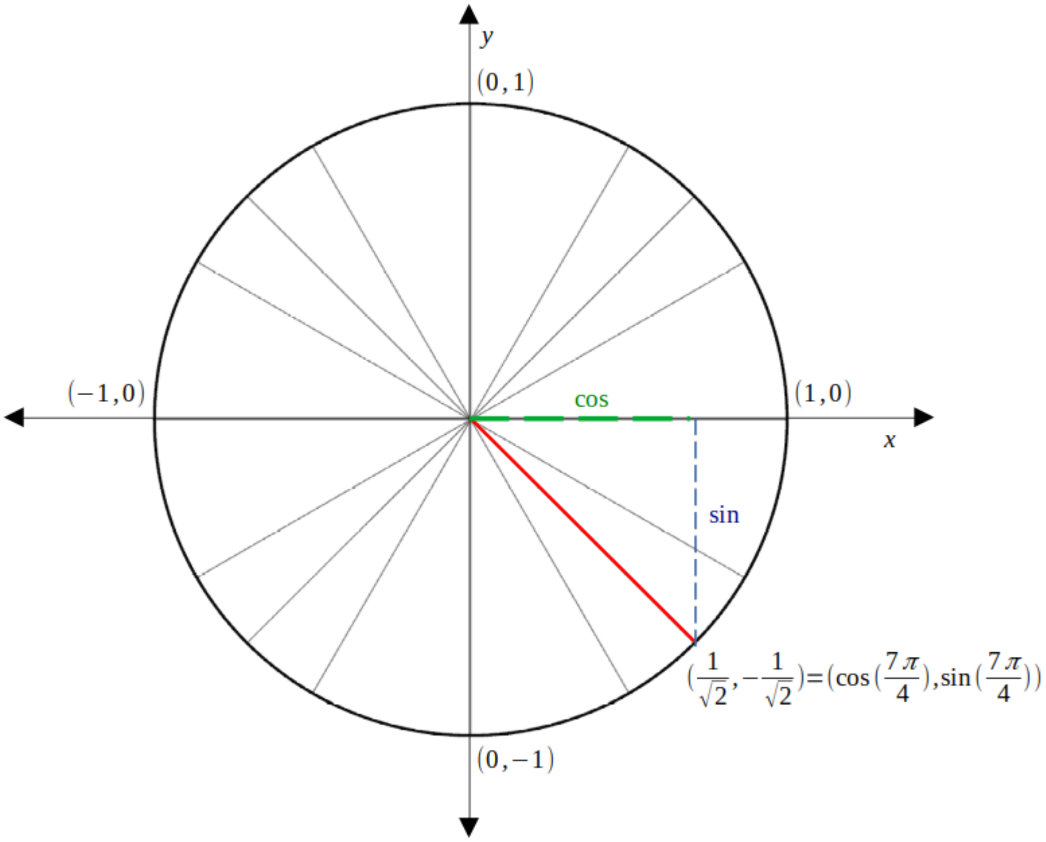
The variable $x \in \mathbb{R}$ corresponds to the (anticlockwise) angle that is enclosed between the horizontal axis and the line to the point $(\cos x, \sin x)$. We will usually use the unit of *radians* to refer to an angle.

Example - Consider the point $p = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ on the unit circle. We can calculate (using high school trigonometry) that the angle between the horizontal axis and the line to the point p is $\frac{\pi}{4}$. Therefore, we have

$$\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = \left(\cos\left(\frac{\pi}{4}\right), \sin\left(\frac{\pi}{4}\right)\right).$$

For the point $\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$, the (anticlockwise) angle is $2\pi - \frac{\pi}{4} = \frac{7\pi}{4}$, by symmetry, and so

$$\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right) = \left(\cos\left(\frac{7\pi}{4}\right), \sin\left(\frac{7\pi}{4}\right)\right).$$



y
(0,1)
(−1,0)
cos
(1,0)
x
sin
$(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}) = (\cos(\frac{7\pi}{4}), \sin(\frac{7\pi}{4}))$
(0,−1)

Figure 17: Illustration of the example above.

The functions $\sin x$ and $\cos x$ are defined for all $x \in \mathbb{R}$, see the illustration below.

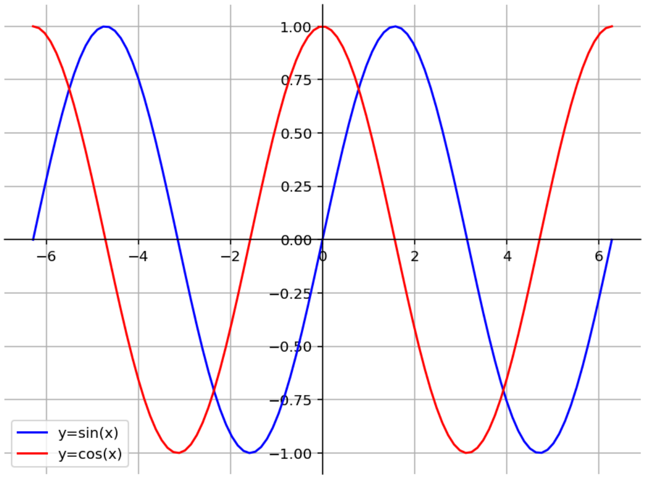


Figure 18: Graphical representation of sine and cosine.

The sine and cosine functions are *periodic*. In particular, we always have

$$\sin(x + 2\pi) = \sin x, \quad \text{and} \quad \cos(x + 2\pi) = \cos x.$$

This means that we have infinitely many choices for representing a point on the unit circle using sine and cosine functions. For instance, recalling an earlier example, we have

$$\begin{aligned}\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = \left(\cos\left(\frac{\pi}{4}\right), \sin\left(\frac{\pi}{4}\right)\right) &= \left(\cos\left(\frac{9\pi}{4}\right), \sin\left(\frac{9\pi}{4}\right)\right) \\ &= \left(\cos\left(\frac{17\pi}{4}\right), \sin\left(\frac{17\pi}{4}\right)\right) \\ &= \dots\end{aligned}$$

However, we will typically use the simplest possible representation of a point on the unit circle, with the angle in the interval $[0, 2\pi)$. In particular, **every point p on the unit circle can be expressed *uniquely* in the form $p = (\cos x, \sin x)$ for some $x \in [0, 2\pi)$.**

By the periodicity of sine and cosine, as well as observing that the graph of one function can be obtained by shifting the other, we obtain the following simple calculation rules.

$$\begin{aligned}\sin(-x) &= -\sin x \\ \sin\left(x + \frac{\pi}{2}\right) &= \cos x \\ \sin(x + \pi) &= -\sin x \\ \sin(x + 2\pi) &= \sin x \\ \cos(-x) &= \cos x \\ \cos\left(x + \frac{\pi}{2}\right) &= -\sin x \\ \cos(x + \pi) &= -\cos x \\ \cos(x + 2\pi) &= \cos x.\end{aligned}$$

Some important values of the sine and cosine functions are listed in the following table.

ϕ	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{3\pi}{4}$	$\frac{5\pi}{6}$	π
$\sin\phi$	0	$\frac{1}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{3}}{2}$	1	$\frac{\sqrt{3}}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	0
$\cos\phi$	1	$\frac{\sqrt{3}}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{1}{\sqrt{2}}$	$-\frac{\sqrt{3}}{2}$	-1

We can use the results from the table above along with the fact that $\sin(x + \pi) = -\sin(x)$ and $\cos(x + \pi) = -\cos(x)$ to deduce some of the corresponding values in the interval $(\pi, 2\pi)$. In fact, we can deduce most of these trigonometric values by using just a few assumptions, as the following exercise shows.

Exercise - You are given the information that

$$\sin\left(\frac{\pi}{6}\right) = \frac{1}{2}, \quad \sin\left(\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}}, \quad \text{and} \quad \sin\left(\frac{\pi}{3}\right) = \frac{\sqrt{3}}{2}$$

and the following facts about the sine and cosine functions:

$$\begin{aligned}\sin(-x) &= -\sin x \\ \sin\left(x + \frac{\pi}{2}\right) &= \cos x \\ \cos\left(x + \frac{\pi}{2}\right) &= -\sin x.\end{aligned}$$

Using only these facts, prove that

$$\begin{aligned}
\cos\left(\frac{\pi}{6}\right) &= \frac{\sqrt{3}}{2},\\
\cos\left(\frac{\pi}{4}\right) &= \frac{1}{\sqrt{2}},\\
\cos\left(\frac{\pi}{3}\right) &= \frac{1}{2},\\
\sin\left(\frac{2\pi}{3}\right) &= \frac{\sqrt{3}}{2},\\
\sin\left(\frac{3\pi}{4}\right) &= \frac{1}{\sqrt{2}},\\
\sin\left(\frac{5\pi}{6}\right) &= \frac{1}{2}.
\end{aligned}$$

We also recall some important **trigonometric identities**. Most importantly of all, for all $x \in \mathbb{R}$,

$$\sin^2 x + \cos^2 x = 1.$$

We will also need the **trigonometric addition formulae**.

$$\begin{aligned}
\sin(x+y) &= \sin x \cos y + \cos x \sin y,\\
\cos(x+y) &= \cos x \cos y - \sin x \sin y.
\end{aligned}$$

Based on sine and cosine, we can define $\tan x$ by

$$\tan x := \frac{\sin x}{\cos x}, \quad \forall\, x \neq \frac{\pi}{2} + k\pi, k \in \mathbb{Z}.$$

This restriction is necessary, since $\cos\left(\frac{\pi}{2} + k\pi\right) = 0$ for all $k \in \mathbb{Z}$. The function $\tan x$ is not defined at these values. So, tan is a function with domain $\mathbb{R} \setminus \{\frac{\pi}{2} + k\pi : k \in \mathbb{Z}\}$ and codomain $\mathbb{R}$.

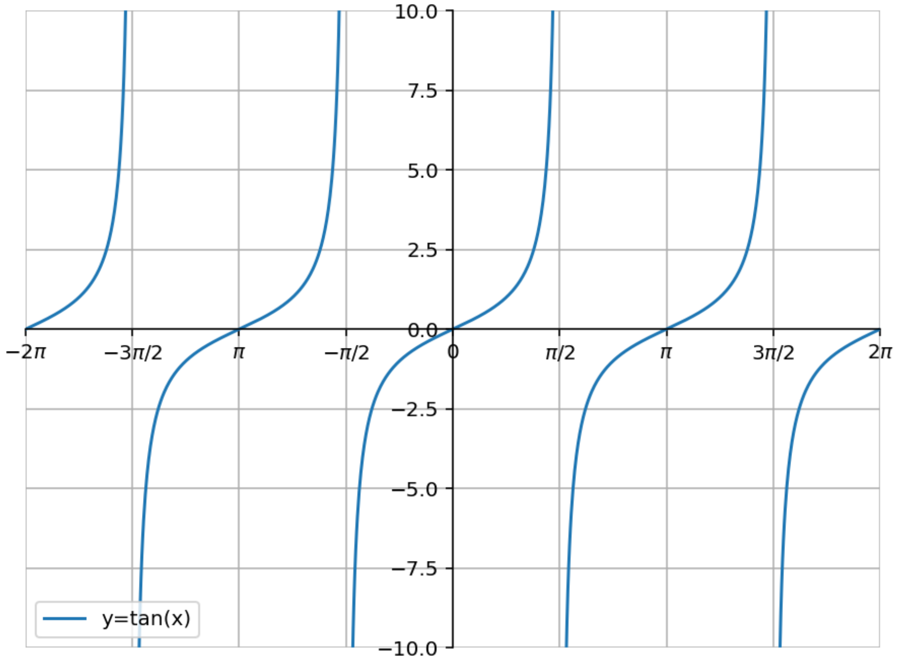


Figure 19: Graphical representation of tan.

Additionally, we can define the inverse trigonometric functions. Let's first consider the inverse of the cosine function. We know from our earlier work (see Theorem 1.24) that a function is invertible if and only if it is a bijection. We need to take a little care, since the cosine function is not a bijection (it is not injective, for instance $\cos 0 = \cos(2\pi) = \cos(4\pi) = \dots$).

However, if we restrict the domain of the cosine function suitably, we can make it into a bijection. In particular, one may see from the graph of the cosine function that the function restricted to the domain $[0, \pi]$ and codomain $[-1, 1]$ is a bijection. We can therefore define the **arccosine** function

$$\arccos : [-1, 1] \to [0, \pi].$$

The function is defined by

$$y = \arccos x \iff x = \cos y \text{ and } y \in [0, \pi].$$

The arccosine function is just a special name for the inverse of the cosine function. We also use the intuitive notation $\cos^{-1}$ for the same function.

By making a similar restriction of the relevant domains, we define the **arcsine** function

$$\arcsin : [-1, 1] \to \left[-\frac{\pi}{2}, \frac{\pi}{2}\right],$$

and the **arctangent** function

$$\arctan : \mathbb{R} \to \left(-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

via

$$y = \arcsin x \iff x = \sin y \text{ and } y \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right],$$
$$y = \arctan x \iff x = \tan y \text{ and } y \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

1.9 Complex numbers

In section 1.4, we discussed how we extend our number system gradually from the set of natural numbers to the set of real numbers in order to have the tools to solve certain equations. However, at a certain point in mathematics we cannot answer all of the questions that arise using only real numbers. This point is reached when we want to find a solution of the equation

$$x^2 = -1, \tag{29}$$

which has no solution $x \in \mathbb{R}$.

We are not satisfied with this situation, and so we introduce a new element to solve the equation. Define

$$i := \sqrt{-1}.$$

Then, by definition, the equation (29) has two solutions, namely $\pm i$. This extension of the real numbers leads us to the (field of) complex numbers.

Definition 1.47. *The set of all* ***complex numbers*** *is defined by*

$$\mathbb{C} := \{x + iy : x, y \in \mathbb{R}\}.$$

We often use z to denote a generic element of $\mathbb{C}$ (so $z = x + iy$ for some $x, y \in \mathbb{R}$).

For a complex number $z = x + iy$ we call x the ***real part*** *of z and y the* ***imaginary part****. We write $Re(z) = x$ and $Im(z) = y$.*

Let $z = x + iy \in \mathbb{C}$. We define the ***complex conjugate*** *of z by*

$$\bar{z} := x - iy.$$

Note that the imaginary part of a complex number is in fact a real number!

The representation of the term $z = x + iy$ is called the **canonical representation**. There are alternative ways to express a generic complex numbers, as we shall see later.

Let's first discuss how to do some simple calculations with complex numbers. We will always follow the same rules that we are familiar with for making calculations with real numbers. Let $z = x + iy$ and $w = u + iv$ be complex numbers (where x, y, u and v are real numbers). Addition is straightforward, as we collect the real and imaginary parts together:

$$z + w = (x + iy) + (u + iv) = (x + u) + i(y + v).$$

For multiplication, we follow familiar rules for expanding multiplication with brackets, simplifying whenever possible by making use of the fact that $i^2 = -1$. For example,

$$zw = (x+iy)(u+iv) = xu+i(yu+xv)+i^2yv = xu+i(yu+xv)-yv = (xu-yv)+i(yu+xv). \tag{30}$$

A useful consequence of (30) is that $z\bar{z}$ is always a non-negative real number. Indeed, let $z = x + iy$ be a complex number. Then

$$z\bar{z} = (x + iy)(x - iy) = x^2 + y^2.$$

We stated above that $\mathbb{C}$ is a field. Indeed the operations of addition and multiplication are associative, commutative, distributive. The additive and multiplicative identity elements are the same as they are for the real numbers. The additive inverse of $z = x + iy$ is naturally $-z = -x - iy$. It remains to verify the existence of a multiplicative inverse for $z \neq 0$. We simply define

$$z^{-1} = \frac{1}{x+iy}.$$

We should check that this really is a complex number (i.e. we should check that it can be expressed in the canonical form). For this, we can make use of complex conjugate to transform the denominator into a real number. Indeed,

$$z^{-1} = \frac{1}{x+iy} = \frac{1}{x+iy}\frac{x-iy}{x-iy} = \frac{x-iy}{x^2+y^2} = \frac{x}{x^2+y^2} - i\frac{y}{x^2+y^2}.$$

Hence, $\mathbb{C}$ with these operations is also a field.

Example - Consider $z = 4 + 2i$. Then z is a complex number with $Re(z) = 4$, $Im(z) = 2$, $\bar{z} = 4 - 2i$ and

$$z^{-1} = \frac{4}{4^2+2^2} - i\frac{2}{4^2+2^2} = \frac{1}{5} - i\frac{1}{10}.$$

Exercise - Show that, for any $z_1, z_2 \in \mathbb{C}$,

$$\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2} \text{ and } \overline{z_1 z_2} = \bar{z}_1\bar{z}_2.$$

Show that, for any $z \in \mathbb{C}$,

$$\overline{(\bar{z})} = z.$$

Example - Recall that i is just a symbol for $\sqrt{-1}$. We can use this symbol to express square roots of other negative real numbers. For example, we have

$$\sqrt{-4} = \sqrt{(-1)(4)} = \sqrt{-1}\sqrt{4} = 2i.$$

Moreover, for arbitrary positive real number a, we have

$$\sqrt{-a} = i(\sqrt{a}).$$

We can even calculate the square root of a complex number, and this will always be another complex number!

Example - We would like to calculate $\sqrt{i}$. Given the sentence before this example, let us assume that this is another complex number $z = x + iy$ (with $x, y \in \mathbb{R}$). We have $z^2 = i$, which means that

$$i = (x+iy)^2 = x^2 - y^2 + i(2xy). \tag{31}$$

If we compare the real and imaginary parts of both sides of (31), we arrive at a system of two equations:

$$\begin{aligned} x^2 - y^2 &= 0, \\ 2xy &= 1. \end{aligned}$$

We seek to solve this for x and y. The first equation tells us that either $x = y$ or $x = -y$. However, if $x = -y$ then we have a contradiction with the second equation (since the left hand side is not positive), and so it must be the case that $x = y$.

The second equation then becomes $2x^2 = 1$, and so either $x = \frac{1}{\sqrt{2}}$ or $x = -\frac{1}{\sqrt{2}}$. It follows that

$$\left(\frac{1}{\sqrt{2}} + i\frac{1}{\sqrt{2}}\right)^2 = i$$

and

$$\left(-\frac{1}{\sqrt{2}} - i\frac{1}{\sqrt{2}}\right)^2 = i$$

That is,

$$\sqrt{i} = \pm\left(\frac{1}{\sqrt{2}} + i\frac{1}{\sqrt{2}}\right). \tag{32}$$

Let's emphasise again, that even though we have introduced a new number i to our universe, we are still using all of the same rules for calculations that we know from our mathematical education, including all of the basic rules that we have been repeating since early school days. The complex numbers are just an extension of what we already know, we should not think of them as being something completely new. For instance, if we set all of the imaginary parts to be equal to zero, all we have done so far in this section is to make some trivial observations about real numbers.

One can identify an element of the set $\mathbb{C}$ with a tuple (x, y) of real numbers. Each complex number $z = x + iy$ can therefore be illustrated as a point in the plane, which is called the **complex plane**. The coordinate axes are the called real and imaginary axis.

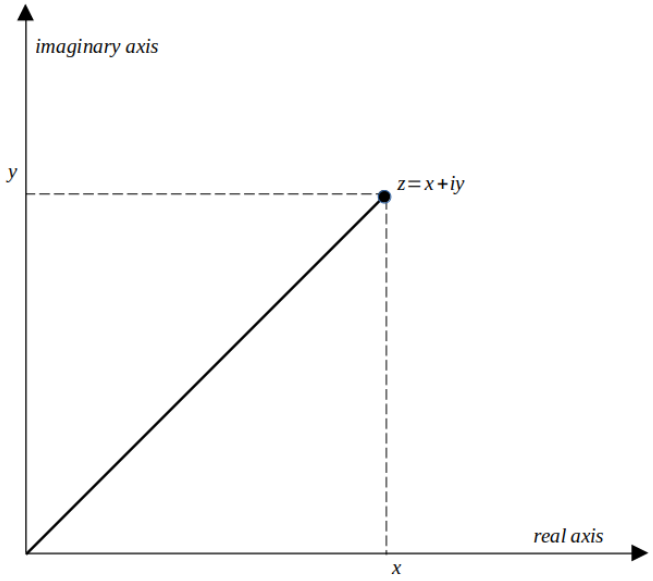


Figure 20: Illustration of the complex plane.

Exercise - Think about where to draw the complex conjugate $\bar{z}$. What about $-z$?

We define the absolute value of a complex number $z = x + iy$ to be the length of the straight line from $(0, 0)$ to (x, y). Here is a formulaic version of this definition.

Definition 1.48. *Let $z = x + iy \in \mathbb{C}$. We define the* ***absolute value*** *of z to be*

$$|z| := \sqrt{z\bar{z}} = \sqrt{x^2 + y^2}.$$

In particular, note that $|z|$ is always a real number, since $z\bar{z}$ is always a positive real number.

We have used the same notation here as we used for the absolute value of a real number in Section 1.8. This is very deliberate! The definition of the absolute value for a complex number is an extension of the previous definition over $\mathbb{R}$. Indeed, if z is a real number (i.e. the imaginary part is equal to 0), then according to Definition 1.48,

$$|z| = \sqrt{x^2 + y^2} = \sqrt{z^2}.$$

This is equal z if $z \geq 0$ and is equal to $-z$ otherwise, which matches the definition given in (24).

Several calculation rules for the absolute value are just the same as for the absolute value over $\mathbb{R}$. The next lemma is an analogue (and extension) of Lemma 1.44.

Lemma 1.49. *The absolute value function satisfies the following properties.*

1. *For all $z \in \mathbb{C}$, $|z| \geq 0$.*
2. $|z| = 0 \iff z = 0$.
3. *For all $z \in \mathbb{C}$, $|z| \geq Re(z)$ and $|z| \geq -Re(z)$. Also,*
$$Re(z) = |z| \iff z \in \mathbb{R} \text{ and } z \geq 0.$$
4. *For all $z \in \mathbb{C}$, $|z| \geq Im(z)$ and $|z| \geq -Im(z)$. Also,*
$$Im(z) = |z| \iff z = iy, y \in \mathbb{R}, \text{ and } y \geq 0.$$
5. *For all $z, w \in \mathbb{C}$, $|zw| = |z||w|$.*
6. *For all $z \in \mathbb{C}$, $|z| = |\bar{z}|$.*
7. *For all $z \in \mathbb{C}$, $|z| = |-z|$.*

Proof. 1. This follows immediately from the definition of the absolute value, since we take the positive square root.

2. This is left as an exercise.

3. Write $z = x + iy$. To prove that $|z| \geq Re(z)$ we need to show that
$$|z| = \sqrt{x^2 + y^2} \geq x. \qquad (33)$$
This is true, because y^2 is non-negative.

For the if and only if statement, let us first verify the "$\Leftarrow$" direction. If z is real and non-negative then $y = 0$ and so $x \geq 0$. Then
$$|z| = \sqrt{x^2 + y^2} = \sqrt{x^2} = |x| = x = Re(z),$$

as required. For the "$\Rightarrow$" direction, suppose that $Re(z) = |z|$. That is,

$$\sqrt{x^2 + y^2} = x.$$

Then it must be the case that $y^2 = 0$ (otherwise we would have $\sqrt{x^2 + y^2} > x$. Also, since $\sqrt{x^2 + y^2} \geq 0$, it must be the case that $x \geq 0$, as required.

4. This is left as an exercise.

5. Write $z = x + iy$ and $w = u + iv$. It follows from the formula for multiplication of complex numbers given in (30), along with the definition of the absolute value, that

$$\begin{aligned}
|zw| &= \sqrt{(xu - yv)^2 + (yu + xv)^2} \\
&= \sqrt{x^2u^2 + y^2v^2 - 2xuyv + y^2u^2 + x^2v^2 + 2yuxv} \\
&= \sqrt{x^2u^2 + y^2v^2 + y^2u^2 + x^2v^2} \\
&= \sqrt{(x^2 + y^2)(v^2 + u^2)} \\
&= \sqrt{x^2 + y^2}\sqrt{v^2 + u^2} \\
&= |z||w|.
\end{aligned}$$

6. This is left as an exercise.

7. This is left as an exercise.

□

The triangle inequality is also valid for the absolute value function, and it is also a very important and useful result.

Theorem 1.50. *Let $z, w \in \mathbb{C}$. Then we have*

$$|z + w| \leq |z| + |w|. \tag{34}$$

We also have

$$|z + w| = |z| + |w| \iff z\bar{w} \textit{ is a real number and } z\bar{w} \geq 0. \tag{35}$$

We also have

$$\big||z| - |w|\big| \leq |z - w|. \tag{36}$$

Proof. We will prove (34) and (35) simultaneously. First observe that

$$|z + w|^2 = (z + w)(\overline{z + w}) = (z + w)(\bar{z} + \bar{w}).$$

Here we have used the fact that $\overline{z + w} = \bar{z} + \bar{w}$, which was an exercise on page 50. It then follows that

$$|z + w|^2 = z\bar{z} + (z\bar{w} + \bar{z}w) + w\bar{w} = |z|^2 + (z\bar{w} + \bar{z}w) + |w|^2. \tag{37}$$

We can calculate directly that

$$z\bar{w} + \bar{z}w = 2Re(z\bar{w}). \tag{38}$$

Indeed, for any $a \in \mathbb{C}$, we have

$$a + \overline{a} = 2Re(a). \tag{39}$$

Now apply (39) with $a = z\overline{w}$. Since $\bar{z}w = \overline{z\overline{w}}$, this proves (38).

Combining (38) with (37) and then applying points 3, 5 and 6 from Lemma 1.49, we have

$$\begin{aligned} |z+w|^2 = |z|^2 + (z\bar{w} + \bar{z}w) + |w|^2 &= |z|^2 + 2Re(z\overline{w}) + |w|^2 \\ &\leq |z|^2 + 2|z\overline{w}| + |w|^2 \\ &= |z|^2 + 2|z||\overline{w}| + |w|^2 \\ &= |z|^2 + 2|z||w| + |w|^2 \\ &= (|z| + |w|)^2. \end{aligned}$$

This proves (34). Note also that there is only one inequality in this sequence, which was an application of Lemma 1.49, point 3. If we apply the additional information from Lemma 1.49, point 3, we see that this inequality is an equality if and only if $z\bar{w} \in \mathbb{R}$ and $z\bar{w} \geq 0$. This proves (35).

We now turn to the proof of (36). After relabelling the variables and rearranging the expression, it follows from (34) that

$$|a + b| - |b| \leq |a| \tag{40}$$

for all $a, b \in \mathbb{C}$. Applying (40) with $a = z - w$ and $b = w$, we get

$$|z| - |w| \leq |z - w|. \tag{41}$$

Applying (40) with $a = z - w$ and $b = -z$, we get

$$|w| - |z| = |-w| - |-z| \leq |z - w|. \tag{42}$$

In the equality above, we have used point 7 of Lemma 1.49.

Finally, note that $|z| - |w|$ is a real number, and therefore the absolute value $\big||z| - |w|\big|$ is equal to either $|z| - |w|$ or $-(|z| - |w|) = |w| - |z|$. In either of these cases, (41) or (42) give the required bound $\big||z| - |w|\big| \leq |z - w|$.

□

Once again, the triangle inequality (inequality 34) also implies that

$$|a - b| = |a - c + c - b| \leq |a - c| + |c - b|, \tag{43}$$

for all $a, b, c \in \mathbb{C}$.

The triangle inequality in the form $|a - b| \leq |a - c| + |c - b|$ can be interpreted geometrically as saying that the fastest route from a to b is to go directly, rather than travelling via the point c. This interpretation also gives some justification for why we call this the triangle inequality.

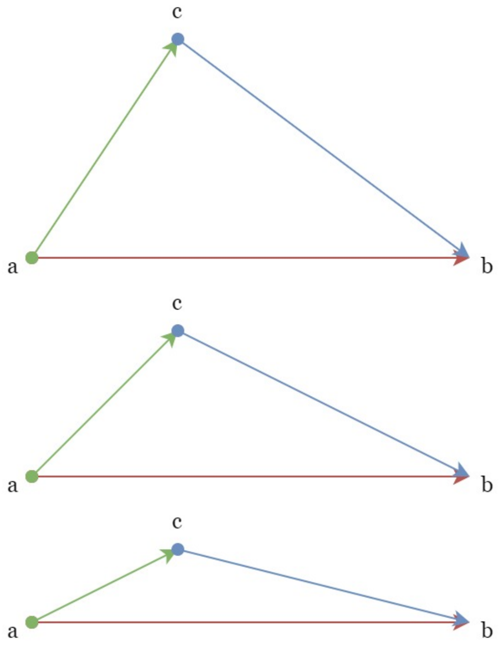


Figure 21: The Triangle Inequality. A visualization of equation 43.

The following table summarises the representation and visualisation of complex numbers in the complex plane.

$\mathbb{C}$	The plane $\mathbb{R}^2$
$z = x + iy$	point with coordinates (x, y)
$Re(z)$	x-coordinate of the corresponding point
$Im(z)$	y-coordinate of the corresponding point
The set of real numbers	points on the x-axis
The set of imaginary numbers	points on the y-axis
$-z$	z after rotation by 180 degrees centred at the origin
$\bar{z}$	z reflected in the x-axis
$\lvert z\rvert$	the distance between z and the origin
$\lvert z_1 - z_2\rvert$	the distance between z_1 and z_2
$z_1 + z_2$	vector addition of the corresponding points
$\lvert z_1 + z_2\rvert \leq \lvert z_1\rvert + \lvert z_2\rvert$	the triangle inequality

Instead of using the Cartesian coordinates (x, y) for a complex number $z = x+iy$, one can also switch to **polar coordinates** (r, ϕ). We can identify a complex number by its “direction” (or angle) from the origin, and its distance from the origin. Once we know these two properties of a complex number, we know exactly what the number is.

For example, suppose that we are told that a complex number z has absolute value $|z| = 2$ and makes an angle of $\frac{\pi}{4}$ (i.e. 45 degrees) anticlockwise from the real axis (i.e. the x-axis). We can use some simple trigonometry to calculate the canonical form $z = x + iy$. Indeed, we have

$$\sin(\pi/4) = \frac{y}{2}, \quad \cos(\pi/4) = \frac{x}{2},$$

and since $\sin(\pi/4) = \cos(\pi/4) = \frac{\sqrt{2}}{2}$, it follows that

$$z = \sqrt{2} + \sqrt{2}i = 2\left(\frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}\right).$$

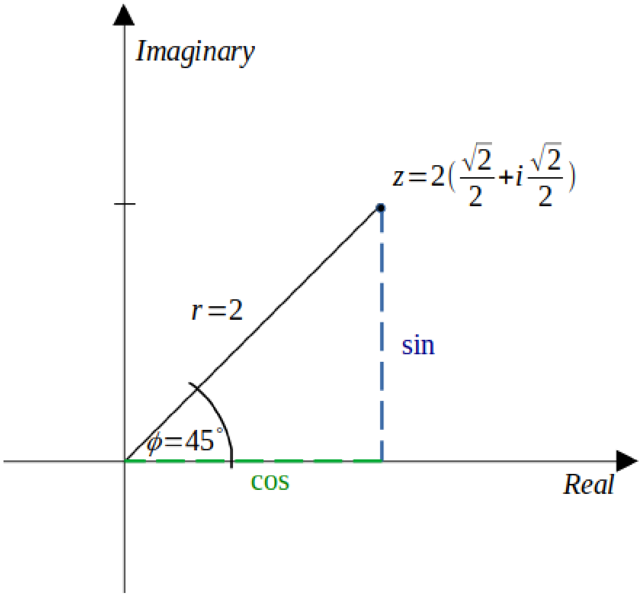


Figure 22: Illustration of the point z determined by its polar coordinates.

The choice of the factorisation may seem a little strange here, but there is a reason for this, as should become clear soon.

Recall from Section 1.8 that every point on the unit circle $\{(x, y) : x^2 + y^2 = 1\}$ can be written uniquely as $(x, y) = (\cos\phi, \sin\phi)$ for some $\phi \in [0, 2\pi)$. Putting this fact into the context of the complex plane $\mathbb{C}$, it follows that every complex number z such that $|z| = 1$ can be expressed uniquely as

$$z = \cos\phi + i\sin\phi, \tag{44}$$

for some $\phi \in [0, 2\pi)$. This ϕ tells us the direction of a "unit" complex number. For a generic complex number $z \neq 0$ (i.e. not necessarily with $|z| = 1$), we introduce a dilate r which tells us the distance between z and the origin (i.e. the value of $|z|$).

To summarise, we can uniquely express an arbitrary nonzero complex number z in the form

$$z = r(\cos\phi + i\sin\phi),$$

for some $r \in \mathbb{R}_+$ and $\phi \in [0, 2\pi)$. The values r and ϕ have a geometric meaning; r is the distance of z from the origin (the **absolute value** or **magnitude** of z) and ϕ is the (anticlockwise) angle determined by the real axis and z. We call ϕ the **argument** of z.

An easy, useful and compact way of writing complex numbers with the form of (44) can be obtained by using **Euler's formula**, which states that, for all $\phi \in \mathbb{R}$,

$$e^{i\phi} := \cos\phi + i\sin\phi. \tag{45}$$

Note that $e^{i\phi}$ can at the moment only be understood as a symbol for the right hand side above, and it is already useful as such. However, we will see later that this is really an equation, i.e., we will also define e^z for any $z \in \mathbb{C}$ in a way that is consistent with the case when z is real, and show that Euler's formula is true for this definition of a complex exponential.

Examples - By using Euler's formula and some familiar trignometric values, we can immediately see that

$$\begin{aligned} e^{i\frac{\pi}{2}} &= \cos\left(\frac{\pi}{2}\right) + i\sin\left(\frac{\pi}{2}\right) = i, \\ e^{i\pi} &= \cos(\pi) + i\sin(\pi) = -1, \\ e^{i\frac{3\pi}{2}} &= \cos\left(\frac{3\pi}{2}\right) + i\sin\left(\frac{3\pi}{2}\right) = -i, \\ e^{i2\pi} &= \cos(2\pi) + i\sin(2\pi) = 1. \end{aligned}$$

Additionally, we obtain from these values, together with the periodicity of the trigonometric functions (or standard calculation rules for exponentials), that for $k \in \mathbb{Z}$,

$$e^{ik\pi} = \begin{cases} 1 & \text{if } k \text{ is even} \\ -1 & \text{if } k \text{ is odd} \end{cases}.$$

Using Euler's formula we can write every complex number (in its polar form) as

$$z = re^{i\phi}. \tag{46}$$

The polar representation in this form is particularly useful when it comes to multiplication and powers of complex numbers. Given two complex numbers $z_1 = r_1 e^{i\phi_1}$ and $z_2 = r_2 e^{i\phi_2}$, we obtain

$$z_1 \cdot z_2 = (r_1 e^{i\phi_1})(r_2 e^{i\phi_2}) = r_1 r_2 e^{i(\phi_1+\phi_2)}.$$

This formula is often easier to deal with and more intuitive than the formula for multiplication of complex numbers in canonical form given in (30).

From this formula (with $z = w$) and induction we obtain **de Moivre's formula** for powers z^n of $z = r(\cos\phi + i\sin\phi) = re^{i\phi}$ with $n \in \mathbb{N}$:

$$z^n = r^n e^{in\phi} = r^n(\cos(n\phi) + i\sin(n\phi)).$$

Example - As an example we calculate $(1+i)^{42}$. We set $z := 1+i$. In order to use de Moivre's formula, we must first express z in to its trigonometric (polar) form. We obtain

$$z = \sqrt{2}e^{i\frac{\pi}{4}}.$$

You should verify this in detail! By de Moivre's formula and the periodicity of the sine and cosine functions, we get

$$\begin{aligned} z^{42} = (\sqrt{2})^{42}\left(\cos\left(42\frac{\pi}{4}\right) + i\sin\left(42\frac{\pi}{4}\right)\right) &= 2^{21}\left(\cos\left(\frac{\pi}{2}\right) + i\sin\left(\frac{\pi}{2}\right)\right) \\ &= 2^{21}i. \end{aligned}$$

Exercise - Suppose that $z, w \in \mathbb{C} \setminus \{0\}$. Show that $|z + w| = |z| + |w|$ if and only if z and w have the same argument. (Hint: use Theorem 1.50.)

It is an important result in mathematics, that complex numbers are really all we need to solve polynomial equations. In fact, the *Fundamental theorem of algebra* (in one of its variants) even gives the precise answer for the number of solutions of polynomial equations, the number of solutions to a polynomial degree d polynomial equation with complex coefficients is exactly d. We do not discuss this in detail here, but illustrate with an example.

Example - We seek to find all solutions $x \in \mathbb{C}$ to the equation.

$$x^2 + (1 - i)x - i = 0.$$

This is a quadratic equation, or in other words, a polynomial of degree 2. So, there should be in general be two solutions $x \in \mathbb{C}$. We can find them by using the *quadratic formula*, which implies that

$$x = \frac{i - 1 \pm \sqrt{(1-i)^2 + 4i}}{2} = \frac{i - 1 \pm \sqrt{2i}}{2} = \frac{i - 1 \pm \sqrt{2}\sqrt{i}}{2}.$$

Recalling from (32) that

$$\sqrt{i} = \left(\frac{1}{\sqrt{2}} + i\frac{1}{\sqrt{2}}\right),$$

we conclude that the solutions to our equation are

$$x = \frac{i - 1 + \sqrt{2}\left(\frac{1}{\sqrt{2}} + i\frac{1}{\sqrt{2}}\right)}{2} = i$$

and

$$x = \frac{i - 1 + \sqrt{2}\left(-\frac{1}{\sqrt{2}} - i\frac{1}{\sqrt{2}}\right)}{2} = -1.$$

1.10 Vectors and norms

For the final section of this introductory chapter, we will discuss some important concepts related to **vectors**. Let $d \in \mathbb{N}$. A vector is a tuple

$$\mathbf{v} = (v_i)_{i=1}^d = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{pmatrix} \in \mathbb{C}^d.$$

The elements v_i are complex numbers. The **dimension** of $\mathbf{v}$ is d. If all of the v_i are equal to zero then we call $\mathbf{v}$ the **zero vector**, which is denoted as $\mathbf{0}$.

We define the **addition** and **scalar multiplication** of vectors **component-wise**. That is, for two vectors

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{pmatrix},$$

and a number $\lambda \in \mathbb{C}$, we define

$$\mathbf{u} + \mathbf{v} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{pmatrix} = \begin{pmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_d + v_d \end{pmatrix} \quad \text{and} \quad \lambda\mathbf{u} = \begin{pmatrix} \lambda u_1 \\ \lambda u_2 \\ \vdots \\ \lambda u_d \end{pmatrix}.$$

Note that it is important for vector addition that the vectors have the same dimension. If the dimensions of the two vectors do not agree, then their sum is not defined. The term “scalar” is used to distinguish this multiplication from the other notions of multiplication with vectors to be discussed later.

Here we use the field of complex numbers $\mathbb{C}$ to build complex vectors $\mathbf{v} \in \mathbb{C}^d$. Note that we can easily also consider only real vectors $\mathbf{v} \in \mathbb{R}^d$, as real vectors are a special case of complex vectors. However, since all definitions here work directly in the complex case, and this will be needed later, we define it in the more general context and comment on necessary changes when needed. Moreover, note that we could define vectors, and the corresponding operations, in a much more general context, as long as the operations for the components are well-defined. This will be discussed much later when we come to *vector spaces.*

For real and complex numbers, we defined the absolute value to give a notion of how “large” a number is. We will now do something similar for vectors.

Definition 1.51. *Let $\boldsymbol{v} = (v_i)_{i=1}^d \in \mathbb{C}^d$. Then, we define the* ***Euclidean norm*** *or* ***length*** *of $\boldsymbol{v}$ by the formula*

$$\|\boldsymbol{v}\|_2 := \sqrt{\sum_{i=1}^d |v_i|^2}.$$

Moreover, for two vectors $\boldsymbol{u} = (u_i)_{i=1}^d$ *and* $\boldsymbol{v} = (v_i)_{i=1}^d \in \mathbb{C}^d$ *we define the* ***inner product*** *or* ***dot product*** *of* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *by the formula*

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle := \sum_{i=1}^d u_i \bar{v}_i.$$

With this, we have the useful representation

$$\|\boldsymbol{v}\|_2 = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle}.$$

For real vectors $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$, these definitions simplify to

$$\|\mathbf{v}\|_2 := \sqrt{\sum_{i=1}^d v_i^2}$$

and

$$\langle \mathbf{u}, \mathbf{v} \rangle := \sum_{i=1}^d u_i v_i.$$

First of all, let us note that the notion of the Euclidean norm really does generalise the concept of the absolute value of a complex number discussed earlier. To see this, consider the case of a one-dimensional vector $\mathbf{v} \in \mathbb{C}^1$. This "vector" is in reality just a complex number $v_1 = x + iy$, with some $x, y \in \mathbb{R}$. Then applying Definition 1.51, we see that

$$\|\mathbf{v}\|_2 := \sqrt{\sum_{i=1}^d |v_i|^2} = \sqrt{|v_1|^2} = |v_1|.$$

We use the subscript 2 here, because we will study also other norms later in the Mathematics for AI syllabus. Note that some authors use the notation $\|\cdot\|$ for the Euclidean norm (i.e. they remove the subscript), which emphasises its role as generalization of the absolute value of a real or complex number.

The inner product is formally a mapping $\langle \cdot, \cdot \rangle : \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}$.

Exercise - Prove that the inner product satisfies the following properties.

1. **Positive definiteness**: for all $\mathbf{u} \in \mathbb{C}^d$, $\langle \mathbf{u}, \mathbf{u} \rangle \in \mathbb{R}$ and $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$. Also,

$$\langle \mathbf{u}, \mathbf{u} \rangle = 0 \iff \mathbf{u} \text{ is the zero vector.}$$

2. **Linearity in the first argument**: for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{C}^d$ and $\lambda \in \mathbb{C}$

$$\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle \tag{47}$$

and

$$\langle \lambda \mathbf{u}, \mathbf{v} \rangle = \lambda \langle \mathbf{u}, \mathbf{v} \rangle. \tag{48}$$

3. **Conjugate symmetry**: for all $\mathbf{u}, \mathbf{v} \in \mathbb{C}^d$,

$$\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$$

We can use the previous three properties of the inner product to prove the following.

Lemma 1.52. *1. For all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^d$ and $\mu, \lambda \in \mathbb{C}$,*

$$\langle \lambda \boldsymbol{u} + \mu \boldsymbol{v}, \boldsymbol{w} \rangle = \lambda \langle \boldsymbol{u}, \boldsymbol{w} \rangle + \mu \langle \boldsymbol{v}, \boldsymbol{w} \rangle.$$

2. For all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^d$ and $\mu, \lambda \in \mathbb{C}$,

$$\langle \boldsymbol{u}, \lambda \boldsymbol{v} + \mu \boldsymbol{w} \rangle = \overline{\lambda} \langle \boldsymbol{u}, \boldsymbol{v} \rangle + \overline{\mu} \langle \boldsymbol{u}, \boldsymbol{w} \rangle.$$

Proof. 1. By linearity in the first argument,

$$\langle \lambda \mathbf{u} + \mu \mathbf{v}, \mathbf{w} \rangle = \langle \lambda \mathbf{u}, \mathbf{w} \rangle + \langle \mu \mathbf{v}, \mathbf{w} \rangle = \lambda \langle \mathbf{u}, \mathbf{w} \rangle + \mu \langle \mathbf{v}, \mathbf{w} \rangle.$$

2. By conjugate symmetry

$$\langle \mathbf{u}, \lambda \mathbf{v} + \mu \mathbf{w} \rangle = \overline{\langle \lambda \mathbf{v} + \mu \mathbf{w}, \mathbf{u} \rangle}. \tag{49}$$

We then apply part 1 of this lemma, and some basic properties of the complex conjugate (see the exercise on page 51) to obtain

$$\overline{\langle \lambda \mathbf{v} + \mu \mathbf{w}, \mathbf{u} \rangle} = \overline{\lambda \langle \mathbf{v}, \mathbf{u} \rangle + \mu \langle \mathbf{w}, \mathbf{u} \rangle} = \overline{\lambda} \cdot \overline{\langle \mathbf{v}, \mathbf{u} \rangle} + \overline{\mu} \cdot \overline{\langle \mathbf{w}, \mathbf{u} \rangle}. \tag{50}$$

Using conjugate symmetry again and the fact that $\overline{(\overline{z})} = z$, (50) implies that

$$\overline{\langle \lambda \mathbf{v} + \mu \mathbf{w}, \mathbf{u} \rangle} = \overline{\lambda} \cdot \langle \mathbf{u}, \mathbf{v} \rangle + \overline{\mu} \cdot \langle \mathbf{u}, \mathbf{w} \rangle$$

After combining this with (49), the proof is complete. □

Note that, taking $\lambda = \mu = 1$ in part 2 of Lemma 1.52, obtain

$$\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle.$$

This is very similar to the identity 47 from the previous exercise. An analogue of (48) is the identity

$$\langle \mathbf{u}, \lambda \mathbf{v} \rangle = \overline{\lambda} \langle \mathbf{u}, \mathbf{v} \rangle.$$

This is obtained from part 2 of Lemma 1.52 by fixing $\mu = 0$.

Exercise - Prove that, for any $\lambda \in \mathbb{C}$ and $\mathbf{v} \in \mathbb{C}^d$, we have

$$\|\lambda \mathbf{v}\|_2 = |\lambda| \|\mathbf{v}\|_2.$$

We have already seen that the Euclidean norm generalises many features of the aboslute value function to the higher dimensional setting. The next result, the **triangle inequality**, is another important step in this direction.

Theorem 1.53. *For all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^d$,*

$$\|\boldsymbol{u} + \boldsymbol{v}\|_2 \leq \|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2.$$

Note that in the case that $d = 2$ or $d = 3$ the Euclidean norm coincides with the usual intuition of the distance between the origin and the point $\mathbf{v}$, i.e. $\|\mathbf{v}\|$ is the length of the 'direct path' between the origin and $\mathbf{v}$. This makes the triangle inequality appear to be an obvious statement. However, in higher dimensions (particularly when $d > 3$) this is not so clear. We present a proof below.

First we need another inequality, the **Cauchy-Schwarz inequality**, which is one of the most important inequalities in mathematics. It continues to be applied extremely frequently in active research across almost all areas of mathematics, and many great and celebrated results essentially come down to skillful manipulations of the bound.

Theorem 1.54 (Cauchy-Schwarz inequality)**.** *For all* $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^d$,

$$|\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \leq \|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2.$$

Proof. Write

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{pmatrix}.$$

Case 1: Suppose that either $\mathbf{u} = 0$ or $\mathbf{v} = 0$. Then both sides of the inequality are equal to zero, and thus the result is valid.

Case 2: Suppose that $\mathbf{u}, \mathbf{v} \neq 0$. We will make use of the fact that the inequality

$$\frac{a^2 + b^2}{2} \geq ab \tag{51}$$

holds for all $a, b \in \mathbb{R}$. The inequality (51) can be proved by rearranging the inequality $(a-b)^2 \geq 0$.

We define real numbers

$$a_i = \frac{|u_i|}{\|\mathbf{u}\|_2}, \text{ and } b_i = \frac{|v_i|}{\|\mathbf{v}\|_2}.$$

Observe that

$$\sum_{i=1}^{d} a_i^2 = \sum_{i=1}^{d} \frac{|u_i|^2}{\|\mathbf{u}\|_2^2} = \frac{1}{\|\mathbf{u}\|_2^2} \sum_{i=1}^{d} |u_i|^2 = 1.$$

Similarly, $\sum_{i=1}^{d} b_i^2 = 1$. It therefore follows that

$$\sum_{i=1}^{d} \frac{a_i^2 + b_i^2}{2} = 1. \tag{52}$$

We then use the triangle inequality for $\mathbb{C}$ (inequality 34 of Theorem 1.50), along with (51) and (52), to conclude that

$$\begin{aligned}
|\langle \mathbf{u}, \mathbf{v}\rangle| = \left|\sum_{i=1}^{d} u_i \bar{v}_i\right| \leq \sum_{i=1}^{d} |u_i \bar{v}_i| &= \sum_{i=1}^{d} |u_i||v_i| \\
&= \sum_{i=1}^{d} a_i \|\mathbf{u}\|_2 b_i \|\mathbf{v}\|_2 \\
&= \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \sum_{i=1}^{d} a_i b_i \\
&\leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \sum_{i=1}^{d} \frac{a_i^2 + b_i^2}{2} \\
&= \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 .
\end{aligned}$$

□

The Cauchy-Schwarz inequality can now be used to prove the triangle inequality for the Euclidean norm.

Proof of Theorem 1.53. By the properties of the inner product established in Lemma 1.52 and the exercise before it, it follows that for all $\mathbf{u}, \mathbf{v} \in \mathbb{C}^d$,

$$\begin{aligned}
\|\mathbf{u} + \mathbf{v}\|_2^2 &= \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v}\rangle \\
&= \langle \mathbf{u}, \mathbf{u}\rangle + \langle \mathbf{u}, \mathbf{v}\rangle + \langle \mathbf{v}, \mathbf{u}\rangle + \langle \mathbf{v}, \mathbf{v}\rangle \\
&= \langle \mathbf{u}, \mathbf{u}\rangle + \langle \mathbf{u}, \mathbf{v}\rangle + \overline{\langle \mathbf{u}, \mathbf{v}\rangle} + \langle \mathbf{v}, \mathbf{v}\rangle \\
&= \|\mathbf{u}\|_2^2 + \langle \mathbf{u}, \mathbf{v}\rangle + \overline{\langle \mathbf{u}, \mathbf{v}\rangle} + \|\mathbf{v}\|_2^2 \\
&= \|\mathbf{u}\|_2^2 + 2Re\left(\langle \mathbf{u}, \mathbf{v}\rangle\right) + \|\mathbf{v}\|_2^2 .
\end{aligned}$$

We now use Lemma 1.49 (part 3) and the Cauchy-Schwarz Inequality to conclude that

$$\begin{aligned}
\|\mathbf{u} + \mathbf{v}\|_2^2 &= \|\mathbf{u}\|_2^2 + 2Re\left(\langle \mathbf{u}, \mathbf{v}\rangle\right) + \|\mathbf{v}\|_2^2 \\
&\leq \|\mathbf{u}\|_2^2 + 2\left|\langle \mathbf{u}, \mathbf{v}\rangle\right| + \|\mathbf{v}\|_2^2 \\
&\leq \|\mathbf{u}\|_2^2 + 2\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 + \|\mathbf{v}\|_2^2 \\
&= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2 .
\end{aligned}$$

This completes the proof. □

We finally introduce some particularly important vectors which will appear very often in the upcoming considerations. These are the **unit vectors** $\mathbf{e}_1, \ldots, \mathbf{e}_d \in \mathbb{C}^d$, where $\mathbf{e}_k$ is the vector which is zero everywhere except for the kth entry, which is 1. That is,

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots \quad , \mathbf{e}_d = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Note that $\|\mathbf{e}_i\| = 1$ for all $i = 1, \ldots, d$.

One important property of the unit vectors is, that they can be used to represent arbitrary vectors. For this, note that $\lambda \mathbf{e}_k$ with $\lambda \in \mathbb{C}$ is the vector with λ in the kth entry, and zero elsewhere. It is therefore easy to see that

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{pmatrix} = \sum_{i=1}^{d} v_i \mathbf{e}_i.$$

Moreover, the representation of the vector $\mathbf{v}$ in this way is unique. Such a method for representing elements of $\mathbb{C}^d$ uniquely turns out to be very useful and important. A set with such a property is called a **basis** for $\mathbb{C}^d$, and the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ is called the **standard basis**. These concepts will be discussed in greater detail later, when we come to study vector spaces.

2 Matrices and systems of linear equations

In this chapter we want to solve *systems of equations.* That is, we want to find possible values for some variables that fulfill a certain collection of equations. This is one of the most important disciplines in applied mathematics and numerical applications. In particular, we will focus on systems of *linear equations.*

Systems of linear equations are the most frequently occurring type of multivariate problems to solve, and they are also the easiest. Many (even "non-linear") numerical problems can be rewritten as, or approximated by, a system of linear equations. Such systems can be rather large and are usually solved by a computer, but it is up to the user to transfer the problem under consideration to a well-defined linear system. It is therefore indispensable to have a solid understanding of these basic problems.

For one free variable, linear equations are easy to solve. If we want to solve the equation $ax = b$ for some (fixed) $a, b \in \mathbb{R}$, then the unique solution is given by $x = \frac{b}{a}$ if $a \neq 0$. However, if $a = 0$ and $b \neq 0$ then this equation cannot be solved. That is, there is no x satisfying the equation $0 \cdot x = b$. Meanwhile, for $a = b = 0$ the equation $ax = b$ is fulfilled for every $x \in \mathbb{R}$.

Although this situation was straightforward to solve, we still see that there are some subtleties and different cases to consider. In particular, a single linear equation can have a *unique solution*, can have *no solutions*, or can have *infinitely many solutions*, depending on the properties of the equation. It turns out that systems of linear equations also satisfy the same trichotomy.

As we increase the number of variables and equations, things become more complicated. Let us illustrate this with some examples involving two equations and two variables.

Example - Find all solutions $(x_1, x_2) \in \mathbb{R}^2$ to the system of linear equations

$$\begin{aligned} 2x_1 + x_2 &= 1 \\ 6x_1 + 2x_2 &= 2. \end{aligned}$$

We can use the first equation to eliminate one of the variables and reduce this to one equation with one variable. The first equation gives $x_2 = 1 - 2x_1$. Plugging this into the second equation, we obtain $6x_1 + 2(1 - 2x_1) = 2$, which simplifies to $2x_1 = 0$. So, it must be the case that $x_1 = 0$. Plugging this back into the first equation, it follows that $x_2 = 1$, and so the unique solution is $(x_1, x_2) = (0, 1)$.

If we change the system of equations slightly, the set of solutions can change significantly.

Example - Find all solutions $(x_1, x_2) \in \mathbb{R}^2$ to the system of linear equations

$$\begin{aligned} 2x_1 + x_2 &= 1 \\ 6x_1 + 3x_2 &= 2. \end{aligned}$$

Let us try to use the same approach as we used for the previous example. The first equation again gives $x_2 = 1 - 2x_1$. Plugging this into the second equation, we obtain $6x_1 + 3(1 - 2x_1) = 2$, which simplifies to $3 = 2$. It seems that we have reached a nonsensical contradiction!

Indeed, there are no solutions to this system, as the two equations are incompatible. Another way to see this is by multiplying both sides of the first equation by 3. We arrive at the equivalent system

$$\begin{aligned} 6x_1 + 3x_2 &= 3 \\ 6x_1 + 3x_2 &= 2. \end{aligned}$$

A solution to this system would again imply the contradiction $3 = 2$.

Making another small modification changes the story completely again.

Example - Find all solutions $(x_1, x_2) \in \mathbb{R}^2$ to the system of linear equations

$$\begin{aligned} 2x_1 + x_2 &= 1 \\ 6x_1 + 3x_2 &= 3. \end{aligned}$$

Plugging $x_2 = 1 - 2x_1$ into the second equation, we obtain $6x_1 + 3(1 - 2x_1) = 3$, which simplifies to $3 = 3$. This is satisfied for all $x_1 \in \mathbb{R}$.

What is happening in this example is that the two equations are equivalent. This can again be seen by multiplying both sides of the first equation by 3. Therefore, the solutions (x_1, x_2) to this system are the same as the solutions to the equation $2x_1 + x_2 = 1$. We can choose any $x_1 \in \mathbb{R}$ and then choose the corresponding x_2 to satisfy the equation, and so there are infinitely many equations. For instance, $(0, 1)$ and $(1, -1)$ are solutions.

These three examples show that such systems of equations might be quite sensitive to small changes of the parameters (and this was just a two dimensional example). It is therefore desirable to have criteria for a given system of equations to be (uniquely) solvable that can be checked more easily before we start trying to calculate a solution. Moreover, this procedure of eliminating variables becomes less practical for larger systems, and so we would like to develop a more systematic approach that can deal with larger systems efficiently.

The key objects that will be used for developing such an approach are *matrices.*

2.1 Matrices

One may think of matrices as a multi-dimensional analogue of vectors, where we arrange numbers into an array.

Definition 2.1. *Let $m, n \in \mathbb{N}$ and $a_{ij} \in \mathbb{R}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. A (real) $m \times n$* ***matrix*** *is an array given by*

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

In this case we use the notation $A \in \mathbb{R}^{m \times n}$, and call m and n the ***dimensions*** *of A. An $m \times 1$ matrix is called a* ***column vector****. A $1 \times n$ matrix is called a* ***row vector****. If $m = n$, then A is a* ***square matrix****.*

In order to save space, we sometimes use the notation $(a_{ij})_{i,j=1}^{m,n}$ as a shorthand for the matrix A described above. This notation tells us about the dimensions of A, and also that the entry in the ith row and jth column is a_{ij}. Another notation that is sometimes convenient is that $(A)_{ij}$ is used to denote the entry of A in the ith row and jth column. So, for the matrix A defined above, we have $(A)_{ij} = a_{ij}$.

The case of complex matrices, i.e., $a_{ij} \in \mathbb{C}$, can be treated analogously, and we write $A \in \mathbb{C}^{m\times n}$. We can even consider matrices whose entries come from an arbitrary field $\mathbb{F}$. However, in order to give a solid grounding for understanding this new material, we will consider only matrices of real numbers in this course.

We now turn to basic operations of matrices. The first two, namely *scalar multiplication* and *matrix addition*, are simple (and familiar from the corresponding operations for vectors). These operations are carried out *component-wise*, meaning that they are performed in each entry of the matrices individually.

Let $A, B \in \mathbb{R}^{m\times n}$. Write

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

and

$$B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix}.$$

Then $A + B$ is the $m \times n$ matrix

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{pmatrix}.$$

Note that it is important here that the matrices have the **same dimension**. If the dimensions of the two matrices do not agree, then their sum is not defined.

The second operation is **scalar multiplication**, which is the operation which multiplies every entry of the matrix by a fixed scalar. Let $\lambda \in \mathbb{R}$ and

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

Then $\lambda A \in \mathbb{R}^{m\times n}$ is the matrix

$$A = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \dots & \lambda a_{mn} \end{pmatrix}.$$

The term “scalar” is used to distinguish this from the **matrix product**. This is the third essential operation on matrices that we consider. The definition of the product of matrices is a little more complicated, as we do not define this product component-wise.

Let $A \in \mathbb{R}^{m\times p}$ be an $m \times p$ matrix and let $B \in \mathbb{R}^{p\times n}$ be a $p \times n$ matrix. Write

$$A = (a_{ij})_{i,j=1}^{m,p}, \quad B = (b_{ij})_{i,j=1}^{p,n}.$$

The **product** of A and B is the matrix $C = AB \in \mathbb{R}^{m\times n}$ such that $C = (c_{ij})_{i,j=1}^{m,n}$ and

$$c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}.$$

In other words, the entry of AB in the ith row and jth column is $\sum_{k=1}^{p} a_{ik}b_{kj}$.

For this definition to make sense, it is crucial that the dimensions of A and B are correct. In particular, **the number of columns of A must be equal to the number of rows in B** (we may sometimes say that the *inner dimensions* of A and B match).

A helpful way to think about computing the matrix product may be the following: to calculate the ij entry of the product AB, move along the ith row of A and down the jth column of B.

Example - Let $A \in \mathbb{R}^{3\times 2}$ and $B \in \mathbb{R}^{2\times 2}$ be the matrices

$$A = \begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 7 & 9 \\ 8 & 0 \end{pmatrix}.$$

There are 2 columns in A and 2 rows in B. These numbers are the same, and so the matrix AB is defined. In particular AB is a 3×2 matrix. Write

$$AB = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix}.$$

We will fill in the blanks one entry at a time to write C out explicitly. To calculate c_{11}, we consider the first row of A and first column of B.

$$A = \begin{pmatrix} \mathbf{1} & \mathbf{6} \\ 2 & 5 \\ 3 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} \mathbf{7} & 9 \\ \mathbf{8} & 0 \end{pmatrix}.$$

Then

$$c_{11} = 1 \cdot 7 + 6 \cdot 8 = 55.$$

To calculate c_{12}, we consider the first row of A and second column of B.

$$A = \begin{pmatrix} \mathbf{1} & \mathbf{6} \\ 2 & 5 \\ 3 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 7 & \mathbf{9} \\ 8 & \mathbf{0} \end{pmatrix}.$$

Therefore,

$$c_{12} = 1 \cdot 9 + 6 \cdot 0 = 9.$$

We can repeat this process for each entry. We eventually obtain

$$AB = \begin{pmatrix} 55 & 9 \\ 54 & 18 \\ 53 & 27 \end{pmatrix}.$$

On the other hand, the reverse product BA is not defined, because the inner dimensions do not match.

One may observe from the example above that the process of computing an entry in the product of two matrices is similar to that of computing an inner product of two vectors. We formalise this observation below.

Let $A \in \mathbb{R}^{m \times p}$ and $B \in \mathbb{R}^{p \times n}$. The matrix A has m rows, each of which can be viewed as row vectors. We write

$$A = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_m \end{pmatrix},$$

where $\mathbf{a}_i$ is the row vector $\mathbf{a}_i = (a_{i1}, a_{i2}, \dots, a_{ip})$. Similarly, we can write

$$B = \begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_n . \end{pmatrix}$$

Here, $\mathbf{b}_j$ is the column vector

$$\mathbf{b}_j = \begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{pj} \end{pmatrix}.$$

The entry of AB in the ith row and jth column is then equal to

$$\sum_{k=1}^{p} a_{ik} b_{kj} = \langle \mathbf{a}_i, \mathbf{b}_j \rangle.$$

We can also define the **matrix-vector product** of a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$. These kind of products will be particularly useful when it comes down to solving systems of linear equations later! Write

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_m \end{pmatrix}$$

and

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Then $A\mathbf{x} \in \mathbb{R}^m$ is a vector, defined as

$$A\mathbf{x} := \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix} = \begin{pmatrix} \langle \mathbf{a}_1, \mathbf{x} \rangle \\ \langle \mathbf{a}_2, \mathbf{x} \rangle \\ \vdots \\ \langle \mathbf{a}_m, \mathbf{x} \rangle \end{pmatrix}. \tag{53}$$

In the previous chapter, we considered vectors $\mathbf{x} \in \mathbb{R}^n$ for some $n \in \mathbb{N}$. Such a vector may be considered as a matrix in $\mathbb{R}^{n \times 1}$. If we consider a vector to be a matrix in this way, we can consider the product of a matrix $A \in \mathbb{R}^{m \times n}$ with the matrix $\mathbf{x} \in \mathbb{R}^{n \times 1}$ using the definition of matrix product given above. Then the matrix product $A\mathbf{x}$ is exactly the same as the definition of $A\mathbf{x}$ given above in (53). In other words, the matrix vector product is just a special case of the matrix product we have already defined.

Example - Let

$$A = \begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix} \text{ and } \mathbf{x} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}.$$

Then

$$A\mathbf{x} = \begin{pmatrix} 1 \cdot 3 + 6 \cdot 4 \\ 2 \cdot 3 + 5 \cdot 4 \\ 3 \cdot 3 + 4 \cdot 4 \end{pmatrix} = \begin{pmatrix} 27 \\ 26 \\ 25 \end{pmatrix}.$$

As with matrix multiplication, we need to take care to ensure that the matrix-vector product we are considering has the correct dimensions to be defined. For instance, if we instead set

$$A = \begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix} \text{ and } \mathbf{x} = \begin{pmatrix} 3 \\ 4 \\ 0 \end{pmatrix},$$

then $A\mathbf{x}$ is not defined.

We have the following calculation rules, which are reminiscent of rules we established for vectors in the previous section.

Exercise

- Prove that, for all $A, B \in \mathbb{R}^{m \times n}$ and any $\lambda \in \mathbb{R}$,

$$\lambda(A + B) = \lambda A + \lambda B.$$

- Prove that, for all $A \in \mathbb{R}^{m \times p}$, $B \in \mathbb{R}^{p \times n}$ and any $\lambda \in \mathbb{R}$,

$$A(\lambda B) = \lambda AB.$$

- Prove that, for all $A \in \mathbb{R}^{m \times p}$ and $B, C \in \mathbb{R}^{p \times n}$,

$$A(B + C) = AB + AC.$$

Since matrix-vector multiplication is just a special case of matrix multiplication, the folllowing two facts follow immediately from the previous exercise.

- For all $A \in \mathbb{R}^{m\times n}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$A(\mathbf{x}+\mathbf{y}) = A\mathbf{x} + A\mathbf{y}$$

- For all $A \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$,

$$A(\lambda\mathbf{x}) = \lambda A\mathbf{x}.$$

However, **matrix multiplication is not commutative**. Even in the case when both matrices AB and BA are defined and have the same dimensions, it is usually the case that $AB \neq BA$. You can check this yourself by choosing two "random" matrices A and B, with the correct dimensions to ensure that both AB and BA are defined (what condition does this impose on the dimensions?), and computing both AB and BA.

There exist **identity elements** for the operations of matrix addition and multiplication.

Let 0_{mn} denote the matrix in $\mathbb{R}^{m\times n}$ with every entry being equal to zero. Then, for any $A \in \mathbb{R}^{m\times n}$ we have

$$A + 0_{mn} = 0_{mn} + A = A.$$

The multiplicative identity is of more use, and also practical interest. For $n \in \mathbb{N}$, we define the **identity matrix** $I_n \in \mathbb{R}^{n\times n}$ to be the matrix

$$I_n := \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

In words, I_n is the matrix with 1 at every diagonal entry and 0 everywhere else.

Note that the identity is a square matrix, and we may write $I := I_n$ if the dimension is clear.

Exercise - Let $A \in \mathbb{R}^{m\times n}$. Prove that

$$AI_n = A$$

and

$$I_m A = A.$$

Next, we discuss the *transpose* of a matrix. Since the dimensions of a matrix are important, it makes a huge difference if the dimensions of a matrix are $m \times n$ or $n \times m$, and it is quite useful to have a compact notation to switch the rows and columns of a matrix. That is, for a given $m \times n$ matrix $A = (a_{ij})_{i,j=1}^{m,n}$, we define its **transpose**, denoted A^T as the $n \times m$ matrix whose rows are the columns of A. To be more precise, if

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

then

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}.$$

Example - Consider again the matrix

$$A = \begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{pmatrix}.$$

Then

$$A^T = \begin{pmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \end{pmatrix}.$$

The transpose notation is also convenient for distinguishing column vectors and row vectors. Recall that the standard basis unit vectors $\mathbf{e}_k \in \mathbb{R}^n = \mathbb{R}^{n\times 1}$ are the (column) vectors that contain exactly one 1 (in the kth position) and all other entries are zero. The row unit vectors are defined via transposition as $\mathbf{e}_k^T \in \mathbb{R}^{1\times n}$. That is,

$$\mathbf{e}_1^T = (1,0,0,\dots,0),\ \ \mathbf{e}_2^T = (0,1,0,\dots,0),\ \ \dots\ \ \mathbf{e}_n^T = (0,0,0,\dots,0,1).$$

Using these unit vectors, we can write the identity matrix as

$$I_n = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) = \begin{pmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \\ \vdots \\ \mathbf{e}_n^T \end{pmatrix}.$$

With the above considerations and the fact that $I_n A = A I_n = A$, we see that the unit vectors can be used to “extract” the rows and columns from a matrix. For instance, given a matrix $A \in \mathbb{R}^{m\times n}$ of the form

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_m \end{pmatrix},$$

and the row vector $\mathbf{e}_k^T \in \mathbb{R}^{1\times m}$, we can compute that

$$\mathbf{e}_k^T A = \mathbf{a}_k.$$

Similarly, the product $A\mathbf{e}_k$ (with $\mathbf{e}_k \in \mathbb{R}^{n\times 1}$) can be used to extract the kth column from A.

There is one calculation rule related to the transpose, that is sometimes also very useful for computing the product of matrices. We state this in the following lemma.

Lemma 2.2. *Let $m, p, n \in \mathbb{N}$, $A \in \mathbb{R}^{m\times p}$ and $B \in \mathbb{R}^{p\times n}$. Then*

$$(AB)^T = B^T A^T.$$

Proof. Firstly, we note that $B^T A^T$ is well-defined, and is and $n \times m$ matrix, which means that the dimensions of $(AB)^T$ and $B^T A^T$ are the same. Indeed, $B^T \in \mathbb{R}^{n \times p}$ and $A^T \in \mathbb{R}^{p \times m}$, and so the inner dimensions of B^T and A^T agree. The outer dimensions confirm that $B^T A^T \in \mathbb{R}^{n \times m}$.

To prove that $(AB)^T = B^T A^T$, we need to show that each corresponding pair of entries of the two matrices are the same. Recall that, for an arbitrary matrix M, the notation $(M)_{ij}$ is used for the entry of M in the ith row and jth column. We need to show that

$$((AB)^T)_{ij} = (B^T A^T)_{ij} \tag{54}$$

holds for all $1 \leq i \leq n$ and $1 \leq j \leq m$.

Since $(M^T)_{ij} = (M)_{ji}$, it follows that

$$((AB)^T)_{ij} = (AB)_{ji} = \sum_{k=1}^{p} (A)_{jk}(B)_{ki}.$$

On the other hand

$$(B^T A^T)_{ij} = \sum_{k=1}^{p} (B^T)_{ik}(A^T)_{kj} = \sum_{k=1}^{p} (B)_{ki}(A)_{jk}.$$

Comparing the previous two equations, we see that we have proved (54).

□

Some matrices do not change under transposition. A matrix $A \in \mathbb{R}^{n \times n}$ such that $A^T = A$ is called **symmetric**.

Note that symmetric matrices must be square, and we will see later that symmetric matrices have several important properties.

An obvious but important example of a symmetric matrix is the identity matrix. More generally, *diagonal matrices* are always symmetric.

Definition 2.3. *A square matrix $A = (a_{ij})_{i,j}^{n,n}$ is a **diagonal matrix** if there exist $d_1, \ldots, d_n \in \mathbb{R}$ such that*

$$a_{ij} := \begin{cases} d_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

*The numbers d_i are called **diagonal elements** of A and we write $A = diag(d_1, \ldots, d_n)$.*

The last concept related to matrix multiplication that we will need is the *inverse* of a matrix.

Definition 2.4. *Let $A \in \mathbb{R}^{n \times n}$. The **inverse** of A, if it exists, is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that*

$$AA^{-1} = A^{-1}A = I_n.$$

*If an inverse exists, then we call a matrix **invertible**.*

Note that we only considered square matrices in the above definition. Why?

Example - The matrix I_n is invertible and $I_n^{-1} = I_n$.

Example - Let

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

We can verify by direct calculation that A is invertible and

$$A^{-1} = \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}$$

We just need to check that

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This calculation is left to the student to check.

In general, it is not easy to see whether a matrix is invertible or not. For example, we showed above that the matrix

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

is invertible, but the slightly modified matrix

$$\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}$$

is not invertible. We will discuss a way to verify if a matrix is invertible, and how to compute an inverse later in this chapter. However, let us already add here that even if we know that a matrix is invertible, it is usually difficult to compute its inverse. We will come back to this issue later, and present some ways for computing the inverse, at least for small matrices. This inverse will be the ultimate tool to solve certain systems of linear equations. But we will first discuss some more direct, but less powerful ways to calculate solutions.

Recall the field axioms that we stated in Section 1.4. Many of these axioms are also satisfied for matrices. Let us restrict to the case of matrices in $\mathbb{R}^{n\times n}$ for some $n \in \mathbb{N}$. We have a notion of addition and multiplication which satisfy the properties of associativity and distributivity. There exist additive and multiplicative inverses. Addition is commutative, and every matrix has an additive inverse.

On the other hand, multiplication of matrices is *not* commutative, and so $\mathbb{R}^{n\times n}$ is *not* a field. Furthermore, not all matrices in $\mathbb{R}^{n\times n}$ have a multiplicative inverse.

2.2 Systems of linear equations

Throughout this section, there will be the parameters $m, n \in \mathbb{N}$, where n **is the number of unknown variables** and m **is the number of equations** that must be fulfilled. The system of equations we want to solve will be of the following form.

Definition 2.5. *Let $m, n \in \mathbb{N}$ and for all $1 \leq i \leq m$ and $1 \leq j \leq n$, let $a_{ij} \in \mathbb{R}$ and $b_i \in \mathbb{R}$. A* ***system of linear equations*** *is given by*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m. \end{aligned}$$

The x_i with $1 \leq i \leq n$ are called ***variables****, or* ***unknowns****.*

The a_{ij} are called the ***coefficients*** *of the system.*

The matrix $A = (a_{ij})_{i,j=1}^{m,n}$ is called the ***matrix of coefficients*** *of the system.*

The vector

$$\boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

is called the ***right hand side (RHS)*** *of the system.*

If there exist numbers $x_1, \ldots, x_n \in \mathbb{R}$ that fulfill all the equations, then we call the tuple $(x_1, ..., x_n)$ a ***solution*** *to the linear system.*

If there is no solution, then we call the linear system ***inconsistent****.*

Let

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

and

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}.$$

Recall from the previous section that the matrix-vector product $A\mathbf{x}$ is defined as

$$A\mathbf{x} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{pmatrix}.$$

Therefore, the linear system from the previous definition can also be written in shorter form as

$$A\mathbf{x} = \mathbf{b}.$$

Obviously, we are interested in solutions to a linear system. However, as already discussed above, *such systems may have no solutions, a unique solution, or even infinitely many solutions.* We will see that a more detailed analysis of the matrix of coefficients can help us to determine which of these three cases we face. Before we come to this, let us introduce some more notation and discuss some examples.

Definition 2.6. *Given a linear system $A\boldsymbol{x} = \boldsymbol{b}$ with coefficient matrix $A \in \mathbb{R}^{m\times n}$ and RHS $\boldsymbol{b} \in \mathbb{R}^m$, then we denote the set of solutions by*

$$L(A, \boldsymbol{b}) = \{\boldsymbol{x} \in \mathbb{R}^n : A\boldsymbol{x} = \boldsymbol{b}\}.$$

Examples - Let us revisit some examples we considered at the beginning of this chapter. The system

$$\begin{aligned} 2x_1 + x_2 &= 1 \\ 6x_1 + 2x_2 &= 2 \end{aligned}$$

has the unique solution $(x_1, x_2) = (0, 1)$. This is the same thing as saying that the system

$$\begin{pmatrix} 2 & 1 \\ 6 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

has the unique solution

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Therefore, we can write

$$L\left(\begin{pmatrix} 2 & 1 \\ 6 & 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}\right) = \left\{\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right\}.$$

Note that $L(A, \mathbf{b})$ is a *set* and therefore, we need to write $L(A, b) = \{\mathbf{x}\}$ (rather than $L(A, \mathbf{b}) = \mathbf{x}$ if $\mathbf{x}$ is the only solution.

Example - We also considered the system of equations

$$\begin{aligned} 2x_1 + x_2 &= 1 \\ 6x_1 + 3x_2 &= 3. \end{aligned}$$

Since the two equations are identical, the solutions to this system are simply the solutions to the equation $2x_1 + x_2 = 1$. This can be rewritten as $x_2 = 1 - 2x_1$.

We can treat $x_1 \in \mathbb{R}$ as a free variable, and conclude that any point of the form $(\lambda, 1 - 2\lambda)$ such that $\lambda \in \mathbb{R}$ is solution to our system of equations. In summary, we have shown that

$$L\left(\begin{pmatrix} 2 & 1 \\ 6 & 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \end{pmatrix}\right) = \left\{\begin{pmatrix} \lambda \\ 1 - 2\lambda \end{pmatrix} : \lambda \in \mathbb{R}\right\}.$$

Example - Consider the following system of linear equations.

$$\begin{aligned} 2x_1 + x_2 + 3x_3 &= 1 \\ 6x_1 + 3x_2 &= 3 \\ 4x_1 &= 8. \end{aligned} \tag{55}$$

The set of solutions to this system is the same as the set of all $(x_1, x_2, x_3) \in \mathbb{R}^3$ such that

$$\begin{pmatrix} 2 & 1 & 3 \\ 6 & 3 & 0 \\ 4 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 8 \end{pmatrix}$$

The form that this system takes makes it quite simple to solve. The last equation immediately gives $x_1 = 2$. Plugging this into the second equation, we have $12 + 3x_2 = 3$, and so $x_2 = -3$. Plugging these values of x_1 and x_2 into the first equation gives $4 - 3 + 3x_3 = 1$, and so $x_3 = 0$. We conclude that

$$L\left(\begin{pmatrix} 2 & 1 & 3 \\ 6 & 3 & 0 \\ 4 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 8 \end{pmatrix} \right) = \left\{ \begin{pmatrix} 2 \\ -3 \\ 0 \end{pmatrix} \right\}.$$

It would be nice if all systems of linear equations could be solved as easily as (55). While this is not quite the case, it is true that every system of linear equations can be *reduced* to make it a little easier to consider. This is essentially what we will be doing in the next section, when we learn about Gaussian elimination and row reduction.

We now discuss one special case of equations, for which the right-hand side of the system is made up only of zeroes.

Definition 2.7. *Let $A \in \mathbb{R}^{m\times n}$. A linear system of the form*

$$A\boldsymbol{x} = \boldsymbol{0}$$

*is called a **homogeneous** system.*

*For a linear system $A\boldsymbol{x} = \boldsymbol{b}$, we say that the homogeneous system $A\boldsymbol{x} = \boldsymbol{0}$ is the **corresponding homogeneous system.***

For a given matrix $A \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$, it turns out that we can learn a lot about the set of solutions $L(A, \mathbf{b})$ by considering the set of solutions $L(A, \mathbf{0})$ to the corresponding homogenous system. This is the content of the next two lemmas.

Lemma 2.8. *Let $A \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Suppose that there exist $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$ with $\boldsymbol{z} \neq \boldsymbol{0}$ such that $A\boldsymbol{y} = \boldsymbol{b}$, and $A\boldsymbol{z} = \boldsymbol{0}$. Then there exist infinitely many solutions $\boldsymbol{x} \in \mathbb{R}^n$ to the system $A\boldsymbol{x} = \boldsymbol{b}$.*

Proof. Let $\lambda \in \mathbb{R}$ be arbitrary. Then

$$A(\mathbf{y} + \lambda\mathbf{z}) = A\mathbf{y} + \lambda A\mathbf{z} = \mathbf{b} + \lambda\mathbf{0} = \mathbf{b}.$$

Since $\mathbf{z} \neq 0$, it follows that all of the vectors $\mathbf{y} + \lambda\mathbf{z}$ are distinct. As there are infinitely many choices for $\lambda \in \mathbb{R}$, it follows that $A\mathbf{x} = \mathbf{b}$ has infinitely many solutions.

□

Lemma 2.9. *Let $A \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Suppose that the homogeneous system $A\boldsymbol{x} = \boldsymbol{0}$ has only the trivial solution $\boldsymbol{x} = \boldsymbol{0}$. Then there exists at most one solution to the system $A\boldsymbol{x} = \boldsymbol{b}$.*

Proof. This is left as an exercise. □

2.3 Gaussian elimination

Now we make an important observation which allows us to derive an algorithm for solving linear systems by manipulating matrices. We can perform the following operations to a linear system without changing the set of solutions:

- Interchanging any two equations, i.e., changing the order of the equations.
- Multiplying an equation with a scalar $0 \neq \lambda \in \mathbb{R}$.
- Adding a multiple of an equation to another equation.

Take a moment to consider why these three changes do not alter the solution set. The third of these points is a little more difficult than the others, and will be considered in more detail in a forthcoming exercise sheet.

Since every system of linear equations can be written with the help of a matrix, it is natural to consider how the above operations change the corresponding matrix of coefficients of a linear system. We will see that they indeed allow for successive modifications that lead to much simpler matrices, i.e., matrices in *echelon form* and *reduced echelon form*. From such a matrix, we will be able to basically *see* if a corresponding linear system is (uniquely) solvable or not.

Let us start by discussing how the above operations to a linear system $A\mathbf{x} = \mathbf{b}$ affect the corresponding matrix A. However, note already now that these operations also change the RHS $\mathbf{b}$ of a linear system, and this is essential. We will come back to this shortly, but for now we only consider the corresponding matrix of coefficients.

In view of the operations from above that can be used to change a linear system $A\mathbf{x} = \mathbf{b}$ without changing the set of solutions, we see that the matrix A is changed in the following way:

- Interchanging two rows.
- Multiplying a row with a scalar $0 \neq \lambda \in \mathbb{R}$.
- Adding a multiple of a row to another row.

For obvious reasons, these operations are called *row operations*, or sometimes *elementary row operations*. Two matrices A and B are said to be *equivalent* if A can be obtained from B by performing row operations. Note that this definition is symmetric, since one can perform "inverse" row operations to then get from A back to B.

The goal now is to use these operations to simplify the given matrix into *echelon form*. In particular, we look to create as many zeroes in the matrix as possible, and for these zeroes to appear in a certain structured manner. We are ready to give some proper definitions.

Definition 2.10. *Let $C \in \mathbb{R}^{m\times n}$ and let $1 \leq i \leq m$. The* ***leading coefficient*** *of the ith row of C is the first non-zero entry in the row.*

Definition 2.11. *Let $C \in \mathbb{R}^{m\times n}$ be a matrix which has at least one non-zero entry and let c_{ij} denote the entry of C in the ith row and jth column. Let c_{ij_i} denote the leading coefficient of the ith row of C, if such a leading coefficient exists.*

*C is in **row echelon form** if both of the following conditions hold.*

- *There is some $1 \leq k \leq n$ such that the leading coefficient c_{ij_i} exists if and only if $1 \leq i \leq k$. In other words, all zero rows appear at the bottom of the matrix.*
- *For all $1 \leq i < i' \leq k$, $j_i < j_{i'}$. In other words, the leading coefficients move strictly to the right as we move down the rows of C.*

The matrix with every entry being 0 is also in row echelon form.

Note that this definition also implies that the entries below a leading coefficient in the same column are all equal to 0.

Definition 2.12. *Let $C \in \mathbb{R}^{m\times n}$ and let c_{ij} denote the entry of C in the ith row and jth column. Let c_{ij_i} denote the leading coefficient of the ith row of C, if such a leading coefficient exists.*

*C is in **reduced row echelon form** if it is in row echelon form and it also satisfies the following two conditions.*

- *For all $1 \leq i \leq k$, $c_{ij_i} = 1$. In other words, all leading coefficients are equal to 1.*
- *For all $1 \leq i \leq k$, and $1 \leq i' < i$, we have $c_{i'j_i} = 0$. In other words, the entries above a leading coefficient in the same column are all equal to 0.*

The matrix with every entry being 0 is also in reduced row echelon form.

Examples - The following two matrices are in row echelon form:

$$A = \begin{pmatrix} 1 & 0 & 2 & 5 \\ 0 & 0 & -3 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} -2 & 0 & 2 & 3 & -1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 8 & \pi \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The following two matrices are in reduced row echelon form:

$$C = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Consider the matrix

$$E = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This matrix is not in row echelon form. It does not satisfy the required condition that all of the zero rows are at the bottom of the matrix. However, if we reverse the order of the rows, swapping the second row with the fourth one, we obtain the matrix C, which is in reduced row echelon form.

Another matrix which is not in row echelon form is

$$F = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

This is because the leading coefficient of the the third row is not to the right of the leading coefficient of the row above.

Example - Let's see an example of how reduced row echelon form matrices correspond to linear systems that can be easily solved. We use the reduced row echelon form matrix

$$C = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

from an earlier example. We seek to find all solutions $\mathbf{x} \in \mathbb{R}^3$ to the equation

$$C\mathbf{x} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix}.$$

Recalling the notation introduced in the previous section, we want to understand the set

$$L\left(C, \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix}\right)$$

Writing this as a system of linear equations, this becomes

$$\begin{aligned} x_1 + 2x_2 &= 2 \\ x_3 &= 1 \\ 0 &= 0 \\ 0 &= 1. \end{aligned}$$

Since the last equation is never valid, there are no solutions to this system, and therefore

$$L\left(C, \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix}\right) = \emptyset.$$

Let's see what happens when we slightly modify the RHS of this system and consider solutions $\mathbf{x} \in \mathbb{R}^3$ to the equation

$$C\mathbf{x} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

Writing this as a system of linear equations, this becomes

$$\begin{aligned} x_1 + 2x_2 &= 2 \\ x_3 &= 1 \\ 0 &= 0 \\ 0 &= 0. \end{aligned}$$

The last two equations are always satisfied, and can thus be disregarded. So, we need to solve the system. Writing this as a system of linear equations, this becomes

$$\begin{aligned} x_1 + 2x_2 &= 2 \\ x_3 &= 1. \end{aligned}$$

We can treat x_2 as a *free variables.* This means that we allow x_2 to range over all possible values of $\mathbb{R}$, and write the other variables in terms of the free variables (if necessary). We conclude that

$$L\left(C, \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \end{pmatrix}\right) = \left\{ \begin{pmatrix} 2 - 2x_2 \\ x_2 \\ 1 \end{pmatrix} : x_2 \in \mathbb{R} \right\}. \tag{56}$$

Note that there is some choice in how to choose the free variables and therefore in how we express the final form of the solution set. In this case, we could have instead chosen x_1 as the free variable and expressed x_2 in terms of x_1. We could then conclude that

$$L\left(C, \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \end{pmatrix}\right) = \left\{ \begin{pmatrix} x_1 \\ 1 - \frac{1}{2}x_1 \\ 1 \end{pmatrix} : x_1 \in \mathbb{R} \right\}.$$

These two expressions may appear different, but they are just two different descriptions of the same set.

When we are looking to express the solution set for a system given in reduced row echelon form, a convenient method is the following: we can set the free variables to correspond to the columns which do not contain a leading coefficient. This is what we did when writing down the solution set in the form of (56). As we have seen above, there are other ways to choose the free variables, but using the columns without leading coefficients is guaranteed to produce a fairly tidy looking expression for the solution set.

We do not prove the following statement here formally, but note that it is the basis of the considerations below.

Theorem 2.13. *Every matrix can be transformed to reduced row echelon form by performing row operations. Moreover, the reduced row echelon form of a matrix is unique.*

In contrast, a given matrix A can be transformed by row operations into different matrices in (non-reduced) row echelon form. For example, multiplying any row of a row echelon form matrix by 2 leads to another row echelon form matrix. But even then, all row echelon forms of a matrix have the same number of non-zero rows.

Exercise - Prove that any two equivalent matrices in row echelon form have the same number of non-zero rows.

This allows us to state the following definition.

Definition 2.14. *Let $A \in \mathbb{R}^{m\times n}$ be arbitrary. We define the **rank** of A, denoted by $rank(A)$, as the number of non-zero rows in a row echelon form matrix C that is equivalent to A.*

Exercise - Let $A \in \mathbb{R}^{m\times n}$ be an arbitrary matrix. Show that

$$rank(A) \leq \min\{m, n\}.$$

Examples - Let us revisit the six matrices A, B, C, D, E and F from a previous example. Since A, B, C and D are in row echelon form, we can immediately see their ranks by counting the number of non-zero rows. Note that

$$rank(A) = 3, \quad rank(B) = 3, \quad rank(C) = 2, \quad rank(D) = 3.$$

Although E is not in row echelon form, we can easily transform it into row echelon form with row operations, namely by switching the second and fourth row. We obtain the equivalent matrix

$$E' = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Therefore, $rank(E) = 2$. We can transform F into row echelon form by subtracting the second row from the third. We obtain the equivalent row echelon form matrix

$$F' = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Therefore, $rank(F) = 3$.

Let us see some more involved examples of how we can use row operations to transform a matrix into row echelon form and reduced row echelon form.

Example - Let us consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

Our first task is to reduce the matrix to row echelon form. To do this, we need to create zeroes underneath the leading entries. We can do this by subtracting four times the first row from the second. We indicate this procedure as follows.

$$\begin{pmatrix}1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\end{pmatrix} \xrightarrow[R_2 = R_2 - 4R_1]{} \begin{pmatrix}1 & 2 & 3\\ 0 & -3 & -6\\ 7 & 8 & 9\end{pmatrix}$$

Note that the "$R_2 = R_2 - 4R_1$" appearing above is not an actual mathematical equation, but rather a piece of notation for telling us how the row operation has been carried out. There are many slight variants of this notation that appear throughout the literature, so please be aware of this when reading other sources.

Similarly, we obtain a zero in the third entry of the first column by subtracting 7 times the first row.

$$\begin{pmatrix}1 & 2 & 3\\ 0 & -3 & -6\\ 7 & 8 & 9\end{pmatrix} \xrightarrow[R_3 = R_3 - 7R_1]{} \begin{pmatrix}1 & 2 & 3\\ 0 & -3 & -6\\ 0 & -6 & -12\end{pmatrix}$$

We are finished with the first leading entry, but we also need to have zeroes beneath the second leading entry. The row operation which achieves this is $R_3 = R_3 - 2R_2$. With this, we obtain the matrix

$$B = \begin{pmatrix}1 & 2 & 3\\ 0 & -3 & -6\\ 0 & 0 & 0\end{pmatrix}.$$

This matrix is in row echelon form. To transform this into reduced row echelon form, we dilate the second row to ensure that all leading coefficients are equal to 1 (we perform the operation $R_2 = -\frac{1}{3}R_2$).

After that, we still need to create zeroes above all of the leading coefficients. This means that the entry in the first row and second column needs to be zero. The operation $R_1 = R_1 - 2R_2$ achieves this. We summarise these two steps via the following notation.

$$\begin{aligned}\begin{pmatrix}1 & 2 & 3\\ 0 & -3 & -6\\ 0 & 0 & 0\end{pmatrix} &\xrightarrow[R_2 = -\frac{1}{3}R_2]{} \begin{pmatrix}1 & 2 & 3\\ 0 & 1 & 2\\ 0 & 0 & 0\end{pmatrix}\\ &\xrightarrow[R_1 = R_1 - 2R_2]{} \begin{pmatrix}1 & 0 & -1\\ 0 & 1 & 2\\ 0 & 0 & 0\end{pmatrix}\end{aligned}$$

We have finally obtained a matrix in reduced row echelon form. We can therefore say that the reduced row echelon form of A is the matrix

$$C = \begin{pmatrix}1 & 0 & -1\\ 0 & 1 & 2\\ 0 & 0 & 0\end{pmatrix}.$$

The process that we have used above to transform a given matrix into (reduced) row echelon form is called **Gaussian elimination**.

There are many ways to reduce a matrix to row echelon form and reduced row echelon form via row operations, and the choices mainly come down to choosing an order to perform the operations. We can often use intuition to spot certain shortcuts that will simplify the process. We can also state a formal algorithm for doing this, which is essentially implicit in the examples outlined above.

Step 1 - Begin with the leftmost nonzero column. If necessary, use row interchange so that this column's first entry is nonzero.

Step 2 (optional) - Dilate so that the first entry in this column is equal to 1 (this is not essential for reducing to echelon form, but it usually makes the calculations easier)

Step 3 - Use row replacement operations to create zeros in all positions except for the first entry of the column.

Step 4 - Ignore the first row of the matrix, and apply steps 1,2 and 3 to the submatrix that remains. Repeat the process until the matrix is in echelon form (this process will certainly terminate, since any matrix with exactly one row is in echelon form).

At this point we have a matrix in echelon form, but we can extend the algorithm to produce a matrix in reduced echelon form.

Step 5 - Beginning with the rightmost leading entry, use row replacement operations to create zeros above the leading entry. Do this for all of the leading terms, progressing to the left and up.

Step 6 - Use row dilation so that all of the leading entries are changed to 1 (this will not be necessary if we performed step 2 every time).

The order of steps 5 and 6 can be reversed. It is often more practical to perform step 6 before step 5.

Now we discuss how we can solve linear systems by calculating (reduced) row echelon forms of matrices. We consider the linear system $A\mathbf{x} = \mathbf{b}$ with corresponding matrix $A = (a_{ij}) \in \mathbb{R}^{m\times n}$ and RHS

$$b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

Define the **augmented matrix** $(A|\mathbf{b})$ to be the matrix A with an additional row $\mathbf{b}$ added. That is,

$$(A|\mathbf{b}) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} & b_m \end{array}\right).$$

As we outlined at the beginning of this section, applying row operations to a system of linear equations does not change the solution set. Therefore, if the augmented matrix $(C|\mathbf{b}')$ is obtained from $(A|\mathbf{b})$ using only row operations, then

$$L(A, \mathbf{b}) = L(C, \mathbf{b}').$$

Let's see how to perform Gaussian elimination with augmented matrices to solve linear systems in practice.

Example - Consider the linear system $A\mathbf{x} = \mathbf{b}$ where $\mathbf{x} \in \mathbb{R}^2$ is a (vector) variable,

$$A = \begin{pmatrix} 3 & 5 \\ 1 & -1 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 42 \\ 6 \end{pmatrix}.$$

We write the augmented matrix

$$(A|\mathbf{b}) = \left(\begin{array}{cc|c} 3 & 5 & 42 \\ 1 & -1 & 6 \end{array}\right).$$

Next, we reduce the augmented matrix to reduced row echelon form using row operations. This means that we perform row operations to transform the left side of the augmented matrix to reduced row echelon, and in the process record the changes to $\mathbf{b}$ on the right side of the augmented matrix. We obtain

$$\begin{aligned} \left(\begin{array}{cc|c} 3 & 5 & 42 \\ 1 & -1 & 6 \end{array}\right) &\xrightarrow[R_1 \leftrightarrow R_2]{} \left(\begin{array}{cc|c} 1 & -1 & 6 \\ 3 & 5 & 42 \end{array}\right) \\ &\xrightarrow[R_2 = R_2 - 3R_1]{} \left(\begin{array}{cc|c} 1 & -1 & 6 \\ 0 & 8 & 24 \end{array}\right) \\ &\xrightarrow[R_2 = \frac{1}{8}R_2]{} \left(\begin{array}{cc|c} 1 & -1 & 6 \\ 0 & 1 & 3 \end{array}\right) \\ &\xrightarrow[R_1 = R_1 + R_2]{} \left(\begin{array}{cc|c} 1 & 0 & 9 \\ 0 & 1 & 3 \end{array}\right). \end{aligned}$$

It therefore follows that the set of solutions to $A\mathbf{x} = \mathbf{b}$ is equal to the set of solutions to the system

$$\begin{aligned} x_1 + 0x_2 &= 9 \\ 0x_1 + x_2 &= 3. \end{aligned}$$

We have therefore proved that

$$L\left(\begin{pmatrix} 3 & 5 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} 42 \\ 6 \end{pmatrix}\right) = \left\{\begin{pmatrix} 9 \\ 3 \end{pmatrix}\right\}.$$

Example - Consider the linear system $A\mathbf{x} = \mathbf{b}$ where

$$A = \begin{pmatrix} 2 & 1 & -2 \\ 0 & 3 & 6 \\ 2 & 0 & -4 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 5 \\ 3 \\ 4 \end{pmatrix}.$$

We write the augmented matrix

$$(A|\mathbf{b}) = \left(\begin{array}{ccc|c} 2 & 1 & -2 & 5 \\ 0 & 3 & 6 & 3 \\ 2 & 0 & -4 & 4 \end{array}\right).$$

Next, we reduce the augmented matrix to reduced row echelon form using row operations.

$$\begin{aligned}
\left(\begin{array}{ccc|c} 2 & 1 & -2 & 5 \\ 0 & 3 & 6 & 3 \\ 2 & 0 & -4 & 4 \end{array}\right) &\xrightarrow[R_3 = R_3 - R_1]{} \left(\begin{array}{ccc|c} 2 & 1 & -2 & 5 \\ 0 & 3 & 6 & 3 \\ 0 & -1 & -2 & -1 \end{array}\right) \\
&\xrightarrow[R_3 = -R_3]{} \left(\begin{array}{ccc|c} 2 & 1 & -2 & 5 \\ 0 & 3 & 6 & 3 \\ 0 & 1 & 2 & 1 \end{array}\right) \\
&\xrightarrow[R_2 \leftrightarrow R_3]{} \left(\begin{array}{ccc|c} 2 & 1 & -2 & 5 \\ 0 & 1 & 2 & 1 \\ 0 & 3 & 6 & 3 \end{array}\right) \\
&\xrightarrow[R_3 = R_3 - 3R_2]{} \left(\begin{array}{ccc|c} 2 & 1 & -2 & 5 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right) \\
&\xrightarrow[R_1 = R_1 - R_2]{} \left(\begin{array}{ccc|c} 2 & 0 & -4 & 4 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right) \\
&\xrightarrow[R_1 = \frac{1}{2} R_1]{} \left(\begin{array}{ccc|c} 1 & 0 & -2 & 2 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right).
\end{aligned}$$

The left side is of the augmented matrix is now in reduced row echelon form, and so we are ready to finalise the solution. The system is equivalent to

$$\begin{aligned} x_1 - 2x_3 &= 2 \\ x_2 + 2x_3 &= 1. \end{aligned}$$

This system has a free variable, which we set as x_3. We rewrite this system as

$$\begin{aligned} x_1 &= 2 + 2x_3 \\ x_2 &= 1 - 2x_3. \end{aligned}$$

We therefore conclude that

$$L(A, \mathbf{b}) = \left\{ \begin{pmatrix} 2 + 2x_3 \\ 1 - 2x_3 \\ x_3 \end{pmatrix} : x_3 \in \mathbb{R} \right\}.$$

2.3.1 The relationship between $Rank(A)$ and the number of solutions to $A\mathbf{x} = \mathbf{b}$

The next result highlights the convenient simplicity of the reduced row echelon form of a square matrix with full rank.

Lemma 2.15. *Let $A \in \mathbb{R}^{n \times n}$ be a square matrix and let $C \in \mathbb{R}^{n \times n}$ be its reduced row echelon form. Then*

$$rank(A) = n \iff C = I_n.$$

Proof. ($\Leftarrow$) We prove the contrapositive form. Suppose that $rank(A) \neq n$. Then C contains at least one zero row, and thus $C \neq I_n$, as required.

($\Rightarrow$) Suppose that $rank(A) = n$. Then C is a square matrix in reduced echelon form with at least one entry in each row. It must then be the case that the leading coefficients are the diagonal entries of C. Since C is in reduced echelon form, these leading coefficients are all 1. Also, since C is in reduced echelon form, all of the other entries in the same column as the leading coefficients must be zero. It follows that $C = I_n$. □

This leads to the following nice characterisation of square matrices with full rank.

Theorem 2.16. *If $A \in \mathbb{R}^{n\times n}$ is a square matrix with $rank(A) = n$, then the linear system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution for any $\boldsymbol{b} \in \mathbb{R}^n$.*

Proof. Let C be the reduced row echelon form of the matrix A. By Lemma 2.15, $C = I_n$. The set of solutions to $A\mathbf{x} = \mathbf{b}$ is the same as the set of solutions to $C\mathbf{x} = \mathbf{b}'$, where $\mathbf{b}'$ is some fixed vector in $\mathbb{R}^n$. However, since $C = I_n$, the system $C\mathbf{x} = \mathbf{b}'$ has the unique solution $\mathbf{x} = \mathbf{b}'$. □

Let $A \in \mathbb{R}^{m\times n}$ and let C be the reduced echelon form matrix equivalent to A. Note that the linear system $A\mathbf{x} = \mathbf{b}$ is inconsistent, i.e., has no solutions, if and only if there is a zero row in C (from the augmented matrix $(C|\mathbf{b}')$) and the corresponding entry of $\mathbf{b}'$ is not equal to zero.

Since there are exactly m rows, and $k = rank(A)$ of them are non-zero, we see that a linear system $A\mathbf{x} = \mathbf{b}$ is solvable (independent from the RHS $\mathbf{b}$) if $rank(A) = m$.

We can see from the discussion above and the previous few examples, as well as Theorem 2.16, that the rank of a matrix A is very influential in determining whether a system $A\mathbf{x} = \mathbf{b}$ has no solutions, a unique solution, or infinitely many solutions. We summarise some more features of the relationship between rank and the number of solutions in the next theorem.

Theorem 2.17. *Let $A \in \mathbb{R}^{m\times n}$.*

1. *If $rank(A) < m$, then there is some $\boldsymbol{b} \in \mathbb{R}^n$ such that the linear system $A\boldsymbol{x} = \boldsymbol{b}$ has no solutions.*

2. *If $rank(A) < n$, then the homogeneous system $A\boldsymbol{x} = 0$ has infinitely many solutions.*

3. *If $rank(A) < n$, then the system $A\boldsymbol{x} = \boldsymbol{b}$ either has no solutions or has infinitely many solutions.*

Proof.

1. Let C be the reduced row echelon form matrix equivalent to A. Since $rank(A) < m$, the last row of C is a zero row. Let $\mathbf{b}' \in \mathbb{R}^m$ be any vector with a non-zero entry $b'_m \neq 0$ in the final position. Observe that that the system $C\mathbf{x} = \mathbf{b}'$ has no solutions.

   Since A and C are equivalent, we can perform row operations on the augmented matrix $(C|\mathbf{b}')$ to transform it into the equivalent matrix $(A|\mathbf{b})$ (here $\mathbf{b}$ is a vector in $\mathbb{R}^m$ which is obtained by performing the row operations which transform C into A). The solutions to the system $A\mathbf{x} = \mathbf{b}$ are exactly the same as those for the system $C\mathbf{x} = \mathbf{b}'$. Therefore, there are no solutions to $A\mathbf{x} = \mathbf{b}$.

2. Let C be the reduced echelon form matrix equivalent to A. If the rank of A is strictly less than the number of columns, then there must exist a column in C which does not contain a leading coefficient. The corresponding variable can be treated as a free variable.

3. Suppose that $A\mathbf{x} = \mathbf{b}$ has at least one solution. By part 2 of this theorem, $A\mathbf{x} = 0$ has a non-trivial solution (i.e. a solution $\mathbf{x} \neq 0$). Lemma 2.8 then implies that $A\mathbf{x} = \mathbf{b}$ has infinitely many solutions.

□

We can use this to strengthen Theorem 2.16, as follows.

Theorem 2.18. *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix. Then*

$$\mathit{rank}(A) = n \iff \textit{the linear system } A\boldsymbol{x} = \boldsymbol{b} \textit{ has a unique solution for any } \boldsymbol{b} \in \mathbb{R}^n.$$

Proof. The "⇒" direction is Theorem 2.16. The other direction follows from Theorem 2.17. □

2.3.2 Row operations are the same as elementary matrix multiplication

The final key idea of this section to introduce is that *row operations are equivalent to multiplication by certain "elementary" matrices.* This interpretation of the row operations will be useful as we seek to build a basic theory of matrices.

An *elementary matrix* is a matrix which can be obtained from the identity matrix by a single row operation. There are three kinds of elementary matrices

1. A matrix corresponding to row interchange, for example performing the operation $R_2 \leftrightarrow R_1$ gives

$$E_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

2. A matrix corresponding to row dilation, for example performing the operation $R_2 = 3R_2$ gives

$$E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

3. A matrix corresponding to row replacement, for example performing the operation $R_3 = R_3 + 4R_1$ gives

$$E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & 0 & 1 \end{pmatrix}.$$

Furthermore, every row operation that we perform can be restated as matrix multiplication by an elementary matrix. For example, consider the matrix

$$A = \begin{pmatrix} 1 & 0 & -2 \\ -3 & 1 & 4 \\ 2 & -3 & 4 \end{pmatrix}.$$

We would typically start the process of reducing this to echelon form by performing the operation $R_2 = R_2 + 3R_1$. We obtain

$$\begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & -2 \\ 2 & -3 & 4 \end{pmatrix}.$$

However, this is the same thing as left-multiplying by the matrix

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Indeed

$$\begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -2 \\ -3 & 1 & 4 \\ 2 & -3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & -2 \\ 2 & -3 & 4 \end{pmatrix}.$$

The same is true for all elementary row operations. This hints at the following result.

Lemma 2.19. *Let A and B be $m \times n$ matrices and suppose that A and B are equivalent. Then, there exists a sequence of elementary matrices $E_1, E_2, \dots E_k$ such that*

$$B = E_k E_{k-1} \dots E_1 A.$$

Proof. The proof follows immediately from the preceding discussion. □

Note that, since elementary matrices are derived from the identity matrix via a row operations, it follows that they are always square. A useful fact about elementary matrices is that they are always invertible, and their inverses are also elementary matrices.

Lemma 2.20. *If $E \in \mathbb{R}^{n \times n}$ is an elementary matrix then E is invertible and E^{-1} is also an elementary matrix.*

Proof. This is left as an exercise. □

2.4 Matrices as linear transformations

We can use our notion of matrix multiplication to view matrices as a kind of function, or transformation.

Let $A \in \mathbb{R}^{m\times n}$, we say that $T_A : \mathbb{R}^n \to \mathbb{R}^m$ is the *matrix transformation* of A. This function is defined by

$$T_A(\mathbf{x}) = A\mathbf{x}.$$

Example - Let

$$A = \begin{pmatrix} 1 & -3 \\ 3 & 5 \\ -1 & 7 \end{pmatrix}.$$

Then, for an arbitrary vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$$

we have

$$T_A(\mathbf{x}) = \begin{pmatrix} 1 & -3 \\ 3 & 5 \\ -1 & 7 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 - 3x_2 \\ 3x_1 + 5x_2 \\ -x_1 + 7x_2 \end{pmatrix}.$$

Note that $T_A(\mathbf{x})$ is only defined if $\mathbf{x} \in \mathbb{R}^2$.

An important class of functions is the set of linear transformations.

Definition 2.21. *A function $T : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation if*

1. *For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $T(\boldsymbol{x} + \boldsymbol{y}) = T(\boldsymbol{x}) + T(\boldsymbol{y})$,*
2. *For all $\boldsymbol{x} \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$, $T(\lambda\boldsymbol{x}) = \lambda T(\boldsymbol{x})$.*

Theorem 2.22. *Let A be an $m \times n$ matrix. Then $T_A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation.*

Proof. We proved in an earlier exercise that $A(\mathbf{x} + \mathbf{y}) = A\mathbf{x} + A\mathbf{y}$ and $A(\lambda\mathbf{x}) = \lambda A\mathbf{x}$. This immediately implies the theorem. □

The next result shows that there is a direct correspondence between matrices and linear transformations.

Theorem 2.23. *Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then there exists a unique matrix A such that $T = T_A$. In fact, A is the $m \times n$ matrix*

$$A = (T(\boldsymbol{e}_1)\ T(\boldsymbol{e}_2) \dots T(\boldsymbol{e}_n)).$$

Proof. This is left as an exercise. □

Theorem 2.23 can then be used to prove the following result, which will be useful later in this chapter.

Theorem 2.24. *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix. Then*

$$T_A \textit{ is invertible} \iff A \textit{ is invertible.}$$

Proof. Throughout this proof, we use Id to denote the identity function with domain $\mathbb{R}^n$.

($\Leftarrow$) Suppose that A is invertible. Then $AA^{-1} = I_n = A^{-1}A$. Therefore,

$$T_A \circ T_{A^{-1}} = T_{AA^{-1}} = T_{I_n} = Id,$$

and similarly

$$T_{A^{-1}} \circ T_A = T_{A^{-1}A} = T_{I_n} = Id.$$

($\Rightarrow$) Suppose that T_A is invertible, and so there is some function $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T \circ T_A = Id = T_A \circ T$.

Claim. *T is a linear transformation.*

First we will show that the claim implies the theorem, and then we will prove the claim

Claim $\Rightarrow$ Theorem 2.24 - Since T is a linear transformation, Theorem 2.23 implies that $T = T_B$ for some matrix $B \in \mathbb{R}^{n\times n}$. Therefore

$$Id = T_A \circ T = T_A \circ T_B = T_{AB}, \text{ and } Id = T \circ T_A = T_B \circ T_A = T_{BA}.$$

Therefore, for all $\mathbf{x} \in \mathbb{R}^n$, we have

$$AB(\mathbf{x}) = \mathbf{x} \text{ and } BA(\mathbf{x}) = \mathbf{x}.$$

It follows that $AB = I_n = BA$, and so A is invertible with $A^{-1} = B$.

It remains to prove the claim.

Proof of Claim. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ be arbitrary. Since T is the inverse of T_A, we have

$$T_A(T(\mathbf{x})) = \mathbf{x} = T(T_A(\mathbf{x})), \text{ and } T_A(T(\mathbf{y})) = \mathbf{y} = T(T_A(\mathbf{y})).$$

Therefore,

$$\begin{aligned} T(\mathbf{x}+\mathbf{y}) &= T(T_A(T(\mathbf{x})) + T_A(T(\mathbf{y}))) \\ &= T(A(T(\mathbf{x})) + A(T(\mathbf{y}))) \\ &= T(A(T(\mathbf{x}) + T(\mathbf{y}))) \\ &= T(T_A(T(\mathbf{x}) + T(\mathbf{y}))) \\ &= T(\mathbf{x}) + T(\mathbf{y}). \end{aligned}$$

Similarly,

$$\begin{aligned} T(\lambda\mathbf{x}) &= T(\lambda T_A(T(\mathbf{x}))) \\ &= T(\lambda A(T(\mathbf{x}))) \\ &= T(A(\lambda T(\mathbf{x}))) \\ &= T(T_A(\lambda T(\mathbf{x}))) \\ &= \lambda T(\mathbf{x}). \end{aligned}$$

This completes the proof of the claim, and therefore also completes the proof of the theorem. □

□

2.5 Determinants

We now introduce the determinant of a matrix. This is a particularly useful tool for determining if a matrix is invertible or not.

Given a matrix A, let A_{ij} denote the matrix with row i and column j of A removed. For example, let

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 6 & 7 & 8 \\ 1 & -1 & -1 \end{pmatrix}.$$

Then

$$A_{13} = \begin{pmatrix} 6 & 7 \\ 1 & -1 \end{pmatrix}, \quad A_{22} = \begin{pmatrix} 3 & 1 \\ 1 & -1 \end{pmatrix}.$$

We use this to define the determinant recursively.

Definition 2.25. *For a 1×1 matrix A whose only entry is a, we say that the* ***determinant*** *of A is a, and write* $\det(A) = a$.

Suppose that $n \geq 2$ and $A \in \mathbb{R}^{n \times n}$. The ***determinant*** *of A, denoted* $\det(A)$ *is*

$$\sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(A_{1j}).$$

We sometimes omit the brackets and simply write $\det A$. Note that the determinant is only defined for square matrices.

We remark here that there are many equivalent definitions of the determinant that may be found elsewhere in the literature.

Example - For 2×2 matrices, the definition above gives a simple description of the determinant. Let

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Then, according to the definition above, we have

$$\begin{aligned} \det A &= (-1)^{1+1} a_{11} \det(A_{11}) + (-1)^{1+2} a_{12} \det(A_{12}) \\ &= (-1)^{1+1} a_{11} \det(a_{22}) + (-1)^{1+2} a_{12} \det(a_{21}) \\ &= a_{11}a_{22} - a_{12}a_{21}. \end{aligned}$$

Example - Compute the determinant of

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 6 & 7 & 8 \\ 1 & -1 & -1 \end{pmatrix}.$$

Using the definition of the determinant above, this can be written as the sum of three smaller determinants. We obtain

$$\begin{aligned}\det A &= 3\cdot\det\begin{pmatrix}7 & 8\\ -1 & -1\end{pmatrix} - 2\cdot\det\begin{pmatrix}6 & 8\\ 1 & -1\end{pmatrix} + 1\cdot\det\begin{pmatrix}6 & 7\\ 1 & -1\end{pmatrix}\\ &= 3\cdot[7\cdot(-1)-8\cdot(-1)] - 2\cdot[6\cdot(-1)-8\cdot 1] + 1\cdot[6\cdot(-1)-7\cdot 1]\\ &= 3\cdot 1 - 2\cdot(-14) + 1\cdot(-13) = 18.\end{aligned}$$

Another common notation for the determinant of a matrix is to use a pair of vertical lines instead of the usual round brackets for the matrix border (resembling the notation for the absolute value function). So, we may also write

$$\begin{vmatrix}3 & 2 & 1\\ 6 & 7 & 8\\ 1 & -1 & -1\end{vmatrix} = 18.$$

The definition of the determinant is given by expanding along the first row of the matrix. In fact, there is much more flexibility, and we can also expand along any row or column to calculate the determinant.

Theorem 2.26. *Let A be an $n\times n$ matrix, $n\geq 2$. Then, for any $1\leq i\leq n$*

$$\det A = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det A_{ij}$$

and for any $1\leq j\leq n$

$$\det A = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det A_{ij}.$$

We skip the proof of this result because we are running out of time. This result is very useful, in both theory and practice. In particular, it can provide a convenient shortcut for calculating determinants by choosing a row or column that contains many zeroes.

Example - Compute the determinant of

$$\begin{pmatrix}1 & 2 & 100\\ 3 & 5 & \sqrt{2}\\ 0 & 0 & 1\end{pmatrix}.$$

To make things easier, we use the third row. The presence of many zeros in this row makes our calculations quicker.

$$\det A = 0\cdot\det A_{31} - 0\cdot\det A_{32} + 1\cdot\det A_{33} = 1\cdot\det A_{33} = 1\cdot(1\cdot 5-2\cdot 3)) = -1.$$

Another example is the following.

Example - Compute the determinant of

$$\begin{pmatrix}3 & -7 & 8 & 9 & -6\\ 0 & 2 & -5 & 7 & 3\\ 0 & 0 & 1 & 5 & 0\\ 0 & 0 & 2 & 4 & -1\\ 0 & 0 & 0 & -2 & 0\end{pmatrix}.$$

If we used the original definition of the determinant, it would require many computations to calculate $\det A$. However, the many zeroes appearing in the first column can be used to make things easier. We obtain

$$\begin{aligned}\det A &= 3 \cdot \det \begin{pmatrix} 2 & -5 & 7 & 3 \\ 0 & 1 & 5 & 0 \\ 0 & 2 & 4 & -1 \\ 0 & 0 & -2 & 0 \end{pmatrix} \\ &= 3 \cdot 2 \cdot \det \begin{pmatrix} 1 & 5 & 0 \\ 2 & 4 & -1 \\ 0 & -2 & 0 \end{pmatrix} \\ &= 3 \cdot 2 \cdot (-1) \cdot (-1). \det \begin{pmatrix} 1 & 5 \\ 0 & -2 \end{pmatrix} \\ &= 3 \cdot 2 \cdot (-2) = -12.\end{aligned}$$

Definition 2.27. *An $m \times n$ matrix is said to be **upper triangular** if all of the entries below the main diagonal are zero. That is, A is upper triangular if $a_{ij} = 0$ for all $i > j$. A matrix is **lower triangular** if $a_{ij} = 0$ for all $i < j$. A matrix is **triangular** if it is either upper triangular or lower triangular.*

Lemma 2.28. *Let $A = (a_{ij})$ be an $n \times n$ triangular matrix. Then $\det A = a_{11} \cdot a_{22} \cdots a_{nn}$.*

Proof. This is left as an exercise. □

Exercise - Let $A \in \mathbb{R}^{n \times n}$. Show that $\det(A^T) = \det A$.

The next result concerns how row operations change the determinant of a matrix.

Lemma 2.29. *Let $A \in \mathbb{R}^{n \times n}$ with $n \geq 2$. Then,*

1. *If the matrix B is obtained by multiplying one row of A by a scalar $\lambda \in \mathbb{R}$, then $\det B = \lambda \det A$. In particular, $\det(\lambda A) = \lambda^n \det(A)$.*
2. *If the matrix B is obtained by interchanging two rows of A, then $\det B = -\det A$.*
3. *If the matrix B is obtained from A by adding a multiple of one row of A to another row, then $\det B = \det A$.*

Proof. 1. Suppose that B is obtained from A by dilating the ith row by λ. So

$$b_{ij} = \lambda a_{ij}, \ \ \forall \, 1 \leq j \leq n$$

and

$$b_{kj} = a_{kj}, \ \ \forall 1 \leq k, j \leq n \text{ such that } k \neq i.$$

In particular, it follows that $A_{ij} = B_{ij}$ for all $1 \leq j \leq n$. The determinant $\det B$ can be calculated by expanding along the ith row. We obtain

$$\begin{aligned}\det B &= \sum_{j=1}^{n} (-1)^{i+j} b_{ij} \det B_{ij} \\ &= \sum_{j=1}^{n} (-1)^{i+j} \lambda a_{ij} \det B_{ij} \\ &= \sum_{j=1}^{n} (-1)^{i+j} \lambda a_{ij} \det A_{ij} \\ &= \lambda \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij} \\ &= \lambda \det A.\end{aligned}$$

2. Proof by induction on n. The base case $n = 2$ can be verified directly. Suppose that

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

and

$$B = \begin{pmatrix} c & d \\ a & b \end{pmatrix}.$$

Then

$$\det A = ad - bc = -(cb - da) = -\det B.$$

Now let $n \geq 3$ and suppose that the result holds for dimension $(n-1) \times (n-1)$ matrices. Choose a row of B that is the same as the corresponding row of A. Since there are at least 3 rows, and only two rows change, such an unchanged row is guaranteed to exist. The determinant of B can be calculated by expanding along this row (let us say that the ith row is the same in B and A). We obtain

$$\det B = \sum_{j=1}^{n} (-1)^{i+j} b_{ij} \det B_{ij} = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det B_{ij}$$

Now observe that the matrix $B_{ij} \in \mathbb{R}^{(n-1)\times(n-1)}$ is obtained from the matrix A_{ij} by interchanging two rows. Therefore, by the induction hypothesis, $\det B_{ij} = -\det A_{ij}$. Finally,

$$\det B = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det B_{ij} = -\sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij} = -\det A.$$

3. A similar proof by induction argument to the one used for part 2 of this lemma can be used here. This is left as an exercise.

□

Observe that the previous result immediately implies the following handy consequence.

Corollary 2.30. *Let $A \in \mathbb{R}^{n\times n}$ and suppose that B can be obtained from A by row operations. Then*

$$\det A = 0 \iff \det B = 0.$$

Lemma 2.29 can be useful for calculating determinants. The idea is simply to reduce a given matrix to row echelon form (which is a triangular matrix), and to then use Lemma 2.28 to quickly calculate the determinant of the echelon form matrix. We need to keep track of the row operations we carry out, and include this factor in our calculation.

Example - Consider the matrix

$$\begin{pmatrix} 1 & -4 & 2 \\ -2 & 8 & -9 \\ -1 & 7 & 0 \end{pmatrix}.$$

By Lemma 2.29,

$$\det \begin{pmatrix} 1 & -4 & 2 \\ -2 & 8 & -9 \\ -1 & 7 & 0 \end{pmatrix} = \det \begin{pmatrix} 1 & -4 & 2 \\ 0 & 0 & -5 \\ 0 & 3 & 2 \end{pmatrix} = -\det \begin{pmatrix} 1 & -4 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & -5 \end{pmatrix} = (-1)\cdot 1 \cdot 3 \cdot (-5) = 15.$$

In the first step we only performed the operation of adding a multiple of one row to another row ($R_2 = R_2 + 2R_1$ and $R_3 = R_2 + R_1$), so we did not need to introduce a multiplicative factor. In the second step, we switched the sign because we applied a row interchange operation.

The determinant is of special interest, because it can be used to characterise if a matrix is invertible and if a linear system is uniquely solvable. Recalling Theorem 2.16, the latter property is equivalent to the corresponding matrix having full rank.

Lemma 2.31. *Let $A \in \mathbb{R}^{n\times n}$. Then*

$$\det A \neq 0 \iff rank(A) = n.$$

Proof. By Theorem 2.13, A can be transformed into reduced row echelon form C by performing row operations.

($\Rightarrow$) We will prove the contrapositive form. Suppose that $rank(A) \neq n$. Then, by definition, C has a zero row, and in particular, one of the diagonal entries of C is equal to zero. On the other hand, C is an upper triangular matrix, and so Lemma 2.28 implies that $\det C = 0$. Since C is obtained from A via row operations, Corollary 2.30 then implies that $\det A = 0$.

($\Leftarrow$) Suppose that $rank(A) = n$. Then by Lemma 2.15, $C = I_n$. Thus $\det C = 1$, and Corollary 2.30 then implies that $\det A \neq 0$.

□

Lemma 2.32. *Let $A \in \mathbb{R}^{n\times n}$. Then*

$$A \text{ is invertible} \iff rank(A) = n.$$

Proof. By Theorem 2.16,

$$rank(A) = n \iff \forall\, \mathbf{b} \in \mathbb{R}^n, \text{ there exists a unique } \mathbf{x} \in \mathbb{R}^n \text{ such that } A\mathbf{x} = \mathbf{b}. \tag{57}$$

Matrix multiplication can also be considered as a function on vectors. Indeed, recall that we defined the linear transformation $T_A : \mathbb{R}^n \to \mathbb{R}^n$ by

$$T_A(\mathbf{x}) = A\mathbf{x}.$$

The right hand side of the equivalence (57) is equivalent to the function T_A being bijective (for all $\mathbf{b} \in \mathbb{R}^n$, there exists exactly one $\mathbf{x} \in \mathbb{R}^n$ such that $T_A(\mathbf{x}) = \mathbf{b}$). Therefore, by Theorem 1.13 and Theorem 2.24, we conclude that

$$\begin{aligned} rank(A) = n &\iff \forall\, \mathbf{b} \in \mathbb{R}^n, \text{ there exists a unique } \mathbf{x} \in \mathbb{R}^n \text{ such that } A\mathbf{x} = \mathbf{b} \\ &\iff T_A \text{ is a bijection} \\ &\iff T_A \text{ is invertible} \\ &\iff A \text{ is invertible.} \end{aligned}$$

□

Combining the previous two lemmas, we have the following important theorem relating determinants and invertibility.

Theorem 2.33. *Let $A \in \mathbb{R}^{n\times n}$. Then*

$$A \textit{ is invertible} \iff \det A \neq 0.$$

Moreover, we can combine several of the statements we have recently proved into one big statement which shows that several important properties of square matrices are equivalent.

Theorem 2.34. *Let $A \in \mathbb{R}^{n\times n}$. Then the following statements are equivalent.*

1. *A is invertible.*
2. $\det A \neq 0$.
3. *$rank(A) = n$.*
4. *The linear system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution for any $\boldsymbol{b} \in \mathbb{R}^n$.*
5. *A is equivalent to I_n.*

Example - We computed earlier that

$$\begin{vmatrix} 3 & -7 & 8 & 9 & -6 \\ 0 & 2 & -5 & 7 & 3 \\ 0 & 0 & 1 & 5 & 0 \\ 0 & 0 & 2 & 4 & -1 \\ 0 & 0 & 0 & -2 & 0 \end{vmatrix} = -12.$$

It therefore follows from Theorem 2.33 that the matrix

$$\begin{pmatrix} 3 & -7 & 8 & 9 & -6 \\ 0 & 2 & -5 & 7 & 3 \\ 0 & 0 & 1 & 5 & 0 \\ 0 & 0 & 2 & 4 & -1 \\ 0 & 0 & 0 & -2 & 0 \end{pmatrix}$$

is invertible.

The next statement says that the determinant of the product of two matrices is equal to the product of the determinants.

Theorem 2.35. *Let $A, B \in \mathbb{R}^{n\times n}$. Then*

$$\det(AB) = \det(A) \cdot \det(B).$$

We will first prove this theorem in the special case when one of the matrices is elementary.

Lemma 2.36. *Let $B \in \mathbb{R}^{n\times n}$ and let $E \in \mathbb{R}^{n\times n}$ be an elementary matrix. Then*

$$\det(EB) = \det(E) \cdot \det(B).$$

Proof. There are three cases to consider, corresponding to the three row operations that E may represent.

1. Suppose that E is an elementary matrix corresponding to row interchange. Then, since E is obtained from I_n by interchanging two rows and $\det(I_n) = 1$, it follows from Lemma 2.29 that $\det(E) = -1$. On the other hand, the matrix EB is obtained from B by interchanging two rows, and so Lemma 2.29 again implies that $\det(EB) = -\det B = (\det E)(\det B)$.

2. Suppose that E is an elementary matrix corresponding to row dilation. The proof is similar to the first case, and is left as an exercise.

3. Suppose that E is an elementary matrix corresponding to adding a multiple of one row to another row. The proof is similar to the first case, and is left as an exercise.

□

Proof of Theorem 2.35. **Case 1** - Suppose that either $\det(A) = 0$ or $\det(B) = 0$. We consider the case when $\det(A) = 0$ in detail only, as the case when $\det(B) = 0$ can be handled similarly.

It follows from Lemma 2.31 that $rank(A) < n$. Part 1 of Theorem 2.17 then implies that there is some vector $\mathbf{b} \in \mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{b}$ has no solutions.

Suppose for a contradiction that $\det(AB) \neq 0$. Then, by Theorem 2.34, there exists $\mathbf{y} \in \mathbb{R}^n$ such that $(AB)\mathbf{y} = \mathbf{b}$, where $\mathbf{b}$ is the same vector as in the previous paragraph. This is a contradiction, since if we write $B\mathbf{y} = \mathbf{x}$, we see that

$$\mathbf{b} = (AB)\mathbf{y} = A(B\mathbf{y}) = A\mathbf{x},$$

contradicting the fact that $A\mathbf{x} = \mathbf{b}$ has no solutions.

Case 2 - Suppose that $\det(A)$ and $\det(B)$ are both non-zero. Theorem 2.34 then implies that $rank(A) = rank(B) = n$. Then, by Lemma 2.15, it follows that both A and B can be reduced to I_n by elementary row operations.

Since A can be reduced to I_n by a sequence of row operations, this process can be reversed so that I_n is transformed to A by a sequence of row operations. It follows that there is a sequence of elementary matrices $E_1, \dots, E_k$ such that

$$A = E_j \dots E_1 I_n = E_j \dots E_1.$$

Similarly,

$$B = F_k \dots F_1,$$

where $F_1, \dots F_k$ are elementary matrices.

Repeated applications of Lemma 2.36 imply that

$$\begin{aligned} \det(A) &= \det(E_j \dots E_1) \\ &= \det(E_j) \det(E_{j-1} \dots E_1) \\ &\vdots \\ &= \det(E_j) \cdots \det(E_1). \end{aligned} \tag{58}$$

Similarly,

$$\det(B) = \det(F_k) \cdots \det(F_1). \tag{59}$$

Finally,

$$\begin{aligned} \det(AB) &= \det(E_j \dots E_1 F_k \dots F_1) \\ &= \det(E_j) \det(E_{j-1} \dots E_1 F_k \dots F_1) \\ &\vdots \\ &= \det(E_j) \cdots \det(E_1) \cdot \det(F_k) \cdots \det(F_1) \\ &= \det(A) \cdot \det(B). \end{aligned}$$

In the final step, we have used (58) and (59). □

Example

Consider the matrix

$$A = \begin{pmatrix} 15 & 10 & 24 \\ 15 & 22 & 12 \\ 12 & 4 & 33 \end{pmatrix}.$$

It looks like it might be tricky to compute $\det A$, at least without a calculator. We are, fortunately, given the information that

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 3 & 4 & 0 \\ 2 & 0 & 5 \end{pmatrix} \begin{pmatrix} 1 & 2 & 4 \\ 3 & 4 & 0 \\ 2 & 0 & 5 \end{pmatrix}.$$

We can therefore deduce from Theorem 2.35 that

$$\det A = \left(\begin{vmatrix} 1 & 2 & 4 \\ 3 & 4 & 0 \\ 2 & 0 & 5 \end{vmatrix} \right)^2.$$

It remains to compute the easier determinant, which we do here by expanding along the final row:

$$\begin{vmatrix}1 & 2 & 4\\ 3 & 4 & 0\\ 2 & 0 & 5\end{vmatrix} = 2\begin{vmatrix}2 & 4\\ 4 & 0\end{vmatrix} + 5\begin{vmatrix}1 & 2\\ 3 & 4\end{vmatrix} = 2\cdot(-16) + 5\cdot(-2) = -42.$$

Therefore,

$$\det A = (-42)^2 = 1764.$$

We can also use Theorem 2.35, combined with Theorem 2.34, to quickly determine whether a product of two matrices is invertible.

Theorem 2.37. *Let $A, B \in \mathbb{R}^{n\times n}$. Then*

$$AB \text{ is invertible} \iff A \text{ and } B \text{ are both invertible.}$$

Proof.

$$\begin{aligned} AB \text{ is invertible} &\iff \det(AB) \neq 0\\ &\iff (\det A)\cdot(\det B) \neq 0\\ &\iff \det A \neq 0 \text{ and } \det B \neq 0\\ &\iff A \text{ and } B \text{ are both invertible.}\end{aligned}$$

□

Now that we know that the determinant "behaves well" with respect to transposition and multiplication, one might guess that a similar relation also holds for addition. However, this is not the case, and **there is no similar formula for the determinant of the sum of matrices** as the following simple example shows. **Example** Consider the matrices

$$A = \begin{vmatrix}1 & 0\\ 0 & 0\end{vmatrix} \quad \text{and} \quad B = \begin{vmatrix}0 & 0\\ 0 & 1\end{vmatrix},$$

and observe that $A + B = I_2$. Since A and B both contain a zero row, it follows that

$$\det A = 0 = \det B.$$

Since $\det(A + B) = \det I_n = 1$, it follows that

$$\det(A + B) \neq \det A + \det B.$$

Moreover, this example also shows that an analogue of Theorem 2.37, with sums instead of products, is not true.

2.6 Inverse matrices

We now take a closer look at inverse matrices. One of the main motivations for working with the inverse matrix is that it can be used to solve linear systems in a straightforward way. Indeed, since the inverse matrix A^{-1} satisfies $AA^{-1} = A^{-1}A = I_n$, we have that

$$A\mathbf{x} = \mathbf{b} \iff \mathbf{x} = A^{-1}\mathbf{b}. \tag{60}$$

In the previous section, we saw how the determinant can be used to quickly determine whether or not a matrix is invertible. However, so far, we do not have a method for calculating what the inverse is. Developing such a method will be the focus of this section. There are different methods for calculating the inverse of a matrix, and we choose a method which is an extension of the Gaussian elimination tecniques we learned earlier in this chapter.

Suppose that we have a matrix $A \in \mathbb{R}^{n\times n}$ that we know is invertible. Let $\mathbf{c}_i$ denote the ith column of A^{-1}, so

$$A^{-1} = (\mathbf{c}_1 \dots \mathbf{c}_n).$$

Recall that we discussed earlier (see the discussion before Lemma 2.2) that *unit vectors can be used to extract columns from matrices.* In particular,

$$\mathbf{c}_i = A^{-1}\mathbf{e}_i.$$

Then, following the logic of (60) (in other words, left-multplying both sides of the above equation by A), we have

$$A\mathbf{c}_i = \mathbf{e}_i.$$

This shows that calculating $\mathbf{c}_i$ is equivalent to solving the linear system $A\mathbf{x} = \mathbf{e}_i$. In section 2.3, we learnt how to do this by considering the augmented matrix $(A|\mathbf{e}_i)$, and using Gaussian elimination to reduce this to $(I_n|\mathbf{x})$. The resulting right hand side of the resulting augmented matrix is a solution to the system $A\mathbf{x} = \mathbf{b}$.

To use this method to calculate the inverse A^{-1}, we must repeat this procedure for each column of A. However, we can apply the Gaussian elimination procedure to several vectors at once! Hence, we can compute all columns of A^{-1} at once by computing the reduced row echelon form of the augmented matrix

$$\left(\begin{array}{ccc|ccc} a_{11} & \dots & a_{1n} & 1 & & 0 \\ \vdots & & \vdots & & \ddots & \\ a_{n1} & \dots & a_{nn} & 0 & & 1 \end{array}\right).$$

If A is invertible then the reduced echelon form of A is the identity matrix I_n (see Theorem 2.34. Thus, by using Gaussian elimination, we are able to compute

$$(A|I_n) \to (I|A^{-1}).$$

We consolidate the discussion above in the following theorem.

Theorem 2.38. *Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix. Then, the reduced row echelon form of*

$$\left(\begin{array}{ccc|ccc} a_{11} & \dots & a_{1n} & 1 & & 0 \\ \vdots & & \vdots & & \ddots & \\ a_{n1} & \dots & a_{nn} & 0 & & 1 \end{array}\right).$$

has the form

$$\left(\begin{array}{ccc|ccc} 1 & & 0 & a'_{11} & \dots & a'_{1n} \\ & \ddots & & \vdots & & \vdots \\ 0 & & 1 & a'_{n1} & \dots & a'_{nn} \end{array}\right).$$

The matrix on the right hand side of the second augmented matrix above is the inverse of A. That is,

$$A^{-1} = (a'_{ij})_{i,j=1}^n.$$

Example - We want to compute the inverse of

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

By Gaussian elimination

$$\left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \end{array}\right) \xrightarrow[R_2 = R_2 - 3R_1]{} \left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & -2 & -3 & 1 \end{array}\right)$$
$$\xrightarrow[R_1 = R_1 + R_2]{} \left(\begin{array}{cc|cc} 1 & 0 & -2 & 1 \\ 0 & -2 & -3 & 1 \end{array}\right)$$
$$\xrightarrow[R_2 = (-1/2)R_2]{} \left(\begin{array}{cc|cc} 1 & 0 & -2 & 1 \\ 0 & 1 & 3/2 & -1/2 \end{array}\right).$$

It therefore follows from Theorem 2.38 that

$$A^{-1} = \begin{pmatrix} -2 & 1 \\ 3/2 & -1/2 \end{pmatrix}. \tag{61}$$

Note that this conclusion agrees with the claim made on page 79, when we stated without proof that this matrix A is invertible with the inverse matrix as in (61).

There are many opportunities for miscalculations throughout this process, and it is a good idea to check that your solution is correct by checking that $AA^{-1} = I_n$.

Example

We want to compute the inverse of

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 4 & 1 & 8 \\ 0 & 1 & 1 \end{pmatrix}.$$

By Gaussian elimination

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 4 & 1 & 8 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \xrightarrow[R_2 = R_2 - 4R_1]{} \left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & -4 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right)$$
$$\xrightarrow[R_3 = R_3 - R_2]{} \left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & -4 & 1 & 0 \\ 0 & 0 & 1 & 4 & -1 & 1 \end{array}\right)$$
$$\xrightarrow[R_1 = R_1 - 2R_3]{} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -7 & 2 & -2 \\ 0 & 1 & 0 & -4 & 1 & 0 \\ 0 & 0 & 1 & 4 & -1 & 1 \end{array}\right).$$

It therefore follows from Theorem 2.38 that

$$A^{-1} = \begin{pmatrix} -7 & 2 & -2 \\ -4 & 1 & 0 \\ 4 & -1 & 1 \end{pmatrix}.$$

3 Sequences and Series

This chapter is dedicated to formalising the idea of the limiting processes. It forms one of the central ideas of mathematical analysis and defines the basis for essential concepts like continuity, differentiability, integration etc.

For example, consider the following infinite sum of powers of 2:

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots=\sum_{k=0}^{\infty}2^{-k}.$$

As we add more and more terms (i.e. as k gets larger and larger), we get closer and closer to 2, and we can get as close as we want (i.e. arbitrarily close) by adding enough terms. It is intuitive to suggest that

$$\sum_{k=0}^{\infty}2^{-k}=2.$$

We will develop a framework to consider such infinite sums more formally in this chapter. We begin with the definition of a *sequence*.

Definition 3.1. *Let $M \neq \emptyset$ be an arbitrary set, and let $I \subset \mathbb{Z}$ be an infinite set. A* ***sequence*** *in M is a mapping $a : I \to M$. With the notation $a_n := a(n)$, we can write the sequence as $(a_n)_{n\in I}$.*

The ***range*** *of a sequence $(a_n)_{n\in I}$ is the set $\{a_n : n \in I\}$. The domain I of a sequence is called the* ***index set*** *of the sequence.*

In most cases, we consider $I = \mathbb{N}$ or $I = \{K, K+1, \dots\}$ for some $K \in \mathbb{Z}$. In the latter case, we write the sequence as $(a_n)_{n\in I} = (a_n)_{n=K}^{\infty}$. If the index set is clear, we may just write (a_n) for $(a_n)_{n\in I}$.

In the special cases $M = \mathbb{R}$ or $M = \mathbb{C}$ we say that $(a_n)_{n\in I}$ is a **real-valued** or **complex-valued** sequence, respectively. In this course, we will almost always deal with real-valued or complex-valued sequences.

To define a sequence, the most common way is to use an explicit formula, for instance

$$a_n = 2^n \tag{62}$$

or

$$b_n = 1+\frac{1}{n}.$$

We can also define a sequence by **recursion**. This means that we give one (or more) starting value(s) and a rule for how to calculate a new term using previous terms. For example, we can set $a_1 = 2$ and define $a_i = 2a_{i-1}$ for all $i \geq 2$. This is another description of the sequence (62).

Example - Consider the sequence $(a_n)_{n\in\mathbb{N}}$ given by $a_n = \frac{1}{n}$. We can also represent this sequence by listing its elements, that is this sequence is the same thing as the list

$$1, \frac{1}{2}, \frac{1}{3}, \dots.$$

We can immediately observe that the terms of this sequence get very close to 0 as n gets large, but never reach zero. We say that the sequence converges to 0. We will give a formal definition of what this means in the next section.

Example - One of the most famous sequences, which appears in several areas of natural science, is the **Fibonacci sequence**. Here, the recursion depends on the previous two values. The sequence $(F_n)_{n\in\mathbb{N}}$ is defined by

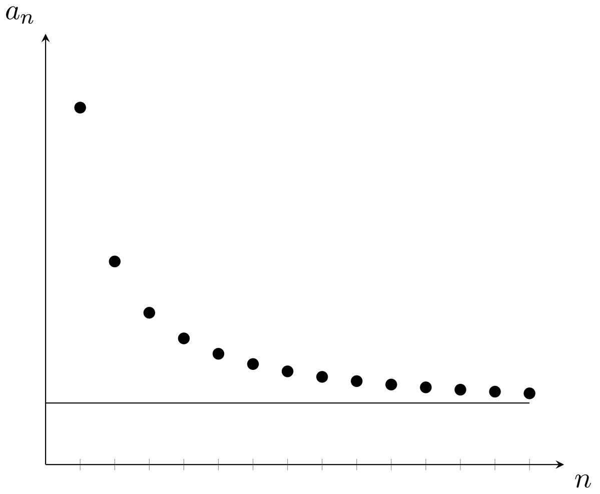
$$F_1 = 1, F_2 = 1, \quad \text{and} \quad F_n = F_{n-1} + F_{n-2} \quad \text{for} \quad n \geq 3.$$

The first values of this sequence are $1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots$. It is an interesting phenomenon that the quotients F_{n+1}/F_n converge to the **golden ratio** $\frac{1+\sqrt{5}}{2}$. (See e.g. Wikipedia for background on the importance of this constant).

Example - Given a sequence a_n, we can consider the new sequence $s_n = \sum_{k=1}^{n} a_k$. We would like to know if s_n approaches a certain number when n goes to infinity. These special sequences are called **series**, and we come back to this later in the chapter.

3.1 Convergence of sequences

The concept of convergence is central to mathematical analysis. Intuitively, it states that the terms of the sequence $(a_n)_{n\in\mathbb{N}}$ approach a limit with growing index n.



To define what it means to converge to something, the notion of a neighbourhood may be useful.

Definition 3.2. *Let $M = \mathbb{R}$ or $M = \mathbb{C}$, $a \in M$ and let $\epsilon > 0$ be a real number. We define the **ϵ-neighbourhood** of a in M by*

$$U_\epsilon(a) := \{x \in M : |x - a| < \epsilon\}.$$

Note that, for $M = \mathbb{R}$, the ϵ-neighbourhood $U_\epsilon(a)$ is just the open interval $(a - \epsilon, a + \epsilon)$. For $M = \mathbb{C}$, $U_\epsilon(a)$ is the disc of radius ϵ around centred at a in the complex plane $\mathbb{C}$.

One can also consider neighbourhoods in a much more general situation. All we need is a notion of distance. For instance, one may define the ϵ-neighbourhood of a point $\mathbf{x} \in \mathbb{R}^d$. We may consider such generalisations later in the course.

We now come to the formal definition of *convergence to a limit a*.

Definition 3.3. *Let $(a_n)_{n\in\mathbb{N}}$ be a complex-valued sequence and $a \in \mathbb{C}$. We say that the sequence $(a_n)_{n\in\mathbb{N}}$* ***converges*** *to a if and only if*

$$\forall\, \epsilon > 0,\, \exists\, n_0 \in \mathbb{N} :\, n \geq n_0 \implies |a_n - a| < \epsilon.$$

An equivalent definition is

$$\forall\, \epsilon > 0,\, \exists\, n_0 \in \mathbb{N} :\, n \geq n_0 \implies a_n \in U_\epsilon(a).$$

We call a the ***limit*** *of the sequence and write*

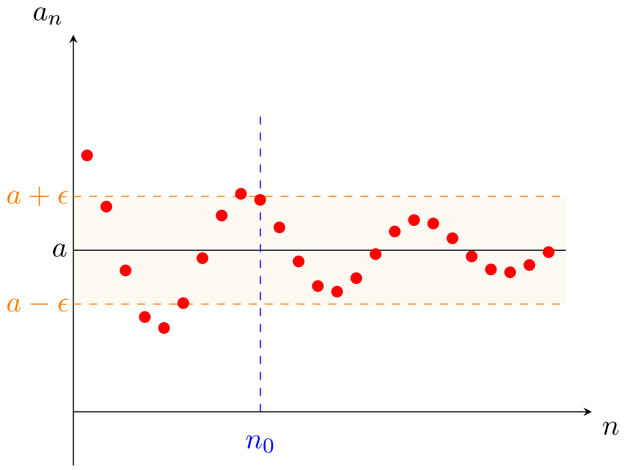
$$a = \lim_{n\to\infty} a_n$$

which is sometimes abbreviated to

$$a_n \to a.$$

The sequence $(a_n)_{n\in\mathbb{N}}$ is called ***convergent*** *if there exists some $a \in \mathbb{C}$ such that $a_n \to a$. Otherwise, the sequence $(a_n)_{n\in\mathbb{N}}$ is called* ***divergent****.*

This definition may appear intimidating to some readers on first viewing. Let's try to draw a picture to illustrate the meaning of the definition.



What this definition says is that, no matter how tiny we choose ϵ to be, the sequence will always eventually stay within a distance ϵ of the limit a.

Note that the limit does not depend on the first terms of a sequence. In particular, we can always disregard finitely many elements of a sequence when considering its limiting behaviour. For instance, if $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are two sequences such that $a_n = b_n$ holds for all but finitely many values of n, and $\lim_{n\to\infty} a_n = a$, then $\lim_{n\to\infty} b_n = a$.

Let us consider some examples.

Example - Consider again the sequence $(a_n)_{n\in\mathbb{N}}$ with $a_n = \frac{1}{n}$. For all $\epsilon > 0$ we can find some $n_0 \in \mathbb{N}$ such that $\frac{1}{n_0} < \epsilon$. This is the Archimedean property (in particular, see Theorem 1.39). Since $\frac{1}{n} \leq \frac{1}{n_0}$ for $n \geq n_0$, we obtain

$$|a_n - 0| = \frac{1}{n} \leq \frac{1}{n_0} < \epsilon$$

for all $n \geq n_0$. Therefore, $a_n \to 0$.

Example - Consider the sequence $(a_n)_{n\in\mathbb{N}}$ with $a_n = (-1)^n$, i.e. the sequence which alternates between 1 and -1. This sequence is divergent. For a proof we assume the opposite, i.e., that (a_n) converges to some $a \in \mathbb{R}$. Now, by the definition of convergence, we have that there exists some n_0 such that

$$a_n \in U_{\frac{1}{2}}(a) \text{ for all } n \geq n_0$$

and thus we have by the triangle inequality,

$$|a_{n+1} - a_n| = |a_{n+1} - a + a - a_n| \leq |a_{n+1} - a| + |a - a_n| < \frac{1}{2} + \frac{1}{2} < 2 \tag{63}$$

for all $n \leq n_0$. (Note that the value $1/2$ is quite arbitrary here, and any choice of $\epsilon < 1$ would work.) However, it is also true for all $n \in \mathbb{N}$ that

$$|a_{n+1} - a_n| = 2.$$

But this contradicts (63), and completes the proof.

Definition 3.4. *Let $(a_n)_{n\in\mathbb{N}}$ be a real sequence such that $\lim_{n\to\infty} a_n = 0$. Then we call $(a_n)_{n\in\mathbb{N}}$ a* ***null sequence****.*

Example - The sequences $\left(\frac{1}{n}\right)_{n\in\mathbb{N}}$ and $(2^{-n})_{n\in\mathbb{N}}$ are null sequences.

Exercise - Show that, for any $c > 0$ the sequence $\left(\frac{1}{n^c}\right)_{n\in\mathbb{N}}$ is a null sequence.

We now consider the notion of boundedness. This is very similar to the notion of boundedness for sets which we considered in Section 1.6.

Definition 3.5. *Let $(a_n)_{n\in\mathbb{N}}$ be a complex valued sequence. We call the sequence* ***bounded*** *(by C) if and only if*

$$\exists\, C > 0 \ : \forall\, n \in \mathbb{N}, |a_n| \leq C.$$

A real-valued sequence $(a_n)_{n\in\mathbb{N}}$ is ***bounded from above*** *if and only if*

$$\exists\, C \in \mathbb{R} : \ \forall\, n \in \mathbb{N}, a_n \leq C,$$

and ***bounded from below*** *if and only if*

$$\exists\, C \in \mathbb{R} : \ \forall\, n \in \mathbb{N}, a_n \geq C.$$

Examples - The sequences

$$((-1)^n)_{n\in\mathbb{N}}$$

and

$$\left(\frac{42}{n}\right)_{n\in\mathbb{N}}$$

are bounded (by 1 and 42 respectively). We also have that $\left((-1)^n\frac{42}{n}\right)_{n\in\mathbb{N}}$ is bounded (by 42). Actually, we easily see from the definition that the (term-wise) product of two bounded sequences is bounded. The triangle inequality shows that the sum of two bounded sequences is also bounded.

Let $\theta \in [0, 2\pi)$ be arbitrary and consider the sequence

$$(e^{ni\theta})_{n\in\mathbb{N}}.$$

This sequence is bounded by 1.

The sequence

$$(\sqrt{n})_{n\in\mathbb{N}}$$

is bounded from below by 0 and is not bounded from above. What about the sequence

$$((-1)^{F(n)}\sqrt{n})_{n\in\mathbb{N}},$$

where $F(n)$ is the nth term in the Fibonnaci sequence?

Next, we make an observation concerning the relationship between the convergence and boundedness of a sequence.

Theorem 3.6. *Let $(a_n)_{n\in\mathbb{N}}$ be a convergent sequence. Then $(a_n)_{n\in\mathbb{N}}$ is bounded.*

Proof. Let $a = \lim_{n\to\infty} a_n$. By the definition of convergence (with $\epsilon = 1$), there exists n_0 such that, for all $n \geq n_0$,

$$|a - a_n| < 1.$$

It follows from the triangle inequality that

$$|a_n| = |(a_n - a) + a| \leq |a_n - a| + |a| < 1 + |a|$$

holds for all $n \geq n_0$. It therefore follows that, for all $n \in \mathbb{N}$,

$$|a_n| \leq \max\{1 + |a|, |a_1|, |a_2|, \ldots, |a_{n_0-1}|\}.$$

□

Note that the sequence $(a_n)_{n\in\mathbb{N}}$ given by $a_n = (-1)^n$ is bounded by 1 but not convergent. Therefore, the opposite implication for the theorem above does not hold in general.

Example - Consider the sequence $(a_n)_{n\in\mathbb{N}}$ with $a_n = \log_2 n$. This sequence is unbounded, and it therefore follows from Theorem 3.6 that the sequence is divergent.

In the previous example, we have seen a sequence that appears to "converge to infinity". We now introduce the terminology of *definite divergence* of a real-valued sequence in order to give a formal description for this kind of situation.

Definition 3.7. *Let $(a_n)_{n\in\mathbb{N}}$ be a real-valued sequence. The sequence $(a_n)_{n\in\mathbb{N}}$ tends to ∞ if and only if*

$$\forall\, C \in \mathbb{R}, \exists\, n_0 \in \mathbb{N} : \forall\, n \geq n_0, a_n \geq C.$$

In this case, we write $\lim_{n\to\infty} a_n = \infty$ or $a_n \to \infty$ and call ∞ the ***improper limit*** *of $(a_n)_{n\in\mathbb{N}}$.*

The sequence $(a_n)_{n\in\mathbb{N}}$ tends to $-\infty$ if and only if

$$\forall\, C \in \mathbb{R}, \exists\, n_0 \in \mathbb{N} : \forall\, n \geq n_0, a_n \leq C.$$

In this case, we write $\lim_{n\to\infty} a_n = -\infty$ or $a_n \to -\infty$ and call $-\infty$ the ***improper limit*** *of $(a_n)_{n\in\mathbb{N}}$.*

If the sequence $(a_n)_{n\in\mathbb{N}}$ tends to ∞ or $-\infty$, it is called ***definitely divergent****.*

Note that definitely divergent sequences are necessarily unbounded. Moreover, we do not have such a concept for complex-valued sequences, as we do not have an order on $\mathbb{C}$.

3.2 Calculation rules for limits

We now study how to determine the limit of more complicated sequences. This always follows the same procedure; either we already know the limit of the sequence under consideration, or one has to split up the sequence into easier parts that can be handled, or split again.

The following result helps us to reduce the task of determining a limit to several smaller and hopefully easier limits.

Theorem 3.8. *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be convergent complex-valued sequences and let $\lambda \in \mathbb{C}$. Let $a = \lim_{n\to\infty} a_n$ and $b = \lim_{n\to\infty} b_n$. Then, we have*

(i) $\lim_{n\to\infty}(a_n + b_n) = a + b$,

(ii) $\lim_{n\to\infty}(\lambda \cdot a_n) = \lambda \cdot a$,

(iii) $\lim_{n\to\infty}(a_n \cdot b_n) = a \cdot b$,

(iv) if $b \neq 0$ and $b_n \neq 0$ for all $n \in \mathbb{N}$, then

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{a}{b}.$$

Proof. (i) Let $\epsilon > 0$ be arbitrary. By the definition of convergence, there exist $m_0, n_0 \in \mathbb{N}$ such that

$$n \geq m_0 \implies |a_n - a| < \frac{\epsilon}{2}$$

and

$$n \geq n_0 \implies |b_n - b| < \frac{\epsilon}{2}.$$

In particular, for all $n \geq \max\{m_0, n_0\}$, we have $|a_n - a|, |b_n - b| < \frac{\epsilon}{2}$. Then, by the triangle inequality

$$|(a_n + b_n) - (a + b)| = |(a_n - a) + (b_n - b)| \leq |a_n - a| + |b_n - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

holds for all $n \geq \max\{m_0, n_0\}$.

(ii) Let $\epsilon > 0$ be arbitrary. By the definition of convergence, there exists $n_0 \in \mathbb{N}$ such that

$$n \geq n_0 \implies |a_n - a| < \frac{\epsilon}{|\lambda|}.$$

Therefore, for all $n \geq n_0$,

$$|\lambda a_n - \lambda a| = |\lambda(a_n - a)| = |\lambda||a_n - a| < |\lambda|\frac{\epsilon}{|\lambda|} = \epsilon.$$

(iii) Since (b_n) is a convergent sequence, it follows from Theorem 3.6 that there is some $C > 0$ such that

$$|b_n| \leq C$$

holds for all $n \in \mathbb{N}$.

Let $\epsilon > 0$ be arbitrary. By the definition of convergence, there exist $m_0, n_0 \in \mathbb{N}$ such that

$$n \geq m_0 \implies |a_n - a| < \frac{\epsilon}{2C}$$

and

$$n \geq n_0 \implies |b_n - b| < \frac{\epsilon}{2|a|}.$$

In particular, for all $n \geq \max\{m_0, n_0\}$, we have both $|a_n - a| < \frac{\epsilon}{2C}$ and $|b_n - b| < \frac{\epsilon}{2|a|}$. Then, by the triangle inequality,

$$\begin{aligned} |a_n b_n - ab| = |a_n b_n - ab_n + ab_n - ab| &\leq |a_n b_n - ab_n| + |ab_n - ab| \\ &= |b_n||a_n - a| + |a||b_n - b| \\ &\leq C|a_n - a| + |a||b_n - b| \\ &< C\frac{\epsilon}{2C} + |a|\frac{\epsilon}{2|a|} = \epsilon. \end{aligned}$$

(iv) We will show that

$$\lim_{n\to\infty} \frac{1}{b_n} = \frac{1}{b}. \tag{64}$$

Once this has been proven, part (iv) of the Theorem follows by an application of part (iii). It remains to prove (64).

Let $\epsilon > 0$ be arbitrary. There exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$|b_n - b| < \frac{1}{2}\min\{|b|, \epsilon|b|^2\}.$$

It follows from Theorem 1.50 and $|b_n - b| < \frac{1}{2}|b|$, that

$$|b_n| > \frac{1}{2}|b|$$

which is equivalent to

$$\frac{1}{|b_n|} < \frac{2}{|b|}$$

holds for all $n \geq n_0$. Therefore, for all $n \geq n_0$

$$\left|\frac{1}{b_n} - \frac{1}{b}\right| = \left|\frac{b - b_n}{bb_n}\right| = \frac{1}{|b|} \cdot \frac{1}{|b_n|} \cdot |b - b_n| < \frac{1}{|b|} \cdot \frac{1}{|b_n|} \cdot \frac{1}{2}\epsilon|b|^2 < \frac{1}{|b|} \cdot \frac{2}{|b|} \cdot \frac{1}{2}\epsilon|b|^2 = \epsilon.$$

□

Example - We can use the previous theorem to calculate the limit of the sequence $(a_n)_{n\in\mathbb{N}}$ given by $a_n = 1 + \frac{\pi}{n}$. Write $a_n = b_n + \pi c_n$ with

$$b_n = 1,\ \forall n \in \mathbb{N} \ \text{ and } \ c_n = \frac{1}{n}.$$

Note that both of the sequences (b_n) and (c_n) are convergent, with limits 1 and 0 respectively. By applying the first and second points of Theorem 3.8, it follows that

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n + \pi \cdot \lim_{n\to\infty} c_n = 1 + \pi \cdot 0 = 1.$$

Example - Consider the sequence $(a_n)_{n\in\mathbb{N}}$ given by

$$a_n = \frac{3n^2+4n+100}{7n^2+13n+\sqrt{n}}.$$

We can use Theorem 3.8 to calculate the value of $\lim_{n\to\infty} a_n$. Divide the numerator and denominator by n^2 to express a_n as

$$a_n = \frac{3+\frac{4}{n}+\frac{100}{n^2}}{7+\frac{13}{n}+\frac{1}{n^{3/2}}}.$$

Let

$$b_n := 3+\frac{4}{n}+\frac{100}{n^2}, \quad c_n = 7+\frac{13}{n}+\frac{1}{n^{3/2}}.$$

It follows from (parts (i) and (ii) of) Theorem 3.8 that the sequences $(b_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ are convergent with

$$\lim_{n\to\infty} b_n = 3, \quad \lim_{n\to\infty} c_n = 7.$$

It then follows from Theorem 3.8(iv) that

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{b_n}{c_n} = \frac{3}{7}.$$

In order to make more use of Theorem 3.8, we need to build up a bigger collection of simple sequences for which we know that they are convergent and what they converge to. Let us start with a lemma that shows how to verify that a sequence is a null sequence by comparison with another null sequence.

Lemma 3.9. *Let $(a_n)_{n\in\mathbb{N}}$ be a complex-valued null sequence and let $c, C > 0$ be arbitrary positive real numbers. Suppose that the sequence $(b_n)_{n\in\mathbb{N}}$ satisfies*

$$|b_n| \leq C|a_n|^c$$

for all but finitely many $n \in \mathbb{N}$. Then $(b_n)_{n\in\mathbb{N}}$ is a null sequence.

Proof. This is left as an exercise. □

Example - Consider the sequence $(b_n)_{n\in\mathbb{N}}$ given by

$$b_n = \frac{10000}{n^{0.00001}}.$$

It follows from Lemma 3.9 that (b_n) is a null sequence. Indeed,

$$b_n = 10000|a_n|^{0.00001}$$

where $a_n = \frac{1}{n}$ and $(a_n)_{n\in\mathbb{N}}$ is a null sequence.

Let us now consider other important "building blocks", i.e., limits that may be considered known from now on, together with the corresponding proofs. The first example is concerned with powers of small complex bases.

Lemma 3.10. *Let $z \in \mathbb{C}$ with $|z| < 1$. Then*

$$\lim_{n\to\infty} z^n = 0.$$

Proof. Let

$$x := \frac{1}{|z|} - 1.$$

Observe that x is a positive real number. Bernoulli's Inequality (or even the Binomial Theorem, see Corollary 1.43) implies that $(1+x)^n \geq 1 + nx$ holds for all $n \in \mathbb{N}$. Therefore,

$$|z^n| = |z|^n = \left(\frac{1}{1+x}\right)^n = \frac{1}{(1+x)^n} \leq \frac{1}{1+nx} < \frac{1}{nx} = \frac{1}{x} \cdot a_n,$$

where $a_n = \frac{1}{n}$. Note that $(a_n)_{n\in\mathbb{N}}$ is a null sequence. An application of Lemma 3.9 (with $c = 1$ and $C = \frac{1}{x} > 0$) implies that $(z^n)_{n\in\mathbb{N}}$ is also a null sequence.

□

Exercise - Let $z \in \mathbb{C}$ with $|z| > 1$. Show that the sequence $(z^n)_{n\in\mathbb{N}}$ is divergent. For what $z \in \mathbb{C}$ with $|z| = 1$ is the sequence $(z^n)_{n\in\mathbb{N}}$ convergent?

The next limit will be a useful building block for establishing the convergence of other sequences later.

Lemma 3.11.

$$\lim_{n\to\infty} \sqrt[n]{n} = 1.$$

Proof. Define

$$x_n = \sqrt[n]{n} - 1.$$

We will show that $(x_n)_{n\in\mathbb{N}}$ is a null sequence. It then follows from Theorem 3.8(i) that

$$\lim_{n\to\infty} \sqrt[n]{n} = \lim_{n\to\infty} (x_n + 1) = 0 + 1 = 1.$$

To show that (x_n) is a null sequence, we use Corollary 1.43 with $m = 2$ to deduce that

$$n = (1 + x_n)^n \geq 1 + \binom{n}{2} x_n^2 = 1 + \frac{n(n-1)}{2} x_n^2.$$

A rearrangement of this inequality gives

$$x_n \leq \sqrt{2} \cdot \frac{1}{\sqrt{n}}.$$

It then follows from Lemma 3.9 that (x_n) is a null sequence. □

We will now use Lemma 3.11 in combination with Theorem 3.8 to give a strengthened version of the previous result in which we allow the argument of the root to grow even faster.

Lemma 3.12. *Let $k \in \mathbb{N}$. Then*

$$\lim_{n\to\infty} \sqrt[n]{n^k} = 1$$

Proof. The proof is by induction on k. The base case $k = 1$ was established in Lemma 3.11.

Now, suppose that the result holds for k. We need to prove that

$$\lim_{n\to\infty} a_n = 1.$$

where

$$a_n = \sqrt[n]{n^{k+1}}.$$

Let $b_n = \sqrt[n]{n}$ and $c_n = \sqrt[n]{n^k}$, and observe that $a_n = b_n \cdot c_n$. Also, by Lemma 3.11 and the induction hypothesis respectively, we have both

$$\lim_{n\to\infty} b_n = 1$$

and

$$\lim_{n\to\infty} c_n = 1.$$

It therefore follows from Theorem 3.8(iii) that

$$\lim_{n\to\infty} a_n = \left(\lim_{n\to\infty} b_n\right) \cdot \left(\lim_{n\to\infty} c_n\right) = 1 \cdot 1 = 1.$$

□

The following result illustrates the fact that exponential growth is faster than polynomial growth.

Lemma 3.13. *Let $z \in \mathbb{C}$ with $|z| > 1$ and let $k \in \mathbb{N}$ be fixed. Then*

$$\lim_{n\to\infty} \frac{n^k}{z^n} = 0.$$

Proof. Our plan for this proof is as follows: we will show that there is some constant C such that

$$\left|\frac{n^k}{z^n}\right| \le C\frac{1}{n} \tag{65}$$

holds for all but finitely many $n \in \mathbb{N}$. It then follows from Lemma 3.9 that $(\frac{n^k}{z^n})$ is a null sequence. It remains to prove (65) for almost all $n \in \mathbb{N}$.

Set $x := |z| - 1 > 0$ and suppose that $n \ge 2k$ which is equivalent to $n - k \ge \frac{n}{2}$. This assumption on n is ok for us, since we are disregarding only finitely many values of n. Apply Corollary 1.43 with this x to obtain

$$\begin{aligned} |z|^n = (1+x)^n &\ge 1 + \binom{n}{k+1} x^{k+1} \\ &= 1 + \frac{n(n-1)\cdots(n-k)}{(k+1)!} x^{k+1} \\ &\ge \frac{(n/2)^{k+1}}{(k+1)!} x^{k+1} \\ &= \frac{n^{k+1}}{(k+1)!2^{k+1}} x^{k+1}. \end{aligned}$$

It therefore follows that, for all $n \geq 2k$,

$$\left|\frac{n^k}{z^n}\right| = \frac{n^k}{|z|^n} \leq n^k \cdot \frac{(k+1)!2^{k+1}}{x^{k+1}n^{k+1}} = \frac{1}{n} \cdot C,$$

where $C = \frac{(k+1)!2^{k+1}}{x^{k+1}}$ is an absolute constant (i.e. it is independent of n). This proves (65), and completes the proof of the lemma. □

The next result gives a very helpful tool for the calculation of difficult limits. This one is helpful when the sequence under consideration can be bounded from above and below by sequences that converge to the same limit.

Theorem 3.14. *(Sandwich Rule) Let $(a_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ be convergent real-valued sequences. Suppose that $(b_n)_{n\in\mathbb{N}}$ is a real-valued sequence and that there exists $n_0 \in \mathbb{N}$ such that*

$$a_n \leq b_n \leq c_n \quad \forall n \geq n_0.$$

Suppose also that

$$\lim_{n\to\infty} a_n = L = \lim_{n\to\infty} c_n.$$

Then $(b_n)_{n\in\mathbb{N}}$ is convergent and

$$\lim_{n\to\infty} b_n = L.$$

Proof. This is left as an exercise. □

Our first application of the Sandwich Rule is to prove a variant of Lemma 3.11.

Lemma 3.15. *Let a be a positive real number. Then*

$$\lim_{n\to\infty} \sqrt[n]{a} = 1.$$

Proof. The lemma holds for trivial reasons if we set $a = 1$. For $a > 1$, observe that, for all $n \geq a$

$$1 \leq \sqrt[n]{a} \leq \sqrt[n]{n}.$$

We have bounded the sequence $(\sqrt[n]{a})_{n\in\mathbb{N}}$ from above and below by two sequences which both converge to 1 (by Lemma 3.11). It therefore follows from the Sandwich Rule that

$$\lim_{n\to\infty} \sqrt[n]{a} = 1.$$

For $0 < a < 1$, let $(b_n)_{n\in\mathbb{N}}$ be the sequence given by

$$b_n := \sqrt[n]{\frac{1}{a}}.$$

Since $1/a > 1$, it follows from what we have just proven that $\lim_{n\to\infty} b_n = 1$. In particular, for all $\epsilon > 0$, there exists n_0 such that, for all $n \geq n_0$,

$$|b_n - 1| < \epsilon.$$

Therefore, since $\sqrt[n]{a} < 1$

$$|\sqrt[n]{a} - 1| = \left|\sqrt[n]{a}\frac{\sqrt[n]{a}-1}{\sqrt[n]{a}}\right| = |\sqrt[n]{a}|\left|\frac{\sqrt[n]{a}-1}{\sqrt[n]{a}}\right| < \left|\frac{\sqrt[n]{a}-1}{\sqrt[n]{a}}\right| = |1 - b_n| < \epsilon.$$

□

Example - We give another application of the Sandwich Rule in this example. Let $x, y > 0$ be fixed real numbers and consider the sequence $(b_n)_{n\in\mathbb{N}}$ where

$$b_n = \sqrt[n]{x^n + y^n}.$$

We will show that

$$\lim_{n\to\infty} b_n = \max\{x, y\}.$$

We may assume (w.l.o.g.) that $x \geq y$ and seek to show that $\lim_{n\to\infty} b_n = x$. Note that

$$x = \sqrt[n]{x^n} \leq \sqrt[n]{x^n + y^n} \leq \sqrt[n]{2x^n} = \sqrt[n]{2}\,\sqrt[n]{x^n} = \sqrt[n]{2} \cdot x.$$

Apply the Sandwich Rule with $a_n = x$ for all n and $c_n = \sqrt[n]{2} \cdot x$. We know from Lemma 3.15 that $\sqrt[n]{2} \to 1$, and it then follows from Theorem 3.8(ii) that $c_n \to x$. The Sandwich Rule then implies that

$$\lim_{n\to\infty} b_n = x.$$

Example - We can also use the Sandwich Rule to prove that

$$\lim_{n\to\infty} \frac{(1+\sin n)^n}{n2^n} = 0.$$

This is a good illustration of the use of the Sandwich Rule, since this sequence is rather complicated and it is not easy to observe a pattern by writing out the first few terms of the sequence.

Observe that

$$0 \leq \frac{(1+\sin n)^n}{n2^n} \leq \frac{(1+1)^n}{n2^n} = \frac{1}{n}.$$

Apply the Sandwich Rule with $a_n = 0$ and $c_n = \frac{1}{n}$. Both of these sequences converge to 0, and thus $\lim_{n\to\infty} \frac{(1+\sin n)^n}{n2^n} = 0$ also.

Exercise - Let $k \in \mathbb{N}$ be fixed. Use the Sandwich Rule to prove that

$$\lim_{n\to\infty} \frac{n!}{(n-k)!n^k} = 1.$$

We conclude this section by giving an analogue of Theorem 3.8 for definitely divergent sequences.

Theorem 3.16. *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be real-valued sequences and let $b, \lambda \in \mathbb{R}$. Then*

(i) $\lim_{n\to\infty} a_n = \infty$ *and* $\lim_{n\to\infty} b_n = \infty \implies \lim_{n\to\infty}(a_n + b_n) = \infty$,

(ii) $\lim_{n\to\infty} a_n = \infty$ *and* $\lim_{n\to\infty} b_n = \infty \implies \lim_{n\to\infty}(a_n b_n) = \infty$,

(iii) $\lim_{n\to\infty} a_n = \infty$ *and* $\lim_{n\to\infty} b_n = b \implies \lim_{n\to\infty}(a_n + b_n) = \infty$,

(iv) $\lim_{n\to\infty} a_n = \infty \implies \lim_{n\to\infty} \frac{\lambda}{a_n} = 0$,

(v) $\lim_{n\to\infty} a_n = \infty$ *and* $\lambda > 0 \implies \lim_{n\to\infty} \lambda a_n = \infty$,

(vi) $\lim_{n\to\infty} a_n = \infty$ *and* $\lambda < 0 \implies \lim_{n\to\infty} \lambda a_n = -\infty$.

Proof. This is left as an exercise. □

In general, one needs to take a little more care with these kinds of rules for definitely divergent sequences, and we do not always have a convenient rule to apply. For instance, there is no easy rule for determining the limit of the sequence $(a_n b_n)_{n\in\mathbb{N}}$ for the case when $a_n \to \infty$ and $b_n \to 0$. If we set $a_n = n$ and $b_n = 2^{-n}$ then Lemma 3.13 informs us that $a_n b_n \to 0$. However, if we set $a_n = n^2$ and $b_n = \frac{1}{n}$, then $a_n b_n = n \to \infty$.

3.3 Monotone sequences

Let's get straight to the key definition of this section.

Definition 3.17. *A real-valued sequence $(a_n)_{n\in\mathbb{N}}$ is called*

- ***increasing*** *if and only if*
$$\forall n \in \mathbb{N},\ a_{n+1} > a_n,$$
- ***non-decreasing*** *if and only if*
$$\forall n \in \mathbb{N},\ a_{n+1} \geq a_n,$$
- ***decreasing*** *if and only if*
$$\forall n \in \mathbb{N},\ a_{n+1} < a_n,$$
- ***non-increasing*** *if and only if*
$$\forall n \in \mathbb{N},\ a_{n+1} \leq a_n.$$

Moreover, we say that a sequence is ***monotone*** *if it is non-increasing or non-decreasing, and* ***strictly monotone*** *if it is either increasing or decreasing.*

Note that, since the definition of monotonicity requires a notion of order, we do not have an analogue of the definition above for complex-valued sequences.

Examples - Many of the sequences that we have discussed so far are monotone. For example $\left(\frac{1}{n}\right)_{n\in\mathbb{N}}$ is decreasing (and thus also strictly monotone). Any sequence of the form $(n^k)_{n\in\mathbb{N}}$ with $k > 0$ is increasing. If $k < 0$ then the sequence is decreasing, and if $k = 0$ then the sequence is constant (and so both non-increasing and non-decreasing). The sequence $((-1)^n)_{n\in\mathbb{N}}$ is not monotone.

For some sequences, a little more work is required to determine whether or not they are monotone. In some cases, a helpful trick is to consider the quotients of consecutive terms of a sequence and show that they are bounded (from above or below) by one. This works because a sequence is increasing (for example) if and only if
$$\frac{a_{n+1}}{a_n} > 1$$
holds for all $n \in \mathbb{N}$.

Lemma 3.18. *The sequence $(a_n)_{n\in\mathbb{N}}$ given by*
$$a_n := \left(1 + \frac{1}{n}\right)^n = \left(\frac{n+1}{n}\right)^n$$
is non-decreasing.

Proof. We consider quotients of successive terms. We will show

$$\frac{a_{n+1}}{a_n} \geq 1$$

holds for all $n \in \mathbb{N}$. Observe that

$$\begin{aligned}\frac{a_{n+1}}{a_n} = \frac{\left(\frac{n+2}{n+1}\right)^{n+1}}{\left(\frac{n+1}{n}\right)^n} &= \left(\frac{(n+2)n}{(n+1)^2}\right)^{n+1} \cdot \frac{n+1}{n} \\ &= \left(\frac{(n+1)^2-1}{(n+1)^2}\right)^{n+1} \cdot \frac{n+1}{n} \\ &= \left(1 - \frac{1}{(n+1)^2}\right)^{n+1} \cdot \frac{n+1}{n}.\end{aligned}$$

An application of Bernoulli's Inequality (Thm. 1.30) with $x = -\frac{1}{(n+1)^2} \geq -1$ yields

$$\frac{a_{n+1}}{a_n} = \left(1 - \frac{1}{(n+1)^2}\right)^{n+1} \cdot \frac{n+1}{n} \geq \left(1 - (n+1)\frac{1}{(n+1)^2}\right)\frac{n+1}{n} = 1.$$

□

In fact, if we were a little more careful in this proof, we could use a strict version of Bernoulli's Inequality that holds for $x > -1$ to prove that the given sequence is decreasing (and hence strictly monotone).

The following result, whose statement and proof is similar to Lemma 3.18, will be used later.

Lemma 3.19. *The sequence $(b_n)_{n\in\mathbb{N}}$ given by*

$$b_n := \left(1 + \frac{1}{n}\right)^{n+1} = \left(\frac{n+1}{n}\right)^{n+1}$$

is non-increasing.

Proof. This is left as an exercise. □

The following result shows that monotonicity is a very helpful property as we only need to check if a monotone sequence is bounded in order to know whether it is convergent or not. Note that boundedness of a sequence is usually much easier to show.

Theorem 3.20 (Monotonicity Principle). *(i) If $(a_n)_{n\in\mathbb{N}}$ is a non-decreasing sequence which is bounded above, then*

$$\lim_{n\to\infty} a_n = \sup\{a_n : n \in \mathbb{N}\} \in \mathbb{R}.$$

(ii) If $(a_n)_{n\in\mathbb{N}}$ is non-increasing sequence which is bounded below, then

$$\lim_{n\to\infty} a_n = \inf\{a_n : n \in \mathbb{N}\} \in \mathbb{R}.$$

(iii) If $(a_n)_{n\in\mathbb{N}}$ is a monotone sequence. Then

$$(a_n)_{n\in\mathbb{N}} \textit{ is convergent } \iff (a_n)_{n\in\mathbb{N}} \textit{ is bounded.} \tag{66}$$

We use the notation $\sup(a_n)$ as a shorthand for $\sup\{a_n : n \in \mathbb{N}\}$, and similarly $\inf(a_n) = \inf\{a_n : n \in \mathbb{N}\}$.

Proof. (i) Suppose that $(a_n)_{n\in\mathbb{N}}$ is non-decreasing sequence which is bounded above. This means that the set $\{a_n : n \in \mathbb{N}\}$ is bounded above. It follows from the Completeness Axiom (Axiom 1.38) that there exists $t \in \mathbb{R}$ such that

$$t = \sup\{a_n : n \in \mathbb{N}\}.$$

Now let $\epsilon > 0$ be arbitrary. It follows from the definition of the supremum that $t - \epsilon$ is not an upper bound for the set $\{a_n : n \in \mathbb{N}\}$. In particular, there exists $n_0 \in \mathbb{N}$ such that $a_{n_0} > t - \epsilon$.

However, since $(a_n)_{n\in\mathbb{N}}$ is non-decreasing, it follows that for all $n \geq n_0$

$$t - \epsilon < a_{n_0} \leq a_n \leq t < t + \epsilon.$$

Hence, $|a_n - t| < \epsilon$ for all $n \geq n_0$, and so by definition $\lim_{n\to\infty} a_n = t$.

(ii) The proof is very similar to part (i), and is left as an exercise.

(iii) We know from Theorem 3.6 that convergent sequences are bounded, which proves the first direction of the implication (66). In order to prove (66), we need to prove the reverse implication. This follows from the previous two parts of this theorem.

Indeed, suppose that $(a_n)_{n\in\mathbb{N}}$ is a bounded sequence which is monotone. By definition of monotonicity, the sequence is either non-decreasing or non-increasing. In the first of these cases, we can use part (i) of this theorem to show that $(a_n)_{n\in\mathbb{N}}$ is convergent and $\lim_{n\to\infty} a_n = t = \sup(a_n)$. Similarly, if $(a_n)_{n\in\mathbb{N}}$ is non-increasing then the result follows from part (ii).

□

Exercise - Let $(a_n)_{n\in\mathbb{N}}$ be a non-decreasing real-valued sequence which is unbounded. Prove that

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

With Theorem 3.20 we see that there are convergent sequences where we do not have to know the limit to verify that it exists. In some cases, we may even define numbers just as limits of specific sequences, because we do not have another (explicit) description. One typical example is **Euler's number**.

Lemma 3.21. *The sequence* $(a_n)_{n\in\mathbb{N}}$ *given by*

$$a_n := \left(1 + \frac{1}{n}\right)^n = \left(\frac{n+1}{n}\right)^n$$

is convergent.

Proof. We have already proven in Lemma 3.18 that this sequence is monotone. To prove this lemma, we will show that $(a_n)_{n\in\mathbb{N}}$ is bounded. We are then done, by Theorem 3.20.

Since $(a_n)_{n\in\mathbb{N}}$ is non-decreasing, it follows from Lemma 3.18 that, for all $n \in \mathbb{N}$,

$$a_n \geq a_1 = 2.$$

It remains to establish an upper bound for a_n. For this, we recall the related sequence $(b_n)_{n\in\mathbb{N}}$, given by

$$b_n := \left(1 + \frac{1}{n}\right)^{n+1} = \left(\frac{n+1}{n}\right)^{n+1}.$$

Observe that, for all $n \in \mathbb{N}$, $a_n \leq b_n$. We also know, from Lemma 45, that $(b_n)_{n\in\mathbb{N}}$ is non-increasing, which implies that, for all $n \in \mathbb{N}$,

$$a_n \leq b_n \leq b_1 = 4.$$

We have found both an upper and lower bound for a_n, which means that the sequence is indeed bounded and the proof is complete.

□

If we take a little more care with the application of Theorem 3.20, we see that this shows that the limit of the sequence given by

$$a_n := \left(1 + \frac{1}{n}\right)^{n} = \left(\frac{n+1}{n}\right)^{n}$$

exists and equals $\sup\{a_n : n \in \mathbb{N}\}$. We define this limit to be *Euler's number*, denoted e. That is,

$$e := \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^{n} = \sup \left(1 + \frac{1}{n}\right)^{n}.$$

3.4 Subsequences

The concepts of the last sections deal with sequences that converge or, in other words, concentrate around a single point. In some cases, however, divergent sequences may also have some points of interest for very large n. An obvious example is $((-1)^n)_{n\in\mathbb{N}}$, which appears to converge towards two different points. Now, we want to formalise the idea of sequences having more than one limit.

Definition 3.22. *Let $(n_1, n_2, n_3, \dots)$ be an (infinite) increasing sequence of natural numbers and let $(a_n)_{n\in\mathbb{N}}$ be a sequence. Then, we call*

$$(a_{n_k})_{k\in\mathbb{N}} = (a_{n_1}, a_{n_2}, \dots)$$

*a **subsequence** of $(a_n)_{n\in\mathbb{N}}$.*

Example - Consider the sequence $(a_n)_{n\in\mathbb{N}}$ given by

$$a_n = (-1)^n.$$

Two notable subsequences are given by taking the odd and even terms of the sequences. That is, we can consider $(n_1, n_2, \dots) = (1, 3, \dots)$ and $(n_1, n_2, \dots) = (2, 4, \dots)$. These subsequences are convergent (in fact, they are constant) with limit -1 and 1 respectively.

Exercise - Suppose that $(a_n)_{n\in\mathbb{N}}$ is a sequence such that $\lim_{n\to} a_n = a$. Show that any subsequence $(a_{n_k})_{k\in\mathbb{N}}$ satisfies

$$\lim_{k\to\infty} a_{n_k} = a.$$

Definition 3.23. *Let $(a_n)_{n\in\mathbb{N}}$ be a complex-valued sequence. We call $a \in \mathbb{C}$ an **accumulation point** of $(a_n)_{n\in\mathbb{N}}$ if there exists a subsequence $(a_{n_k})_{k\in\mathbb{N}}$ with*

$$\lim_{k\to\infty} a_{n_k} = a.$$

The accumulation points of the sequence $((-1)^n)_{n\in\mathbb{N}}$ are -1 and 1. We can also consider some more complicated examples.

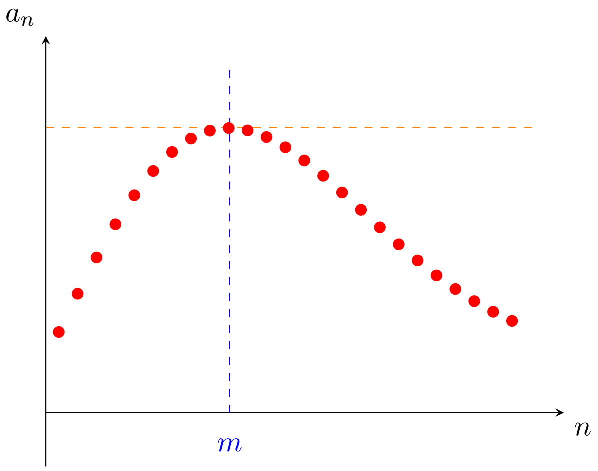
Example - Consider the sequence defined by

$$a_n = \begin{cases} 1 & \text{if } n \text{ is not prime} \\ \frac{1}{n} & \text{if } n \text{ is prime} \end{cases}.$$

The accumulation points of this sequence are 0 and 1. The subsequence $(a_n)_{n\in\mathbb{P}}$, where $\mathbb{P}$ denotes the set of all primes, converges to 0 and the subsequence $(a_n)_{n\notin\mathbb{P}}$ is constant, always taking value 1.

Next we want to show that each bounded sequence has at least one convergent subsequence. This result bears the names of Bolzano and Weierstrass and is an important technical tool for proofs in many areas of analysis.

Theorem 3.24 (Bolzano-Weierstrass Theorem)**.** *Let $(a_n)_{n\in\mathbb{N}}$ be a bounded real-valued sequence. Then $(a_n)_{n\in\mathbb{N}}$ has at least one convergent subsequence.*



Proof. We call m a *peak* of $(a_n)_{n\in\mathbb{N}}$ if

$$a_n < a_m, \; \forall \; n > m.$$

Case 1 - Suppose that there are finitely many peaks $m_1 < m_2 < \cdots < m_l$. Set $n_1 = m_l + 1$. In particular, n_1 is not a peak, and so there exists an integer $n_2 > n_1$ such that $a_{n_2} \geq a_{n_1}$. Also, n_2 is not a peak, and so there exists $n_3 > n_2$ such that $a_{n_3} \geq a_{n_2} \geq a_{n_1}$. We can continue this process to obtain a (infinite) non-decreasing sequence $(a_{n_1}, a_{n_2}, \dots)$ such that $a_{n_1} \leq a_{n_2} \leq a_{n_3} \leq a_{n_4} \dots$. This subsequence is also bounded, because of the assumption that $(a_n)_{n\in\mathbb{N}}$ is bounded. It therefore follows from the Monotonicity Principle (Theorem 3.20) that it is convergent.

Case 2 - Suppose that there are no peaks. The proof is the same as that of Case 1, expect that we set $n_1 = 1$.

Case 3 - Suppose that there are infinitely many peaks $m_1 < m_2 < \dots$. Then the sequence $(a_{m_1}, a_{m_2}, \dots)$ is decreasing. It is also bounded, because of the assumption that $(a_n)_{n\in\mathbb{N}}$ is bounded. It therefore follows again from the Monotonicity Principle (Theorem 3.20) that this subsequence is convergent. □

Example - We can use the Bolzano-Weierstrass Theorem to show that certain sequences contain convergent subquences even in cases when the convergent subsequences are rather difficult to see. For instance, consider the sequence $(a_n)_{n\in\mathbb{N}}$ given by

$$a_n = \frac{n \cdot \cos(3n^2 - 5)}{n+1}.$$

It is not easy to see a pattern in this sequence, with the values of a_n jumping around fairly randomly, somewhere in the range $(-1, 1)$. However, it is not difficult to check that the sequence $(a_n)_{n\in\mathbb{N}}$ is bounded, and so the Bolzano-Weierstrass Theorem tells us that there exists a convergent subsequence.

Note that the converse of the Bolzano-Weierstrass Theorem does not hold, i.e. not every sequence with a convergent subsequence is bounded. One may consider, for example, the sequence $(a_n)_{n\in\mathbb{N}}$ given by

$$a_n = \begin{cases} n & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases}.$$

We finish this section with an extreme example, highlighting the potentially strange behaviour of accumulation points.

Example Consider the sequence $(a_n)_{n\in\mathbb{N}}$, which is a list of all rational numbers in the interval $(0,1)$. One may define this sequence more formally by constructing a bijection between $f : \mathbb{N} \to \mathbb{Q} \cap (0,1)$ (we did this back in Chapter 1) and then setting $a_n = f(n)$.

For every real number $x \in (0,1)$ it is possible to define an infinite sequence of rational numbers which converge to x. Such a sequence can be constructed using part 3 of Theorem 1.39 (and I encourage you to formally define such a sequence).

In other words, the set of accumulation points for this sequence is the whole interval $(0,1)$. This is an uncountable and therefore somewhat "larger" than the actual range of the sequence!

3.5 Cauchy criterion

In this section we introduce *Cauchy sequences* and the *Cauchy criterion* for establishing the convergence of a sequence. The Cauchy criterion is, similarly to the Monotonicity Principle (Theorem 3.20), an important tool to verify that a sequence is convergent without knowing its limit and it will be used several times throughout the remainder of the course.

Definition 3.25. *A complex-valued sequence $(a_n)_{n\in\mathbb{N}}$ is called a **Cauchy sequence** if*

$$\forall \epsilon > 0,\ \exists n_0 \in \mathbb{N} : m, n \geq n_0 \implies |a_n - a_m| < \epsilon.$$

You should compare this definition with the definition of convergence in order to gain better understanding.

We now take a look at two familiar sequences and examine whether or not they are Cauchy.

Example - The sequence $(a_n)_{n\in\mathbb{N}}$ given by

$$a_n = \frac{1}{n}$$

is a Cauchy sequence. To see this, observe that, by the triangle inequality,

$$|a_n - a_m| = \left|\frac{1}{n} - \frac{1}{m}\right| \leq \left|\frac{1}{n}\right| + \left|\frac{1}{m}\right| = \frac{1}{n} + \frac{1}{m}.$$

It therefore follows that, for all $\epsilon > 0$ and for all $m, n > \frac{2}{\epsilon}$, we have $|a_n - a_m| < \epsilon$.

Example - The sequence $(a_n)_{n\in\mathbb{N}}$ given by

$$a_n = (-1)^n$$

is a not Cauchy sequence. Suppose for a contradiction that it is Cauchy. Then, for all $\epsilon > 0$, there is some $n_0 \in \mathbb{N}$ such that $|a_n - a_m| < \epsilon$ holds for all $m, n \geq n_0$. But, if we set $\epsilon = 1$, we obtain a contradiction, since no matter which value we choose for n_0, we have

$$|a_{n_0} - a_{n_0+1}| = 2 > 1.$$

These two examples suggest that the property of being Cauchy may be similar to that of being convergent (since we already know that the first sequence $(1/n)$ is convergent, and that $((-1)^n)$ is not). Indeed, this intuition is correct, as the following important result shows.

Theorem 3.26 (Cauchy criterion). *Let $(a_n)_{n\in\mathbb{N}}$ be a real-valued sequence. Then*

$$(a_n)_{n\in\mathbb{N}} \textit{ is convergent} \iff (a_n)_{n\in\mathbb{N}} \textit{ is Cauchy.}$$

Before proving the Cauchy criterion, we prove a lemma that will be used in the proof.

Lemma 3.27. *Let $(a_n)_{n\in\mathbb{N}}$ be a complex-valued sequence which is Cauchy. Then $(a_n)_{n\in\mathbb{N}}$ is bounded.*

Proof. Apply the definition of a Cauchy sequence with $\epsilon = 1$. It follows that there is some $n_0 \in \mathbb{N}$ such that, for all $m, n \geq n_0$,

$$|a_n - a_m| < 1.$$

Also, by the triangle inequality,

$$|a_n| = |(a_n - a_{n_0}) + a_{n_0}| \leq |a_n - a_{n_0}| + |a_{n_0}| < 1 + |a_{n_0}|.$$

This gives a bound for all $n \geq n_0$. It then follows that, for all $n \in \mathbb{N}$, $|a_n| \leq C$, where

$$C = \max\{|a_1|, |a_2|, \ldots, |a_{n_0-1}|, 1 + |a_{n_0}|\}.$$

□

Proof of the Cauchy Criterion. First, we show that

$$(a_n)_{n\in\mathbb{N}} \text{ is convergent } \implies (a_n)_{n\in\mathbb{N}} \text{ is Cauchy.}$$

Suppose that $(a_n)_{n\in\mathbb{N}}$ is convergent with $a_n \to a$. Let $\epsilon > 0$ be arbitrary. By the definition of convergence, there exists $n_0 \in \mathbb{N}$ such that for all $m, n \geq n_0$,

$$|a_m - a|, |a_n - a| < \frac{\epsilon}{2}.$$

It follows from the triangle inequality that, for all $m, n \geq n_0$,

$$|a_m - a_n| = |(a_m - a) + (a - a_n)| \leq |a_m - a| + |a - a_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Next, we consider the opposite implication. Suppose that $(a_n)_{n\in\mathbb{N}}$ is Cauchy. Lemma 3.27 implies that $(a_n)_{n\in\mathbb{N}}$ is bounded. By the Bolzano-Weierstrass Theorem (Thm. 3.24), $(a_n)_{n\in\mathbb{N}}$ has at least one convergent subsequence.

Let $(a_{n_k})_{n\in\mathbb{N}}$ be a subsequence of $(a_n)_{n\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} a_{n_k} = a.$$

Let $\epsilon > 0$ be arbitrary. Since $(a_{n_k})_{n\in\mathbb{N}}$ tends to a, it follows from the definition of convergence that there is some $k_0 \in \mathbb{N}$ such that, for all $k \geq k_0$,

$$|a_{n_k} - a| < \frac{\epsilon}{2}.$$

Also, since $(a_n)_{n\in\mathbb{N}}$ is Cauchy, it follows that there is some n_0 such that, for all $m, n \geq n_0$,

$$|a_n - a_m| < \frac{\epsilon}{2}.$$

Let $k \in \mathbb{N}$ be any integer such that both $k \geq k_0$ and $n_k \geq n_0$ hold. Then, by the triangle inequality, we have that for all $n \geq n_0$

$$|a_n - a| = |(a_n - a_{n_k}) + (a_{n_k} - a)| \leq |a_n - a_{n_k}| + |a_{n_k} - a| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

□

The Cauchy criterion is particularly useful as a shortcut for proving that certain sequences are convergent, since it is generally easier to verify that a sequence is Cauchy than it is to show that it is convergent. In particular, we do not need to know what the limit of the sequence is when checking that it is Cauchy.

Example - Let $(a_n)_{n\in\mathbb{N}}$ be a recursively defined sequence with $a_1 = 1$ and

$$a_{n+1} = \begin{cases} a_n + \frac{1}{2^n} & \text{if } n \text{ is prime} \\ a_n - \frac{1}{2^n} & \text{if } n \text{ is not prime} \end{cases}.$$

If we write down the first few terms of this sequence, it is not immediately obvious that the sequence is convergent, and it is even more tricky to identify a possible limit. However, without knowing anything about the possible limit of the sequence, we can show that is it Cauchy, as follows.

First, observe that, for all $n \geq 2$,

$$|a_{n+1} - a_n| = \frac{1}{2^n}.$$

Now let $\epsilon > 0$ and suppose that m and n are both larger than n_0. We need make sure that n_0 is sufficiently large for the forthcoming proof, and we will specify the choice of n_0 to make the argument work.

Without loss of generality, we may assume that $m > n \geq n_0$. By the triangle inequality,

$$\begin{aligned} |a_m - a_n| &= |(a_m - a_{m-1}) + (a_{m-1} - a_{m-2}) + \cdots + (a_{n+1} - a_n)| \\ &\leq |a_m - a_{m-1}| + |a_{m-1} - a_{m-2}| + \cdots + |a_{n+1} - a_n| \\ &\leq \frac{1}{2^{m-1}} + \frac{1}{2^{m-2}} + \cdots + \frac{1}{2^n} \\ &= \frac{1}{2^{n-1}} \sum_{k=1}^{m-n} \left(\frac{1}{2}\right)^k \\ &< \frac{1}{2^{n-1}} \leq \frac{1}{2^{n_0-1}}. \end{aligned}$$

In particular, if we set n_0 so that $2^{n_0-1} > \frac{1}{\epsilon}$, it follows that, for all $m, n \geq n_0$,

$$|a_m - a_n| < \epsilon.$$

Formally, we may define $n_0 := \lceil \log_2 \left(\frac{2}{\epsilon}\right) \rceil$.

In summary, we have shown that $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, and thus by the Cauchy criterion, it follows that the sequence is indeed convergent.

3.6 Introduction to series

In this section, we use sequences to build series. A series is really just a special kind of sequence $(s_n)_{n\in\mathbb{N}}$ given by

$$s_n = \sum_{k=1}^{n} a_k,$$

where $(a_n)_{n\in\mathbb{N}}$ is another sequence. The sum of all terms of the sequence $(a_n)_{n\in\mathbb{N}}$, i.e. the limit of the sequence $(s_n)_{n\in\mathbb{N}}$, is one of the main motivations for considering limits at all, and some interesting phenomena appear when it comes to the question of whether such limits exist.

We begin with a more formal definition.

Definition 3.28. *Let $(a_n)_{n\in\mathbb{N}}$ be a complex-valued sequence and*

$$s_n = \sum_{k=1}^{n} a_k.$$

We call s_n the ***nth partial sum*** *of the* ***series***

$$\sum_{k=1}^{\infty} a_k.$$

If the sequence $(s_n)_{n\in\mathbb{N}}$ converges with $\lim_{n\to\infty} s_n = s$ then we say that the series ***converges****. We call s the* ***sum*** *of the series, and write*

$$\sum_{k=1}^{\infty} a_k = s.$$

If a series is not convergent, then it is ***divergent****.*

If $\lim_{n\to\infty} s_n = \pm\infty$ then we write

$$\sum_{k=1}^{\infty} a_k = \pm\infty$$

and say that the series is ***definitely divergent****.*

Note that "**series**" is just another word for an infinite sum of elements of a sequence. Moreover, the notation $\sum_{k=1}^{\infty} a_k$ should be understood as a formal symbol for the limit of the corresponding sequence $(s_n)_{n\in\mathbb{N}}$: it might be a number or $\pm\infty$, but it might also not exist.

We will also sometimes consider series which do not start with index 1, i.e. series of the form $\sum_{k=k_0}^{\infty} a_k$ for some $k_0 \in \mathbb{Z}$. The most common variant we consider is with $k_0 = 0$.

It should also be noted that, if we consider an arbitrary series $\sum_{k=1}^{\infty} a_k$, we should assume that the terms a_k are complex numbers unless stated otherwise. There will be some instances later where we make the additional restriction that $a_k \in \mathbb{R}$.

The definition above states that a series converges if and only if the sequence $(s_n)_{n\in\mathbb{N}}$ of partial sums converges. This implies that we can use the results from the previous sections to analyse series. Moreover, we will see that there are even more tools for working with series. But first let us consider some particularly important examples.

Lemma 3.29 (Geometric series)**.** *Let $q \in \mathbb{C}$ with $|q| < 1$. Then we have that*

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}, \quad \textit{and} \quad \sum_{k=1}^{\infty} q^k = \frac{q}{1-q}. \tag{67}$$

Moreover, we have

$$\sum_{k=0}^{n} q^k = \frac{1-q^{n+1}}{1-q}. \tag{68}$$

Proof. We will first prove (68), and then use it to prove (67). Let $s_n := \sum_{k=0}^{n} q^k$ and consider the equation

$$\begin{aligned}(1-q)s_n &= (1-q)\sum_{k=0}^{n} q^k \\ &= (1-q)(1+q+q^2+\cdots+q^n) \\ &= (1+q+q^2+\cdots+q^n) - (q+q^2+q^3+\cdots+q^{n+1}) \\ &= 1-q^{n+1}.\end{aligned}$$

In the last step, we have simply observed that all but the extreme terms in the two brackets cancel out with one another. This proves (68).

To prove the first part of (67), we make use of some facts about convergence of sequences that we established earlier in this chapter (namely Theorem 3.8 and Lemma 3.10) to see that

$$\sum_{k=0}^{\infty} q^k = \lim_{n\to\infty} s_n = \lim_{n\to\infty} \frac{1-q^{n+1}}{1-q} = \frac{1}{1-q} \cdot \lim_{n\to\infty}(1-q^{n+1}) = \frac{1}{1-q}.$$

The second sum from (67) follows immediately from the first. □

In the proof above, we considered two long sums and observed that almost all of the terms cancelled out. Such arguments are called **telescoping tricks** and sums of this form are called **telescoping sums**. This kind of trick will be used more throughout this chapter.

Example - If we set $q = \frac{1}{2}$ in Lemma 3.29, it follows that

$$\sum_{k=0}^{\infty} \frac{1}{2^k} = 2, \quad \text{and} \quad \sum_{k=1}^{\infty} \frac{1}{2^k} = 1.$$

The next example shows that **not all convergent sequences give rise to convergent series**. This is a very important example of a divergent series.

Lemma 3.30 (Harmonic series)**.** *Consider the sequence $(a_n)_{n\in\mathbb{N}}$ given by $a_n = \frac{1}{n}$. Then, the corresponding series satisfies*

$$\sum_{k=1}^{\infty} \frac{1}{k} = \infty.$$

Proof. We need to show that the sequence of partial sums $s_n = \sum_{k=1}^n \frac{1}{k}$ tends to infinity. We group the terms of the partial sum s_{2^n} and manipulate the sums as follows:

$$\begin{aligned} s_{2^n} &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \cdots + \left(\frac{1}{2^{n-1}+1} + \frac{1}{2^{n-1}+2} + \cdots + \frac{1}{2^n}\right) \\ &\geq 1 + \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + \cdots + 2^{n-1} \cdot \frac{1}{2^n} \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} \cdots + \frac{1}{2} \\ &= 1 + \frac{n}{2}. \end{aligned}$$

Since s_n is an increasing sequence, it follows that, for all $C \in \mathbb{R}$, there exists $n_0 = 2^{\lceil 2C \rceil}$ such that, for all $n \geq n_0$ we have

$$s_n \geq s_{n_0} = s_{2^{\lceil 2C \rceil}} \geq 1 + \frac{\lceil 2C \rceil}{2} > C.$$

Recalling the definition of the improper limit, we see that $s_n \to \infty$. □

The series $\sum_{k=1}^{\infty} \frac{1}{n} = \infty$ is called the **harmonic series**. By contrast with Lemma 3.30, the series $\sum_{k=1}^{\infty} n^{-\alpha}$ converges for any $\alpha > 1$, and so the harmonic series is something of a critical example at which there is a change of behaviour.

Example - In this example, we discuss how the aforementioned telescoping trick can sometimes be a powerful tool for obtaining the precise value of apparently complicated series. We will prove that

$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1.$$

We do not even know that this series is convergent yet. However, we first make the helpful observation that

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}. \tag{69}$$

It therefore follows that

$$s_n = \sum_{k=1}^{n} \frac{1}{k(k+1)} = \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k+1}\right) = \sum_{k=1}^{n} \frac{1}{k} - \sum_{k=2}^{n+1} \frac{1}{k} = 1 - \frac{1}{n+1}.$$

It therefore follows that $\lim_{n\to\infty} s_n = 1$, as required.

In the example above, and in particular in the identity (69), we have used the method of *partial fraction decomposition* to write a fraction with a product in the denominator as a sum of two fractions with more simple denominators.

However, note that it is rare that we can easily compute the sum of a series precisely. Already for the very similar example

$$\sum_{k=1}^{\infty} \frac{1}{k^2}$$

the problem of determining the value of the sum is considerably more difficult. We will use more sophisticated mathematics later in this program to prove that $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. For many other sums, there is just no closed expression.

3.7 Calculation rules and basic properties of series

In this section, we will use some of the theory that we have built for sequences to derive some basic properties about series. We begin with an analogue of Theorem 3.8.

Theorem 3.31. *Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be convergent series and let $c \in \mathbb{C}$. Then we have*

$$\sum_{k=1}^{\infty}(a_k + b_k) = \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k$$

and

$$\sum_{k=1}^{\infty} c \cdot a_k = c \cdot \sum_{k=1}^{\infty} a_k.$$

Proof. This is left as an exercise. □

Example - Recall from the previous section that

$$\sum_{k=1}^{\infty} 2^{-k} = 1 \quad \text{and} \quad \sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1.$$

It therefore follows from Theorem 3.31 that

$$\sum_{k=1}^{\infty} \frac{\pi k(k+1) + 2^k}{k(k+1)2^k} = \sum_{k=1}^{\infty}\left(\frac{\pi}{2^k} + \frac{1}{k(k+1)}\right) = \pi \sum_{k=1}^{\infty} 2^{-k} + \sum_{k=1}^{\infty} \frac{1}{k(k+1)} = \pi + 1.$$

The next result gives a useful application of the Monotonicity Principle (Theorem 3.20).

Theorem 3.32. *Let $(a_n)_{n\in\mathbb{N}}$ be a real-valued sequence such that $a_n \geq 0$ for all $n \in \mathbb{N}$. Let $s_n = \sum_{k=1}^{n} a_k$. Then*

$$(s_n)_{n\in\mathbb{N}} \text{ is bounded} \iff \sum_{k=1}^{\infty} a_k \text{ converges.}$$

Proof. Since $a_k \geq 0$ for all $k \in \mathbb{N}$, it follows that the sequence $(s_n)_{n\in\mathbb{N}}$ is non-decreasing. In particular, this sequence is monotone. It therefore follows from Theorem 3.20(iii) that

$$(s_n)_{n\in\mathbb{N}} \text{ is bounded} \iff (s_n)_{n\in\mathbb{N}} \text{ is convergent .}$$

This proves the theorem. □

Example - We can use the previous result to show that the series $\sum_{k=1}^{\infty} \frac{1}{k!}$ converges. We just need to show that the sequence $s_n = \sum_{k=1}^{n} \frac{1}{k!}$ is bounded. To see this, first observe that $k! \geq 2^{k-1}$ holds for all $k \in \mathbb{N}$. From this and from Lemma 3.29 it follows then that

$$\sum_{k=1}^{n} \frac{1}{k!} \leq \sum_{k=1}^{n} \frac{1}{2^{k-1}} = \sum_{k=0}^{n-1} \frac{1}{2^k} \leq \sum_{k=0}^{\infty} \frac{1}{2^k} = 2.$$

Since $\sum_{k=1}^{\infty} \frac{1}{k!}$ converges, it immediately follows that

$$\sum_{k=0}^{\infty} \frac{1}{k!} = 1 + \sum_{k=1}^{\infty} \frac{1}{k!}$$

and so $\sum_{k=0}^{\infty} \frac{1}{k!}$ also converges. Moreover, one can indeed show that

$$\sum_{k=0}^{\infty} \frac{1}{k!} = e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n . \tag{70}$$

We omit the proof of (70) here, but it will be considered on a forthcoming exercise sheet.

Next, we will apply the Cauchy criterion to the sequence of partial sums to obtain the following result.

Theorem 3.33. *Let $\sum_{k=1}^{\infty} a_k$ be a series. Then $\sum_{k=1}^{\infty} a_k$ is convergent if and only if*

$$\forall \epsilon > 0,\ \exists\, n_0 \in \mathbb{N} \,:\, \forall m > n > n_0,\ \left| \sum_{k=n+1}^{m} a_k \right| < \epsilon. \tag{71}$$

Proof. By definition, the series $\sum_{k=1}^{\infty} a_k$ is convergent if and only if the sequence of partial sums $s_n = \sum_{k=1}^{n} a_k$ converges. By the Cauchy criterion, this sequence converges if and only if it is Cauchy.

But what does it mean for the sequence of partial sums to be Cauchy? By definition, this means that

$$\forall \epsilon > 0,\ \exists\, n_0 \in \mathbb{N} \,:\, \forall m > n > n_0,\ |s_m - s_n| < \epsilon. \tag{72}$$

Observe that $s_m - s_n = \sum_{k=1}^{m} a_k - \sum_{k=1}^{n} a_k = \sum_{n+1}^{m} a_k$. Therefore, the statement (72) is equivalent to the statement (71), which completes the proof. □

This theorem immediately leads to the following simple criterion. In many cases, this is already enough to show that a series is *divergent*.

Corollary 3.34. *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence and suppose that the series $\sum_{k=1}^{\infty} a_k$ converges. Then*

$$\lim_{n\to\infty} a_n = 0.$$

In other words, a series can only be convergent if the corresponding sequence is a null sequence.

Proof. Suppose that the series $\sum_{k=1}^{\infty} a_k$ converges. Then it follows from Theorem 3.33 (setting $m = n + 1$ in (71)) that

$$\forall \epsilon > 0,\ \exists\, n_0 \in \mathbb{N} \,:\, \forall n > n_0,\ |a_{n+1}| < \epsilon.$$

This implies that $\lim_{n\to a_n} = 0$. □

Example - We can immediately use the previous corollary to see that, for any divergent sequence, or any convergent sequence which is not a null sequence, the corresponding series is not convergent. In particular, the series

$$\sum_{k=1}^{\infty}(-1)^k$$

is divergent. Also, for any $m \in \mathbb{N}$,

$$\sum_{k=1}^{\infty}\sqrt[k]{k^m}$$

is divergent.

An important remark is that the converse of Corollary 3.34 does not hold. For instance, as we have shown in Lemma 3.30, the harmonic series is definitely divergent with

$$\sum_{k=1}^{\infty}\frac{1}{n}=\infty,$$

although the sequence of its summands $(1/n)_{n\in\mathbb{N}}$ is a null sequence.

It is sometimes helpful to consider the following stronger form of convergence.

Definition 3.35. *Let $(a_n)_{n\in\mathbb{N}}$ be a complex-valued sequence with the property that there exists $C \in \mathbb{R}$ such that*

$$\sum_{k=1}^{n}|a_k| \leq C, \quad \forall n \in \mathbb{N}.$$

Then we say that the series $\sum_{k=1}^{\infty} a_k$ is ***absolutely convergent****.*

Note that, by Theorem 3.32, the series $\sum_{k=1}^{\infty} a_k < \infty$ is absolutely convergent if and only if the series $\sum_{k=1}^{\infty}|a_k|$ is convergent (which we could have used as an equivalent definition). It is therefore perfectly reasonable to write $\sum_{k=1}^{\infty}|a_k| < \infty$ as a shorthand for the series $\sum_{k=1}^{\infty} a_k$ being absolutely convergent.

We have already shown in Lemma 3.29 that, for $q \in \mathbb{C}$ with $|q| < 1$, the geometric series $\sum_{k=1}^{\infty} q^k$ is convergent. We will now show that it is also absolutely convergent.

Lemma 3.36. *The geometric series $\sum_{k=1}^{\infty} q^k$, with $|q| < 1$, is absolutely convergent.*

Proof. We have

$$\sum_{k=1}^{\infty}|q^k| = \sum_{k=1}^{\infty}|q|^k = \frac{|q|}{1-|q|},$$

where we have just applied Lemma 3.29 again, with $|q|$ in the role of q.

□

Example - The series

$$\sum_{k=1}^{\infty}\frac{(-1)^k}{k}$$

is called the **alternating harmonic series**. This series is *not* absolutely convergent, since

$$\sum_{k=1}^{\infty} \left| \frac{(-1)^k}{k} \right| = \sum_{k=1}^{\infty} \frac{1}{k}$$

is the harmonic series, which is divergent. However, it turns out that the alternating harmonic series is convergent.

The next result shows that absolute convergence is indeed a stronger condition than “mere” convergence.

Theorem 3.37. *If a series $\sum_{k=1}^{\infty} a_k$ is absolutely convergent then it is also convergent.*

Proof. Let $\sum_{k=1}^{\infty} a^k$ be an absolutely convergent series and let $\epsilon > 0$ be arbitrary. By Theorem 3.33 (applied to the series $\sum_{k=1}^{\infty} |a_k|$) and the triangle inequality, there exists $n_0 \in \mathbb{N}$ such that, for all $m > n \geq n_0$,

$$\left| \sum_{k=n+1}^{m} a_k \right| \leq \sum_{k=n+1}^{m} |a_k| = \left| \sum_{k=n+1}^{m} |a_k| \right| < \epsilon.$$

It then follows from a second application of Theorem 3.33 that $\sum_{k=1}^{\infty} a_k$ is convergent.

□

3.8 Convergence tests

We will now discuss several criteria, called *convergence tests*, that can be used to verify if a series is convergent or not. However, note that these test are sometimes inconclusive, i.e., we do not always get a definite answer by applying them, and one may need to apply other techniques.

3.8.1 Comparison Test

The first result shows that we can obtain information about the convergence of a series by comparing it to a series that is known to be convergent (or not).

Theorem 3.38. *Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be two series.*

1. *If $\sum_{k=1}^{\infty} b_k$ is absolutely convergent and $|a_k| \leq |b_k|$ holds for all but finitely many $k \in \mathbb{N}$, then $\sum_{k=1}^{\infty} a_k$ is also absolutely convergent.*
2. *If $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are real-valued sequences such that $0 \leq b_k \leq a_k$ holds for all $k \in \mathbb{N}$, and $\sum_{k=1}^{\infty} b_k = \infty$, then $\sum_{k=1}^{\infty} a_k = \infty$.*

Proof. 1. It follows from the hypothesis that there exists $k_0 \in \mathbb{N}$ such that,

$$|a_k| \leq |b_k|, \quad \forall\, k \geq k_0.$$

Since $\sum_{k=1}^{\infty} b_k$ converges absolutely, it follows that the sequence of partial sums $s_n = \sum_{k=1}^{n} |a_k|$ is bounded. Indeed, for any $n \geq k_0$, we have

$$\begin{aligned}\sum_{k=1}^{n} |a_k| &= \sum_{k=1}^{k_0-1} |a_k| + \sum_{k=k_0}^{n} |a_k| \\ &\leq \sum_{k=1}^{k_0-1} |a_k| + \sum_{k=k_0}^{n} |b_k| \\ &\leq \sum_{k=1}^{k_0-1} |a_k| + \sum_{k=k_0}^{\infty} |b_k| \in \mathbb{R}.\end{aligned}$$

It then follows from Theorem 3.32 that $\sum_{k=1}^{\infty} |a_k|$ converges.

2. Let

$$s_n = \sum_{k=1}^{n} a_n, \quad t_n = \sum_{k=1}^{n} b_n$$

denote the sequences of partial sums of the two series. Since $b_n \geq 0$, it follows from Theorem 3.32 that the sequence (t_n) is not bounded. But also, $s_n \geq t_n$ for all $n \in \mathbb{N}$, and so it must also be the case that the sequence (s_n) is not bounded. It then follows from the exercise after the proof of Theorem 3.20 that $s_n \to \infty$.

□

Example - We will now use Theorem 3.38 to prove that the series $\sum_{k=1}^{\infty} \frac{1}{k^2}$ is (absolutely) convergent. For any $k \in \mathbb{N}$, we have $k+1 \leq 2k$ and thus

$$\frac{1}{k^2} = \frac{k+1}{k} \cdot \frac{1}{k(k+1)} \leq 2 \cdot \frac{1}{k(k+1)}.$$

Since both sides of this inequality are non-negative, it follows that

$$\left| \frac{1}{k^2} \right| \leq \left| \frac{2}{k(k+1)} \right|.$$

Note that the series $\sum_{k=1}^{\infty} \frac{2}{k(k+1)}$ is absolutely convergent. Indeed, we showed earlier (see the example after the proof of Lemma 3.30) that the sequence $\sum_{k=1}^{\infty} \frac{1}{k(k+1)}$ converges, and it then follows from Theorem 3.31 that $\sum_{k=1}^{\infty} \frac{2}{k(k+1)}$ is also convergent, and moreover

$$\sum_{k=1}^{\infty} \frac{2}{k(k+1)} = 2 \sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 2.$$

Since the corresponding sequence consists of non-negative real numbers, it follows that $\sum_{k=1}^{\infty} \frac{2}{k(k+1)}$ is also absolutely convergent.

Apply the first part of Theorem 3.38 with $a_k = \frac{1}{k^2}$ and $b_k = \frac{2}{k(k+1)}$. It follows that the series $\sum_{k=1}^{\infty} \frac{1}{k^2}$ is absolutely convergent.

Exercise - Use Theorem 3.38 to prove that

- $\sum_{k=1}^{\infty} \frac{1}{k^c}$ is absolutely convergent for all $c \geq 2$, and
- $\sum_{k=1}^{\infty} \frac{1}{k^c} = \infty$ for all $c \leq 1$.

3.8.2 Root Test

Next, we present the Root Test.

Theorem 3.39. *Let $\sum_{k=1}^{\infty} a_k$ be a series.*

1. *If there exists a real number $c < 1$ such that*

$$\sqrt[k]{|a_k|} \leq c$$

holds for all but finitely many $k \in \mathbb{N}$, then $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

2. *Conversely, if*

$$\sqrt[k]{|a_k|} \geq 1$$

holds for infinitely many $k \in \mathbb{N}$, then $\sum_{k=1}^{\infty} a_k$ is divergent.

Proof. 1. The series $\sum_{k=1}^{\infty} c^k$ is absolutely convergent (see Lemma 3.36). Also, by the hypothesis of the theorem $|a_k| \leq c^k = |c^k|$ holds for all but finitely many $k \in \mathbb{N}$. Therefore, by part 1 of Theorem 3.38, $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

2. It follows from the condition that $|a_k| \geq 1$ for infinitely many $k \in \mathbb{N}$. In particular, the sequence $(a_k)_{k\in\mathbb{N}}$ is not a null sequence. It then follows from Corollary 3.34 that the series $\sum_{k=1}^{\infty} a_k$ does not converge.

□

The root test is usually most helpful when a the kth term of the corresponding sequence involves a kth power.

Example - Consider the series

$$\sum_{k=1}^{\infty} \sin(k)\frac{k^{100}}{2^{k/2}}.$$

This looks like a complicated series, and if we try to calculate the first few terms of the corresponding sequence, they are rather large! However, we can use the comparison test and the root test to prove that the series is convergent.

Note that, for all $k \in \mathbb{N}$,

$$\left|\sin(k)\frac{k^{100}}{2^{k/2}}\right| \leq \left|\frac{k^{100}}{2^{k/2}}\right|$$

Therefore, by part 1 of Theorem 3.38 it will be sufficient to prove that the series

$$\sum_{k=1}^{\infty} \frac{k^{100}}{2^{k/2}}$$

is absolutely convergent. We use the root test. Observe that

$$\sqrt[k]{\left|\frac{k^{100}}{2^{k/2}}\right|} = \frac{1}{\sqrt{2}}\sqrt[k]{k^{100}}.$$

We proved in Lemma 3.12 that $\lim_{n\to\infty} \sqrt[k]{k^{100}} = 1$. In particular, there is some $k_0 \in \mathbb{N}$ such that, for all $k \geq k_0$, $\sqrt[k]{k^{100}} \leq 1.1$. It follows that, for all $k \geq k_0$

$$\sqrt[k]{\left|\frac{k^{100}}{2^{k/2}}\right|} = \frac{1}{\sqrt{2}}\sqrt[k]{k^{100}} \leq 1.1 \cdot \frac{1}{\sqrt{2}} < 1.$$

The root test implies that

$$\sum_{k=1}^{\infty} \frac{k^{100}}{2^{k/2}}$$

is absolutely convergent, as required.

Example - Consider the series

$$\sum_{k=1}^{\infty} \frac{k^{k/4}}{3^{2+3k}}.$$

For the root test, we study the terms

$$\sqrt[k]{\left|\frac{k^{k/4}}{3^{2+3k}}\right|} = \frac{k^{1/4}}{27 \cdot \sqrt[k]{9}}.$$

Recall from Lemma 3.15 that $\lim_{k\to\infty} \sqrt[k]{9} = 1$. Therefore, there is some k_0 such that, for all $k \geq k_0$, we have $\sqrt[k]{9} \leq 2$. So, for all $k \geq k_0$,

$$\sqrt[k]{\left|\frac{k^{k/4}}{3^{2+3k}}\right|} \geq \frac{k^{1/4}}{54}.$$

We see that, for all k sufficiently large, the right hand side of the inequality above is at least 1. Indeed, for all $k \geq \max\{k_0, 54^4\}$, we have

$$\sqrt[k]{\left|\frac{k^{k/4}}{3^{2+3k}}\right|} \geq \frac{k^{1/4}}{54} \geq 1.$$

It follows from part 2 of Theorem 3.39 that the series

$$\sum_{k=1}^{\infty} \frac{k^{k/4}}{3^{2+3k}}.$$

is divergent.

The following corollary gives us a helpful repackaging of the root test.

Corollary 3.40. *Let $(a_k)_{k\in\mathbb{N}}$ be a sequence.*

1. *If*

$$\lim_{k\to\infty} \sqrt[k]{|a_k|} < 1$$

then the series $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

2. *If*

$$\lim_{k\to\infty} \sqrt[k]{|a_k|} > 1$$

then the series $\sum_{k=1}^{\infty} a_k$ is divergent.

Proof. 1. Since

$$\lim_{k\to\infty} \sqrt[k]{|a_k|} < 1$$

it follows that there is some $c < 1$ such that $\sqrt[k]{|a_k|} < c$ holds for all k sufficiently large. It then follows from part 1 of Theorem 3.39 that the series $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

2. If

$$\lim_{k\to\infty} \sqrt[k]{|a_k|} > 1$$

then it follows that $\sqrt[k]{|a_k|} > 1$ holds for all k sufficiently large. Part 2 of Theorem 3.39 then implies that the series $\sum_{k=1}^{\infty} a_k$ is divergent.

□

3.8.3 Ratio Test

The next test is based on quotients of successive terms of the series.

Theorem 3.41. *Let $\sum_{k=1}^{\infty} a_k$ be a series.*

1. *If there exists some real number $c < 1$ such that, for all but finitely many $k \in \mathbb{N}$,*

$$a_k \neq 0, \quad \text{and} \quad \left|\frac{a_{k+1}}{a_k}\right| \leq c,$$

then $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

2. *Conversely, if for all but finitely many $k \in \mathbb{N}$,*

$$a_k \neq 0, \quad \text{and} \quad \left|\frac{a_{k+1}}{a_k}\right| \geq 1,$$

then $\sum_{k=1}^{\infty} a_k$ is divergent.

Proof. 1. There exists $k_0 \in \mathbb{N}$ such that, for all $k \geq k_0$,

$$\left|\frac{a_{k+1}}{a_k}\right| \leq c.$$

It follows by induction that, for all $m \in \mathbb{N}$,

$$|a_{k_0+m}| \leq c^m |a_{k_0}|.$$

Let

$$b_n := c^n \cdot \frac{|a_{k_0}|}{c^{k_0}}.$$

The series $\sum_{k=1}^{\infty} b_k$ is absolutely convergent. This follows from Theorem 3.31 and Lemma 3.36. Also, for all $k \geq k_0$,

$$|a_k| = |a_{m+k_0}| \leq c^m |a_{k_0}| = c^{k-k_0} |a_{k_0}| = |b_k|.$$

It follows from part 1 of Theorem 3.38 that $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

2. There exists $k_0 \in \mathbb{N}$ such that, for all $k \geq k_0$,

$$\left|\frac{a_{k+1}}{a_k}\right| \geq 1.$$

It follows by induction that, for all $k \geq k_0$,

$$|a_k| \geq |a_{k_0}|.$$

In particular, the sequence $(a_k)_{k \in \mathbb{N}}$ is not a null sequence. It therefore follows from Corollary 3.34 that the series $\sum_{k=1}^{\infty} a_k$ is divergent.

□

The following corollary gives us a helpful repackaging of the ratio test.

Corollary 3.42. *Let $(a_k)_{k\in\mathbb{N}}$ be a sequence.*

1. *If*
$$\lim_{k\to\infty}\left|\frac{a_{k+1}}{a_k}\right| < 1$$
then the series $\sum_{k=1}^{\infty} a_k$ is absolutely convergent.

2. *If*
$$\lim_{k\to\infty}\left|\frac{a_{k+1}}{a_k}\right| > 1$$
then the series $\sum_{k=1}^{\infty} a_k$ is divergent.

Proof. This is left as an exercise. □

Example We can use the ratio test to prove that the series $\sum_{k=1}^{\infty} a_k$ given by
$$a_k = \frac{k^3}{k!}$$
is convergent. Indeed, for all $k \in \mathbb{N}$
$$\left|\frac{a_{k+1}}{a_k}\right| = \left|\frac{(k+1)^3}{(k+1)!}\cdot\frac{k!}{k^3}\right| = \frac{1}{k+1}\cdot\left(\frac{k+1}{k}\right)^3 \leq \frac{8}{k+1}.$$
The last inequality is just an application of the fact that $\frac{k+1}{k} \leq 2$ holds for all $k \in \mathbb{N}$. It therefore follows that, for all $k \geq 15$ we have
$$\left|\frac{a_{k+1}}{a_k}\right| \leq \frac{1}{2}.$$
The ratio test implies that the series $\sum_{k=1}^{\infty} \frac{k^3}{k!}$ is convergent.

Example - As was hinted at earlier, the ratio test does not always provide a definite answer regarding the convergence or divergence of a series. To see this, let's try and apply the ratio test for the series $\sum_{n=1}^{\infty} \frac{1}{n}$ (which we already know is divergent, see Lemma 3.30).

Observe that
$$\left|\frac{a_{k+1}}{a_k}\right| = \frac{k}{k+1}.$$
Since the sequence $\frac{k}{k+1}$ tends to 1, it follows that neither of the two conditions in Theorem 3.41 hold, and we do not get any information about the series $\sum_{n=1}^{\infty} \frac{1}{n}$ by this method.

4 Continuous functions and limits

4.1 Definition and examples

In this chapter, we discuss the notion of a function being continuous. The informal version of the definition of a continuous function is that its graph can be drawn without taking the pen off the page. While this is very helpful to keep in mind as a guide, this definition is insufficient for dealing with some more complicated functions, and so we will need a more formal approach.

We will also discuss the limit of a function. We have discussed limits of sequences in the previous chapter, and the notion of the limit of a function is an extension of this idea. Indeed, the definition of the limit of a function involves sequences.

Let us begin by giving the definition of a continuous function. We focus on the case of real-valued functions, since this allows for easier visualisations. However, most of the results in this chapter extend to complex-valued functions without significant changes.

Definition 4.1. *Let $D \subset \mathbb{R}$, $x_0 \in D$ and $f : D \to \mathbb{R}$. We call f* ***continuous at*** *x_0 if and only if, for all sequences $(x_n)_{n\in\mathbb{N}}$ in D such that $x_n \to x_0$, we have that $\lim_{n\to\infty} f(x_n)$ exists and*

$$\lim_{n\to\infty} f(x_n) = f(x_0).$$

If $U \subset D$ and f is continuous at x for all $x \in U$, then we say that f is ***continuous on*** *U. If f is continuous on the whole domain D, then we just say that f is* ***continuous****. If a function is not continuous, then we say it is* ***discontinuous****.*

Let us try and understand this definition by looking at some examples.

Example - For any $c \in \mathbb{R}$ and any $D \subset \mathbb{R}$, the constant function $f : D \to \mathbb{R}$ given by

$$f(x) = c, \quad \forall\, x \in D$$

is continuous. Indeed, let $x_0 \in D$ and suppose that we have a sequence $(x_n)_{n\in\mathbb{N}}$ in D with $x_n \to x_0$. Then

$$f(x_0) = c = \lim_{n\to\infty} c = \lim_{n\to\infty} f(x_n).$$

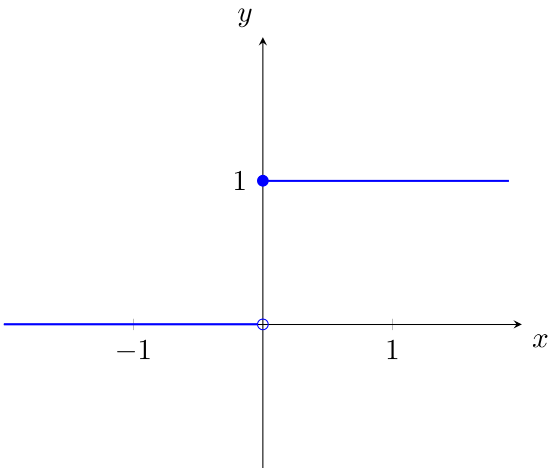
Example - The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x$ is continuous. Indeed, let $x_0 \in \mathbb{R}$ and suppose that we have a sequence $(x_n)_{n\in\mathbb{N}}$ with $x_n \to x_0$. Then

$$f(x_0) = x_0 = \lim_{n\to\infty} x_n = \lim_{n\to\infty} f(x_n).$$

Example - The quadratic function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ is continuous. Indeed, let $x_0 \in \mathbb{R}$ and suppose that we have a sequence $(x_n)_{n\in\mathbb{N}}$ with $x_n \to x_0$. Then

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_n^2 = \left(\lim_{n\to\infty} x_n\right)^2 = x_0^2 = f(x_0).$$

For the second equality above we have used part (iii) of Theorem 3.8.



Example - The prototype of a discontinuous function is the *Heaviside function* $H : \mathbb{R} \to \mathbb{R}$, defined by

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}.$$

Our intuition tells us that we cannot draw this graph without taking the pen off the page at $x = 0$, and so H is not continuous at 0. Let us verify this formally using the definition. Consider the sequence $x_n = -\frac{1}{n}$, and note that $x_n \to 0$. Also, $H(x_n) = 0$ for all $n \in \mathbb{N}$. Therefore,

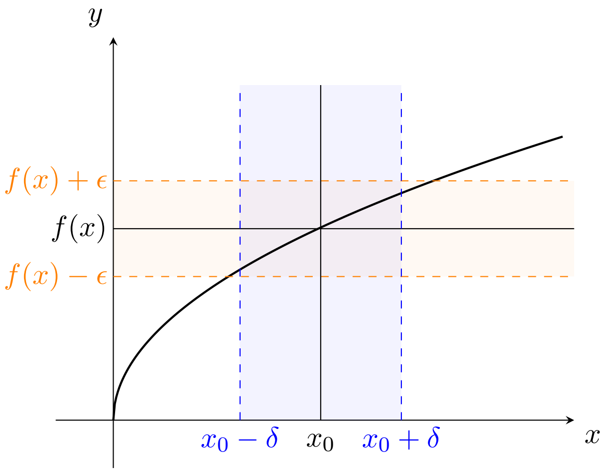
$$\lim_{n\to\infty} H(x_n) = \lim_{n\to\infty} 0 = 0 \neq 1 = H(0).$$

This shows that H is not continuous at 0.

Next, we present an alternative description of what it means for a function to be continuous at x_0. In many textbooks, this is used as the definition of continuity and, in many examples, this is the characterisation of continuity that will be most useful.

Theorem 4.2. *Let $D \subset \mathbb{R}$, $x_0 \in D$ and $f : D \to \mathbb{R}$. Then f is continuous at x_0 if and only if*

$$\forall \epsilon > 0, \exists \delta > 0 : |x_0 - x| < \delta \text{ and } x \in D \implies |f(x) - f(x_0)| < \epsilon. \tag{73}$$



The statement above is one of the most important and fundamental ones in formal mathematics, and is sometimes called the $\epsilon - \delta$ **criterion**. It is also a classically intimidating statement that is a necessary part of university mathematics education!

What does it mean? It says that, if x gets very close to x_0, then $f(x)$ gets very close to $f(x_0)$. Moreover, by taking x to be sufficiently close to x_0, we can make $f(x_0)$ and $f(x)$ as close as we like.

Proof of Theorem 4.2. First we show that, if f is continuous at x_0 then (73) holds. So suppose that f is continuous and, for the sake of contradiction, that (73) does not hold. That is,

$$\exists \, \epsilon > 0 : \forall \delta > 0, \, \exists \, x \in D, |x - x_0| < \delta \text{ and } |f(x) - f(x_0)| \geq \epsilon. \tag{74}$$

We use this to define a sequence $(x_n)_{n \in \mathbb{N}}$ such that $x_n \to x_0$ but $f(x_n) \not\to f(x_0)$. This will give the intended contradiction.

Indeed, for $n \in \mathbb{N}$, define $\delta_n := \frac{1}{n}$. Apply the statement (74) with $\delta = \delta_n$, for all $n \in \mathbb{N}$. We obtain a sequence (x_n) in D such that, for all $n \in \mathbb{N}$,

$$|x_n - x_0| < \frac{1}{n} \tag{75}$$

and

$$|f(x_n) - f(x_0)| \geq \epsilon. \tag{76}$$

It follows from (75) that $x_n \to x_0$. However, it follows from (76) that $f(x_n) \not\to f(x_0)$, which contradicts our assumption that f is continuous.

For the other direction, assume that (73) holds. Let $\epsilon > 0$ be arbitrary and suppose that (x_n) is a sequence with $x_n \to x_0$. Then, there exists $\delta > 0$ such that, for all $x \in D$

$$|x - x_0| < \delta \implies |f(x) - f(x_0)| < \epsilon.$$

Also, since $x_n \to x_0$, it follows from the definition of convergence of sequences that there exists $n_0 \in \mathbb{N}$ such that

$$n \geq n_0 \implies |x_n - x_0| < \delta \implies |f(x_n) - f(x_0)| < \epsilon.$$

We have thus shown that $f(x_n) \to f(x_0)$, and the proof is complete. □

Example - The quadratic function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ is continuous. We have already checked this using the definition, but now we will prove this using the $\epsilon - \delta$ criterion.

Let $\epsilon > 0$ and let $x_0 \in \mathbb{R}$ be arbitrary. We need to find $\delta > 0$ such that

$$|x - x_0| < \delta \implies |x^2 - x_0^2| < \epsilon.$$

Typically when applying the $\epsilon - \delta$ criterion, we will begin by treating δ as if it were a variable, and then fix it at the end of the proof to make the argument work. The value of δ can depend on ϵ and x_0 (but not on x). We will start by bounding the thing we want to bound (in this case $|x^2 - x_0^2|$), and try to find a way to state this bound in terms of $|x - x_0|$. We can then apply the assumption that $|x - x_0| < \delta$.

In this case, observe that

$$|x^2 - x_0^2| = |x - x_0||x + x_0| < \delta \cdot |x + x_0| = \delta \cdot |(x - x_0) + 2x_0| \leq \delta(\delta + 2|x_0|).$$

We need to make a choice of δ which gives $\delta(\delta + 2|x_0|) \leq \epsilon$. This inequality is easier to deal with if the second term of the product does not involve δ. Therefore, we note that, as long as $\delta \leq 1$, we have

$$\delta(\delta + 2|x_0|) \leq \delta(1 + 2|x_0|).$$

We have now reduced the problem to choosing δ such that $\delta(1 + 2|x_0|) \leq \epsilon$, which rearranges as $\delta \leq \frac{\epsilon}{1+2|x_0|}$.

We have imposed two requirements on δ; we need $\delta \leq 1$ and $\delta \leq \frac{\epsilon}{1+2|x_0|}$. Therefore, we may choose

$$\delta = \min \left\{ 1, \frac{\epsilon}{1 + 2|x_0|} \right\},$$

and the proof is complete.

Note that the choice of δ in the previous example does indeed depend on the fixed point x_0 at which we are checking that the function is continuous. This dependence is really necessary in this example, and we will revisit this issue in more detail later.

Exercise - Using Theorem 4.2, prove that the absolute value function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = |x|$ is continuous.

As our intuition suggests, the exponential function is continuous.

Theorem 4.3. *Let $f : \mathbb{R} \to \mathbb{R}_+$ be the function given by $f(x) = a^x$, where $a \in (0, \infty)$. Then f is continuous.*

(Recall that the set $\mathbb{R}_+$ mentioned in the previous statement is defined as $\mathbb{R}_+ := [0, \infty)$.)

Proof. The case when $a = 1$ is trivial, since then f becomes a constant function. In what follows, we will consider the case when $a > 1$. For the case when $0 < a < 1$, a slight modification of the argument gives the same result (I encourage you to write down the proof for this case rigorously).

We will use the $\epsilon - \delta$ criterion. Let $x_0 \in \mathbb{R}$ and $\epsilon > 0$ be arbitrary. We need to show that there exists $\delta > 0$ such that

$$|x_0 - x| < \delta \implies |a^{x_0} - a^x| < \epsilon.$$

We will fix $\delta > 0$ later in the proof to ensure that all of the inequalities we need are valid.

Note that

$$|a^{x_0} - a^x| = |a^{x_0}(1 - a^{x - x_0})| = a^{x_0}|1 - a^{x - x_0}|.$$

The assumption that $a > 1$ implies that the function f is strictly increasing. Therefore, we have

$$|x - x_0| < \delta \implies -\delta < x - x_0 < \delta \implies a^{-\delta} < a^{x - x_0} < a^{\delta}.$$

By considering the two possible values for the sign of $1 - a^{x - x_0}$, we see that[1]

$$|1 - a^{x - x_0}| < \max\{1 - a^{-\delta}, a^{\delta} - 1\}.$$

[1] In fact, the second term here is always the larger of the two, but we do not want to prove this here.

Therefore,

$$|a^{x_0} - a^x| = a^{x_0}|1 - a^{x-x_0}| < a^{x_0} \max\{1 - a^{-\delta}, a^\delta - 1\}.$$

We now need to choose δ sufficiently small so as to ensure that $a^{x_0} \max\{1 - a^{-\delta}, a^\delta - 1\} \leq \epsilon$. There are two inequalities we need here, namely

$$1 - a^{-\delta} \leq \frac{\epsilon}{a^{x_0}} \tag{77}$$

and

$$a^\delta - 1 \leq \frac{\epsilon}{a^{x_0}}. \tag{78}$$

The first requirement (77) can be rearranged to give

$$\delta \leq -\log_a \left(1 - \frac{\epsilon}{a^{x_0}}\right).$$

The second requirement (78) can be rearranged into the form

$$\delta \leq \log_a \left(1 + \frac{\epsilon}{a^{x_0}}\right).$$

Therefore, we can ensure that both (77) and (78) are satisfied by setting

$$\delta := \min\left\{-\log_a \left(1 - \frac{\epsilon}{a^{x_0}}\right), \log_a \left(1 + \frac{\epsilon}{a^{x_0}}\right)\right\}.$$

□

Theorem 4.4. *The trigonometric functions $f(x) = \sin(x)$ and $f(x) = \cos(x)$ are both continuous.*

The proof of Theorem 4.4 will be considered on the next exercise sheet, with some helpful hints provided there.

Example - The root function $f : [0, \infty) \to \mathbb{R}$ defined by $f(x) = \sqrt{x}$ is continuous. We will use the $\epsilon - \delta$ criterion to verify this.

Let $\epsilon > 0$ and $x_0 \in [0, \infty)$ be arbitrary. Observe that

$$|\sqrt{x} - \sqrt{x}_0|^2 = (\sqrt{x} - \sqrt{x_0})^2 = x + x_0 - 2\sqrt{xx_0} \leq x + x_0 - 2\min\{x, x_0\} = |x - x_0|.$$

Therefore, setting $\delta = \epsilon^2$, it follows that, for all x such that $|x - x_0| < \delta$, we have

$$|\sqrt{x} - \sqrt{x}_0| \leq |x - x_0|^{1/2} \leq \delta^{1/2} = \epsilon.$$

We could also have used the following theorem to prove that the root function is continuous.

Theorem 4.5. *Let I be an interval and let $D \subset \mathbb{R}$. Let $f : I \to D$ be an invertible function. If f is continuous on I, then the inverse function f^{-1} is continuous on D.*

To prove Theorem 4.5, we need to develop a little more theory. A proof will be given later in this chapter.

4.2 Calculation rules for continuous functions

Next, we want to establish some calculation rules for continuous functions. These rules allow us to prove continuity of complicated functions by proving that its (hopefully easier) 'building blocks' are continuous.

First of all, let us give a fairly obvious definition concerning the sum, product, ratio and dilate of functions.

Definition 4.6. *Let $D \subset \mathbb{R}$, $f, g : D \to \mathbb{R}$ and $c \in \mathbb{R}$. Then $f + g : D \to \mathbb{R}$ is a function defined by*

$$(f + g)(x) = f(x) + g(x), \quad \forall x \in D.$$

Similarly, $f \cdot g : D \to \mathbb{R}$ is defined by $(f \cdot g)(x) = f(x) \cdot g(x)$ and $c \cdot f : D \to \mathbb{R}$ is defined by $(c \cdot f)(x) = c \cdot f(x)$.

Additionally, suppose that $g(x) \neq 0$ for all $x \in D$. Then $\frac{f}{g} : D \to \mathbb{R}$ is defined by

$$\left(\frac{f}{g}\right)(x) = \frac{f(x)}{g(x)}.$$

Theorem 4.7. *Let $D \subset \mathbb{R}$. Suppose that $f, g : D \to \mathbb{R}$ are continuous at $x_0 \in D$ and $c \in \mathbb{R}$. Then $f + g$, $f \cdot g$ and $c \cdot f$ are continuous at x_0.*

In addition, if $g(x) \neq 0$ for all $x \in D$ then $\frac{f}{g}$ is continuous at x_0.

Proof. We will just prove the case of $f + g$ being continuous. The other statements can be proved similarly (and I encourage you to write down formal proofs).

Let (x_n) be a sequence with $x_n \to x_0$. We need to check that $\lim_{n\to\infty}(f + g)(x_n)$ exists and

$$\lim_{n\to\infty}(f + g)(x_n) = (f + g)(x_0).$$

This is indeed true, since

$$\lim_{n\to\infty}(f+g)(x_n) = \lim_{n\to\infty}(f(x_n)+g(x_n)) = \lim_{n\to\infty} f(x_n)+\lim_{n\to\infty} g(x_n) = f(x_0)+g(x_0) = (f+g)(x_0).$$

The second equality above is an application of part (i) of Theorem 3.8, and the third equality uses the assumption that f and g are continuous at x_0. □

Example - We can immediately use Theorem 4.7 to prove that polynomials are continuous functions. Indeed, let $p : \mathbb{R} \to \mathbb{R}$ be defined by

$$p(x) = \sum_{k=0}^{n} c_k x^k$$

where $c_0, \ldots, c_n \in \mathbb{R}$. We will now show that $p(x)$ is continuous.

We have already seen that constant functions and the identity are continuous on $\mathbb{R}$. Applying Theorem 4.7 repeatedly (in particular, using the fact that the product of two continuous functions is continuous), it follows that the function $g(x) = x^k$ is continuous for all $k \in \mathbb{N}$. It then follows from another application of Theorem 4.7 that $h(x) = c_k x^k$ is a continuous function. Finally, adding together these continuous functions and applying the theorem again, we get the result.

Example - It follows from the previous example and Theorem 4.5 that $f(x) = \sqrt[k]{x}$ is continuous on $[0,\infty)$ for all $k \in \mathbb{N}$. Indeed, $f(x)$ is the inverse of the polynomial function $g : [0,\infty) \to [0,\infty)$ given by $g(x) = x^k$. Since g is continuous, it follows from Theorem 4.5 that its inverse f is also continuous.

Example - We can use Theorems 4.7 and 4.4 to prove that $f(x) = \tan(x)$ is continuous for all x not of the form $x = \frac{\pi}{2} + k\pi$, where $k \in \mathbb{Z}$. This follows from the fact that $\tan x = \frac{\sin x}{\cos x}$.

Exercise - Use Theorem 4.7 to prove that the reciprocal function $h : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ given by $h(x) = 1/x$ is continuous.

The next result shows that the composition of two continuous functions is also continuous. See page 18 for a reminder of the definition of the composition of functions.

Theorem 4.8. *Let $D, E \subset \mathbb{R}$. Suppose that $f : D \to E$ is continuous at $x_0 \in D$, and that $g : E \to \mathbb{R}$ is continuous at $y_0 = f(x_0)$. Then $g \circ f$ is continuous at x_0.*

Proof. Let (x_n) be an arbitrary sequence in D such that $x_n \to x_0$. We need to show that $\lim_{n\to\infty}(g \circ f)(x_n) = (g \circ f)(x_0)$.

Let $y_n = f(x_n)$. Since f is continuous at x_0, it follows that the sequence (y_n) converges to y_0. Indeed,

$$\lim_{n\to\infty} y_n = \lim_{n\to\infty} f(x_n) = f(x_0) = y_0.$$

It therefore follows from the fact that g is continuous at y_0 that

$$\lim_{n\to\infty} g(f(x_n)) = \lim_{n\to\infty} g(y_n) = g(y_0) = g(f(x_0)),$$

as required. □

Example - A function $f : \mathbb{R} \to \mathbb{R}$ of the form

$$f(x) = a_0 + \sum_{k=1}^{n} (a_k \sin(kx) + b_k \cos(kx)),$$

where $a_k, b_k \in \mathbb{R}$ for all $k \in \mathbb{N} \cup \{0\}$ and $n \in \mathbb{N}$, is called a **trignometric polynomial**. It follows from Theorems 4.4, 4.7 and 4.8 that f is continuous. We will consider trigonometric polynomials in much more detail when we come to the topic of Fourier series later in the course.

Lemma 4.9. *Let $f : D \to \mathbb{R}$ be continuous. The function $|f| : D \to \mathbb{R}$ given by*

$$(|f|)(x) = |f(x)|$$

is continuous.

Proof. Note that $|f| = g \circ f$, where $g(x) = |x|$ is the absolute value function. It was an exercise in the previous section to show that the absolute value function is continuous. An application of Theorem 4.8 completes the proof. □

Lemma 4.10. *Let $f, g : D \to \mathbb{R}$ be continuous. Then the function* $\min\{f, g\} : D \to \mathbb{R}$ *given by*

$$(\min\{f, g\})(x) = \min\{f(x), g(x)\}$$

is continuous. Similarly, $\max\{f, g\} : D \to \mathbb{R}$ *is continuous.*

Proof. Observe that, for any $a, b \in \mathbb{R}$,

$$\min\{a, b\} = \frac{a + b - |a - b|}{2}$$

and

$$\max\{a, b\} = \frac{a + b + |a - b|}{2}.$$

We can check these identities using case distinction (please do this).

Therefore,

$$(\min\{f, g\})(x) = \min\{f(x), g(x)\} = \frac{f(x) + g(x) - |f(x) - g(x)|}{2}.$$

That is the same as saying that

$$\min\{f, g\} = \frac{f + g - |f - g|}{2}.$$

It follows from Theorem 4.7 and Lemma 4.9 that this function is continuous. A similar argument works for the case of $\max\{f, g\}$. □

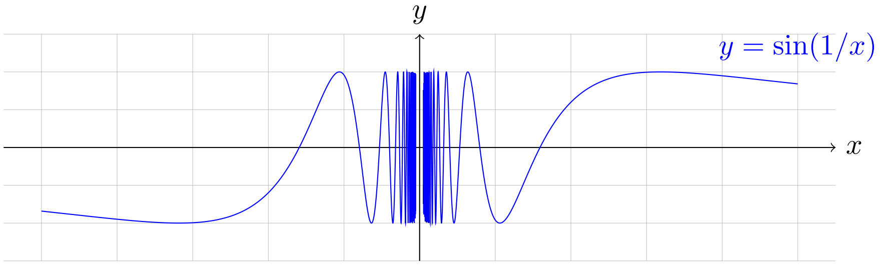
Example - Let $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ be given by

$$f(x) = \sin\left(\frac{1}{x}\right).$$

Since f is formed by composition of continuous functions, it follows from Theorem 4.8 that f is continuous. Indeed, $f = g \circ h$ where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \setminus \{0\}$ are given by

$$g(x) = \sin x, \quad h(x) = \frac{1}{x}.$$

This is an interesting example, since the heuristic that a function is continuous if it can be drawn without taking the pen of the page is not so easy to apply here. The behaviour of the function becomes very unstable as x gets close to 0, and it is very difficult to give a good drawing of this function.



4.3 Limits of functions

In this section, we will formally define the notion of the limit of a function, building on our earlier work for understanding the limit of a sequence. We first need the notion of an accumulation point, which is a point for which the limit may be possible to define. We will see later that this allows us to define the limit of a function as it approaches a point *outside* of the domain.

Definition 4.11. *Let $D \subset \mathbb{R}$ be a non-empty set. We call $x_0 \in \mathbb{R} \cup \{-\infty, \infty\}$ an **accumulation point** of D if there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in D such that*

$$x_0 = \lim_{n\to\infty} x_n \quad \textit{and} \quad x_n \neq x_0 \quad \forall\, n \in \mathbb{N}.$$

In particular, note that this definition allows for the possibility that ∞ and $-\infty$ are accumulation points.

In general, accumulation points for a set D can be quite varied. Not all points of D are automatically accumulation points, and not all accumulation points are in D. An extreme example is the set

$$M := \left\{ \frac{1}{n} : n \in \mathbb{N} \right\}.$$

Observe that 0 is an accumulation point of M, even though $0 \notin M$. This is because we can construct a sequence in M which has limit 0 but is never equal to 0, such as the sequence $(1/n)_{n\in\mathbb{N}}$. However, none of the elements of M are accumulation points of M. For example, $1/100 \in M$, but there are no sequences in M which converge to $1/100$ which never take this value. So, the set M is completely disjoint from its accumulation points!

The reason for this "unusual" example is that the set M consists of discrete or *isolated* points, and that the elements of M are in some sense quite sparse. However, such sets are usually not very helpful for constructing interesting continuous functions, and so we will not spend much time worrying about them. We will mostly be interested in the case when D is an interval, or perhaps when D is an interval with some points removed. Let us take a closer look at the accumulation points for these two cases.

Examples - Let $D = (a, b)$. The set of accumulation points for D is the closed interval $[a, b]$. If $D = (a, \infty)$ then the set of accumulation points is $[a, \infty) \cup \{\infty\}$.

Example - Let $c \in \mathbb{R}$ and $D = \mathbb{R} \setminus \{c\}$. Note that c is an accumulation point for D (for example, we can consider the sequence $a_n = c + \frac{1}{n}$). The set of accumulation points for D is $\mathbb{R} \cup \{-\infty, \infty\}$.

Exercise - What is the set of all accumulation points for $\mathbb{Q}$?

We now can give a precise definition of the limit of a function.

Definition 4.12. *Let $D \subset \mathbb{R}$, $f : D \to \mathbb{R}$, $y \in \mathbb{R} \cup \{-\infty, \infty\}$ and let $x_0 \in \mathbb{R} \cup \{-\infty, \infty\}$ be an accumulation point of D. We call y the* ***limit of** f **as** x **tends to*** *x_0 if, for any sequence $(x_n)_{n\in\mathbb{N}}$ in D such that*

$$x_n \to x_0 \quad \text{and} \quad x_n \neq x_0 \ \forall\, n \in \mathbb{N}$$

we have

$$\lim_{n\to\infty} f(x_n) = y.$$

In this case, we write

$$\lim_{x\to x_0} f(x) = y.$$

The purpose of the definition above is to capture the behaviour of the function f as we get infinitesimally close to x_0. Let us see a few examples to gain a better understanding.

Example - Consider the identity function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x$. Every $c \in \mathbb{R}$ is an accumulation point, and so we can consider the possibility that $\lim_{x\to c} f(x)$ exists. Guided by our intuition, we expect that

$$\lim_{x\to c} f(x) = c.$$

Let's check that this is indeed true. Let $(x_n)_{n\in\mathbb{N}}$ be an arbitrary sequence such that $x_n \to c$ and $x_n \neq c$ for all $n \in \mathbb{N}$. Then

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_n = c.$$

Since this sequence was arbitrary, we have verified that $\lim_{x\to c} f(x) = c$.

Since ∞ is an accumulation point for the domain of this function, we can also consider the limit $\lim_{x\to\infty} f(x)$. This is equal to ∞, as one expects. Indeed, for any sequence $(x_n)_{n\in\mathbb{N}}$ such that $x_n \to \infty$, we have

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_n = \infty.$$

Example - In the last example, we considered the limit of a function f as x tends to a value outside of the domain, namely as x tends to ∞. We can also define the limit of a function as x tends to a real number value outside of the domain. Consider for instance the function $f : \mathbb{R} \setminus \{1\} \to \mathbb{R}$ given by

$$f(x) = \frac{x^2 - 1}{x - 1}, \quad \forall\, x \in \mathbb{R} \setminus \{1\}.$$

Since 1 is an accumulation point of the domain, we can consider the limit $\lim_{x\to 1} f(x)$. Let $(x_n)_{n\in\mathbb{N}}$ be any sequence such that $x_n \to 1$ and $x_n \neq 1$ for all $n \in \mathbb{N}$. Then,

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} \frac{x_n^2 - 1}{x_n - 1} = \lim_{n\to\infty} \frac{(x_n - 1)(x_n + 1)}{x_n - 1} = \lim_{n\to\infty} (x_n + 1) = 2.$$

Example - Note that the limit of a function at a given point is not always defined. In order for the limit to be defined at x_0, we need that

$$\lim_{n\to\infty} f(x_n)$$

is the same for *all* sequences $x_n \to x_0$ (with the additional condition that $x_n \neq x_0$ for all $n \in \mathbb{N}$). To see this, consider once again the Heaviside function

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}.$$

The limit of H as x tends to 0 is not defined. To see this, consider the sequences $(1/n)_{n\in\mathbb{N}}$ and $(-1/n)_{n\in\mathbb{N}}$. Both of these sequences converge to 0 and never take the value 0. However,

$$\lim_{n\to\infty} H(1/n) = \lim_{n\to\infty} 1 = 1$$

and

$$\lim_{n\to\infty} H(-1/n) = \lim_{n\to\infty} 0 = 0.$$

So, we see that there is no valid choice of y in Definition 4.12 for the Heaviside function.

Note that Definition 4.12 does not make any mention of the value $f(x_0)$. So, the limit of f as we approach x_0 does not necessarily have any relation to $f(x_0)$. However, if f is continuous at $x_0 \in D$ and x_0 is an accumulation point of D, then this limit must be equal to $f(x_0)$.

Theorem 4.13. *Let $D \subset \mathbb{R}$, $f : D \to \mathbb{R}$ and let $x_0 \in D$ be an accumulation point of D. Then*

$$f \text{ is continuous at } x_0 \iff \lim_{x\to x_0} f(x) = f(x_0).$$

Proof. This is left as an exercise. □

Example - Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^3 + 2x^2 - 1$. This is a polynomial, and so it is continuous. It therefore follows immediately from Theorem 4.13 that

$$\lim_{x\to 2} f(x) = f(2) = 15.$$

Moreover, for any $a \in \mathbb{R}$,

$$\lim_{x\to a} f(x) = f(a).$$

It can be the case that we have a function f which is not continuous at x_0, but for which the limit as x tends to x_0 is defined. If this situation occurs, then Theorem 4.13 tells us that

$$\lim_{x\to x_0} f(x) \neq f(x_0).$$

Consider for instance, the function $f : \mathbb{R} \to \mathbb{R}$ given by

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

We would like to determine the value of $\lim_{x\to 0} f(x)$. Let $(x_n)_{n\in\mathbb{N}}$ be an arbitrary sequence such that $x_n \to 0$ and $x_n \neq 0$ for all $n \in \mathbb{N}$. Then,

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} 0 = 0.$$

Therefore,

$$\lim_{x\to 0} f(x) = 0 \neq 1 = f(0).$$

Example - Consider the reciprocal function $f : \mathbb{R}\setminus\{0\} \to \mathbb{R}$ given by $f(x) = \frac{1}{x}$. The purpose of this example is to show that the limit

$$\lim_{x\to 0} f(x)$$

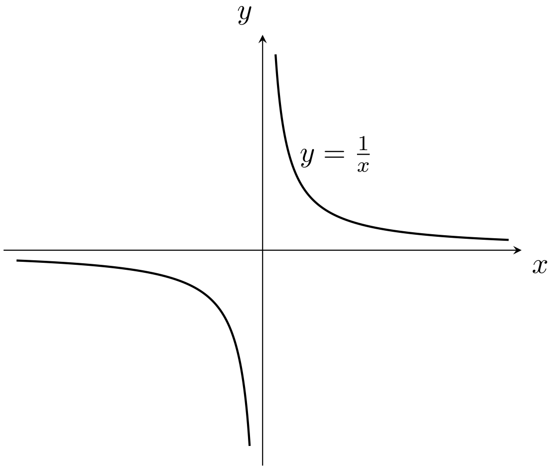
is not defined. To see this, consider the sequences $(1/n)_{n\in\mathbb{N}}$ and $(-1/n)_{n\in\mathbb{N}}$. Both of these sequences converge to 0 and never take the value 0. However,

$$\lim_{n\to\infty} f(1/n) = \lim_{n\to\infty} n = \infty$$

and

$$\lim_{n\to\infty} f(-1/n) = \lim_{n\to\infty} -n = -\infty.$$

Since these two sequences have different limits, it follows that $\lim_{x\to 0} f(x)$ is not defined.



Exercise - Let $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ be given by $f(x) = \frac{1}{x^2}$. Calculate $\lim_{x\to 0} f(x)$, if it exists.

Recall that Theorem 3.8 allows us to quickly calculate limits of sequences that are formed as sums and products of convergent sequences. The next result gives an analogue of Theorem 3.8 for limits of functions.

Theorem 4.14. *Let $D \subset \mathbb{R}$, $f, g : D \to \mathbb{R}$ and suppose that x_0 is an accumulation point of D. Suppose that there exist $A, B \in \mathbb{R}$ such that*

$$\lim_{x\to x_0} f(x) = A \quad \text{and} \quad \lim_{x\to x_0} g(x) = B.$$

Then, we have

1. $\lim_{x\to x_0} f(x) + g(x) = A + B$,
2. $\lim_{x\to x_0} f(x) \cdot g(x) = A \cdot B$, *and*
3. *if* $B \neq 0$ *then* $\lim_{x\to x_0} \frac{f(x)}{g(x)} = \frac{A}{B}$.

Proof. This is left as an exercise. □

Example - Consider

$$\lim_{x\to 3}(x^2-1)(e^x-\sqrt{12x}).$$

Using the previous theorem, we can split this limit up as

$$\lim_{x\to 3}(x^2-1)(e^x-\sqrt{12x}) = \left(\lim_{x\to 3}x^2-1\right)\left((\lim_{x\to 3}e^x)-(\lim_{x\to 3}\sqrt{12x})\right).$$

Since each of the three component functions are continuous, we can use Theorem 4.13 to conclude that

$$\lim_{x\to 3}(x^2-1)(e^x-\sqrt{12x}) = 8(e^3-6)\approx 112.684.$$

The next result shows that limits behave nicely with respect to the composition of functions.

Theorem 4.15. *Let $D, E \subset \mathbb{R}$, $g : D \to E$ and $h : E \to \mathbb{R}$. Let x_0 be an accumulation point of D and let $y_0 := \lim_{x\to x_0} g(x) \in E$. Suppose that h is continuous at y_0. Then*

$$\lim_{x\to x_0} h\circ g(x) = h(y_0).$$

Perhaps an easier way to remember the conclusion of Theorem 4.15 is to write it as

$$\lim_{x\to x_0} h\circ g(x) = h\left(\lim_{x\to x_0} g(x)\right).$$

That is, the limit can be taken either inside or outside the composition operation, giving the same outcome.

Proof. We need to show that, for all $x_n \to x_0$ such that $x_n \neq x_0$ for all $n \in \mathbb{N}$, we have

$$\lim_{n\to\infty} h\circ g(x_n) = h(y_0).$$

Let (y_n) be the sequence given by defining $y_n = g(x_n)$. Then, since $y_0 = \lim_{x\to x_0} g(x)$, it follows from the definition of the limit that

$$\lim_{n\to\infty} y_n = \lim_{n\to\infty} g(x_n) = y_0.$$

Since h is continuous at y_0, it follows from the definition of continuity that

$$\lim_{n\to\infty} h(g(x_n)) = \lim_{n\to\infty} h(y_n) = h(y_0),$$

as required.

□

Example - Note that the previous theorem does not require the function g to be continuous at x_0, or even for $f(x_0)$ to be defined. Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} \sin(x^2) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0. \end{cases}$$

This can be written as a composition $f = h \circ g$ where $h(x) = \sin x$ and

$$g(x) = \begin{cases} x^2 & \text{if } x \neq 0 \\ \frac{\pi}{2} & \text{if } x = 0 \end{cases}.$$

The function h is continuous everywhere, and

$$\lim_{x \to 0} g(x) = 0.$$

It therefore follows from Theorem 4.15 that

$$\lim_{x \to x_0} f(x) = \lim_{x \to x_0} h \circ g(x) = h\left(\lim_{x \to x_0} g(x)\right) = h(0) = 0.$$

4.4 One-sided limits

We have seen earlier in this chapter, when considering jump functions such as the Heaviside function, that the limiting behaviour of the function can be different at depending on which side we approach from. This motivates the definition of **one-sided limits** of functions.

Definition 4.16. *Let $D \subset \mathbb{R}$ and $f : D \to \mathbb{R}$. Let $x_0 \in \mathbb{R}$ be an accumulation point of $D_+ := D \cap (x_0, \infty)$. We say that $y \in \mathbb{R} \cup \{-\infty, \infty\}$ is the* ***right-sided limit*** *of f as $x \to x_0$ if, for every sequence $(x_n) \subset D_+$ such that $x_n \to x_0$ we have*

$$\lim_{n\to\infty} f(x_n) = y.$$

We use the notation

$$\lim_{x\to x_0^+} f(x) = y.$$

Let x_0 be an accumulation point of $D_- := D \cap (-\infty, x_0)$. We say that y is the ***left-sided limit*** *of f as $x \to x_0$ if, for every sequence $(x_n) \subset D_-$ such that $x_n \to x_0$ we have*

$$\lim_{n\to\infty} f(x_n) = y.$$

We use the notation

$$\lim_{x\to x_0^-} f(x) = y.$$

Note that the assumption that x_0 is an accumulation point of D_+ just means that there are points in D that are to the right of x_0 and arbitrarily close to x_0, i.e., there is a sequence in D_+ converging to x_0. That x_0 is an accumulation point of D_- means that there are points in D that are to the left of x_0 and arbitrarily close to it.

Example - Consider again the Heaviside function $H : \mathbb{R} \to \mathbb{R}$ given by

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}.$$

Our intuition informs us that

$$\lim_{x\to 0^+} H(x) = 1$$

and

$$\lim_{x\to 0^-} H(x) = 0$$

Let's check the first of these claims. Observe that, in this situation, we have $\mathbb{R}_+ = \mathbb{R} \cap (0, \infty) = (0, \infty)$. Suppose that (x_n) is a sequence in $(0, \infty)$ such that $x_n \to x_0$. Then

$$\lim_{n\to\infty} H(x_n) = \lim_{n\to\infty} 1 = 1.$$

Example - Consider again the reciprocal function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ given by $f(x) = \frac{1}{x}$. We saw earlier in this section that the limit of this function as x tends to zero is not defined. However, the left and right sided limits are defined.

To see that the right sided limit is defined, let (x_n) be an arbitrary sequence in $(0,\infty)$ such that $x_n \to 0$. Then,

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} \frac{1}{x_n} = \infty.$$

To see that this limit really is equal to ∞, we should recall what it means for a sequence to tend to ∞. We need to check that

$$\forall\, C \in \mathbb{R},\, \exists\, n_0 \in \mathbb{N} : \forall\, n \geq n_0,\, \frac{1}{x_n} \geq C.$$

This is indeed true. Since $x_n \to 0$, it follows that, for all $C \in \mathbb{R}$, there exists n_0 such that, for all $n \geq n_0$ we have

$$x_n = |x_n - 0| \leq \frac{1}{C}.$$

A rearrangement of this inequality gives $\frac{1}{x_n} \geq C$, as required.

We therefore have

$$\lim_{x\to 0^+} f(x) = \infty.$$

Similarly, we can check that

$$\lim_{x\to 0^-} f(x) = -\infty.$$

The previous examples hint at the following result.

Theorem 4.17. *Let $D \subset \mathbb{R}$, $f : D \to \mathbb{R}$ and let $x_0 \in \mathbb{R}$ be an accumulation point of both $D \cap (x_0,\infty)$ and $D \cap (-\infty, x_0)$. Then*

$$\lim_{x\to x_0} f(x) \text{ exists} \iff \lim_{x\to x_0^+} f(x) \text{ and } \lim_{x\to x_0^-} f(x) \text{ exist and are equal.}$$

In this case,

$$\lim_{x\to x_0} f(x) = \lim_{x\to x_0^+} f(x) = \lim_{x\to x_0^-} f(x).$$

Proof. First, let us assume that $\lim_{x\to x_0} f(x)$ exist. Then, by definition, there is some $y \in \mathbb{R} \cup \{-\infty, \infty\}$ such that every sequence (x_n) with $x_n \to x_0$ satisfies

$$\lim_{n\to\infty} f(x_n) = y. \tag{79}$$

It immediately follows that every sequence (x_n) to the left of x_0 also satisfies (79), and so

$$\lim_{x\to x_0^-} f(x) = y.$$

The same argument, considering only sequences converging to x_0 from the right, shows that

$$\lim_{x\to x_0^+} f(x) = y.$$

To prove the converse, let us assume that

$$\lim_{x\to x_0^+} f(x) = \lim_{x\to x_0^-} f(x) = y,$$

for some $y \in \mathbb{R} \cup \{-\infty, \infty\}$. Let x_n be an arbitrary sequence in D such that $x_n \to x_0$ and $x_n \neq x_0$ for all $n \in \mathbb{N}$. We need to show that

$$\lim_{n \to \infty} f(x_n) = y.$$

Let us assume that, for infinitely many $k \in \mathbb{N}$ we have $x_k < x_0$ and for infinitely many $j \in \mathbb{N}$ we have $x_j > x_0$ (the case when the sequence has only finitely many terms on one side of x_0 can be handled separately, and this is left to the student to check). We use this to split the sequence (x_n) into two disjoint subsequences $(x_{j_m})_{m \in \mathbb{N}}$ and $(x_{k_m})_{m \in \mathbb{N}}$ such that, for all $m \in \mathbb{N}$,

$$x_{j_m} < x_0 \text{ and } x_{k_m} > x_0.$$

Case 1 - Assume that $y \in \mathbb{R}$. We need to show that, for all $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that

$$n \geq n_0 \implies |f(x_n) - y| < \epsilon. \tag{80}$$

Since $\lim_{x \to x_0^-} f(x) = y$, it follows that there is some m_0 such that, for all $m \geq m_0$,

$$|f(x_{j_m}) - y| < \epsilon.$$

Similarly, since $\lim_{x \to x_0^+} f(x) = y$ it follows that there is some m_0' such that, for all $m \geq m_0'$,

$$|f(x_{k_m}) - y| < \epsilon.$$

It follows that there is some $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$, we have $|f(x_n) - y| < \epsilon$. In particular, set $n_0 := \max\{j_{m_0}, k_{m_0'}\}$. This proves the intended statement (80) and completes the proof for this case.

Case 2 - Suppose that $y = \infty$. We need to show that, for all $C \in \mathbb{R}$, there exists n_0 such that

$$n \geq n_0 \implies f(x_n) > C. \tag{81}$$

Since $\lim_{x \to x_0^-} f(x) = \infty$, it follows that there is some m_0 such that, for all $m \geq m_0$,

$$f(x_{j_m}) > C.$$

Similarly, since $\lim_{x \to x_0^+} f(x) = \infty$ it follows that there is some m_0' such that, for all $m \geq m_0'$,

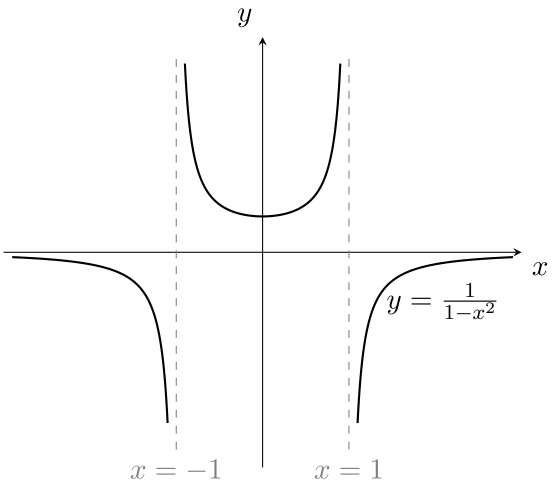
$$f(x_{k_m}) > \epsilon.$$

It follows that there is some $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$, we have $f(x_n) > C$. In particular, set $n_0 := \max\{j_{m_0}, k_{m_0'}\}$. This proves the intended statement (81) and completes the proof for this case.

Case 3 - Suppose that $y = -\infty$. Then the proof from case 2 works with some small adjustments. □

Example - Consider the function $f : \mathbb{R} \setminus \{-1, 1\} \to \mathbb{R}$ given by

$$f(x) = \frac{1}{1 - x^2} = \frac{1}{(1-x)(1+x)}.$$



This function is continuous at every point of its domain. Therefore, by Theorem 4.13 and Theorem 4.17, we have

$$\frac{1}{1-x_0^2} = \lim_{x \to x_0} f(x) = \lim_{x \to x_0^+} f(x) = \lim_{x \to x_0^-} f(x)$$

for all $x \in \mathbb{R} \setminus \{-1, 1\}$. What about the case of $x_0 = \pm 1$? We can get a pretty good idea about the limits at these points by making a sketch of the function.

This sketch suggests that

$$\lim_{x \to -1^-} f(x) = -\infty \tag{82}$$

$$\lim_{x \to -1^+} f(x) = \infty \tag{83}$$

$$\lim_{x \to 1^-} f(x) = \infty \tag{84}$$

$$\lim_{x \to 1^+} f(x) = -\infty. \tag{85}$$

We formally check that the first of these one-sided limits is correct (I encourage you to write down a similarly formal proof for the remaining 3 one-sided limits). To prove

$$\lim_{x \to -1^-} f(x) = -\infty$$

we need to check that, for any sequence $x_n \in (-\infty, -1)$ such that $x_n \to -1$, we have

$$\forall\, C \in \mathbb{R}, \exists\, n_0 \in \mathbb{N} : \forall\, n \geq n_0,\ f(x_n) \leq C. \tag{86}$$

We will in fact check that the following statement is true:

$$\forall\, C \geq 3, \exists\, n_0 \in \mathbb{N} : \forall\, n \geq n_0,\ -f(x_n) \geq C. \tag{87}$$

I encourage you to carefully consider why (87) implies (86).

Let $C \geq 3$ be given. Since $x_n \to -1$, it follows that, for all $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$|x_n + 1| = |x_n - (-1)| < \epsilon. \tag{88}$$

Consequently, by the triangle inequality,

$$|x_n - 1| = |x_n + 1 - 2| \leq |x_n + 1| + |-2| < \epsilon + 2. \tag{89}$$

It therefore follows that, for all $n \geq n_0$,

$$|f(x_n)| = \left|\frac{1}{(1-x_n)(1+x_n)}\right| = \frac{1}{|1-x_n|} \cdot \frac{1}{|1+x_n|} > \frac{1}{\epsilon(2+\epsilon)}. \tag{90}$$

Now, observe that $x_n < -1$ implies that $f(x_n)$ is negative. Indeed, this follows from the fact that, in this range, $1 - x_n$ is positive and $1 + x_n$ is negative. Therefore, (90) can be rewritten as

$$-f(x_n) > \frac{1}{\epsilon(2+\epsilon)}. \tag{91}$$

Since this is valid for all $\epsilon > 0$, we are free to make a choice ϵ which gives the intended inequality $-f(x_n) \geq C$. We set

$$\epsilon = \frac{1}{3C}.$$

Note that, because of the assumption that $C \geq 3$, it follows that $\epsilon \leq 1$, and so (91) gives us

$$-f(x_n) > \frac{1}{\epsilon(2+\epsilon)} \geq \frac{1}{3\epsilon} = C,$$

as required.

Finally, it follows from (82) and (83), along with Theorem 4.17, that

$$\lim_{x \to -1} f(x)$$

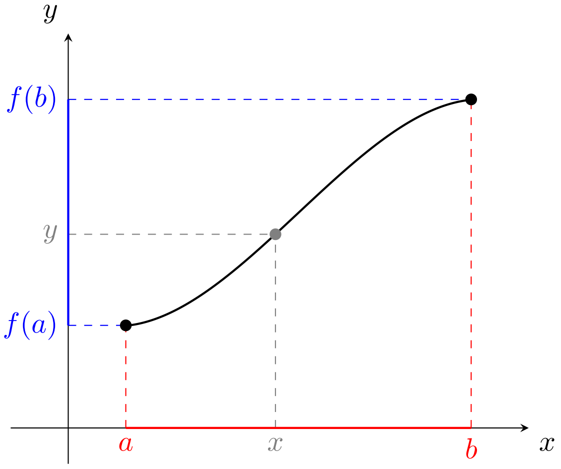
does not exist. Similarly, it follows from (84) and (85) that

$$\lim_{x \to 1} f(x)$$

does not exist.

4.5 The Intermediate Value Theorem

In this section we will discuss the Intermediate Value Theorem, which states that a continuous function on a closed interval attains all values between the function values at the two endpoints of the interval. We will use the Intermediate Value Theorem to deduce some interesting theoretical consequence, including a proof that the range of a continuous function on a closed interval is also a closed interval.



Theorem 4.18 (Intermediate Value Theorem)**.** *Let a and b be real numbers such that $a < b$ and $I = [a, b]$. Let $f : I \to \mathbb{R}$ be a continuous function. Then, for every*

$$y \in [\min\{f(a), f(b)\}, \max\{f(a), f(b)\}]$$

there exists $x \in I$ such that $f(x) = y$.

This statement is perhaps slightly easier to digest if we make the simplifying assumption that $f(a) \le f(b)$. We then conclude that, for every $y \in [f(a), f(b)]$ there exists $x \in I$ such that $f(x) = y$.

Proof. First of all, observe that the case when $f(a) = f(b)$ follows immediately, since the only value of y we need to consider is $y = f(a) = f(b)$.

We may henceforth assume that $f(a) \neq f(b)$. In fact, we may assume without loss of generality that $f(a) < f(b)$. The case when $f(a) > f(b)$ can be handled similarly, with some small changes to the forthcoming proof.

The goal of the proof is construct a convergent sequence (x_n) in I such that $x_n \to x \in I$ and $f(x) = y$. We will do this by strategically and repeatedly cutting the interval I in half.

We can formally define the sequence x_n by an algorithm. We will in fact simultaneously define three interdependent sequences, (a_n), (b_n) and (x_n). Set $a_1 = a$ and $b_1 = b$. For $n \in \mathbb{N}$, let x_n be the midpoint

$$x_n := \frac{a_n + b_n}{2}.$$

The definition of a_{n+1} and b_{n+1} depends on whether $f(x_n)$ is greater than or smaller than y.

- If $f(x_n) \geq y$ then we set
$$a_{n+1} := a_n \text{ and } b_{n+1} = x_n.$$
- If $f(x_n) < y$ then we set
$$a_{n+1} := x_n \text{ and } b_{n+1} = b_n.$$

With this, we get two sequences (a_n) and (b_n) in I such that
$$f(a_n) \leq y \leq f(b_n), \quad \forall n \in \mathbb{N}. \tag{92}$$
Also, the sequences (a_n) and (b_n) are both monotone and bounded. Theorem 3.20 then tells us that both sequences are convergent.

Next observe that
$$b_n - a_n = \frac{b-a}{2^{n-1}}.$$
Indeed, $b_1 - a_1 = b - a$, and in each step we halve the length of the interval. Therefore,
$$\lim_{n\to\infty} b_n - \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n - a_n = \lim_{n\to\infty} \frac{b-a}{2^{n-1}} = 0,$$
and it follows that
$$\lim_{n\to\infty} b_n = \lim_{n\to\infty} a_n := x.$$
Since f is continuous, we have
$$\lim_{n\to\infty} f(b_n) = \lim_{n\to\infty} f(a_n) = f(x).$$
Note also that $x \in I$. Indeed, suppose for a contradiction that $x \notin I$. Then we have a sequence (a_n) in I such that $a_n \neq x$ for all $n \in \mathbb{N}$ and $a_n \to x$. By definition, this means that x is an accumulation point of I. But we know that the set of accumulation points for I is also I, and so $x \in I$, a contradiction.

Finally, we can use the Sandwich Rule for sequences to complete the proof. Indeed, by (92) and applying the Sandwich Rule with the constant sequence $y_n = y$, we have that
$$y = \lim_{n\to\infty} y_n = \lim_{n\to\infty} f(b_n) = \lim_{n\to\infty} f(a_n) = f(x).$$

□

Example - The Intermediate Value Theorem is a useful tool for showing that certain equations have solutions in a given interval. For example, we can use the Intermediate Value Theorem to prove that the equation $e^x = 4 + x^3$ has a solution $x \in [-2,-1]$. Setting $f(x) = e^x - x^3 - 4$, this is the same thing as showing that there is some $x \in [-2,-1]$ such that
$$f(x) = 0.$$
f is a continuous function. Also,
$$f(-2) = e^{-2} + 4 > 0, \quad \text{and} \quad f(-1) = e^{-1} - 3 < 0.$$
Apply the Intermediate Value Theorem with $y = 0 \in [e^{-1} - 3, e^{-2} + 4]$. Then there is some $x \in [-2,-1]$ such that $f(x) = y$, as required.

Exercise - Let f be a polynomial with degree 3. Use the Intermediate Value Theorem to prove that there is some $x \in \mathbb{R}$ such that $f(x) = 0$.

Here is a nice corollary of the Intermediate Value Theorem.

Corollary 4.19. *Let $f : [0,1] \to [0,1]$ be a continuous function. Then there exists $x \in [0,1]$ such that $f(x) = x$.*

Proof. Consider the continuous function $g : [0,1] \to [0,1]$ given by

$$g(x) = f(x) - x.$$

Note that

$$g(0) = f(0) \geq 0$$

and

$$g(1) = g(1) - 1 \leq 0.$$

Therefore, by the Intermediate Value Theorem, there exists $x \in [0,1]$ such that

$$g(x) = 0.$$

It immediately follows that $f(x) = x$. □

The next result in this section will be the **Extreme Value Theorem**, which states that a continuous function on a closed interval attains its extreme values. We first need to properly define what minimal and maximal values of a function are.

Definition 4.20. *Let $D \subset \mathbb{R}$ and $f : D \to \mathbb{R}$. Then f has a* ***(global) minimum*** *at $x_0 \in D$ if*

$$f(x) \geq f(x_0) \ \ \forall x \in D,$$

and f has a ***(global) maximum*** *at $x_0 \in D$ if*

$$f(x) \leq f(x_0) \ \ \forall x \in D.$$

The value $f(x_0)$ is called a ***minimum/maximum*** *of f, or an* ***extreme value****, or an* ***extremum****.*

We use the word "global" in this definition because we will later introduce the related concept of local extrema.

Example - Consider the function $f : [-1,1] \to \mathbb{R}$ given by $f(x) = x^2$. Then, for all $x \in [-1,1]$,

$$f(x) \leq 1 = f(1) = f(-1).$$

Therefore, f has a global maximum at -1 and 1. Also,

$$f(x) \geq 0 = f(0)$$

for all $x \in [-1,1]$. Therefore, f has a global minimum at 0.

Example - In this example, we show how a very small change to the function changes the nature of the extrema. Consider the function $f : (-1,1) \to \mathbb{R}$ given by $f(x) = x^2$. Note that we have only changed the domain from the previous example, which is now an open interval. However, in this case, there is no global maximum. Indeed, for all $x \in (-1,1)$, there exists $y \in (-1,1)$ such that $f(y) > f(x)$.

With these examples in mind, we are ready to state the Extreme Value Theorem, which states that a continuous function on a closed interval has a global maximum and minimum.

Theorem 4.21 (Extreme Value Theorem)**.** *Let I be a closed interval and let $f : I \to \mathbb{R}$ be a continuous function. Then there exists $m, M \in I$ such that*

$$f(m) = \inf f(I), \quad \textit{and} \quad f(M) = \sup f(I).$$

As the example immediately before this statement shows, it is crucial that I is a closed interval in the statement of Theorem 4.21

Proof. We will only show that there exists $M \in I$ such that

$$f(M) = \sup\{f(x) : x \in I\}.$$

The task of showing that f achieves its infimum can be handled similarly.

Recall that

$$f(I) = \{y \in \mathbb{R} : \exists\, x \in I,\ f(x) = y\}.$$

Case 1 - Suppose that $f(I)$ is bounded from above. Then $\sup f(I) = T \in \mathbb{R}$ and by Definition 1.37, we have that

$$\forall\, \epsilon > 0,\ \exists\, y \in f(I)\ :\ y > T - \epsilon.$$

We can use this to construct a sequence (y_n) in $f(I)$ such that $y_n \to M$. This implies that there exist (x_n) in I such that $f(x_n) = y_n$.

It follows from the Bolzano-Weierstrass Theorem that (x_n) has a convergent subsequence $(x_{n_k})_{k\in\mathbb{N}}$, with

$$\lim_{k\to\infty} x_{n_k} = x_0.$$

Since the set of accumulation points for the interval I is also I, it must be the case that $x_0 \in I$. The definition of continuity then implies that

$$\lim_{k\to\infty} y_{n_k} = \lim_{k\to\infty} f(x_{n_k}) = f(x_0).$$

On the other hand, since $y_n \to M$, and so it follows that $y_{n_k} \to M$ (see the exercise on page 130). We conclude that

$$T = \lim_{k\to\infty} y_{n_k} = \lim_{k\to\infty} f(x_{n_k}) = f(x_0).$$

That is, $T \in f(I)$, as required.

Case 2 - Suppose that $f(I)$ is unbounded. This case is left as an exercise. Hint: similarly to the proof of case 1 above, start by showing that, under this assumption, there is a sequence y_n in $f(I)$ such that $y_n \to \infty$. Then repeat the proof from case 1 until you reach a contradiction, thus showing that this case cannot exist. □

As we mentioned earlier, Theorem 4.21 is not valid when I is an open interval. I encourage you to consider the following question: which part of the proof above breaks down in the case when I is an open interval?

By combining the Intermediate Value Theorem and the Extreme Value Theorem, we obtain a nice corollary.

Corollary 4.22. *Let I be a closed interval and let $f : I \to \mathbb{R}$ be a continuous function. Then $f(I) = [\min_{x\in I} f(x), \max_{x\in I} f(x)]$.*

In particular, **continuous functions send closed intervals to closed intervals**.

Proof. Write $m = \inf\{f(x) : x \in I\}$ and $M = \sup\{f(x) : x \in I\}$. By the Extreme Value Theorem, there exist $t_1, t_2 \in I$ such that

$$f(t_1) = m \text{ and } f(t_2) = M.$$

Now, apply the Intermediate Value Function to the function f restricted to the closed interval $[\min\{t_1, t_2\}, \max\{t_1, t_2\}]$. It follows that, for all $y \in [m, M]$, there exists $x \in [\min\{t_1, t_2\}, \max\{t_1, t_2\}] \subset I$ such that $f(x) = y$. □

We will now give another application of the Intermediate Value Theorem, which roughly states that a function on an interval that is both continuous and bijective cannot change direction. To state this formally, we need the definition of a monotone function.

Definition 4.23. *Let I be an interval and let $f : I \to \mathbb{R}$. We call f **increasing** if*

$$\forall x_1, x_2 \in I,\, x_1 < x_2 \implies f(x_1) < f(x_2).$$

*We call f **decreasing** if*

$$\forall x_1, x_2 \in I,\, x_1 < x_2 \implies f(x_1) > f(x_2).$$

If we replace we have the weaker condition that

$$\forall x_1, x_2 \in I,\, x_1 < x_2 \implies f(x_1) \leq f(x_2).$$

*'$<$' by '$\leq$' or '$>$' by '$\geq$', then we say f is **non-decreasing**. Similarly, if we replace '$>$' by '$\geq$' we say that f is **non-increasing**.*

*A function is said to be **monotone** if it is either increasing, non-decreasing, decreasing or non-increasing.*

*A function is said to be **strictly monotone** if it is either increasing or decreasing.*

We sometime use the terms **strictly increasing** for increasing, in order to further emphasise that the inequality $f(x_1) < f(x_2)$ is strict. Similarly, we sometimes use the term **strictly decreasing** for decreasing.

Our next major goal is to prove the following result.

Theorem 4.24. *Let I be an interval, let $D \subset \mathbb{R}$ and suppose that $f : I \to D$ is continuous and bijective. Then f is strictly monotone.*

To prove Theorem 4.24, we will make use of the following lemma, which we sometimes call the **chain condition**.

Lemma 4.25. *Let I be an interval, let $D \subset \mathbb{R}$ and suppose that $f : I \to D$ is continuous and bijective. Suppose that $a, b, c \in I$ satisfy $a < b < c$. Then either*

$$f(a) < f(b) < f(c) \tag{93}$$

or

$$f(a) > f(b) > f(c) \tag{94}$$

hold.

Proof. Suppose for a contradiction that neither of (93) or (94) occur. Then either

$$f(b) > f(a), f(c) \tag{95}$$

or

$$f(b) < f(a), f(c) \tag{96}$$

hold. We will just deal with the situation of (95). The case (96) can be handled similarly.

Suppose that (95) occurs. Then there exists some y such that

$$f(a), f(c) < y < f(b). \tag{97}$$

Apply the Intermediate Value Theorem for the function f restricted to the domain $[a, b]$. It follows that there exists $x_1 \in [a, b]$ such that $f(x_1) = y$. Moreover, we see that $x_1 \in (a, b)$, since $f(a) = y$ or $f(b) = y$ would contradict (97).

Similarly, by applying the Intermediate Value Theorem for the function f restricted to the domain $[b, c]$, it follows that there is some $x_2 \in (b, c)$ such that $f(x_2) = y$. But then $f(x_2) = f(x_1)$. Also, since $x_1 \in (a, b)$ and $x_2 \in (b, c)$, it follows that $x_1 \neq x_2$. This contradicts the assumption that f is bijective. □

We now move to the proof of Theorem 4.24.

Proof of Theorem 4.24. Fix two arbitrary elements $a_0, b_0 \in I$ such that $a_0 < b_0$. We make the additional assumption that $f(a_0) < f(b_0)$. We will show that f is strictly increasing. (For the case when $f(a_0) > f(b_0)$, we can use essentially the same argument to show that f is strictly decreasing.)

Claim. 1. *If $x \in I$ and $x < a_0$ then $f(x) < f(a_0)$.*

2. *If $x \in I$ and $x > a_0$ then $f(x) > f(a_0)$.*

Proof of Claim. 1. Suppose that $x < a_0$, and so we have. Then

$$x < a_0 < b_0.$$

It follows from Lemma 4.25, and the assumption that $f(b_0) > f(a_0)$, that

$$f(x) < f(a_0) < f(b_0).$$

In particular, this proves the intended inequality $f(x) < f(a_0)$.

2. Suppose that $x > a_0$ and $f(x) < f(a_0)$. If $x = b_0$ then we have a contradiction, as $f(x) = f(b_0) > f(a_0)$. There are then two cases two consider, according to whether $x < b_0$ or $x > b_0$.

   Case A - Suppose that $x < b_0$, and so
   $$a_0 < x < b_0.$$
   Lemma 4.25, and the assumption that $f(b_0) > f(a_0)$, imply that
   $$f(a_0) < f(x) < f(b_0).$$
   In particular, this proves the intended inequality $f(x) > f(a_0)$.

   Case B - Suppose that $x > b_0$, and so
   $$a_0 < b_0 < x.$$
   Lemma 4.25, and the assumption that $f(b_0) > f(a_0)$, imply that
   $$f(a_0) < f(b_0) < f(x).$$
   In particular, this proves the intended inequality $f(x) > f(a_0)$.

□

We are now ready to complete the proof of Theorem 4.24. Suppose that $x_1, x_2 \in I$ satisfy $x_1 < x_2$. We need to show that
$$f(x_1) < f(x_2). \tag{98}$$
Note first of all that if either of x_1 or x_2 is equal to a_0 then (98) immediately follows from the previous claim. We can therefore henceforth assume that $x_1, x_2 \neq a_0$.

Case 1 - Suppose that $x_1 < x_2 < a_0$. By the first point of Lemma 4.25, we have
$$f(x_1), f(x_2) < f(a_0).$$
Suppose for a contradiction that $f(x_1) > f(x_2)$ (note that $f(x_1) = f(x_2)$ is not possible because of the assumption that f is a bijection). Then we have
$$f(x_2) < f(x_1) < f(a_0)$$
which contradicts Lemma 4.25.

Case 2 - Suppose that $x_1 < a_0 < x_2$. Then parts 1 and 2 of the previous claim respectively imply that
$$f(x_1) < f(a_0)$$
and
$$f(x_2) > f(a_0).$$
Combining these two inequalities gives the intended bound (98).

Case 3 - Suppose that $a_0 < x_1 < x_2$. Then part 2 of the claim implies that
$$f(x_1), f(x_2) > f(a_0).$$

Suppose for a contradiction that $f(x_1) > f(x_2)$. Then

$$f(x_1) > f(x_2) > f(a_0).$$

This contradicts Lemma 4.25, which completes the proof.

□

We now have enough theory to give a proof of Theorem 4.5. To simplify things slightly, we will prove a slightly modified statement where we assume that the domain of f is an open interval. For convenience, the result is below.

Theorem 4.26. *Let I be an open interval and let $D \subset \mathbb{R}$. Let $f : I \to D$ be an invertible function. If f is continuous on I, then the inverse function f^{-1} is continuous on D.*

Proof. By Theorem 4.24, f is strictly monotone. We assume that f is strictly increasing (the case when f is strictly decreasing can be handled simlarly).

Claim. *$f^{-1} : D \to I$ is strictly increasing.*

Proof. Let $x, y \in D$ with $x < y$ and suppose for a contradiction that $f^{-1}(x) \geq f^{-1}(y)$. Then, since f is increasing, it follows that

$$x = f(f^{-1}(x)) \geq f(f^{-1}(y)) = y,$$

contradicting the assumption that $x < y$. □

Now let $b \in D$ be arbitrary, and we will show that f^{-1} is continuous at b. So, for all $\epsilon > 0$, we need to show that there exists $\delta > 0$ such that

$$y \in D \text{ and } |y - b| < \delta \implies |f^{-1}(y) - f^{-1}(b)| < \epsilon.$$

Let $a = f^{-1}(b) \in I$. Since I is an open, a is not an endpoint of I, and so there exists ϵ' such that $0 < \epsilon' < \frac{\epsilon}{2}$ and

$$(a - \epsilon', a + \epsilon') \subset I.$$

Write

$$b_1 := f(a - \epsilon'), \quad b_2 = f(a + \epsilon').$$

Since f is strictly increasing,

$$b_1 < b < b_2.$$

Set $\delta := \min\{b - b_1, b_2 - b\}$. Then, making use of the claim, we conclude that

$$\begin{aligned} |y - b| < \delta \implies b_1 < y < b_2 &\implies f^{-1}(b_1) < f^{-1}(y) < f^{-1}(b_2) \\ &\implies a - \epsilon' < f^{-1}(y), f^{-1}(b) < a + \epsilon' \\ &\implies |f^{-1}(y) - f^{-1}(b)| < 2\epsilon' < \epsilon. \end{aligned}$$

□

4.6 Uniform continuity

In this section, we consider the following strengthened form of continuity.

Definition 4.27. *Let $D \subset \mathbb{R}$ and let $f : D \to \mathbb{R}$. We say that f is* ***uniformly continuous*** *if, for all sequences $(x_n), (y_n)$ in D with $\lim_{n\to\infty} |x_n - y_n| = 0$, we have that*

$$\lim_{n\to\infty} |f(x_n) - f(y_n)| = 0.$$

Example - Let $a, b \in \mathbb{R}$ be arbitrary and consider the linear function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = ax + b$. This function is uniformly continuous. Indeed, for any sequences $(x_n), (y_n)$ in $\mathbb{R}$ with $\lim_{n\to\infty} |x_n - y_n| = 0$, we have that

$$\lim_{n\to\infty} |f(x_n) - f(y_n)| = \lim_{n\to\to} |ax_n + b - (ay_n + b)| = |a| \lim_{n\to\infty} |x_n - y_n| = 0.$$

Let's first verify that a function being uniformly continuous does indeed imply that it is continuous.

Theorem 4.28. *Let $D \subset \mathbb{R}$ and let $f : D \to \mathbb{R}$ be uniformly continuous. Then f is continuous.*

Proof. Let $x_0 \in D$ be arbitrary and let (x_n) be a sequence in D such that $x_n \to x_0$. We need to show that $f(x_n) \to f(x_0)$.

Indeed, applying the definition of uniform continuity with $y_n = x_0$ for all n, we see that $\lim_{n\to\infty} |x_n - x_0| = 0$ and hence

$$\lim_{n\to\infty} |f(x_n) - f(x_0)| = 0.$$

This is equivalent to the statement that $\lim_{n\to\infty} f(x_n) = f(x_0)$. □

Example - The purpose of this example is to show that uniform continuity is a strictly stronger property than continuity. That is, we will give an example of a function that is continuous but not uniformly continuous.

Indeed, let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^2$. We have seen earlier in this chapter that f is continuous. To show that it is not uniformly continuous, consider the sequences (x_n) and (y_n) defined by

$$x_n := n + \frac{1}{n}, \quad y_n := n.$$

Then,

$$\lim_{n\to\infty} |x_n - y_n| = \lim_{n\to\infty} \frac{1}{n} = 0.$$

However,

$$\lim_{n\to\infty} |f(x_n) - f(y_n)| = \lim_{n\to\infty} \left| \left(n + \frac{1}{n}\right)^2 - n^2 \right| = \lim_{n\to\infty} 2 + \frac{1}{n^2} = 2 \neq 0.$$

This shows that f is not uniformly continuous.

Another way to see the difference between continuity and uniform continuity is to consider the following statement and compare it with Theorem 4.2.

Theorem 4.29. *Let $D \subset \mathbb{R}$ and $f : D \to \mathbb{R}$. Then, f is uniformly continuous if and only if,*

$$\forall\, \epsilon > 0,\, \exists \delta > 0 \,:\, |x-y| < \delta \text{ and } x, y \in D \implies |f(x) - f(y)| < \epsilon. \tag{99}$$

By contrast, we can use Theorem 4.2 (changing the notation x_0 to y for a convenient comparison) to state that f is continuous if and only if

$$\forall\, y \in D\,, \forall\, \epsilon > 0,\, \exists \delta > 0 \,:\, |x-y| < \delta \text{ and } x \in D \implies |f(x) - f(y)| < \epsilon. \tag{100}$$

The difference between (99) and (100) here is in the order of the introduction of the variables. In the statement (100), the variable y is introduced earlier, which allows for the possibility that δ depends on y, whereas in the statement (99), the parameter δ comes first, and so is independent of y.

Proof of Theorem 4.29. We omit the proof of this statement for the purpose of saving some time. For those interested to read the full proof, see page 154 of Mario Ullrich's notes from the previos version of this course. □

Exercise - Use Theorem 4.29 to prove that the absolute value function is uniformly continuous.

Uniform continuity will be an important tool later in this course, but it is not always straightforward to show that a given function is uniformly continuous. However, the following theorem shows that, if a continuous function is defined on a closed interval, then it is guaranteed to be uniformly continuous.

Theorem 4.30. *Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Then f is uniformly continuous.*

In particular, by combining this result with Theorem 4.28, we see that, for $f : [a, b] \to \mathbb{R}$,

$$f \text{ is continuous} \iff f \text{ is uniformly continuous}\,.$$

Proof. Let f be continuous and suppose for a contradiction that f is not uniformly continuous. By the definition of uniform continuity, this means that there exist sequences (x_n) and (y_n) in $[a, b]$ with $\lim_{n\to\infty} |x_n - y_n| = 0$ and some $\epsilon > 0$ such that,

$$|f(x_n) - f(y_n)| \geq \epsilon \tag{101}$$

holds for infinitely many $n \in \mathbb{N}$. Let J be the set of all $n \in \mathbb{N}$ such that (101) holds and consider the subsequence $(x_n)_{n\in J}$. This is a bounded sequence, and so the Bolzano-Weierstrass Theorem implies that there exists a further subsequence $(x_{n_k})_{k\in\mathbb{N}}$ which is convergent. Write

$$x_0 := \lim_{k\to\infty} x_{n_k}.$$

It is important to note that $x_0 \in [a, b]$. Indeed, if this were not true, we would have a sequence in $[a, b]$ which never takes the value x_0 and converges to x_0, which implies that x_0 is an accumulation point of $[a, b]$. But the accumulation points of a closed interval are the same as the interval itself, which is a contradiction. This is the only point in the proof where we use the fact that our domain is a closed interval.

Note that the way that these nested subsequences are defined ensures that

$$|f(x_{n_k}) - f(y_{n_k})| \geq \epsilon \tag{102}$$

holds for all $k \in \mathbb{N}$.

It follows from the triangle inequality that, for all $k \in \mathbb{N}$,

$$|x_0 - y_{n_k}| \leq |x_0 - x_{n_k}| + |x_{n_k} - y_{n_k}|.$$

Since both of the absolute values on the right side of this inequality tend to zero, it follows from the Sandwich Rule that

$$\lim_{k\to\infty} |x_0 - y_{n_k}| = 0.$$

We now use the definition of continuity and the fact that f and the absolute value function are continuous at $x_0 \in [a, b]$ to conclude that

$$\begin{aligned} \lim_{k\to\infty} |f(x_{n_k}) - f(y_{n_k})| = \left| \lim_{k\to\infty} (f(x_{n_k}) - f(y_{n_k})) \right| &= \left| f\left(\lim_{k\to\infty} x_{n_k} \right) - f\left(\lim_{k\to\infty} y_{n_k} \right) \right| \\ &= |f(x_0) - f(x_0)| = 0 \end{aligned}$$

However, we know that (102) holds for all $k \in \mathbb{N}$, which contradicts the claim that

$$\lim_{k\to\infty} |f(x_{n_k}) - f(y_{n_k})| = 0.$$

This gives us the intended contradiction and completes the proof. □

Example - We saw earlier that the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ is not uniformly continuous. However, if we restrict the domain to a closed interval $[a, b]$, the function is continuous, and it is therefore also uniformly continuous by Theorem 4.30.

5 Differential calculus

5.1 Definition and basic examples

In this chapter we want to introduce and study derivatives of real-valued functions. For functions defined on the real line, as will be the case throughout this chapter, one may think of the derivative as the slope of the tangent line attached to a point of the graph of the function. Similarly to our "definition" of continuity in terms of drawing the function without taking the pen off the page, this is a useful intuition that helps one to understand the derivative. However, we also need a precise definition of a derivative to handle cases where a visualisation is not possible or helpful.

We will use derivatives to identify minima and maxima of a function, and to determine whether a function is increasing or decreasing at a given point. We will use derivatives to help us calculate complicated limits, via l'Hospital's rule. Finally, we will study higher derivatives and use them to approximate functions by polynomials, using the Taylor polynomial.

Let us begin with a precise definition of a derivative of a function.

Definition 5.1. *Let $I = (a, b)$, $x_0 \in I$, and $f : I \to \mathbb{R}$. We call f* ***differentiable at*** *x_0 if*

$$\lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0}. \tag{103}$$

exists and is a real number. In this case we call this limit the ***derivative of*** *f* ***at*** *x_0, and denote this limit as $f'(x_0)$ (or $\frac{d}{dx} f(x_0)$, or $\frac{df}{dx}(x_0)$).*

If the limit is not defined at x_0, then f is ***not differentiable*** *at x_0.*

If f is differentiable at every point of its domain I, we call f ***differentiable****. The function $f' : I \to \mathbb{R}$ is called the* ***derivative*** *of f.*

Here are some important remarks about this definition.

- Note that we give two different and equivalent formulations of the limit in (103). It is important to understand why these two limits are the same. This is a special case of the fact that
$$\lim_{x\to x_0} f(x) = \lim_{h\to 0} f(x_0 + h). \tag{104}$$
I encourage you to formally check that (104) is indeed always true (one should also check that, if one side of (104) is not defined, then neither is the other side).

- In most calculations, we will use the first representation of the derivative, that is
$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$
This is slightly easier to work with in practice. Since we often view the derivative as a function, we often make a small modification to the notation here, writing
$$f'(x) = \lim_{h\to 0} \frac{f(x + h) - f(x)}{h}.$$

The second representation of the derivative, namely

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

is typically more useful in the context of proofs.

- The quantity $\frac{f(x+h)-f(x)}{h}$ is important enough to be given a name; we call it the **difference quotient**.
- We restrict our attention to functions that are defined on an open interval $I = (a, b)$. This is because the endpoints of an interval sometimes need more care. However, we could also replace I with $\mathbb{R}$ in the above definition, since this domain also has no endpoints. For the rest of this chapter, we will use the term **open interval** to decribe any continuous subset of $\mathbb{R}$ which does not contain an endpoint. In particular, an open interval can take any of the following forms:

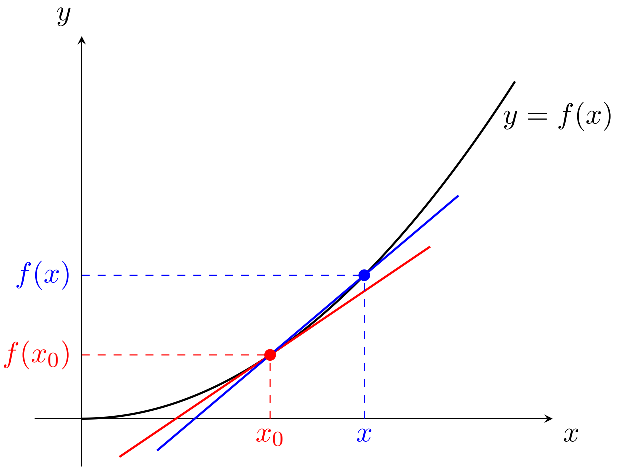
$$(a, b),\ (a, \infty),\ (-\infty, b),\ (-\infty, \infty).$$

- We could also give a version of this definition for a more general domain D, but we skip this extra detail in order to keep things grounded and reduce technicalities.

To see how the derivative is related to the slope, we will use the second representation of the derivative from (103). Consider the quantity

$$\frac{f(x) - f(x_0)}{x - x_0}.$$

Geometrically, this measures the slope between the points $(x_0, f(x_0))$ and $(x, f(x))$. Now consider what happens as x gets closer and closer to x_0. We see that this quantity estimates the direction of the function at x.



Let us now consider a few basic examples of the derivative, making use of the definition.

Example - Let $c \in \mathbb{R}$ be arbitrary and consider the constant function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = c$ for all $x \in I$. Then, for all $x \in \mathbb{R}$,

$$f'(x) = 0.$$

Indeed,

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}=\lim_{h\to 0}\frac{c-c}{h}=\lim_{h\to 0}0=0.$$

Example - Let $n\in\mathbb{N}$ and let $f(x)=x^n$. Then f is differentiable at every $x\in\mathbb{R}$, and

$$f'(x)=nx^{n-1}.$$

Indeed, for an arbitrary (fixed) $x\in\mathbb{R}$ we use the binomial theorem to calculate that

$$\begin{aligned}\frac{f(x+h)-f(x)}{h}&=\frac{(x+h)^n-x^n}{h}\\&=\binom{n}{1}x^{n-1}+\binom{n}{2}x^{n-2}h+\cdots+\binom{n}{n-1}xh^{n-2}+h^{n-1}.\end{aligned}$$

Computing the limit as h goes to zero, we see that

$$\begin{aligned}\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}&=\lim_{h\to 0}\left[\binom{n}{1}x^{n-1}+\binom{n}{2}x^{n-2}h+\cdots+\binom{n}{n-1}xh^{n-2}+h^{n-1}\right]\\&=nx^{n-1}.\end{aligned}$$

Example - Let $n\in\mathbb{N}$ and let $f:(0,\infty)\to\mathbb{R}$ be given by $f(x)=\frac{1}{x^n}$. Then f is differentiable and

$$f'(x)=\frac{-n}{x^{n+1}}$$

for all $x\in(0,\infty)$.

Indeed, similarly to the previous example, we use the binomial theorem to calculate that

$$\begin{aligned}\frac{f(x+h)-f(x)}{h}&=\frac{1}{h}\left(\frac{1}{(x+h)^n}-\frac{1}{x^n}\right)\\&=\frac{1}{h}\left(\frac{x^n-(x+h)^n}{(x+h)^nx^n}\right)\\&=\frac{-1}{(x+h)^nx^n}\left(\frac{(x+h)^n-x^n}{h}\right)\\&=\frac{-1}{(x+h)^nx^n}\left(\binom{n}{1}x^{n-1}+\binom{n}{2}x^{n-2}h+\cdots+\binom{n}{n-1}xh^{n-2}+h^{n-1}\right).\end{aligned}$$

Therefore, taking limits and applying Theorem 4.14, we conclude that

$$\begin{aligned}\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}&=\lim_{h\to 0}\frac{-1}{(x+h)^nx^n}\\&\quad\cdot\lim_{h\to 0}\left[\binom{n}{1}x^{n-1}+\binom{n}{2}x^{n-2}h+\cdots+\binom{n}{n-1}xh^{n-2}+h^{n-1}\right]\\&=\frac{-1}{x^{2n}}\cdot nx^{n-1}=\frac{-n}{x^{n+1}}.\end{aligned}$$

Note that, combining the previous two examples, we see that for all $n\in\mathbb{Z}\setminus\{0\}$ the function $f:(0,\infty)\to\mathbb{R}$ given by $f(x)=x^n$ is differentiable with derivative

$$f'(x)=nx^{n-1}.$$

Example - Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = \sin x$. Then f is differentiable and

$$f'(x) = \cos x$$

for all $x \in \mathbb{R}$.

To prove that this is indeed true, we will use the trigonometric identity (this is the "product-to-sum" formulae, which can be derived from the angle sum formulae for the sine function)

$$\sin x - \sin y = 2\cos\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right).$$

It follows that

$$\frac{f(x+h)-f(x)}{h} = \frac{\sin(x+h)-\sin x}{h} = \frac{2\cos\left(x+\frac{h}{2}\right)\sin\left(\frac{h}{2}\right)}{h} = \cos\left(x+\frac{h}{2}\right)\cdot\frac{\sin\left(\frac{h}{2}\right)}{\frac{h}{2}}.$$

Taking limits of both sides, using the fact that $\lim_{t\to 0}\frac{\sin t}{t} = 1$ (see Exercise 43 on Sheet 8), and applying Theorem 4.14, we obtain

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \lim_{h\to 0}\cos\left(x+\frac{h}{2}\right)\cdot\frac{\sin\left(\frac{h}{2}\right)}{\frac{h}{2}} = \lim_{h\to 0}\cos\left(x+\frac{h}{2}\right).$$

Finally, since cos is continuous, it follows that

$$\lim_{h\to 0}\cos\left(x+\frac{h}{2}\right) = \cos x.$$

Exercise - Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = \cos x$. Show that f is differentiable and

$$f'(x) = -\sin x.$$

The next differentiable function that we will consider is the exponential function $f(x) = e^x$. We will use the following representation of the exponential function without proof (see page 140 of Mario Ullrich's lecture notes for a sketch of a proof of this):

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n.$$

Before calculating the derivative of e^x, we first need the following inequality, which is also of independent interest.

Lemma 5.2. *For all $x < 1$,*

$$1 + x \leq e^x \leq \frac{1}{1-x}.$$

In addition, the lower bound $e^x \geq 1 + x$ holds for all $x \in \mathbb{R}$.

Proof. Fix $x < 1$. Recall that Bernoulli's inequality (Theorem 1.30) states that, for all $y \geq -1$ and all $n \in \mathbb{N}$,

$$(1+y)^n \geq 1 + ny.$$

The lower bound - Apply Bernoulli's inequality with $y = \frac{x}{n}$, which is valid for all $n \geq -x$. In particular, for all sufficiently large $n \in \mathbb{N}$, we have

$$\left(1 + \frac{x}{n}\right)^n \geq 1 + x.$$

Therefore,

$$e = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n \geq 1 + x.$$

The upper bound - Apply Bernoulli's inequality with $y = \frac{-x}{n}$. Since $x < 1$, this application of Bernoulli's inequality is valid for all $n \in \mathbb{N}$. It follows that

$$\left(1 + \frac{(-x)}{n}\right)^n \geq 1 - x,$$

and so

$$\frac{1}{e^x} = e^{-x} \geq 1 - x.$$

Since $x < 1$, this inequality can be rearranged to give

$$\frac{1}{1-x} \geq e^x.$$

□

With this, we can compute the derivative of e^x.

Theorem 5.3. *Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = e^x$. Then f is differentiable and $f'(x) = e^x$ for all $x \in \mathbb{R}$.*

Proof. We compute the difference quotient

$$\frac{f(x+h) - f(x)}{h} = \frac{e^{x+h} - e^x}{h} = e^x \cdot \frac{e^h - 1}{h}. \tag{105}$$

By Lemma 5.2, we have

$$1 + h \leq e^h \leq \frac{1}{1-h},$$

for all $h \in \mathbb{R}$ with $|h| < 1$. This implies that

$$\frac{1}{1+|h|} \leq \frac{e^h - 1}{h} \leq \frac{1}{1-|h|}$$

holds for any h such that $|h| < 1$. (To check this properly, you should do a case analysis, considering the possibility that h is positive or negative.) It then follows from the Sandwich Rule that

$$\lim_{h\to 0} \frac{e^h - 1}{h} = 1.$$

Finally, recalling (105), we conclude that

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} e^x \cdot \frac{e^h - 1}{h} = e^x.$$

□

As part of this proof, we established that

$$\lim_{h\to 0}\frac{e^h-1}{h}=1. \tag{106}$$

This is an interesting example of an *indeterminate limit* where both the numerator and denominator tend to zero. Later in this chapter we will introduce l'Hospital's rule, which is gives a convenient way to prove (106) and other similarly tricky limits.

Example - We will now consider an example of a function that is not differentiable. Let $f:\mathbb{R}\to\mathbb{R}$ be given by $f(x)=|x|$. In particular, we will show that this function is not differentiable at 0.

To achieve this, we need to show that the limit

$$\lim_{h\to 0}\frac{f(0+h)-f(0)}{h}=\lim_{h\to 0}\frac{|h|}{h}$$

is not defined. This is a situation that we should be familiar with from our work in the previous chapter. Our task is to find two sequences (x_n) and (y_n) which converge to 0 such that

$$\lim_{n\to\infty}\frac{x_n}{|x_n|}$$

and

$$\lim_{n\to\infty}\frac{y_n}{|y_n|}$$

are not the same. We can simply set $x_n=\frac{1}{n}$ and $y_n=\frac{-1}{n}$ and see that

$$\lim_{n\to\infty}\frac{x_n}{|x_n|}=\lim_{n\to\infty}1=1$$

and

$$\lim_{n\to\infty}\frac{y_n}{|y_n|}=\lim_{n\to\infty}-1=-1.$$

Now that we have formally proven that the absolute value function is not differentiable at 0, let us consider *why* the definition fails for this function. If we look at the graph of $f(x)=|x|$, we see that there is a sharp point at $x=0$, where the direction of the curve changes suddenly. Roughly speaking, for a function to be differentiable at a point, it must be the case that the direction close to that point varies smoothly.

The previous example shows that there exist functions which are continuous but not differentiable. In other words, the implication

$$f \text{ is continuous} \implies f \text{ is differentiable}$$

is false. However, the reverse implication is valid, as the next result shows.

Theorem 5.4. *Let $x_0\in(a,b)$ and let $f:(a,b)\to\mathbb{R}$ be differentiable at x_0. Then f is continuous at x_0.*

Proof. Since f is differentiable at x_0, the limit

$$\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}$$

exists and is a real number. To show that f is continuous at x_0, we need to show that $\lim_{x\to x_0} f(x) = f(x_0)$. This is the same thing as showing that

$$\lim_{x\to x_0} f(x) - f(x_0) = 0. \tag{107}$$

However,

$$\begin{aligned}
\lim_{x\to x_0} f(x) - f(x_0) &= \lim_{x\to x_0} f(x) - f(x_0) \cdot \frac{x - x_0}{x - x_0} \\
&= \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} \cdot (x - x_0) \\
&= \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} \cdot \lim_{x\to x_0} (x - x_0).
\end{aligned}$$

In the last inequality above we have used Theorem 4.14 (this is where we make use of the fact that the limit $\lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}$ exists and is a real number). Therefore,

$$\lim_{x\to x_0} f(x) - f(x_0) = f'(x_0) \cdot 0 = 0,$$

which proves the intended equation (107) □

5.2 Calculation rules for differentiable functions

This chapter will follow a similar pattern to the previous two chapters. We have already begun establishing derivatives for some simple and important functions in the previous section, and we can view these derivatives as "building blocks". In this section, we will develop some basic rules for calculating derivatives for more complicated function which are built from these building blocks. In this way, we will be able to compute derivatives of sums, products and compositions of the simple functions we have already discussed. So, for example, the tools in this section will allow us to calculate the derivative of the function

$$f(x) = \cos(x^2 - 3\sqrt{x}) \cdot e^{(\sin x)^2}.$$

We begin with the most simple calculation rule for derivatives, which concerns the derivative of the sum of two functions.

Theorem 5.5. *Let I be an interval and let $f, g : I \to \mathbb{R}$ be functions that are differentiable at $x_0 \in I$. Then $(f+g)'(x_0)$ exists and*

$$(f+g)'(x_0) = f'(x_0) + g'(x_0).$$

Also, for any $c \in \mathbb{R}$, we have

$$(c \cdot f)'(x_0) = c \cdot f'(x_0).$$

Proof. The proof follows from the analogous calculation rules for limits, and it left for the student to verify. □

Example - A simple application of Theorem 5.5 gives the derivative of a polynomial

$$f(x) = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0.$$

We already know that the derivative of x^k is kx^{k-1}, and therefore

$$f'(x) = c_n n x^{n-1} + c_{n-1}(n-1)x^{n-2} + \cdots + c_2 2x + c_1.$$

The calculation rule for products is slightly less straighforward.

Theorem 5.6 (Product Rule)**.** *Let I be an interval and let $f, g : I \to \mathbb{R}$ be differentiable at $x_0 \in I$. Then $(fg)'(x_0)$ exists and*

$$(fg)'(x_0) = f(x_0)g'(x_0) + f'(x_0)g(x_0).$$

In short $(fg)' = f'g + fg'$.

Proof. Note that

$$\begin{aligned}(fg)'(x) &= \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} \\ &= \lim_{h\to 0} f(x+h)\frac{g(x+h) - g(x)}{h} + \lim_{h\to 0} g(x)\frac{f(x+h) - f(x)}{h},\end{aligned}$$

where we have simply "added zero" (in the form $0 = -f(x+h)g(x) + f(x+h)g(x)$) in the latter equality. Since the derivatives $f'(x_0)$ and $g'(x_0)$ are defined and real, and also f is continuous at x_0 (by Theorem 5.4) we can apply Theorem 4.14 to conclude that

$$(fg)'(x) = \lim_{h\to 0} f(x+h)\frac{g(x+h)-g(x)}{h} + \lim_{h\to 0} g(x)\frac{f(x+h)-f(x)}{h} = f(x)g'(x) + g(x)f'(x),$$

as required. □

Example - Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = x^2 \sin x.$$

Then, by the product rule, this function is differentiable and

$$f'(x) = (2x) \cdot \sin x + (x^2) \cdot (\cos x) = x(2\sin x + x\cos x).$$

The **chain rule** is crucial for dealing with the composition of two functions.

Theorem 5.7 (Chain Rule)**.** *Let I amd J be open intervals and let $f : I \to J$ and $g : J \to \mathbb{R}$ be functions with f differentiable at $x_0 \in I$ and g differentiable at $f(x_0) \in J$. Then $(g\circ f)'(x_0)$ exists and*

$$(g \circ f)'(x_0) = g'(f(x_0)) \cdot f'(x_0).$$

Proof. Set $y_0 := f(x_0)$ and define a function $h : J \to \mathbb{R}$ by

$$h(y) = \begin{cases} \frac{g(y)-g(y_0)}{y-y_0} & \text{if } y \neq y_0 \\ g'(y_0) & \text{if } y = y_0 \end{cases}.$$

By the definition of the derivative $g'(y_0)$, h is continuous at y_0, and therefore, by Theorem 4.8, $h \circ f$ is continuous at x_0. Also,

$$g(y) - g(y_0) = h(y)(y - y_0) \tag{108}$$

holds for all $y \in J$ (this can be checked by a quick case analysis). Therefore, applying (108) with $y = f(x)$, we see that

$$\begin{aligned} \lim_{x\to x_0} \frac{g(f(x)) - g(f(x_0))}{x - x_0} &= \lim_{x\to x_0} \frac{h(f(x))(f(x) - f(x_0))}{x - x_0} \\ &= \lim_{x\to x_0} h(f(x)) \cdot \lim_{x\to x_0} \frac{(f(x) - f(x_0))}{x - x_0} \\ &= h(y_0) \cdot f'(x_0) \\ &= g'(y_0)f'(x_0) = g'(f(x_0)) \cdot f'(x_0). \end{aligned}$$

In the second/third equality above, we used the fact that $h \circ f$ is continuous at x_0.

□

Example - Consider the function $f : \left(\frac{-\pi}{2}, \frac{-\pi}{2}\right) \to \mathbb{R}$ given by $f(x) = \tan x$. Since

$$f(x) = \sin x \cdot \frac{1}{\cos x}$$

we can use the product rule to compute that

$$f'(x) = \cos x \cdot g(x) + \sin x \cdot g'(x)$$

where

$$g(x) = \frac{1}{\cos x}.$$

This can then be rewritten as

$$f'(x) = 1 + \sin x \cdot g'(x) \tag{109}$$

It remains to calculate the derivative of g. Since g is a composition of two functions, we will use the chain rule for this. Indeed,

$$g(x) = (j \circ h)(x)$$

where

$$h(x) = \cos x, \quad j(x) = \frac{1}{x}.$$

Note that the conditions of Theorem 5.7 are satisfied, since h is differentiable on $\left(\frac{-\pi}{2}, \frac{-\pi}{2}\right)$ and j is differentiable at $h(x) > 0$. Therefore,

$$g'(x) = j'(h(x)) \cdot h'(x) = -(\cos x)^{-2} \cdot (-\sin(x)) = \frac{\sin x}{(\cos x)^2}.$$

Inserting this expression into (109) yields

$$f'(x) = 1 + \sin x \cdot g'(x) = 1 + \frac{(\sin x)^2}{(\cos x)^2} = 1 + (\tan x)^2.$$

Many textbooks and lecture notes on differentiation present the **quotient rule** as a useful tool for this kind of situation, where we need to compute $(f/g)'$. This is essentially just the product rule applied where one of the products is a function involving a reciprocal. Personally, I do not like using this rule, as it is one more thing to remember. But, in case you prefer, here is the rule in short:

$$\left(\frac{f}{g}\right) = \frac{f'g - fg'}{g^2}.$$

Example - Let us once again consider the function we mentioned at the beginning of the chapter, that is $f : (0, \infty) \to \mathbb{R}$ given by

$$f(x) = \cos(x^2 - 3\sqrt{x}) \cdot e^{(\sin x)^2}.$$

We can write this as

$$f(x) = (g_1 \circ g_2)(x) \cdot (g_3 \circ g_4 \circ g_5)(x),$$

where $g_1, \dots, g_5 : (0, \infty) \to \mathbb{R}$ are defined by

$$g_1(x) = \cos x, \;\; g_2(x) = x^2 - 3\sqrt{x}, \;\; g_3(x) = e^x, \;\; g_4(x) = x^2, \;\; g_5(x) = \sin x.$$

By the product rule and the chain rule, we calculate that

$$\begin{aligned}
f'(x) &= (g_1 \circ g_2)'(x) \cdot (g_3 \circ g_4 \circ g_5)(x) + (g_1 \circ g_2)(x) \cdot (g_3 \circ g_4 \circ g_5)'(x) \\
&= g_1'(g_2(x)) \cdot g_2'(x) \cdot e^{(\sin x)^2} + \cos(x^2 - 3\sqrt{x}) \cdot g_3'(g_4 \circ g_5(x)) \cdot (g_4 \circ g_5)'(x) \\
&= -\sin(x^2 - 3\sqrt{x}) \cdot \left(2x + \frac{3}{2}x^{-1/2}\right) \cdot e^{(\sin x)^2} + \cos(x^2 - 3\sqrt{x}) \cdot e^{(\sin x)^2} \cdot g_4'(g_5(x)) \cdot g_5'(x) \\
&= -\sin(x^2 - 3\sqrt{x}) \cdot \left(2x + \frac{3}{2}x^{-1/2}\right) \cdot e^{(\sin x)^2} + \cos(x^2 - 3\sqrt{x}) \cdot e^{(\sin x)^2} \cdot 2\sin x \cdot \cos x.
\end{aligned}$$

Next, we consider the derivative of the inverse of a function, which can be formulated in terms of the derivative of the original function.

Theorem 5.8. *Let I be an open interval and suppose that $f : I \to D$ is continuous and bijective. Let $x_0 \in I$ and suppose that f is differentiable at x_0 with $f'(x_0) \neq 0$. Then $f^{-1} : D \to I$ is differentiable at $y_0 = f(x_0)$ and*

$$(f^{-1})'(y_0) = \frac{1}{f'(x_0)} = \frac{1}{f'(f^{-1}(y_0))}.$$

In short, this theorem tells us that

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

Proof. By Theorem 4.26, we know that f^{-1} is continuous at y_0.

To study the differentiability of y_0, we need to calculate the limit

$$\lim_{y \to y_0} \frac{f^{-1}(y) - f^{-1}(y_0)}{y - y_0}.$$

Therefore, let (y_n) be an arbitrary sequence in D such that $y_n \to y_0$ and $y_n \neq y_0$ for all n. Write $x_n = f^{-1}(y_n)$. Since f^{-1} is continuous at y_0, it follows that $x_n \to x_0$ Then,

$$\begin{aligned}
\lim_{n\to\infty} \frac{f^{-1}(y_n) - f^{-1}(y_0)}{y_n - y_0} &= \lim_{n\to\infty} \frac{x_n - x_0}{f(x_n) - f(x_0)} \\
&= \lim_{n\to\infty} \frac{1}{\frac{f(x_n)-f(x_0)}{x_n - x_0}} \\
&= \frac{1}{\lim_{n\to\infty} \frac{f(x_n)-f(x_0)}{x_n-x_0}} \\
&= \frac{1}{f'(x_0)}
\end{aligned}$$

where the penultimate equality uses Theorem 4.14 and the fact that $\lim_{n\to\infty} \frac{f(x_n)-f(x_0)}{x_n-x_0} = f'(x_0) \neq 0$. □

Example - We can use Theorem 5.8 to calculate the derivative of the natural logarithm function $f : (0, \infty) \to \mathbb{R}$ where $f(x) = \ln x$. Note that $f = g^{-1}$ where $g : \mathbb{R} \to (0, \infty)$ is the exponential function $g(x) = e^x$. Therefore, for any $x \in (0, \infty)$,

$$f'(x) = \frac{1}{g(\ln x)} = \frac{1}{x}.$$

Example - We can use the previous example to calculate that, for any $a \in \mathbb{R}$, the function $f : (0, \infty) \to \mathbb{R}$ given by $f(x) = x^a$ is differentiable with

$$f'(x) = ax^{a-1}.$$

Indeed, we can write

$$f(x) = x^a = e^{a \ln x}$$

and then use the chain rule to conclude that

$$f'(x) = e^{a \ln x} \cdot \frac{a}{x} = ax^{a-1}.$$

5.3 Higher derivatives and continuous differentiability

By iteratively applying the process of calculating derivatives, we arrive at the notion of **higher order derivatives**.

Definition 5.9. *Let I be an interval, let $f : I \to \mathbb{R}$ be differentiable and suppose that its derivative f' is differentiable at $x_0 \in I$. Then we say that f is **twice differentiable** at x_0 and write*

$$f''(x_0) = (f')'(x_0).$$

This procedure can be repeated as long as derivatives exist, and so we can define the nth derivative of a function inductively by

$$f^{(n)}(x_0) := (f^{(n-1)})'(x_0).$$

*If $f^{(n)}(x_0)$ exists then we say that f is n**-times differentiable** at x_0. If the function $f^{(n)}(x)$ is continuous at x_0 then we say that f is n**-times continuously differentiable** at x_0. If a function f is 1-time continuously differentiable, we simply say that it is **continuously differentiable**.*

Example - Let $n \in \mathbb{N}$ be fixed and consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^n$. Then we can calculate the higher derivatives of f as follows:

$$\begin{aligned}
f'(x) &= nx^{n-1}, \\
f^{(2)}(x) &= n(n-1)x^{n-2}, \\
f^{(3)}(x) &= n(n-1)(n-2)x^{n-3}, \\
&\vdots \\
f^{(n-1)} &= n(n-1)\cdots 2x = n!x, \\
f^{(n)} &= n!, \\
f^{(m)} &= 0 \ \forall m \in \mathbb{N} \text{ such that } m > n.
\end{aligned}$$

Since all of these derivatives are continuous functions, it follows that f is m-times continuously differentiable over $\mathbb{R}$ for all $m \in \mathbb{N}$.

Example - The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \sin x$ is n-times continuously differentiable for all $n \in \mathbb{N}$. Indeed,

$$\begin{aligned}
f'(x) &= \cos x, \\
f^{(2)}(x) &= -\sin x, \\
f^{(3)}(x) &= -\cos x, \\
f^{(4)} &= \sin x, \\
f^{(5)} &= \cos x, \\
&\vdots
\end{aligned}$$

The derivatives follow a periodic pattern, which can be summarised as follows:

$$f^{(n)}(x) = \begin{cases} \cos x & \text{if there exists } k \in \mathbb{N} \cup \{0\} \text{ such that } n = 4k+1 \\ -\sin x & \text{if there exists } k \in \mathbb{N} \cup \{0\} \text{ such that } n = 4k+2 \\ -\cos x & \text{if there exists } k \in \mathbb{N} \cup \{0\} \text{ such that } n = 4k+3 \\ \sin x & \text{if there exists } k \in \mathbb{N} \cup \{0\} \text{ such that } n = 4k \end{cases}.$$

Example - So far, all of the derivatives that we have considered were also continuous functions, and so all of the functions we have considered are continuously differentiable over their whole domain. Let us see an example where this is not the case, i.e. where the derivative is defined on the whole domain, but is not continuous.

The classic example of such a function is the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} x^2 \sin(1/x) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}.$$

To calculate the derivative at 0, we need to consider the limit

$$\lim_{h \to 0} \frac{f(h) - f(0)}{h} = \lim_{h \to 0} \frac{h^2 \sin(1/h)}{h} = \lim_{h \to 0} h \sin(1/h).$$

Since $-1 \leq \sin x \leq 1$, we can use the sandwich rule to conclude that this limit is equal to zero. Therefore, $f'(0)$ exists and

$$f'(0) = 0.$$

For any $x \in \mathbb{R} \setminus 0$, we can use the chain rule and the product rule to calculate that

$$f'(x) = 2x \sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right).$$

Therefore, f is differentiable on $\mathbb{R}$ with $f'(x) : \mathbb{R} \to \mathbb{R}$ given by the formula

$$f'(x) = \begin{cases} 2x \sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}.$$

However, the function f' is **not** continuous at 0. In particular, the limit

$$\lim_{x \to 0} f'(x)$$

does not exist. To see that this limit does not exist, consider the null sequence (x_n) given by $x_n = \frac{1}{\pi n}$. Then

$$\begin{aligned} f'(x_n) &= 2x_n \sin\left(\frac{1}{x_n}\right) - \cos\left(\frac{1}{x_n}\right) \\ &= -\cos(\pi n) \\ &= -(-1)^n. \end{aligned}$$

This shows that $\lim_{n \to \infty} f'(x_n)$ does not exist, and thus by definition $\lim_{x \to 0} f'(x)$ does not exist.

5.4 Global and local extrema

In this section, we discuss a very important property of derivatives, which is that they can be used to calculate extreme points of a function. In particular, the extreme points for a differentiable function must be zeroes of the derivative function. This is frequently of practical use, particularly in optimisation problems that frequently arise in the real world.

For convenience, we begin by restating the definition of global extrema (see Definition 4.20).

Definition 5.10. *Let $D \subset \mathbb{R}$ and $f : D \to \mathbb{R}$. Then f has a* ***(global) minimum*** *at $x_0 \in D$ if*

$$f(x) \geq f(x_0) \;\; \forall\, x \in D,$$

and f has a ***(global) maximum*** *at $x_0 \in D$ if*

$$f(x) \leq f(x_0) \;\; \forall\, x \in D.$$

The point x_0 is called a ***(global) minimum/maximum point*** *or an* ***extreme point****. The value $f(x_0)$ is called a* ***minimum/maximum*** *of f, or an* ***extreme value****, or an* ***extremum****.*

We also need a *local* version of this notion, whereby the value of the function is larger/smaller than everything nearby.

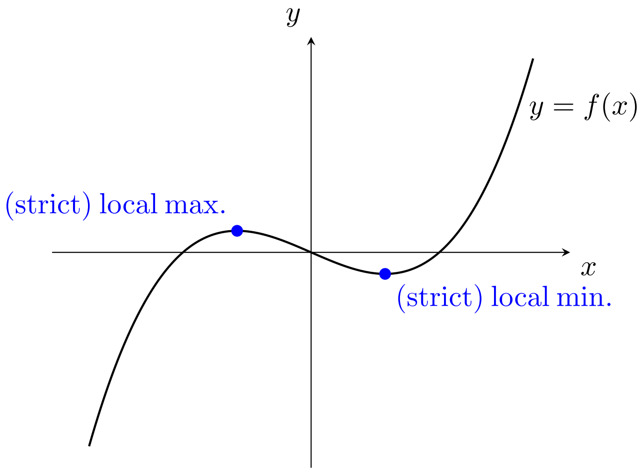
Definition 5.11. *Let $D \subset \mathbb{R}$ and $f : D \to \mathbb{R}$. Then f has a* ***local minimum*** *at $x_0 \in D$ if there exists $\epsilon > 0$ such that*

$$f(x) \geq f(x_0) \;\; \forall\, x \in D \cap (x_0 - \epsilon, x_0 + \epsilon),$$

and a ***strict local minimum*** *if $f(x) > f(x_0)$ for all $x \in D \cap (x_0 - \epsilon, x_0 + \epsilon)$. Analogously, f has a* ***local maximum*** *at $x_0 \in D$ if there exists $\epsilon > 0$ such that*

$$f(x) \leq f(x_0) \;\; \forall\, x \in D \cap (x_0 - \epsilon, x_0 + \epsilon)$$

and a ***strict local maximum*** *if $f(x) < f(x_0)$ for all $x \in D \cap (x_0 - \epsilon, x_0 + \epsilon)$. The point x_0 is called a* ***local minimum/maximum point*** *or a* ***local extreme point****. The value $f(x_0)$ is called a* ***minimum/maximum*** *of f, or an* ***extreme value****, or an* ***extremum****.*



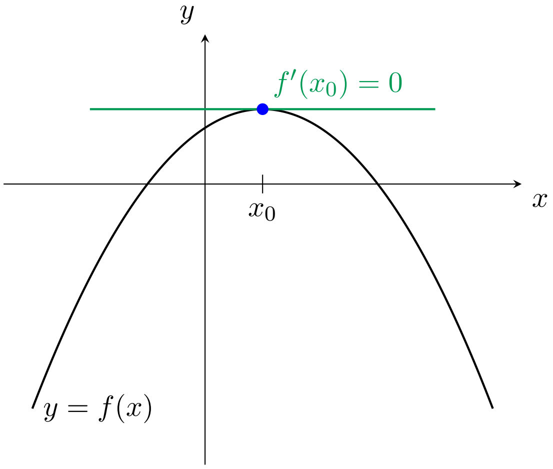
It follows immediately that global extreme points are also local extreme points. However, the reverse is not true, and a function can have local extreme points which are not global extreme points.

Consider the function $f(x) = x^3 - x^2$. We can plot this using whatever software/tools we like (I used desmos.com), and we see that this function has a local minimum but no global minimum. We will use derivatives to determine that this local extreme point occurs at $x_0 = \frac{2}{3}$, but we first need to develop some theory.

A crucial connection between derivatives and extreme points is given by the following theorem.

Theorem 5.12 (Necessary condition for an extreme point)**.** *Let I be an open interval and let $f : I \to \mathbb{R}$ be a function such that x_0 is a local extreme point of f and f is differentiable at x_0. Then*

$$f'(x_0) = 0$$



Proof. We will just deal with the case when f has a local minimum at x_0, and the case when x_0 is a maximum point can be handled similarly.

Since x_0 is a local minimum, there exists $\epsilon > 0$ such that

$$f(x) \geq f(x_0)$$

holds for all $x_0 - \epsilon < x < x_0 + \epsilon$. In particular, for all $x \in (x_0, x_0 + \epsilon)$,

$$\frac{f(x) - f(x_0)}{x - x_0} \geq 0.$$

Therefore, it must be the case that

$$\lim_{x \to x_0+} \frac{f(x) - f(x_0)}{x - x_0} \geq 0.$$

Similarly, for all $x \in (x_0 - \epsilon, x_0)$,

$$\frac{f(x) - f(x_0)}{x - x_0} \leq 0,$$

and hence

$$\lim_{x\to x_0-}\frac{f(x)-f(x_0)}{x-x_0}\leq 0.$$

On the other hand, we assume that $f'(x_0)$ exists, and so these two one sided limits must be the same. The only possibility is that

$$\lim_{x\to x_0-}\frac{f(x)-f(x_0)}{x-x_0}=0=\lim_{x\to x_0+}\frac{f(x)-f(x_0)}{x-x_0},$$

and therefore

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

□

Since zeroes of the derivative are so important to understanding extreme points, we give them a special name. Any x_0 such that $f'(x_0)=0$ is called a **stationary point** of f. We can then summarise Theorem 5.12 informally as follows:

$$x_0 \text{ is a local extreme point } \implies x_0 \text{ is a stationary point.} \tag{110}$$

Since the only global extreme points are also local extreme points, this can be expanded to say that

$$x_0 \text{ is a global extreme point } \implies x_0 \text{ is a local extreme point } \implies x_0 \text{ is a stationary point.}$$

Example - Consider again the function $f:\mathbb{R}\to\mathbb{R}$ given $f(x)=x^3-x^2$. The derivative is

$$f'(x)=3x^2-2x.$$

To find all stationary points, we solve $f'(x)=0$. This gives

$$x(3x-2)=0$$

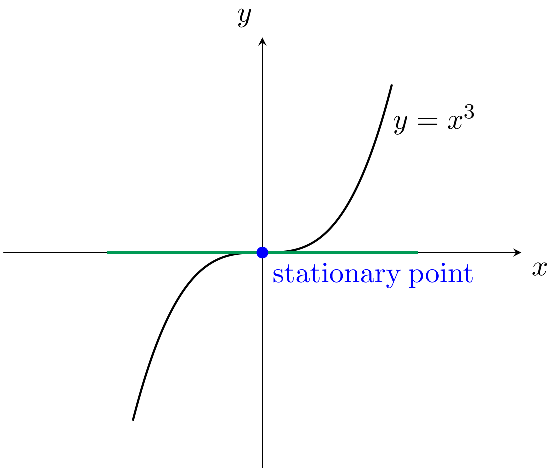
and so the stationary points are $x=0$ and $x=\frac{2}{3}$. We see from our picture that f has a strict local maximum at 0 and a strict local minimum at $\frac{2}{3}$ (we will deal with this a little more rigorously later).

Example - The purpose of this example is to show that the converse of (110) is false, that is,

$$x_0 \text{ is a stationary point } \implies x_0 \text{ is a local extreme point}$$

is **false**. Consider the function $f:\mathbb{R}\to\mathbb{R}$ given by $f(x)=x^3$. The derivative is $f'(x)=3x^2$, and therefore the only stationary point of f is 0. However, 0 is not a local extreme point of f, as we can see by drawing a graph of the function.

Although Theorem 5.12 helps us to determine the potential extreme points of a function, it does not give us a complete picture, and we will would like to have a method for determining whether a stationary point really is an extreme point, and if so, what kind of extreme point it is. For this, we will use the second derivative.



Theorem 5.13 (Second derivative test)**.** *Let I be an open interval, $x_0 \in I$, and suppose that $f : I \to \mathbb{R}$ is twice continuously differentiable at x_0. Suppose also that $f'(x_0) = 0$. Then,*

$$f''(x_0) > 0 \implies f \textit{ has a strict local minimum at } x_0,$$

and

$$f''(x_0) < 0 \implies f \textit{ has a strict local maximum at } x_0.$$

The proof of Theorem 5.13 is postponed until the next section, by which time we will have developed more theory to allow for a shorter proof.

Example - Let us consider again the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = x^3 - x^2.$$

This function satisfies the conditions of Theorem 5.13, since its second derivative can be calculated as

$$f''(x) = 6x - 2,$$

which is a continuous function from $\mathbb{R}$ to $\mathbb{R}$.

We have already seen that the stationary points of f are 0 and $\frac{2}{3}$. Note that

$$f''(0) = -2 < 0$$

and

$$f''\left(\frac{2}{3}\right) = 2 > 0.$$

Therefore, the second derivative test informs us that f has a strict local maximum at 0 and a strict local minimum at $\frac{2}{3}$. This matches up with our earlier intuition, based on the graph of the function.

Examples - It is important to emphasise that this theorem does not give us any information if $f''(x_0) = 0$.

- In the case of the function $f(x) = x^3$, we have $f'(x) = 3x^2$ and $f''(x) = 6x$, and so 0 is a stationary point of f for which $f''(0) = 0$. In this case, 0 is neither a local maximum or a local minumum, since we can find negative and positive values of $f(x)$ arbitrarily close to 0.

- In the case of the function $g(x) = x^4$, we have $g'(x) = 4x^3$ and $g''(x) = 12x^2$, and so 0 is a stationary point of g for which $g''(0) = 0$. In this case, g has a global minimum at 0.

- In the case of the function $h(x) = -x^4$, we have $h'(x) = -4x^3$ and $h''(x) = -12x^2$, and so 0 is a stationary point of h for which $h''(0) = 0$. In this case, h has a global minimum at 0.

5.5 Mean Value Theorem and l'Hospital's rule

The main new result of this section will be the Mean Value Theorem, which guarantees that the derivative of function attains a certain value over a closed interval, based only on the value taken by the function at the endpoints. We can think of this theorem as saying the slope of the graph of a function must be average somewhere in the interval. This will then be used to prove l'Hospital's rule, which is a useful tool for dealing with more difficult limits of rational expressions.

As a first step towards the Mean Value Theorem, we prove Rolle's Theorem, which is really a special case of the Mean Value Theorem where the function takes the same value at the beginning and end of the interval.

Theorem 5.14 (Rolle's Theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b), and suppose that $f(a) = f(b)$. Then there exists $x \in (a, b)$ such that $f'(x) = 0$.*

Geometrically the above theorem states that the graph of f has at least one point where the tangent is horizontal.

Proof. If f is a constant function then $f'(x) = 0$ for all $x \in (a, b)$ and the proof is immediate.

We can henceforth assume that f is not a constant function. Since f is continuous on $[a, b]$, we know from the Extreme Value Theorem 4.21 that f has a global (and hence also local) minimum at $m \in [a, b]$ and a global (and hence also local) maximum at $M \in [a, b]$. Since the function is not constant, $m \neq M$. Also, since $f(a) = f(b)$, it cannot be the case that both of the endpoints are extreme points, and so, without loss of generality, we assume that $m \in (a, b)$.

It then follows from Theorem 5.12 that $f'(m) = 0$. □

Example - Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \cos\left(x^2 - \frac{\pi}{2}\right)$. We can use Rolle's Theorem to deduce that there is some value $x \in (\sqrt{\pi}, 2\sqrt{\pi})$ such that $f'(x) = 0$.

Indeed, the function restricted to the interval $[\sqrt{\pi}, 2\sqrt{\pi}]$ is continuous and is differentiable on $(\sqrt{\pi}, 2\sqrt{\pi})$ (by the chain rule). Also,

$$f(\sqrt{\pi}) = \cos\left(\frac{\pi}{2}\right) = 0 = \cos\left(\frac{7\pi}{2}\right) = f(2\sqrt{\pi}).$$

We do not even need to calculate the derivative in order to establish the existence of the solution to $f'(x) = 0$. This could be practically useful in the case when a function has a very complicated derivative and it is not easy to solve $f'(x) = 0$.

The purpose of the next example is to highlight that we must be careful when applying the theorems we prove in this course, and in particular we need to take care that all of the conditions of the theorem are satisfied. In this case, it is important for Rolle's Theorem that the function is differentiable on the whole (open) interval.

Example - Consider the function $f : [-1, 1] \to \mathbb{R}$ given by $f(x) = |x|$. If we are careless with Rolle's Theorem, we may observe that $f(-1) = 1 = f(1)$ and conclude that there is some $x \in (-1, 1)$ such that $f'(x) = 0$. However, this is false, and for every point $x \in (-1, 1)$ for which $f'(x)$ exists, we either have $f'(x) = -1$ or $f'(x) = 1$.

The fallacy in the reasoning above is that this function does not satisfy the hypotheses of Theorem 5.14 since f is not differentiable at 0.

It is also important that the function is continuous on the whole closed interval $[a, b]$, as the following exercise illustrates.

Exercise - Give an example of a function $f : [a, b] \to \mathbb{R}$ that is differentiable on (a, b), satisfies $f(a) = f(b)$, and such that

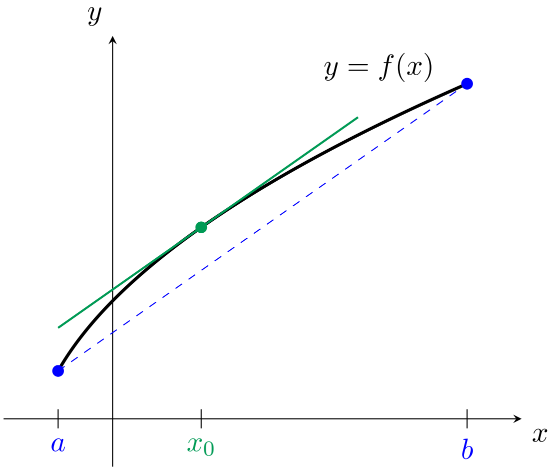
$$f'(x) \neq 0$$

for all $x \in (a, b)$

We are now ready to state and prove the Mean Value Theorem.

Theorem 5.15 (Mean Value Theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $x \in (a, b)$ such that*

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



Note that Rolle's Theorem immediately follows from the Mean Value Theorem. However, we will use Rolle's Theorem to prove the Mean Value Theorem.

Proof. Define a function $h : [a, b] \to \mathbb{R}$ by the formula

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

It follows from the basic properties of continuity and differentiability (and the corresponding properties of the function f) that h is continuous on $[a, b]$ and differentiable on (a, b). Furthermore, we can calculate that

$$h(a) = h(b),$$

since both $h(a)$ and $h(b)$ are equal to $f(a)$. It therefore follows from Rolle's Theorem that there exists $x \in (a, b)$ such that $h'(x) = 0$. On the other hand

$$0 = h'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

and this rearranges to give

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

as required. □

As with Rolle's Theorem, a nice feature of the Mean Value Theorem is that it allows us to deduce information about the derivative of a function just by knowing two endpoint values of the function, as well as the fact the function is continuous and differentiable.

Example - Consider the function $f : [-1, 2]$ given by

$$f(x) = 2x^3 - x + 1.$$

This function is continuous on the whole domain and differentiable on $(-1, 2)$, and so the Mean Value Theorem tells us that there is some $x \in (-1, 2)$ such that

$$f'(x) = \frac{f(2) - f(-1)}{2 - (-1)} = \frac{7 - 0}{3} = \frac{7}{3}.$$

The Mean Value Theorem also has nice theoretical consequences, and here is one such corollary, which confirms our intuition that we can use derivatives to see if a function is increasing or decreasing.

Corollary 5.16. *Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Suppose also that $f'(x) > 0$ for all $x \in (a, b)$. Then f is (strictly) increasing.*

Proof. Let $c, d \in [a, b]$ be arbitrary elements such that $c < d$. We need to show that

$$f(c) < f(d). \tag{111}$$

Apply the Mean Value Theorem for the function f restricted to the interval $[c, d]$. It follows that there exists $x \in (c, d)$ such that

$$f'(x) = \frac{f(d) - f(c)}{d - c}.$$

On the other hand, $f'(x) > 0$ (by the hypothesis of the corollary) and $d - c > 0$. It therefore must be the case that $f(d) - f(c) > 0$, which proves the intended inequality (111) □

The converse of Corollary 5.16 is *almost* true, but not quite. This is made more precise in the following exercise.

Exercise -

- Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b), and suppose that f is non-decreasing. Prove that, for all $x \in (a, b)$,

  $$f'(x) \geq 0$$

- Give an example of a function $f : [a, b] \to \mathbb{R}$ which is continuous and strictly increasing on $[a, b]$ and differentiable on (a, b), but for which there is some $x \in (a, b)$ with

  $$f'(x) \leq 0.$$

Note that, if we combine the first part of this exercise with Corollary 5.16, we have the following convenient equivalence.

Corollary 5.17. *Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then*

$$f \text{ is non-decreasing} \iff f'(x) \geq 0 \ \forall x \in (a, b).$$

Proof. The $\Rightarrow$ direction is precisely the first part of the previous exercise. The $\Leftarrow$ direction can be proved by slightly modifying the proof of Corollary 5.16. □

In order to prove l'Hospital's Rule, it will be useful to have the following generalisation of the Mean Value Theorem, which is sometimes known as **Cauchy's Mean Value Theorem**.

Theorem 5.18. *Let $f, g : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $x \in (a, b)$ such that*

$$f'(x)(g(b) - g(a)) = g'(x)(f(b) - f(a)) \tag{112}$$

In particular, if $g'(x) \neq 0$ for all $x \in (a, b)$, then there exists $x \in (a, b)$ such that

$$\frac{f'(x)}{g'(x)} = \frac{f(b) - f(a)}{g(b) - g(a)}. \tag{113}$$

Note that Theorem 5.15 immediately follows from Theorem 5.18 by setting $g(x) = x$ and applying (113).

Proof. **Case 1** - Suppose that $g(a) = g(b)$. It follows from Rolle's Theorem that there exists $x \in (a, b)$ with $g'(x) = 0$. With this x, we see that both sides of (112) are equal to zero, as required. Since the condition that $g'(x) \neq 0$ for all $x \in (a, b)$ is violated, we do not need to verify (113) in this case.

Case 2 - The proof is similar to the proof of the Mean Value Theorem. Define the function $h : [a, b] \to \mathbb{R}$ by the formula

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

It follows from the basic properties of continuity and differentiability (and the corresponding properties of the functions f and g) that h is continuous on $[a, b]$ and differentiable on (a, b). Furthermore, we can calculate that

$$h(a) = h(b),$$

since both $h(a)$ and $h(b)$ are equal to $f(a)$. It therefore follows from Rolle's Theorem that there exists $x \in (a, b)$ such that $h'(x) = 0$. On the other hand

$$0 = h'(x) = f'(x) - \frac{f(b) - f(a)}{g(b) - g(a)}g'(x),$$

and this rearranges to give (112).

To prove (113) from (112), simply use the knowledge that $g'(x) \neq 0$ to divide both sides of (112) through by $g'(x)$ and rearrange accordingly. □

This puts us in a good position to prove l'Hospital's rule. This is a tool which is helpful for calculating limits of the form "$\frac{0}{0}$" or "$\frac{\infty}{\infty}$". Consider for instance the limit

$$\lim_{x\to 0}\frac{e^x-1}{x},$$

which we encountered earlier in the proof of Theorem 5.3. The easiest way to calculate such a limit might be to apply Theorem 4.14, making use of the fact that "the limit of the ratio is the ratio of the limits", to conclude that

$$\lim_{x\to 0}\frac{e^x-1}{x}=\frac{\lim_{x\to 1}e^x-1}{\lim_{x\to 1}x}=\frac{0}{0}.$$

However, this method gives us some nonsense, and the reason for this is that our application of Theorem 4.14 was invalid as the limit of the denominator is 0.

We need a new tool to deal with this situation and other similar ones, and this is provided by l'Hospital's Rule.

Theorem 5.19. *Let $I=(a,b)$ and $x_0\in[a,b]$. Let $f,g:I\setminus\{x_0\}\to\mathbb{R}$ be differentiable on $I\setminus\{x_0\}$ and assume that $g'(x)\neq 0$ for all $x\in I\setminus\{x_0\}$. Furthermore, assume that one of the following situations arises:*

- $\lim_{x\to x_0}f(x)=\lim_{x\to x_0}g(x)=0$*, or*
- $\lim_{x\to x_0}f(x)=\pm\infty$ *and* $\lim_{x\to x_0}g(x)=\pm\infty$*.*

Then

$$\lim_{x\to x_0}\frac{f(x)}{g(x)}=\lim_{x\to x_0}\frac{f'(x)}{g'(x)}\tag{114}$$

if the limit on the right hand side exists (including the possibility that the limit on the right hand side is $\pm\infty$).

Proof.

- Suppose that $\lim_{x\to x_0}f(x)=\lim_{x\to x_0}g(x)=0$. First, note that we can extend the domains of f and g to include x_0 by defining

  $$f(x_0)=g(x_0)=0.$$

  Since $\lim_{x\to x_0}f(x)=0$, it follows that the new extended function f is continuous at x_0, and the same is true for g.

  Now, let (x_n) be an arbitrary sequence in I such that $x_n\neq x_0$ for all $n\in\mathbb{N}$, and $x_n\to x_0$. We will show that

  $$\lim_{n\to\infty}\frac{f(x_n)}{g(x_n)}=\lim_{x\to x_0}\frac{f'(x)}{g'(x)},\tag{115}$$

  which will then imply the intended result (114).

  For each term x_n in the sequence, apply Theorem 5.18 for the functions f and g restricted to the interval $[x_0,x_n]$ if $x_n>x_0$ (or $[x_n,x_0]$ if $x_n<x_0$). It follows that there exists $y_n\in I$ such that $0<|y_n-x_0|<|x_n-x_0|$ and

  $$\frac{f'(y_n)}{g'(y_n)}=\frac{f(x_n)-f(x_0)}{g(x_n)-g(x_0)}=\frac{f(x_n)}{g(x_n)}.\tag{116}$$

Since $|y_n - x_0| < |x_n - x_0|$ and $x_n \to x_0$, it follows that $y_n \to x_0$. Therefore, taking the limit as n goes to infinity for both sides of (116), we obtain

$$\lim_{x \to x_0} \frac{f'(x)}{g'(x)} = \lim_{n \to \infty} \frac{f'(y_n)}{g'(y_n)} = \lim_{n \to \infty} \frac{f(x_n)}{g(x_n)}.$$

This establishes (115) and completes the proof of the first point of the theorem.

- Suppose that $\lim_{x \to x_0} f(x) = \pm\infty$ and $\lim_{x \to x_0} g(x) = \pm\infty$. In this case, we can apply the first part of the theorem to the functions $1/f$ and $1/g$ to get the result.

□

Limits which take the two forms "$\frac{0}{0}$" or "$\frac{\infty}{\infty}$" described in the statement of l'Hospital's rule are examples of limits with *indeterminate form.*

Example - Let's begin by using l'Hospital's rule to solve the aforementioned limit

$$\lim_{x \to 0} \frac{e^x - 1}{x}.$$

The numerator and denominator both tend to 0 and so l'Hospital's rule is applicable. It follows that

$$\lim_{x \to 0} \frac{e^x - 1}{x} = \lim_{x \to 0} \frac{e^x}{1} = 1.$$

This agrees with (106), which we proved earlier using a different method.

Example - On an earlier exercise sheet (Exercise 43 on Sheet 8), you proved that

$$\lim_{x \to 0} \frac{\sin x}{x} = 1.$$

We can prove this again now using l'Hospital's rule. Indeed, since the numerator and denominator both tend to zero, we have

$$\lim_{x \to 0} \frac{\sin x}{x} = \lim_{x \to 0} \frac{\cos x}{1} = \frac{1}{1} = 1.$$

Example - Consider the function $f : (0, \infty) \to \mathbb{R}$ given by $f(x) = x \ln x$. Then,

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} \frac{g(x)}{h(x)}$$

where $g, h : (0, \infty) \to \mathbb{R}$ are given by

$$g(x) = \ln x, \;\; h(x) = \frac{1}{x}.$$

Note that $\lim_{x \to 0} g(x) = -\infty$ and $\lim_{x \to 0} h(x) = \infty$, Therefore, we can use l'Hospital's rule to calculate that

$$\lim_{x \to 0} x \ln x = \lim_{x \to 0} \frac{g'(x)}{h'(x)} = \lim_{x \to 0} \frac{x^{-1}}{-x^{-2}} = \lim_{x \to 0} -x = 0.$$

Example - Sometimes we need to use l'Hospital's rule multiple times to calculate a given limit. For instance,

$$\lim_{x\to 0}\frac{x-\sin x}{x^3}=\lim_{x\to 0}\frac{1-\cos x}{3x^2}=\lim_{x\to 0}\frac{\sin x}{6x}=\lim_{x\to 0}\frac{\cos x}{6}=\frac{1}{6}.$$

The following question concerning l'Hospital's rule was raised in a lecture: if we are in the situation where l'Hospital's rule may be applicable, but the limit

$$\lim_{x\to x_0}\frac{f'(x)}{g'(x)}$$

does not exist, does this imply that the limit

$$\lim_{x\to x_0}\frac{f(x)}{g(x)}$$

also does not exist? The purpose of the next example is to show that the answer to this question is "no".

Example - Consider the limit

$$\lim_{x\to\infty}\frac{\sin x+x}{x}.$$

It is not too difficult to use the definition of the limit and the fact that $|\sin x|\leq 1$ to verify directly that

$$\lim_{x\to\infty}\frac{\sin x+x}{x}=1.$$

On the other hand, this also appears to be a candidate for an application of l'Hospital's rule, since the numerator and demominator both tend to infinity. Therefore,

$$\lim_{x\to\infty}\frac{\sin x+x}{x}=\lim_{x\to\infty}\frac{\cos x+1}{1},$$

if the limit on the right hand side of the above equation exists. However,

$$\lim_{x\to\infty}\frac{\cos x+1}{1}$$

does not exist, since the numerator continues to oscilate between 0 and 2.

Finally, we can use l'Hospital's rule to prove Theorem 5.13, the second derivative test. The statement is restated below for convenient reading.

Theorem 5.20 (Second derivative test)**.** *Let I be an open interval, $x_0\in I$, and suppose that $f:I\to\mathbb{R}$ is twice continuously differentiable at x_0. Suppose also that $f'(x_0)=0$. Then,*

$$f''(x_0)>0 \implies f \text{ has a strict local minimum at } x_0, \tag{117}$$

and

$$f''(x_0)<0 \implies f \text{ has a strict local maximum at } x_0. \tag{118}$$

Proof. We will only prove (117), and leave it to the student to check that (118) can be proved by a similar argument.

Suppose that $f'(x_0) = 0$ and $f''(x_0) > 0$. We need to show that x_0 is a strict local minimum, i.e. that there exists $\epsilon > 0$ such that, for all $x \in (x_0 - \epsilon, x_0 + \epsilon)$,

$$f(x) > f(x_0).$$

Since f is differentiable at x_0, it is also continuous at x_0 and so $\lim_{x\to x_0} f(x) - f(x_0) = 0$. Then, by l'Hospital's rule,

$$\lim_{x\to x_0} \frac{f(x) - f(x_0)}{(x - x_0)^2} = \lim_{x\to x_0} \frac{f'(x)}{2(x - x_0)} = \frac{1}{2} \lim_{x\to x_0} \frac{f'(x) - f'(x_0)}{x - x_0} = \frac{1}{2} f''(x_0) > 0.$$

Since this limit is positive, it follows that

$$\frac{f(x) - f(x_0)}{(x - x_0)^2} > 0 \tag{119}$$

holds for all x sufficiently close to x_0. But the denominator on the left hand side of (119) is positive, and so it must be the case that

$$f(x) - f(x_0) > 0$$

holds for all x sufficiently close to x_0. This is equivalent to the statement that there exists $\epsilon > 0$ such that, for all $x \in (x_0 - \epsilon, x_0 + \epsilon)$,

$$f(x) > f(x_0).$$

□

5.6 Convexity

The second derivative helps us to determine the shape of the graph of a function f. If every line connecting two points on the curve lies above the curve then we say that f is **convex**. If every such line lies below the curve then we say that f is **concave**. A more formal version of this definition is given now.

Definition 5.21. *Let I be an interval and $f : I \to \mathbb{R}$. We say that f is* ***convex*** *if, for all $x_0, x_1 \in I$ with $x_0 < x_1$, and all $\lambda \in (0,1)$,*

$$f((1-\lambda)x_0 + \lambda x_1) \leq (1-\lambda)f(x_0) + \lambda f(x_1). \tag{120}$$

We say that f is ***concave*** *if, for all $x_0, x_1 \in I$ and all $\lambda \in (0,1)$*

$$f((1-\lambda)x_0 + \lambda x_1) \geq (1-\lambda)f(x_0) + \lambda f(x_1). \tag{121}$$

We can also make a small modification of the above definition to define functions that are **strictly convex** and **strictly concave**. In these cases, the inequalities (120) and (121) should be strict inequalities.

Let us take some time to understand the geometry of inequality (120), and why this really captures the notion of convexity described in the first paragraph of this section. As λ varies from 0 to 1, the quantity

$$(1-\lambda)x_0 + \lambda x_1$$

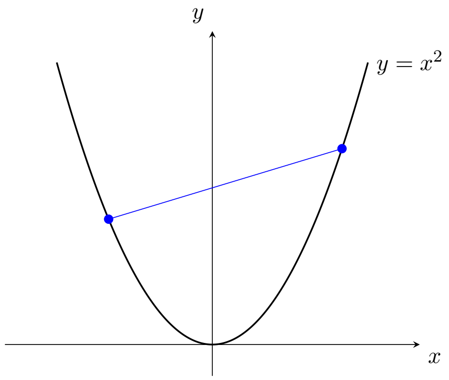
varies from x_0 to x_1, moving continuously and linearly from left to right. Therefore, the left hand side of (120) describes the value (or height) of the function f at some point in between x_0 and x_1.

On the other hand, as λ varies from 0 to 1, the quantity

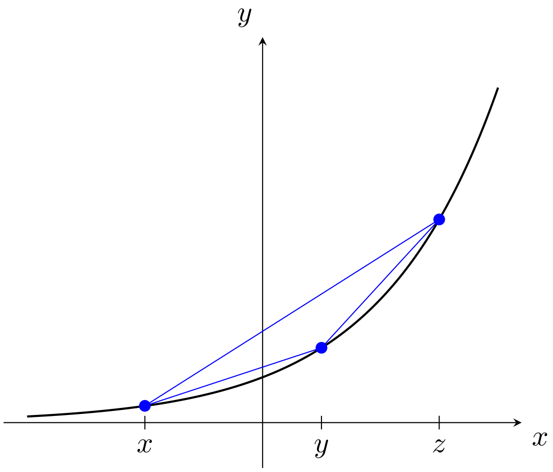
$$(1-\lambda)f(x_0) + \lambda f(x_1)$$

varies from $f(x_0)$ to $f(x_1)$, moving continuously and linearly. In other words, this value describes the height of the height of the straight line from $(x_0, f(x_0))$ to $(x_1, f(x_1))$ at some point in between x_0 and x_1.

Example - Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$. If we draw (a segment of) a straight line between any two points on the curve, this line is strictly above the curve, and so the function is convex. In fact, this function is strictly convex, since the segment of the line only touches the curve at the endpoints.



The next lemma gives us a useful property of convex functions, and will be used later in this section to prove the forthcoming Theorem 5.23.



Lemma 5.22. *Let I be an interval and suppose that $f : I \to \mathbb{R}$ is convex. Let $x, y, z \in I$ such that $x < y < z$. Then*

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x} \leq \frac{f(z) - f(y)}{z - y}.$$

Proof. We will just prove the first bound from Lemma 5.22, that is,

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x}. \tag{122}$$

We rearrange this to make $f(y)$ the subject, and so we see that (122) is equivalent to

$$f(y) \leq \frac{(y - x)(f(z) - f(x))}{z - x} + f(x) = \frac{f(x)(z - y) + f(z)(y - x)}{z - x}. \tag{123}$$

Since $x < y < z$, we can write $y = (1 - \lambda)x + \lambda z$ for some $\lambda \in (0, 1)$, and a rearrangement of this gives us that

$$\lambda = \frac{y - x}{z - x},$$

and thus

$$1 - \lambda = \frac{z - y}{z - x}.$$

Therefore, by the definition of convexity applied with this choice of λ, we obtain that

$$f(y) \leq (1 - \lambda)f(x) + \lambda f(z) = \frac{z - y}{z - x}f(x) + \frac{y - x}{z - x}f(z) = \frac{f(x)(z - y) + f(z)(y - x)}{z - x},$$

which proves the intended bound (123). □

For functions that are twice differentiable, we can determine convexity by studying the second derivative.

Theorem 5.23. *Let I be an open interval and suppose that $f : I \to \mathbb{R}$ is twice differentiable. Then*

- *f is convex $\iff$ $f''(x) \geq 0$ for all $x \in I$, and*
- *f is concave $\iff$ $f''(x) \leq 0$ for all $x \in I$.*

Proof. We will only prove the first part of Theorem 5.23, concerning the case when f is convex. For the case of concave f, a similar argument works.

We first prove the "$\Rightarrow$" direction. Suppose that f is convex. We will show that f' is non-decreasing, and then use Corollary 5.17 to conclude that $f''(x) \geq 0$ for all $x \in I$.

It remains to show that f' is non-decreasing. That is, for arbitrary $x_0, x_1 \in I$ such that $x_0 < x_1$, we need to show that

$$f'(x_0) \leq f'(x_1). \tag{124}$$

Let $x \in I$ be an element satisfying $x_0 < x < x_1$. It follows from Lemma 5.22 that

$$\frac{f(x) - f(x_0)}{x - x_0} \leq \frac{f(x_1) - f(x_0)}{x_1 - x_0} \leq \frac{f(x_1) - f(x)}{x_1 - x}. \tag{125}$$

It follows from the first part of (125) that

$$\lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0} \leq \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

However, as f is differentiable at x_0, it must be the case that the one-sided limit above is equal to $f'(x_0)$ (this is a consequence of Theorem 4.17). So, we have

$$f'(x_0) \leq \frac{f(x_1) - f(x_0)}{x_1 - x_0}. \tag{126}$$

Similarly, the second inequality in (125) implies that

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} \leq \lim_{x \to x_1^-} \frac{f(x_1) - f(x)}{x_1 - x} = f'(x_1).$$

Combining this with (126) gives the intended inequality (124).

Now we turn to the "$\Leftarrow$" direction. Suppose that $f''(x) \geq 0$ for all $x \in I$. Let $x_0, x_1 \in I$ be arbitrary. We need to prove that

$$f((1-\lambda)x_0 + \lambda x_1) \leq (1-\lambda)f(x_0) + \lambda f(x_1) \tag{127}$$

holds for all $\lambda \in (0,1)$. Without loss of generality, we assume that $x_0 < x_1$, and so $x := (1-\lambda)x_0 + \lambda x_1$ satisfies

$$x_0 < x < x_1.$$

Applying the Mean Value Theorem for the function restricted to the intervals $[x_0, x]$ and $[x, x_1]$, it follows that there exists $y_1 \in (x_0, x)$ and $y_2 \in (x, x_1)$ such that

$$f'(y_1) = \frac{f(x) - f(x_0)}{x - x_0}$$

and

$$f'(y_2) = \frac{f(x_1) - f(x)}{x_1 - x}.$$

Since $f'' \geq 0$, Corollary 5.17 implies that f' is non-decreasing, and so

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(y_1) \leq f'(y_2) = \frac{f(x_1) - f(x)}{x_1 - x}$$

A rearrangement of the above inequality yields

$$f(x)(x_1 - x_0) \leq f(x_1)(x - x_0) + f(x_0)(x_1 - x). \tag{128}$$

It also follows from the definition of $x := (1 - \lambda)x_0 + \lambda x_1$ that

$$\begin{aligned} x - x_0 &= \lambda(x_1 - x_0) \\ x_1 - x &= (1 - \lambda)(x_1 - x_0). \end{aligned}$$

Inserting these two expressions into (128) and dividing through by $x_1 - x_0 > 0$, we conclude that

$$f(x) \leq \lambda f(x_1) + (1 - \lambda) f(x_0),$$

which gives (127) and completes the proof.

□

5.7 Taylor polynomials

In this section, we will use higher derivatives to approximate familiar functions (such as the exponential function and the trigonometric functions) by polynomials. What is particularly strong about this estimate is that we only need to know about the values of the derivative $f^{(i)}(x_0)$ for some fixed x_0 and all $1 \leq i \leq n$ in order to get an estimate for the behaviour of $f(x)$ for x close to x_0. Depending on certain properties of the function and its derivatives, this estimate can be very accurate.

One of our main motivations for developing the theory of this section is to prove the beautiful formula

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!},$$

thereby giving a description of the exponential function as an infinite sum of polynomials.

The fundamental result of this section is Taylor's Theorem.

Theorem 5.24 (Taylor's Theorem)**.** *Let I be an open interval, $x_0 \in I$, and let $f : I \to \mathbb{R}$ be $(n+1)$-times differentiable for some $n \in \mathbb{N}$. Then, for all $x \in I$, there exists $y \in [\min\{x_0, x\}, \max\{x_0, x\}]$ (i.e. there exists y in between x and x_0) such that*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \frac{f^{(n+1)}(y)}{(n+1)!}(x - x_0)^{n+1}. \tag{129}$$

Definition 5.25. *With the objects f, I, x_0, x, y and n as in the statement of Theorem 5.24, we call*

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

*the **Taylor polynomial of f at order n (at x_0)**. The term*

$$R_n(x) := \frac{f^{(n+1)}(y)}{(n+1)!}(x - x_0)^{n+1}$$

*is called the **remainder** of the Taylor polynomial.*

This statement looks rather complicated, and so let us make some effort to understand what it is telling us.

The theorem does give us a precise valuation for $f(x)$, but this is not particularly helpful, since we do not know enough information about the value of y, and thus we do not have an easy way to determine the remainder $R_n(x)$. However, the theorem is very useful for giving an estimation for $f(x)$ by a polynomial $T_n(x)$, particularly if we have some information about the (higher) derivatives which allows us to control the size of the remainder term. In particular, if we have a good upper bound for $|f^{(n+1)}(y)|$ in the given range then $T_n(x)$ becomes an accurate estimate for $f(x)$. We will study some examples to give a better understanding of this statement and its uses.

Before we prove Theorem 5.24, let us consider a subtle technical point. Does this statement makes sense in the case when $x = x_0$? It should do, as we have not excluded this possibility from the statement.

Indeed, we use the common convention that $0^0 = 1$, and then, for the case when $x = x_0$, (129) reduces to $f(x) = f(x_0)$, since all of the terms in the sum with $k \neq 0$ vanish. We will also use the convention that $0^0 = 1$ in the proof, which starts now.

Proof of Theorem 5.24. The discussion immediately before the proof verifies the theorem for the case when $x = x_0$. Let $x \in I \setminus \{x_0\}$ be arbitrary. We assume without loss of generality that $x > x_0$ (this assumption is made only to make some of the notation simpler). Define the function $g : [x_0, x] \to \mathbb{R}$ by the formula

$$g(t) := f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(t)}{k!}(x-t)^k - \frac{m}{(n+1)!}(x-t)^{n+1},$$

where $m \in \mathbb{R}$ is chosen to ensure that $g(x_0) = 0$. (The precise value of m is not really important in the definition, but we can write it down by making m the subject of the formula $g(x_0) = 0$. This gives

$$m := f(x) \cdot \frac{(n+1)!}{(x-x_0)^{n+1}} - \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} \cdot (x-x_0)^{k-n-1}(n+1)!$$

for the value of m.) It follows from the definition of g (and the convention that $0^0 = 1$) that $g(x) = f(x) - f(x) = 0$. Together with the fact that $g(x_0) = 0$, we see that the function g satisfies the conditions of Rolle's Theorem, and so there exists $y \in (x_0, x)$ such that $g'(y) = 0$.

On the other hand, we can use the product rule to calculate the derivative $g'(t)$ (let us emphasise that t is the variable here, while x and x_0 can be regarded as constants). After some convenient cancellation of terms, we obtain that

$$\begin{aligned} g'(t) &= -\sum_{k=0}^{n} \frac{f^{(k+1)}(t)}{k!}(x-t)^k + \sum_{k=0}^{n} \frac{f^{(k)}(t)}{k!} \cdot k(x-t)^{k-1} + \frac{m}{n!}(x-t)^n \\ &= -\sum_{k=0}^{n} \frac{f^{(k+1)}(t)}{k!}(x-t)^k + \sum_{k=1}^{n} \frac{f^{(k)}(t)}{(k-1)!}(x-t)^{k-1} + \frac{m}{n!}(x-t)^n \\ &= -\frac{f^{(n+1)}(t)}{n!}(x-t)^n + \frac{m}{n!}(x-t)^n. \end{aligned}$$

Therefore,

$$0 = g'(y) = -\frac{f^{(n+1)}(y)}{n!}(x-y)^n + \frac{m}{n!}(x-y)^n$$

Since $y \neq x$, we can divide through by $\frac{n!}{(x-y)^n}$ and get

$$m = f^{(n+1)}(y).$$

Plugging this value of m back into the definition of g and using the fact that $g(x_0) = 0$, we conclude that

$$0 = g(x_0) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k - \frac{f^{(n+1)}(y)}{(n+1)!}(x-x_0)^{n+1}.$$

A rearrangement of this gives the intended formula (129). □

We mentioned earlier that Taylor's Theorem is particularly useful when we have control over the higher derivatives. One such instance is that of polynomials, since a polynomial of degree n satisfies $f^{(k)}(x) = 0$ for all $k \geq n+1$ and all $x \in \mathbb{R}$. This allows us to rewrite polynomials by "shifting the origin".

Corollary 5.26. *Let $p : \mathbb{R} \to \mathbb{R}$ be a polynomial of degree n. Then, for all $x, x_0 \in \mathbb{R}$,*

$$p(x) = \sum_{k=0}^{n} \frac{p^{(k)}(x_0)}{k!}(x - x_0)^k.$$

Proof. This is left as an exercise. □

Example - Let

$$f(x) = x^4 - 2x^2 + x + 10.$$

Apply the previous corollary with $x_0 = 1$. Then

$$f(x) = \sum_{k=0}^{4} \frac{f^{(k)}(1)}{k!}(x-1)^k.$$

We need to calculate some derivatives:

$$\begin{aligned} f^{(1)}(x) &= 4x^3 - 4x + 1 \\ f^{(2)}(x) &= 12x^2 - 4 \\ f^{(3)}(x) &= 24x \\ f^{(4)}(x) &= 24. \end{aligned}$$

Therefore,

$$\begin{aligned} f^{(0)}(1) &= 10 \\ f^{(1)}(1) &= 1 \\ f^{(2)}(1) &= 8 \\ f^{(3)}(1) &= 24 \\ f^{(4)}(1) &= 24, \end{aligned}$$

and

$$f(x) = \sum_{k=0}^{4} \frac{f^{(k)}(1)}{k!}(x-1)^k = 10 + (x-1) + 4(x-1)^2 + 4(x-1)^3 + (x-1)^4.$$

For a more general description of the accuracy of the Taylor polynomial given in Theorem 5.24, the following corollary is convenient. Note that we must restrict our domain to be a bounded open interval in this statement.

Corollary 5.27. *Let $I = (a,b)$, $x_0 \in I$, and let $f : I \to \mathbb{R}$ be $(n+1)$-times differentiable. Suppose also that there is some constant M such that $|f^{(n+1)}(x)| \leq M$ for all $x \in I$. Let $T_n(x)$ be as defined in Definition 5.25 (recall that this quantity depends on the choice of x_0). Then, for all $x \in I$,*

$$|f(x) - T_n(x)| \leq \frac{M(b-a)^{n+1}}{(n+1)!}.$$

Proof. The result follows immediately from the statement of Theorem 5.24, making use of the fact that $|x - x_0| \leq b - a$ holds for all $x, x_0 \in I$. □

We will apply Corollary 5.27 now to express the exponential function as an infinite sum. In the application, we will use the following exercise concerning a certain null sequence.

Exercise - Show that, for any fixed constant $c \in \mathbb{R}$,

$$\lim_{n\to\infty} \frac{c^n}{n!} = 0.$$

Corollary 5.28. *For any $x \in \mathbb{R}$,*

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Proof. Fix $x \in \mathbb{R}$ (which means we can regard x as a constant) and consider the function $f : (-2|x|, 2|x|) \to \mathbb{R}$ given by $f(x) = e^x$. Apply the previous corollary with $x_0 = 0$, and note that

$$T_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}.$$

Since $f(x)$ is strictly increasing and its derivatives are all also the same as $f(x)$, it follows that

$$|f^{(n+1)}(x)| \leq e^{2|x|}$$

holds for all x in the domain of f. Therefore, by Corollary 5.27

$$0 \leq |f(x) - T_n(x)| \leq \frac{e^{2|x|}(4|x|)^{n+1}}{(n+1)!}.$$

It follows from the previous exercise that

$$\lim_{n\to\infty} \frac{e^{2|x|}(4|x|)^{n+1}}{(n+1)!} = 0.$$

It then follows from the Sandwich Rule that

$$\lim_{n\to\infty} |f(x) - T_n(x)| = 0.$$

This is equivalent to the statement that

$$\lim_{n\to\infty} T_n(x) = f(x),$$

and the limit on the right is by definition equal to $\sum_{k=0}^{\infty} \frac{x^k}{k!}$. □

We can use a similar argument to give a formula for the cosine function as an infinite sum.

Corollary 5.29. *For any $x \in \mathbb{R}$,*

$$\cos(x) = \sum_{k=0}^{\infty} (-1)^k \cdot \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots.$$

Proof. Fix $x \in \mathbb{R}$ (which means we can regard x as a constant) and consider the function $f : (-2|x|, 2|x|) \to \mathbb{R}$ given by $f(x) = \cos x$. We will apply Corollary 5.27 with $x_0 = 0$. To see how this application works, we need to study the higher derivatives of $f(x)$. Observe that

$$\begin{aligned} f^{(1)}(x) &= -\sin x, \\ f^{(2)}(x) &= -\cos x, \\ f^{(3)} &= \sin x, \\ f^{(4)} &= \cos x, \\ f^{(5)} &= -\sin x \\ &\vdots \end{aligned}$$

Therefore, the values of $f^{(k)}(0)$ follows the periodic pattern "$1, 0, -1, 0$". It follows that

$$T_n(x) = \sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k \cdot \frac{x^{2k}}{(2k)!}.$$

It follows from the basic properties of the sine and cosine functions that

$$|f^{(n+1)}(x)| \leq 1$$

holds for all x in the domain of f. Therefore, by Corollary 5.27

$$0 \leq |f(x) - T_n(x)| \leq \frac{(4|x|)^{n+1}}{(n+1)!}.$$

It follows from the previous exercise that

$$\lim_{n \to \infty} \frac{(4|x|)^{n+1}}{(n+1)!} = 0.$$

It then follows from the Sandwich Rule that

$$\lim_{n \to \infty} |f(x) - T_n(x)| = 0.$$

This is equivalent to the statement that

$$\lim_{n \to \infty} T_n(x) = f(x),$$

and the limit on the right is by definition equal to $\sum_{k=0}^{\infty} (-1)^k \cdot \frac{x^{2k}}{(2k)!}$. □

6 Basic Integration Theory

6.1 Introduction and open sets

After working hard to calculate derivatives of certain functions in the previous chapter, a natural follow-up question is to ask if (at least in some cases) there is also an *inverse operation*. In other words:

Given some function f, is there a differentiable function F such that $F' = f$?

Such a function will be called an *antiderivative* of f.

In contrast with the definition of the derivative, which is unique if the function is differentiable, there are some problems with this definition. In particular, we will see that, if an antiderivative exists, it is not unique. (One can easily see, e.g., that all linear functions of the form $F(x) = x + c$ with $c \in \mathbb{R}$, are antiderivatives of the constant function $f(x) = 1$.) For other functions, the antiderivative just does not exist. However, we will see that an antiderivative exists for all continuous functions, and even more. This statement is one part of the *fundamental theorem of calculus*.

The other part of the theorem is concerned with the *integral of a function over an interval*. This notion has a geometric definition: for a non-negative function $f : [a, b] \to [0, \infty)$, the integral $\int_a^b f(x)\,dx$ is the area bounded by the x-axis, the graph of f, and the two vertical lines $x = a$ and $x = b$. That is, the area of the set

$$\{(x, y) \in \mathbb{R}^2 : a \leq x \leq b, 0 \leq y \leq f(x)\}.$$

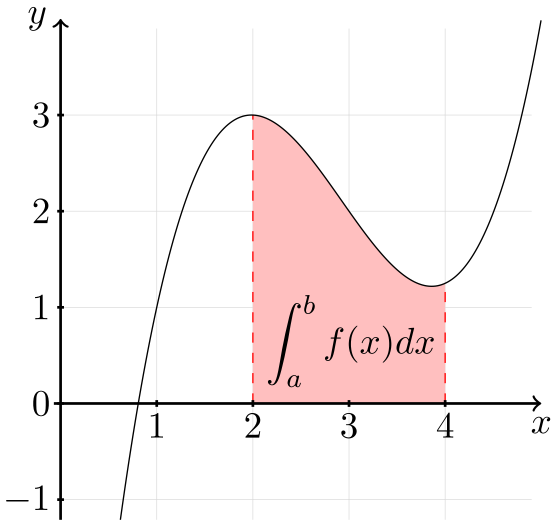


Figure 23: The (definite) integral of a function corresponds to the area "between" the graph of the function and the x-axis. In this example we have $a = 2$ and $b = 4$.

Clearly, we have to define precisely what this means. In contrast to derivatives, which were closely connected to the slope at a point, local extrema, and other local properties of a function, the integral is a global quantity. However, we will see later that both concepts are very much related. In particular, if $f : [a, b] \to \mathbb{R}$ is continuous, then there exists an antiderivative F of f and

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

This is (the second part of) the *fundamental theorem of calculus*. This establishes a connection between the geometrically defined integral and the analytically defined antiderivative!

In this chapter we will first discuss antiderivatives of functions. We will learn some techniques for calculating antiderivatives. Then we introduce a basic definition of an integral, and prove the fundamental theorem of calculus, which gives a rather easy way of computing integrals (or areas). We will also discuss some limitations of this (too naive) approach. If time permits, Mathematics for AI 3 will see us turn to the more powerful Lebesgue integral to overcome some of these limitations.

In the previous chapter, we had restricted our attention to functions whose domain is an open interval, in order to avoid issues with differentiation which can arise at the endpoint of a set. We would like to slightly broaden our universe of functions in this chapter, and so we consider a generalisation of the open interval, namely *open sets*.

Definition 6.1. *Let $\Omega \subset \mathbb{R}$. We call Ω an* ***open set*** *if*

$$\forall x \in \Omega, \exists \epsilon > 0 : U_\epsilon(x) \subset \Omega,$$

where $U_\epsilon(x) = (x - \epsilon, x + \epsilon)$ (we have seen this notation earlier; recall Definition 3.2). That is, no matter how close x comes to the boundary of Ω we can always construct a small interval around x which stays inside Ω.

Let $\Omega^c := \mathbb{R} \setminus \Omega$ denote the ***complement*** *of Ω. Then Ω is said to be a* ***closed set*** *if Ω^c is open.*

Open intervals (a, b) and $\Omega = \mathbb{R}$ are open. Also sets of the form (a, ∞), and their unions, are open. Therefore, also $\mathbb{R} \setminus \{0\} = (-\infty, 0) \cup (0, \infty)$ is open. Similarly, we obtain that closed intervals $[a, b]$ are closed since $[a, b]^c = \mathbb{R} \setminus [a, b] = (-\infty, a) \cup (b, \infty)$ is open. Therefore, singleton sets $\{a\}$, that contain only one element, are also closed.

When we consider open sets in the forthcoming material, you can think of these, for simplicity, as open (or closed) interval, or unions of them. In concrete examples, the open sets we consider are usually either $\mathbb{R}$ or an open interval, perhaps with a finite number of points removed.

6.2 Antiderivatives

Definition 6.2. *Let Ω be an open set. If $F : \Omega \to \mathbb{R}$ is a differentiable function such that*

$$F'(x) = f(x) \qquad \forall x \in \Omega,$$

then we call F an ***antiderivative*** *of f. We also use the notation*

$$F = \int f(x)\, dx = \int f\, dx$$

to say that F is a antiderivative of f, and call f the ***integrand****.*

This definition seems easy to handle since we are experienced in computing derivatives. However, when we compute the derivative of a differentiable function, we always ended up with a (unique) function, and we had a pointwise criterion for deciding if a function is differentiable. This is now different since we want to find a function F, but we only have information about its derivative $F' = f$. This is not enough to end up with a unique antiderivative F. To see this, note that knowing the slope in each point does not give any information about the function values at all. This is because a function with the same derivative might be at any "height". Let us write this down mathematically. For any functions f and F , and $c \in \mathbb{R}$, we have that

$$F \text{ is an antiderivative of } f \Leftrightarrow F + c \text{ is an antiderivative of } f.$$

For this, we only used that the derivative of a constant function equals zero. In particular, if a function has an antiderivative, then it has infinitely many.

The notation we use here could potentially be slightly confusing. The correct meaning of $F = \int f(x)\, dx$ is just "F is an antiderivative of f", which is not an actual equality, but the derivatives of both sides have to coincide everywhere. Indeed, since $F + c$ is also an antiderivative of f, we can write

$$F = \int f(x)\, dx = F + c.$$

However, this does not mean that $F = F + c$! They are of course different functions.

Let us start with the easy example of the exponential function e^x, which does not change under differentiation.

Example - For the exponential function, we know that $(e^x)' = e^x$, and therefore that $F(x) = e^x$ is *one possible* antiderivative, i.e.,

$$e^x = \int e^x\, dx.$$

However, if one asks for *all* antiderivatives of e^x, then we have to take $F(x) = e^x + c$ for arbitrary $c \in \mathbb{R}$, i.e.,

$$e^x + c = \int e^x\, dx.$$

In most applications, it is enough to know just one of the antiderivatives, and therefore we mostly omit the constant c. However, keep in mind that an antiderivative is not unique.

Example - Now consider $f(x) = \frac{1}{x}$ on $\mathbb{R} \setminus \{0\}$. Let us show that

$$\int \frac{1}{x}\, dx = \ln |x|.$$

First of all, we know from our work in the previous chapter (see page 189) that the derivative of $\ln x$ is equal to $1/x$ on its whole domain, that is, for all $x > 0$. Hence, we can take the antidervative to be $F(x) = \ln(x)$ for $x > 0$. But $\ln(x)$ is not defined for $x < 0$ and therefore, it is not obvious how to define $F(x)$ in this range such that $F'(x) = 1/x$. However, it is easy to verify that, for $x < 0$, the function $\ln |x| = \ln(-x)$ is well defined and $(\ln(-x))' = \frac{1}{-x} \cdot (-1) = 1/x$. This proves the claim.

This is already an example that shows, that it might be hard to find the antiderivative of a given function, but it is easy to verify that a function is an antiderivative. (Hint: Always double check your antiderivative by calculating its derivative!)

Next we provide a list of antiderivatives, many of which will be used throughout the rest of the chapter. All of them follow by differentiating the right hand side. (Do this again as an exercise!)

$$\begin{aligned}
\int a^x\, dx &= \frac{a^x}{\ln a}, \quad a > 0, a \neq 1. \\
\int x^a\, dx &= \frac{x^{a+1}}{a+1} \\
\int \frac{1}{x}\, dx &= \ln |x| \\
\int \cos x\, dx &= \sin x \\
\int \sin x\, dx &= -\cos x \\
\int \frac{1}{\cos^2 x}\, dx &= \tan x \\
\int \frac{1}{\sin^2 x}\, dx &= -\frac{1}{\tan x} \\
\int \frac{1}{1+x^2}\, dx &= \tan^{-1} x
\end{aligned}$$

All the antiderivatives $\int f\, dx$ above exist on the whole domain where f is defined.

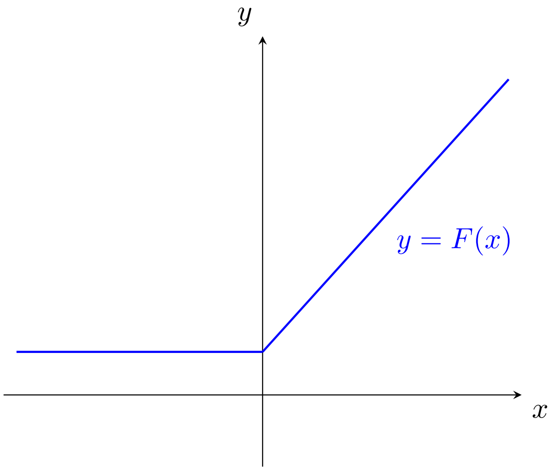
However, not all functions have an antiderivative, as the following example shows.

Example - Consider again the Heaviside function $f : \mathbb{R} \to \mathbb{R}$, with $f(x) = 1$ for $x \geq 0$, and $f(x) = 0$ for $x < 0$.

Suppose that an antiderivative F exists. Then we must have:

- $F(x) = c$, for $x < 0$ (this is necessary to ensure that $F'(x) = 0$ for $x < 0$), and
- $F(x) = x + b$ for $x > 0$, for some $b \in \mathbb{R}$ (this is necessary to ensure that $F'(x) = 1$ for all $x > 0$).

It remains to consider $x = 0$. Since F has to be differentiable, it has to be continuous (recall Theorem 5.4), and we obtain that $b = c$. However, it is not difficult to use the formal definition of the derivative to check that F cannot be differentiable at 0. If we limit ourselves to non-rigorous intuition, we can see the problem here; F has a kink at $x = 0$.



6.3 Calculation rules for antiderivatives

We will now present some rules that are useful when we want to find the antiderivative of a complicated function, which is composed of some elementary functions, like the ones given in the previous section. This is similar to the approach taken in each of the last 3 chapters.

However, and unfortunately, although we were able to determine the derivative for nearly any combination of 'easy' functions, it is much harder to find an antiderivative. In fact, a very common strategy is to guess an antiderivative, and then to verify it by calculating its derivative. For this, one clearly needs to be well-practiced in calculating derivatives. Moreover, it is sometimes just impossible to determine a closed formula for the antiderivative, even for 'easy looking' functions like e^{-x^2}.

The first calculation rule for antiderivatives, which directly follows from the corresponding rules for derivatives, is linearity.

Lemma 6.3. *Let Ω be an open set and $F, G : \Omega \to \mathbb{R}$ such that $F = \int f(x)\,dx$ and $G = \int g(x)\,dx$. Then for all $\alpha, \beta \in \mathbb{R}$,*

$$\alpha F + \beta G = \int (\alpha f(x) + \beta g(x))\,dx.$$

Proof. We only have to verify that the derivative of the function on the left equals the one in the integral on the right. By Theorem 5.5, we have

$$(\alpha F + \beta G)' = \alpha F' + \beta G' = \alpha f + \beta g,$$

since $F' = f$ and $G' = g$.

□

Now let us see some examples.

Example - We have

$$\int (4x^3 + 6x^2)\,dx = 4\int x^3\,dx + 6\int x^2\,dx = 4\left(\frac{x^4}{4}\right) + 6\left(\frac{x^3}{3}\right) = x^4 + 2x^3.$$

Example - We have

$$\int (\sqrt{x} + x)^2\,dx = \int (x + x^2 + 2x^{3/2})\,dx = \frac{x^2}{2} + \frac{x^3}{3} + 4\frac{x^{5/2}}{5}.$$

Example - In some cases, one may need some modifications of the integrand to bring it to the right form:

$$\begin{aligned}\int \frac{x^2 - x^4}{1 - x^4}\,dx = \int \frac{x^2 - 1 + 1 - x^4}{1 - x^4}\,dx = \int \frac{x^2 - 1}{1 - x^4} + 1\,dx &= \int \frac{x^2 - 1}{(1 - x^2)(1 + x^2)} + 1\,dx \\ &= (-1)\int \frac{1}{1 + x^2}\,dx + \int 1\,dx \\ &= x - \tan^{-1} x\end{aligned}$$

In the same way as we used linearity of differentiation above, we can also deduce rules for antiderivatives by utilising other calculation rules for derivatives that we studied in chapter 5.

Recall that the *product rule* (Theorem 5.6) states that

$$(fg)' = f'g + fg'. \tag{130}$$

whenever the derivatives exist. Putting this into the language we have been building in this section, this statement says that fg is an antiderivative of $f'g + fg'$. By rearranging this, we obtain the rule of **integration by parts**.

Lemma 6.4 (Integration by Parts)**.** *Let f and g be differentiable functions. Then*

$$\int f'g\,dx = fg - \int fg'\,dx. \tag{131}$$

Again, we should remember that any "equality" involving antiderivatives is not a real equality, but a mathematical statement. What (131) really says is "an antiderivative of $f'g$ is given by $fg - F$, where F is any antiderivative of fg'".

Let us see a few examples of how this rule is typically applied. Usually, the idea is to make the integrand fg' on the right hand side simpler. So, we often look to choose g in such a way that its derivative becomes easier to deal with than g itself. An example of a function with this property is $g(x) = x$.

Example - Consider

$$\int x\sin x\,dx.$$

We can take $g(x) = x$ and $f(x) = -\cos x$. Therefore, $g'(x) = 1$ and $f'(x) = \sin x$. We get

$$\int x\sin x\,dx = -x\cos x - \int(-\cos x)\,dx = -x\cos x + \int \cos x\,dx = -x\cos x + \sin x.$$

When higher degree polynomials appear in the product of the integrand, we can often deal with them via repeated applications of integration by parts.

Example - Consider

$$\int x^2e^x\,dx.$$

We can take $g(x) = x^2$ and $f(x) = e^x$. Therefore, $g'(x) = 2x$ and $f'(x) = e^x$. We get

$$\int x^2e^x\,dx = e^xx^2 - 2\int e^xx\,dx. \tag{132}$$

We still have an integral that we cannot immediately compute. However, it is another good candidate for integration by parts, and the second application will make things even simpler. We now define $g(x) = x$ and $f(x) = e^x$ and calculate that

$$\int e^xx\,dx = e^xx - \int e^x\,dx = e^xx - e^x = e^x(x-1). \tag{133}$$

Inserting (133) into (132) and rearranging gives

$$\int x^2 e^x \, dx = e^x (x^2 - 2x + 2).$$

The key to this example is that x^2 vanishes after taking enough derivatives. The same idea can be extended to higher degree polynomials, with more applications of integration by parts required as the degree increases.

Exercise - Calculate

$$\int (x^3 - 3x + 1) \sin x \, dx.$$

This degree reduction does not occur if we consider the product of a trigonometric and an exponential function. However, in such cases, it may happen that we reach the same integral again after some steps of partial integration. This can be used to calculate the antiderivative.

Example - We try to calculate

$$\int e^x \cos x \, dx.$$

Take $f'(x) = f(x) = e^x$ and $g(x) = \cos x$. Then

$$\int e^x \cos x \, dx = e^x \cos x - \int e^x(-\sin x) \, dx = e^x \cos x + \int e^x \sin x \, dx. \tag{134}$$

Now we use similar application of integration by parts to evaluate $\int e^x \sin x$. We get

$$\int e^x \sin x \, dx = e^x \sin x - \int e^x \cos x \, dx. \tag{135}$$

Combining (134) and (135) gives

$$\int e^x \cos x \, dx = e^x \cos x + e^x \sin x - \int e^x \cos x \, dx.$$

A rearrangement of this gives us our solution

$$\int e^x \cos x \, dx = \frac{e^x}{2}(\cos x + \sin x).$$

The next example is a little trickier, as it utilises a convenient cancellation occurring in the product fg'.

Example - Assume we want to compute $\int \ln x \, dx$. Since we know the derivative of $\ln x$, i.e. $(\ln(x))' = 1/x$, we may choose $g(x) = \ln(x)$ above. Additionally, we choose $f(x) = x$, which satisfies $f'(x) = 1$. We obtain

$$\int \ln x \, dx = \int 1 \cdot \ln x \, dx = x \ln x - \int x \frac{1}{x} \, dx = x \ln x - x = x(\ln x - 1).$$

The next rule we want to employ is the chain rule (Theorem 5.7). For this recall that for two differentiable functions F and g, we have

$$(F \circ g)'(x) = g'(x) \cdot F'(g(x)) \tag{136}$$

If F is a antiderivative of f, then this shows us that $F \circ g$ is an antiderivative of $(f \circ g) \cdot g'$. This is called the *substitution rule*.

Lemma 6.5 (Integration by Substitution)**.** *Let $F = \int f \, dx$ and let g be a differentiable function. Then*

$$F(g(x)) = \int g'(x) \cdot f(g(x)) \, dx.$$

Let us again discuss some examples to understand this rule.

Example - Assume we want to calculate

$$\int x^6 \cos(x^7 + 1) \, dx$$

(This may also be done by making several applications of integration by parts, but that would take ages.) We may observe something useful here which suggests the use of integration by substitution; the derivative of the $x^7 + 1$ in the cosine is similar to x^6, which also appears in the integrand as part of the product. Let's write $g(x) = x^7 + 1$. Then, we have $g'(x) = 7x^6$. If we now write $f(x) = \cos(x)$ and $F(x) = \sin x$, we obtain

$$\begin{aligned} \int x^6 \cos(x^7 + 1) \, dx = \frac{1}{7} \int 7x^6 \cos(x^7 + 1) \, dx = \frac{1}{7} \int g'(x) f(g(x)) \, dx &= \frac{1}{7} F(g(x)) \\ &= \frac{\sin(x^7 + 1)}{7}. \end{aligned}$$

Example - Consider the integral

$$\int (x^3 - 1)^6 x^2 \, dx.$$

If we are paying close attention, we will notice that the derivative of $x^3 - 1$ appears in the integrand. Therefore, we set $g(x) = x^3 - 1$ and $f(x) = x^6$. This gives $g'(x) = 3x^2$ and $F(x) = \int f(x) \, dx = \frac{1}{7} x^7$. Therefore,

$$\int (x^3 - 1)^6 x^2 \, dx = \frac{1}{3} \int (x^3 - 1)^6 \cdot (3x^2) \, dx = \frac{1}{3} F(g(x)) = \frac{1}{3} \cdot \frac{1}{7} \cdot (x^3 - 1)^7 = \frac{(x^3 - 1)^7}{21}.$$

There is one particularly useful rule that follows directly from the substitution rule.

Corollary 6.6. *Let $h : \Omega \to \mathbb{R} \setminus \{0\}$ be a differentiable function. Then*

$$\int \frac{h'(x)}{h(x)} \, dx = \ln |h(x)|$$

Proof. Apply Integration by Substitution with $f(x) = \frac{1}{x}$ and $g(x) = h(x)$. Recall that

$$\int \frac{1}{x} \, dx = \ln |x|.$$

Therefore, we can take $F(x) = \ln |x|$. We then have

$$\int \frac{h'(x)}{h(x)} \, dx = \int h'(x) \cdot f(h(x)) \, dx = F(h(x)) = \ln |h(x)|.$$

□

Examples - Here is an application of Corollary 6.6.

$$\int \frac{x}{1+x^2}\,dx = \frac{1}{2}\int \frac{2x}{1+x^2}\,dx = \frac{1}{2}\ln|\underbrace{1+x^2}_{\geq 1>0}| = \frac{1}{2}\ln(1+x^2).$$

Another application is

$$\int \tan x\,dx = \int \frac{\sin x}{\cos x}\,dx = -\int \frac{-\sin x}{\cos x}\,dx = -\ln|\cos x|.$$

Let us finally recall again, that an antiderivative is not unique. It is just a function, whose derivative satisfies something. However, it can be unique, if we know in advance that it satisfies additional conditions. In particular, one function value is enough.

For example, suppose that we are looking for an antiderivative F of e^x, such that $F(0) = 0$. Our antiderivative must have the form

$$F(x) = e^x + c$$

for some $c \in \mathbb{R}$. With the extra information that $F(0) = 0$, we can determine the correct choice of c. Simply solve the equation $F(0) = 0$ for c. That is, solve

$$e^0 + c = 0.$$

We must have $c = -1$, and so the unique antiderivative satisfying this additional constraint is

$$F(x) = e^x - 1.$$

This is a special case of a so-called *initial value problem* (or *boundary value problem*): for given $f : I \to \mathbb{R}$, $x_0 \in I$ and $y_0 \in \mathbb{R}$, we want to find F such that $F' = f$ and $F(x_0) = y_0$.

Suggested exercise - Suppose that F_1 and F_2 are antiderivatives of f. Prove that the function $F_1 - F_2$ is constant.

6.4 A first definition of the integral - The Riemann Integral

As noted in the beginning of this chapter, antiderivatives are very much connected to the integral of a function. Recall that, for a given function $f : \mathbb{R} \to [0, \infty)$, and a given interval $[a, b]$, the integral $\int_a^b f(x)\,dx$ is the *area* between the x-axis and the graph of f, bounded on each side by the vertical lines $x = a$ and $x = b$. That is, the area of the set

$$\{(x, y) \in \mathbb{R}^2 : a < x < b, 0 < y < f(x)\},$$

see Figure 23.

We first have to define precisely what this means. In particular, we have to clarify which functions are integrable, and which functions (or sets) do not allow for the definition of a meaningful area. We also want our definition of the interval to allow for the possibility that the range of $f([a, b])$ contains negative values.

We now introduce one possibility of defining an integral; namely the Riemann integral. Actually, this is probably the most simple and straightforward definition, which is its main advantage. The disadvantage is that it cannot deal with some more complicated functions. We will (hopefully) consider a more sophisticated notion in Math for AI 3, which can deal with integrals for a broader class of functions.

Let us consider a continuous function $f : \Omega \to \mathbb{R}$ and a closed interval $[a, b]$. (Note that f is therefore bounded on $[a, b]$ by Corollary 4.22.) If we now want to calculate the area between the graph of f and the x-axis, then we could divide the interval $[a, b]$ into equal subintervals, and approximate the area in this subinterval just by the area of a suitable rectangle. There are many reasonable choices. For example, Figure 24 shows the (bad) approximation of the integral by using:

1. the smallest rectangle containing the area under the curve, and
2. the largest rectangle contained in the area under the curve,

when we divide the interval into only four equal subintervals.

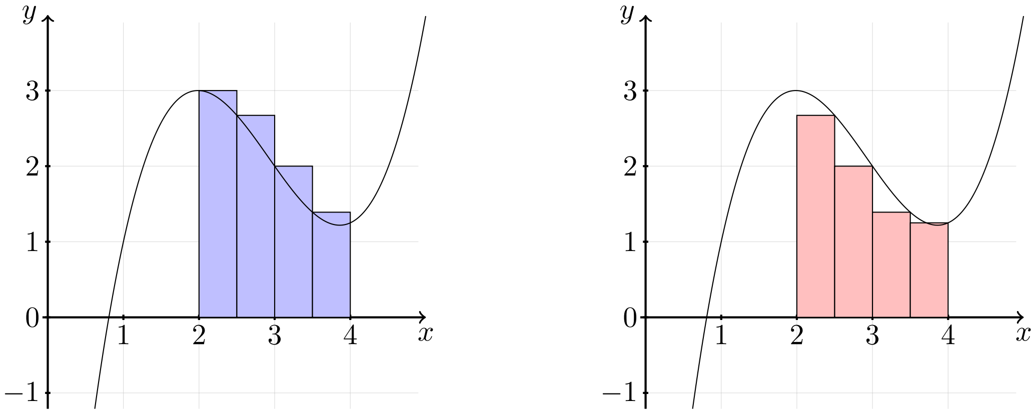


Figure 24: Estimating via the smallest rectangle containing the area under the curve, and the largest rectangle contained in the area under the curve. Both with stepsize 0.5.

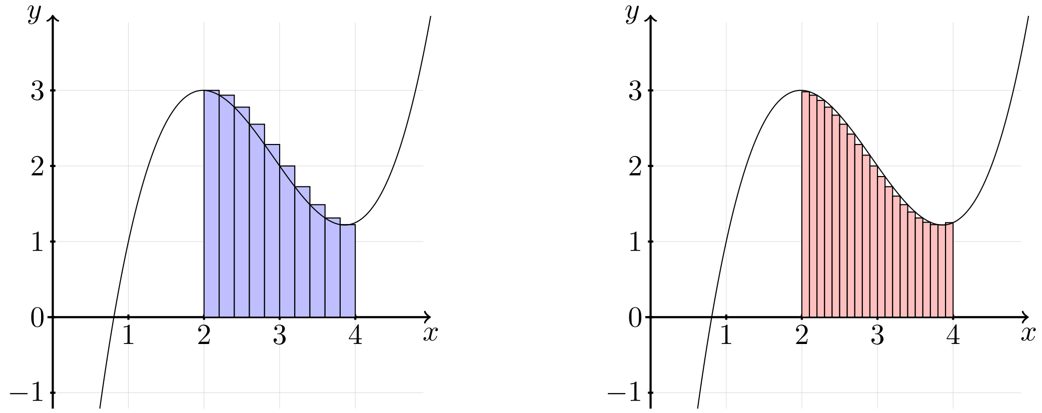


Figure 25: The error of the approximation becomes smaller with decreasing stepsize (left: stepsize 0.2; right: stepsize 0.1).

Although the resulting approximation of the integral might be quite different, this difference gets smaller and smaller if we increase the number of subintervals, see Figure 25.

This suggests that it is actually unimportant which of the rectangles we take, and we will prove that this is in fact the case when we consider *continuous functions on a closed interval.*

Therefore, for simplicity, we may just as well use the rectangles whose height is given by the left endpoint of the subinterval. Note that, in the case of monotonically decreasing functions (such as the previous two diagrams), these are the same as the upper rectangles.

The above reasoning also makes sense for possibly negative functions. In this case, the area between f and the x-axis is counted negatively. In particular, if the integral of a function is zero, then this only means that the area above the x-axis equals the area below.

Let us express this via formulae: assume we divide the interval $I = [a, b]$, which has length $b - a$, into $n \in \mathbb{N}$ subintervals of equal length, which are therefore all of length $\frac{b-a}{n}$. That is, we use the partition

$$\begin{aligned}[a,b] &= \left[a, a + \frac{b-a}{n}\right] \cup \left[a + \frac{b-a}{n}, a + 2\frac{b-a}{n}\right] \cup \dots \cup \left[a + (n-1)\frac{b-a}{n}, b\right] \\ &= \bigcup_{k=0}^{n-1} \left[a + k\frac{b-a}{n}, a + (k+1)\frac{b-a}{n}\right]\end{aligned}$$

Note that in the special case $[a, b] = [0, 1]$, this partition has the simpler form

$$[0,1] = \bigcup_{k=0}^{n-1} \left[\frac{k}{n}, \frac{k+1}{n}\right].$$

For the purpose of illustration, let us stick to the case $[a, b] = [0, 1]$. To approximate the integral of a function $f : [0, 1] \to \mathbb{R}$ first consider the first subinterval $[0, \frac{1}{n}]$. In this interval, we approximate the area below the graph by the area of the rectangle $[0, \frac{1}{n}] \times [0, f(0)]$, which is clearly $\frac{f(0)}{n}$. We then consider the second subinterval $[\frac{1}{n}, \frac{2}{n}]$, and approximate the area in this subinterval by the rectangle $[\frac{1}{n}, \frac{2}{n}] \times [0, f(\frac{1}{n})]$, which has area $\frac{f(1/n)}{n}$. We repeat this process and add up the areas of all these rectangles. Let $Q_n(f)$ denote the sum of the areas of these rectangles. We have

$$Q_n(f) = \frac{1}{n}\sum_{k=0}^{n-1} f\left(\frac{k}{n}\right). \tag{137}$$

Similarly, if we repeat the same argument for functions $f : [a,b] \to \mathbb{R}$ on an arbitrary closed interval $[a,b]$, we obtain the sum

$$\frac{b-a}{n}\sum_{k=0}^{n-1} f\left(a + k\frac{b-a}{n}\right). \tag{138}$$

To assign the function f its integral over $[a,b]$, it remains to show that these sums converge if we make n larger and larger. To keep things simple, we only deal with the case $[a,b] = [0,1]$. The general case can be proven using the same argument, making some small modifications along the way.

Moreover, as we already discussed above, our choice of the left endpoint to determine the height of the rectangles was somewhat arbitrary, and we have to justify that it is indeed irrelevant. For this, we show that one might also take the smallest or the largest of these rectangles in each subinterval, and the result would still be the same. Roughly speaking, we will show that the most different possible choices for an approximation by rectangles still converge to the same value.

To this end, define the *lower sums*

$$L_n(f) = \frac{1}{n}\sum_{k=0}^{n-1} \min\left\{f(x) : x \in \left[\frac{k}{n}, \frac{k+1}{n}\right]\right\}$$

and the *upper sums*

$$U_n(f) = \frac{1}{n}\sum_{k=0}^{n-1} \max\left\{f(x) : x \in \left[\frac{k}{n}, \frac{k+1}{n}\right]\right\}.$$

Note that the minima and maxima exist, since f is continuous on the closed intervals. This follows from Theorem 4.21.

Observe that

$$L_n(f) \le Q_n(f) \le U_n(f).$$

So, if $L_n(f)$ and $U_n(f)$ converge to the same value, then, by the sandwich rule, $Q_n(f)$ also converges to that value.

It remains to show that these two sequences do indeed converge to the same value, and this is the content of the next lemma.

Lemma 6.7. *Let $f : [0,1] \to \mathbb{R}$ be continuous. Then,*

$$\lim_{n\to\infty} L_n(f) = \lim_{n\to\infty} U_n(f).$$

In particular, both limits exist. Therefore, the sequence $(Q_n(f))_{n\in\mathbb{N}}$ also converges to the same limit.

Proof. To prove that the limits of $L_n(f)$ and $U_n(f)$ are equal, we will show that the difference

$$U_n(f) - L_n(f)$$

converges to zero. That is, we will show that for all $\epsilon > 0$ there is some $n_0 \in \mathbb{N}$ such that $|U_n(f) - L_n(f)| < \epsilon$ for all $n \geq n_0$.

First of all, note that f is continuous on a closed interval, and therefore uniformly continuous. This follows from Theorem 4.30.

Let us now fix some $\epsilon > 0$. By the uniform continuity of f, we obtain that there is some $\delta > 0$ such that

$$|x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon. \tag{139}$$

Now take $n_0 = \lceil 1/\delta \rceil + 1$. In particular, n_0 is an integer which depends on ϵ, and $n_0 > 1/\delta$. Observe that, for all $n \geq n_0$

$$1/n \leq 1/n_0 < \delta.$$

For each k, the interval $\left[\frac{k}{n}, \frac{k+1}{n}\right]$ has length $1/n$. Therefore, for any $x, y \in \left[\frac{k}{n}, \frac{k+1}{n}\right]$, we have

$$|x - y| \leq 1/n < \delta.$$

It therefore follows from (139) that

$$|f(x) - f(y)| < \epsilon.$$

Now consider $U_n(f) - L_n(f)$. This is equal to

$$\frac{1}{n}\sum_{k=0}^{n-1}\left(\max_{x \in \left[\frac{k}{n}, \frac{k+1}{n}\right]} f(x) - \min_{y \in \left[\frac{k}{n}, \frac{k+1}{n}\right]} f(y)\right) < \frac{1}{n}\sum_{k=0}^{n-1} \epsilon = \epsilon.$$

This shows that $\lim_{n\to\infty} U_n(f) - L_n(f)$ is equal to zero, which completes the proof.

□

By the above lemma combined with the Sandwich Rule, it does not matter which specific points we choose in the respective intervals. We always obtain the same limit. Therefore, we can choose our favourite points in each interval and *define* the integral of a function as the limit of the given average of the function values. We will use the left endpoints, which gives rise to the sequence $Q_n(f)$. In other words, recalling the definition of Q_n for the domain $[0,1]$ from (137), we have proven that

$$\lim_{n\to\infty} Q_n = \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} f\left(\frac{k}{n}\right)$$

exists.

Returning to the more general case of a function $f : [a,b] \to \mathbb{R}$, the sequence $Q_n(f)$ is replaced by the quantity in (138). We arrive at the following (well defined!) definition.

Definition 6.8. *Let $f : \Omega \to \mathbb{R}$ be continuous, and consider an interval $[a,b] \subset \Omega$. Then, we define by*

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \frac{b-a}{n}\sum_{k=0}^{n-1} f\left(a + k\frac{b-a}{n}\right)$$

the ***(definite) integral of f over*** *$[a,b]$. We call a and b the* ***limits of the integral****.*

The definition of the integral is not a very practical one when it comes to dealing with concrete examples. The involved limit is usually hard to determine. However, we will see in the following section that the integral can be given in terms of the antiderivative of a function. This is also the typical way of calculating integrals, and justifies the similarity of the notations. But always bear in mind that antiderivatives are functions (more precisely, classes of functions), whereas integrals are just a number.

Example Let us consider the function $f : [0, 1] \to \mathbb{R}$ given by $f(x) = x$. One does not need advanced mathematics to see that the integral (i.e. the area below the graph) is $1/2$, since it is just half of the area of the square with side-length 1. Let us see if this fits our definition. Since f is continuous, its integral is given by

$$\begin{aligned}\int_0^1 f(x)\,dx = \int_0^1 x\,dx = \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1}\frac{k}{n} = \lim_{n\to\infty}\frac{1}{n^2}\sum_{k=0}^{n-1} k = \lim_{n\to\infty}\frac{1}{n^2}\frac{n(n-1)}{2} \\ = \lim_{n\to\infty}\frac{n-1}{2n} = \frac{1}{2}.\end{aligned}$$

In the calculations above, we have used the Gauss Summation Formula (see Theorem 1.29).

From the definition of the integral as a limit, we obtain that it satisfies a list of rules, like linearity. Most of them may even already be clear from the "graphical definition". We state them without proof.

Lemma 6.9. *Let $f, g : \Omega \to \mathbb{R}$ be continuous functions, and let $[a, b] \subset \Omega$. Then,*

1. *$f = 0$ on $[a, b] \Rightarrow \int_a^b f(x)\,dx = 0$.*
2. *$\int_a^a f(x)\,dx = 0$ (i.e. the area under a curve over an interval of length zero is zero).*
3. *$\int_a^b f(x) + g(x)\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$ (Linearity).*
4. *$\int_a^b \lambda f(x)\,dx = \lambda \int_a^b f(x)\,dx$ for all $\lambda \in \mathbb{R}$.*
5. *for all $c \in [a, b]$, $\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$.*
6. *$f(x) \le g(x)$ for all $x \in [a, b] \Rightarrow \int_a^b f(x)\,dx \le \int_a^b g(x)\,dx$.*
7. *$f(x) \ge 0$ for all $x \in [a, b]$ and $[c, d] \subset [a, b] \Rightarrow \int_c^d f(x)\,dx \le \int_a^b f(x)\,dx$.*

Let us finally state some remarks on difficulties with and variants of the above definition.

Non-continuous functions. Continuity seems to be an unnecessary assumption for the considerations above. For example, Definition 6.8 makes perfect sense for indicator functions of intervals. For a given set A, the *indicator function*, which is denoted by χ_A, is defined by the formula

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}.$$

The function χ_A is not continuous over $\mathbb{R}$ if A is a proper subset of $\mathbb{R}$.

However, if we consider a closed interval $[a, b] \subset [0, 1]$ and apply Definition 6.8, then it is not hard to verify (do it please!) that

$$\int_0^1 \chi_{[a,b]}(x)\, dx = b - a,$$

which equals the true area. We will comment on such "piecewise functions" in detail later in this chapter.

However, we cannot simply remove the condition that the function is continuous and still have a reasonable definition that works for all functions, as the following example shows. If we consider the *Dirichlet function* $\chi_{\mathbb{Q}}$ instead, i.e., the indicator function of the rational numbers, then, with our definition, we would obtain

$$\int_0^1 \chi_{\mathbb{Q}}(x)\, dx = 1,$$

because all the function values we compute are at rational points. And therefore, by Lemma 6.9

$$\int_0^1 \chi_{\mathbb{R}\setminus\mathbb{Q}}(x)\, dx = \int_0^1 \chi_{\mathbb{R}}(x) - \int_0^1 \chi_{\mathbb{Q}}(x) = 1 - 1 = 0$$

However, this is unsatisfactory (and it goes against our intuition), since there are more irrational than rational numbers. One can check that Lemma 6.7 fails for this function. This shows that we have to be careful how we define an integral.

To solve this issue (at least partially) one must be more elaborate and think about a definition of an *integrable functions*. One could then define the integral for a much larger class of functions. We will consider a more powerful definition, namely the Lebesgue integral, in Math for AI 3. We should again keep in mind that any generalisation of Definition 6.8 should lead to the same result when applied to a continuous function.

6.5 The fundamental theorem of calculus

We now turn to the fundamental theorem of calculus, which provides us with an easier way of determining integrals, and which shows the existence of antiderivatives for continuous functions. Recall that a differentiable function F is an antiderivative of f if and only if $F' = f$. We start with the following additional result, which is of independent interest.

Theorem 6.10 (Mean value theorem for definite integrals)**.** *Let $f, g : [a, b] \to \mathbb{R}$ be continuous functions, and assume that $g(x) \geq 0$ for all $x \in [a, b]$. Then, $\int_a^b f(x)g(x)\,dx$ exists and there exists some $y \in [a, b]$ such that*

$$\int_a^b f(x)g(x)\,dx = f(y)\int_a^b g(x)\,dx. \tag{140}$$

In particular, we have (if we take $g(x) = 1$ for all $x \in [a, b]$) that for some $y \in [a, b]$.

$$\frac{1}{b-a}\int_a^b f(x)\,dx = f(y).$$

Proof. The fact that that integral

$$\int_a^b f(x)g(x)\,dx$$

exists follows from the definition of the integral and the fact that the function fg is continuous on $[a, b]$, since f and g are continuous in this range (this was part of Theorem 4.7).

Since f is continuous, it attains its extrema on $[a, b]$ (by Theorem 4.21). Let us denote them by

$$m := \min_{x\in[a,b]} f(x)$$

and

$$M := \max_{x\in[a,b]} f(x).$$

It follows that, for all $x \in [a, b]$,

$$mg(x) \leq f(x)g(x) \leq Mg(x).$$

Therefore, by Lemma 6.9, parts 6 and 4,

$$m\int_a^b g(x)\,dx \leq \int_a^b f(x)g(x)\,dx \leq M\int_a^b g(x)\,dx$$

We define $I := \int_a^b g(x)\,dx$ and obtain

$$m \cdot I \leq \int_a^b f(x)g(x)\,dx \leq M \cdot I. \tag{141}$$

Case 1 - Suppose that $I = 0$. It then follows from (141) that

$$\int_a^b f(x)g(x)\,dx = 0.$$

It therefore follows that, for any $y \in [a, b]$

$$\int_a^b f(x)g(x)\,dx = f(y)\int_a^b g(x)\,dx,$$

since both sides are equal to zero for any $y \in [a, b]$.

Case 2 - Suppose that $I \neq 0$. In particular, this means that $I > 0$ (this is another instance of part 6 of Lemma 6.9, using the fact that $g(x) \geq 0$ for all $x \in [a, b]$). Then we divide the inequality (141) by I and get

$$m \leq \frac{1}{I}\int_a^b f(x)g(x)\,dx \leq M.$$

Due to the Intermediate Value Theorem (Theorem 4.18), f attains every value in the interval $[m, M]$ (i.e., between its extreme values). In particular, f attains the value

$$\frac{1}{I}\int_a^b f(x)g(x)\,dx.$$

That is, there is some $y \in [a, b]$ such that

$$f(y) = \frac{1}{I}\int_a^b f(x)g(x).$$

A rearrangement of this expression gives (140).

□

This is all we need to formulate the main result of this section. For this, we define

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx$$

whenever $b < a$. (Note that we had previously defined the left hand side only for $a < b$.)

Theorem 6.11 (Fundamental theorem of calculus)**.** *Let f be continuous on some interval $I \subset \mathbb{R}$, and $a \in I$. Then, the function $F : I \to \mathbb{R}$ defined by*

$$F(x) = \int_a^x f(z)\,dz \tag{142}$$

is an antiderivative of f , i.e., $F' = f$. Moreover, for any $a, b \in I$ and any antiderivative G of f, we have

$$\int_a^b f(x)\,dx = G(b) - G(a)$$

and we write

$$[G]_a^b := G(b) - G(a).$$

Proof. First we show that F as given is an antiderivative of f. We need to show that $F'(x) = f(x)$ for all $x \in I$. Therefore we calculate the derivative of F, by considering its difference quotient. Let $x \in I$ be arbitrary. For $h \neq 0$, we have

$$\begin{aligned}\frac{F(x+h)-F(x)}{h} &= \frac{1}{h}\left(\int_a^{x+h} f(z)\,dz - \int_a^x f(z)\,dz\right)\\ &= \frac{1}{h}\left(\int_x^{x+h} f(z)\,dz\right),\end{aligned}$$

where the last identity uses Lemma 6.9, part 5.

By the mean value theorem for definite integrals (Theorem 6.10), there is some $y_h \in [x, x+h]$ such that

$$\frac{1}{h}\int_x^{x+h} f(z)\,dz = f(y_h).$$

So, we have

$$\frac{F(x+h)-F(x)}{h} = f(y_h). \tag{143}$$

This is valid for any $h > 0$. We repeat this process, taking h to be smaller and smaller. As h tends to zero, y_h tends to x (more formally, the Sandwich Rule implies that y_h tends to x). That is,

$$\lim_{h\to 0} y_h = x.$$

Since f is continuous, it then follows that

$$\lim_{h\to 0} f(y_h) = f(x).$$

Taking the limit of (143) as h goes to zero, we have

$$F'(x) = \lim_{h\to 0}\frac{F(x+h)-F(x)}{h} = \lim_{h\to 0} f(y_h) = f(x).$$

This proves that $F' = f$.

For the second part, let us first test that it is valid for the antiderivative (142) that we have just defined. We plug in a and b and obtain

$$F(b) - F(a) = \int_a^b f(z)\,dz - \int_a^a f(z)\,dz = \int_a^b f(z)\,dz.$$

Moreover, note that different antiderivatives differ only by a constant. Therefore, if G is another antiderivative, it must take the form $G(x) = F(x) + c$. Therefore,

$$G(b) - G(a) = (F(b)+c) - (F(a)+c) = F(b) - F(a) = \int_a^b f(y)\,dy,$$

as required.

□

With this very important and powerful theorem we can calculate many integrals easily (at least if you know many antiderivatives). Let us reconsider an earlier example.

Example - Consider the function $f : [0,1] \to \mathbb{R}$ given by $f(x) = x$. We already used the definition of the integral to show that

$$\int_0^1 x\,dx = \frac{1}{2}.$$

This may also be shown by considering the function $F(x) = \frac{x^2}{2}$, which is an antiderivative of f. We therefore have from the fundamental theorem of calculus that

$$\int_0^1 x\,dx = F(1) - F(0) = \frac{1}{2}.$$

We now consider some more complicated examples, which may be difficult to compute only with the definition of the integral as a limit.

Example - We proved in the earlier in this chapter that $F(x) = e^x(x^2 - 2x + 2)$ is an antiderivative of the function $f(x) = x^2 e^x$. Therefore

$$\int_0^1 x^2 e^x = F(1) - F(0) = e - 2.$$

Example - We proved earlier in this chapter that $F(x) = x(\ln x - 1)$ is an antiderivative of the function $f(x) = \ln(x)$, where both f and F are defined for all $x > 0$. Therefore

$$\int_1^e \ln(x) = F(e) - F(1) = e(\ln e - 1) - 1(\ln 1 - 1) = 1.$$

We will finally present (again) the rules for integration, that we already discussed in the section about antiderivatives. Although they can be directly deduced from there, we state them again for definite integrals to clarify their meaning.

Let us start with integration by parts for definite integrals.

Lemma 6.12. *Let f and g be continuous and differentiable on a closed interval $[a,b]$. Then,*

$$\int_a^b f'(x)g(x)\,dx = [fg]_a^b - \int_a^b f(x)g'(x)\,dx.$$

Recall that the notation $[h]_a^b$ is shorthand for $h(b) - h(a)$.

Proof. Recall from the product rule (130) that fg is an antiderivative of $f'g + fg'$. Therefore, by the fundamental theorem of calculus,

$$[fg]_a^b = \int_a^b f'(x)g(x) + f(x)g'(x)\,dx.$$

By Lemma 6.9, part 3, it follows that

$$[fg]_a^b = \int_a^b f'(x)g(x)\,dx + \int_a^b f(x)g'(x)\,dx,$$

and a rearrangement of this completes the proof. □

Example - If we want to calculate

$$\int_0^\pi x \sin x \, dx$$

we set $f'(x) = \sin x$ and $g(x) = x$. This implies that $g'(x) = 1$ and $f(x) = -\cos x$. Lemma 6.12 yields

$$\begin{aligned}\int_0^\pi x \sin x \, dx &= [-x \cos x]_0^\pi - \int_0^\pi -\cos x \, dx \\ &= [-x \cos x]_0^\pi + \int_0^\pi \cos x \, dx \\ &= (-\pi \cos \pi) + [\sin x]_0^\pi \\ &= \pi.\end{aligned}$$

Next we consider again integration by substitution, see Lemma 6.5.

Lemma 6.13. *Let $I = [a, b]$, f be continuous and g be continuous and differentiable on I. Then,*

$$\int_a^b g'(x) f(g(x)) \, dx = \int_{g(a)}^{g(b)} f(y) \, dy$$

Note that, usually, we have a (complicated looking) integral like the one on the left and want to transform it to a more easy one, like the one on the right. We will come to some examples soon.

Proof. First of all, we recall the chain rule, see (136),

$$(F \circ g)'(x) = g'(x) \cdot F'(g(x))$$

Apply this for some antidervative F of f (i.e. with $F' = f$). Such an antiderivative exists since f is continuous (this is part of the statement of the Fundamental Theorem of Calculus).

Then, we use the fact that $(F \circ g)(x)$ is an antiderivative of $g'(x) \cdot F'(g(x))$, and the Fundamental Theorem of Calculus, to get

$$\int_a^b g'(x) \cdot f(g(x)) \, dx = \int_a^b g'(x) \cdot F'(g(x)) \, dx = (F \circ g)(b) - (F \circ g)(a) = F(g(b)) - F(g(a)).$$

Another application of the fundamental theorem of calculus (this time using the fact that F is an antiderivative of f) gives

$$\int_{g(a)}^{g(b)} f(y) \, dy = F(g(b)) - F(g(a)).$$

Combining these two identities, the proof is complete.

□

Let us see some examples.

Example - Consider the integral

$$\int_0^\pi \sin(2x)\,dx.$$

We can set $g(x) = 2x$ and $f(x) = \sin x$. Then $g'(x) = 2$, and we can apply Lemma 6.13 as follows:

$$\int_0^\pi \sin(2x)\,dx = \frac{1}{2}\int_0^\pi 2\sin(2x)\,dx = \frac{1}{2}\int_0^{2\pi} \sin(y)\,dy$$
$$= \frac{1}{2}[-\cos y]_0^{2\pi} = \frac{1}{2}(-\cos(2\pi) - (-\cos(0)) = 0.$$

Example - Consider the integral

$$\int_1^2 \frac{1}{42-2x}\,dx.$$

We can set $g(x) = 42 - 2x$ and $f(x) = \frac{1}{x}$. Then we can apply Lemma 6.13, to get

$$\int_1^2 \frac{1}{42-2x}\,dx = -\frac{1}{2}\int_1^2 -2\cdot\frac{1}{42-2x}\,dx = -\frac{1}{2}\int_{40}^{38} \frac{1}{y}\,dy = \frac{1}{2}[\ln|y|]_{38}^{40} = \frac{1}{2}(\ln(40) - \ln(38)).$$

Let us finish this section with a **warning** regarding a typical mistake.

Example - Consider the function $f(x) = \frac{1}{x^2}$, and let us try to compute

$$\int_{-1}^1 f(x)\,dx.$$

When we try to use the strategies we used so far, and are not careful enough, then we might use the antiderivative of f, which we know to be $F(x) = \frac{-1}{x}$, and apply the fundamental theorem of calculus to calculate that

$$\int_{-1}^1 f(x)\,dx = F(1) - F(-1) = -1 - 1 = -2.$$

However, this is clearly wrong since the integral of a positive function must be positive. But where is the mistake? The problem is that f is not continuous on $[-1, 1]$, which is necessary for the fundamental theorem. In fact, the function is not even defined at $x = 0$. This shows that it is important to verify all assumptions before we apply such a theorem. Otherwise, we can end up proving nonsense.

6.6 Improper integrals

So far we discussed how to calculate integrals (or areas below graphs), whenever the function and the corresponding interval is bounded. However, one might imagine that we can also calculate integrals of functions over unbounded intervals if the function $f(x)$ is “small enough” for large x. By a similar reasoning, we can integrate functions with a pole at the boundary, i.e., functions that diverge to infinity at the boundary of the interval, if the divergence is “fast enough”. We discuss both cases.

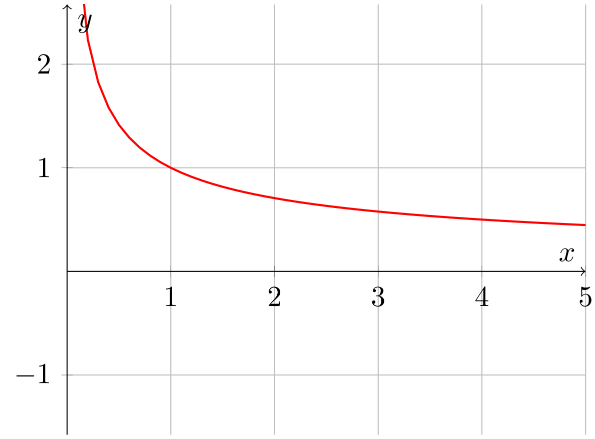


Figure 26: The function $x \mapsto \frac{1}{\sqrt{x}}$ has a pole, but its divergence is “fast enough”.

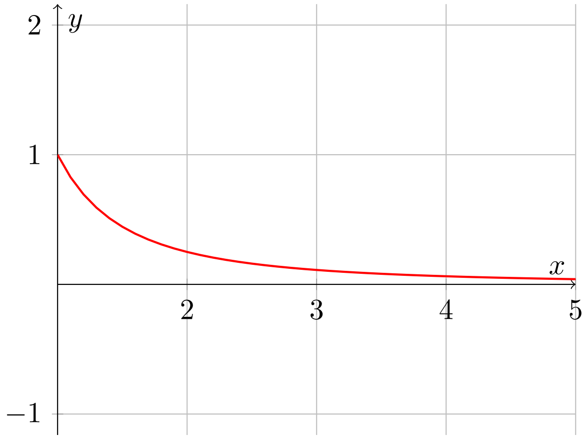


Figure 27: The function $x \mapsto \frac{1}{x^2}$ is “small enough”.

Let us start with unbounded intervals. In this case, we define the integrals

$$\int_a^\infty f\,dx := \lim_{b\to\infty} \int_a^b f\,dx \tag{144}$$

$$\int_{-\infty}^b f\,dx := \lim_{a\to-\infty} \int_a^b f\,dx \tag{145}$$

whenever the integrals and limits on the right hand side exist. If these limits exist and are finite, we then say that the integrals on the left *converge*.

We also define

$$\int_{-\infty}^\infty f\,dx := \int_{-\infty}^0 f\,dx + \int_0^\infty f\,dx. \tag{146}$$

However, note that this sum is not guaranteed to be defined, even if both of the improper integrals on the right hand side are defined. For instance (and speaking a little roughly here), if $\int_{-\infty}^0 f\,dx = \infty$ and $\int_0^\infty f\,dx = -\infty$, the the sum

$$\int_{-\infty}^0 f\,dx + \int_0^\infty f\,dx = \infty - \infty$$

is not defined. At the end of this section, we will discuss an alternative definition for the integral on the left hand side of (146).

From the fundamental theorem of calculus, we know how to compute the finite integrals on the right hand side of (144) and (145). That is, if F is an antiderivative of f (on the corresponding interval), then

$$\int_a^b f\,dx = F(b) - F(a)$$

To compute the above limits, it is therefore enough to compute the limits for the antiderivative. That is, if we denote

$$F(-\infty) = \lim_{a\to-\infty} F(a) \quad F(\infty) = \lim_{b\to\infty} F(b),$$

then we obtain

$$\int_a^\infty f\,dx := F(\infty) - F(a)$$
$$\int_{-\infty}^b f\,dx := F(b) - F(-\infty),$$

whenever the corresponding limits exist and are finite.

Very roughly and informally, we can pretend that the values $\pm\infty$ are numbers and apply the fundamental theorem of calculus, provided that the corresponding limits $F(\infty)$ and $F(-\infty)$ exist and are finite. In the case when say $F(\infty) = \pm\infty$, we can still work in this way by adopting the convention that the "equations"

$$\infty + c = \infty, \quad -\infty + c = -\infty.$$

are valid for any $c \in \mathbb{R}$.

Let us see some examples.

Example - Consider $f(x) = \frac{1}{x^\alpha} = x^{-\alpha}$ on $[1,\infty)$, where $\alpha > 1$. An antiderivative of f is given by

$$F(x) = \frac{1}{1-\alpha} x^{1-\alpha}.$$

Then by the definitions above, we have

$$\int_1^\infty x^{-\alpha} = F(\infty) - F(1) = \lim_{b\to\infty} F(b) - F(1) = \left(\lim_{b\to\infty} \frac{1}{1-\alpha} b^{1-\alpha}\right) - \frac{1}{1-\alpha} = \frac{1}{\alpha-1}.$$

In the above calculations, it is crucial that $\alpha > 1$. For $\alpha < 1$, the same function F as above is an antiderivative. We make small modifications to the above calculations, and obtain

$$\int_1^\infty x^{-\alpha} = F(\infty) - F(1) = \lim_{b\to\infty} F(b) - F(1) = \left(\lim_{b\to\infty} \frac{1}{1-\alpha} b^{1-\alpha}\right) - \frac{1}{1-\alpha} = \infty - \frac{1}{1-\alpha}.$$

That is, for $\alpha < 1$, the integral

$$\int_1^\infty x^{-\alpha}$$

diverges to ∞. We can indeed write that

$$\int_1^\infty x^{-\alpha} = \infty.$$

Exercise - What about when $\alpha = 1$? Does the integral

$$\int_1^\infty x^{-1}$$

converge or diverge?

Exercise - Show that

$$\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx = \pi.$$

In the above arguments, we see that it is not necessary for the antiderivative to be defined at the limits of the integral. (Note that a function on $\mathbb{R}$ is never defined at $\pm\infty$; we can only compute limits.) This can also be used if a function is defined up to but not including a boundary point. For simplicity, let us restrict ourselves to the case of half-open intervals. That is, we consider the integral $\int_a^b f(x)\,dx$ for functions of the form $f : (a, b] \to \mathbb{R}$ or $f : [a, b) \to \mathbb{R}$.

Consider for example the function $f(x) = x^{-1/2}$ on $(0, 1]$. This function is continuous on $(0, 1]$, and therefore has an antiderivative on $(0, 1]$ (in particular, we can take $F(x) = 2x^{1/2}$ in this range). However, if we try to compute the integral

$$\int_0^1 \frac{1}{x^{1/2}}\,dx$$

by using the fundamental theorem directly, we would need an antiderivative of f which is defined at 0. However, $f(0)$ is not defined, so it makes no sense to ask for a function whose derivative equals f at 0, i.e., there cannot be an antiderivative at 0.

However, we might guess that the integral is $F(1) - F(0) = 2$. This guess turns out to be correct. Let us now tackle this question more formally.

Consider a continuous function $f : (a, b] \to \mathbb{R}$, and let $F : (a, b] \to \mathbb{R}$ be one of its antiderivatives. We then define

$$F(a) = \lim_{x \to a^+} F(x),$$

if this limit exists. Again, if the limits exist, we then define

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

One can say that we replace F by its continuous extension to $[a, b]$, when it exists, and then use this extension in the fundamental theorem.

Similarly, to deal with the case of functions defined on a half-open interval $[a, b)$, we define

$$F(b) = \lim_{x \to b^-} F(x).$$

Returning to our example $f(x) = x^{-1/2}$, with antiderivative $F(x) = 2x^{1/2}$, we have

$$\int_0^1 \frac{1}{x^{1/2}}\,dx = F(1) - F(0) = 2 - 0 = 2.$$

Example - Consider the integral

$$\int_0^1 \ln x\, dx$$

We saw earlier in this chapter that $F(x) = x(\ln x - 1)$ is an antiderivative of $\ln x$. Therefore,

$$\int_0^1 \ln x\, dx = F(1) - F(0) = F(1) - \lim_{x\to 0^+} F(x).$$

It remains to calculate $\lim_{x\to 0^+} F(x)$. This can be done using l'Hospital's rule, as follows:

$$\lim_{x\to 0^+} x(\ln x - 1) = \lim_{x\to 0^+} \frac{\ln x - 1}{1/x} = \lim_{x\to 0^+} \frac{x^{-1}}{-x^{-2}} = \lim_{x\to 0^+} -x = 0.$$

Therefore,

$$\int_0^1 \ln x\, dx = F(1) - \lim_{x\to 0^+} F(x) = F(1) - 0 = -1.$$

Exercise - Let $\alpha \in \mathbb{R}$. Calculate the integral

$$\int_0^1 x^\alpha\, dx,$$

for all values of α for which the integral exists.

An alternative definition for the integral over the whole real line

Following an interesting discussion during the previous lecture, I am adding some additional comments about the definition in (146). An alternative definition for this integral could be given by

$$\int_{-\infty}^{\infty} f\, dx = \lim_{a\to\infty} \int_{-a}^{a} f\, dx. \tag{147}$$

The definitions look very similar; perhaps they are even equivalent?

In fact, the definitions are not the same, and this becomes apparent when one considers *antisymmetric* functions. For instance, one may compare the value of

$$\int_{-\infty}^{\infty} 2x\, dx$$

given by the two definitions. An antiderivative is given by $F(x) = x^2$. The original definition (146) then gives us

$$\begin{aligned}\int_{-\infty}^{\infty} 2x\, dx &= \int_{-\infty}^{0} 2x\, dx + \int_0^{\infty} 2x\, dx \\ &= \lim_{a\to-\infty} [x^2]_a^0 + \lim_{a\to\infty} [x^2]_0^\infty \\ &= -\infty + \infty.\end{aligned}$$

The expression $-\infty + \infty$ is meaningless, and so, according to the definition given in (146), the integral $\int_{-\infty}^{\infty} 2x\, dx$ is not defined.

On the other hand, evaluating this integral using the alternative definition (147), we obtain

$$\int_{-\infty}^{\infty} 2x\,dx = \lim_{a\to\infty}\int_{-a}^{a} 2x\,dx = \lim_{a\to\infty}[x^2]_{-a}^{a} = \lim_{a\to\infty} a^2 - (-a)^2 = 0.$$

This feels like a better answer, as it takes into account the cancellation between the positive area on the positive side of the y-axis with the negative area on the other side.

However, despite this difference, there is a connection between the two possible definitions. One can check that, whenever (146) gives a finite value, (147) gives the same outcome. So, one can view the definition in (147) as an extension of that given by (146).

Despite the advantages of the definition given by (147), we use (146) as the definition for the integral between $-\infty$ and ∞, as this is consistent with the accepted definition for the *improper integral.*

6.7 Piecewise continuous functions

Let us finally comment on functions, which have some desired properties only *piecewise*.

Definition 6.14. *Let $I = [a, b]$. We say that a function $f : I \to \mathbb{R}$ is piecewise continuous if and only if there exist a finite number of points $x_1, \ldots, x_m \in I$ such that*

- *f is continuous on every subinterval $[a, x_1), (x_m, b]$ and (x_k, x_{k+1}) for $k = 1, \ldots m - 1$.*
- *the limits $\lim_{x \to x_k^-} f(x)$ and $\lim_{x \to x_k^+} f(x)$ exist and are finite.*

We call $x_1, \ldots, x_m$ the (finite) discontinuities of f.

A simple example of a function for which such piecewise considerations might be necessary is the indicator function of an interval $[c, d] \subset \mathbb{R}$. That is,

$$\chi_{[c,d]}(x) := \begin{cases} 1 & \text{if } x \in [c, d] \\ 0 & \text{otherwise} \end{cases} .$$

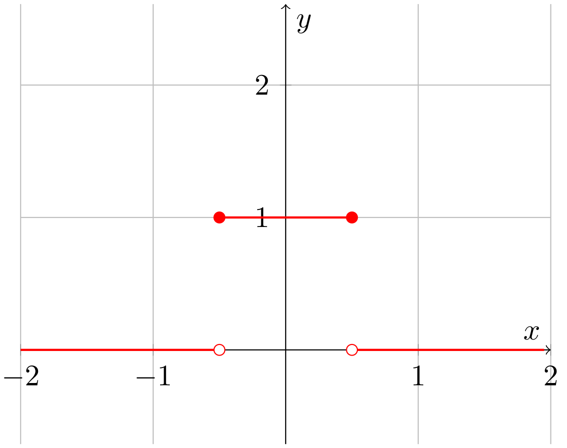


Figure 28: The rectangle function $\sqcap(x) := \chi_{[-\frac{1}{2},\frac{1}{2}]}(x)$

However, one might also think about other piecewise defined functions, like

$$f(x) := \begin{cases} -x^2 & \text{if } x < 0 \\ 2x^2 + 1 & \text{if } x \in [0, 1] \\ x & \text{if } x > 1 \end{cases} .$$

These functions are clearly not continuous on $\mathbb{R}$. However, when restricting to the individual "pieces" of the functions then they are continuous. Note also that the one-sided limits at the discontinuities exist and are finite. Therefore, both functions are piecewise continuous.

Now, to compute the integral of such piecewise countinuous functions, we can just split the integral into the corresponding parts and then use the respective rules for calculating integrals. That is, if $f : [a, b] \to \mathbb{R}$ is a piecewise continuous function with discontinuities at $x_1, \cdots, x_m$, then we use

$$\int_a^b f(x)\,dx = \int_a^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \cdots + \int_{x_{m-1}}^{x_m} f(x)\,dx + \int_{x_m}^b f(x)\,dx.$$

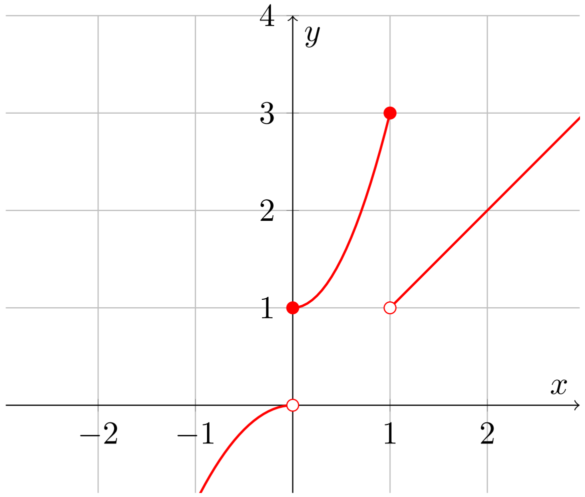


Figure 29: Previous example function

Example - For the function $\chi_{[c,d]}$ defined above, we have

$$\begin{aligned}
\int_{-\infty}^{\infty} \chi_{[c,d]}(x)\,dx &= \int_{-\infty}^{c} \chi_{[c,d]}(x)\,dx + \int_{c}^{d} \chi_{[c,d]}(x)\,dx + \int_{d}^{\infty} \chi_{[c,d]}(x)\,dx \\
&= \int_{-\infty}^{c} 0\,dx + \int_{c}^{d} 1\,dx + \int_{d}^{\infty} 0\,dx \\
&= [x]_c^d = d - c.
\end{aligned}$$

This agrees with the answer that we get by drawing the graph of the function and simply observing that the area under the graph is $d - c$.

However, note that the subintervals (x_k, x_{k+1}) are open intervals. Therefore, formally, we need to treat the integrals as improper integrals. The assumption about the one-sided limits ensures that these integrals always exist.

Example - Recall the piecewise continuous function introduced earlier:

$$f(x) := \begin{cases} -x^2 & \text{if } x < 0 \\ 2x^2 + 1 & \text{if } x \in [0,1] \\ x & \text{if } x > 1 \end{cases} .$$

Let us calculate

$$\int_{-1}^{2} f(x)\,dx.$$

We split this integral into 3 parts.

$$\begin{aligned}\int_{-1}^{2} f(x)\,dx &= \int_{-1}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{2} f(x)\,dx\\ &= \int_{-1}^{0} -x^2\,dx + \int_{0}^{1} 2x^2+1\,dx + \int_{1}^{2} x\,dx\\ &= -\frac{1}{3}\left[x^3\right]_{-1}^{0} + \left[\frac{2}{3}x^3+x\right]_{0}^{1} + \frac{1}{2}\left[x^2\right]_{1}^{2}\\ &= -\frac{1}{3}(0+1) + \left(\frac{5}{3}-0\right) + \frac{1}{2}(4-1) = \frac{17}{6}.\end{aligned}$$