University of New England

School of Science and Technology

MTHS120

Calculus and Linear Algebra 1

Lecture Notes

Trimester 1, 2024

Contents

1 Sets and functions

Sometimes mathematics is called the "language of science", which is certainly one of its functions. The language of mathematics is "set theory". For our purposes a naive approach to set theory is sufficient. It will allow us to use the terminology and notations of sets, which are ubiquitous in modern mathematics.

The basic statement is: *Sets* are collections of *elements*. This statement introduces two new notions, namely sets and elements, and a relation between them, namely 'the element belongs to the set' or 'the set contains the element'. We can define a set by listing its elements. Think of a set as a container and the elements as its content. Any object can be an element of a set. In fact, we can even form sets that are collections of other sets. However, a set can never be an element of itself.

Usually, we denote sets by upper case letters. We list the elements of a set in curly brackets. The order of the listed elements does not matter.

♦Example. Define a set A by $A = \{1, 2, 3\}$. This is the set of the first three positive integers. In this case $1, 2, 3$ are the elements of A. We write

$$1 \in A, \quad 2 \in A, \quad 3 \in A, \quad 4 \notin A.$$

A set can have infinitely many elements. In this case it is impossible to list its elements and one needs other ways for stating what the elements of the set are. The number of elements of a set is called its *cardinality*. We denote the cardinality of a set A by $|A|$. At this stage we will only discuss the cardinality of finite sets, leaving the fascinating subject of infinite cardinalities aside.

The following infinite sets are particularly important and this is why we denote them by special letters:

- The set of natural numbers (non-negative integers) $\mathbb{N} = \{0, 1, 2, \dots\}$. The dots indicate that we continue the pattern suggested by the three listed elements.
- The set of integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$. Again, the dots at both sides indicate that we continue the pattern.
- The set of positive integers $\mathbb{Z}_+ = \{1, 2, 3, \dots\}$.
- The set of rational numbers (=fractions) $\mathbb{Q} = \{\frac{p}{q} \mid p \in \mathbb{Z}, q \in \mathbb{Z}_+\}$. The notation we just used means that the set consists of elements of a certain form, namely placeholders p and q for numerator and denominator, separated by a horizontal bar (the familiar way of writing a fraction) and then, following

the vertical bar | we specify from which set the placeholders p and q can be taken.[1]

We will say that two sets A and B are equal if each element of A is also an element of B and vice versa. In this case we write $A = B$.

♦Example. $\{1, 2, 3\} = \{3, 2, 1\}$, since the order in which we list the elements does not matter. $\{1, 2, 3\} \neq \{1, 2, 4\}$, because 4 does not belong to the first set (and 3 does not belong to the second set).

Equal sets have the same cardinality but sets with equal cardinality do not have to be equal. E.g. $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$ have the same cardinality $|A| = |B| = 3$ but they are not equal.

We say that A is a subset of B and write $A \subseteq B$ if any element of A is also an element of B. It follows that $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.

If A is not a subset of B we write $A \not\subseteq B$.

We also write $A \subset B$ if $A \subseteq B$ and $A \neq B$. In this case we say that A is a *proper* subset of B.

♦Example. $\{1, 2\} \subset \{1, 2, 3\}$.

There is a special set that has no elements (think of an empty container). This set is called the *empty set* and denoted by $\varnothing$. The empty set is a subset of any set A: $\varnothing \subseteq A$. The cardinality of $\varnothing$ is $|\varnothing| = 0$.

Sets that contain exactly one element are sometimes called *singletons*.

We can construct new sets from given sets using several *set operations*.

Union of sets. The union of two sets A and B is the set C that contains precisely the elements of A or B. We write $C = A \cup B$. Formally,

$$c \in A \cup B \text{ if and only if } c \in A \text{ OR } c \in B.$$

Notice that the logic "OR" is not exclusive. The OR statement is true if the first part is true or the second part is true or both parts are true.

♦Example. $\{1, 2, 3\} \cup \{2, 3, 4\} = \{1, 2, 3, 4\}$.

[1]NB. The definition of rational numbers given here is not entirely correct. It would suggest that $\frac{1}{2}$ and $\frac{2}{4}$ are different elements of the set, however they represent the same rational number.

♠ *Exercises 1.* Show that, for any two sets A and B, $A \subseteq A \cup B$.

Proof. For any element $a \in A$ it is the case that $a \in A$ or $a \in B$. Therefore, $a \in A \cup B$. □

Intersection of sets. The intersection of two sets A and B is the set C that contains precisely the elements common to A and B. We write $C = A \cap B$. Formally,

$$c \in A \cap B \text{ if and only if } c \in A \text{ AND } c \in B.$$

♦Example. $\{1, 2, 3\} \cap \{2, 3, 4\} = \{2, 3\}$.

♠ *Exercises 2.* Show that, for any two sets A and B, $A \cap B \subseteq A$.

Difference of sets. The difference of two sets A and B is the set C that contains precisely the elements that belong to A but do not belong to B. We write $C = A \setminus B$. Formally,

$$c \in A \setminus B \text{ if and only if } c \in A \text{ AND } c \notin B.$$

♦Example. $\{1, 2, 3\} \setminus \{2, 3, 4\} = \{1\}$.

♠ *Exercises 3.* Show that, for any two sets A and B, $A \setminus B \subseteq A$.

Cartesian product of sets. The Cartesian[2] product of two sets A and B is the set C that consists precisely of the pairs (a, b), where $a \in A$ and $b \in B$. We write $C = A \times B$. Formally,

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

♦Example. $\{1, 2, 3\} \times \{a, b\} = \{(1, a), (2, a), (3, a), (1, b), (2, b), (3, b)\}$.

If A and B are finite sets of cardinality m and n respectively, then the cardinality of $A \times B$ is mn.

♠ *Exercises 4.* Is it true of false that $A \times B = B \times A$ for two sets A and B? Justify your answer.

[2]In honour of the French mathematician and philosopher René Descartes, the inventor of coordinates.

Functions. Functions (or mappings) assign to each element of a set X, called the domain, an element of a set Y, called the codomain of the function. Think of the elements of the domain as inputs of some procedure, which produces some output belonging to the codomain Y. Functions are usually denoted by lower case letters f, g, h etc. or Greek letters φ, ψ etc. We write

$$f(x) = y$$

if the function assigns the output y to the input x.

♦Example. Consider the following function with domain $X = \mathbb{Z}$ and codomain $Y = \mathbb{Z}$, i.e. domain and codomain are both the set of integers. Consider the rule of assigning to $x \in \mathbb{Z}$ its square $f(x) = x^2$. We write

$$f\colon \mathbb{Z} \to \mathbb{Z}$$

when we want to emphasise that f is a function from the domain $\mathbb{Z}$ into the codomain $\mathbb{Z}$. We write

$$f\colon x \mapsto f(x) = x^2$$

when we want to emphasise that the function assigns to each input its square.

The input of a function is called its *argument* or independent variable. The output of a function is called the *value* (for a given argument) or the dependent variable. In the example above the value of the function was produced by a simple formula. Most of the functions we encounter in this unit will be based on algebraic formulae.

Note: A function assigns to each element x of the domain one and only one value $y = f(x)$. However the function may produce the same value for different arguments, as in the example above. Indeed, for the different inputs $x = 1$ and $x = -1$ the function f takes the value $y = 1$. Not all elements of the codomain have to be actual values of the function. Again, in the example above $y = -1$ cannot be the value for any argument x, because the square of any integer is non-negative.
We call the set of all $x \in X$ such that $f(x) = y$ the *preimage* of y. The preimage of y is denoted by $f^{-1}(y)$. We may also refer to the elements of $f^{-1}(y)$ as the preimages of y.

For a function $f\colon X \to Y$ we define the *range* as the subset of the codomain that consists of the elements that are actually values of f for some arguments. In our example from above the range R consists of all perfect squares $R = \{0, 1, 4, 9, 16, \dots\}$.

Functions can be given in various ways. One way to define a function $f\colon X \to Y$ is by its graph, which is the subset of the Cartesian product $S \subseteq X \times Y$ that consists of all pairs (x, y) such that $x \in X$ and $y = f(x)$. Formally,

$$S = \{(x, y) \mid x \in X, y = f(x)\}.$$

If the domain X of a function $f\colon X \to Y$ is finite we can list the pairs $(x, f(x))$ in a table.

♦Example. Consider the function $f\colon \{1,2,3,4\} \to \mathbb{Z}$ such that $f(x) = 2x - 1$. We can table this function by

x	1	2	3	4
$f(x)$	1	3	5	7

This is more efficient than listing the graph $S = \{(1,1),(2,3),(3,5),(4,7)\}$. The range of this function is the set $R = \{1,3,5,7\}$.

Functions are one of the main objects in this unit. They are extremely useful in modelling quantities that depend in a deterministic way on other quantities. Some quantities can be easily measured, e.g. time, length or force, whereas other quantities may not be easily accessible. There can be known relations between such quantities, e.g. from physical or economic laws. Such relations often allow us to represent the unaccessible quantity as a function of an easily accessible quantity. The following is a simple example. Imagine you have a box with a large amount of screws of the same mass $12g$ and the empty box weighs $216g$. Then the total mass x of the box containing the screws is

$$x = 216 + 12y,$$

where y is the number of screws. Instead of tediously counting the screws we can quickly weigh the box with the screws and compute the number of screws as the function

$$y = \frac{x - 216}{12} = \frac{x}{12} - 18.$$

Here it makes sense to chose the domain to be the integers that are greater than or equal to 216 and divisible by 12. For our model we can restrict the domain also from above, say by $12,216$ if we know that there is no way that more than 1000 screws fit into the box. For the codomain we can choose the non-negative integers (say, smaller than or equal to 1000).

We conclude this lecture by some other important notions.

A function $f\colon X \to Y$ is called *surjective* (or sometimes a "function onto") if the range $R = Y$. Some textbook authors consider only surjective functions. This is not very practicable because it may be difficult to find the exact range of a function and for many purposes we do not need to know it. We can always turn a function into a surjective function by shrinking the codomain Y to its subset R.

♠ *Exercises 5.* Give an example of a surjective function and an example of a non-surjective function.

A function $f\colon X \to Y$ is called *injective* (or sometimes a "1-to-1 function") if it assigns to different arguments $a \neq b \in X$ different values, i.e. $f(a) \neq f(b)$.

♠ *Exercises 6.* Give an example of an injective function and an example of a non-injective function.

Functions that are both injective and surjective are called *bijective.*

Compositions of functions

If $f\colon X \to Y$ is a function with codomain Y and $g\colon Y \to Z$ is a function with domain Y we can form the composition $g \circ f\colon X \to Z$ ("f followed by g") which assigns to $x \in X$ the value $g(f(x)) \in Z$. This assignment can be expressed by

$$x \mapsto y = f(x) \mapsto z = g(y) = g(f(x)).$$

♦Example. The function $h(x) = 2x+1\colon \mathbb{R} \to \mathbb{R}$ can be expressed as the composition of $y = f(x) = 2x\colon \mathbb{R} \to \mathbb{R}$ and $z = g(y) = y + 1\colon \mathbb{R} \to \mathbb{R}$.

Inverse functions. Sometimes problems can be solved by swapping the roles of the independent and the dependent variable. In fact, in the example above we started with a relation where the number y of screws was the input variable and the weight x was the output variable

$$x = 216 + 12y$$

and we solved this equation for y, so that y became the new output variable

$$y = \frac{x}{12} - 18.$$

The two functions are inverse to each other. In general, for a function $y = f(x)$ we find the inverse function (if it exists) by solving the equation for x, i.e.

$$x = g(y),$$

where $g(y)$ is some expression (formula) of y. It is common to denote the independent variable again by x and the dependent variable by y so that the resulting inverse function is

$$y = g(x).$$

When we inverted the function the roles of domain and codomain have also swapped. For the inverse function we need to find for each element $y \in Y$ a unique $x \in X$ such that $y = f(x)$. The inverse function assigns now to y that unique $x = g(y)$. It is easy to see that the necessary and sufficient condition for the existence of such unique x is bijectivity of f. Indeed, for any y from the codomain there exists at

least one $x \in X$ such that $f(x) = y$ if and only if any $y \in Y$ is in the range, i.e. f is surjective. Such x is unique if and only if different x correspond to different $y = f(x)$, i.e. f is injective. The inverse function of a function f is often denoted by f^{-1}. The inverse function of f^{-1} is again f. We have

$$f^{-1} \circ f(x) = x \text{ for all } x \in X \text{ and } f \circ f^{-1}(y) = y \text{ for all } y \in Y.$$

Please don't confuse this notation with $\frac{1}{f}$, which may occur when f takes numerical values and we consider the composition of the function f followed by taking the reciprocal of the value of f.

♦Example. Let $f\colon \mathbb{Q} \to \mathbb{Q}$ be given by $y = f(x) = \frac{3}{2}x + \frac{1}{4}$. This function is bijective because, for *any* $y \in \mathbb{Q}$ there is a *unique* $x \in \mathbb{Q}$ such that $f(x) = y$, namely

$$x = \frac{2}{3}y - \frac{1}{6}.$$

The expression on the right hand side defines the inverse function

$$g(y) = \frac{2}{3}y - \frac{1}{6}.$$

Renaming the new input variable by x and the new output variable by y yields

$$y = f^{-1}(x) = \frac{2}{3}x - \frac{1}{6}.$$

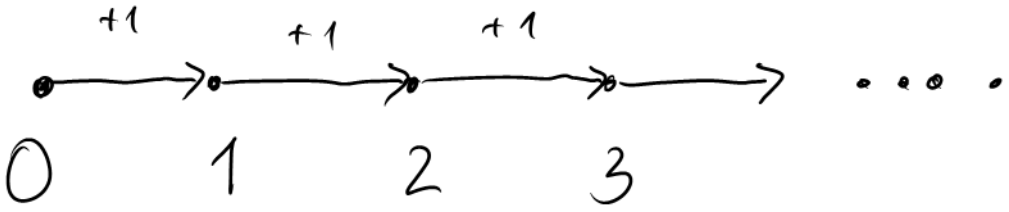
2 Numbers

We assume that you are familiar with natural numbers, integers, rational numbers, the rules of arithmetic operations and the ordering of numbers.

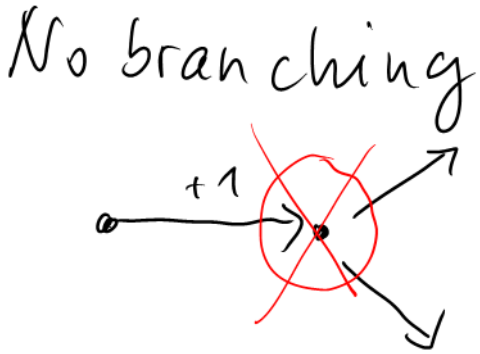
The natural numbers are defined by Peano's axioms:

1. 0 is a natural number
2. Each natural number has exactly one successor.
3. Each natural number, except 0 is the successor of exactly one natural number. 0 is not the successor of any natural number
4. (Principle of mathematical induction) A statement is true for all natural numbers if we can show that it is true for $n = 0$ (the first natural number) and that it is true for any n, assuming that it is true for its predecessor $n - 1$.

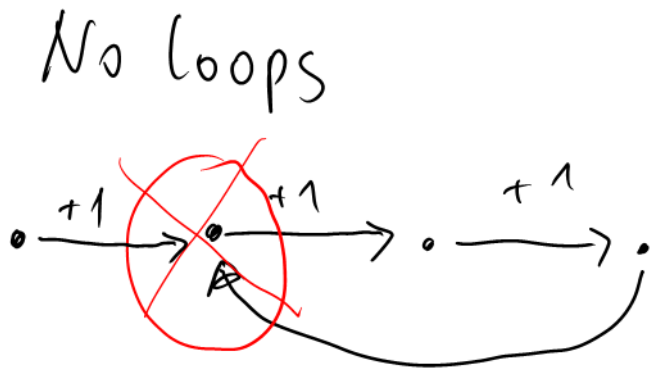
These axioms reflect the structure of natural numbers as a connected graph with one root, no branching and no loops:



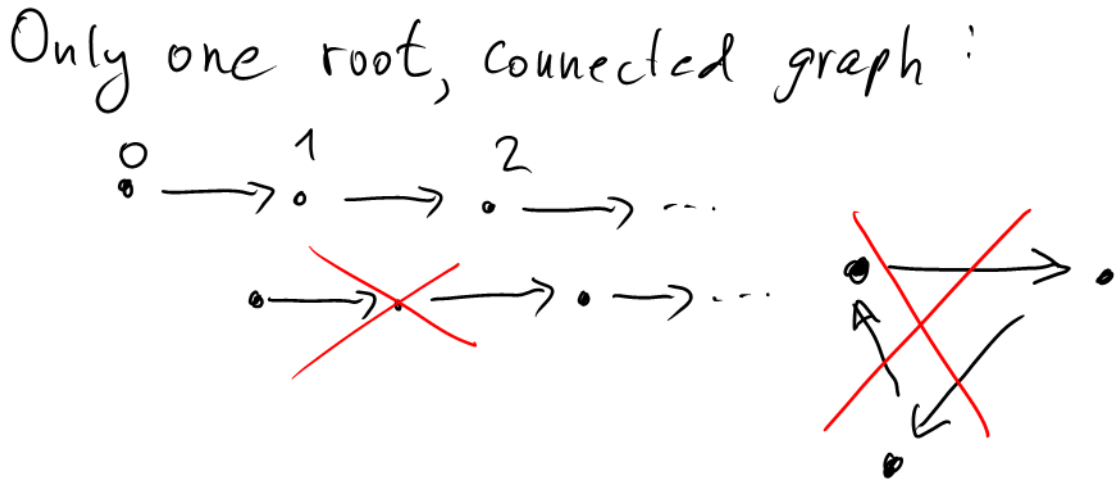
Each natural number has exactly one successor, i.e. no branching:



Each natural number, except 0 is the successor of exactly one number, i.e. no loops:



0 is not the successor of any number, i.e. 0 is the only root of the graph, and each number is connected to 0 by a sequence of consecutive successors, i.e. the graph is connected.



We will use the method of induction to prove properties of sequences.

Regarding rational numbers we assume that you are familiar with the arithmetic operations addition and multiplication, which result in rational numbers. We also assume familiarity with the ordering relation of rational numbers. Below we list the basic arithmetic and ordering properties of rational numbers:

- For each $a, b, c \in \mathbb{Q}$ we have $a + b = b + a$, $(a + b) + c = a + (b + c)$, $0 + a = a + 0 = a$, $ab = ba$, $(ab)c = a(bc)$, $1a = a1 = a$, $(a + b)c = ac + bc$.
- For each $a \in \mathbb{Q}$ there exists $y \in \mathbb{Q}$ such that $a + y = 0$.
- For each $a(\neq 0) \in \mathbb{Q}$ there exists $y \in \mathbb{Q}$ such that $ay = 1$ (the reciprocal of a).
- For every $a, b \in \mathbb{Q}$ one and only one of

$$a > b, \ a = b, \ a < b$$

is true.

- If $a > b$ and $b > c$, then $a > c$.
- If $a > b$, then $a + c > b + c$.
- If $a > b$ and $c > 0$, then $ac > bc$.

We will use these properties as axioms, i.e., we will take them for granted and use them in future proofs.

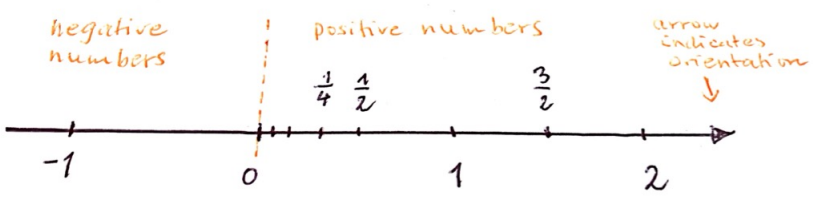
♦Example. Show that the product of two positive rational numbers is a positive rational number.

Proof. Let a and c be two positive rational numbers, i.e., $a > 0$ and $c > 0$. Now we just use the last property in the list above with $b = 0$. We multiply both sides of the valid inequality $a > 0$ with the positive number c to get $ac > 0$, as required. □

♦Example. Show that the average $c = \frac{a+b}{2}$ of two rational numbers is a rational number.

Proof. The average is a combination of addition $a + b$ followed by multiplication by the rational number $\frac{1}{2}$, so the result is a rational number. □

We will often refer to numbers as points on the number line. The number line is an oriented straight line with a choice of a point 0 (called the origin) and a point 1 (usually at the right of 0 if the line is oriented from the left to the right, which is indicated by an arrow pointing to the right). The distance between the points 0 and 1 defines a scale on the number line (=1 length unit). Then any positive number $x > 0$ can be marked as a point of distance x (length units) from 0 to the right, any negative number $x < 0$ can be marked as a point of distance $-x$ (length units) from 0 to the left and the number 0 corresponds to the point 0 on the number line. If a number b is greater than a number a then the position of b on the number line is to the right of a.



The rational numbers are "dense" on the number line in the following sense.

Proposition 1. *For any two distinct rational numbers a and b there are infinitely many rational numbers between them.*

Proof. First we show that between a and b there is a least one other rational number, namely its average

$$c = \frac{a+b}{2}.$$

It is clear (see the example above) that the average of two rational numbers is again a rational number.

Without loss of generality let us assume that $a < b$. The case $a > b$ can be handled in a similar way. We show that $a < c < b$. We know that $b - a > 0$ and need to show that $c - a > 0$. Indeed,

$$c - a = \frac{a+b}{2} - a = \frac{b-a}{2} > 0.$$

Similarly,

$$b - c = b - \frac{a+b}{2} = \frac{b-a}{2} > 0.$$

Denote $c = c_1$, Now we can find a number c_2 between a and c_1 (and hence between a and b) which is distinct from c_1, namely the average of a and c_1. By continuing this procedure we construct an infinite sequence of numbers $c_1, c_2, c_3, \ldots$, where c_{n+1} is the average of a and c_n. All the these numbers are different and between a and b because

$$b > c_1 > c_2 > c_3 > \cdots > a. \qquad \square$$

♠ *Exercises 7.* Show that the distance between the numbers a and c_n from the Proof above equals $\frac{b-a}{2^n}$.

Despite the density property of rational numbers it turns out that there are "gaps" in the number line, in the sense that there are points that do not represent rational numbers. It was known to the ancient Greeks at the time of Pythagoras that the length c of the diagonal of a square of side length 1 cannot be expressed as a rational number[3]. A rigorous proof of this fact will be given in Number theory Pmth338 in year 3. It relies on the plausible fact that any integer has a unique factorisation into primes (up to the order of the prime factors). Let's take this fact for granted and assume that the length c is rational, i.e.

$$c = \frac{p}{q}$$

for some integers p, q. According to Pythagoras's theorem we have $c^2 = 1^2 + 1^2 = 2$, hence

$$c^2 = 2 = \frac{p^2}{q^2}$$

[3]For the time being I avoid calling $c = \sqrt{2}$.

or

$$2q^2 = p^2.$$

Being a perfect square, the number p^2 on the right hand side of the equation above contains an even number of prime factors 2. On the other hand, the number $2q^2$ contains an odd number of prime factors 2, namely the even number of factors 2 from q^2 and one additional factor 2. This contradiction shows that our initial assumption that c was rational cannot be true. This is an example of an indirect proof, also known as a "proof by contradiction".

In order to fill the gaps on the number line we extend the rational numbers to the larger set of *real* numbers. We will not give a rigorous construction of the real numbers but rather develop some technical tools needed for understanding calculus. Geometrically, a real number is a point on the number line. Using decimals we can get as close to any point as we wish. E.g. the number c with $c^2 = 2$ from above is between 1 and 2, so 1 is an approximation with an error ≤ 1. Getting more precise, we could show that c is between 1.4 and 1.5, so $1.4 = \frac{7}{5}$ is an approximation with an error $\leq \frac{1}{10}$. Continuing this procedure we can get arbitrarily close to c by a sequence of rational numbers. So, roughly speaking, we can think of a real number as a decimal with "infinitely many places". This rough notion will become more precise (as the limit of a series) after we sufficiently advance in our understanding of calculus.

The set of real numbers is denoted by $\mathbb{R}$. We can define addition and multiplication of real numbers so that the result will be again a real number. Real numbers satisfy the same arithmetical and ordering properties as the rational numbers listed above:

- For each $a, b, c \in \mathbb{R}$ we have $a + b = b + a,\ (a + b) + c = a + (b + c)$, $0 + a = a + 0 = a,\ ab = ba,\ (ab)c = a(bc),\ 1a = a1 = a,\ (a + b)c = ac + bc.$
- For each $a \in \mathbb{R}$ there exists $y \in \mathbb{R}$ such that $a + y = 0$.
- For each $a(\neq 0) \in \mathbb{R}$ there exists $y \in \mathbb{R}$ such that $ay = 1$ (the reciprocal of a).
- For every $a, b \in \mathbb{R}$ one and only one of

  $$a > b,\ a = b,\ a < b$$

  is true.
- If $a > b$ and $b > c$, then $a > c$.
- If $a > b$, then $a + c > b + c$.
- If $a > b$ and $c > 0$, then $ac > bc$.

Real numbers that are not rational are called irrational numbers.

♠ *Exercises 8.* Show that the sum $a + b$ of a rational number a and an irrational number b is irrational.

♠ *Exercises 9.* Given an irrational number b. For which rational numbers a is the product ab rational?

We introduce the following absolute value function:

$$abs(x) = |x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0. \end{cases}$$

This function is defined for all real numbers x and takes non-negative real values. Geometrically, it expresses the distance from x to the origin 0 on the number line. This function can also be used to express the distance between two real numbers $a, b \in \mathbb{R}$:

$$dist(a, b) = |a - b|.$$

For a given positive number ε we will say that x is ε-close to a if

$$|x - a| < \varepsilon.$$

In this case we will also say that x is within an ε-neighbourhood of a.

♦Example. $x = \frac{5}{4}$ is $\frac{1}{2}$-close to $a = 1$; $x = \frac{3}{2}$ is not $\frac{1}{2}$-close to $a = 1$

♠ *Exercises 10.* Show that x is ε-close to a if and only if a is ε-close to x, for any $x, a \in \mathbb{R}$.

♠ *Exercises 11.* Show that the inequality $|x - a| < \varepsilon$ is equivalent to the two simultaneous inequalities

$$x < a + \varepsilon, \qquad a - \varepsilon < x.$$

The notion of ε-neighbourhood is fundamental in calculus and we will refer to it regularly.

We prove some useful properties regarding absolute values.

Proposition 2. *The absolute value has the following properties:*

1. $|a| \geq 0$ *for all* $a \in \mathbb{R}$ *and* $|a| = 0$ *if and only if* $a = 0$.
2. $|ab| = |a| \cdot |b|$, *for all* $a, b \in \mathbb{R}$.

3. $|a|^2 = a^2$, *for all* $a \in \mathbb{R}$.

4. $|a+b| \leq |a| + |b|$, *for all* $a, b \in, \mathbb{R}$. *This is the so-called* triangle inequality.

5. $|a-b| \geq ||a| - |b||$, *for all* $a, b \in \mathbb{R}$. *This is called the* reverse triangle inequality.

Proof. The first three properties are consequences of the definition and are left as exercises.

The following observation will be useful in the proof. For any number $a \in \mathbb{R}$ we have $a \leq |a|$ and $-a \leq |a|$. We consider two cases: If $a \geq 0$ then $a = |a|$ and $-a \leq 0 \leq a$. If $a < 0$ then $a < 0 \leq |a|$ and $-a = |a|$.

To prove (4) we consider two separate case. First assume $a + b \leq 0$. Then

$$\begin{aligned} |a+b| &= -(a+b) \\ &= -a - b \\ &= -a + (-b) \\ &\leq |a| + |b|, \end{aligned}$$

as $-a \leq |a|$ and $-b \leq |b|$. On the other hand, if $a + b \geq 0$ then

$$\begin{aligned} |a+b| &= a + b \\ &\leq |a| + |b|. \end{aligned}$$

And we have the required property.

We use (4) to prove (5). Note that

$$|a| = |(a-b) + b| \leq |a-b| + |b|,$$

so that

$$|a-b| \geq |a| - |b|.$$

Similarly,

$$|b| = |(b-a) + a| \leq |b-a| + |a|,$$

so that

$$|b-a| = |a-b| \geq |b| - |a|.$$

Again we consider the two possible cases: If $|a| - |b| \geq 0$ the first inequality means

$$|a-b| \geq |a| - |b| = ||a| - |b||.$$

If $|a| - |b| < 0$ then the second inequality means

$$|a-b| \geq |b| - |a| = -(|a| - |b|) = ||a| - |b||$$

So we have (5),

$$|a - b| \geq ||a| - |b||.$$

□

Before we formulate a further fundamental property of the set of real numbers we need to introduce the notion of *boundedness*.

Definition 1. *A subset $S \subseteq \mathbb{R}$ is called* bounded above *if there exists a number K such that, for any $x \in S$*

$$x \leq K.$$

In this case the number K is called an upper bound of the set S.
A subset $S \subseteq \mathbb{R}$ is called bounded below *if there exists a number k such that, for any $x \in S$*

$$k \leq x.$$

In this case the number k is called a lower bound of the set S.
A subset $S \subseteq \mathbb{R}$ is called bounded *if it is at the same time bounded below and above.*

In the statements in the definition above the wording "there exists" and "for all" appeared several times. It is convenient to use the shorthands $\exists$ for "there exists" and $\forall$ for "for all". Then the statement of boundedness above becomes:

$$\exists K \text{ such that } \forall x \in S, x \leq K.$$

Notice that whenever K is an upper bound of a set S then any number $K' > K$ is also an upper bound of S. Similarly, whenever k is a lower bound of a set S then any number $k' < k$ is also a lower bound of S.

If a set $S \subseteq \mathbb{R}$ contains an element which is an upper (lower bound) then this element is called the *maximum* (*minimum*) of S.

Proposition 3. *If a set $S \subseteq \mathbb{R}$ has a maximum M (minimum m) then M (m) is unique.*

Proof. Assume that S has two maxima M and M'. Since $M \in S$ we must have $M \leq M'$. Similarly, we also have $M' \leq M$. Therefore, $M = M'$.

The proof of the uniqueness of m is analogous. □

The following sets are called intervals.
For given $a \leq b$, the closed interval $[a, b]$ is defined as

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

the open interval (a, b) is defined as

$$(a, b) = \{x \in \mathbb{R} \mid a < x < b\}.$$

We can also define semiclosed intervals

$$[a, b) = \{x \in \mathbb{R} \mid a \leq x < b\} \text{ and } (a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}.$$

Notice that the ε-neighbourhood of a point $a \in \mathbb{R}$ is the open interval

$$(a - \varepsilon, a + \varepsilon).$$

Geometrically, intervals are segments on the number line. The endpoints are included or excluded depending on the type of the interval.

♦Example. All intervals defined above are bounded. In each case a is a lower bound and b is an upper bound. Closed intervals have the minimum a and the maximum b. Open intervals don't have a minimum or a maximum. Semiclosed intervals have either a maximum or a minimum but not both.

NB. Notice that the bounded sets from the example above are not finite, but contain infinitely many elements.

♠ *Exercises 12.* Show that the intersection of two intervals is again an interval.

Sometimes the sets $\mathbb{R}$ and the rays

$$(a, \infty), \quad [a, \infty), \quad (-\infty, b), \quad (-\infty, b]$$

are referred to as unbounded intervals.

♠ *Exercises 13.* Show that any finite set is bounded.

In view of the ambiguity of lower and upper bounds we can ask for making them as sharp as possible. More precisely, we may ask for a least upper bound and a largest lower bound.

Definition 2. *Given a set $S \subseteq \mathbb{R}$. We define the supremum (or least upper bound) of S to be the number $s = \sup S$ such that*

1. *s is an upper bound of S*
2. *for any upper bound K of S we have $s \leq K$*

We define the infimum (or largest lower bound) of S to be the number $t = \inf S$ such that

1. *t is a lower bound of S*

2. *for any upper lower bound k of S we have $k \leq t$.*

We show that infimum and supremum are unique. Indeed, let s and s' be two suprema of the set S. Then both s and s' are upper bounds and $s \leq s'$ and $s' \leq s$. Therefore, $s = s'$. In an analogous way one can show that $\inf S$ is unique.

The supremum of a set S is denoted by $\sup S$ and the infimum of S is denoted by $\inf S$.

♦Example. Use the definition to show that $\inf(a, b) = a$ and $\sup(a, b) = b$.

For any $x \in (a, b)$ we have $a < x$, so a is a lower bound. We need to show that there is no lower bound larger than a. Assume that there is a lower bound $a' > a$. We can assume that $a' \leq b$ because otherwise b would be a lower bound and we could replace a' by b. Now, $a < \frac{a+a'}{2} < a' \leq b$, hence $\frac{a+a'}{2} \in (a, b)$ and $a' > \frac{a+a'}{2}$, which contradicts the assumption that a' was a lower bound.

The proof of $\sup(a, b) = b$ is analogous. □

Now we are ready to formulate the *axiom of completeness* of the real numbers:

Any non-empty subset $S \subseteq R$ that is bounded above has a supremum.

Notice that the rational numbers do not satisfy the axiom of completeness. Indeed, consider the set $S \subseteq \mathbb{Q}$ of all rational numbers x such that $x^2 < 2$. The supremum of this set within the real numbers is the number c with $c^2 = 2$, which is not a rational number. Now, any rational number smaller than c is not an upper bound of S and for any rational number r greater than c there is a rational number $r' < r$ which is still an upper bound. Try to prove the statements in the preceding sentence. Try to prove first that for any two non-negative real numbers $a < b$ it follows $a^2 < b^2$.

The axiom of completeness makes sure that there are "enough" real numbers to fill the "gaps". To make sure that there are not "too many" real numbers we need another axiom, called the Archimedean axiom:

For any two positive real numbers x, y there is a natural number n such that $nx > y$.

This means that by adding a positive real number x sufficiently many times to itself we can make it larger than any other positive real number. The consequences of the Archimedean axiom are further explored in the Appendix. In particular, we

can show the stronger density property of the rational numbers:

For any two *real* numbers $a < b$ there infinitely many rational numbers x such that $a < x < b$.

♠ *Exercises 14.* Show that the Archimedean axiom is automatically satisfied for rational numbers, i.e., for any two positive rational numbers x, y there is a natural number n such that $nx > y$.

♠ *Exercises 15.* Use the Archimedean axiom to show that for any positive real number ε there exists a rational number $\frac{m}{n}$ such that $0 < \frac{m}{n} < \varepsilon$.

3 Sequences

In this lesson we introduce and study the notion of sequences. One of our aims is to establish a relation between sequences and the supremum of a set, as defined in the previous lecture.

Definition 3. *A sequence is a function* $f\colon \mathbb{N} \to \mathbb{R}$. *Instead of* $f(n)$ *we often use index notation* $f(n) = a_n$ *and we represent the sequence as an infinite ordered list of the numbers* $(a_0, a_1, a_s, \dots)$. *Other common notations are* $(a_n)_{n=0}^{\infty}$ *or just* (a_n).

Sequences are used in mathematical modelling for recording consecutive measurements in regular time steps over a potentially unrestricted period.

A sequence may (or may not) follow the pattern given by a formula. Although there are ways to analyse "big" measurement data using computers, having a single formula makes the analysis much easier and often reveals underlying laws. Examples of sequences defined by a formula are

1. $(n^2)_{n=0}^{\infty} = (0, 1, 4, 9, \dots)$ – the sequence of squares
2. $(2, 3, 5, 7, 11, \dots)$ – the sequence of primes
3. $(\frac{1}{n})_{n=1}^{\infty}$ – the sequence of reciprocals
4. $(2n)_{n=0}^{\infty}$ – the sequence of even natural numbers
5. $(2n+1)_{n=0}^{\infty}$ – the sequence of odd natural numbers
6. $((-1)^n)_{n=0}^{\infty} = (1, -1, 1, \dots)$ – the sequence of alternating 1 and -1.

The following two types of sequences are particularly important in modelling and have special names:

1. Sequences of the form $c_n = an + b$, where a, b are some given numbers (parameters) are called *arithmetic progressions*.
2. Sequences of the form $c_n = aq^n$, where $a \neq 0$ is a real parameter and $q \neq 1$ is a positve real parameter, are called *geometric progressions*.

In analysing a sequence (e.g. a sequence of measurements) we are interested in the following properties:

A sequence $(a_n)_{n=0}^{\infty}$ is called

1. constant if $\exists c \in \mathbb{R}$ such that $\forall n$, $a_n = c$,
2. positive (non-negative) if $\forall n$, $a_n > 0$ ($a_n \geq 0$),
3. negative (non-positive) if $\forall n$, $a_n < 0$ ($a_n \leq 0$),
4. (strictly) increasing if $\forall n$, $a_{n+1} \geq a_n$ ($a_{n+1} > a_n$),
5. (strictly) decreasing if $\forall n$, $a_{n+1} \leq a_n$ ($a_{n+1} < a_n$).

We can form new sequences from given sequences in several ways:

1. We scale the sequence (b_n) by a real number a to produce the sequence (c_n) with $c_n = ab_n$.

2. We add two sequences (a_n) and (b_n) to produce the sequence (c_n) with $c_n = a_n + b_n$.

For any sequence $(a_n)_{n=0}^{\infty} = (a_0, a_1, a_2, \dots)$ we can form the *derived sequence* [4] (also known as "sequence of consecutive differences")

$$(a'_n)_{n=0}^{\infty} = (a_{n+1} - a_n)_{n=0}^{\infty} = (a_1 - a_0, a_2 - a_1, a_3 - a_2, \dots).$$

We have used the notation (a'_n) for the derived sequence. The "prime" (dash) or other modifications of a letter like $\hat{a}$ ('a hat') or $\tilde{a}$ ('a tilde') are used in mathematics as a notation for a different, but usually a related object[5].

♠ *Exercises 16.* Show that the derived sequence of the sum of two sequences (a_n) and (b_n) equals to the sum of their derived sequences (a'_n) and (b'_n), and that the derived sequence of (ka_n) is the derived sequence (a'_n) scaled by k.

Proposition 4. *A sequence $(a_n)_{n=0}^{\infty}$ is constant if and only if its derived sequence is identically 0.*
A sequence $(a_n)_{n=0}^{\infty}$ is (strictly) increasing if and only if its derived sequence is non-negative (positive).
A sequence $(a_n)_{n=0}^{\infty}$ is (strictly) decreasing if and only if its derived sequence is non-positive (negative).

Proof. We prove the first statement. It is clear that the derived sequence of a constant sequence $a_n = c$ is $a'_n = a_{n+1} - a_n = c - c = 0$.

[4] Students who have encountered calculus before may see here some analogy with the derivative of a function. Notice that the "formula" of a derived sequence is different from the "formula" of the derivative of a function, even if sequence and function are given by the same "formula". The derivative of $f(x) = x^2$ is $f'(x) = 2x$, whereas the derived sequence of (n^2) is $(2n + 1)$.

[5] Later in this unit we will use the notation f' for the derivative of the function f.

We prove the converse by induction. We have $a_0 = c$ just by denoting a_0 by c. Now, assuming that $a_{n-1} = c$ we conclude that $a_n = a_{n-1} + a'_{n-1} = a_{n-1} + 0 = c$. This concludes the proof by induction.

The proof of the second statement is straight forward: Clearly, $a_{n+1} \geq a_n$ if and only if $a'_n = a_{n+1} - a_n \geq 0$. The same argument works in the case of strict inequalities and also for the proof of the third statement. □

Let us compute the derived sequences for some list of standard sequences:

1. The derived sequence for an arithmetic progression $c_n = an+b$ is the constant sequence $c'_n = a$. In fact, a sequence is an arithmetic progression if and only if its derived sequence is constant.

2. The derived sequence for a geometric progression $c_n = aq^n$ is the geometric progression $c'_n = a(q-1)q^n$. In particular, for q=2, the derived sequence coincides with the original geometric progression.

3. The derived sequence of the quadratic sequence $a_n = n^2$ is $a'_n = (n+1)^2 - n^2 = 2n+1$ (which is an arithmetic progression). More generally, the derived sequence of $a_n = n^p$, where p is a positive integer, is given by a polynomial formula

$$a'_n = (n+1)^p - n^p = \sum_{k=0}^{p-1} \binom{p}{k} n^k = pn^{p-1} + \cdots .$$

This polynomial is of order $p-1$ with leading term pn^{p-1}.

♠ *Exercises 17.* Compute the derived sequence for the sequence of reciprocals $a_n = \frac{1}{n}$. Here we assume that $n \geq 1$.

Another interesting problem is to find a sequence (a_n) with a given derived sequence (b_n). We have already solved this problem for the identically zero derived sequence. The resulting constant sequence involved an arbitrary parameter $c = a_0$. It is a general phenomenon that a_0 in the resulting sequence is arbitrary. In general we have

$$\begin{aligned} a_0 &= c \\ a_1 &= c + b_0 \\ &\vdots \\ a_n &= a_{n-1} + b_{n-1} = b_{n-1} + \cdots + b_0 + c = c + \sum_{\nu=0}^{n-1} b_\nu \end{aligned}$$

The sequence (s_n) with terms

$$s_n = \sum_{\nu=0}^{n} b_\nu$$

is called the sequence of partial sums of the sequence (b_n). Finding the original sequence for a given sequence is closely related to finding the sequence of partial sums:

$$a_n = c + s_{n-1}.$$

Notice the correspondence between the n-th term of the sequence (a_n) and the $(n-1)$-st term of (s_n).

In general, it can be very challenging to find a formula for the partial sums of a given sequence. We can use the formulae of the derived sequences from above to compute some partial sums.

The partial sums of the sequence of natural numbers

$$s_{n-1} = \sum_{\nu=0}^{n-1} \nu = 1 + \cdots + (n-1) = \frac{n(n-1)}{2}$$

are given by the well-known Gaussian formula. We can also compute it using our derived sequences from above:

The derived sequence of (n^2) is $(2n+1)$ and the derived sequence of (n) is (1). Subtracting the corresponding terms yields the derived sequence $(2n)$ of $(n^2 - n)$. Dividing by 2 shows that the derived sequence of $(\frac{n(n-1)}{2})$ is (n) and hence, the sequence (a_n) of partial sums of (n) is

$$s_{n-1} = \sum_{k=0}^{n-1} k = a_n + c = \frac{n(n-1)}{2} + c.$$

The constant c can be determined from $a_0 = 0$, hence $c = 0$. This yields the Gaussian formula

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

The same method works for all power sequences.

♠ *Exercises 18.* Find the sequence of partial sums for $b_n = n^2$.

Let us find the partial sums for the geometric progression:

$$s_{n-1} = a_n = \sum_{\nu=0}^{n-1} q^\nu.$$

We know that the derived sequence of (q^n) is $(q-1)q^n$, hence the derived sequence of $(\frac{q^n}{q-1})$ is (q^n). It follows that the partial sums of (q^n) are

$$s_{n-1} = a_n = \frac{q^n}{q-1} + c,$$

where $c = -\frac{1}{q-1}$ can be determined from $a_1 = 1$. This yields the well-known formula for partial sums of the geometric progression

$$\sum_{\nu=0}^{n-1} q^\nu = \frac{q^n - 1}{q-1} = \frac{1-q^n}{1-q}. \tag{1}$$

4 Limits of sequences

A sequence (of measurements) can, in its long-term behaviour, approach and stay close to a constant *limit*. In this lecture we will give the notion of limit a precise meaning.

Let (a_n) be a sequence. We would refer to the number L as the limit of the sequence if the terms a_n of the sequence get and stay arbitrarily close to L if n is sufficiently large. We have earlier introduced the notion of ε-closeness to quantify how close a_n to L is. Getting arbitrarily close means that, no matter how small a positive ε we chose, we want $|a_n - L| < \varepsilon$ as soon as n is large enough, i.e. n is greater than some number N, which depends on ε. The smaller ε we choose, the larger N becomes. Sometimes we will be able to express N as a function of ε but sometimes we will just be able to show that such N exists. Using the quantifiers $\forall$ and $\exists$ this can be expressed formally as:

$$\forall \varepsilon > 0 \quad \exists N \text{ such that } \forall n > N, |a_n - L| < \varepsilon. \tag{2}$$

This is a rather complex logical construct.

♠ *Exercises 19.* Show that the statement (2) is equivalent to the statement:

$$\forall \varepsilon > 0 \quad |a_n - L| < \varepsilon \text{ holds for all but finitely many numbers } n.$$

♠ *Exercises 20.* Show that the statement (2) is equivalent to the statement: $\forall \varepsilon > 0$ the set $\{n \mid |a_n - L| \geq \varepsilon\}$ is bounded above.

♠ *Exercises 21.* Formulate the negation of the statement (2), i.e. express that L is not the limit of the sequence (a_n).

We say that the sequence (a_n) is *convergent* if it has a limit. We write

$$\lim_{n \to \infty} a_n = L.$$

In this case we also say that the sequence (a_n) tends (or converges) to L. If a sequence does not have a limit then it is called *divergent*.

Notice that altering finitely many terms in a sequence (a_n) neither changes its convergence or divergence nor its limit in the case of convergence. Indeed, finitely many changes affect only terms a_n with $n \leq N_0$ for some number N_0. Then the statement of convergence to a limit L remains true (or remains false) by replacing N with $\max(N, N_0)$.

We illustrate the notion of limit in an example.

♦Example. Consider the sequence of reciprocals $(\frac{1}{n})$. This sequence is strictly decreasing and positive and it appears that its terms get and stay close to 0:

$$\left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\right).$$

We conjecture that the limit might be $L = 0$. For each positive ε we have to find N such that

$$\forall \varepsilon > 0 \quad \exists N \text{ such that } \forall n > N, \left|\frac{1}{n} - 0\right| < \varepsilon.$$

A good strategy is to start from the end. We want to achieve that $\left|\frac{1}{n}\right| < \varepsilon$. This is equivalent to

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

So all we need to do is to find an integer N that is greater than or equal to the real number $\frac{1}{\varepsilon}$. Then $n > N \geq \frac{1}{\varepsilon}$ implies $n > \frac{1}{\varepsilon}$ which in turn makes sure that

$$\left|\frac{1}{n}\right| < \varepsilon,$$

as required. In this case we can even give a formula for N as a function of ε

$$N = \left\lceil \frac{1}{\varepsilon} \right\rceil,$$

where $\lceil x \rceil$ is the "ceiling" function that assigns to a real number x the smallest integer that is greater than or equal to x.

♦Example. Consider the sequence of negative powers of 2, $a_n = 2^{-n}$,

$$\left(1, \frac{1}{2}, \frac{1}{4}, \dots\right).$$

This sequence is also positive and decreasing and seems to approach 0. To confirm this we need to show that

$$\forall \varepsilon > 0 \quad \exists N \text{ such that } \forall n > N, \left|2^{-n} - 0\right| < \varepsilon.$$

Again we start from the end,

$$\left|2^{-n} - 0\right| = 2^{-n} < \varepsilon$$

is equivalent to $2^n > \frac{1}{\varepsilon}$. We could solve this inequality for n if we knew the $\log_2$ function and if we knew that it is strictly increasing and therefore respects inequalities. Instead we follow another approach. We show that $2^n > n$, which is equivalent to $2^{-n} < \frac{1}{n}$, by induction: We start with $n = 1$, $2^1 = 2 > 1$ is correct. Assume, for

$n \geq 2$, $2^{n-1} > n-1$. In particular, then also $2^{n-1} > n-1 \geq 1$. Adding the two inequalities gives

$$2^{n-1} + 2^{n-1} > n - 1 + 1,$$

that is

$$2^n > n,$$

as required. Now, the choice made in the previous example also works here: If $n > N \geq \frac{1}{\varepsilon}$ then

$$2^n > n > N \geq \frac{1}{\varepsilon}$$

and hence

$$2^{-n} < \varepsilon.$$

This example shows that sometimes choosing a much larger N than the optimal one allows us to simplify the computations. Here we chose $N = \left\lceil \frac{1}{\varepsilon} \right\rceil$, rather than the smaller $N = \lceil -\log_2 \varepsilon \rceil$. There is no general recipe for this approach.

♠ *Exercises 22.* Show that the limit of a constant sequence $a_n = c$ exists and equals c.

♦Example. Most sequences have no limit. We demonstrate this for the alternating sequence $a_n = (-1)^n$, that is $(1, -1, 1, -1, \dots)$. The difficulty of showing that there is no limit is that we have to negate the limit statement for any candidate L:

$$\forall L \in \mathbb{R} \quad \exists \varepsilon > 0 \quad \forall N \quad \exists n \geq N \qquad |a_n - L| \geq \varepsilon.$$

We choose $\varepsilon = 1$. If there was a limit L then for any $m = n+1, n > N$ we would have

$$|a_m - L| < 1 \text{ and } |L - a_n| = |a_n - L| < 1.$$

By the triangle inequality we would have then

$$2 = |a_{n+1} - a_n| \leq |a_{n+1} - L| + |L - a_n| < 2,$$

which is a contraditction.

Using the "ε - N" technique we prove some properties of limits.

Proposition 5. *1. If a sequence (a_n) converges to the limit L then the sequence of absolute values $(|a_n|)$ converges to $|L|$.*
2. If the sequence (a_n) is convergent then it is bounded, i.e. there exists a number M such that $\forall n \in \mathbb{N}$, $|a_n| < M$.
3. If a sequence (a_n) has a limit $L > 0$ ($L < 0$) then $\exists N_0$ such that, for $n > N_0$, $a_n > \frac{L}{2}$ ($a_n < \frac{L}{2}$).

Proof. We prove the first statement. We know that

$$\forall \varepsilon > 0 \quad \exists N \text{ such that } \forall n > N, |a_n - L| < \varepsilon.$$

By the reverse triangle inequality,

$$||a_n| - |L|| \leq |a_n - L|.$$

Therefore, choosing for any given ε the same N yields

$$||a_n| - |L|| \leq |a_n - L| < \varepsilon,$$

as required.

For the second statement, let $\lim_{n\to\infty} a_n = L$ and, consequently, $\lim_{n\to\infty} |a_n| = |L|$. Then for $\varepsilon = 1$ there exists N_0 such that, for $n > N_0$,

$$||a_n - |L|| < 1,$$

hence, by the triangle inequality,

$$|a_n| = |(a_n - L) + L| \leq |a_n - L| + |L| < |L| + 1.$$

Now we just take M as the maximum of the finitely many numbers $|a_0|, \ldots, |a_{N_0-1}|, |L|+1$.

Finally, for the third statement, let $\lim_{n\to\infty} a_n = L > 0$. Then for $\varepsilon = \frac{L}{2}$ there exists N_0 such that, for $n > N_0$,

$$|a_n - L| < \frac{L}{2},$$

hence

$$\frac{L}{2} < a_n (< \frac{3L}{2}),$$

as required. The proof in the case of $L < 0$ is analogous. □

♠ *Exercises 23.* Show that the limit of a sequence (if it exists) is unique.

Theorem 1. *Let (a_n) and (b_n) be convergent sequences with limits K and L respectively.*

1. *The sum and the difference of the sequences $(a_n \pm b_n)$ are convergent and have the limit $K \pm L$.*
2. *The scaled sequence (ca_n) is convergent and has the limit cK.*
3. *The product of the sequences $a_n b_n$ is convergent and has the limit KL.*

4. *If $L \neq 0$ then the quotient of the sequences $\frac{a_n}{b_n}$ is convergent and has the limit $\frac{K}{L}$. In particular, there exists a number N_0, such that, for $n > N_0$, $b_n \neq 0$ and the quotient sequence is well defined for $n > N_0$.*

Proof. We know that

$$\forall \varepsilon > 0 \quad \exists N_1 \text{ such that } \forall n > N_1, |a_n - K| < \varepsilon.$$

$$\forall \varepsilon > 0 \quad \exists N_2 \text{ such that } \forall n > N_2, |b_n - L| < \varepsilon.$$

The notations N_1 and N_2 reflect that the numbers N can be different for the different sequences. In both statements we can replace N_1 and N_2 by their maximum $N = \max(N_1, N_2)$. The same statements remain true if we replace ε by the smaller number $\frac{\varepsilon}{2}$. In this case just N becomes larger. We have

$$\forall \frac{\varepsilon}{2} > 0 \quad \exists N \text{ such that } \forall n > N, |a_n - K| < \frac{\varepsilon}{2} \text{ and } |b_n - L| < \frac{\varepsilon}{2}.$$

Now, for $n > N$,

$$|a_n + b_n - K - L| \leq |a_n - K| + |b_n - L| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which proves that

$$\lim_{n\to\infty} a_n + b_n = K + L.$$

The proof of

$$\lim_{n\to\infty} a_n - b_n = K - L$$

is analogous.

We leave the proof of statement 2. as an exercise. (In fact, this is a special case of statement 3.)

Let's prove statement 3. We need to show that

$$\forall \varepsilon > 0 \quad \exists N \text{ such that } \forall n > N, |a_n b_n - KL| < \varepsilon.$$

We start again from the inequality at the end:

$$|a_n b_n - KL| < \varepsilon$$

in order to find what conditions we need to impose on N. We can rewrite

$$a_n b_n - KL = a_n b_n - a_n L + a_n L - KL = a_n(b_n - L) + (a_n - K)L$$

and apply the triangle inequality

$$|a_n b_n - KL| \leq |a_n||b_n - L| + |a_n - K||L|.$$

Our strategy is to make both $|a_n||b_n - L|$ and $|a_n - K||L|$ smaller than $\frac{\varepsilon}{2}$. Since $\lim_{n\to\infty} b_n = L$ and $|a_n| < M$ for some positive constant M,

$$\forall \varepsilon > 0 \quad \exists N_1 \text{ such that } \forall n > N_1, |b_n - L| < \frac{\varepsilon}{2M}, \text{ hence } |a_n||b_n - L| < \frac{\varepsilon}{2}.$$

If $L = 0$ we need not worry about the second term. If $L \neq 0$,

$$\forall \varepsilon > 0 \quad \exists N_2 \text{ such that } \forall n > N_2, |a_n - K| < \frac{\varepsilon}{2|L|}, \text{ hence } |a_n - K||L| < \frac{\varepsilon}{2}.$$

Let $N = \max(N_1, N_2)$. Then adding the two inequalities yields

$$\forall \varepsilon > 0 \quad \exists N \text{ such that } \forall n > N, |a_n b_n - KL| \leq |a_n||b_n - L| + |a_n - K||L| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

The proof of statement 4. is similar to the proof of statement 3. but even more technically involved. We need to show that

$$\forall \varepsilon > 0 \quad \exists N \text{ such that } \forall n > N, \left| \frac{a_n}{b_n} - \frac{K}{L} \right| < \varepsilon.$$

We have

$$\frac{a_n}{b_n} - \frac{K}{L} = \frac{a_n L - b_n K}{b_n L} = \frac{a_n L - KL + KL - b_n K}{b_n L} = \frac{(a_n - K)L - (b_n - L)K}{b_n L}$$

hence

$$\left| \frac{a_n}{b_n} - \frac{K}{L} \right| \leq \frac{|a_n - K||L| + |b_n - L||K|}{|b_n||L|}.$$

According to 3rd statement of Proposition 5., for $\varepsilon > 0$ we can choose N_1 such that, for $n > N_1$,

$$|b_n| > \frac{|L|}{2}.$$

Now, we choose N_2 such that, for $n > N_2$,

$$|a_n - K| < \frac{|L|\varepsilon}{4}$$

and we choose N_3 such that, for $n > N_3$

$$|b_n - L| < \frac{L^2 \varepsilon}{4K}$$

if $K \neq 0$ (otherwise, we need not worry about the second term). Let $N = \max(N_1, N_2, N_3)$. Then for $n > N$ we have

$$\left| \frac{a_n}{b_n} - \frac{K}{L} \right| < \varepsilon,$$

as required.

The theorem above allows us to show convergence and to compute limits of more complicated sequences, as in the example below.

♦Example. Decide whether the sequence $a_n = \frac{3n^2+n}{2n^2-1}$ converges and if so, compute the limit.

Solution. First we divide numerator and denominator by n^2, which yields

$$a_n = \frac{3 + \frac{1}{n}}{2 - \frac{1}{n^2}}.$$

Now the denominator tends to 2 because $\frac{1}{n^2} = \frac{1}{n}\frac{1}{n}$ and the limit of the constant sequence 2 is 2. Similarly, the numerator sequence tends to 3. It follows that the sequence converges to $\frac{3}{2}$.

5 Limit and supremum

In this section we investigate the relation between the notions of limit and supremum and prove some new criteria for convergence.

Proposition 6. *Let $S \subset \mathbb{R}$ be a nonempty bounded set and $\alpha = \sup S$. Then there is an increasing sequence (a_n) of elements of S that converges to α.*

Proof. Since α is the lowest upper bound of S, $\alpha - \frac{1}{n} < \alpha$ is not an upper bound for any natural number n. Therefore, for any $n \in \mathbb{N}$ there is an element $b_n \in S$ such that $b_n > \alpha - \frac{1}{n}$. We show that the sequence of (b_n) chosen in this way converges to α. Indeed,

$$\forall \varepsilon \quad \text{choose } N = \left\lceil \frac{1}{\varepsilon} \right\rceil \text{ then } \forall n > N, \alpha - \frac{1}{N} < \alpha - \frac{1}{n} < b_n \leq \alpha.$$

Since $N \geq \frac{1}{\varepsilon}$ and $\varepsilon > 0$, the latter inequality implies

$$\alpha - \varepsilon < b_n < \alpha + \varepsilon,$$

which is equivalent to

$$|b_n - \alpha| < \varepsilon.$$

This shows that the sequence (b_n) converges to α. Finally, we modify (b_n) so that it still converges to α but becomes increasing. Let $a_0 = b_0$. Assume that we have already constructed $a_0, \ldots, a_{n-1}$. Then $a_n = \max\{a_{n-1}, b_n\}$. It follows that $a_n \geq a_{n-1}$, thus (a_n) is increasing. On the other hand,

$$\forall \varepsilon, \quad N = \left\lceil \frac{1}{\varepsilon} \right\rceil \text{ then } \forall n > N, \alpha - \varepsilon < b_n \leq a_n \leq \alpha, \text{ i.e. } |a_n - \alpha| < \varepsilon \qquad \square.$$

The Proposition above means that although the supremum α of a set S may not belong to S it can be approached as the limit of a sequence of elements from S. This gives another approach to real numbers, namely as limits of sequences.

The following converse of the Proposition is a useful criterion for convergence without the need of conjecturing the actual limit.

Theorem 2. *Any increasing bounded above sequence (a_n) converges and*

$$\lim_{n \to \infty} a_n = \sup\{a_n\}.$$

Proof. Since the sequence (as a set of real numbers) is not empty and bounded it has a supremum $\sup\{a_n\} = \alpha$. We show that this is the limit of the sequence.

Indeed, clearly $a_n \leq \alpha$. On the other hand, for any $\varepsilon > 0$, $\alpha - \varepsilon$ is not an upper bound, so there is some N such that $a_N > \alpha - \varepsilon$. Since the sequence is increasing, this implies that, for all $n > N$, $a_n > \alpha - \varepsilon$. Collecting this together yields

$$\forall \varepsilon > 0 \quad \exists N \text{ such that } \forall n > N, |a_n - \alpha| < \varepsilon,$$

as required. □

♦Example. Consider the sequence $a_n = (1 + \frac{1}{n})^n$. We show that it is increasing and bounded.

We have

$$\begin{aligned} a_n = (1 + \frac{1}{n})^n &= 1 + \frac{n}{n} + \frac{n(n-1)}{2!n^2} + \cdots + \frac{n!}{n!n^n} \\ &= 1 + 1 + \frac{1}{2!}(1 - \frac{1}{n}) + \cdots + \frac{1}{n!}(1 - \frac{1}{n}) \cdots (1 - \frac{n-1}{n}) \end{aligned}$$

and

$$\begin{aligned} a_{n+1} = (1 + \tfrac{1}{n+1})^{n+1} &= 1 + \tfrac{n+1}{n+1} + \tfrac{(n+1)n}{2!(n+1)^2} + \cdots + \tfrac{(n+1)!}{(n+1)!(n+1)^{n+1}} \\ &= 1 + 1 + \frac{1}{2!}(1 - \frac{1}{n+1}) + \cdots + \frac{1}{n!}(1 - \frac{1}{n+1}) \cdots (1 - \frac{n-1}{n+1}) \\ &\quad + \frac{1}{(n+1)!}(1 - \frac{1}{n+1}) \cdots (1 - \frac{n}{n+1}). \end{aligned}$$

Now, the terms in a_n are smaller than or equal to the corresponding terms in a_{n+1} and a_{n+1} has an extra positive term, which makes $a_{n+1} > a_n$.

In order to show that the sequence is bounded above we notice that

$$\begin{aligned} a_n = (1 + \frac{1}{n})^n &= 1 + \frac{n}{n} + \frac{n(n-1)}{2!n^2} + \cdots + \frac{n!}{n!n^n} \\ &< 1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{n!} < 1 + 1 + \frac{1}{2} + \cdots + \frac{1}{2^{n-1}} = 1 + \frac{1 - 2^{-n}}{1 - \frac{1}{2}} < 1 + 2 = 3. \end{aligned}$$

We have used the formula for partial sums of a geometric progression (1). It follows that the sequence (a_n) converges. In fact, it converges to an irrational real number denoted by e (in honour of Leonhard Euler)

$$\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = \mathrm{e} \approx 2.71828.$$

We will encounter the number e many times in the future.

The following plausible comparison principle will lead to another technique of finding the limit of a sequence.

Theorem 3 (Comparison principle)**.** *Let (a_n) and (b_n) be sequences with $a_n \leq b_n$ (for all but finitely many n) and such that*

$$\lim_{n\to\infty} a_n = L \text{ and } \lim_{n\to\infty} b_n = K,$$

then $L \leq K$.

Proof. We have

$$\begin{aligned} \forall \varepsilon > 0 \quad & \exists N_1 \text{ such that } \forall n > N_1, L - \varepsilon < a_n < L + \varepsilon \\ \forall \varepsilon > 0 \quad & \exists N_2 \text{ such that } \forall n > N_2, K - \varepsilon < b_n < K + \varepsilon \end{aligned}$$

Let $N = \max(N_1, N_2)$. It follows that for any $n > N$

$$0 \leq b_n - a_n < K + \varepsilon - L + \varepsilon = K - L + 2\varepsilon.$$

In other words,

$$L - K < 2\varepsilon,$$

no matter what positive ε we choose. So $L - K$ is smaller than any positive number, hence it is smaller than or equal to 0. This proves the claim $L \leq K$. □

NB. Even if the strict inequalities $a_n < b_n$ hold in the theorem above, the limits may not satisfy the strong inequality as the following example shows: Let $a_n = 0$ and $b_n = \frac{1}{n}$. Then $a_n < b_n$, but

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = 0.$$

We prove now the *Squeeze theorem* (also known as squeezing principle).

Theorem 4 (Squeeze theorem)**.** *Let (a_n), (b_n) and (c_n) be sequences with $a_n \leq b_n \leq c_n$ (for all but finitely many n) and such that*

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n = L,$$

then (b_n) also converges and

$$\lim_{n\to\infty} b_n = L.$$

Proof. Similar to the proof of the theorem above, we have

$$\begin{aligned} \forall \varepsilon > 0 \quad & \exists N \text{ such that } \forall n > N, L - \varepsilon < a_n < L + \varepsilon \\ \forall \varepsilon > 0 \quad & \exists N \text{ such that } \forall n > N, L - \varepsilon < c_n < L + \varepsilon \end{aligned}$$

Combining the inequalities yields, for $n > N$,

$$L - \varepsilon < a_n \leq b_n \leq c_n < L + \varepsilon,$$

that is,

$$|b_n - L| < \varepsilon,$$

as required. □

♦Example. Let $b_n = \frac{\sin n}{n}$. Here we assume that you are familiar with the sine function and its basic properties from high school. In particular,

$$-1 \leq \sin n \leq 1,$$

since the opposite site to an angle in a right triangle cannot be longer than the hypotenuse. Therefore,

$$a_n = \frac{-1}{n} \leq \frac{\sin n}{n} \leq c_n = \frac{1}{n}.$$

Now,

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n = 0$$

and therefore, by the Squeeze theorem,

$$\lim_{n\to\infty} \frac{\sin n}{n} = 0.$$

At the end of this lecture we investigate a sequence of rational numbers that tends to $\sqrt{2}$, that is the real number the square of which is 2. The sequence we consider is given in a *recursive* way: the term a_{n+1} is given as a formula of a_n. Let $a_0 = 1$ (or any other positive number). Then we define

$$a_{n+1} = \frac{a_n + 2}{a_n + 1} = 1 + \frac{1}{a_n + 1}.$$

The first five terms of the sequence are:

$$a_0 = 1, \quad a_1 = \frac{3}{2} = 1.5, \quad a_2 = \frac{7}{5} = 1.4, \quad a_3 = \frac{17}{12} \approx 1.417, \quad a_4 = \frac{41}{29} \approx 1.4138$$

This sequence is not monotone (neither increasing nor decreasing). We will show that it oscillates around $\sqrt{2}$ with the even terms undershooting and the odd terms overshooting $\sqrt{2}$. We use induction, $a_0 = 1 < \sqrt{2}$ (since $1^2 < 2$). If $a_n < \sqrt{2}$ then

$$a_{n+1} = 1 + \frac{1}{a_n + 1} > 1 + \frac{1}{\sqrt{2} + 1} = \frac{2 + \sqrt{2}}{\sqrt{2} + 1} = \sqrt{2}.$$

Similarly, if $a_n > \sqrt{2}$,

$$a_{n+1} = 1 + \frac{1}{a_n + 1} < 1 + \frac{1}{\sqrt{2} + 1} = \frac{2 + \sqrt{2}}{\sqrt{2} + 1} = \sqrt{2}.$$

Next we show that the sequence of distances $\alpha_n = |a_n - \sqrt{2}|$ tends to 0. Indeed,

$$\left|a_{n+1} - \sqrt{2}\right| = \left|\frac{a_n + 2}{a_n + 1} - \sqrt{2}\right| = \left|\frac{a_n(1-\sqrt{2}) + 2 - \sqrt{2}}{a_n + 1}\right|$$
$$= \frac{\sqrt{2} - 1}{a_n + 1}\left|\sqrt{2} - a_n\right| < \frac{1}{2}\left|a_n - \sqrt{2}\right|.$$

We have used $\frac{1}{a_n+1} < 1$ and $\sqrt{2} - 1 < \frac{1}{2}$ (since $\sqrt{2} < \frac{3}{2}$, since $2 < \frac{9}{4}$). Therefore,

$$\alpha_{n+1} < \frac{1}{2}\alpha_n,$$

hence

$$\alpha_n < \alpha_0 2^{-n}.$$

It follows

$$0 \leq \alpha_n \leq \alpha_0 2^{-n}$$

and by the squeeze theorem $\lim_{n\to\infty} \alpha_n = 0$.

If we knew that the sequence (a_n) converges we could compute the limit L as follows: Let n pass to infinity in

$$a_{n+1} = \frac{a_n + 2}{a_n + 1}.$$

This gives

$$L = \frac{L + 2}{L + 1},$$

which is equivalent to

$$L^2 = 2.$$

Since all terms of a_n are positive, the only option is $L = \sqrt{2}$.

It is also instructive to look at this example in the following way: Interpret the even terms a_{2n} of the sequence as the left ends and the consecutive odd terms a_{2n+1} as the right ends of the intervals $I_n = [a_{2n}, a_{2n+1}]$. Then $\sqrt{2}$ belongs to all intervals I_n. The sequence of intervals I_n is *nested* in the sense that

$$I_0 \supseteq I_1 \supseteq I_2 \supseteq \cdots .$$

On the other hand, the lengths of those intervals $\beta_n = a_{2n+1} - a_{2n}$ tends to zero. Indeed,

$$\beta_n = \frac{a_{2n} + 2}{a_{2n} + 1} - a_{2n} = \frac{2 - a_{2n}^2}{a_{2n} + 1} \to 0.$$

This means that the sequence of intervals I_n is *contracting*. The axiom of completeness of the real numbers can be replaced by the statement:

Any sequence of contracting nested closed intervals has exactly one common element.

In general, we can describe real numbers as the common element of a sequence of contracting nested closed intervals, as we have done above for $\sqrt{2}$.

6 Infinity as a limit

We say that a sequence (a_n) has the limit ∞ if for any (large) number M only finitely many members of the sequence are smaller than M. We can assume that M is a natural number. Formally we can express this by

$$\forall M \quad \exists N \text{ such that } \forall n > N, a_n > M.$$

In this case we write

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

Infinity is not a number and cannot be treated as such. A sequence that tends to infinity is divergent.

♦Example. $\lim_{n\to\infty} n = \infty$, indeed, for any M there exists N (namely, $N = M$) such that, for $n > M$,

$$a_n = n > N = M.$$

We say that a sequence (b_n) tends to $-\infty$ if $a_n = -b_n$ tends to ∞.

♠ *Exercises 24.* Show that for any sequence (a_n) which tends to ∞ the sequence

$$b_n = \frac{1}{a_n}$$

tends to 0.

The following sequences tend to infinity: $(cn + b)$, where $c > 0$; n^2; (cr^n), where $c > 0$ and $r > 1$; $(n!)$, where $n! = 1 \cdot 2 \cdots n$; (n^n). It turns out that some sequences tend to infinity faster than others. Before we investigate this further we prove

Proposition 7. *If a sequence (a_n) is increasing and unbounded above then*

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

Proof. Since the sequence (a_n) is unbounded

$$\forall M \quad \exists N \text{ such that } a_N > M.$$

Since (a_n) is increasing, $\forall n > N$,

$$a_n \geq a_N.$$

This combines into

$$\forall M \quad \exists N \text{ such that } \forall n > N, a_n > M,$$

that is

$$\lim_{n\to\infty} a_n = \infty. \qquad \square$$

It is easy to show that the sequences above are increasing and we leave this as an exercise. We show that the sequences are unbounded and hence tend to infinity.

1. Let $a_n = cn + b$ with $c > 0$. Then for an arbitrary number M the inequality

$$a_n = cn + b > M$$

is equivalent to

$$n > \frac{M - b}{c}.$$

2. Let $a_n = n^2$. We could also try and solve the inequality $a_n = n^2 > M$. This would involve the notion of square root, which we discuss later. Instead, we use the crude estimate $n^2 > n$, hence $a_n > n$. Therefore, $a_n > M$ for $n > M$.

3. Let $a_n = cr^n$, with $c > 0$ and $r > 1$. Again we could try and solve $a_n = cr^n > M$. This would require the notion of logarithm, which again we leave for later. We use a proof by "contradiction". Namely, we assume that the sequence is bounded and by correct mathematical reasoning derive an obviously wrong statement. This will prove that our assumption was wrong and the sequence is, in fact, unbounded.

If (a_n) was bounded above it would have a supremum s. By the definition of the supremum there must be some N such that

$$a_N > \frac{s}{r}$$

since $\frac{s}{r} < s$ is not an upper bound of the sequence. But, now

$$a_{N+1} = ra_N > s,$$

which contradicts to s being the supremum. This proves that (a_n) is unbounded.

4. The unboundedness of the sequences $n!$ and n^n can be shown by the same method as for n^2.

Sequences that tend to infinity are used in Computer Science to describe how fast an algorithm works. Problems that are solved by computer algorithms often depend on some complexity parameter n, e.g., the problem of sorting n objects, or the problem of optimising a path for visiting n places. The time a_n needed to perform such algorithm, of course, increases with increasing parameter n. Algorithms with linear or polynomial time, e.g., $a_n = 3n^2 + 5$ can be expected to work well for large n, whereas algorithms with exponential time, e.g., $a_n = 1.01^n$ may exceed computational resources for large n. This topic will be discussed further in Amth140.

At the end of this lecture we demonstrate that the geometric progression $a_n = cr^n$ with $c > 0$, $r > 1$ grows so much faster than the linear progression $b_n = \alpha n + \beta$ with $\alpha > 0$ that the sequence of ratios $\frac{a_n}{b_n}$ still tends to ∞.

We know that the derived sequence of a_n is $a'_n = \frac{c}{r-1}r^n$. It follows that

$$a_n = a_0 + a'_0 + \cdots + a'_{n-1} \leq c + n\frac{c}{r-1}r^{n-1}.$$

Therefore,

$$\frac{a_n}{b_n} \leq \frac{c + n\frac{c}{r-1}r^{n-1}}{\alpha n + \beta} = \frac{\frac{c}{n}}{\alpha + \frac{\beta}{n}} + \frac{c}{(r-1)(\alpha + \frac{\beta}{n})}r^{n-1}$$

which for large n is approximately

$$\frac{c}{r(r-1)\alpha}r^n$$

and still grows like a geometric progression with $r > 1$.

If the ratio of two sequences $\frac{a_n}{b_n}$ tends to ∞ then the reciprocal ratio $\frac{b_n}{a_n}$ tends to zero. The latter statement can be expressed by the notation

$$b_n = o(a_n),$$

which is pronounced b_n is little-o of a_n. Using this notation we can say that a sequence (b_n) tends to 0 if $b_n = o(1)$. The little-o (and big-O[6]) notations are very convenient but its discussion and use will be postponed to Amth140 and MTHS130.

[6] A sequence (b_n) is big-O of (a_n), written as $b_n = O(a_n)$, if the sequence of ratios $\frac{b_n}{a_n}$ is bounded. E.g., $b_n = n \sin n$ is big-O of $a_n = 2n + 1$. If a sequence (b_n) is bounded we write $b_n = O(1)$.

7 Functions

In this unit we consider functions defined on a subset X of the real numbers $\mathbb{R}$. The domain X usually is an open, closed or semi-closed interval, a union of those, the entire set $\mathbb{R}$ or rays

$$(a,\infty) = \{x \in \mathbb{R} \mid x > a\}, \quad [a,\infty) = \{x \in \mathbb{R} \mid x \geq a\},$$
$$(-\infty,a) = \{x \in \mathbb{R} \mid x < a\}, \quad (-\infty,a] = \{x \in \mathbb{R} \mid x \leq a\}.$$

Mostly we will not pay too much attention to the codomain and assume it to be $\mathbb{R}$.

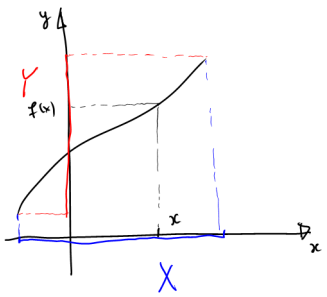
Usually the functions we consider are given by one or several algebraic formulae. An algebraic formula $f(x)$ may make sense for some arguments x and not for others. E.g., the formula

$$f(x) = \frac{1}{x}$$

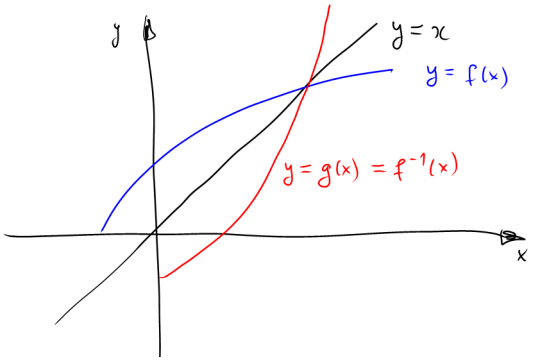
does not make sense for $x = 0$ but for all other real numbers x. We say that $f(x)$ is well defined for $x \neq 0$. Often we will assume that the domain of a function given by a formula $f(x)$ is the largest subset of $\mathbb{R}$ where the formula is well defined. We will call this set the *natural domain* of $f(x)$. E.g., the natural domain of $f(x) = \frac{1}{x}$ is

$$X = (-\infty,0) \cup (0,\infty).$$

The graph of a function $f\colon X \to Y$, where $X, Y \subseteq \mathbb{R}$, can be sketched in an xy-coordinate system.



If f has an inverse function $g\colon Y \to X$ then its graph is the reflection of the graph of f about the bisector $y = x$.



You have already encountered most of the functions below.

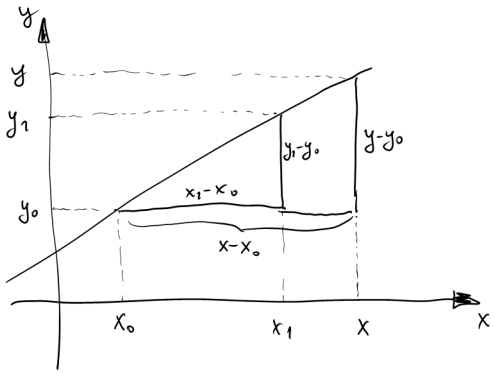
Linear functions. A linear (or affine[7]) function is defined by a formula

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

where m and b are given real parameters. The natural domain X of any linear function is $\mathbb{R}$. The parameter m is called the slope. If $m = 0$ the function is constant. The graph of the function passes through the point $(0, b)$ on the y-axis. Therefore the parameter b is called the y-intercept. The graph passes through the origin $(0, 0)$ if and only if $b = 0$. We show that the graph of a linear function is a straight line. Let (x_0, y_0) and (x_1, y_1) be the coordinates of two points of the graph and let (x, y) be the coordinates of another point of the graph. Then

$$\frac{y - y_0}{x - x_0} = m = \frac{y_1 - y_0}{x_1 - x_0}. \tag{3}$$

This shows that any point (x, y) of the graph lies on the hypothenuse of similar right triangles with vertices (x_0, y_0), (x, y_0), (x, y), i.e. on the line formed by those hypotenuses (see sketch below).



Linear functions are used to model processes where a quantity changes at a constant rate, e.g. a motion with constant velocity. The slope m is a measure of the rate of change (velocity).

[7]In Calculus this kind of functions is called linear functions, but in Linear Algebra a function is called linear only if $b = 0$. The term affine functions is used if $b \neq 0$.

If $m \neq 0$ the function $f(x) = mx + b$ has an inverse, which is also a linear function

$$f^{-1}(x) = \frac{1}{m}x - \frac{b}{m}.$$

Since every real number x has a preimage $f^{-1}(x)$ the range of f consists of all real numbers.

If $m = 0$ the range of $f(x) = b$ consists of the single value b. In this case $f(x)$ has no inverse.

Definition 4. *A function f is called (strictly) increasing if for any two arguments $x_1 < x_2$ from the domain $f(x_1) \leq f(x_2)$ ($f(x_1) < f(x_2)$).*
A function f is called (strictly) decreasing if for any two arguments $x_1 < x_2$ from the domain $f(x_1) \geq f(x_2)$ ($f(x_1) > f(x_2)$).
A function is called monotone if it is either increasing or decreasing.

Proposition 8. *If a function is strictly monotone, i.e. strictly increasing or strictly decreasing, then it is injective.*

Proof. Assume $f(x)$ is strictly increasing (the case of strict decrease is completely analogous). We have to show that for two different inputs $x_1 \neq x_2$ the outputs are also different. Without loss of generality, $x_1 < x_2$. Then $f(x_1) < f(x_2)$, in particular, $f(x_1) \neq f(x_2)$. □

A linear function is strictly increasing (decreasing) if $m > 0$ ($m < 0$). Indeed, assume $m > 0$, then

$$\begin{aligned} x_1 &< x_2 \\ mx_1 &< mx_2 \\ mx_1 + b &< mx_2 + b \\ f(x_1) &< f(x_2). \end{aligned}$$

The case $m < 0$ is analogous.

Power functions. The simplest power functions are given by a formula $f(x) = x^n$, where $n > 1$ is a natural number. For $n = 2$, the function $f(x) = x^2$ is called the square function and, for $n = 3$, the function $f(x) = x^3$ is called the cubic function. The natural domain of the power functions with natural exponent n is the set of all real numbers $\mathbb{R}$. The behaviour of those power functions depends on whether n is odd or even. Therefore we consider these cases separately.

Power functions with even $n = 2m$ have the property $f(-x) = f(x)$, since

$$(-x)^{2m} = ((-x)^2)^m = x^{2m}.$$

Functions with this property are called even functions. More precisely,

Definition 5. *A function $f\colon X \to \mathbb{R}$ is called even, if for any $x \in X$ also $-x \in X$ and $f(-x) = f(x)$ for all $x \in X$.*
A function $f\colon X \to \mathbb{R}$ is called odd, if for any $x \in X$ also $-x \in X$ and $f(-x) = -f(x)$ for all $x \in X$.

The graph of an even function is mirror-symmetric with respect to the y-axis, the graph of an odd function is point-symmetric with respect to the origin.

♠ *Exercises 25.* Show that for any odd function defined on a domain that contains 0, $f(0) = 0$.

Even functions are not injective (unless $X = \{0\}$). In particular, power functions with even exponent are never injective because, e.g. $f(-1) = f(1)$.

Power functions with even exponent $n = 2m$ take only non-negative values since

$$f(x) = x^{2m} = (x^2)^m \geq 0.$$

Therefore the range R is a subset of $[0, \infty)$. In fact, $R = [0, \infty)$. This is a highly non-trivial fact, the proof of which requires more advanced methods of calculus.

♠ *Exercises 26.* Show that the graph of the function $y = x^2$ is a parabola with focus $(0, \frac{1}{4})$ and directrix $y = -\frac{1}{4}$, i.e. show that each point of the graph has the same distance to the focus and to the directrix.

If we restrict the domain of the function $f(x) = x^{2m}$ to $X = [0, \infty)$ the function becomes strictly increasing and hence injective.

Proposition 9. *For any natural number n and for any pair of non-negative numbers $0 \leq x_1 < x_2$ we have $x_1^n < x_2^n$.*

Proof. We use induction on n starting with $n = 1$. The statement is tautological for $n = 1$. Assume

$$x_1^n < x_2^n.$$

Then

$$x_1^n x_1 < x_2^n x_1 < x_2^n x_2,$$

thus

$$x_1^{n+1} < x_2^{n+1},$$

as required. □

If we restrict the codomain also to $Y = [0, \infty)$ the function becomes surjective, thus

$$f\colon [0, \infty) \to [0, \infty)$$

has an inverse. The inverse function of the power function $f(x) = x^{2m}$ is called the $2m$-th root $g(x) = f^{-1}(x) = \sqrt[2m]{x}$ and maps the domain $[0, \infty)$ onto the codomain $[0, \infty)$.

Restricting f to the domain $X^- = (-\infty, 0]$ also renders f injective (strictly decreasing) and gives another inverse function

$$g^-(x) = -g(x) = -\sqrt[2m]{x} \colon [0, \infty) \to (-\infty, 0].$$

NB. The even roots always take non-negative values. The equation $x^{2m} = a$ has two solutions, namely $x^+ = \sqrt[2m]{a}$ and $x^- = -\sqrt[2m]{x}$.

We consider now the power functions $f(x) = x^n$ with odd $n = 2m + 1$. These functions are odd, since

$$(-x)^{2m+1} = ((-x)^2)^m(-x) = -x^{2m}x = -x^{2m+1}.$$

Power functions with odd n are strictly increasing throughout their natural domain $\mathbb{R}$. We have proved that, for $0 \leq x_1 < x_2$ it is the case that $x_1^n < x_2^n$. For $x_1 < 0 \leq x_2$ we clearly have $x_1^n < 0 \leq x_2^n$. If both x_1 and x_2 are negative, we have

$$\begin{aligned} x_1 &< x_2 \\ -x_1 &> -x_2 \\ (-x_1)^n &> (-x_2)^n \\ -x_1^n &> -x_2^n \\ x_1^n &< x_2^n. \end{aligned}$$

Again we state the non-trivial fact, that the range of an odd power function equals $\mathbb{R}$ without proof. It follows that the odd power functions

$$f(x) = x^{2m+1} \colon \mathbb{R} \to \mathbb{R}$$

are invertible. The inverse functions are the $2m + 1$-st root functions

$$g(x) = \sqrt[2m+1]{x} \colon \mathbb{R} \to \mathbb{R}.$$

They are also odd functions.

It is common to denote the inverse functions of the power functions $f(x) = x^n$ by $g(x) = x^{\frac{1}{n}}$. This is consistent with the usual power rules

$$g(f(x)) = (x^n)^{\frac{1}{n}} = x^1 = x.$$

This allows us to define power functions with (positive) rational exponent by

$$f(x) = x^{\frac{p}{q}} = (\sqrt[q]{x})^p.$$

The natural domain of those functions is $[0, \infty)$. They have the property $f(0) = 0$ and $f(1) = 1$.

Power functions with negative rational exponents are defined by the power rule $x^{-1} = \frac{1}{x}$. We need to exclude $x = 0$ from their natural domain. We adopt the notation $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$. For negative integer exponents the functions

$$f(x) = x^{-n} = \frac{1}{x^n} \colon \mathbb{R}^* \to \mathbb{R}^*$$

are even for even n and odd for odd n.

♠ *Exercises 27.* Show that the graph of the function $f(x) = \frac{1}{x}$ is a hyperbola with foci $F_1(\sqrt{2}, \sqrt{2})$ and $F_2(-\sqrt{2}, -\sqrt{2})$, i.e. show that the difference of the distances $|PF_1| - |PF_2|$ from each point P of the graph to the foci is constant $(2\sqrt{2})$.

The function $f(x) = \frac{1}{x} \colon \mathbb{R}^* \to \mathbb{R}^*$ is bijective and inverse to itself. For negative odd exponents the function $f(x) = x^{-2m+1} \colon \mathbb{R}^* \to \mathbb{R}^*$ is bijective and the inverse is

$$g(x) = x^{-\frac{1}{2m+1}} = \frac{1}{\sqrt[2m+1]{x}} \colon \mathbb{R}^* \to \mathbb{R}^*.$$

For negative even exponents the function $f(x) = x^{-2m} \colon \mathbb{R}^* \to \mathbb{R}^*$ is not injective and takes only positive values. It becomes bijective after restricting domain and codomain to $(0, \infty)$. The inverse function is

$$g(x) = x^{-\frac{1}{2m}} = \frac{1}{\sqrt[2m]{x}} \colon (0, \infty) \to (0, \infty).$$

Polynomial functions. A polynomial function is given by a formula

$$f(x) = \sum_{k=0}^{n} a_k x^k = a_n x^n + \cdots + a_0$$

where $a_0, \ldots, a_n$ are real parameters, called the coefficients of the polynomial. We assume that $a_n \neq 0$. In this case we call a_n the leading coefficient and n the order of the polynomial. A polynomial of the form $a_n x^n$ is called a monomial. The natural domain of a polynomial function is the set of all real numbers $\mathbb{R}$. In general, polynomials are neither injective nor surjective.

A number x_0 from the domain is called a zero of the function $f(x)$ if $f(x_0) = 0$. Zeros of polynomial functions are also called the roots of the polynomial. The zeros

of a function are the points of intersection of the graph with the x-axis. If x_0 is a root of a polynomial of $f(x)$, the polynomial factors into

$$f(x) = (x - x_0)g(x)$$

where $g(x)$ is a polynomial of order less by 1 than the order of f[8]. This shows that a polynomial of order n has at most n roots. It can have less than n or even no roots.

♠ *Exercises 28.* Give an example of a polynomial that has no roots.

♦Example. The polynomial $f(x) = x^3 - 2x^2 + x - 2$ has a root $x_0 = 2$. Factorising $f(x) = (x - 2)g(x)$ can be done by long division.

$$\begin{array}{rl|llllll} & x^2 & + & 1 & & & & \\ \hline x-2\ ) & x^3 & - & 2x^2 & + & x & - & 2 \\ & x^3 & - & 2x^2 & & & & \\ \hline & & & 0 & + & x & - & 2 \\ & & & & & x & - & 2 \\ \hline & & & & & & & 0 \end{array}$$

Rational functions. Rational functions are ratios of polynomials

$$f(x) = \frac{p(x)}{q(x)}$$

where $p(x)$ and $q(x)$ are polynomials and $q(x)$ is not the zero polynomial. The natural domain of rational functions is the set of real numbers, excluding the roots of the polynomial in the denominator. Polynomial functions are particular cases of rational functions with $q(x) = 1$. The functions $f(x) = x^{-n}$ are also particular cases of rational functions.

[8]The proof of this fact is based on long division of polynomials.

8 Transcendental functions

Transcendental functions form an important class of non-rational, non-algebraic functions. Important examples of transcendental functions covered in this unit are exponential functions, logarithmic functions and the trigonometric functions and their inverses. We will study these functions in some detail later in this unit. At this stage it is sufficient to know how the exponential functions are defined for rational arguments. For any positive real number $a \neq 1$ (called the base) and any rational exponent $x = \frac{p}{q}$ we can define

$$a^x : x \mapsto \sqrt[q]{a}^p.$$

In fact, the exponential functions can be defined throughout $\mathbb{R}$ and will be strictly increasing if $a > 1$ and strictly decreasing if $a < 1$. The range is the set of positive numbers. The proofs of these facts require more advanced knowledge in calculus and will be given in MTHS130.

You need to be familiar with the basic rules

$$a^{x+y} = a^x a^y, \quad (a^x)^y = a^{xy},$$

for rational x, y. These identities remain true for real x, y.

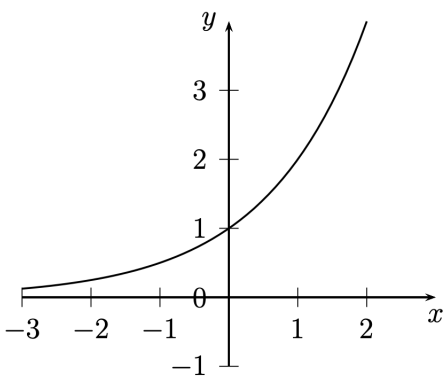
If we take for granted that the exponential functions are strictly increasing and onto $\mathbb{R}^+$ we can define the inverse functions

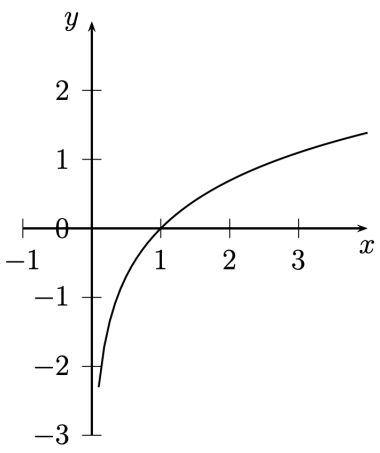
$$\log_a : \mathbb{R}^+ \to \mathbb{R}.$$

Thus $\log_a x$ is the number y such that $a^y = x$. From the rules for the exponential functions we get

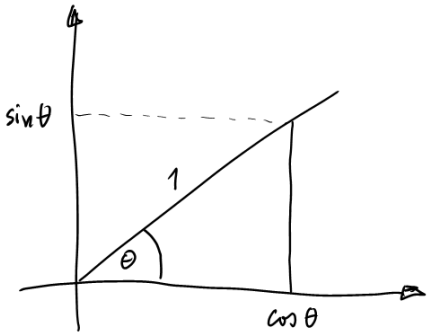
$$\log_a xy = \log_a x + \log_a y, \quad \log_a x^c = c \log_a x.$$

Below are the graphs of $y = 2^x$ and $y = \log_{\mathrm{e}} x$ (where $\mathrm{e} \approx 2.7$ the Euler number)





The trigonometric functions $\sin\theta$ and $\cos\theta$ are defined as follows. Consider the half line that forms an angle of θ (measured in radians) with the x-axis. Then $\sin\theta$ is the y coordinate and $\cos\theta$ the x coordinate of the intersection point of this half line with the unit circle. Hence $\sin\theta$ and $\cos\theta$ are the lengths of the opposite and adjacent sides of a right triangle with hypothenuse of length 1.



It readily follows from Pythagoras's theorem that

$$\sin^2\theta + \cos^2\theta = 1$$

for any angle θ. We will use this identity many times. Notice that

$$\sin\theta = \cos\left(\frac{\pi}{2} - \theta\right)$$

since sin and cos interchange as adjacent and opposite sides interchange.

It is a characteristic feature for trigonometric functions that they are periodic:

$$\sin(\theta + 2\pi) = \sin\theta, \quad \cos(\theta + 2\pi) = \cos\theta.$$

This follows from the fact that adding 2π (or integer multiples of 2π) to an angle gives the same half-line.

We will need the so-called addition theorem for sin and cos:

$$\sin(\theta + \phi) = \sin\theta\cos\phi + \cos\theta\sin\phi,$$
$$\cos(\theta + \phi) = \cos\theta\cos\phi - \sin\theta\sin\phi.$$

The functions sin and cos are defined for any real number and take values in the interval $[-1, 1]$. The periodicity precludes them from being injective. However when we restrict sin to the domain $[-\frac{\pi}{2}, \frac{\pi}{2}]$ it becomes strictly increasing with range $[-1, 1]$. This is plausible from the geometric definition of sin and will be formally proven by means of calculus later in this unit.

This so restricted function has an inverse

$$\arcsin\colon [-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}].$$

Similarly, cos is strictly decreasing if restricted to $[0, \pi]$ with range $[-1, 1]$. This restricted function also has an inverse

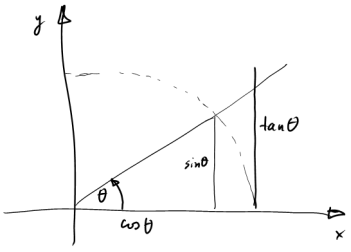
$$\arccos\colon [-1, 1] \to [0, \pi].$$

The function sin is odd, whereas cos is even.

Another important trigonometric function is tan which is defined as

$$\tan\theta = \frac{\sin\theta}{\cos\theta}.$$

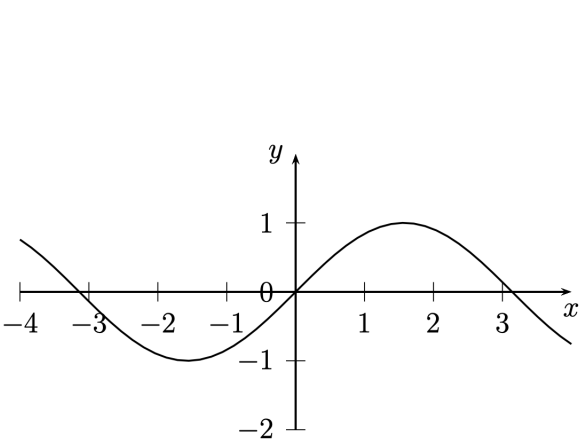
tan is defined for all real numbers except numbers of the form $\theta = \frac{\pi}{2} + k\pi$, where k in an integer. These are the zeros of cos. The range of tan is the set of all real numbers. This follows from the interpretation of $\tan\theta$ as the length (with positive or negative sign) of the segment of the tangent to the unit circle at $(0, 1)$ (whence the name tangent) between the point $(0, 1)$ and the intersection with the half-line determined by the θ.

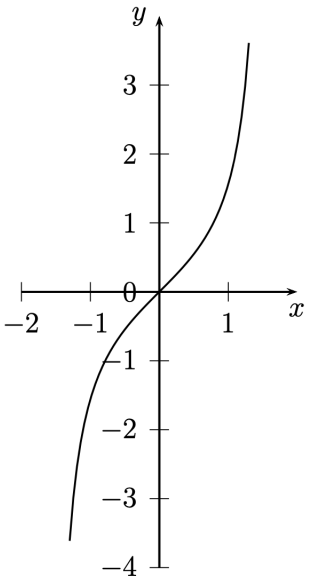


tan is an odd function. If restricted to $(-\frac{\pi}{2}, \frac{\pi}{2})$ it is strictly increasing with range $\mathbb{R}$. This function has an inverse

$$\arctan\colon \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2}).$$

Below are the graphs of $y = \sin x$ and $y = \tan x$.





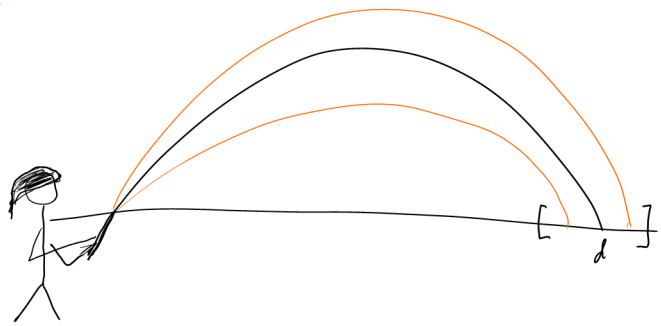
9 Continuity of functions

For the topic of continuity you may have in mind the following example:

♦Example. A firefighter shoots a water jet keeping the hose at an angle α. Assume that the water jet has an initial velocity v, that it moves along a parabolic trajectory and hits the ground at the distance d. The following formula models the relation between the angle α and the distance d

$$d = \frac{v^2}{g} \sin 2\alpha,$$

where g is the gravitation constant. Assume that the initial velocity is $20\frac{m}{sec}$ and $g = 10\frac{m}{sec^2}$. For $\alpha = 15°$ the distance $d = 20m$. The firefighter wants to extinguish a fire that extends from $15m$ to $25m$. In what range should he vary the angle?



In practice we often need to compute a function for an argument that might be subject to some error. Such errors may occur in measurements. Also when we use a calculator we can only enter numbers with a relatively small number of digits. We rely then on the assumption that the value of the function at a "nearby" argument is "close enough" to the result we want. The concept of continuity will help us to understand when such assumption is justified.

Assume we want to compute a function f for an argument x_0 but are only able to compute it for a nearby number x.

Our basic question is:

Can we control the error in our computation by choosing x close enough to x_0?

First of all we need to define the notion of "close enough". We measure how close a number a to number b is by their distance $|a - b|$. Saying that the distance

between a and b is smaller than some small but positive number ε means that

$$|a - b| < \varepsilon.$$

This inequality can be rewritten without the absolute value as two inequalities

$$-\varepsilon < a - b, \qquad \text{and } a - b < \varepsilon.$$

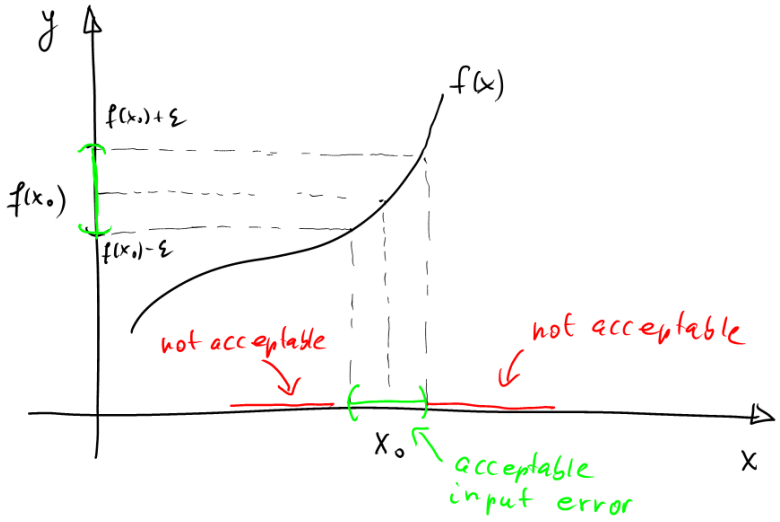
Thus we want to make sure that

$$|f(x) - f(x_0)| < \varepsilon$$

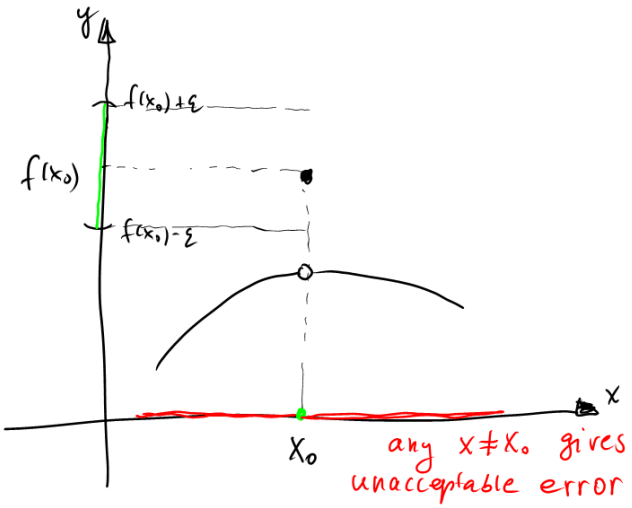
where ε is a small, but positive (acceptable) error. Now we can reformulate our basic question:

Can we achieve a desired precision $|f(x) - f(x_0)| < \varepsilon$ of the function f by making $|x - x_0|$ smaller than some positive number δ (which, of course, depends on ε)?

In the sketches below we illustrate this concept for a function where such control is possible:



and for a function where it is not possible:



Let us look at some examples:

♦Example.

1. $f(x) = mx + b$ (with $m \neq 0$). We want to make

$$|f(x) - f(x_0)| < \varepsilon.$$

Let us analyse this inequality. It is equivalent to

$$|mx + b - (mx_0 + b)| = |m(x - x_0)| = |m||x - x_0| < \varepsilon.$$

This inequality is satisfied as soon as

$$|x - x_0| < \frac{\varepsilon}{|m|}.$$

This is exactly the kind of condition on x we wanted. We just need to take x such that

$$|x - x_0| < \delta = \frac{\varepsilon}{|m|}$$

to guarantee a precision of ε in the output. Since ε can be arbitrarily small, we can achieve any desired precision by an appropriate choice of δ. In this example we could express the optimal δ as a simple function of ε.

2. $f(x) = b$, i.e., f is a constant function. Then

$$|f(x) - f(x_0)| = |b - b| = 0 < \varepsilon,$$

no matter what $x_0, \varepsilon, \delta, x$ we choose.

3. $f(x) = x^2$. We want
$$|f(x) - f(x_0)| < \varepsilon$$
This is equivalent to
$$|x^2 - x_0^2| = |x - x_0||x + x_0| < \varepsilon.$$
To satisfy this inequality we choose δ to make the factor $|x - x_0|$ small and take care that the other factor $|x + x_0|$ does not become too big. Thus the choice of δ is subject to several conditions. First we stipulate $\delta < 1$. Then
$$|x + x_0| = |x - x_0 + 2x_0| \leq |x - x_0| + 2|x_0| \leq 2|x_0| + 1.$$
Then by choosing $\delta < \dfrac{\varepsilon}{2|x_0| + 1}$ we make sure that
$$|x^2 - x_0^2| = |x - x_0||x + x_0| \leq \frac{\varepsilon}{2|x_0| + 1}(2|x_0| + 1) < \varepsilon$$
as required. This choice of δ is not optimal.

4. $f(x) = |x|$. This function coincides with the linear function $f(x) = x$ for $x > 0$ and with $f(x) = -x$ for $x < 0$. Therefore the only point that requires attention is $x_0 = 0$. But then
$$||x| - |x_0|| = ||x| - |0|| = |x| < \varepsilon$$
is a consequence of $|x - 0| = |x| < \delta$ with $\delta = \varepsilon$.

The examples motivate the following definition.

A function $f : X \to \mathbb{R}$ is continuous at some point $x_0 \in X$ if for any positive number ε there exists a positive number δ (that depends on x_0 and ε) such that for all $x \in X$ the condition $|x - x_0| < \delta$ implies $|f(x) - f(x_0)| < \varepsilon$.

A formal short way to write this statement uses the quantifiers $\forall$ and $\exists$.

$\forall \varepsilon > 0 \quad \exists \delta > 0$ such that $\forall x \in X$ with $|x - x_0| < \delta$ we have $|f(x) - f(x_0)| < \varepsilon$

or even shorter,
$$\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall \quad x \in X \colon |x - x_0| < \delta, \quad |f(x) - f(x_0)| < \varepsilon.$$

This is a rather complex definition. The following objects are involved: the function f which is tested for continuity at a point x_0 of the domain, the desired

output precision ε and the necessary input precision δ (which depends on ε). Proving continuity means to find δ with the required properties for any given $\varepsilon > 0$. δ depends on ε. The choice of δ is ambiguous: we can always replace δ by a smaller positive number and the statement will still be true.

Now we investigate the questions: What does it mean that a function is discontinuous (=not continuous) at x_0? Are there such functions?

First of all, the notion of continuity (or discontinuity) only makes sense for points x_0 from the domain of the function. If $x_0 \in X$ the negation of the statement that defines continuity is:

There is some $\varepsilon > 0$ such that for any $\delta > 0$ the condition $|x - x_0| < \delta$ does not imply $|f(x) - f(x_0)| < \varepsilon$, i.e. there exists some x such that $|f(x) - f(x_0)| \geq \varepsilon$ although $|x - x_0| < \delta$.

In other words, there is some output precision that cannot be achieved, no matter how precise the input is. This negation can be formally derived by swapping $\forall$ and $\exists$ and negating the final statement:

$$\exists \varepsilon > 0 \text{ such that } \forall \delta > 0 \quad \exists x \in X \text{ with } |x - x_0| < \delta \text{ such that } |f(x) - f(x_0)| \geq \varepsilon.$$

The following function is not continuous at $x_0 = 0$

$$f(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ 0 & \text{for } x < 0. \end{cases}$$

This function is important in physics and technology. It models jumps from one state into another. This function is defined for $x_0 = 0$. We can approach $x_0 = 0$ from the left as close as we want, the value $f(x) = 0$ will stay far from $f(0) = 1$, i.e., we can't achieve an output precision ε that is smaller than 1. Formally: $\exists \varepsilon > 0$ (namely $\varepsilon = \frac{1}{2}$) such that $\forall \delta > 0$ $\exists x \in \mathbb{R}$ with $|x - x_0| < \delta$ (namely $x = -\frac{\delta}{2}$) such that $|f(x) - f(x_0)| = 1 \geq \varepsilon = \frac{1}{2}$.

Disproving continuity means to find some particular ε and, no matter how small we choose δ, to find an argument x (depending on δ) with $|x - x_0| < \delta$ and $|f(x) - f(x_0)| \geq \varepsilon$.

Roughly speaking, a function is continuous if we can draw its graph in one go, without gaps.

♠ *Exercises 29.*

1. Prove that $y = f(x) = \sqrt{x}$ is continuous at $x_0 = 0$.

2. Show that the function

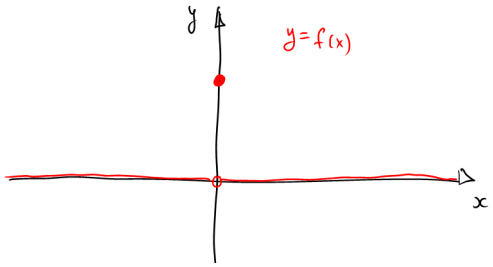
$$y = f(x) = \begin{cases} \sin \frac{1}{x} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0 \end{cases}$$

is not continuous at 0. Hint. Use that $\sin \frac{1}{x} = 1$ for $x = \dfrac{1}{\frac{\pi}{2} + 2k\pi}$.

10 Limits of functions

The following function is discontinuous at $x_0 = 0$:

$$f(x) = \begin{cases} 1 & \text{for } x = 0 \\ 0 & \text{for } x \neq 0. \end{cases}$$



We can find some $\varepsilon > 0$ (namely $\varepsilon = \frac{1}{2}$) such that for any $\delta > 0$ there exist x with $|x < \delta|$ and $|f(x) - f(0)| > \frac{1}{2}$ (namely $x = \frac{\delta}{2}$).

This discontinuity seems to result from someone having put the "wrong" value at $x_0 = 0$. We can make f continuous by redefining the function at the single point $x_0 = 0$ to $f(0) = 0$. We will now investigate the following problem:

When can a function be made continuous at x_0 by just redefining $f(x_0)$? What is the "right" value $f(x_0)$?

Assume some number a is this right value. Then

$\forall \varepsilon > 0\, \exists \delta > 0$ such that $\forall x \in X$ with $|x - x_0| < \delta$ implies $|f(x) - f(x_0)| < \varepsilon$

must hold with a instead of $f(x_0)$. But if $f(x_0)$ was the "wrong" value then $|f(x_0) - a|$ is a fixed positive number even if $|x - x_0| = 0$. Hence we cannot require that $|f(x) - a| < \varepsilon$ holds for $x = x_0$. This gives us the definition of a as the number such that

$\forall \varepsilon > 0 \quad \exists \delta > 0$ such that $\forall x \in X$ with $0 \neq |x - x_0| < \delta$ it is true that $|f(x) - a| < \varepsilon$.

This is the definition of the *limit* of f as x approaches x_0. This is expressed by the following notation

$$\lim_{x \to x_0} f(x) = a.$$

The notion of limit allows us reformulate the definition of continuity:

A function $f : X \to \mathbb{R}$ is continuous at $x_0 \in X$ if and only if $\lim_{x \to x_0} f(x) = f(x_0)$.

This gives us at once a powerful tool for computing limits: If f is known to be continuous at x_0 the limit $\lim_{x\to x_0} f(x)$ can be obtained by evaluating $f(x_0)$. This motivates the strategy to study continuous functions.

♦Example.

1.
$$\lim_{x\to x_0} mx + b = mx_0 + b$$

2.
$$\lim_{x\to x_0} x^2 = x_0^2.$$

Finding limits "from first principles", i.e., just using the formal definition, can be tedious. We will derive rules that allow us to compute new limits from already known limits. These rules are analogous to the rules for limits of sequences and the proofs are also similar. We have

Theorem 5. *Let f and g be two functions and*

$$\lim_{x\to x_0} f = a \text{ and } \lim_{x\to x_0} g = b$$

then

(a) $\lim_{x\to x_0} f + g = a + b$

(b) $\lim_{x\to x_0} f \cdot g = a \cdot b$

(c) $\lim_{x\to x_0} \frac{f}{g} = \frac{a}{b}$ *if* $b \neq 0$.

Proof.

(a) This just says the "limit of a sum is the sum of the limits".

Let $\varepsilon > 0$ be given. Now both f and g have well defined limits as $x \to x_0$ so we know there exists a number δ such that

$$|f(x) - a| < \frac{\varepsilon}{2} \text{ and } |g(x) - b| < \frac{\varepsilon}{2}$$

whenever $0 < |x - x_0| < \delta$. Then

$$\begin{aligned} |[f(x) + g(x)] - (a + b)| &= |[f(x) - a] + [g(x) - b]| \\ &\leq |[f(x) - a]| + |[g(x) - b]|, \end{aligned}$$

by the triangle inequality. Combining this with our previous inequalities we have

$$\begin{aligned} |[f(x)+g(x)]-(a+b)| &\leq |f(x)-a|+|g(x)-b| \\ &< \frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon, \end{aligned}$$

whenever $0<|x-x_0|<\delta$.

This completes the proof of (a).

(b) The proof of part (b) is a bit more intricate. Firstly we observe that

$$f(x)g(x)-ab=(f(x)-a)g(x)+(g(x)-b)a$$

so by the triangle inequality.

$$(*) \quad |f(x)g(x)-ab| \leq |f(x)-a|\,|g(x)|+|g(x)-b|\,|a|\,.$$

We need to make the right hand side (and hence the left hand side) of the above inequality smaller than any given positive ε.

Now f and g both have well-defined limits at $x=x_0$, so for any $1\geq\bar{\varepsilon}>0$ we can find $\delta>0$ such that

$$|f(x)-a|<\bar{\varepsilon} \text{ and } |g(x)-b|<\bar{\varepsilon},$$

whenever $0<|x-x_0|<\delta$.

In particular,

$$|g(x)|\leq|b|+\bar{\varepsilon}\leq|b|+1.$$

Choose $\bar{\varepsilon}=\min\left\{1,\dfrac{\varepsilon}{2(|a|+1)},\dfrac{\varepsilon}{2(|b|+1)}\right\}$.

Then

$$\begin{aligned} |f(x)g(x)-ab| &\leq |f(x)-a|\,|g(x)|+|g(x)-b|\,|a| \\ &\leq \bar{\varepsilon}(|b|+1)+\bar{\varepsilon}|a|\leq\varepsilon \end{aligned}$$

whenever $0<|x-x_0|<\delta$. This proves (b).

(c) The proof is similar to that of (b) and is left as an exercise.

□

It follows immediately that for any rational function $f(x)=\dfrac{p(x)}{q(x)}$, where $p(x)$ and $q(x)$ are any polynomials

$$\lim_{x\to x_0} f(x)=f(x_0)=\frac{p(x_0)}{q(x_0)},$$

if x_0 is not a root of $q(x)$.

Notice, that a limit A of a function f as $x \to x_0$ might not exist. In this case the function cannot be made continuous by suitable definition of $f(x_0)$.

♠ *Exercises 30.* Prove statement c) from the theorem above.

The following Corollary is an immediate consequence of the theorem above.

Corollary 1. *If two functions f and g are defined on the same domain X and they are both continuous at $x_0 \in X$ then the functions*

(a) $f + g$ and $f - g$

(b) fg are continuous at x_0.

(c) The function $\dfrac{f}{g}$ is continuous at x_0 if $g(x_0) \neq 0$.

Proof. We prove only part (b). Part (a) and (c) are analogous. The function fg is continuous at x_0 if

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

We have

$$\lim_{x \to x_0} f(x)g(x) = \lim_{x \to x_0} f(x) \lim_{x \to x_0} g(x) = f(x_0)g(x_0),$$

as required. □

One-sided limits. The existence of the limit requires that $f(x)$ approaches the same value when x approaches x_0 from either side. The notion of one-sided limits allows us to investigate the behaviour of a function when x approaches x_0 either from the left or right hand side, i.e. either staying smaller or bigger than x_0. Here are the precise definitions:

A function $f : X \to \mathbb{R}$ has a left-sided limit

$$\lim_{x \to x_0^-} f(x) = A$$

if $\forall \varepsilon > 0$ $\exists \delta > 0$ such that $0 < x_0 - x < \delta$ implies $|f(x) - A| < \varepsilon$

The additional condition $0 < x_0 - x$ relaxes the statement by ignoring all x that are greater than or equal to x_0. Notice the superscript $-$ at x_0 that indicates the left-sided limit. Analogously,

A function $f : X \to \mathbb{R}$ has a right-sided limit

$$\lim_{x \to x_0^+} f(x) = A$$

if $\forall \varepsilon > 0$ $\exists \delta > 0$ *such that* $0 < x - x_0 < \delta$ *implies* $|f(x) - A| < \varepsilon$

The only change to left-sided limits is that $0 < x_0 - x$, i.e. $x < x_0$ became $0 < x - x_0$, i.e. $x > x_0$.

The following theorem relates limits to one-sided limits.

Theorem 6. *The limit* $\lim_{x \to x_0} f(x) = A$ *exists if and only if both one-sided limits also exist and equal* A*. Consequently, a function* $f : X \to \mathbb{R}$ *is continuous at* $x_0 \in X$ *if and only if*

$$\lim_{x \to x_0^-} f(x) = \lim_{x \to x_0^+} f(x) = f(x_0).$$

Proof. If the limit $\lim_{x \to x_0} f(x) = A$ then

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } -\delta < x - x_0 < \delta, x \neq x_0 \text{ implies } |f(x) - A| < \varepsilon$$

then the weaker statements (for a smaller set of x)

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } 0 < x - x_0 < \delta \text{ implies } |f(x) - A| < \varepsilon$$

and

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } -\delta < x - x_0 < 0 \text{ implies } |f(x) - A| < \varepsilon$$

are also true, i.e. both one-sided limits exist and equal A.

Vice versa, the existence of both one-sided limits and their equality to A means

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } 0 < x - x_0 < \delta_1 \text{ implies } |f(x) - A| < \varepsilon$$

and

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } -\delta_2 < x - x_0 < 0 \text{ implies } |f(x) - A| < \varepsilon.$$

Notice that δ_1 and δ_2 in the statement can be different for a given ε. However, their minimum

$$\delta = \min\{\delta_1, \delta_2\}$$

satisfies both statements, which combine into

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } -\delta < x - x_0 < \delta, x \neq x_0 \text{ implies } |f(x) - A| < \varepsilon,$$

as required. □

The arithmetic rules for one-sided limits are the same as for limits. The proofs are analogous.

♠ *Exercises 31.* Prove that for an even function $f(x)$

$$\lim_{x\to 0^+} f(x) = \lim_{x\to 0^-} f(x)$$

and for an odd function $g(x)$

$$\lim_{x\to 0^+} g(x) = -\lim_{x\to 0^-} g(x).$$

The equality statement implies that one one-sided limit exists if and only the other exists.

We show here that our definition of convergence is equivalent to the following statement:

$$\lim_{x\to a} f(x) = A$$

if and only if for ANY sequence $\{x_n\}$ such that $x_n \neq a$ and $\lim_{n\to\infty} x_n = a$, $\lim_{n\to\infty} f(x_n) = A$. First we show that

$$\lim_{n\to\infty} x_n = a \text{ and } \lim_{x\to a} f(x) = A$$

implies

$$\lim_{n\to\infty} f(x_n) = A.$$

Indeed, assume that $\lim_{n\to\infty} x_n = a$. Then

$$\forall \delta > 0 \quad \exists N \in \mathbb{N} \text{ such that } \forall n > N, |x_n - a| < \delta.$$

On the other hand

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } \forall\, x \text{ such that } 0 \neq |x - a| < \delta, |f(x_n) - A| < \varepsilon.$$

Since $x_n \neq a$ and for $n > N$, $|x_n - a| < \delta$ it follows $|f(x_n) - A| < \varepsilon$, as required.

Now we show that if A is not the limit of $f(x)$ as x approaches a then there exists a sequence $\{x_n\}$ such that $f(x_n)$ does not approach A as $n \to \infty$. Indeed,

$$\exists \varepsilon > 0 \quad \forall \delta > 0 \quad \exists 0 \neq |x - a| < \delta \text{ such that } |f(x) - A| \geq \varepsilon$$

Now, fix some ε which exists according to the statement above. Then for any $\delta = \frac{1}{n}$ there exists x_n such that

$$0 \neq |x_n - a| < \delta \text{ and } |f(x_n) - A| \geq \varepsilon.$$

It follows that

$$\lim_{n\to\infty} x_n = a. \text{ Why?}$$

but

$$\lim_{n\to\infty} f(x_n) \neq A. \text{ Why?}$$

Limits as $x \to \pm\infty$. A function can model measurements related to a long term process. Again we may be interested in understanding their long-term behaviour. Similar to a sequence, the measured quantities of that process may approach some limit.

Similar, to our previous notion of a limit, we would call a the limit of a function $f(x)$ as x tends to ∞ if we can make $f(x)$ as close to a as we wish by choosing x big enough. As before we express closeness of $f(x)$ to a by saying $|f(x) - a| < \varepsilon$. Largeness of x can be expressed by saying that x is greater than some (big number) L.

Formally:

We say $f(x)$ *has the limit* a *as* $x \to \infty$, *or* $\lim_{x\to\infty} f(x) = a$ *as* $x \to \infty$, *if* $\forall \varepsilon > 0$ $\exists L > 0$ *such that* $\lvert f(x) - a\rvert < \varepsilon$ *whenever* $x > L$.

All this says is that if you give me any $\varepsilon > 0$ I can find a number $L > 0$ such that $|f(x) - a|$ is smaller than ε whenever $x > L$.

Similarly:

We say $f(x)$ *has the limit* a *as* $x \to -\infty$, *or* $\lim_{x\to-\infty} f(x) = a$ *as* $x \to -\infty$, *if* $\forall \varepsilon > 0$ $\exists L > 0$ *such that* $\lvert f(x) - a\rvert < \varepsilon$ *whenever* $x < -L$.

If a function has a limit a as x tends to $-\infty$ or to $+\infty$ we say that the horizontal line $y = a$ is a horizontal asymptote. Horizontal asymptotes are useful in sketching the graph of the function because for large (positive or negative) x the graph of f is very close to the asymptote.

♦Example. Prove that $\dfrac{1}{\sqrt{x}} \to 0$ as $x \to \infty$.

Solution. Let $\varepsilon > 0$ be given. We have to find an $L > 0$ such that $\left|\dfrac{1}{\sqrt{x}}\right| < \varepsilon$ whenever $x > L$. This is pretty easy in this case. The desired inequality

$$\left|\frac{1}{\sqrt{x}}\right| < \varepsilon$$

is equivalent to

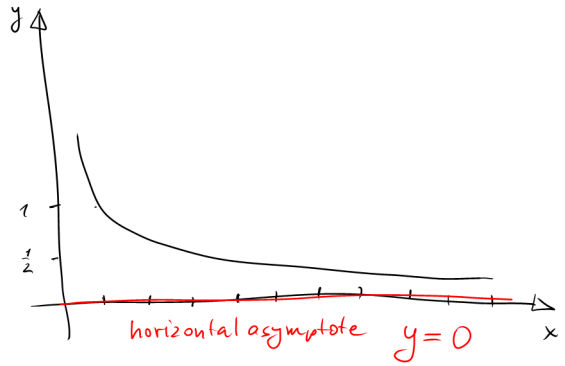
$$x > \frac{1}{\varepsilon^2}$$

Hence it suffices to take $L = \frac{1}{\varepsilon^2}$. Then $x > L$ guarantees

$$\left| \frac{1}{\sqrt{x}} \right| < \varepsilon.$$

□

See a rough sketch of the graph of $f = \dfrac{1}{\sqrt{x}}$ below



♦Example. Prove that $\dfrac{\sqrt{x+1} - \sqrt{x-1}}{x} \to 0$ as $x \to \infty$.

Solution. The numerator is the difference of two square roots, each of which grows as $x \to \infty$. The behaviour of the difference is not obvious. The trick here is to make the troublesome numerator rational (i.e. to get rid of the square roots),

$$\begin{aligned} \left(\sqrt{x+1} - \sqrt{x-1}\right)\left(\sqrt{x+1} + \sqrt{x-1}\right) &= (x+1) - (x-1), \text{ “difference of squares”.} \\ &= 2 \end{aligned}$$

Therefore

$$\begin{aligned} \frac{\sqrt{x+1} - \sqrt{x-1}}{x} &= \frac{\sqrt{x+1} - \sqrt{x-1}}{x} \cdot \frac{\sqrt{x+1} + \sqrt{x-1}}{\sqrt{x+1} + \sqrt{x-1}} \\ &= \frac{2}{x\left(\sqrt{x+1} + \sqrt{x-1}\right)}. \end{aligned}$$

It's now clear that this expression tends to zero as $x \longrightarrow \infty$. We need to prove it formally.

We note that for $x > 1$,

$$\sqrt{x+1} + \sqrt{x-1} > 1,$$

so

$$\frac{1}{\sqrt{x+1} + \sqrt{x-1}} < 1.$$

Thus we have

$$\frac{\sqrt{x+1}-\sqrt{x-1}}{x} = \frac{2}{x\left(\sqrt{x+1}+\sqrt{x-1}\right)} < \frac{2}{x}.$$

So for $x > L > \max(1, \frac{2}{\varepsilon})$, we have

$$\left|\frac{\sqrt{x+1}-\sqrt{x-1}}{x}\right| < \frac{2}{x} < \frac{2}{L} = \varepsilon,$$

We are done. □

♠ *Exercises 32.* Prove that $\lim\limits_{x\to\infty} \dfrac{1}{1+x^2} = 0.$

Poles. Consider a function $f(x)$ that is not defined for some number x_0 but on some interval $(x_0, b]$, $[a, x_0)$ or some punctured interval $[a, b] \setminus \{x_0\}$. If both one-sided limits exist and they are equal we can stipulate $f(x_0) = \lim_{x\to x_0} f(x)$ so that f becomes continuous.

♦Example. Let $f(x) = \frac{x^2-1}{x+1}$. The natural domain of this function does not include $x_0 = -1$ because this would require division by zero. However, for $x \neq -1$ the formula can be replaced by $\frac{x^2-1}{x+1} = \frac{(x-1)(x+1)}{x+1} = x - 1$, which gives a function that is defined and continuous for any $x \in \mathbb{R}$. Therefore,

$$\lim_{x\to -1} f(x) = -2.$$

By stipulating $f(-1) = -2$ we make $f(x)$ continuous at $x_0 = -1$.

Other possible scenarios include the option that the function grows or decays unboundedly as x approaches x_0. Let's look at an example.

♦Example. Consider $f(x) = \frac{1}{x}$. This function is not defined for $x_0 = 0$. For small positive x the function takes large positive values and for small negative x it takes large negative values. We say

$$\lim_{x\to 0^+} f(x) = \infty.$$

The precise formal meaning of this statement is

$$\forall M > 0 \quad \exists \delta > 0 \text{ such that } \forall x\colon 0 < x < \delta, f(x) > M.$$

If we approach $x_0 = 0$ from the left we get

$$\lim_{x\to 0^-} f(x) = -\infty$$

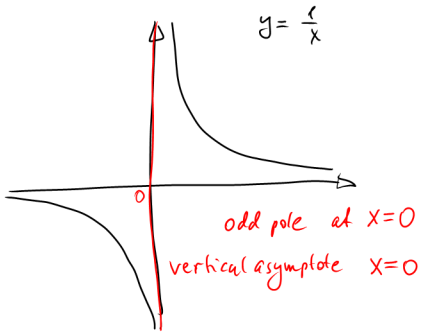
which can be formally expressed as

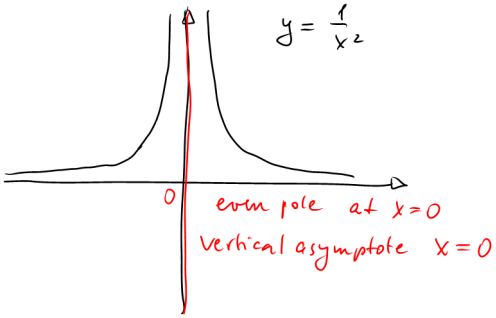
$$\forall M > 0 \quad \exists \delta > 0 \text{ such that } \forall -\delta < x < 0, f(x) < -M.$$

We say that the function $f(x) = \frac{1}{x}$ has an odd pole at $x_0 = 0$ since both one-sided limits are infinity with different sign.

For the function $g(x) = \frac{1}{x^2}$ both one-sided limits as $x \to 0$ are $+\infty$. In this case we have an even pole.

If a function has the limit $\pm\infty$ as $x \to x_0$ the vertical line $x = x_0$ is called a vertical asymptote. Vertical asymptotes help in sketching the graph of a function. See sketches below.





11 Continuity of elementary functions

Elementary functions are polynomial functions, trigonometric functions, exponential functions (which still need to be defined), their inverses, sums, products, ratios and compositions. The aim of this lecture is to show that elementary functions are continuous throughout their natural domains. Before we state and prove the corresponding theorem we will establish some more properties of continuous functions.

If $f\colon X \to Y$ is a function and $g\colon Y' \to Z$ is another function, such that the range of f is contained in the domain Y' of g. Then we can define the *composition* $g \circ f\colon X \to Z$ by

$$g \circ f = g(f(x)),$$

that is by applying first f on x and then applying g on the output $f(x)$.

♦Example. Let $f\colon \mathbb{R} \to \mathbb{R}$ be given by $f(x) = 2x + 1$ and $g\colon \mathbb{R} \to \mathbb{R}$ be given by $g(y) = y^2$. Then the composition $g \circ f(x) = (2x + 1)^2$.

Theorem 7. *Let $f\colon X \to Y$ and $g\colon Y' \to Z$ be two functions with range$(f) \subseteq Y'$. Assume that f is continuous at x_0 and g is continuous at $y_0 = f(x_0)$. Then $g \circ f$ is continuous at x_0*

Proof. The idea of the proof is simple. If we can control the precision of the output $g(y)$ by the precision of the input of the outside function g, which is at the same time the output of the inside function f and if we can control this output of the inside function by its input x we can control the precision of $g(f(x))$ by the precision of x. Formally, we need to show that

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } \forall |x - x_0| < \delta, |g \circ f(x) - g \circ f(x_0)| < \varepsilon.$$

Since g is continuous at y_0 we have

$$\forall \varepsilon > 0 \quad \exists \gamma > 0 \text{ such that } \forall |y - y_0| < \gamma, |g(y) - g(y_0)| < \varepsilon.$$

Since f is continuous at x_0 we have

$$\forall \gamma > 0 \quad \exists \delta > 0 \text{ such that } \forall |x - x_0| < \delta, |f(x) - f(x_0)| < \gamma.$$

Now, for any $\varepsilon > 0$ we find γ such that $|y - y_0| < \gamma$ implies $|g(y) - g(y_0)| < \varepsilon$. For this γ we now find δ such that for all $|x - x_0| < \delta$ we have $|f(x) - f(x_0)| = |y - y_0| < \gamma$ and hence $|g \circ f(x) - g \circ f(x_0)| = |g(y) - g(y_0)| < \varepsilon$, as required. □

This theorem implies the following corollary on limits of compositions.

Corollary 2. *Let f be a function with*

$$\lim_{x\to x_0} f(x) = a$$

(x_0 does not have to belong to the domain of f) and let $g(y)$ be a function that is continuous at a. Then

$$\lim_{x\to x_0} g\circ f(x) = g(a).$$

Proof. If we define (or redefine) f so that $f(x_0) = a$, the so amended function f becomes continuous at x_0. By the theorem above now $g \circ f$ is continuous at x_0, hence

$$\lim_{x\to x_0} g\circ f(x) = g\circ f(x_0) = g(\lim_{x\to x_0} f(x)) = g(a). \quad \square$$

We say that a function f is continuous on a subset A of its domain X if it is continuous for any $x \in A$. We list some properties of functions that are continuous on closed intervals $[a, b] \subset \mathbb{R}$.

Theorem 8. *Let $f\colon [a, b] \to \mathbb{R}$ be a function, which is continuous on the closed interval $[a, b]$. Then*

1. *f is bounded, that is, there exists a number K such that $|f(x)| \leq K$ for all $x \in [a, b]$.*
2. *f assumes its minimum and maximum, that is, there exist x_{min} and x_{max} such that for all $x \in [a, b]$*

$$m = f(x_{min}) \leq f(x) \leq f(x_{max}) = M.$$

3. *(Intermediate Value Theorem [IVT]) f assumes all intermediate values $k \in [m, M]$, that is, for any $k \in [m, M]$ there exists $x_k \in [a, b]$ such that*

$$f(x_k) = k.$$

The proof of this theorem is a topic of MTHS130. It relies on the completeness property of the real numbers and would not be true if we stayed with rational numbers. Another short way of stating the theorem is the following.

Theorem 9. *If $f\colon [a, b] \to \mathbb{R}$ is a function, which is continuous on the closed interval $[a, b]$ then its range is a closed interval $[m, M]$.*

The following example is an application of the theorem above.

♦Example. Show that the range of $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$ is the set of all non-negative numbers.

The square function is continuous on any interval $[0, n]$ for any $n \in \mathbb{N}$. Now, 0 and n^2 are in the range and, by the theorem, all intermediate values in $[0, n^2]$ belong to the range. Since $n^2 \geq n$ for $n \geq 1$ and n can be chosen arbitrarily large, we see that any real number ≥ 0 is in the range. On the other hand, the range does not contain negative numbers, since squares are non-negative.

♠ *Exercises 33.* The Dirichlet function is defined on $X = [0, 1]$ as

$$f(x) = \begin{cases} 1 \text{ if } x \text{ is rational} \\ 0 \text{ if } x \text{ is irrational.} \end{cases}$$

Show that the Dirichlet function is nowhere continuous.

Let's now return to our aim to prove continuity of the elementary functions.

Any rational function $r(x) = \frac{p(x)}{q(x)}$ can be obtained from the continuous functions $f(x) = x$ and the constant function $f(x) = a$ by arithmetic operations and therefore is continuous on their natural domain, i.e. where $q(x) \neq 0$.

We discuss now the continuity of the inverse functions (where exist). We have seen that strictly monotone functions are injective. For continuous functions the converse is also true.

Theorem 10. *An injective, continuous function $f : [a, b] \to \mathbb{R}$ on a closed interval is strictly monotone.*

Proof. Injectivity implies $f(a) \neq f(b)$. Assume $f(a) < f(b)$ (Otherwise we can consider $-f$). We prove that f is strictly increasing. Choose any $x_1 \in (a, b)$. We show by contradiction that $f(x_1) < f(b)$. If $f(x_1) \geq f(b)$ then $f(b)$ is an intermediate value between $f(a)$ and $f(x_1)$ and, due to the IVT, there must be a point c between a and x_1 with $f(c) = f(b)$ which contradicts injectivity. In the same way one proves $f(a) < f(x_1)$.

Now choose $x_2 \in (x_1, b)$. The same argument from above applied with x_1, x_2, b instead of a, x_1, b yields $f(x_1) < f(x_2)$ as required. □

♠ *Exercises 34.* Show that the conclusion of the theorem above is also valid if f is an injective, continuous function on an open or semiclosed interval (a, b), $[a, b)$, $(b, a]$, or on a ray (a, ∞), $[a, \infty)$, $(-\infty, b)$, $(-\infty, b]$ or on $\mathbb{R}$.

Before we formulate and prove the next theorem on continuity of inverse functions we take a more geometric view on continuity of a function $f\colon X \to Y$ at some point $c \in X$. For any subset V of the codomain Y we define the *preimage* of V as the set

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

For this definition it does not matter whether f is invertible or not.

For $c \in X$ let $d = f(x)$. For $\varepsilon > 0$ let V be the intersection of the codomain V and the ε-neighbourhood of d, i.e.,

$$V = (d - \varepsilon, d + \varepsilon) \cap Y.$$

Now continuity means that $f^{-1}(V)$ contains a δ-neighbourhood of c, i.e., an interval

$$(c - \delta, c + \delta),$$

which is entirely mapped into V. In other words, no matter how we vary the input within that δ-neighbourhood, the corresponding outputs will stay in V, ε-close to $d = f(c)$.

Theorem 11. *Let f be a strictly monotone, continuous function on an open interval (a, b) (a, b can be finite or $\pm\infty$). Then the inverse function $g = f^{-1}$ is continuous on its domain.*

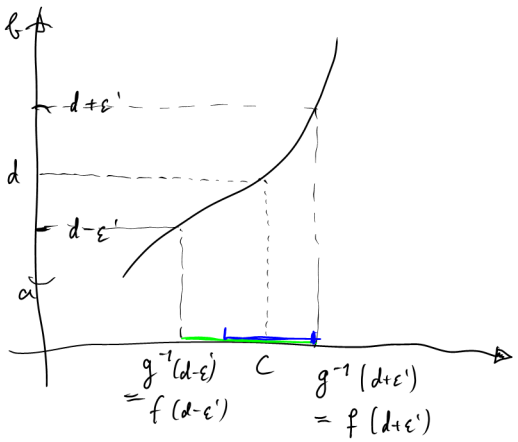
Proof. Without loss of generality, assume that f is strictly increasing. The case of strict decrease is analogous, or, alternatively, we may just consider $-f$ instead of f.

Let c be in the range of f, thus in the domain of g and let $d = g(c)$. We can shrink any given ε to ε' so that $[d - \varepsilon', d + \varepsilon'] \subset (a, b)$. Let

$$V = (d - \varepsilon', d + \varepsilon') \cap (a, b) = (d - \varepsilon', d + \varepsilon').$$

Now,

$$g^{-1}(V) = f(V) = (f(d - \varepsilon'), f(d + \varepsilon')).$$



The set $g^{-1}(V)$ is an open interval, which contains $f(d) = c$. It may not be a symmetric δ-neighbourhood of c but it contains such δ-neighbourhood with

$$\delta = \min\{f(d + \varepsilon') - c, c - f(d - \varepsilon')\}. \qquad \square$$

This Theorem shows, in particular, that the functions

$$\sqrt[m]{x}\colon [0,\infty) \to [0,\infty),$$

being inverse to the strictly increasing functions

$$x^m\colon [0,\infty) \to [0,\infty),$$

are continuous on their domains.

Trigonometric functions and their inverses. First we show that $\sin x$ is continuous at $x_0 = 0$, that is,

$$\lim_{x\to 0} \sin x = 0.$$

Since sin is an even function it suffices to show that

$$\lim_{x\to 0^+} \sin x = 0.$$

We will use the following form of the squeezing principle for limits of functions:

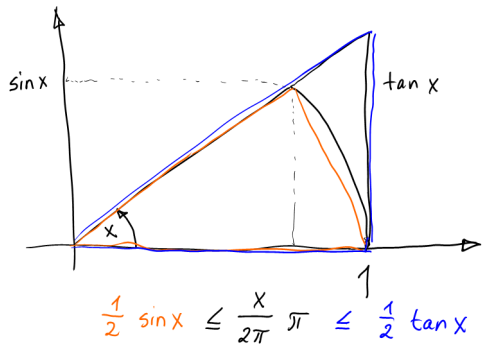
Theorem 12. *Let f, g, h be functions defined in a punctured neighbourhood $(a,b)\setminus\{x_0\}$ or one-sided neigbourhood (x_0,b) or (a,x_0) of x_0 and such that on the respective domain*

$$f(x) \le g(x) \le h(x).$$

If the limits/ right limits / left limits of f and h as x approaches x_0 exist and are equal to L then the limit/ right limit / left limit of g as x approaches x_0 also exists and equals L.

♠ *Exercises 35.* The proof is similar to the proof of the Squeeze theorem for sequences and is left as an exercise.

Comparing the area A of the triangle with vertices $(0,0)$, $(1,0)$, $(\cos x, \sin x)$, the area B of the sector of the unit circle formed by the angle x and the area C of the triangle with vertices $(0,0)$, $(1,0)$, $(1,\tan x)$ (see sketch below)



gives

$$0 \leq A = \frac{1}{2}\sin x \leq B = \frac{x}{2\pi}\pi \leq C = \frac{1}{2}\tan x.$$

It follows, for $x \in [0, \frac{\pi}{2}]$,

$$0 \leq \sin x \leq x \leq \tan x. \tag{4}$$

At this stage we only need the squeezing inequalities

$$0 \leq \sin x \leq x$$

which, by the squeezing principle give

$$\lim_{x \to 0^+} \sin x = 0$$

and hence

$$\lim_{x \to 0} \sin x = 0.$$

It follows that

$$\lim_{x \to 0} \cos x = \lim_{x \to 0} \sqrt{1 - \sin^2 x} = 1 = \cos 0,$$

which shows that cos is also continuous at $x_0 = 0$.

Now, for any x_0,

$$\lim_{x \to x_0} \sin x = \lim_{y \to 0} \sin(x_0 + y) = \lim_{y \to 0} \sin x_0 \cos y + \cos x_0 \sin y = \sin x_0.$$

This proves that $\sin x$ is continuous throughout $\mathbb{R}$. It follows that

$$\cos x = \sin(\frac{\pi}{2} - x)$$

is continuous throughout $\mathbb{R}$ and that

$$\tan x = \frac{\sin x}{\cos x}$$

is continuous on its natural domain, which excludes the zeros of cos, i.e., $\frac{\pi}{2} + k\pi$, where $k \in \mathbb{Z}$.

It follows from Theorem 11 that

$$\begin{aligned} \arcsin &: [-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}] \\ \arccos &: [-1, 1] \to [0, \pi] \\ \arctan &: \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2}) \end{aligned}$$

are continuous functions.

Exponential and Logarithmic functions. We have not given a rigorous definition of exponential and logarithmic functions yet and we also defer this to a time when we have more advanced techniques available. Strictly speaking, we have only defined exponential functions $f(x) = a^x$ for rational x. We need to fill the "gaps" left by irrational numbers in such a way that the resulting function is continuous. There are many ways of doing this. One way to define the exponential function with base e is

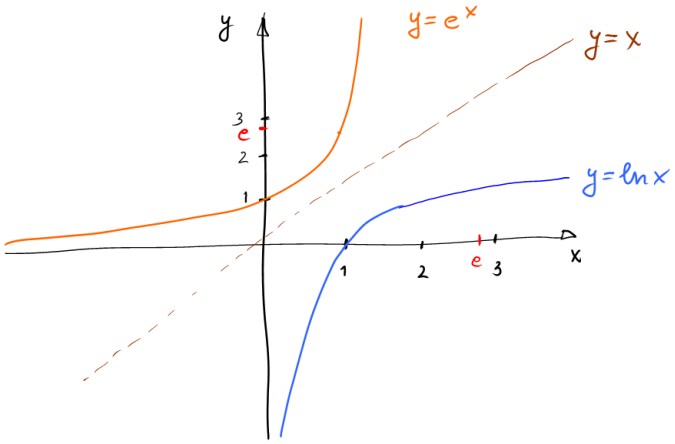
$$\mathrm{e}^x = \lim_{n\to\infty} (1 + \frac{x}{n})^n.$$

Although, we will not pursue this here. This function is defined for any real x, it is strictly increasing and continuous and it has a continuous inverse (called 'natural logarithm')

$$\ln\colon (0,\infty) \to \mathbb{R}.$$

The so defined functions satisfy the identities

$$\mathrm{e}^{x+y} = \mathrm{e}^x\,\mathrm{e}^y, \quad \mathrm{e}^{kx} = (\mathrm{e}^x)^k, \quad \ln(xy) = \ln x + \ln y, \quad \ln x^k = k \ln x.$$



Exponential functions with arbitrary base a ($a > 0, a \neq 1$) can be defined by

$$a^x = \mathrm{e}^{(\ln a)x}.$$

They are strictly increasing for $a > 1$ and strictly decreasing for $a < 1$ and they are continuous throughout $\mathbb{R}$.

12 Rates of change, derivatives and differentials of functions

In this lecture we start Differential Calculus. Consider a function $y = f(x)$. It can be viewed as a model of some process where a measurable quantity y changes over time x, e.g. y can be the distance travelled by a car at time x. We want to measure how rapidly the change takes place, e.g., how fast we travel. To do so we first pick two instants of time x_0 and x_1 and we compute the change of function f:

$$y_1 - y_0 = f(x_1) - f(x_0).$$

Now the ratio

$$m_{av} = \frac{y_1 - y_0}{x_1 - x_0}$$

measures the change relative to the time elapsed and is called the *average rate of change* on the interval $[x_0, x_1]$. In our example this would be the average velocity of the car during the time interval $[x_0, x_1]$.

Processes with constant rates of change are modelled by linear functions

$$y = f(x) = mx + b.$$

Indeed, if

$$\frac{y - y_0}{x - x_0} = m$$

for all $(x, y = f(x))$ then

$$y = f(x) = m(x - x_0) + y_0 = mx + b,$$

where y_0 is the initial value of $f(x)$ at the initial time x_0.

If a process $y = f(x)$ is taking place with non-constant velocity we can still compute the average rate of change during the interval $[x_0, x_1]$ by the same formula

$$m_{av} = \frac{y_1 - y_0}{x_1 - x_0}.$$

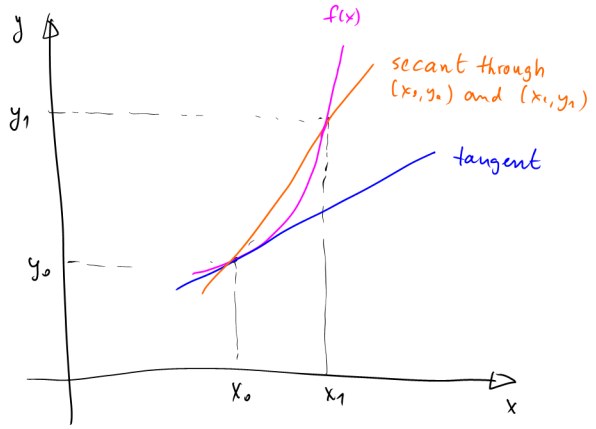
This is the rate of change that would return the same result if the process would unfold at a constant rate of change. Geometrically, m_{av} is the slope of the straight line passing through the points with coordinates $(x_0, y_0 = f(x_0))$ and $(x_1, y_1 = f(x_1))$. Such straight line through two points on the graph is called a *secant*.

If we are interested in the *instantaneous rate of change*, e.g., the velocity shown by the speedometer of your car at a particular instant, we need to make the interval

$[x_0, x_1]$ 'very small'. Using the mathematical technique of limits we can make this interval approach the length 0. We define the instantaneous rate of change

$$m = \lim_{x_1 \to x_0} \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

If this limit exists we say that the function f is *differentiable* at x_0 and we call m the *derivative* of the function $f(x)$ at x_0. The process of computing the derivative is called *differentiation*. Geometrically, m is the slope of the tangent line to the graph at x_0. This is illustrated in the picture below.



♦Example. The motion of free fall of an object caused by gravity is modelled by

$$y = h - ct^2,$$

where c, h are constants t is the time and y is the height at the time t. We compute the instantaneous velocity at the time t_1 as the limit of

$$\frac{h - ct_2^2 - (h - ct_1^2)}{t_2 - t_1} = \frac{-c(t_2^2 - t_1^2)}{t_2 - t_1} = \frac{-c(t_2 + t_1)(t_2 - t_1)}{t_2 - t_1},$$

as t_2 approaches t_1, while never being equal to t_1. Under this assumption we can cancel the factor $t_2 - t_1$, which yields

$$\lim_{t_2 \to t_1} \frac{-c(t_2 + t_1)(t_2 - t_1)}{t_2 - t_1} = \lim_{t_2 \to t_1} -c(t_2 + t_1) = -2ct_1.$$

The equation of the tangent line to the graph of $y = f(x)$ at the point (x_0, y_0) is

$$y = \ell(x) = y_0 + m(x - x_0).$$

The linear function that describes the tangent line approximates the function $f(x)$ in some neighbourhood of x_0 in the following sense: The 'error term' of the

approximation, i.e., the difference between $f(x)$ and the linear approximation $\ell(x)$ is

$$E(x) = f(x) - \ell(x) = f(x) - f(x_0) - m(x - x_0)$$

We have

$$\lim_{x \to x_0} \frac{E(x)}{x - x_0} = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} - m = 0.$$

This means that the error term $E(x)$ tends to zero faster than $x - x_0$, thus it can be neglected if x is close enough to x_0. We can express this using the o-notation as

$$E(x) = o(x - x_0).$$

Approximation of (complicated) non-linear functions by (simple) linear functions is the essence of differential calculus.

We will often write $\Delta x = x - x_0$, called the *increment of the argument* (it is also common to use h instead of Δx), and $\Delta f = f(x) - f(x_0)$ or $\Delta y = y - y_0$, called the increment of the function. Then the derivative at x_0 equals

$$m = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x}$$

and the function f can be expressed as

$$f(x) = f(x_0) + m\,\Delta x + E(x)$$

or

$$\Delta f = m\,\Delta x + E(x).$$

The linear function $m\,\Delta x$ as a function of the variable Δx is denoted

$$df = m\,\Delta x \tag{5}$$

and is called the *differentia*l of f at x_0.

The best linear approximation of a linear function is the linear function itself. For $f(x) = mx + b$ we get

$$df = m\,\Delta x.$$

In particular, for $f(x) = x$ we have

$$df = dx = \Delta x.$$

This allows us to rewrite the differential (5) of a function as

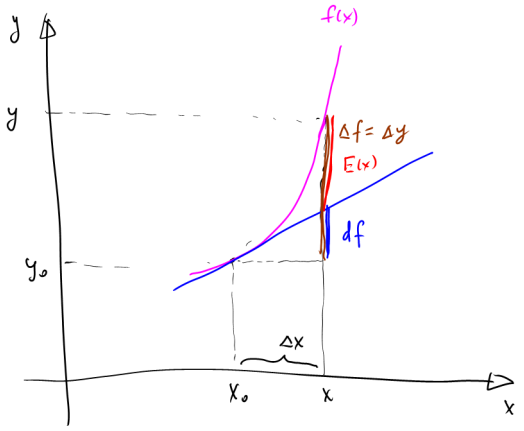
$$df = m\,dx,$$

and justifies the common notation for the derivative

$$m = \frac{df}{dx}.$$

Another common notation for the derivative of the function f at x_0 is $f'(x_0)$.

See the picture below for the geometric meaning of the differential.



♦Example. The derivative of the linear function $f(x) = mx + b$ at any point x_0 is m. Indeed,

$$f'(x_0) = \lim_{x \to x_0} \frac{\Delta f}{\Delta x} = \lim_{\Delta x \to 0} \frac{m\Delta x}{\Delta x} = m.$$

The differential at x_0 is $df = m\,dx$.

♦Example. The derivative of the function $g(x) = x^2$ at the point x_0 is $2x_0$. Indeed,

$$g'(x_0) = \lim_{x \to x_0} \frac{\Delta g}{\Delta x} = \lim_{x \to x_0} \frac{x^2 - x_0^2}{x - x_0} = \lim_{x \to x_0} \frac{(x - x_0)(x + x_0)}{x - x_0} = 2x_0.$$

The differential at x_0 is $dg = 2x_0\,dx$.

♦Example. Show that the function $f(x) = \sqrt[3]{x}$ is not differentiable at $x_0 = 0$.

Solution. We show that

$$\lim_{\Delta x \to 0} \frac{f(\Delta x) - f(0)}{\Delta x} = \infty.$$

Indeed,

$$\frac{f(\Delta x) - f(0)}{\Delta x} = \frac{\Delta x^{1/3}}{\Delta x} = \Delta x^{-\frac{2}{3}}.$$

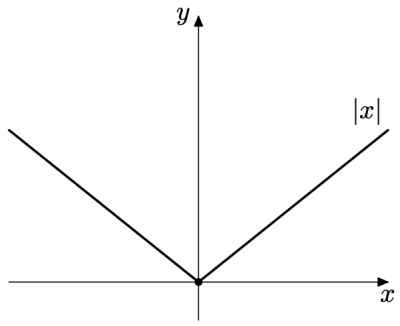
$$\forall M \quad \exists \delta = M^{-\frac{3}{2}} \text{ such that } 0 \neq |\Delta x| < \delta = M^{-\frac{3}{2}} \text{ implies } \frac{f(\Delta x) - f(0)}{\Delta x} > M.$$

Geometrically, this means that the graph of the cubic root function has a vertical tangent at 0.

♦Example. Show that $f(x) = |x|$ is not differentiable at $x = 0$.

Solution.

From the graph of $f(x) = |x|$ it is clear that there is a cusp at $x = 0$ and that the graph does not have a well-defined tangent there.



From our formal definition we have

$$\begin{aligned} f'(0) = \lim_{h \longrightarrow 0} \frac{f(0+h) - f(0)}{h - 0} &= \lim_{h \longrightarrow 0} \frac{|h| - 0}{h - 0} \\ &= \lim_{h \longrightarrow 0} \frac{|h|}{h}. \end{aligned}$$

However the sign of $\dfrac{|h|}{h}$ depends on the sign of h,

$$\frac{|h|}{h} = \begin{cases} 1, & h > 0 \\ -1, & h < 0. \end{cases}$$

So, we have

$$\lim_{h \longrightarrow 0^+} \frac{|h|}{h} = +1$$

and

$$\lim_{h \longrightarrow 0^-} \frac{|h|}{h} = -1.$$

Hence, the limit $\lim\limits_{h \longrightarrow 0} \dfrac{|h|}{h}$ does not exist and $f(x) = |x|$ is not differentiable at $x = 0$. □

One important point about this example is that $f(x) = |x|$ is continuous at $x = 0$. So continuity certainly **does not** imply differentiability. We would, however, expect the converse to be true.

We conclude this lecture by showing that differentiability of a function f at some point x_0 implies that the function is continuous at x_0. Indeed, differentiability means

that

$$f(x) - f(x_0) = f'(x_0)\Delta x + o(\Delta x).$$

The right hand side clearly tends to 0 as Δx tends to zero. Therefore,

$$\lim_{\Delta x \to 0} f(x) - f(x_0) = 0,$$

where $\Delta x = x - x_0$, which is equivalent to

$$\lim_{x \to x_0} f(x) = f(x_0),$$

i.e., continuity of f at x_0.

13 Derivatives of elementary functions

Our next aim is to compute the derivatives of power functions, trigonometric and exponential functions and to establish rules for differentiating sums, products, quotients and compositions of functions. This reduces the differentiation of elementary functions to the application of algebraic rules, rather than dealing with limits.

First we introduce the derivative of a function as a new function. If a function $f\colon X \to \mathbb{R}$ is differentiable at each point $x \in X$ we can form a new function $f'\colon X \to \mathbb{R}$ which assigns to each x the derivative of f at x. This function f' is also called the derivative of f. We will also use the notation introduced in the previous lecture

$$f' = \frac{df}{dx}.$$

If a function $y = f(t)$ models a process, where the independent variable is the time t then the derivative of f is often denoted by $\dot{f}(t)$ or just by $\dot{y}$. This is very common in physics.

Let's start by computing the **derivative of the power function** $f(x) = x^p$, where p is a natural number. We have to compute

$$f'(x) = \lim_{\Delta x \to 0} \frac{(x+\Delta x)^p - x^p}{\Delta x}.$$

According to the binomial formula

$$(x+\Delta x)^p = \sum_{k=0}^{p} \binom{p}{k} x^{p-k} \Delta x^k = x^p + p x^{p-1} \Delta x + o(\Delta x).$$

The expression $o(\Delta x)$ consists of finitely many terms with a factor Δx of power at least two. So, even after dividing it by Δx all terms have still a factor Δx and will tend to 0 as Δx tends to 0. It follows

$$\frac{(x+\Delta x)^p - x^p}{\Delta x} = \frac{p x^{p-1} \Delta x}{\Delta x} + \frac{o(\Delta x)}{\Delta x}$$

and

$$f'(x) = \lim_{\Delta x \to 0} \frac{p x^{p-1} \Delta x}{\Delta x} + \frac{o(\Delta x)}{\Delta x} = p x^{p-1}.$$

The notation $E(\Delta x) = o(\Delta x)$ is not an equality of two functions but merely expresses that $E(\Delta x)$ has a certain property, namely

$$\lim_{\Delta x \to 0} \frac{E(\Delta x)}{\Delta x} = 0.$$

This leads to the following simple rules. Let $E_1(\Delta x)$ and $E_2(\Delta x)$ be $o(\Delta x)$ and $F(\Delta x)$ be any bounded function. Then

1. $E_1(\Delta x)+E_2(\Delta x) = o(\Delta x)$, which can be expressed as $o(\Delta x)+o(\Delta x) = o(\Delta x)$.
2. $E_1(\Delta x)\cdot E_2(\Delta x) = o(\Delta x)$, which can be expressed as $o(\Delta x)\cdot o(\Delta x) = o(\Delta x)$. (In fact, $o(\Delta x)\cdot o(\Delta x) = o(\Delta x^2)$)
3. $E_1(\Delta x)\cdot F(\Delta x) = o(\Delta x)$, which can be expressed as $o(\Delta x)\cdot F(\Delta x) = o(\Delta x)$.

Next we compute the derivative of $f(x) = \sin x$. We use the addition formula

$$\sin(x + \Delta x) = \sin x \cos \Delta x + \cos x \sin \Delta x.$$

It follows

$$\frac{\Delta f}{\Delta x} = \frac{\sin x \cos \Delta x + \cos x \sin \Delta x - \sin x}{\Delta x} = \sin x \frac{\cos \Delta x - 1}{\Delta x} + \cos x \frac{\sin \Delta x}{\Delta x}.$$

We compute the limits

$$\lim_{\Delta x \to 0} \frac{\sin \Delta x}{\Delta x} = 1 \text{ and } \lim_{\Delta x \to 0} \frac{\cos \Delta x - 1}{\Delta x} = 0.$$

The inequalities (4) can be reformulated as

$$\cos \Delta x \leq \frac{\sin \Delta x}{\Delta x} \leq 1.$$

Since

$$\lim_{\Delta x \to 0} \cos \Delta x = \lim_{\Delta x \to 0} 1 = 1$$

the squeeze theorem implies

$$\lim_{\Delta x \to 0} \frac{\sin \Delta x}{\Delta x} = 1.$$

We have

$$\frac{\cos \Delta x - 1}{\Delta x} = \frac{(\cos \Delta x - 1)(\cos \Delta x + 1)}{\Delta x(\cos \Delta x + 1)} = \frac{\cos^2 \Delta x - 1}{\Delta x(\cos \Delta x + 1)} = -\frac{\sin^2 \Delta x}{\Delta x(\cos \Delta x + 1)}$$

and hence

$$\lim_{\Delta x \to 0} \frac{\cos \Delta x - 1}{\Delta x} = \lim_{\Delta x \to 0} -\frac{\sin \Delta x}{\Delta x} \frac{\sin \Delta x}{(\cos \Delta x + 1)} = -1 \cdot 0 = 0.$$

This can be expressed as

$$\cos \Delta x - 1 = o(\Delta x).$$

Now,

$$f'(x) = \lim_{\Delta x \to 0} \sin x \frac{\cos \Delta x - 1}{\Delta x} + \cos x \frac{\sin \Delta x}{\Delta x} = \cos x.$$

The differential $d\sin$ at 0 is $d\sin = 1 \cdot dx$. This gives the approximation

$$\sin \Delta x \approx \Delta x$$

for small Δx. Bear in mind that the angle x has to be measured in radians!

A similar approximation holds for tan:

$$\tan \Delta x \approx \Delta x.$$

For completeness we also give a somewhat handwavy computation of the derivative of exponential functions, leaving a rigorous treatment for later. We assume the addition formula for exponential functions $f(x) = a^x$

$$a^{x+\Delta x} = a^x \cdot a^{\Delta x}.$$

Then

$$\frac{a^{x+\Delta x} - a^x}{\Delta x} = a^x \frac{a^{\Delta x} - 1}{\Delta x}.$$

The expression on the RHS is a product where the first factor does not depend on Δx, so it behaves like a constant for the limit as $\Delta x \to 0$. The second factor does not depend on x and tends to a constant (if the limit exists). In fact, the limit

$$\lim_{\Delta x \to 0} \frac{a^{\Delta x} - 1}{\Delta x}$$

does exist and equals $\ln a$. (We will prove this later.) Hence,

$$f'(x) = a^x \ln a.$$

In particular, the derivative of e^x is $\mathrm{e}^x \ln \mathrm{e} = \mathrm{e}^x$.

We will now establish some algebraic rules for derivatives of sums, products, quotients, compositions and inverse functions.

Sum rule. Let f and g be functions that have derivatives at some point x_0. Then $f \pm g$ also has a derivative at x_0 and

$$(f \pm g)'(x_0) = f'(x_0) \pm g'(x_0).$$

Proof.

$$\begin{aligned}(f \pm g)'(x_0) &= \lim_{\Delta x \to 0} \frac{(f \pm g)(x_0 + \Delta x) - (f \pm g)(x_0)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \pm \lim_{\Delta x \to 0} \frac{g(x_0 + \Delta x) - g(x_0)}{\Delta x} = f'(x_0) \pm g'(x_0).\end{aligned}$$

Product rule. Let f and g be functions that have derivatives at some point x_0. Then fg also has a derivative at x_0 and

$$(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0).$$

Notice that the derivative of a product is NOT the product of the derivatives.

Proof. We give a proof that demonstrates the advantage of the $o(\Delta x)$ formalism.

$$f(x) = f(x_0) + f'(x_0)\Delta x + o(\Delta x)$$
$$g(x) = g(x_0) + g'(x_0)\Delta x + o(\Delta x)$$

Multiplication yields

$$\begin{aligned} f(x)g(x) = f(x_0)g(x_0) + [f(x_0)g'(x_0) + f'(x_0)g(x_0)]\Delta x \\ + [f(x_0) + f'(x_0)\Delta x]o(\Delta x) + [g(x_0) + g'(x_0)\Delta x]o(\Delta x) \\ + f'(x_0)g'(x_0)(\Delta x)^2 + o(\Delta x). \end{aligned}$$

(Here we have used that $o(\Delta x) \cdot o(\Delta x) = o(\Delta x)$.) We need to show that

$$[f(x_0)+f'(x_0)\Delta x]o(\Delta x)+[g(x_0)+g'(x_0)\Delta x]o(\Delta x)+f'(x_0)g'(x_0)(\Delta x)^2+o(\Delta x) = o(\Delta x).$$

This is clearly the case since

1. $[f(x_0) + f'(x_0)\Delta x]$ is bounded and hence $[f(x_0) + f'(x_0)\Delta x]o(\Delta x) = o(\Delta x)$,
2. $[g(x_0) + g'(x_0)\Delta x]$ is bounded and hence $[g(x_0) + g'(x_0)\Delta x]o(\Delta x) = o(\Delta x)$,
3. $\lim_{\Delta x \to 0} \dfrac{f'(x_0)g'(x_0)(\Delta x)^2}{\Delta x} = \lim_{\Delta x \to 0} f'(x_0)g'(x_0)\Delta x = 0$,
4. all four terms in the expression above are $o(\Delta x)$ and therefore, so is the sum.

Quotient Rule. Let f and g be functions that have derivatives at some point x_0 and assume that $g'(x_0) \neq 0$. Then $\frac{f}{g}$ also has a derivative at x_0 and

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{(g(x_0))^2}.$$

Proof. We show that

$$\lim_{\Delta x \to 0} \frac{\frac{f(x_0+\Delta x)}{g(x_0+\Delta x)} - \frac{f(x_0)}{g(x_0)}}{\Delta x} = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{(g(x_0))^2}.$$

We have

$$\begin{aligned} \frac{f(x_0 + \Delta x)}{g(x_0 + \Delta x)} - \frac{f(x_0)}{g(x_0)} &= \frac{f(x_0 + \Delta x)g(x_0) - g(x_0 + \Delta x)f(x_0)}{g(x_0 + \Delta x)g(x_0)} \\ &= \frac{f(x_0 + \Delta x)g(x_0) - f(x_0)g(x_0) - (g(x_0 + \Delta x)f(x_0) - f(x_0)g(x_0))}{g(x_0 + \Delta x)g(x_0)}. \end{aligned}$$

Now,

$$\lim_{\Delta x\to 0}\frac{\frac{f(x_0+\Delta x)}{g(x_0+\Delta x)}-\frac{f(x_0)}{g(x_0)}}{\Delta x}=$$
$$\lim_{\Delta x\to 0}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}\frac{g(x_0)}{g(x_0+\Delta x)g(x_0)}-\frac{g(x_0+\Delta x)-g(x_0)}{\Delta x}\frac{f(x_0)}{g(x_0+\Delta x)g(x_0)}$$
$$=\frac{f'(x_0)g(x_0)-f(x_0)g'(x_0)}{(g(x_0))^2}\qquad\square$$

Chain rule. Chain rule is the rule for differentiating compositions of functions. Let $F(x)=f\circ g(x)=f(g(x))$ be the composition of two functions f,g such that g is differentiable at x_0 and f is differentiable at $y_0=g(x_0)$.Then $F(x)$ is differentiable at x_0 and $F'(x_0)=f'(y_0)g'(x_0)$.

The chain rule becomes very natural if stated in terms of the differentials

$$d(f\circ g)=df\circ dg.$$

The differential of a composition is the composition of the differentials. Composing the two differentials (which are linear functions) means to multiply their slopes.

Proof. The idea of the proof is just to plug in the differential plus error term of the inside function into the differential plus error term of the outside function. We write

$$y=g(x)=g(x_0)+g'(x_0)\Delta x+E_g(\Delta x),$$

where $E_g(\Delta x)$ is the error term such that

$$\lim_{\Delta x\to 0}\frac{E_g(\Delta x)}{\Delta x}=0. \tag{6}$$

We define

$$\alpha(\Delta x)=\begin{cases}\frac{E_g(\Delta x)}{\Delta x} & \text{if } \Delta x\neq 0\\ 0 & \text{if } \Delta x=0.\end{cases}$$

Then

$$y=g(x)=g(x_0)+g'(x_0)\Delta x+\alpha(\Delta x)\cdot\Delta x$$

and

$$\Delta y=g(x)-g(x_0)=g'(x_0)\Delta x+\alpha(\Delta x)\cdot\Delta x. \tag{7}$$

Similarly,

$$f(y)=f(y_0)+f'(y_0)\Delta y+E_f(\Delta y)$$

where

$$\lim_{\Delta y\to 0}\frac{E_f(\Delta y)}{\Delta y}=0.$$

We define

$$\beta(\Delta y) = \begin{cases} \frac{E_f(\Delta y)}{\Delta y} & \text{if } \Delta y \neq 0 \\ 0 & \text{if } \Delta y = 0. \end{cases}$$

Notice that the so defined function β is continuous at 0. We have

$$f(y) = f(y_0) + f'(y_0)\Delta y + \alpha(\Delta y) \cdot \Delta y. \tag{8}$$

Now plugging (7) into (8) yields

$$f(g(x)) = f(g(x_0))+f'(g(x_0))[g'(x_0)\Delta x+\alpha(\Delta x)\cdot\Delta x]+\beta(\Delta y)[g'(x_0)\Delta x+\alpha(\Delta x)\cdot\Delta x],$$

which is equivalent to

$$F(x) = f(x_0) + f'(g(x_0))g'(x_0)\Delta x + [\alpha(\Delta x) + \beta(\Delta y)g'(x_0) + \beta(\Delta y)\alpha(\Delta x)]\Delta x.$$

All we need is to show that the term in square brackets

$$[\alpha(\Delta x) + \beta(\Delta y)g'(x_0) + \beta(\Delta y)\alpha(\Delta x)]$$

tends to 0 as Δx tends to zero. By (6)

$$\lim_{\Delta x \to 0} \alpha(\Delta x) = 0.$$

The only somewhat delicate point to show that

$$\lim_{\Delta x \to 0} \beta(\Delta y(\Delta x)) = 0.$$

We know that

$$\lim_{\Delta y \to 0} \beta(\Delta y) = 0.$$

but now Δy is a function of Δx (defined in (7)). The function $\Delta y(\Delta x)$ is continuous at 0 and $\Delta y(0) = 0$. By Corollary 2,

$$\lim_{\Delta x \to 0} \beta(\Delta y(\Delta x)) = \beta(0) = 0.$$

Therefore, also the terms

$$\beta(\Delta y)g'(x_0) + \beta(\Delta y)\alpha(\Delta x)$$

tend to 0 as $\Delta x \to 0$, as required. □

♠ *Exercises 36.* You may come across the following short "proof" of the chain rule:

$$F' = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta y}\frac{\Delta y}{\Delta x}.$$

Since $\lim_{\Delta x\to 0}\Delta y = 0$, making Δy small as Δx becomes small, we have

$$\lim_{\Delta x\to 0}\frac{\Delta f}{\Delta y} = \lim_{\Delta y\to 0}\frac{\Delta f}{\Delta y} = f'(y_0),$$

hence

$$F' = \lim_{\Delta y\to 0}\frac{\Delta f}{\Delta y}\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x} = f'(y_0)g'(x_0).$$

Find the mistake. Hint. Consider the case when the inside function $g(x)$ is constant.

Derivative of the inverse function. Let $f(x)$ be a continuous function that is differentiable at x_0 with derivative $f'(x_0)$. Assume that $f(x_0) = y_0$ and $f'(x_0) \neq 0$. Assume that f has an inverse g in some neighbourhood of y_0. Then the inverse function $g(y)$ is differentiable at y_0 and $g'(y_0) = \frac{1}{f'(x_0)}$.

Proof. We have

$$g'(y_0) = \lim_{\Delta y\to 0}\frac{\Delta x}{\Delta y} = \frac{1}{\lim_{\Delta y\to 0}\frac{\Delta y}{\Delta x}}.$$

Since f is continuous and has an inverse, it is strictly monotone. Therefore $\Delta y \neq 0$ iff $\Delta x \neq 0$. Since $\Delta y \to 0$ as $\Delta x \to 0$,

$$\lim_{\Delta y\to 0}\frac{\Delta y}{\Delta x} = \lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x} = f'(x_0).$$

Therefore,

$$g'(y_0) = \frac{1}{f'(x_0)} = \frac{1}{f'(g(y_0))}. \qquad \square$$

The rules above show that all elementary functions are differentiable in their natural domains (except for inverse functions of functions with vanishing derivative). We compute the derivatives of the most common functions.

1. Polynomials: $f(x) = \sum_{k=0}^{n} a_k x^k$, $f'(x) = \sum_{k=1}^{n} k a_k x^{k-1}$. The derivative of a polynomial is a polynomial of order decreased by 1.

2. Roots: $g(y) = \sqrt[p]{y} = y^{\frac{1}{p}}$ is the inverse function of $f(x) = x^p$. We have $f'(x) = px^{p-1}$.Therefore,

$$g'(y) = \frac{1}{px^{p-1}} = \frac{1}{p}x^{1-p} = \frac{1}{p}y^{\frac{1-p}{p}} = \frac{1}{p}y^{\frac{1}{p}-1}.$$

3. Trigonometric functions: We know that for $f(x) = \sin x$, $f'(x) = \cos x$. Now, $g(x) = \cos x = \sin(\frac{\pi}{2} - x)$ implies

$$g'(x) = -\cos(\frac{\pi}{2} - x) = -\sin x.$$

For $h(x) = \tan x = \frac{\sin x}{\cos x}$ the quotient rule yields

$$h'(x) = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = 1 + \tan^2 x = \frac{1}{\cos^2 x}.$$

4. Inverse sine: $g(y) = \arcsin y$. Then

$$g'(y) = \frac{1}{\cos x} = \frac{1}{\sqrt{1 - \sin^2 x}} = \frac{1}{\sqrt{1 - y^2}}.$$

5. Inverse tangent: $g(y) = \arctan y$. Then

$$g'(y) = \frac{1}{1 + \tan^2 x} = \frac{1}{1 + y^2}.$$

6. Logarithmic functions: $g(y) = \log_a y$ is the inverse function of $f(x) = a^x$ with $f'(x) = a^x \ln a$. Now,

$$g'(y) = \frac{1}{a^x \ln a} = \frac{1}{x \ln a}.$$

Alternatively, we can compute

$$g'(y) = \lim_{\Delta y \to 0} \frac{\log_a(y + \Delta y) - \log_a y}{\Delta y}$$

We have

$$\frac{\log_a(y + \Delta y) - \log_a y}{\Delta y} = \frac{1}{y} \frac{y}{\Delta y} \log_a(1 + \frac{\Delta y}{y}) = \frac{1}{y} \log_a \left(1 + \frac{\Delta y}{y}\right)^{\frac{y}{\Delta y}}.$$

Knowing that $\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = \mathrm{e}$, we guess that

$$\lim_{\Delta y \to 0} \log_a \left(1 + \frac{\Delta y}{y}\right)^{\frac{y}{\Delta y}} = \lim_{u \to \infty} \left(1 + \frac{1}{u}\right)^u = \mathrm{e},$$

with $u = \frac{y}{\Delta y}$. Therefore,

$$g'(y) = \frac{1}{y} \log_a \mathrm{e} = \frac{1}{y \ln a}.$$

Using this result one could compute the derivative of $f(x) = a^x$.

7. Powers with arbitrary exponents: Let $f(x) = x^p$ where $x > 0$ and p is an arbitrary real number. Then $f(x) = e^{p \ln x}$. According to chain rule,

$$f'(x) = e^{p \ln x} \frac{p}{x} = p x^{p-1}.$$

Thus, the power rule derived above extends to arbitrary powers.

14 Monotone Functions and Concavity

Increase and Decrease

For functions that are differentiable on some interval monotonicity is closely related to the sign of the derivative.

Theorem 13. *If f is increasing/decreasing in some neighbourhood of a point x_0 and if f is differentiable at x_0 then*

$$f'(x_0) \geq 0 \quad / \quad f'(x_0) \leq 0.$$

Proof. We restrict to the case when f is increasing. Then for $\Delta x > 0$, $\Delta f \geq 0$ and for $\Delta x < 0$, $\Delta f \leq 0$. In both cases

$$\frac{\Delta f}{\Delta x} \geq 0$$

and therefore

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} \geq 0. \qquad \square$$

Theorem 14. *Suppose that the function f is continuous on $[a, b]$ and differentiable on (a, b).*

(a) If $f'(x) > 0$ $\forall x \in (a, b)$ then f is strictly increasing on $[a, b]$.

(b) If $f'(x) < 0$ $\forall x \in (a, b)$ then f is strictly decreasing on $[a, b]$.

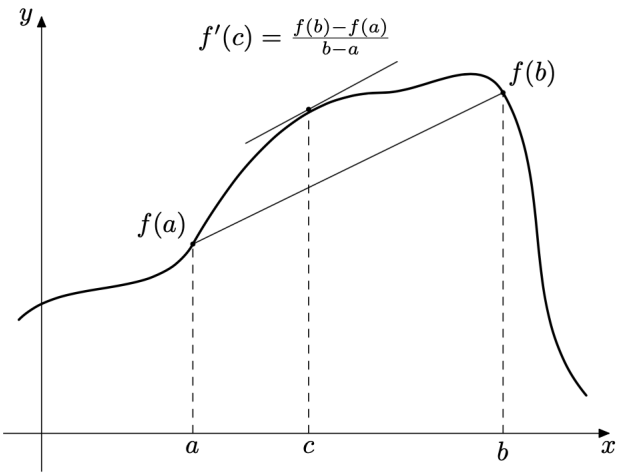
(c) If $f'(x) = 0$ $\forall x \in (a, b)$ then f is constant on $[a, b]$.

In the proof of this theorem we will use the Mean Value Theorem of Differential Calculus. We formulate this plausible theorem here, but defer its proof to MTHS130.

Theorem 15. *Suppose that the function f is continuous on $[a, b]$ and differentiable on (a, b). Then there exists a point $c \in (a, b)$ such that*

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

Geometrically, the theorem means that there is a point $c \in (a, b)$ such that the tangent to the graph of f at $(c, f(c))$ has the same slope as the secant through the points $(a, f(a))$ and $(b, f(b))$. See the picture below.



Proof of Theorem 14.

(a) Let x_1, x_2 be two points on $[a, b]$ with $x_1 < x_2$. We have to show that $f'(a) > 0$ implies $f(x_1) < f(x_2)$. Note that all assumptions of the Mean Value Theorem are satisfied on $[a, b]$ and, in particular, on the subinterval $[x_1, x_2]$. We have

$$f'(c) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}, \quad \text{for some} \quad c \in (x_1, x_2).$$

Now $x_2 - x_1 > 0$ and $f'(c) > 0$ so

$$f(x_2) - f(x_1) = (x_2 - x_1)f'(c) > 0,$$

as required.

(b) and (c) can be proved using the same techniques. □

There is also a 'weak' version' of Theorem 14: If we replace $f'(x) > 0$ by the weaker condition $f'(x) \geq 0$ then we can still conclude that the function is (weakly) increasing, and if $f'(x) \leq 0$ then the function is (weakly) decreasing.

♦Example. Determine the intervals on which the following functions are strictly increasing or strictly decreasing

(a) $f(x) = x^4$ (b) $f(x) = x^2 - 5x + 6$ (c) $f(x) = \mathrm{e}^x$ (d) $f(x) = x \ln x$.

Solution.

(a) $f(x) = x^4$ is defined and differentiable throughout $\mathbb{R}$. We have

$$f'(x) = 4x^3,$$

so $f'(x) > 0$ for $x > 0$ and $f'(x) < 0$ for $x < 0$. That is,

$$\begin{aligned} &x^4 \text{ is strictly increasing for } x \geq 0, \text{ and} \\ &x^4 \text{ is strictly decreasing for } x \leq 0. \end{aligned}$$

(b) $f(x) = x^2 - 5x + 6$ is defined and differentiable throughout $\mathbb{R}$. We have

$$f'(x) = 2x - 5,$$

so $f'(x) > 0$ for $x > \dfrac{5}{2}$ and $f'(x) < 0$ for $x < \dfrac{5}{2}$. That is

$$\begin{aligned} &x^2 - 5x + 6 \text{ is strictly increasing on } \left[\frac{5}{2}, \infty\right), \text{ and} \\ &x^2 - 5x + 6 \text{ is strictly decreasing on } \left(-\infty, \frac{5}{2}\right]. \end{aligned}$$

(c) The derivative of $f(x) = \mathrm{e}^x$ is $\mathrm{e}^x > 0$. Therefore the function e^x is strictly increasing throughout $\mathbb{R}$.

(d) $f(x) = x \ln x$ is defined and differentiable on $(0, \infty)$. We have

$$\begin{aligned} f'(x) &= 1 \cdot \ln x + x \cdot \frac{1}{x}, \quad \text{by the product rule.} \\ \text{i.e., } f'(x) &= 1 + \ln x. \end{aligned}$$

So f is strictly increasing when $1 + \ln x > 0$ i.e., $\ln x > -1$. Since $\mathrm{e}^a > \mathrm{e}^b$ if and only if $a > b$ (by part (c)), $x = \mathrm{e}^{\ln x} > \mathrm{e}^{-1}$, so we finally have $f(x)$ is strictly increasing if

$$x \geq \mathrm{e}^{-1} = \frac{1}{\mathrm{e}}.$$

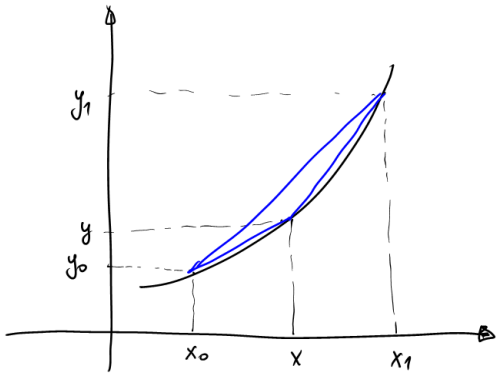
A similar argument leads to the conclusion that $f(x)$ is strictly decreasing when $0 < x \leq \dfrac{1}{\mathrm{e}}$. □

♠ *Exercises 37.* a. Show that the function $f(x) = \frac{1}{x-5}$ is strictly decreasing on the intervals $(-\infty, 5)$ and $(5, \infty)$ but not on $\mathbb{R} \setminus \{5\}$.
b. Show that the function $f(x) = x^3$ is strictly increasing on $(-\infty, 0]$ and $[0, \infty)$ and hence on $\mathbb{R}$.

Concavity

The concept of concavity of a function reflects the increase of the slope of the tangents. As the tangents become steeper the graph of the function "bends up". We

formalise this in the following way. Let $f\colon [a,b] \to \mathbb{R}$ be a function. Let $x_0 < x < x_1$ be three points in $[a,b]$ and $y_0 = f(x_0)$, $y = f(x)$ and $y_1 = f(x_1)$. We call f *concave up*[9] on the interval $[a,b]$ if for any such choice of x_0, x, x_1 the slope of the secant through $P_0(x_0, y_0)$ and $P(x,y)$ is not greater than the slope of the secant through $P(x,y)$ and $P(x_1, y_1)$. This is equivalent to saying that the secant through $P_0(x_0, y_0)$ and $P_1(x_1, y_1)$ lies above the graph of f. See sketch below



We formalise the statements from above. The equation of the secant through P_0 and P_1 is

$$y_{sec} = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0.$$

Therefore, the statement that the secant lies above the graph means

$$f(x) = y \leq y_{sec} = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0.$$

This is equivalent to

$$\begin{aligned}
y &\leq \frac{x_1 - x}{x_1 - x_0} y_0 + \frac{x - x_0}{x_1 - x_0} y_1 \\
y\frac{x_1 - x_0}{x_1 - x} &\leq y_0 + \frac{x - x_0}{x_1 - x} y_1 \\
y(1 + \frac{x - x_0}{x_1 - x}) &\leq y_0 + \frac{x - x_0}{x_1 - x} y_1 \\
y - y_0 &\leq \frac{x - x_0}{x_1 - x}(y_1 - y) \\
\frac{y - y_0}{x - x_0} &\leq \frac{y_1 - y}{x_1 - x}.
\end{aligned}$$

The latter inequality means that the slope of the secant through P_0 and P is not greater than the slope of the secant through P and P_1.

So far, no assumptions on the differentiability of the function f have been made. We have just compared some quantities derived from three points on the graph of

[9] The notion "concave up" is the same as "convex".

f. We assume now that the function is concave up and differentiable on the interval (a, b). Let now $x_0 < x < x^* < x_1$ be four points in (a, b) and y_0, y, y^*, y_1 be the corresponding values of f. Then

$$\frac{y - y_0}{x - x_0} \leq \frac{y^* - y}{x^* - x} \text{ and } \frac{y^* - y}{x^* - x} \leq \frac{y_1 - y^*}{x_1 - x^*}.$$

Passing $x_0 \to x$ in the left inequality yields

$$f'(x) \leq \frac{y^* - y}{x^* - x}$$

and passing $x_1 \to x^*$ in the right inequality yields

$$\frac{y^* - y}{x^* - x} \leq f'(x^*),$$

thus, for $x < x^*$, $f'(x) \leq f'(x^*)$, that is, f' is increasing.

♠ *Exercises 38.* Using the notation above, show that the strong inequalities $\frac{y-y_0}{x-x_0} < \frac{y_1-y}{x_1-x}$ imply that f' is strictly increasing.

The converse statement below is again a consequence of the MVT.

Theorem 16. *If f is differentiable on (a, b) and f' is increasing then f is concave up.*

Proof. Let $x_0 < x < x_1$ be three points in the interval (a, b). Then, by the MTV, there exist $c_1 \in (x_0, x)$ and $c_2 \in (x, x_1)$ such that

$$\frac{y - y_0}{x - x_0} = f'(c_1) \leq f'(c_2) = \frac{y_1 - y}{x_1 - x}. \quad \square$$

If a function f is two times differentiable, that is, its derivative f' has a derivative itself then we can use this second derivative to detect increase of f'. The second derivative is denoted by

$$f''(x) = \frac{d^2 f}{dx^2}.$$

Derivatives of even higher order can be defined as long as the derivatives are still differentiable. They are denoted by f''', f^{IV} etc. (Roman numbers as superscripts), or by $f^{(4)}, f^{(5)}, \ldots f^{(n)}$ (Arabic numbers in parentheses), or by $\frac{d^3 f}{dx^3}, \ldots, \frac{d^n f}{dx^n}$.

We have the following criterion.

Theorem 17. *If f is differentiable twice on (a, b) then f is concave up if and only if $f''(x) \geq 0$.*

The concept of *concavity down* is similar, just change the $\leq$ by $\geq$. Thus a function is concave down if the slopes of the secants are decreasing, or, if the function is differentiable, the derivative is decreasing, or, if the function is differentiable twice, the second derivative is non-positive.

What happens between the concave up and concave down portions of a graph? The point at which a graph changes from being concave up to concave down (or vice versa) is known as a point of *inflection*. If the function has a continuous second derivative on an interval where the graph changes from being concave up to concave down then we must have $f''(c) = 0$ where c is the point of inflection.

♦Example. Determine the open intervals on which the following functions are concave up or concave down.

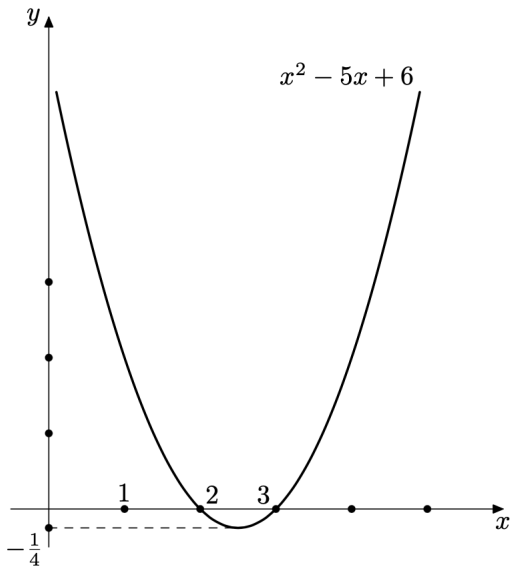
(a) $f(x) = x^2 - 5x + 6$ (b) $f(x) = 2x^4 - 3x^2$

(c) $f(x) = x \ln x$ (d) $f(x) = xe^{-x}$.

Solution

(a)
$$\begin{aligned} f(x) &= x^2 - 5x + 6 \\ f'(x) &= 2x - 5 \\ f''(x) &= 2 > 0. \end{aligned}$$

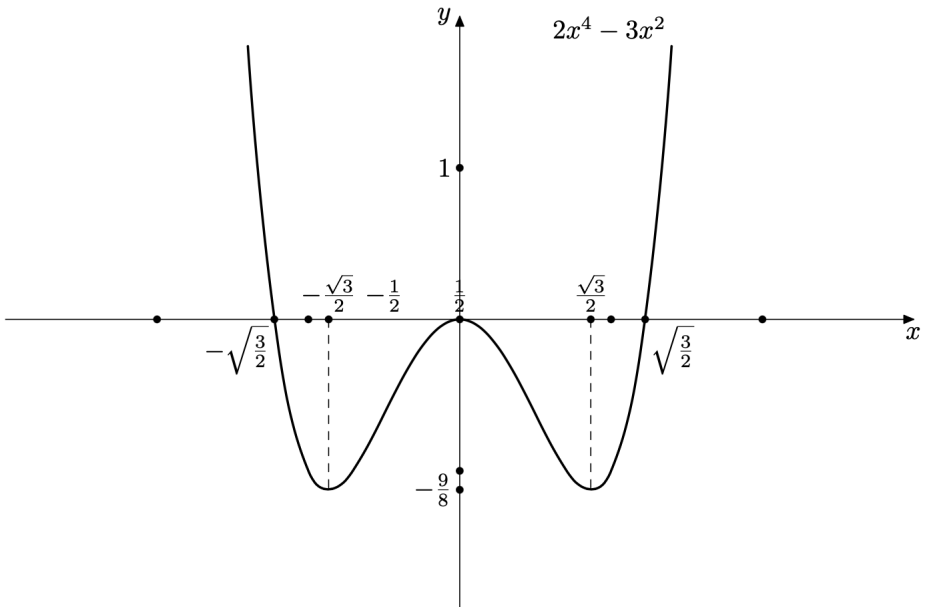
The function is concave up throughout $\mathbb{R}$.



(b)

$$\begin{aligned} f(x) &= 2x^4 - 3x^2 \\ f'(x) &= 8x^3 - 6x \\ f''(x) &= 24x^2 - 6 \\ &= 24\left(x^2 - \frac{1}{4}\right). \end{aligned}$$

So we have $f''(x) > 0$ for $|x| > \frac{1}{2}$, i.e., $x > \frac{1}{2}$ or $x < -\frac{1}{2}$, and $f''(x) < 0$ for $|x| < \frac{1}{2}$. That means f is concave up for $x \in (-\infty, -\frac{1}{2}) \cup \left(\frac{1}{2}, \infty\right)$ and concave down for $x \in \left(-\frac{1}{2}, \frac{1}{2}\right)$.

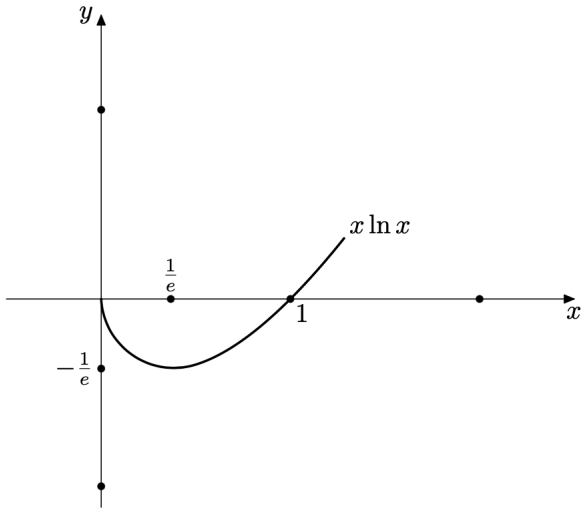


There are points of inflection at $x = \pm\frac{1}{2}$.

(c)

$$\begin{aligned} f(x) &= x \ln x \\ f'(x) &= \ln x + 1 \\ f''(x) &= \frac{1}{x}. \end{aligned}$$

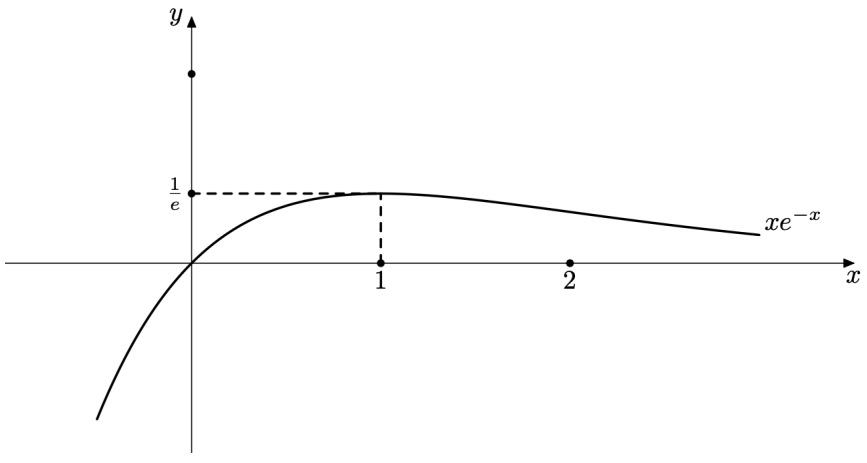
Now the natural domain for $f(x)$ is $(0, \infty)$. On this domain $f''(x) = \frac{1}{x}$ is positive so $f(x)$ is concave up on its entire domain.



(d)

$$\begin{aligned} f(x) &= xe^{-x} \\ f'(x) &= e^{-x} - xe^{-x} \\ f''(x) &= xe^{-x} - 2e^{-x} \\ &= e^{-x}(x-2). \end{aligned}$$

Now e^a is positive for any real number a, so the sign of $f''(x)$ is determined by the sign of $(x-2)$. We have f concave up for $x \in (2, \infty)$ and concave down for $x \in (-\infty, 2)$. There is a point of inflection at $x = 2$.



♠ *Exercises 39.*

1. Give the proofs for parts (b) and (c) of Theorem 14.

2. Determine the intervals on which the following functions are (i) increasing or decreasing and (ii) concave up or concave down.

(a) $\sin x$ (b) $e^{-\frac{x^2}{2}}$ (c) $x^3 - 9x^2 + 24x$

(d) $\ln(x^2 - x + 1)$ (e) $\tan x$, on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

3. (a) Show that $f(x) = e^x - x$ is an increasing function on $[0, \infty)$. Hence conclude that

$$e^x \geq 1 + x, \quad \text{for} \quad x \geq 0.$$

(b) Show that

$$e^x \geq 1 + x + \frac{1}{2}x^2, \quad \text{for } x \geq 0.$$

(c) Prove that

$$e^x \geq 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!},$$

for any $n \in \mathbb{N}$ and $x \geq 0$.

[Hint: Use induction.]

15 Application: Optimisation

When we model some process by a function we may be interested in finding the input(s) for which the function takes its minimal or maximal value. In general, we would like to find the absolute (or global) extrema of the function f, that is the arguments for which the function attains the maximal or minimal values: x_{max} such that

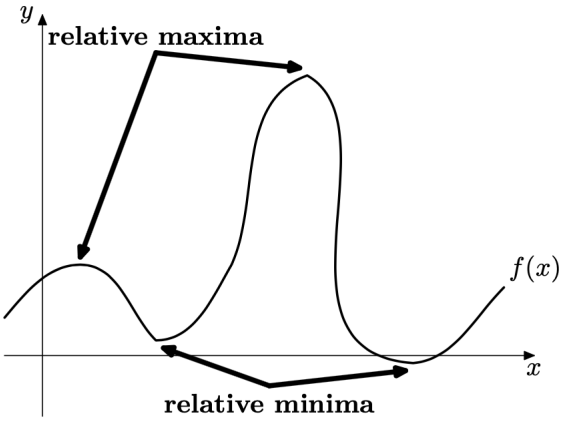
$$f(x) \leq f(x_{max})$$

for any x from the domain and x_{min} such that

$$f(x) \geq f(x_{min})$$

for any x from the domain of f. Such minimum or maximum may or may not exist and if it exists, it may not be unique. We know, by Theorem 8, that any continuous function on a closed interval $[a, b]$ attains its extrema. In order to find the absolute extrema the following notion of relative extrema is very useful.

A function f is said to have a relative (or local) maximum at a point $c \in \mathbb{R}$ if there is a neighbourhood (a, b), with $c \in (a, b)$ such that $f(c)$ is the maximum of f on (a, b), i.e., $f(c) \geq f(x)$, $x \in (a, b)$.
Similarly, a function f is said to have a relative or local minimum at $c \in \mathbb{R}$ if there is a neighbourhood (a, b), with $c \in (a, b)$, such that $f(c)$ is the minimum of f on (a, b), i.e., $f(c) \leq f(x)$, $x \in (a, b)$.
If $f(x)$ has either a relative maximum or relative minimum at $x = c$ then f is said to have a relative extremum at $x = c$.



Relative extrema are points where the function is changing from increasing to decreasing or vice versa. This is why these points are often referred to as turning points. If the function is differentiable at the extremum $x = c$ then we must have $f'(c) = 0$. Indeed, if f has a relative maximum $f(c)$ at $x = c$ then there is a neighbourhood $(a, b) \ni c$ such that

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

for $x \in (a, b)$. Therefore,

$$\frac{f(x) - f(c)}{x - c} \geq 0$$

for $x < c$ and

$$\frac{f(x) - f(c)}{x - c} \leq 0$$

for $x > c$. If we pass x to c from the left, i.e., for $x < c$ we conclude that $f'(c) \geq 0$, whereas passing x to c from the right yields the conclusion $f'(c) \leq 0$. This is only possible if $f'(c) = 0$, as claimed. Geometrically, this means that the graph of f has a horizontal tangent at its local extremum, provided the function is differentiable at c. This observation allows us to look for local extrema by solving the algebraic equation

$$f'(x) = 0. \tag{9}$$

The solutions of equation (9) are called *stationary points*.

Of course, the other possibility is that the function is not differentiable at the extremum. We combine stationary points and points of non-differentiability into the notion of *critical points*.

♦Example. Find the critical points of the following functions and indicate if the point is stationary or not.

(a) $x^2 - 4x + 1$ (b) $\sqrt{1 - x^2}$ (c) $x\,\mathrm{e}^{-x}$ (d) $|x|$ (e) $x \ln x$ (f) x^4.

Solution (a) $f(x) = x^2 - 4x + 1$ is differentiable throughout $\mathbb{R}$ and we have $f'(x) = 2(x-2)$. So there is just one critical point, the stationary point $x = 2$.

(b) $f(x) = \sqrt{1-x^2}$ is defined and continuous on $[-1, 1]$, it is differentiable only on $(-1, 1)$ as $f'(x) = \frac{-x}{\sqrt{1-x^2}}$. So f is nondifferentiable at the two points $x = \pm 1$ of its domain. At the points where f is differentiable, i.e., on $(-1, 1)$, we have $f'(x) = 0$ at $x = 0$. So altogether we have three critical points, $x = -1, 0, +1$, one of which is a stationary point, $x = 0$.

(c) $f(x) = x\,\mathrm{e}^{-x}$ is differentiable throughout $\mathbb{R}$, $f'(x) = \mathrm{e}^{-x}(1-x)$. So the only critical points are stationary points. We find that $f'(x) = 0$ when $x = 1$. There is one critical point, the stationary point $x = 1$.

(d) $f(x) = |x|$ is continuous throughout $\mathbb{R}$ and differentiable on $(-\infty, 0) \cup (0, \infty)$ (i.e. everywhere except $x = 0$). On $(0, \infty)$, $|x| = x$ so $f'(x) = 1$, there can be no critical points on $(0, \infty)$. Similarly, on $(-\infty, 0)$, $f(x) = |x| = -x$ so $f'(x) = -1$, there can be no critical points on $(-\infty, 0)$. So there is just one critical point, the point of nondifferentiability $x = 0$.

(e) $f(x) = x \ln x$ is defined and continuous on $(0, \infty)$. We have

$$f'(x) = 1 + \ln x \ ,$$

f is differentiable on $(0, \infty)$. So $f(x)$ is differentiable at all points of its domain, then the only critical points are stationary points. We have that

$$f'(x) = 1 + \ln x = 0 \quad \text{when} \quad x = \mathrm{e}^{-1} = \frac{1}{\mathrm{e}}.$$

So the only critical point is the stationary point $x = \frac{1}{\mathrm{e}}$.

(f) $f(x) = x^4$ is differentiable throughout $\mathbb{R}$. The only critical points are stationary points. We have

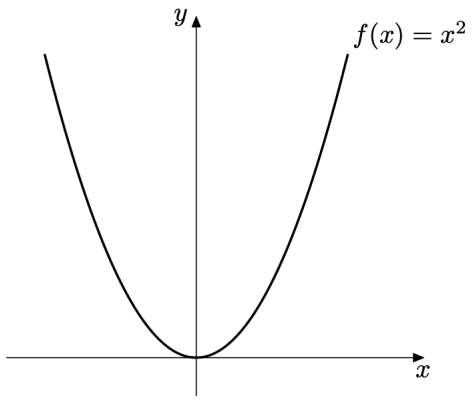
$$f'(x) = 4x^3 = 0 \quad \text{when} \quad x = 0.$$

So there is just one critical point, the stationary point $x = 0$.

We have:

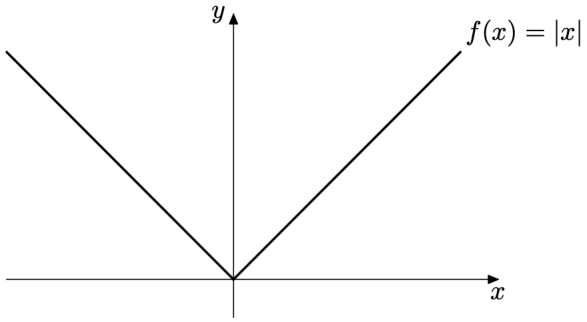
Theorem 18. *If $x = c$ is a relative extremum of a function f then $x = c$ must be a critical point of f.*

♦Example. (a)



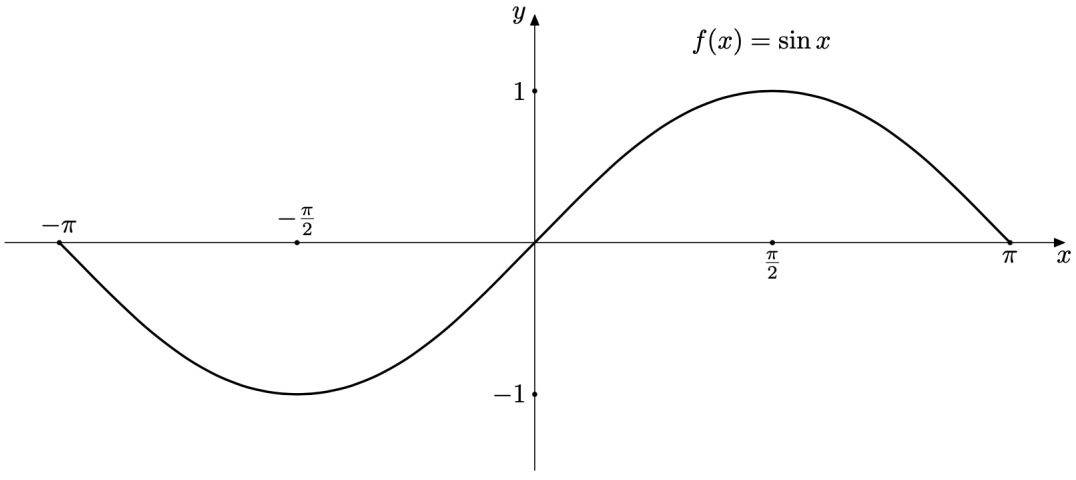
f has a relative minimum at $x = 0$, a stationary point.

(b)



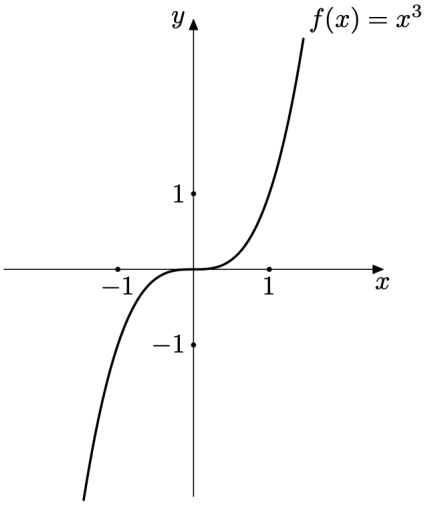
f has a relative minimum at the point of non differentiability $x = 0$.

(c)



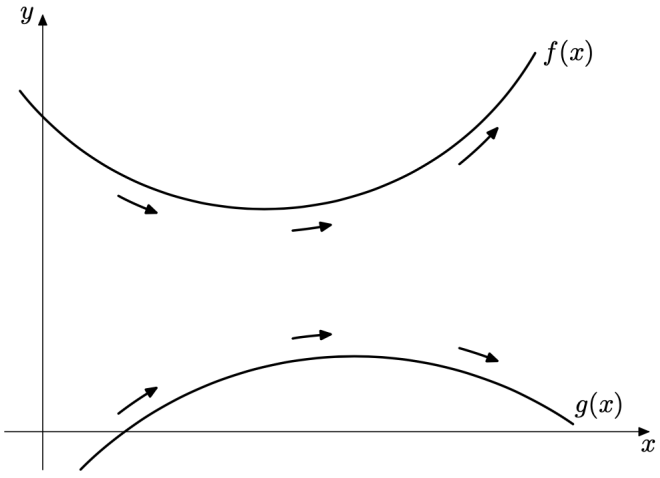
$f(x) = \sin x,\ x \in [-\pi, \pi]$ has a relative minimum at $x = -\frac{\pi}{2}$ (a stationary point) and a relative maximum at $x = \frac{\pi}{2}$ (a stationary point).

(d)



$f(x) = x^3$, no relative extrema. Note however, that we do have $f'(0) = 0$.

Notice in all those cases where f is differentiable in a neighbourhood of the extremum, the derivative $f'(x)$ changes sign when passing the extremum.



This observation is generalised in the next theorem, which gives us a sufficient criterion fir the point being an extremum. This theorem is called the *First derivative test.*

Theorem 19 (First derivative test)**.** *Let f be continuous at a critical point $x = c$, suppose also that f is differentiable in some neighbourhood $(a, c) \cup (c, b)$, $a < b$.*

(a) If, on some neighbourhood (a, b) of c, we have $f'(x) \geq 0$ for $x \in (a, c)$ and $f'(x) \leq 0$ for $x \in (c, b)$ then f has a relative maximum at $x = c$.

(b) If, on some neighbourhood (a, b) of c, we have $f'(x) \leq 0$ for $x \in (a, c)$ and $f'(x) \geq 0$ for $x \in (c, b)$ then f has a relative minimum at $x = c$.

Proof. (a) We have $f'(x) \geq 0$ on (a, c) so f is increasing on this interval. So if $x \in (a, c)$ we have $x < c$ and therefore $f(x) \leq f(c)$. In a similar fashion, on (c, b), if $f'(x) \leq 0$ then f is decreasing; we have for $x \in (c, b)$ that $x > c$ so $f(x) \leq f(c)$. Hence, for any $x \in (a, b)$ we have $f(c) \geq f(x)$, showing that $x = c$ is a local maximum.

(b) can be proved in the same way. □

♦Example. Find the relative maxima and minima for the following functions.

(a) $x^3 - 3x^2 + 3x$ (b) x^4 (c) $x \ln x$ (d) $x\,\mathrm{e}^{-x}$.

Solution (a) $f(x) = x^3 - 3x^2 + 3x$ is differentiable on all of $\mathbb{R}$, so the only critical points are stationary points. We have $f'(x) = 3(x^2 - 2x + 1)$, therefore $f'(x) = 0$, when $x = 1$. Now we can write $f'(x) = 3(x - 1)^2$, so $f'(x)$ does not change sign as we pass the critical point $x = 1$. The function does not have any local maxima or minima.

(b) $f(x) = x^4$ is differentiable throughout $\mathbb{R}$, so extrema will occur at stationary points. We have

$$f'(x) = 4x^3 = 0, \text{ when } x = 0.$$

For $x < 0$, $f'(x) = 4x^3 < 0$ and, for $x > 0$, we have $f'(x) = 4x^3 > 0$. So, by our first derivative test, f has a relative minimum at $x = 0$.

(c) $f(x) = x \ln x$ is defined on $(0, \infty)$ and differentiable on $(0, \infty)$. We have

$$f'(x) = \ln x + 1 = 0$$

when $x = \frac{1}{\mathrm{e}}$. Moreover,

$$f'(x) < 0$$

for $x < \frac{1}{\mathrm{e}}$ and

$$f'(x) > 0$$

for $x > \frac{1}{\mathrm{e}}$. Therefore, f has a relative minimum at $x = \frac{1}{\mathrm{e}}$.

(d) $f(x) = x\,\mathrm{e}^{-x}$ is differentiable throughout $\mathbb{R}$. We have

$$f'(x) = \mathrm{e}^{-x}(1 - x) = 0$$

when $x = 1$. Moreover,

$$f'(x) > 0$$

for $x < 1$ and

$$f'(x) < 0$$

for $x > 1$. Therefore, f has a relative maximum at $x = 1$.

Testing whether or not $f'(x)$ changes sign can at times be quite cumbersome. There is another test for extrema, the second derivative test, which in many cases is easy to implement. The second derivative test combines our understanding of concavity with the results of this lecture to produce a simple test.

Theorem 20 (Second Derivative Test)**.** *Assume that f be differentiable in some neighbourhood of c and twice differentiable at $x = c$.*

(a) If $f'(c) = 0$ and $f''(c) > 0$ then $x = c$ is a relative minimum for f.

(b) If $f'(c) = 0$ and $f''(c) < 0$ then $x = c$ is a relative maximum for f.

(c) If $f'(c) = 0$ and $f''(c) = 0$ then $x = c$ may be a maximum or a minimum or neither of them.

Proof. We prove (b) leaving (a) and (c) as exercises.

(b) We have $f'(c) = 0$ and $f''(c) < 0$. We can choose a small neighbourhood of c, say (a, b), such that $\frac{f'(x)-f'(c)}{x-c} < 0$ for all $x \in (a, b)$. This means that the function $f'(x)$ is strictly positive on (a, c) and hence f is strictly increasing. Similarly, for $x \in (c, b)$ we have $x > c$ and so $f'(x)$ is strictly negative and hence f becomes strictly decreasing. We have shown that $x = c$ is a local maximum for f. □

Despite its popularity, the second derivative test has two major drawbacks: 1. As will be seen in the examples below, the second derivative test may be inconclusive. 2. Computing the second derivative, e.g., for rational functions, can be very tedious.

♦Example. Repeat the previous example using the second derivative test.

Solution

(a)

$$\begin{aligned} f(x) &= x^3 - 3x^2 + 3x \\ f'(x) &= 3x^2 - 6x + 3 \\ f''(x) &= 6x - 6 \end{aligned}$$

Notice f is twice differentiable throughout $\mathbb{R}$. There was one stationary point at $x = 1$, we have $f''(1) = 6 \times 1 - 6 = 0$. The test is inconclusive, we cannot avoid looking at the sign of $f'(x)$ in this case.

(b) $f(x) = x^4$. Here $f'(x) = 4x^3$ and $f''(x) = 12x^2$. The only stationary point is $x = 0$ and $f''(0) = 0$, so again the second derivative test is inconclusive.

(c) $f(x) = x \ln x$. Here $f'(x) = 1 + \ln x$ and $f''(x) = \frac{1}{x}$. From which we see that f is twice differentiable on $(0, \infty)$. We found one stationary point at $x = \frac{1}{\mathrm{e}}$; we have $f''(\frac{1}{\mathrm{e}}) = \mathrm{e} > 0$. Which confirms that f has a local minimum at $x = \frac{1}{\mathrm{e}}$.

(d)

$$\begin{aligned} f(x) &= x\,\mathrm{e}^{-x} \text{ so that} \\ f'(x) &= \mathrm{e}^{-x}(1 - x) \text{ and} \\ f''(x) &= -\,\mathrm{e}^{-x}(2 - x) \end{aligned}$$

The function is twice differentiable throughout $\mathbb{R}$. We found one stationary point at $x = 1$, and

$$f''(1) = -\,\mathrm{e}^{-1} < 0,$$

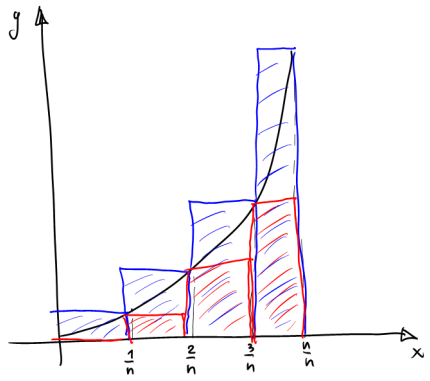
which verifies the local maximum at $x = 1$. □

16 Integration

The concept of differentiation was motivated by finding the instantaneous rate of change of some process. It involves the function and a point from the domain. The motivation of the concept of integration is the converse: we want to compute the accumulation of some quantity with given rate of change over some interval from the domain. This involves a function and an interval. Examples are the accumulation of distance travelled with given velocity function over an interval of time or the area accumulated under the graph of a function over some interval on the x-axis. We start with an example of the latter kind.

♦Example. Find the area A enclosed by the x-axis, the parabola $y = f(x) = x^2$ and the vertical line $x = 1$.

Solution. We sandwich the region under the parabola between two simpler figures, so that the area A of that region can be estimated from above and below. See the sketch below.



The region under the parabola is included into the blue staircase figure that consists of n rectangular bars of basis length $\frac{1}{n}$ and height $f(\frac{k}{n})$, where k is the number of the bar. Notice that $f(\frac{k}{n})$ is the maximum of the function on the interval $[\frac{k-1}{n}, \frac{k}{n}]$ since the function f is increasing. The area $\bar{A}_n$ of this staircase figure is $\geq A$ and equals

$$\bar{A}_n = \sum_{k=1}^{n} \frac{1}{n} \left(\frac{k}{n}\right)^2 = \frac{1}{n^3} \sum_{k=1}^{n} k^2.$$

The red staircase figure is included into the region under the parabola and therefore its area $\underline{A}_n \leq A$. The bars have also basis length $\frac{1}{n}$ but height $f(\frac{k-1}{n})$, which is the minimum of the function on the interval $[\frac{k-1}{n}, \frac{k}{n}]$. We have

$$\underline{A}_n = \sum_{k=1}^{n} \frac{1}{n} \left(\frac{k-1}{n}\right)^2 = \frac{1}{n^3} \sum_{k=1}^{n-1} k^2.$$

The sum $\sum_{k=1}^{n} k^2$ is the partial sum of the sequence of squares. The formula

$$s_n = \sum_{k=1}^{n} k^2 = \frac{n}{6}(2n+1)(n+1)$$

can be verified by induction, but can also be derived as discussed earlier in this unit: We have the derived sequences

$$\begin{aligned} a_n = n^3 \quad &\Longrightarrow \quad a'_n = a_{n+1} - a_n = 3n^2 + 3n + 1 \\ b_n = n^2 \quad &\Longrightarrow \quad b'_n = b_{n+1} - b_n = 2n + 1 \\ c_n = n \quad &\Longrightarrow \quad c'_n = c_{n+1} - c_n = 1 \end{aligned}$$

Combining these together gives

$$s_{n-1} = \frac{1}{3}(a_n - \frac{3}{2}b_n + \frac{1}{2}c_n) = \frac{n}{6}(2n-1)(n-1)$$

Hence

$$s_n = \frac{n}{6}(2n+1)(n+1).$$

Since $s_1 = 1 = 1^2$, no additional constant needs to be added.

Now

$$\bar{A}_n = \frac{1}{6n^2}(2n+1)(n+1), \quad \underline{A}_n = \frac{1}{6n^2}(2n-1)(n-1).$$

We have

$$\underline{A}_n \leq A \leq \bar{A}_n$$

and

$$\lim_{n\to\infty} \underline{A}_n = \lim_{n\to\infty} \bar{A}_n = \frac{1}{3}.$$

By the squeeze theorem if follows $A = \frac{1}{3}$. This method had already been used by ancient Greek mathematician Archimedes. In modern notation we write

$$\int_0^1 x^2 dx = \frac{1}{3}.$$

This is the (definite Riemann) integral of x^2 from 0 to 1.

We can use the same method to find the area $A(b)$ enclosed by the parabola $y = x^2$, the x-axis and the vertical line $x = b$. In this case

$$\begin{aligned} \bar{A}_n &= \frac{b}{n}\sum_{k=1}^{n}\left(\frac{bk}{n}\right)^2 = \frac{b^3}{n}\sum_{k=1}^{n}\left(\frac{k}{n}\right)^2 = \frac{b^3}{6n^2}(2n+1)(n+1) \\ \underline{A}_n &= \frac{b}{n}\sum_{k=1}^{n}\left(\frac{b(k-1)}{n}\right)^2 = \frac{b^3}{n}\sum_{k=1}^{n-1}\left(\frac{k}{n}\right)^2 = \frac{b^3}{6n^2}(2n-1)(n-1) \end{aligned}$$

We get

$$A(b) = \frac{b^3}{3}.$$

We write

$$\int_0^b x^2 dx = \frac{b^3}{3}.$$

This is the integral of x^2 from 0 to b with variable upper limit b. We can interpret this as a new function $F(b)$ that tells us how the area accumulates under the parabola as we travel from 0 to b. The value of the integrand $y = x^2$ gives the rate at which this accumulation occurs.

♠ *Exercises 40.* Apply the same procedure as above to the function $f(x) = x$. Compare the function $F(b)$ with the elementary formula for the area of a triangle enclosed by $y = x$, the x-axis and the vertical line $x = b$.

We generalise this approach now to arbitrary bounded functions $f\colon [a, b] \to \mathbb{R}$. Let $x_k = a + \frac{k}{n}(b - a)$. Then

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b \tag{10}$$

is an equidistant partition of the interval $[a, b]$. We call

$$\underline{A}_n = \sum_{k=1}^{n} \frac{b-a}{n} \inf_{x \in [x_{k-1}, x_k]} f(x)$$

the lower *Darboux sum* for $f(x)$ subject to the partition (10). If f is non-negative then $\underline{A}_n$ is a lower bound of the area enclosed by the graph of $f(x)$, the x-axis and the vertical lines $x = a$ and $x = b$.

Similarly, we define the upper Darboux sums:

$$\bar{A}_n = \sum_{k=1}^{n} \frac{b-a}{n} \sup_{x \in [x_{k-1}, x_k]} f(x),$$

which give an upper bound of the area enclosed by the graph of $f(x)$, the x-axis and the vertical lines $x = a$ and $x = b$.

If the limits

$$\lim_{n\to\infty} \underline{A}_n \text{ and } \lim_{n\to\infty} \bar{A}_n$$

both exist and are equal then this common limit is called the *definite Riemann integral* of $f(x)$ from a to b, denoted by

$$\int_a^b f(x)\, dx.$$

This integral assigns to an interval $[a, b]$ and an integrand $f(x)$ a number.

The integral always exists if the integrand is continuous on the interval of integration $[a, b]$. We outline a sketch of the proof, which relies on a result to be established in Pmth331 in year 3.

Theorem 21. *If $f\colon [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ then it is uniformly continuous, that is*

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } \forall x, x' \in [a, b] \text{ with } |x - x'| < \delta, |f(x) - f(x')| < \varepsilon.$$

Here the constant δ depends only on ε but not on x' as in the case of common continuity at x'.

Theorem 22. *If $f\colon [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ then the definite integral*

$$\int_a^b f(x)\, dx$$

exists.

Proof. We have two sequences of Darboux sums, namely $\underline{A}_n$ and $\bar{A}_n$. Clearly,

$$(b-a) \inf_{x\in[a,b]} f(x) \le \underline{A}_n \le \bar{A}_n \le (b-a) \sup_{x\in[a,b]} f(x).$$

We show that $\bar{A}_n - \underline{A}_n \to 0$ as $n \to \infty$. Indeed,

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \text{ such that } \forall x, x' \in [a, b] \text{ with } |x - x'| < \delta, |f(x) - f(x')| < \varepsilon.$$

If we choose $n > \frac{b-a}{\delta}$ then

$$\sup_{[x_{k-1},x_k]} f(x) - \inf_{[x_{k-1},x_k]} f(x) \le \varepsilon$$

and hence

$$\bar{A}_n - \underline{A}_n = \frac{b-a}{n} \sum_{k=1}^{n} \Big(\sup_{x\in[x_{k-1},x_k]} f(x) - \inf_{x\in[x_{k-1},x_k]} f(x) \Big) \le \varepsilon(b-a).$$

If the sequence $\underline{A}_n$ was increasing and $\bar{A}_n$ was decreasing we would have a contracting family of nested intervals, which would contract to the common limit. Unfortunately, this is not the case. However, a little trick helps. Consider the subsequences $\underline{A}_{2^n}$ and $\bar{A}_{2^n}$. Then each consecutive partition is obtained from the previous partition by adding the mid points. That is, the interval $[x_{k-1}, x_k]$ becomes bisected into $[x_{k-1}, \frac{x_k - x_{k-1}}{2}]$ and $[\frac{x_k - x_{k-1}}{2}, x_k]$. Now

$$\inf_{x\in[x_{k-1},\frac{x_k-x_{k-1}}{2}]} f(x) \ge \inf_{x\in[x_{k-1},x_k]} f(x), \qquad \inf_{x\in[\frac{x_k-x_{k-1}}{2},x_k]} f(x) \ge \inf_{x\in[x_{k-1},x_k]} f(x).$$

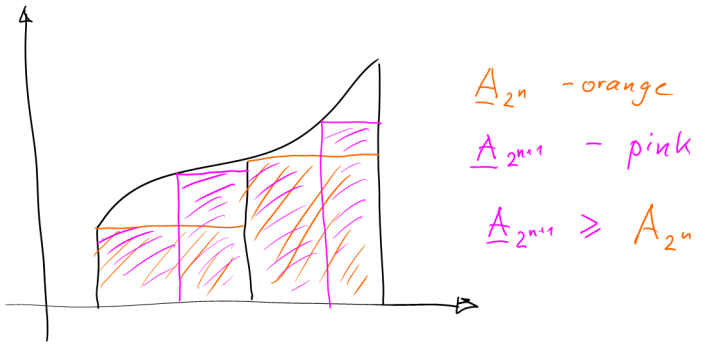
and therefore

$$\inf_{x\in[x_{k-1},\frac{x_k-x_{k-1}}{2}]} f(x)\frac{b-a}{2^{n+1}} + \inf_{x\in[\frac{x_k-x_{k-1}}{2},x_k]} f(x)\frac{b-a}{2^{n+1}} \geq \inf_{x\in[x_{k-1},x_k]} f(x)\frac{b-a}{2^n}$$

hence

$$\underline{A}_{2^{n+1}} \geq \underline{A}_{2^n}.$$

This is illustrated in the picture below.



Analogously,

$$\bar{A}_{2^{n+1}} \leq \bar{A}_{2^n}.$$

The closed intervals $[\underline{A}_{2^n}, \bar{A}_{2^n}]$ are nested and contract to the common limit of the ends, which is

$$\int_a^b f(x)\,dx, \quad \square$$

♠ *Exercises 41.* Give an example of a function for which $\overline{A_3} > \overline{A_2}$. Hint. Consider the function

$$f(x) = \begin{cases} 0 & \text{for } x \in [0, \frac{1}{2}) \\ 1 & \text{for } x \in [\frac{1}{2}, 1]. \end{cases}$$

One can prove that for any continuous function $f\colon [a,b] \to \mathbb{R}$ and any sequence of partitions

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$

the so-called *Riemann sums*

$$\sum_{k=1}^{n} f(x_k^*)(x_k - x_{k-1}) = \sum_{k=1}^{n} f(x_k^*)\Delta x_k,$$

where $x_k^* \in [x_{k-1}, x_k]$ is any intermediate point and $\Delta x_k = x_k - x_{k-1}$, tend to the integral

$$\int_a^b f(x)\,dx,$$

as long as the maximal length of the partition intervals $\max_{k=1,\dots,n} \Delta x_k$ tends to 0.

♠ *Exercises 42.* Compute the integral from a to b of a constant function $f(x) = k$.

♦Example. Show that the Dirichlet function does not have a defined Riemann integral from 0 to 1.

Solution. For any partition $0 = x_0 < x_1 < \cdots < x_n = 1$ we have

$$\underline{A}_n = \sum_{k=1}^{n} 0 \cdot \Delta x_k = 0,$$

since between any $x_{k-1} < x_k$ there is an irrational number. Therefore,

$$\lim_{n\to\infty} \underline{A}_n = 0.$$

On the other hand

$$\bar{A}_n = \sum_{k=1}^{n} 1 \cdot \Delta x_k = 1,$$

since between any $x_{k-1} < x_k$ there is a rational number. Therefore,

$$\lim_{n\to\infty} \bar{A}_n = 1.$$

Since the limits are different, the definite integral does not exist.

17 Properties of the definite integral

We derive some properties of the definite integral.

1. Linearity of the integral. If $f(x)$ and $g(x)$ are continuous on $[a, b]$ and c is any constant, then

(i) $\displaystyle\int_a^b [f(x) + g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx,$

(ii) $\displaystyle\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx.$

The plausibility of (i) can be seen by considering a Riemann sum for $f(x) + g(x)$ over $[a, b]$. We get $\displaystyle\sum_{k=1}^{n} [f(x_k^*) + g(x_k^*)]\,\Delta x_k$ which can be split up as

$$\sum_{k=1}^{n} [f(x_k^*) + g(x_k^*)]\,\Delta x_k = \sum_{k=1}^{n} f(x_k^*)\Delta x_k + \sum_{k=1}^{n} g(x_k^*)\Delta x_k.$$

Taking limits, the first sum tends to $\int_a^b [f(x) + g(x)]dx$ and the others tend to $\int_a^b f(x)\,dx$ and $\int_a^b g(x)\,dx$, respectively.

The argument for (ii) is similar.

2. Comparison principle.

(i) If $f(x) \le g(x)$ for $a \le x \le b$ then

$$\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx.$$

For any partition the upper Darboux sum for f will be less than or equal to the corresponding upper Darboux sum for g. By passing to the limit we get the desired relation for the integral.

(ii) If $m \le f(x) \le M$ for all x in $[a, b]$, then

$$m(b - a) \le \int_a^b f(x)\,dx \le M(b - a).$$

This follows from (i) because if $m \le f(x) \le M$ then

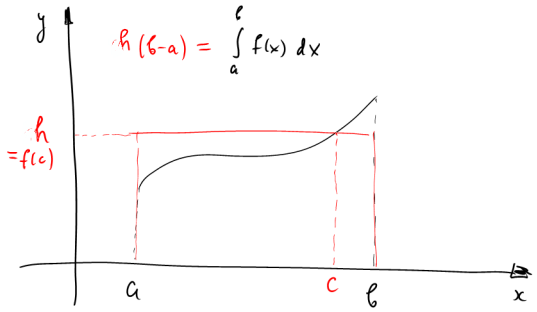
$$\int_a^b m\,dx \le \int_a^b f(x)\,dx \le \int_a^b M\,dx$$

i.e. $m(b - a) \le \int_a^b f(x)\,dx \le M(b - a)$.

(iii) The following *Mean Value Theorem of Integral Calculus* is a consequence of (ii). If f is continuous on $[a,b]$ and m and M are the minimum and maximum of f respectively, then there exists $c \in [a,b]$ such that

$$\int_a^b f(x)\,dx = f(c)(b-a).$$

This follows from the intermediate value theorem for the continuous function $(b-a)f$. For non-negative functions the Mean Value theorem has the following geometric interpretation: The area under the curve of the function equals the area of the rectangle with basis $b-a$ and height $h = f(c)$.



(iv) If a function $f(x)\colon [a,b] \to \mathbb{R}$ is integrable then $|f(x)|$ is integrable as well and

$$\left|\int_a^b f(x)\,dx\right| \leq \int_a^b |f(x)|\,dx.$$

If f is bounded and has only finitely many discontinuities then this is also true for $|f|$ and then $|f|$ is integrable. It would be a bit more involved to show that $|f|$ is integrable just assuming that f is. The inequality follows now from

$$-|f(x)| \leq f(x) \leq |f(x)|$$

and hence

$$-\int_a^b |f(x)| \leq \int_a^b f(x) \leq \int_a^b |f(x)|$$

that is

$$\left|\int_a^b f(x)\,dx\right| \leq \int_a^b |f(x)|\,dx.$$

(v) In applications we often work with approximate data, i.e., instead of a function f we work with a function g that is close to f in the sense that $|f(x)-g(x)| \leq \delta$ for any $x \in [a,b]$. It follows then from (ii) and (iv) that the definite integrals of f and g are close to each other. More precisely,

$$\left|\int_a^b f(x)dx - \int_a^b g(x)dx\right| = \left|\int_a^b f(x) - g(x)\,dx\right|$$
$$\leq \int_a^b |f(x)-g(x)|dx \leq \delta(b-a).$$

Notice that this means continuity of the application of the definite integral (as a function) on an integrable function (as the argument). Indeed, we can make

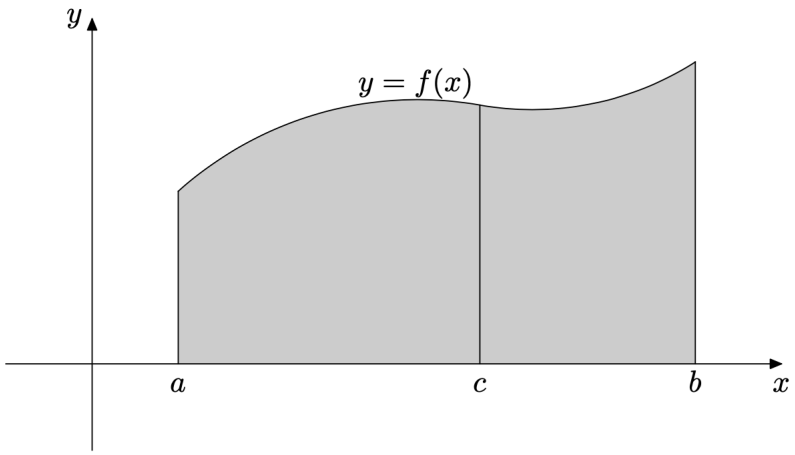
$$\left| \int_a^b f(x)dx - \int_a^b g(x)dx \right| < \varepsilon$$

by choosing f and g such that they are δ-close, where $\delta = \frac{\varepsilon}{b-a}$. This is in contrast to differentiation: The derivatives of f and g need not be close no matter how close f and g are to each other. We will return to this topic in MTHS130.

3. Splitting of the interval of integration

If $f(x)$ is continuous on an interval containing a, b, c then

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx \tag{11}$$



In the diagram, for $a < c < b$, the total area from a to b is the sum of two areas: the area from a to c and the area from c to b.

We define

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx.$$

Then the relation 11 is true even if c does not lie between a and b. In particular, also

$$\int_a^a f(x)\,dx = -\int_a^a f(x)\,dx = 0.$$

It is another consequence of splitting the interval of integration that a function $f\colon [a,b] \to \mathbb{R}$ is also integrable if it has finitely many points of discontinuity, as long

as it is bounded (say $|f| \leq K$). We just need to split the interval into subintervals so that the common length of the subintervals that contain discontinuities is smaller than $\frac{\varepsilon}{2K}$. Then the contribution of those subintervals to the definite integral can be made smaller than ε and therefore does not contribute to the integral.

Miscellaneous Worked Examples.

(i) Which, if any, of the expressions

$$\int_1^3 \frac{x^2}{1+x}\,dx,\ \int_1^3 \frac{z^2}{1+z}\,dz,\ \int_1^3 \frac{u^2}{1+u}\,du$$

are equal?

Answer: They are all equal since they differ only in the dummy variables.

(ii) True or false: $\displaystyle\int_1^2 xf(x)\,dx = x\int_1^2 f(x)\,dx.$

Answer: False.

(iii) True or false: $\displaystyle\int_1^2 5f(x)\,dx = 5\int_1^2 f(x)\,dx.$

Answer: True. (A constant can be moved out to the front of an integral but not a function.)

(iv) True or false:

(a) $\displaystyle\int_a^b [x+f(x)]\,dx = \int_a^b x\,dx + \int_a^b f(x)\,dx.$

(b) $\displaystyle\int_a^b xf(x)\,dx = \int_a^b x\,dx \int_a^b f(x)\,dx.$

(c) $\displaystyle\int_a^b \alpha f(x)\,dx = \int_a^b \alpha\,dx \int_a^b f(x)\,dx.$

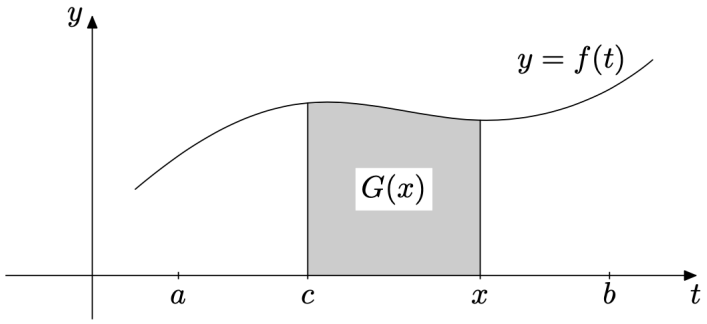
(d) $\displaystyle\int_0^1 \left(\sum_{k=1}^n x^k\right) dx = \sum_{k=1}^n \int_0^1 x^k dx.$

Answers: True, False, False, True.

18 The Fundamental Theorem of Calculus

The computation of integrals as limits of Riemann sums is very tedious and usually requires a formula for the partial sums, which may not be readily available. The Fundamental Theorem of Calculus relates integration and differentiation as inverse operations and gives us a convenient tool to compute integrals by algebraic rules that are based on inverting differentiation. Although methods of integration and differentiation were already known in ancient India and Greece, the relation between them was only discovered by Newton and Leibniz in the 17th century.

In the following discussion we will consider definite integrals as functions of the upper limit of integration. Therefore we want to reserve the variable x for the upper limit of integration, so we will use another letter t for the "dummy" variable within the integral.



Consider the function

$$G(x) = \int_c^x f(t)\,dt,$$

where $f(t)$ is continuous on $[a, b]$ and c, and x are in $[a, b]$. Geometrically, $G(x)$ is the area under the curve $y = f(t)$, above the t-axis and between c and x. We think of c as fixed and of x as moving. Thus $G(x)$ describes the process of accumulation of area under the curve as x changes. Intuitively, the function f gives the rate of change of this accumulation. We make this intuition precise by calculating the derivative of the function $G(x)$. Let $x + h$ also lie in $[a, b]$. Then

$$\begin{aligned} G(x+h) - G(x) &= \int_c^{x+h} f(t)\,dt - \int_c^x f(t)\,dt \\ &= \int_x^{x+h} f(t)\,dt, \qquad \text{using the splitting property,} \\ &= f(x^*)[(x+h) - x] = f(x^*)h, \end{aligned}$$

by the MVT for integrals, where x^* is between x and $x+h$. Hence

$$\frac{G(x+h)-G(x)}{h} = f(x^*), \qquad h \neq 0.$$

We now take the limit as $h \longrightarrow 0$. The number x^* tends to x as $h \longrightarrow 0$ and, by the continuity of f, the number $f(x^*)$ tends to $f(x)$ as $h \longrightarrow 0$. Thus

$$G'(x) = \lim_{h\to 0} \frac{G(x+h)-G(x)}{h} = f(x).$$

We have just proved the first part of the **Fundamental Theorem of Calculus** (FTC):

Theorem 23. *The function $G(x) = \int_c^x f(t)dt$ is differentiable (in particular continuous) on (a,b) and*

$$\boxed{\frac{d}{dx}\int_c^x f(t)\,dt = f(x).}$$

For the second part of the Fundamental Theorem of Calculus we need the notion of a primitive of a function. Any function $F(x)$, such that $F'(x) = f(x)$ is called a *primitive* or *antiderivative* of $f(x)$. Thus

$$G(x) = \int_c^x f(t)dt$$

is an antiderivative of $f(x)$.

If $F(x)$ is an antiderivative of f then, for any constant C, $F(x)+C$ is also an antiderivative of f. The converse is also true if the domain of f is an interval. Indeed, let F_1 and F_2 be two antiderivatives of f. Then

$$\frac{d}{dx}(F_1 - F_2) = f - f = 0.$$

By Theorem 14 part (c) then $F_1 - F_2$ is a constant C, thus $F_1 = F_2 + C$. Notice that, if the domain of f is not connected (i.e., it consists of two or more intervals) then one can add different constants at different intervals.

The *set of all* antiderivatives of a function f is called the indefinite integral of f. We write

$$\int f(x)\,dx.$$

♦Example.

$$\int x^2\,dx = \frac{1}{3}x^3 + C$$

because

$$\frac{d}{dx}\left(\frac{1}{3}x^3 + C\right) = x^2.$$

We can now formulate the second part of the Fundamental Theorem of Calculus, which is our main tool for computing definite integrals.

Theorem 24. *Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function and let $F(x)$ be any antiderivative of f. Then*

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

Instead of $F(b) - F(a)$ we also write

$$[F(x)]_a^b \text{ or } F(x)\big|_a^b.$$

Proof. Let $F(x)$ be an arbitrary antiderivative of f. Then

$$F(x) = G(x) + C$$

where

$$G(x) = \int_a^x f(t)\,dt$$

and C is some constant. Then

$$\int_a^b f(x)\,dx = G(b) - G(a) = G(b) + C - (G(a) + C) = F(b) - F(a). \quad \square$$

♦Example. Compute $\int_1^3 e^x dx$.

Since $\frac{d}{dx}e^x = x$, we find

$$\int e^x dx = e^x + C.$$

By the FTC then

$$\int_1^3 e^x dx = e^x\big|_1^3 = e^3 - e.$$

This method becomes more efficient the more antiderivatives we know.

19 Indefinite Integrals

Any known derivative and any rule of differentiation gives us an antiderivative or a rule of (indefinite) integration. Any rule for integration can be verified by differentiating it.

1. Power functions. The rule

$$\frac{d}{dx}x^p = px^{p-1}$$

is equivalent to

$$\frac{d}{dx}\frac{1}{p}x^p = x^{p-1}$$

and, for $n = p - 1$,

$$\frac{d}{dx}\frac{1}{n+1}x^{n+1} = x^n.$$

Therefore,

$$\boxed{\int x^n\,dx = \tfrac{1}{n+1}x^{n+1} + C, \text{ if } n \neq -1.}$$

If n is negative the domain of the integrand is not connected and the constant C can be chosen different in $(-\infty, 0)$ and $(0, \infty)$.

For the case $n = -1$ we recall that

$$\frac{d}{dx}\ln x = \frac{1}{x}.$$

Since the natural domain of $\ln x$ is $(0, \infty)$ this gives the indefinite integral of $f(x) = \frac{1}{x}$ only for positive x. However, by the chain rule

$$\frac{d}{dx}\ln(-x) = -\frac{1}{-x} = \frac{1}{x}.$$

This can be combined into

$$\frac{d}{dx}\ln|x| = \frac{1}{x},$$

for $x \neq 0$, which is the natural domain of $f(x)$. Hence,

$$\boxed{\int \tfrac{1}{x}dx = \ln|x| + C.}$$

2. Trigonometric functions and their inverses. From the table of derivatives we get

$$\int \sin x\,dx = -\cos x + C$$
$$\int \cos x\,dx = \sin x + C$$
$$\int \frac{1}{\cos^2 x}dx = \tan x + C$$
$$\int \frac{1}{\sqrt{1-x^2}}dx = \arcsin x + C$$
$$\int \frac{1}{1+x^2}dx = \arctan x + C.$$

We will add more indefinite trig integrals to this table as more advanced techniques of integration become available.

3. Exponential functions. From $\frac{d}{dx}a^x = a^x \ln a$ we get

$$\boxed{\int a^x\,dx = \tfrac{a^x}{\ln a} + C.}$$

4. Linearity. It follows from the linearity of definite integrals that, for any two integrable functions f and g and any constant k,

$$\int (f+g)\,dx = \int f(x)\,dx + \int g(x)\,dx$$
$$\int kf(x)\,dx = k\int f(x)\,dx$$

5. Substitution rule. There is no simple rule for integrating products of functions. For certain products the chain rule of differentiation gives an integration rule. Chain rule results in products of the form $f(g(x))\cdot g'(x)$, where one factor is a composite function and the other factor is the derivative of the inside function. In this case

$$\boxed{\int f(g(x))g'(x)\,dx = F(g(x)) + C,}$$

where F is an antiderivative of the outside function f.

Since $g'(x)\,dx$ is nothing but the differential of g we can write

$$\int f(g(x))g'(x)\,dx = \int f(g)\,dg,$$

meaning that we just integrate f as a function of the independent variable g.

♦Example. Compute the indefinite integral $\int 2xe^{x^2}\,dx$.

Solution. The integrand is a product of the form $f(g(x))g'(x)$ with $f(y) = e^y$, $g(x) = x^2$ and $g'(x) = 2x$. Therefore,

$$\int 2xe^{x^2}\,dx = e^{x^2} + C.$$

We can verify by differentiation using chain rule

$$\frac{d}{dx}(e^{x^2} + C) = 2xe^{x^2}.$$

Sometimes, a minor modification can transform a product into the required form as in the example below:

♦Example. Compute the indefinite integral $\int x^2 \sin(x^3)\,dx$.

Solution. In this case x^2 is not exactly the derivative of the inside function x^3, but multiplying by 3 fixes this and can be easily compensated by division by 3.

$$\int x^2 \sin(x^3)\,dx = \frac{1}{3}\int 3x^2 \sin(x^3)\,d = -\frac{1}{3}\cos(x^3) + C.$$

Again we verify this by differentiation using chain rule

$$\frac{d}{dx}(-\frac{1}{3}\cos(x^3) + C) = -\frac{1}{3}3x^2(-\sin(x^3)) = x^2 \sin x^3.$$

The following two special cases occur often:

$$\boxed{\int f(ax+b)\,dx = \tfrac{F(ax+b)}{a} + C,}$$

where F is an antiderivative of f and $g(x) = ax + b$ is a linear function.

♦Example. $\int e^{ax}dx = \frac{1}{a}e^{ax} + C$

The other special case is:

$$\boxed{\int \tfrac{g'(x)}{g(x)}dx = \ln|g(x)| + C.}$$

Here the outside function is $f(y) = \frac{1}{y}$ with antiderivative $F(y) = \ln|y|$. This approach works for quotients, where the numerator is the derivative of the denominator.

♦Example. $\int \frac{2x}{x^2-1}dx = \ln|x^2-1| + C$.

This approach also yields the antiderivative of $\tan x = \frac{\sin x}{\cos x}$. Here the derivative of the denominator is the negative of the numerator.

$$\int \frac{\sin x}{\cos x}dx = -\int \frac{-\sin x}{\cos x}dx = -\ln|\cos x| + C.$$

Thus,

$$\boxed{\int \tan x\, dx = -\ln|\cos x| + C.}$$

The substitution rule is a consequence of the same chain rule from differentiation from a slightly different point of view. Sometimes we encounter integration problems where the product structure of the integrand is hidden or the integrand is not a product but an expression that involves some inner function $u = g(x)$. If $g(x)$ is invertible on the integration interval, then $x = h(u)$ and $dx = h'(u)du$ and we can transform the integral in the following way

$$\int f(x)dx = \int f(h(u))h'(u)du.$$

To use this rule:

(i) Make a choice for $x = h(u)$.

(ii) Calculate $dx = \frac{dx}{du}\, du$.

(iii) Convert to a u integral and evaluate.

(iv) Convert back to an expression in x using $u = g(x)$.

This rule is the most sophisticated integration rule. The right choice of substitution is often a matter of trial and error, or luck (or experience). Sometimes substitution rule does not lead to the solution but to another integration problem that can be solved by other methods.

Example. Find $\displaystyle\int \frac{x}{\sqrt{x-1}}\, dx$ by substituting $x = u + 1$.

Solution. Here we expect a simplification by the substitution $x - 1 = u$. Hence $x = u + 1$ and $dx = du$.

$$\begin{aligned}\int \frac{x}{\sqrt{x-1}}\,dx &= \int \frac{u+1}{\sqrt{u}}\,du \\ &= \int (u^{\frac{1}{2}} + u^{-\frac{1}{2}})du \\ &= \frac{2}{3}u^{\frac{3}{2}} + 2u^{\frac{1}{2}} + C \\ &= \frac{2}{3}(x-1)^{\frac{3}{2}} + 2(x-1)^{\frac{1}{2}} + C.\end{aligned}$$

We will learn more methods of integration in MTHS130, namely integration by parts (which is the counterpart of the product rule of differentiation) and integration of rational functions, using the algebraic theory of partial fractions.

Worked Examples.

(i) Find $\displaystyle\int x\sqrt{x+1}\,dx$.

Solution. The substitution $u = \sqrt{x+1}$ will remove the $\sqrt{}$ sign. Since $x = u^2 - 1$ we have $dx = 2u\,du$ and

$$\begin{aligned}\int x\sqrt{x+1}\,dx &= \int (u^2-1)u2u\,du \\ &= 2\int (u^4 - u^2)\,du \\ &= 2\left(\frac{u^5}{5} - \frac{u^3}{3}\right) + C \\ &= \frac{2}{5}(x+1)^{5/2} - \frac{2}{3}(x+1)^{3/2} + C.\end{aligned}$$

(ii) Use the substitution $x = a\sin\theta$ to evaluate $\displaystyle\int_0^{a/2} \frac{dx}{\sqrt{a^2-x^2}}$, $a > 0$.

Solution. Let $x = a\sin\theta$, $dx = a\cos\theta\,d\theta$. Also,

$$\begin{aligned}\sqrt{a^2-x^2} = \sqrt{a^2 - a^2\sin^2\theta} &= a\sqrt{1-\sin^2\theta} \\ &= a\cos\theta.\end{aligned}$$

We also need to change the limits of integration accordingly. When $\theta = 0$, $x = 0$, and when $\theta = \frac{\pi}{6}$, $x = \frac{a}{2}$. Hence,

$$\int_0^{a/2} \frac{dx}{\sqrt{a^2-x^2}} = \int_0^{\pi/6} \frac{a\cos\theta}{a\cos\theta}\,d\theta = \frac{\pi}{6}.$$

Remark. A substitution $x = a\sin\theta$ is worth a try where the integrand has a term $\sqrt{a^2-x^2}$. As in the above working, this term is changed to $\cos\theta$ and the resulting integrand may be easier to evaluate.

(iii) Find $\displaystyle\int \frac{x\,dx}{\sqrt{1+x^2}}$.

Solution: Put $u = 1 + x^2$, $du = 2x\,dx$. Thus $x\,dx = \frac{1}{2}du$, and

$$\int \frac{x\,dx}{\sqrt{1+x^2}} = \int \frac{du}{2\sqrt{u}} = \sqrt{u} + C = \sqrt{1+x^2} + C.$$

(iv) Find $\displaystyle\int_0^4 2x\sqrt{9+x^2}\,dx$.

Solution: Let $u = 9 + x^2$, $du = 2x\,dx$. When $x = 0$, $u = 9$, and when $x = 4$, $u = 25$. Hence

$$\int_0^4 2x\sqrt{9+x^2}\,dx = \int_4^{25} u^{\frac{1}{2}}\,du = \left[\frac{2}{3}u^{3/2}\right]_4^{25} = \frac{2}{3}\left(5^3 - 2^3\right) = \frac{234}{3}.$$

(v) Justify $\displaystyle\int 2x\,\cos(x^2+1)\,dx = \sin(x^2+1) + C$.

Justification: $\displaystyle\frac{d}{dx}\sin(x^2+1) = \cos(x^2+1)\cdot 2x.$

(vi) Find $\displaystyle\int (3x^2+2)(x^3+2x+1)^{1/2}\,dx$.

Solution: Put $u = x^3 + 2x + 1$, $du = (3x^2+2)\,dx$,

$$I = \int u^{1/2}du = \frac{2}{3}u^{3/2} + C = \frac{2}{3}(x^3+2x+1)^{3/2} + C.$$

Justification: $\displaystyle\frac{d}{dx}\frac{2}{3}(x^3+2x+1)^{3/2} = (x^3+2x+1)^{1/2}(3x^2+2).$

(vii) Find $\displaystyle\int_1^2 \frac{dx}{x^2\sqrt{5x^2-4}}$ by making the substitution $x^2 = \frac{1}{u}$. (This example is a bit difficult, but it illustrates the point that once you have hit upon a substitution to try, the methods are the same as in earlier easier examples.)

Solution:

$$x^2 = \frac{1}{u}, \quad u = \frac{1}{x^2},$$
$$x = u^{-1/2}, \quad dx = -\frac{1}{2}u^{-3/2}du.$$

When $x = 1$, $u = 1$, and when $x = 2$, $u = \frac{1}{4}$.

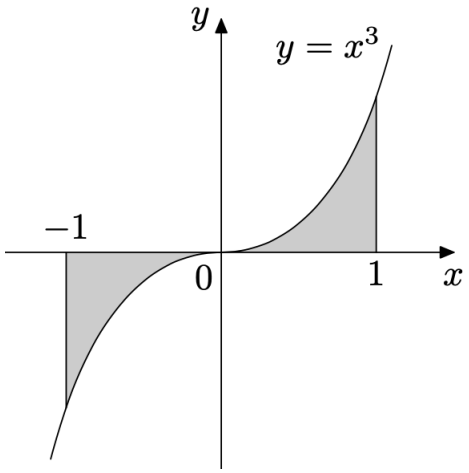
$$\begin{aligned}
\int_1^2 \frac{dx}{x^2\sqrt{5x^2-4}} &= \int_1^{\frac{1}{4}} \frac{-\frac{1}{2}u^{-3/2}du}{\frac{1}{u}\sqrt{\frac{5}{u}-4}} \\
&= -\frac{1}{2}\int_1^{\frac{1}{4}} \frac{du}{\sqrt{5-4u}} \\
&= \left[\frac{1}{4}(5-4u)^{\frac{1}{2}}\right]_1^{\frac{1}{4}} \\
&= \frac{1}{4}\left[\sqrt{4}-1\right] = \frac{1}{4}.
\end{aligned}$$

20 Applications of Integral calculus

1. Areas of Plane Regions

We saw before that $\int_a^b f(x)\,dx$ gives the area under $f(x)$ between $x = a$ and $x = b$ if $f(x) \geq 0$ for $a \leq x \leq b$. If $f(x) \leq 0$ for $a \leq x \leq b$ then the integral is negative and its absolute value gives the area between the curve, the x-axis and the vertical lines $x = a$ and $x = b$.

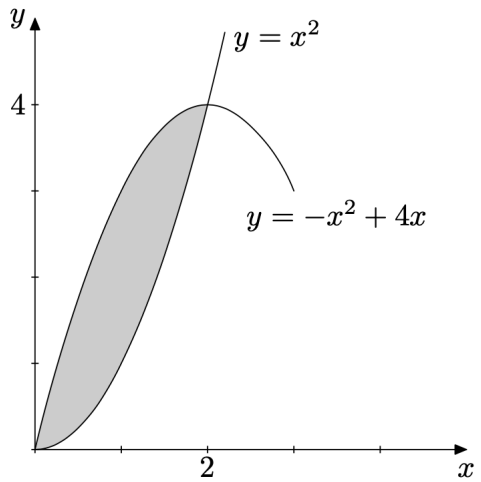
Example: Find the area enclosed by the curve $y = x^3$, the x-axis, and the lines $x = -1$ and $x = 1$.



Solution: Note that $x^3 \leq 0$ for $x \leq 0$, and $x^3 \geq 0$ for $x \geq 0$, so the required area is

$$\begin{aligned} A &= \left| \int_{-1}^{0} x^3 \, dx \right| + \int_0^1 x^3 \, dx \\ &= \left| \left[\frac{x^4}{4} \right]_{-1}^{0} \right| + \left[\frac{x^4}{4} \right]_0^1 \\ &= \left| \frac{-1}{4} \right| + \frac{1}{4} \\ &= \frac{1}{2}. \end{aligned}$$

Example: Find the area bounded by the curves $y = x^2$ and $y = -x^2 + 4x$.



Solution: The curves intersect at $(0, 0)$ and (2,4). The curve $y = -x^2 + 4x$ lies above the curve $y = x^2$ for $0 \leq x \leq 2$.

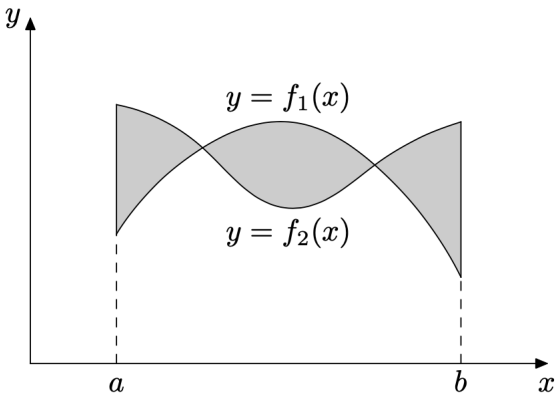
Since $\int_0^2 (-x^2 + 4x)\,dx$ gives the area between the x-axis and $y = -x^2 + 4x$, while $\int_0^2 x^2\,dx$ gives the area between the x-axis and $y = x^2$, the required area is

$$\begin{aligned} A &= \int_0^2 (-x^2 + 4x)\,dx - \int_0^2 x^2\,dx \\ &= \int_0^2 (-2x^2 + 4x)\,dx \\ &= \left[-\frac{2}{3}x^3 + 2x^2\right]_0^2 \\ &= \frac{8}{3}. \end{aligned}$$

In general, if $y = f_1(x)$ and $y = f_2(x)$ intersect at x_1 and x_2, the area enclosed between the curves (between x_1 and x_2) is

$$A = \int_{x_1}^{x_2} |f_1(x) - f_2(x)|\,dx,$$

and this quantity is positive.



More generally, the area enclosed between the curves $y = f_1(x)$, $y = f_2(x)$ and the lines $x = a$ and $x = b$ is

$$A = \int_a^b |f_1(x) - f_2(x)|\,dx.$$

Using this, we offer an alternative solution to the first example...

Example. Find the area enclosed by the curve $y = x^3$, the x-axis, and the lines $x = -1$ and $x = 1$ (refer to diagram on previous page).

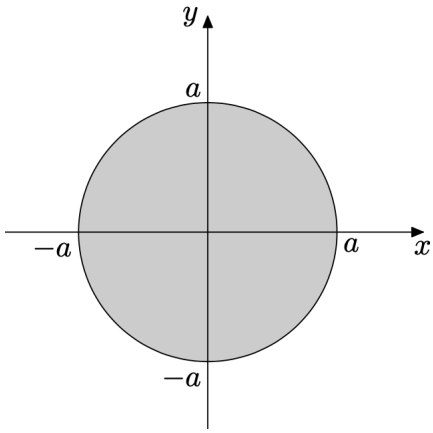
Solution. Putting $f_1(x) = x^3$, $f_2(x) = 0$, $a = -1$ and $b = 1$ in the equation above,

and noting that $|x^3| = -x^3$ when $x \leq 0$, the required area is

$$\begin{aligned} A &= \int_{-1}^{1} \left|x^3\right| \, dx \\ &= \int_{-1}^{0} -x^3 \, dx + \int_{0}^{1} x^3 \, dx \\ &= \left[\frac{-x^4}{4}\right]_{-1}^{0} + \left[\frac{x^4}{4}\right]_{0}^{1} \\ &= \frac{1}{2}. \end{aligned}$$

Example (Area of a Circle). The circle centred at the origin and with radius a has the equation $x^2 + y^2 = a^2$. How do we find its area using integration? The upper half of the circle has the equation $y = \sqrt{a^2 - x^2}$ so the required area is

$$A = 2\int_{-a}^{a} \sqrt{a^2 - x^2}\, dx.$$



The trick here is the substitution $x = a\sin\theta$, $dx = a\cos\theta\, d\theta$. When $\theta = \dfrac{-\pi}{2}$, $x = -a$, and when $\theta = \dfrac{\pi}{2}$, $x = a$. Hence

$$\begin{aligned} A &= 2\int_{-\pi/2}^{\pi/2} \sqrt{a^2 - a^2\sin^2\theta}\, a\cos\theta\, d\theta \\ &= 2a^2 \int_{-\pi/2}^{\pi/2} \cos^2\theta\, d\theta. \end{aligned}$$

The term $\cos^2\theta$ can be changed to a term involving $\cos 2\theta$ (which is easier to integrate) by using the rule $\cos 2\theta = 2\cos^2\theta - 1$. Hence $2\cos^2\theta = 1 + \cos 2\theta$ and

$$\begin{aligned} A &= a^2 \int_{-\pi/2}^{\pi/2} (1 + \cos 2\theta)\, d\theta \\ &= a^2 \left[\theta + \frac{\sin 2\theta}{2}\right]_{-\pi/2}^{\pi/2} \\ &= a^2 \left[\theta\right]_{-\pi/2}^{\pi/2} \qquad (\text{since} \quad \sin\pi = \sin(-\pi) = 0) \\ &= a^2 \left(\frac{\pi}{2} - \left(\frac{-\pi}{2}\right)\right) \\ &= \pi a^2. \end{aligned}$$

2. Finding Volumes

Cavalieri's principle

Cavalieri's principle states that if two bodies have the same height and their cross sections by horizontal planes at each height level have equal area then the two bodies have equal volume.

This principle is traditionally used to determine volumes by elementary means.

To prove Cavalieri's principle we choose a partition $h_0, h_1, \ldots, h_N$ of the height interval $[a, b]$ and slice the bodies by horizontal planes at heights h_i and h_{i-1}. The slices are approximately cylinders of height $\Delta h_i = h_i - h_{i-1}$. Their volume is approximately the product of the area of a cross section $A(h_i^*)$ and the height Δh_i. Hence the volume of the body is approximately

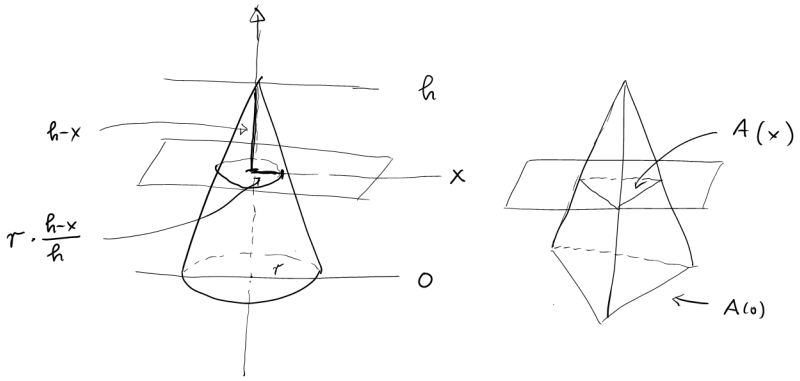
$$\sum_{i=1}^{N} A(h_i^*)\Delta h_i.$$

The exact value is again given by the integral

$$\int_a^b A(h)dh$$

which depends only on the area of the cross sections.

♦Example. **The volume of a cone.** Consider a cone of height h based on a disk of radius r. Then the cross-section at height x is a disk of radius $\frac{h-x}{h}r$, hence its areas is $A(x) = (\frac{h-x}{h})^2 A(0) = (\frac{h-x}{h})^2 \cdot 2\pi r^2$. Compare this to a pyramid of the same height h based on a triangle of area $2\pi r^2$. Then the area of a cross-section at height x will be $A(x) = (\frac{h-x}{h})^2 A(0)$ as for the cone.



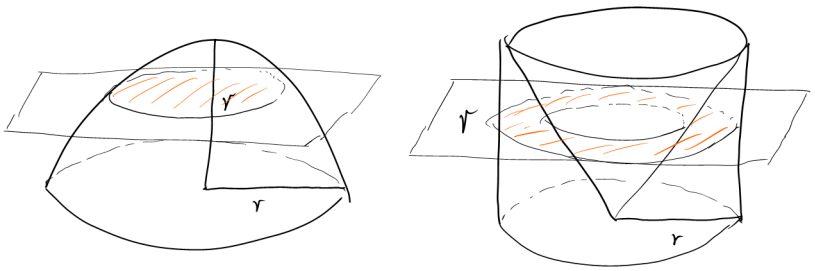
Hence, according to Cavalieri's principle the volume of the cone equals to the volume of the pyramid which is known to be $\frac{h}{3}A(0)$[10]. We find the volume of the

[10]This can be found by cutting a prism of height h with triangular base into three pyramids of equal volume.

cone

$$V = \frac{\pi r^2 h}{3}.$$

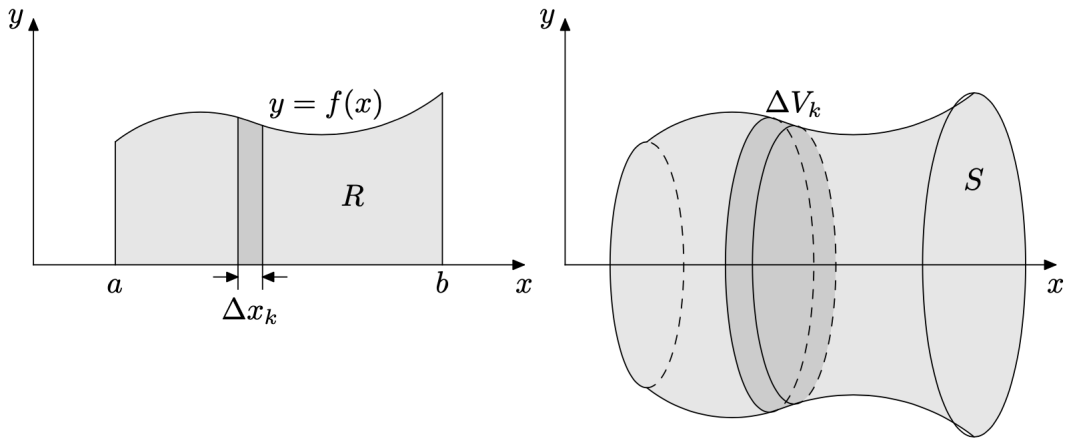
The volume of a hemisphere (and hence a sphere) of radius r. Consider a cylinder of radius r and height r from which an upside-down cone of radius r and height r is cut out. Then any cross section by a horizontal plane at height x is an annulus of outer radius r and inner radius x, hence its area equals $\pi(r^2 - x^2)$. On the other hand a cross section of the hemisphere by a horizontal plane at height x is a disk of radius $\sqrt{r^2 - x^2}$, hence its area equals $\pi(r^2 - x^2)$ as well.



Using Cavalieri's principle we conclude that the volume of the hemisphere equals the volume of the cylinder (πr^3) minus the volume of the removed cone ($\frac{1}{3}\pi r^3$). Hence the volume of the hemisphere is $\frac{2}{3}\pi r^3$.

Cones and spheres are solids of revolution. We will reprove the obtained formulae as more general formulae for the volumes of solids of revolution.

Rotation about the x-axis



Suppose the region R in the xy-plane under the curve $y = f(x)$ between $x = a$ and $x = b$ is rotated 360° around the x-axis to give the solid of revolution S.

Partition $[a, b]$ by points x_k and let $\Delta x_k = x_k - x_{k-1}$. The strip of width Δx_k when rotated about the x-axis produces a disc of width Δx_k and radius approximately $f(x_k)$. The volume of this disc is approximately

$$\Delta V_k \approx \pi \left[f(x_k)\right]^2 \Delta x_k.$$

Summing up we get

$$V \approx \sum_{k=1}^{n} \pi \left[f(x_k)\right]^2 \Delta x_k.$$

This is a Riemann sum where $x_k^* = x_k$, the right endpoint. For continuous f the sum on the right hand side approaches the integral $\displaystyle\int_a^b \pi \left[f(x)\right]^2 dx$, as $\max \Delta x_k \longrightarrow 0$. Hence the volume of V of the region S is given by

$$\begin{aligned} V &= \pi \int_b^a \left[f(x)\right]^2 \, dx \\ &= \pi \int_a^b y^2 \, dx. \end{aligned}$$

Example (Volume of a cone, alternatively): Let $y = mx$, $0 \leq x \leq h$. Rotating this line segment about the x-axis gives a cone of hight h and radius $r = mh$.

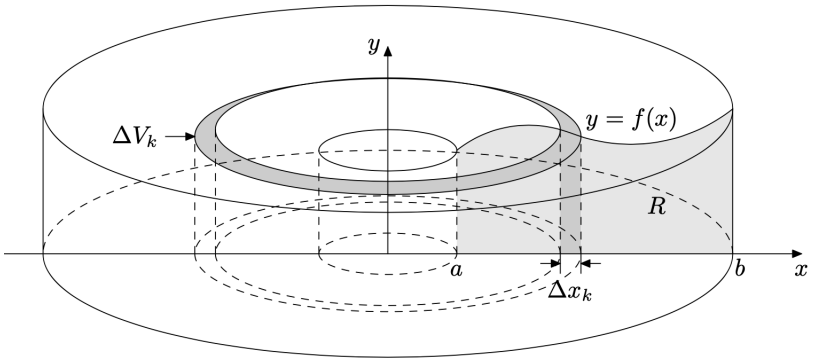
Solution:

$$V = \pi \int_0^h (mx)^2 \, dx = m^2 \pi \left[\frac{x^3}{3}\right]_0^h = \frac{\pi m^2 h^3}{3} = \frac{\pi r^2 h}{3}.$$

Example (Volume of a sphere, alternatively): Let $f(x) = \sqrt{a^2 - x^2}$ for $-a \leq x \leq a$. If this curve is rotated about the x-axis, we obtain a sphere of radius a whose volume is

$$\begin{aligned} V &= \pi \int_{-a}^{a} [f(x)]^2 \, dx \\ &= \pi \int_{-a}^{a} \left(a^2 - x^2\right) \, dx \\ &= \pi \left[a^2 x - \frac{x^3}{3}\right]_{-a}^{a} \\ &= \pi \left[\left(a^2 a - \frac{a^3}{3}\right) - \left(a^2(-a) - \frac{(-a)^3}{3}\right)\right] \\ &= \pi \left(\frac{2a^3}{3} + \frac{2a^3}{3}\right) \\ &= \frac{4}{3}\pi a^3. \end{aligned}$$

Rotation about the y-axis



Consider the same region R as before, but this time rotate it around the y-axis. The strip with width Δx_k sweeps out a cylindrical shell whose outer radius is x_k. Hence its outer circumference is $2\pi x_k$. The shell has width Δx_k and height $f(x_k)$. Hence its volume is approximately

$$\Delta V_k \approx \pi x_k^2 f(x_k) - \pi(x_k - \Delta x_k)^2 f(x_k) \approx 2\pi x_k f(x_k) \Delta x_k.$$

Here we have used that for small Δx_k the expression $(\Delta x_k)^2$ is so small that it can be neglected.

Adding these together we get

$$V \approx \sum_{k=1}^{n} 2\pi x_k f(x_k) \Delta x_k.$$

For continuous f, this sum has a limit as $\max \Delta x_k \longrightarrow 0$. The required volume is

$$\begin{aligned} V &= \int_a^b 2\pi x f(x)\, dx \\ &= 2\pi \int_a^b xy\, dx. \end{aligned}$$

Example (Volume of a cone, third version): Let $y = h - \frac{h}{r}x,\ 0 \leq x \leq r$. Rotating the triangle formed by this line segment and the coordinate axes about the y-axis gives a cone of hight h and radius r.

Solution:

$$V = 2\pi \int_0^r x(h - \frac{h}{r}x)\, dx = 2\pi h \left[\frac{x^2}{2} - \frac{x^3}{3r}\right]_0^r = 2\pi h \frac{r^2}{6} = \frac{\pi r^2 h}{3}.$$

21 The natural logarithm

Definition of the natural logarithm

The function $f(t) = \frac{1}{t}$ is a continuous for $t > 0$ and hence the function

$$F(x) = \int_1^x \frac{1}{t} dt \tag{12}$$

is well defined for $x > 0$. Without having a rigorous notion of the logarithm we gave a handwavy argument that the function $F(x) = \ln x$. We give now a rigorous treatment of logarithmic and exponential functions by adopting the function (12) as the definition of the natural log function. Then we prove that it satisfies the properties we expect from $\ln x$.

Properties of the Logarithm

(i) **Derivative of $\ln x$:**

$$\frac{d}{dx} \ln x = \frac{1}{x}.$$

This follows from the definition of $\ln x$ and the Fundamental Theorem of Calculus, $\frac{d}{dx} \int_a^x f(t)\, dt = f(x)$.

(ii) **Logarithm of 1:**

$$\ln 1 = \int_1^1 \frac{dt}{t} = 0.$$

(iii) **Logarithm of a product:**

Let $x, b > 0$. Then $\frac{d}{dx}(\ln bx) = \frac{1}{bx} \cdot b = \frac{1}{x}$. Thus $\frac{d}{dx}(\ln bx - \ln x) = \frac{1}{x} - \frac{1}{x} = 0$ and so $\ln bx - \ln x = c$. Taking $x = 1$ shows that $c = \ln b$. Thus $\ln bx - \ln x = \ln b$. Taking $x = a > 0$ gives

$$\ln ab = \ln a + \ln b.$$

(iv) **Logarithm of $\frac{1}{x}$:**

$$\begin{aligned} \frac{d}{dx} \ln \frac{1}{x} &= \frac{1}{1/x} \cdot \left(-\frac{1}{x^2} \right) \quad \text{(by the chain rule)} \\ &= -\frac{1}{x}. \end{aligned}$$

Hence $\frac{d}{dx}(\ln x + \ln \frac{1}{x}) = 0$ and $\ln x + \ln \frac{1}{x} = c$. Putting $x = 1$ shows that $c = 0$. Hence

$$\ln \frac{1}{x} = -\ln x \qquad (x > 0).$$

The last two results can be combined to give

$$\begin{aligned}\ln\left(\frac{a}{b}\right) &= \ln a + \ln\frac{1}{b}\\ &= \ln a - \ln b.\end{aligned}$$

(v) **Logarithm of a power:**

$$\ln x^n = n \ln x.$$

This may be proved by induction, using (iii).

(vi) **Monotonicity and concavity:**

Since

$$\begin{aligned}\frac{d}{dx}\ln x &= \frac{1}{x} \;>\; 0, \quad \text{for } x > 0,\\ \text{and } \frac{d^2}{dx^2}\ln x &= -\frac{1}{x^2} \;<\; 0, \quad \text{for } x > 0,\end{aligned}$$

we see that $\ln x$ is increasing and concave (downwards).

(vii) **Asymptotic behaviour:** We show that

$$\lim_{x\longrightarrow\infty} \ln x = \infty,$$

and

$$\lim_{x\longrightarrow 0^+} \ln\, x = -\infty.$$

For the first statement we need

$$\forall M > 0 \qquad \exists N > 0 \text{ such that } \forall x > N, \ln x > M.$$

Indeed, for $M > 0$ let $M' = \lceil \frac{M}{\ln 2} \rceil$ (that is, M' is the smallest integer that is greater than or equal to $\frac{M}{\ln 2}$). Let $N = 2^{M'}$. Then for $x > N$ we have

$$\begin{aligned}\ln x > \ln N = \ln 2^{M'} &\text{ because } \ln \text{ is strictly increasing,}\\ \ln 2^{M'} = M' \ln 2 \geq M &\text{ because of (v).}\end{aligned}$$

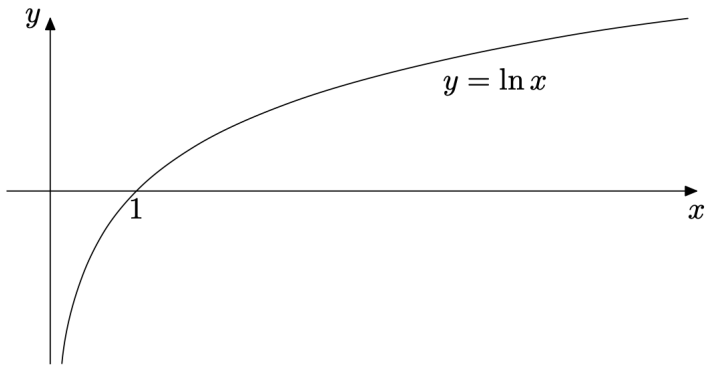
Thus $\ln x > M$, as required.

The second statement follows from this:

$$\lim_{x\to 0^+} \ln x = \lim_{t\to\infty} \ln\frac{1}{t} = -\lim_{t\to\infty} \ln t = -\infty.$$

Since $\ln x$ assumes arbitrarily large positive and arbitrarily large negative values and since it is continuous on $\mathbb{R}^+$ it also assumes all intermediate values. We conclude that the range of $\ln x$ is $\mathbb{R}$.

(viii) **Graph of $\ln x$:**



(ix) **The function $\ln|x|$:**

$$\ln|x| = \begin{cases} \ln x & \text{if } x > 0, \\ \ln(-x) & \text{if } x < 0. \end{cases}$$

If $x < 0$,

$$\frac{d}{dx}\ln|x| = \frac{d}{dx}\ln(-x) = \frac{1}{-x}(-1) = \frac{1}{x}.$$

Thus $\dfrac{d}{dx}\ln|x| = \dfrac{1}{x}$ for all $x \neq 0$, and

$$\int \frac{dx}{x} = \ln|x| + C$$

is valid for negative as well as positive x.

Remark. As discussed earlier, two primitives of a given function differ by a function whose derivative is zero. When we were looking for primitives of functions on an interval then we could conclude that any function whose derivative is zero must be a constant. This is not true when the integrand is defined on disjoint intervals as for $f(x) = \frac{1}{x}$ since we could choose different constants in different components. Here and later on we will interpret the integration "constant" C as a function whose derivative is zero, hence it can assume different values at different components of the domain.

(x) **Exponential functions.** Since $\ln x\colon (0,\infty) \to \mathbb{R}$ is strictly increasing (and hence injective) and also surjective, it has an inverse. We denote this inverse by $e^x\colon \mathbb{R} \to (0,\infty)$. This inverse function is well-defined throughout $\mathbb{R}$ and takes positive values. It is strictly increasing. Its derivative can be computed by the inverse function rule: Let $y = e^x$ and $x = \ln y$. Then

$$\frac{d}{dx}e^x = \frac{1}{\frac{d}{y}\ln x} = \frac{1}{\frac{1}{y}} = y = e^x.$$

The so-defined exponential function satisfies the rules

$$e^{x+y} = e^x \cdot e^y,$$

since

$$e^{x+y} = e^{\ln e^x + \ln e^y} = e^{\ln(e^x e^y)} = e^x e^y.$$

For $a > 0$, we define

$$a^x = e^{x \ln a}.$$

It follows from this definition that

$$e^{xy} = (e^y)^x.$$

Indeed, let $e^y = a$, i.e., $y = \ln a$ then

$$e^{xy} = e^{x \ln a} = a^x = (e^y)^x.$$

♠ *Exercises 43.* Show that for $a > 0$, and any $x, y \in \mathbb{R}$

$$\begin{aligned} a^{x+y} &= a^x \cdot a^y \\ a^{xy} &= (a^y)^x \\ \frac{d}{dx} a^x &= a^x \ln a \\ \int a^x dx &= \frac{a^x}{\ln a} + C. \end{aligned}$$

Worked Examples

(i) $\displaystyle\int_{-8}^{-5} \frac{dx}{x} = [\ln|x|]_{-8}^{-5} = \ln 5 - \ln 8 = \ln\left(\frac{5}{8}\right).$

(ii) $\displaystyle\int \tan x \, dx = -\int \frac{-\sin x}{\cos x} \, dx = -\ln|\cos x| + C.$

(Since $\cos x$ is sometimes negative it is necessary to include the absolute value sign.)

(iii) $\displaystyle\int \frac{2x+5}{x^2+5x+6} \, dx$: put $u = x^2 + 5x + 6$, $du = (2x+5)\,dx$,

$$\begin{aligned} \text{Integral} &= \int \frac{du}{u} = \ln|u| + C \\ &= \ln|x^2 + 5x + 6| + C. \end{aligned}$$

(iv) $\displaystyle\int \frac{\cos x}{2+\sin x}\,dx$: put $u = 2+\sin x$, $du = \cos x dx$,

$$\begin{aligned}\text{Integral} &= \int \frac{du}{u} = \ln u + C \\ &= \ln(2+\sin x) + C.\end{aligned}$$

(Since $2+\sin x > 0$ we do not need the absolute value sign.)

(v) Differentiate $\ln(x+\sqrt{1+x^2})$ and simplify. Write out the corresponding integration formula.

Let $y = \ln u$ where $u = x+\sqrt{1+x^2}$. Now $\dfrac{du}{dx} = 1 + \dfrac{2x}{2\sqrt{1+x^2}}$, so

$$\begin{aligned}\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} &= \frac{1}{u}\left(1+\frac{x}{\sqrt{1+x^2}}\right) \\ &= \frac{1}{x+\sqrt{1+x^2}}\left(\frac{\sqrt{1+x^2}+x}{\sqrt{1+x^2}}\right) \\ &= \frac{1}{\sqrt{1+x^2}}.\end{aligned}$$

Hence

$$\int \frac{dx}{\sqrt{1+x^2}} = \ln\left(x+\sqrt{1+x^2}\right) + C.$$

In particular,

$$\begin{aligned}\int_0^1 \frac{dx}{\sqrt{1+x^2}} &= \left[\ln(x+\sqrt{1+x^2})\right]_0^1 \\ &= \ln(1+\sqrt{2}).\end{aligned}$$

(vi) Differentiate with respect to x :

$$y = \ln\left(\frac{\sqrt{1+x^2}}{2x}\right).$$

Here it is better to use the rules for the logarithm to expand the expression before differentiating.

$$\begin{aligned}y &= \ln\sqrt{1+x^2} - \ln 2x \\ &= \frac{1}{2}\ln(1+x^2) - \ln 2 - \ln x. \\ \frac{dy}{dx} &= \frac{1}{2}\cdot\frac{2x}{1+x^2} - 0 - \frac{1}{x} \\ &= \frac{x}{1+x^2} - \frac{1}{x}.\end{aligned}$$

22 Approaching linear algebra

We start with a simple "predator-prey" model. Such models are used to describe the development of populations of two species, the predator and the prey, e.g. owls and rats. We can record the number of individuals (or their density on some region or the probability of sightings) at some instant of time in a column

$$\begin{bmatrix} o \\ r \end{bmatrix} = \begin{bmatrix} 20 \\ 3000 \end{bmatrix}.$$

If we consider a more complicated model with three or more species the column will have more entries. We will call a column of n real numbers a *vector*. The set of all vectors (with n entries) is denoted by $\mathbb{R}^n$. For $n = 2$ or $n = 3$ the entries of a vector can be interpreted as coordinates of a point on the 2-dimensional plane or in 3-dimensional space. We will get back to this point of view and to resulting applications of vectors in geometry later. At this stage we notice that it makes sense to *add* vectors component-wise, e.g. if we want to compute the population vector of a larger region from the vectors of the smaller subregions. We can also *scale* the vector of a population density component-wise by the size of a region to find the population vector of the region.

According to our model the population vector

$$\begin{bmatrix} O \\ R \end{bmatrix}$$

(notice the upper case letters) one year later follows the rule

$$\begin{aligned} O =& 0.9o + 0.002r \\ R =& -6o + 1.1r \end{aligned} \tag{13}$$

The coefficient 0.9 in the first equation expresses the rate (births−deaths) at which the owl population develops without the presence of rats. The owls prey on rats and the presence of rats increases their reproduction, which is expressed by the coefficient 0.002. In the second equation the negative coefficient -6 expresses the detrimental effect of the presence of owls on the population of rats (one owl eats 6 rats), whereas the coefficient 1.1 shows the strong reproduction of rats in the absence of owls.

If we consider the situation where O and R are given, the two equations (13) become a system of two linear equations and two unknowns o and r.

♠ *Exercises 44.* If the population vector is given by $O = 20$ and $R = 1040$ what was the population vector one year before, i.e. solve the system of linear equations

$$\begin{aligned} 0.9o + 0.002r =& 20 \\ -6o + 1.1r =& 1040. \end{aligned} \tag{14}$$

More generally, we will be interested in *systems of m linear equations with n unknowns*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad &= \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{15}$$

We think of the mn numbers $a_{11}, \ldots, a_{mn}$ as given numbers. They are called the *coefficients* of the system. The numbers $b_1, \ldots, b_m$ are also given and can be considered as a column vector with m entries. We refer to the column of the numbers $b_1, \ldots, b_m$ as the *right hand side* of the system. The numbers $x_1, \ldots, x_n$ are the unknowns, which also can be considered as the entries of a column vector. We can also write the coefficients of the system as a table of m rows and n columns. Such table is called an $m \times n$ *matrix*, it is usually denoted by an upper case letter and written as

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

A matrix with an equal number of rows and columns $m = n$ is called a *square matrix*.

The subscripts (or indices) of the entries show the position of each entry within the matrix. The first index always refers to the row and the second one to the column. We can scale a matrix by a factor component-wise and we can add matrices as long as they have the same size, i.e. the same number of rows and columns:

$$cA = c \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} ca_{11} & ca_{12} & \cdots & ca_{1n} \\ ca_{21} & ca_{22} & \cdots & ca_{2n} \\ \vdots & \vdots & & \vdots \\ ca_{m1} & ca_{m2} & \cdots & ca_{mn} \end{bmatrix}.$$

$$\begin{aligned} A + B &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix}. \end{aligned}$$

Multiplication of matrices is less obvious. We can multiply a matrix A with a matrix B only if the the number of columns of A equals to the number of rows of B, i.e., A is an $m \times k$-matrix and B is a $k \times n$-matrix. The result AB is an $m \times n$-matrix, i.e., it has as many rows as the first factor and as many columns as the second factor. In general, $AB \neq BA$ (which anyway only makes sense for square matrices.) The definition of the matrix product is as follows: The entries $c_{\mu\nu}$ of the $m \times n$-matrix $C = AB$ equal

$$c_{\mu\nu} = a_{\mu 1} b_{1\nu} + \cdots + a_{\mu k} b_{k\nu} = \sum_{\kappa=1}^{k} a_{\mu\kappa} b_{\kappa\nu}.$$

♦Example.

$$\begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 16 & 19 \\ 29 & 36 \end{bmatrix}.$$

♠ *Exercises 45.* Find an example of two 2×2-matrices A and B such that $AB \neq BA$.

Using the product of matrices we can rewrite the system of linear equations (15) as

$$AX = B$$

where A is the $m \times n$-matrix of coefficients and

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad B = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

are interpreted as $n \times 1$ and $m \times 1$ matrices, respectively.

We call a system of linear equations *homogeneous* if $B = 0$ and *inhomogeneous* otherwise. Any homogeneous system has at least one solution, namely $X = 0$. This solution is called the *trivial solution*.

With any $m \times n$-matrix A we can associate a mapping $A\colon \mathbb{R}^n \to \mathbb{R}^m$ by

$$Y = A(X) = AX.$$

This mapping is *linear* in the following sense: For any $U, V \in \mathbb{R}^n$ and $c \in \mathbb{R}$

$$A(U + V) = A(U) + A(V)$$
$$A(cU) = cA(U).$$

The proof is a direct verification. The entry y_μ of $Y = A(U + V)$ is

$$y_\mu = \sum_{\nu=1}^{n} a_{\mu\nu}(u_\nu + v_\nu) = \sum_{\nu=1}^{n} a_{\mu\nu}(u_\nu) + \sum_{\nu=1}^{n} a_{\mu\nu}(v_\nu)$$

which is clearly the sum of the ν-th entryies of $A(U)$ and $A(V)$, respectively. We just used distributivity of real numbers. The verification of the second statement is similar and left as an exercise.

We can now investigate when a linear mapping

$$Y = AX$$

is injective or surjective. Injectivity means that the system of linear equations

$$AX = B$$

has *at most* one solution, no matter what $B \in \mathbb{R}^m$ we choose. Interestingly, we can test for injectivity by just using $B = 0$.

Proposition 10. *The linear mapping $Y = AX$ is injective if and only if the homogeneous system $AX = 0$ has only the trivial solution.*

Proof. It is clear that the mapping cannot be injective if the homogeneous system has two solutions, namely the trivial solution and a nontrivial solution. Vice versa, if the mapping is not injective, i.e. for some B there are different solutions $U \neq V$ such that

$$AU = AV = B.$$

Then

$$A(U - V) = AU - AV = B - B = 0.$$

Since $U - V \neq 0$ this is a nontrivial solution. □

Surjectivity means that for any $B \in \mathbb{R}^m$ the linear system

$$AX = B$$

has *at least* one solution. In any case we need to develop a technique that allows us to solve an arbitrary system of linear equations and/or allows us to understand the structure of such solution. This technique is called *Gaussian elimination*.

23 Gaussian elimination

The usual strategy for solving an equation is to replace it by an equivalent equation that has the same solution. By repeating such manipulations we aim at an explicit expression that equals to the unknown variable. We use the same strategy to solve a system of (linear) equations.

It turns out that the following manipulations do not change the solutions and that they are sufficient to solve the system:

1. interchanging any two equations
2. scaling (both sides of) any equation by the same non-zero factor
3. adding a multiple of any one equation to another equation

It is clear that any solution of the system also satisfies the modified system. On the other hand, all these manipulations can be undone by the same kind of procedure. This shows that, vice versa, any solution of the modified system also solves the original one.

Let's demonstrate this in a specific example. We want to solve the following system of simultaneous linear equations for the unknowns x, y and z.

$$\begin{aligned} x + y + z &= 1 \\ x + y - z &= -1 \\ -x - 2y + z &= 2. \end{aligned}$$

Step 1. We add the -1-fold of the first equation to the second equation and we add the first equation to the last equation. These are manipulations of the 3rd kind. This yields

$$\begin{aligned} x + y + z &= 1 \\ -2z &= -2 \\ -y + 2z &= 3. \end{aligned}$$

Now the variable x occurs only in the first equation and has been eliminated from the second and third equation. We could now consider the second and third equation as a system of two equations and two unknowns, which is an easier problem of lower complexity.

Step 2. Swapping the second and the third equations yields

$$\begin{aligned} x + y + z &= 1 \\ -y + 2z &= 3 \\ -2z &= -2. \end{aligned}$$

The resulting system has a triangular form, where the first equation involves all variables x, y, z, the second equation depends only on y, z and the last equation depends only on z. We could now solve the new system by going backwards from the bottom to the top. However we keep applying our procedures to obtain the explicit solutions.

Step 3. Scaling the second equation by -1 and the third equation by $-\frac{1}{2}$ gives

$$\begin{aligned} x + y + z &= 1 \\ y - 2z &= -3 \\ z &= 1. \end{aligned}$$

Now the coefficient at x, y, z in the first, second and third equation respectively is 1. From this we read already that $z = 1$.

Step 4. Adding double of the third equation to the second equation and subtracting the third from the first equation gives

$$\begin{aligned} x + y &= 0 \\ y &= -1 \\ z &= 1. \end{aligned}$$

From this we read that $y = -1$ and $z = 1$.

Step 5. Subtracting the second equation from the first equation yields

$$\begin{aligned} x &= 1 \\ y &= -1 \\ z &= 1. \end{aligned}$$

Our system of three equations and three unknowns has the unique solution $x = 1$, $y = -1$, $z = 1$.

We now want to "abstract" this process and in doing so get an algorithm for solving general linear systems.

Firstly, we note that we could write our system of equations as a matrix, the *augmented matrix* – rows indicating the equation, columns the coefficient of the

unknown and a final column for the right hand sides. The prevents us from having to write the variables x, y, z in each modified system again and again.

$$\left.\begin{array}{rrr|r} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ -1 & -2 & 1 & 2 \end{array}\right.$$

Of course, we have to remember that column 1 represents the x's, column 2 the y's and column 3 the z's.

The three procedures for manipulating the system translate into the equivalent procedures for the augmented matrix

Any pair of rows may be interchanged.

Any row may be multiplied by a non-zero number.

Any multiple of a row may be added (or subtracted) from another.

How then do we use these three *elementary row operations* to arrive at the solution? Well, if the solution looks like "x = number", "y = number" and "z = number", the matrix for these three equations would be

$$\left.\begin{array}{ccc|l} 1 & 0 & 0 & \text{“number”} \\ 0 & 1 & 0 & \text{“number”} \\ 0 & 0 & 1 & \text{“number”}. \end{array}\right.$$

So if we can use our elementary row operations to reduce the augmented matrix to this form we can read the solution from the right most column.

Let's solve our problem using this technique. The best way to do this is to be methodical: start with column 1, get a 1 as the first entry then try to get the zeros for the entries beneath, go to column 2 and repeat the process with a 1 as second entry and so on. We start with,

$$\left.\begin{array}{rrr|r} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ -1 & -2 & 1 & 2. \end{array}\right.$$

We will indicate the operation performed using R_i to stand for row i with the first R indicating the row on which the operation is performed. For example, $R_1 - 2R_2$

means an operation performed on row 1, the new row 1 is the old row 1 minus twice row 2. Row 2 remains unchanged.

$$\begin{array}{r} \\ R_2 - R_1 \\ \\ \end{array}\left|\begin{array}{rrr|r} 1 & 1 & 1 & 1 \\ 0 & 0 & -2 & -2 \\ -1 & -2 & 1 & 2 \end{array}\right. \qquad \begin{array}{r} \\ \\ R_3 + R_1 \end{array}\left|\begin{array}{rrr|r} 1 & 1 & 1 & 1 \\ 0 & 0 & -2 & -2 \\ 0 & -1 & 2 & 3 \end{array}\right.$$

$$\begin{array}{r} \\ \text{(interchange) } R_2 \leftrightarrow R_3 \\ \\ \end{array}\left|\begin{array}{rrr|r} 1 & 1 & 1 & 1 \\ 0 & -1 & 2 & 3 \\ 0 & 0 & -2 & -2 \end{array}\right. \qquad \begin{array}{r} \\ R_1 + R_2 \\ \\ \end{array}\left|\begin{array}{rrr|r} 1 & 0 & 3 & 4 \\ 0 & -1 & 2 & 3 \\ 0 & 0 & -2 & -2 \end{array}\right.$$

$$\begin{array}{r} \\ \\ -\frac{1}{2}R_3 \end{array}\left|\begin{array}{rrr|r} 1 & 0 & 3 & 4 \\ 0 & -1 & 2 & 3 \\ 0 & 0 & 1 & 1 \end{array}\right. \qquad \begin{array}{r} \\ R_1 - 3R_3 \\ \\ \end{array}\left|\begin{array}{rrr|r} 1 & 0 & 0 & 1 \\ 0 & -1 & 2 & 3 \\ 0 & 0 & 1 & 1 \end{array}\right.$$

$$\begin{array}{r} \\ -R_2 \\ \\ \end{array}\left|\begin{array}{rrr|r} 1 & 0 & 0 & 1 \\ 0 & 1 & -2 & -3 \\ 0 & 0 & 1 & 1 \end{array}\right. \qquad \begin{array}{r} \\ R_2 + 2R_3 \\ \\ \end{array}\left|\begin{array}{rrr|r} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1. \end{array}\right.$$

We are done, we read the answer from the right-most column; $x = 1, y = -1, z = 1$.

This technique of *Gauss-Jordan elimination* can be used on linear systems of any size. The technique will always produce all solutions, if such a solution exists (we'll say more about this below).

♦Example. Solve the following system of linear equations

$$\begin{aligned} 2x_1 + x_2 - x_3 + x_4 &= 6 \\ x_1 + x_3 + 3x_4 &= 4 \\ 3x_1 - 2x_2 + x_3 &= 2 \\ x_1 + x_2 - x_4 &= -2 \end{aligned}$$

Solution.

The augmented matrix for the system is

$$\left[\begin{array}{rrrr|r} 2 & 1 & -1 & 1 & 6 \\ 1 & 0 & 1 & 3 & 4 \\ 3 & -2 & 1 & 0 & 2 \\ 1 & 1 & 0 & -1 & -2 \end{array}\right]$$

We proceed as before, working column by column.

$$\begin{array}{r|cccc|c} \frac{1}{2}R_1 & 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 3 \\ & 1 & 0 & 1 & 3 & 4 \\ & 3 & -2 & 1 & 0 & 2 \\ & 1 & 1 & 0 & -1 & -2 \end{array} \qquad \begin{array}{r|cccc|c} & 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 3 \\ R_2 - R_1 & 0 & -\frac{1}{2} & \frac{3}{2} & \frac{5}{2} & 1 \\ & 3 & -2 & 1 & 0 & 2 \\ & 1 & 1 & 0 & -1 & -2 \end{array}$$

$$\begin{array}{r|cccc|c} & 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 3 \\ & 0 & -\frac{1}{2} & \frac{3}{2} & \frac{5}{2} & 1 \\ R_3 - 3R_1 & 0 & \frac{-7}{2} & \frac{5}{2} & -\frac{3}{2} & -7 \\ & 1 & 1 & 0 & -1 & -2 \end{array} \qquad \begin{array}{r|cccc|c} & 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 3 \\ & 0 & -\frac{1}{2} & \frac{3}{2} & \frac{5}{2} & 1 \\ & 0 & -\frac{7}{2} & \frac{5}{2} & -\frac{3}{2} & -7 \\ R_4 - R_1 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{3}{2} & -5 \end{array}$$

$$\begin{array}{r|cccc|c} & 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 3 \\ -2R_2 & 0 & 1 & -3 & -5 & -2 \\ & 0 & \frac{-7}{2} & \frac{5}{2} & \frac{-3}{2} & -7 \\ & 0 & \frac{1}{2} & \frac{1}{2} & \frac{-3}{2} & -5 \end{array} \qquad \begin{array}{r|cccc|c} R_1 - \frac{1}{2}R_2 & 1 & 0 & 1 & 3 & 4 \\ & 0 & 1 & -3 & -5 & -2 \\ & 0 & \frac{-7}{2} & \frac{5}{2} & \frac{-3}{2} & -7 \\ & 0 & \frac{1}{2} & \frac{1}{2} & \frac{-3}{2} & -5 \end{array}$$

$$\begin{array}{r|cccc|c} & 1 & 0 & 1 & 3 & 4 \\ & 0 & 1 & -3 & -5 & -2 \\ R_3 + \frac{7}{2}R_2 & 0 & 0 & -8 & -19 & -14 \\ & 0 & \frac{1}{2} & \frac{1}{2} & \frac{-3}{2} & -5 \end{array} \qquad \begin{array}{r|cccc|c} & 1 & 0 & 1 & 3 & 4 \\ & 0 & 1 & -3 & -5 & -2 \\ & 0 & 0 & -8 & -19 & -14 \\ R_4 - \frac{1}{2}R_2 & 0 & 0 & 2 & 1 & -4 \end{array}$$

$$\begin{array}{r|cccc|c} & 1 & 0 & 1 & 3 & 4 \\ & 0 & 1 & -3 & -5 & -2 \\ -\frac{1}{8}R_3 & 0 & 0 & 1 & \frac{19}{8} & \frac{7}{4} \\ & 0 & 0 & 2 & 1 & -4 \end{array} \qquad \begin{array}{r|cccc|c} R_1 - R_3 & 1 & 0 & 0 & \frac{5}{8} & \frac{9}{4} \\ & 0 & 1 & -3 & -5 & -2 \\ & 0 & 0 & 1 & \frac{19}{8} & \frac{7}{4} \\ & 0 & 0 & 2 & 1 & -4 \end{array}$$

$$\begin{array}{r|cccc|c} & 1 & 0 & 0 & \frac{5}{8} & \frac{9}{4} \\ R_2 + 3R_3 & 0 & 1 & 0 & \frac{17}{8} & \frac{13}{4} \\ & 0 & 0 & 1 & \frac{19}{8} & \frac{7}{4} \\ & 0 & 0 & 2 & 1 & -4 \end{array} \qquad \begin{array}{r|cccc|c} & 1 & 0 & 0 & \frac{5}{8} & \frac{9}{4} \\ & 0 & 1 & 0 & \frac{17}{8} & \frac{13}{4} \\ & 0 & 0 & 1 & \frac{19}{8} & \frac{7}{4} \\ R_4 - 2R_3 & 0 & 0 & 0 & \frac{-15}{4} & \frac{-15}{2} \end{array}$$

$$\begin{array}{r|cccc|c} & 1 & 0 & 0 & \frac{5}{8} & \frac{9}{4} \\ & 0 & 1 & 0 & \frac{17}{8} & \frac{13}{4} \\ & 0 & 0 & 1 & \frac{19}{8} & \frac{7}{4} \\ -\frac{4}{15}R_4 & 0 & 0 & 0 & 1 & 2 \end{array} \qquad \begin{array}{r|cccc|c} R_1 - \frac{5}{8}R_4 & 1 & 0 & 0 & 0 & 1 \\ & 0 & 1 & 0 & \frac{17}{8} & \frac{13}{4} \\ & 0 & 0 & 1 & \frac{19}{8} & \frac{7}{4} \\ & 0 & 0 & 0 & 1 & 2 \end{array}$$

$$\begin{array}{r|cccc|c} & 1 & 0 & 0 & 0 & 1 \\ R_2 - \frac{17}{8}R_4 & 0 & 1 & 0 & 0 & -1 \\ & 0 & 0 & 1 & \frac{19}{8} & \frac{7}{4} \\ & 0 & 0 & 0 & 1 & 2 \end{array} \qquad \begin{array}{r|cccc|c} & 1 & 0 & 0 & 0 & 1 \\ & 0 & 1 & 0 & 0 & -1 \\ R_3 - \frac{19}{8}R_4 & 0 & 0 & 1 & 0 & -3 \\ & 0 & 0 & 0 & 1 & 2 \end{array}$$

Our solution can be read from the last column of the final augmented matrix:

$$x_1 = 1, x_2 = -1, x_3 = -3, x_4 = 2. \qquad \square$$

Existence of Solutions

For two-dimensional systems, i.e. systems in two independent variables x and y, a linear equations can be represented as a straight line in $\mathbb{R}^2$. For example,

$$\begin{aligned} x+y &= 2 \\ x+2y &= 5, \end{aligned}$$

represents a pair of lines intersecting in the point $(x,y)=(-1,3)$. This intersection point is the solution of the system. Of course a pair of straight lines need not intersect by one point – they can be parallel or coincide. If they coincide then all points (x,y) on the line will satisfy the system. For example,

$$\begin{aligned} x+2y &= 5 \\ -3x-6y &= -15. \end{aligned}$$

The second equation is simply a multiple of the first one and all points $(x, \frac{1}{2}(5-x))$, for any x, solve the system. Two distinct parallel lines give equations, which have no point in common, the equations are *inconsistent* – there is no solution. For example

$$\begin{aligned} x-3y &= 2 \\ x-3y &= 6. \end{aligned}$$

are inconsistent. The fact that the equations are inconsistent is easily discovered if we are using Gauss-Jordan. Of course, it's obvious in this case but for a large system it can be far from obvious.

$$\begin{array}{cc|c} 1 & -3 & 2 \\ 1 & -3 & 6 \end{array} \qquad \begin{array}{r} \\ R_2 - R_1 \end{array} \qquad \begin{array}{cc|c} 1 & -3 & 2 \\ 0 & 0 & 4 \end{array} \longleftarrow \text{ inconsistency.}$$

The last line in the resulting augmented matrix corresponds to the equation

$$0 \cdot x + 0 \cdot y = 4$$

which cannot hold, no matter what x and y are.

♦Example. Solve the following set of linear equations,

$$\begin{aligned} x-y-3z &= -3 \\ 3x+y-z &= -5 \\ x+2y+3z &= 0. \end{aligned}$$

Solution.

$$\begin{array}{rcccc} & 1 & -1 & -3 & -3 \\ & 3 & 1 & -1 & -5 \\ & 1 & 2 & 3 & 0 \end{array} \qquad \begin{array}{rcccc} & 1 & -1 & -3 & -3 \\ R_2 - 3R_1 & 0 & 4 & 8 & 4 \\ & 1 & 2 & 3 & 0 \end{array}$$

$$\begin{array}{rcccc} & 1 & -1 & -3 & -3 \\ & 0 & 4 & 8 & 4 \\ R_3 - R_1 & 0 & 3 & 6 & 3 \end{array} \qquad \begin{array}{rcccc} & 1 & -1 & -3 & -3 \\ \frac{1}{4}R_2 & 0 & 1 & 2 & 1 \\ & 0 & 3 & 6 & 3 \end{array}$$

$$\begin{array}{rcccc} R_1 + R_2 & 1 & 0 & -1 & -2 \\ & 0 & 1 & 2 & 1 \\ & 0 & 3 & 6 & 3 \end{array} \qquad \begin{array}{rcccc} & 1 & 0 & -1 & -2 \\ & 0 & 1 & 2 & 1 \\ R_3 - 3R_2 & 0 & 0 & 0 & 0 \end{array}$$

The last row is equivalent to the *tautological* (useless) equation

$$0 \cdot x + 0 \cdot y + 0 \cdot z = 0.$$

This is as far as we can go with this system. □

We consider now a system of m equations in n unknowns,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m. \end{aligned}$$

Which can be written in matrix form as

$$AX = B.$$

The augmented matrix is $(A|B)$. To solve the system (or to show it to be inconsistent) we would attempt to row reduce the augmented matrix. We notice that in all our examples, even in the inconsistent cases, we could always row reduce the augmented matrix to "stairstep" form

$$(*) \qquad \left[\begin{array}{ccc|c} 0 & \ldots & 0 & * \\ & & & \\ & & & \\ & & & \\ & & & \end{array}\right] \qquad \begin{array}{l} \\ * \\ * \\ * \\ \ldots \end{array}$$

where all the entries under the steps are zero, all corner (or *pivot*) entries marked with a $*$ are non-zero and all other entries are arbitrary. Note the steps descend one

row at a time whereas the span of each step may be more than one column. Such a matrix is known as a *row-echelon* matrix.

♦Example. The following are row-echelon matrices

$$\left[\begin{array}{cc} 1 & 5 \\ \hline 0 & 0 \end{array}\right], \quad \left[\begin{array}{cc} 1 & 10 \\ \hline 0 & 1 \end{array}\right], \quad \left[\begin{array}{ccc} 0 & 1 & 3 \\ \hline 0 & 0 & 2 \\ \hline 0 & 0 & 0 \end{array}\right]$$

$$\left[\begin{array}{cccccc} 0 & 0 & 1 & 5 & 6 & 7 \\ \hline 0 & 0 & 0 & 2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 4 \end{array}\right] \qquad \left[\begin{array}{ccc} 0 & 15 & 2 \\ \hline 0 & 0 & 1 \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

As you might have guessed every matrix can be turned into a row-echelon matrix.

Theorem 25. *By means of elementary row operations any non-zero matrix can be reduced to row-echelon form.*

Proof. Let $A = (a_{ij})$ be a $m \times n$ matrix. If A is the zero matrix we are done. Assume now that A has at least one non-zero column. From the left this first non-zero column must contain at least one non-zero element. By interchanging rows, if necessary, we can make sure that the first (top-most) element of the first non-zero column is non-zero. So A will have been transformed into a matrix of the form

$$B = \begin{bmatrix} 0 & \dots & 0 & b_{11} & b_{12} & \dots & b_{1n} \\ 0 & \dots & 0 & b_{21} & b_{22} & \dots & b_{2n} \\ & \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix}$$

with $b_{11} \neq 0$.

Performing $R_2 - \frac{b_{21}}{b_{11}}R_1, \dots, R_m - \frac{b_{m1}}{b_{11}}R_1$ yields

$$\left[\begin{array}{ccccccc} 0 & \dots & 0 & b_{11} & b_{12} & \dots & b_{1n} \\ \hline 0 & \dots & 0 & 0 & c_{22} & \dots & c_{2n} \\ & \vdots & & \vdots & \vdots & & \vdots \\ 0 & & 0 & 0 & c_{m2} & \dots & c_{mn} \end{array}\right].$$

where $c_{kj} = b_{kj} - \frac{b_{k1}}{b_{11}}b_{1j}$, for $k \geq 2$. Now apply the same process to the submatrix

$$\begin{bmatrix} c_{22} & \dots & c_{2n} \\ \vdots & & \vdots \\ c_{m2} & & c_{mn} \end{bmatrix}.$$

So after no more than m steps of this process we will arrive at a row-echelon matrix. □

If you now look a little closer at the row reductions we have performed you will see that they all have two other things in common, aside from being in row-echelon form. Firstly, the non-zero corner entries are all 1's. Secondly, every entry above each corner 1 is zero. A row echelon matrix with these two additional properties is called a *reduced row-echelon matrix* or *Hermite matrix*.

♦Example. The following are reduced row-echelon matrices

$$\left[\begin{array}{cc} 1 & 0 \\ \hline 0 & 1 \end{array}\right], \quad \left[\begin{array}{cccccc} 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \quad \left[\begin{array}{ccc} 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right].$$

Using the same method of proof as above we can now easily prove the following theorem.

Theorem 26. *Every non-zero matrix can, by means of elementary row operations, be transformed to a reduced row-echelon matrix.*

In fact, our solution method – Gauss-Jordan elimination is just the process of reducing a matrix to reduced row-echelon form.

Once we have reduced the augmented matrix to row-echelon form we can analyse the possible solutions of the system. Before drawing a conclusion whether the system has exactly one, infinitely many or no solutions we delete all zero rows (if there are any) at the bottom of the augmented matrix in row-echelon form. Now we have the following possibilities:

- The last non-zero equation has a non-zero coefficient. Then there is either a unique solution or an infinite number of solutions. The latter occurs if the row echelon form has any step of span greater than 1.
- The last equation reads $0 = c$, where $c \neq 0$. There is no solution, the equations are inconsistent.

One point we should make here is that it is straight forward to solve a system of equations once you have it in row-echelon form. The method simply involves systematically "back substituting" from the last equation. This technique is known as *Gaussian elimination*; on some occasions it may be quicker than Gauss-Jordan elimination.

♦Example. Use Gaussian elimination to solve the following system

$$\begin{aligned} x + 2y + 3z &= -1 \\ 3x + y + 2z &= 2 \\ 2x + 3y + z &= 0 \end{aligned}$$

Solution

The augmented matrix is

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & -1 \\ 3 & 1 & 2 & 2 \\ 2 & 3 & 1 & 0 \end{array}\right].$$

We now reduce to row-echelon form.

$$R_2 - 3R_1 \left[\begin{array}{ccc|c} 1 & 2 & 3 & -1 \\ 0 & -5 & -7 & 5 \\ 2 & 3 & 1 & 0 \end{array}\right] \qquad R_3 - 2R_1 \left[\begin{array}{ccc|c} 1 & 2 & 3 & -1 \\ 0 & -5 & -7 & 5 \\ 0 & -1 & -5 & 2 \end{array}\right]$$

$$\frac{-1}{5}R_2 \left[\begin{array}{ccc|c} 1 & 2 & 3 & -1 \\ 0 & 1 & \frac{7}{5} & -1 \\ 0 & -1 & -5 & 2 \end{array}\right] \qquad R_3 + R_2 \left[\begin{array}{ccc|c} 1 & 2 & 3 & -1 \\ 0 & 1 & \frac{7}{5} & 1 \\ 0 & 0 & \frac{-18}{5} & 1 \end{array}\right]$$

The matrix is in row-echelon form the equations are now

$$\begin{aligned} x + 2y + 3z &= -1 \\ y + \frac{7}{5}z &= 1 \\ \frac{-18}{5}z &= 1. \end{aligned}$$

From the last equation, $z = -\frac{5}{18}$. Substituting into the second equation gives $y = 1 - \frac{7}{5} \times \left(-\frac{5}{18}\right) = \frac{25}{18}$. Finally, substituting these values for y and z into the first equation gives $x = -\frac{53}{18}$. □

24 Square systems and determinants

Consider a system of two linear equations and two unknowns

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

By subtracting the a_{12}-fold of the second equation from the a_{22}-fold of the first equation we get

$$(a_{11}a_{22} - a_{21}a_{12})x_1 = b_1a_{22} - b_2a_{12}$$

hence

$$x_1 = \frac{b_1a_{22} - b_2a_{12}}{a_{11}a_{22} - a_{21}a_{12}}$$

if $a_{11}a_{22} - a_{21}a_{12} \neq 0$.

Similarly, we find

$$x_2 = \frac{b_2a_{11} - b_1a_{21}}{a_{11}a_{22} - a_{21}a_{12}}$$

if $a_{11}a_{22} - a_{21}a_{12} \neq 0$. The numerator and denominator expressions have a similar structure: they are differences of products. It is convenient to write this in the form

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}.$$

This expression is called the *determinant* of the 2×2 matrix $A = (a_{ij})$. Then

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}} \quad \text{and} \quad x_2 = \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}.$$

This is Cramer's rule.

In a similar way, one can define the determinant of a 3×3 matrix

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31}.$$

With some effort one can verify that Cramer's rule is also valid for linear systems of 3 variables and 3 unknowns. The solution of

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

is

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}, \quad x_2 = \frac{\begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_2 & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}} \quad \text{and} \quad x_3 = \frac{\begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}.$$

Determinants "determine" whether a square system of n equations and n unknowns have a unique solution. This can be expressed in terms of the *inverse matrix* . The matrix B is inverse to the matrix A if

$$AB = BA = I$$

where I is the identical matrix that has 1 along the main diagonal and zeros otherwise, i.e.

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

We write $B = A^{-1}$. Inverse matrices can be used to solve square systems of linear equations

$$A\boldsymbol{x} = \boldsymbol{b}.$$

Multiplication by A^{-1} gives

$$\boldsymbol{x} = A^{-1}A\boldsymbol{x} = A^{-1}\boldsymbol{b}.$$

Finding the inverse matrix is in general not easier than solving the system itself.

The determinant indicates whether an inverse matrix exists: the matrix A has an inverse if and only if the determinant of A is different from zero. For 2×2 matrices there is a simple formula which illustrates this. For

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

the inverse matrix is

$$A^{-1} = \frac{1}{\det A} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

if $\det A \neq 0$.

We will study determinants and the algebra of matrices in more detail in MTHS130 and Pmth213.

25 Applications in Geometry

Vectors have become a powerful tool in geometry, physics and other applications. It is common to denote a vector in this context by lower case boldface letters. The entries of a 2-vector

$$\boldsymbol{v} = \begin{bmatrix} x \\ y \end{bmatrix}$$

can be interpreted as the coordinates (x, y) of a point in the Cartesian plane and the entries of a 3-vector

$$\boldsymbol{w} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

can be viewed as the coordinates x, y, z of a point in three-dimensional space.

We have already interpreted linear equations of two variables:

$$ax + by = c$$

as equations of straight lines, if at least one of a, b is different from 0. If $b \neq 0$ this is the line

$$y = -\frac{a}{b}x + \frac{c}{b}$$

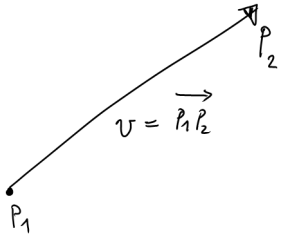
with slope $-\frac{a}{b}$ and y-intercept $\frac{c}{b}$. If $b = 0$ the equation gives a vertical line through through the point with coordinates $x = \frac{c}{a}$, $y = 0$. Now solving systems of linear equations with two variables means to find all points where the corresponding lines intersect.

It is preferable to change the geometric view on vectors slightly: To each point P on the plane or in space we can assign the unique parallel translation that shifts the origin O to P. We use the notation $\boldsymbol{v} = \overrightarrow{OP}$. We consider now vectors in 3-dimensional space. For vectors in the 2-dimensional plane just ignore the third entry. The parallel translation by a vector

$$\boldsymbol{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

of a point $P_1(a_1, b_1, c_1)$ is the point $P_2(a_2, b_2, c_2)$ with coordinates $a_2 = a_1 + x$, $b_2 = b_1 + y$, $c_2 = c_1 + z$. The vector $\boldsymbol{v}$ is completely determined if we know the result P_2 of the translation of any point P_1. Namely,

$$\boldsymbol{v} = \overrightarrow{P_1P_2} = \begin{bmatrix} a_2 - a_1 \\ b_2 - b_1 \\ c_2 - c_1 \end{bmatrix}.$$



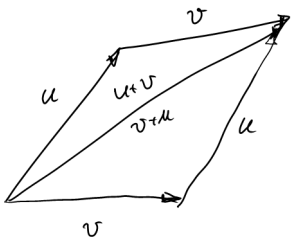
The advantage of the interpretation of a vector as a translation is that we can "add" and "scale" translations. Adding the translations $\boldsymbol{u}$ and $\boldsymbol{v}$ just means to perform them consecutively. If $\boldsymbol{u} = \overrightarrow{P_1P_2}$ and $\boldsymbol{v} = \overrightarrow{P_2P_3}$ then

$$\boldsymbol{u} + \boldsymbol{v} = \overrightarrow{P_1P_2} + \overrightarrow{P_2P_3} = \begin{bmatrix} a_2 - a_1 \\ b_2 - b_1 \\ c_2 - c_1 \end{bmatrix} + \begin{bmatrix} a_3 - a_2 \\ b_3 - b_2 \\ c_3 - c_2 \end{bmatrix} = \begin{bmatrix} a_3 - a_1 \\ b_3 - b_1 \\ c_3 - c_1 \end{bmatrix} = \overrightarrow{P_1P_3}.$$

The addition of vectors in their algebraic form is just component-wise addition:

$$\boldsymbol{u} + \boldsymbol{v} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ u_3 + v_3 \end{bmatrix}.$$

It is clear that this addition is commutative, i.e. $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$. Geometrically, vector addition can be performed by the so-called "parallelogram rule":

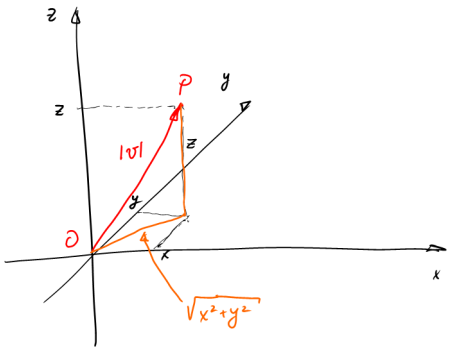


In its geometric interpretation a vector is characterised by

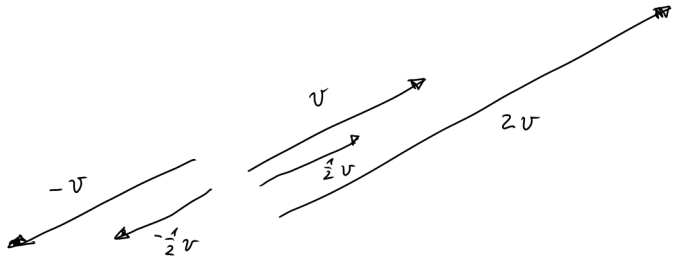
- its length (or norm or magnitude), i.e. a non-negative number equal to the distance by which points are shifted
- its direction, given by a straight line parallel to which the translation is performed
- its orientation, i.e. one of the two ways of moving along a line, given by two points on the line, labelled 'initial' and 'terminal'.

Vector quantities that feature magnitude, direction and orientation are very common in physics, e.g. velocity, force etc.

The length of the vector $\boldsymbol{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ is denoted by $\|\boldsymbol{v}\|$ and equals $\|\boldsymbol{v}\| = \sqrt{x^2 + y^2 + z^2}$.



Scaling of a vector $\boldsymbol{v}$ by a scalar (= number) c does not affect its direction, but changes the length $\|\boldsymbol{v}\|$ to $\|c\boldsymbol{v}\| = |c|\|\boldsymbol{v}\|$ and reverses the orientation iff $c < 0$.

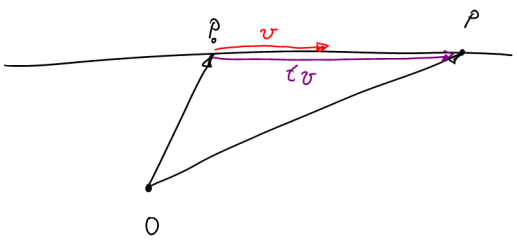


We notice that two vectors are parallel if and only if one can be obtained from the other by scaling. In this case we also call the two vectors *collinear*.

We apply the geometric version of vectors to describe straight lines in two- and three-dimensional space and, actually, in spaces of any dimension as the trajectory of a particle moving with constant vector velocity. Let $\boldsymbol{r} = \overrightarrow{OP}$ be the vector that shifts the origin O to an arbitrary point P on the trajectory. Furthermore, let P_0 be some given point on the trajectory and let $\boldsymbol{v}$ be the velocity vector. Then $\overrightarrow{P_0P}$ is a scalar multiple of $\boldsymbol{v}$, i.e.

$$\overrightarrow{P_0P} = t\boldsymbol{v},$$

where the parameter t can be interpreted as the time lapsed since the particle passed



the position P_0.

Hence

$$\boldsymbol{r}(t) = \overrightarrow{OP} = \overrightarrow{OP_0} + t\boldsymbol{v} = \boldsymbol{a} + t\boldsymbol{v}, \tag{16}$$

where the vectors $\boldsymbol{a} = \overrightarrow{OP_0}$ and $\boldsymbol{v}$ are given.

In the 2-dimensional plane the vector equation (16) is equivalent to the parametric equations

$$\begin{aligned} x &= a_1 + v_1 t \\ y &= a_2 + v_2 t \end{aligned}$$

where

$$\boldsymbol{r} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

These parametric equations give a vertical line if $v_1 = 0$. If $v_1 \neq 0$ we can determine t from the first equation and plug the resulting expression into the second equation:

$$y = a_2 + \frac{v_2}{v_1}(x - a_1) = \frac{v_2}{v_1}x + a_2 - \frac{v_2}{v_1}a_1,$$

which is the usual slope-intercept equation for a non-vertical line.

In 3-dimensional space the parametric equation has three components

$$\begin{aligned} x &= a_1 + v_1 t \\ y &= a_2 + v_2 t \\ z &= a_3 + v_3 t \end{aligned}$$

where

$$\boldsymbol{r} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

26 Linear combinations, linear independence and bases

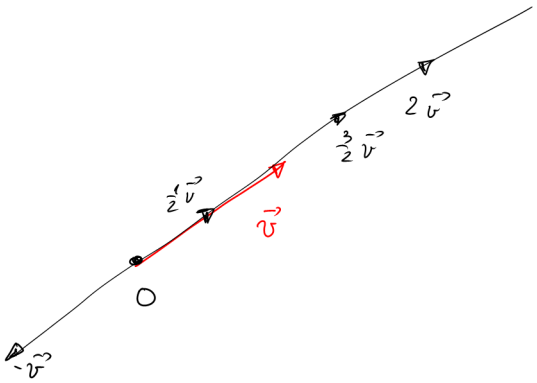
Often we are interested in a subset of $\mathbb{R}^n$, such as the subset of all solutions of a system of linear equations with n unknowns. Some of these sets have a special structure of a *linear subspace*. Before we give a precise definition and geometric interpretation of linear subspaces we introduce the important notion of linear combinations.

For a number of vectors $\boldsymbol{v_1}, \ldots, \boldsymbol{v_k}$ and the scalars $t_1, \ldots, t_k$ we define the *linear combination* as the vector

$$t_1\boldsymbol{v_1} + \cdots + t_k\boldsymbol{v_k}.$$

The set of all linear combinations of the vectors $\boldsymbol{v_1}, \ldots, \boldsymbol{v_k}$ is called the *span* of $\boldsymbol{v_1}, \ldots, \boldsymbol{v_k}$.

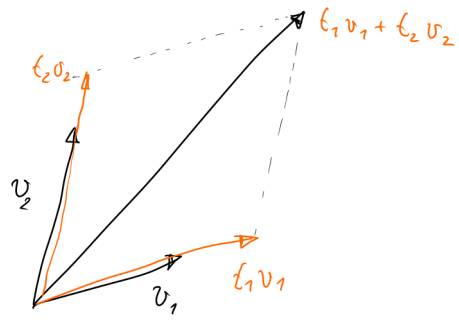
♦Example. If $k = 1$ and $\boldsymbol{v_1}$ is a non-zero vector then $t_1\boldsymbol{v_1}$ is a parallel vector with scaled length and possibly reversed orientation. The span of $\boldsymbol{v_1}$ is a straight line passing through the origin.



If $k = 2$ we have to consider two cases. If $\boldsymbol{v_1}, \boldsymbol{v_2}$ are collinear, e.g. $\boldsymbol{v_2} = r\boldsymbol{v_1}$, then the linear combinations

$$t_1\boldsymbol{v_1} + t_2\boldsymbol{v_2} = (t_1 + rt_2)\boldsymbol{v_1}$$

form just a straight line through the origin, as above. If $\boldsymbol{v_1}, \boldsymbol{v_2}$ are not collinear their linear combinations span a two-dimensional plane passing through the origin.



We define now the notion of a *linear subspace* V of $\mathbb{R}^n$ as a non-empty subset, which contains all linear combinations of its elements.

♦Example. 1. The subset that consists of the zero vector $\{\mathbf{0}\}$ is a linear subspace.

2. Any straight line passing through the origin is a linear subspace. For such straight line we can choose $P_0 = O$, hence $\boldsymbol{a} = \mathbf{0}$, so that the parametric equation becomes

$$\boldsymbol{r} = t\boldsymbol{v}.$$

Therefore, this straight line is the linear subspace spanned by the vector $\boldsymbol{v}$.

3. A straight line (more generally, any set) that does not contain the zero vector $\mathbf{0}$ is not a linear subspace. Indeed, a linear subspace V is, by definition, not empty. Let $\boldsymbol{v} \in V$. Then the linear combination $\mathbf{0} = 0 \cdot \boldsymbol{v} \in V$.

4. A plane passing through the origin spanned on two non-collinear vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. It can be described by a parametric equation that depends on two parameters s and t by

$$\boldsymbol{r} = s\boldsymbol{u} + t\boldsymbol{v}.$$

An arbitrary plane in $\mathbb{R}^3$ can be described by a parametric equation

$$\boldsymbol{r} = \boldsymbol{a} + s\boldsymbol{u} + t\boldsymbol{v}$$

where $\boldsymbol{a} = \overrightarrow{OP_0}$ and P_0 is some known point of the plane and s, t are arbitrary real parameters. The components of the parametric vector equation are

$$\begin{aligned} x &= a_1 + su_1 + tv_1 \\ y &= a_2 + su_2 + tv_2 \\ z &= a_3 + su_3 + tv_3 \end{aligned}$$

We can try and determine s, t from the first two equations

$$\begin{aligned} su_1 + tv_1 =& x - a_1 \\ su_2 + tv_2 =& y - a_2. \end{aligned}$$

If $u_1v_2 - u_2v_1 \neq 0$ then there is a unique solution

$$s = \frac{\begin{vmatrix} x - a_1 & v_1 \\ x - a_2 & v_2 \end{vmatrix}}{\begin{vmatrix} u_1 & v_1 \\ u_2 & v_2 \end{vmatrix}} = \frac{v_2}{u_1v_2 - u_2v_1}(x - a_1) - \frac{v_1}{u_1v_2 - u_2v_1}(y - a_2) = \alpha_1 + \beta_1 x + \gamma_1 y$$

$$t = \frac{\begin{vmatrix} u_1 & x - a_1 \\ u_2 & x - a_2 \end{vmatrix}}{\begin{vmatrix} u_1 & v_1 \\ u_2 & v_2 \end{vmatrix}} = -\frac{u_2}{u_1v_2 - u_2v_1}(x - a_1) + \frac{u_1}{u_1v_2 - u_2v_1}(y - a_2) = \alpha_2 + \beta_2 x + \gamma_2 y.$$

Plugging this into the third equation gives the equation of the plane

$$z = Ax + By + D, \tag{17}$$

where $A = \beta_1 u_3 + \beta_2 v_3$, $B = \gamma_1 u_3 + \gamma_2 v_3$, $D = a_3 + \alpha_1 u_3 + \alpha_2 v_3$.

A more general equation (without the condition $u_1v_2 - u_2v_1 \neq 0$) for a plane in $\mathbb{R}^3$ is

$$Ax + By + Cz + D = 0, \tag{18}$$

where A, B, C, D are real constants and at least one of A, B, C is different from 0. Notice that equation (18) is equivalent to equation (17) if $C \neq 0$. Dividing by C yields

$$z = -\frac{A}{C}x - \frac{B}{C}y - \frac{D}{C}.$$

We can now interpret a system of m linear equations with three unknowns geometrically as the problem of finding the common points of m planes in $\mathbb{R}^3$. Two equations give two planes which may intersect in a line or be parallel. If the planes are parallel the pair of equations will be inconsistent, if they coincide all points on the plane satisfy the pair of equations. If we have a third plane (third equation) then there are a number of possibilities, for distinct planes,

- two (or all three) of the planes are distinct and parallel – no common intersection, inconsistent.
- all planes distinct and non-parallel but they intersect pairwise by three parallel straight lines – inconsistent.

- all three planes intersect in a single straight line – we can solve for two of the unknowns in terms of the third. – Infinitely many solutions forming a straight line.

- the three planes intersect in a single point – a unique solution.

Linear subspaces of $\mathbb{R}^n$ contain infinitely many elements (except for $\{\mathbf{0}\}$) but can be described as the spans of finitely many vectors. A minimal set that spans a linear subspace is called *basis* of the subspace. Being minimal implies that the spanning set does not contain vectors that are linear combinations of the other elements. This is captured by the following definition:

A set of vectors $\boldsymbol{v_1}, \ldots, \boldsymbol{v_k}$ is called linearly independent if the vector equation

$$t_1\boldsymbol{v_1} + \cdots + t_k\boldsymbol{v_k} = \mathbf{0}$$

is only satisfied if $t_1 = \cdots = t_k = 0$, i.e it has only the trivial solution. Otherwise, the set is called linearly dependent. This can be interpreted as a homogeneous system of n equations and k unknowns in the following way: Write the k columns $\boldsymbol{v_1}, \ldots, \boldsymbol{v_k}$ as the columns of a matrix

$$V = [\boldsymbol{v_1}, \ldots, \boldsymbol{v_k}] = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nk} \end{bmatrix}.$$

Then

$$t_1\boldsymbol{v_1} + \cdots + t_k\boldsymbol{v_k} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nk} \end{bmatrix} \begin{bmatrix} t_1 \\ \vdots \\ t_k \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

is indeed a homogeneous systems of linear equations and we have linear independence if and only if the trivial solution $t_1 = \cdots = t_k = 0$ is the only solution.

♦Example. The vectors

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } \boldsymbol{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

are linearly independent because the system

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

has only the trivial solution.

The vectors

$$\boldsymbol{v}_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \text{ and } \boldsymbol{v}_2 = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

are linearly dependent because the system

$$\begin{bmatrix} 2 & 4 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

has the non-trivial solution $t_1 = -2$, $t_2 = 1$. Notice that $\boldsymbol{v}_2$ is a multiple of $\boldsymbol{v}_1$.

In a linearly dependent set there exists at least one vector, which is the linear combination of the remaining ones. Indeed, if one of the $t_1 \ldots, t_k$ is different from zero then the corresponding vector is a linear combination of the others. E.g. if $t_1 \neq 0$ then

$$\boldsymbol{v_1} = -\frac{t_2}{t_1}\boldsymbol{v_2} - \cdots - \frac{t_k}{t_1}\boldsymbol{v_k}$$

Although a linear subspace can have many different bases, their cardinality is always the same and is called the dimension of the subspace. This topic will be studied in more detail in MTHS130 and Pmth213.

The notion of linear subspaces helps us to better understand the structure of the solution set of a homogeneous system of m linear equations with n unknowns

$$A\boldsymbol{x} = \boldsymbol{0}.$$

In fact, this solution set is a linear subspace of $\mathbb{R}^n$. For any solutions $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$ any linear combination is also a solution since

$$A(t_1\boldsymbol{x_1} + \cdots + t_k\boldsymbol{x_k}) = t_1 A\boldsymbol{x_1} + \cdots + t_k A\boldsymbol{x_k} = 0.$$

In MTHS130 we will show that the dimension d of this subspace equals $n-r$, where r is the dimension of the subspace of $\mathbb{R}^m$ spanned by the rows of A, thus $n-m \leq d \leq n$. One can expect that each equation brings down the dimension by 1 starting from n. However this is only the case if the equations are linearly independent. Rows of A that are linearly dependent from other rows can be deleted from the system without changing the space of solutions. Geometrically, the solution space can be interpreted as a d-dimensional plane that passes through the origin $\boldsymbol{0}$.

The solution set of a non-homogeneous system

$$A\boldsymbol{x} = \boldsymbol{b}$$

with $\boldsymbol{b} \neq 0$ is never a linear subspace. (Because it does not contain $\boldsymbol{0}$.) However, it can be interpreted as a d-dimensional plane passing through a point that corresponds to a single particular solution. Assume that $\boldsymbol{x_{part}}$ is a particular solution

of the inhomogeneous system and $\boldsymbol{x_0}$ is an arbitrary solution of the corresponding homogeneous system (i.e., same A and $\boldsymbol{b} = \boldsymbol{0}$.) That is

$$A\boldsymbol{x_{part}} = \boldsymbol{b}$$

and

$$A\boldsymbol{x_0} = \boldsymbol{0}.$$

Then $\boldsymbol{x_{part}} + \boldsymbol{x_0}$ is also a solution because

$$A(\boldsymbol{x_{part}} + \boldsymbol{x_0}) = A\boldsymbol{x_{part}} + A\boldsymbol{x_0} = \boldsymbol{b} + \boldsymbol{0} = \boldsymbol{b}.$$

Any solution $\boldsymbol{x}$ is like this, because the difference $\boldsymbol{x} - \boldsymbol{x_{part}}$ is a solution of the homogeneous system:

$$A(\boldsymbol{x} - \boldsymbol{x_{part}}) = A\boldsymbol{x} - A\boldsymbol{x_{part}} = \boldsymbol{b} - \boldsymbol{b} = \boldsymbol{0}.$$

♦Example. Consider the homogeneous system

$$\begin{aligned} 2x + 3y - z &= 0 \\ 3x + 2y - z &= 0. \end{aligned}$$

Gaussian elimination reduces this system to

$$\begin{aligned} x + \frac{3}{2}y - \frac{1}{2}z &= 0 \\ y - \frac{1}{5}z &= 0. \end{aligned}$$

For any arbitrary value of z we find the parametric solution $x = 5z$, $y = 5z$, $z = z$. The solution space is 1-dimensional and spanned on the vector

$$\begin{bmatrix} 5 \\ 5 \\ 1 \end{bmatrix}.$$

The geometric interpretation of the solution is the straight line

$$\boldsymbol{r} = t \begin{bmatrix} 5 \\ 5 \\ 1 \end{bmatrix}.$$

Consider now the inhomogeneous system

$$\begin{aligned} 2x + 3y - z &= 2 \\ 3x + 2y - z &= 3. \end{aligned}$$

Gaussian elimination yields

$$\begin{aligned} x + \frac{3}{2}y - \frac{1}{2}z &= 1 \\ y - \frac{1}{5}z &= 0. \end{aligned}$$

Since we need only one particular solution we can put $z = 0$. Then $y = 0$ and $x = 1$. This yields the general solution

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 5 \\ 5 \\ 1 \end{bmatrix}.$$

Geometrically, this is a straight line passing through the point $P_0(1, 0, 0)$:

$$\boldsymbol{r} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 5 \\ 5 \\ 1 \end{bmatrix}.$$

27 Complex numbers

We have seen that vectors can be added component-wise and that this addition satisfies the same properties as the addition of numbers.

It is much more tricky to multiply vectors. A completely satisfactory approach exists only in $\mathbb{R}^2$.

We want to define $(x, y) \cdot (u, v)$. A first naive attempt suggests component-wise multiplication

$$(x, y) * (u, v) = (xu, yv).$$

The so defined product would be commutative, associative and distributive. Then $(1, 1)$ would be the neutral element: $(1, 1) * (x, y) = (x, y)$. The problem of such definition is that too many vectors have no multiplicative inverse: We cannot divide by a vector if one of its components is zero.

It turns out that a much better definition is

$$(x, y) \cdot (u, v) = (xu - yv, xv + yu).$$

This product is commutative. Interchanging (x, y) and (u, v) does not affect the result. The vector $(1, 0)$ serves as the neutral element, i.e.

$$(x, y)(1, 0) = (x, y).$$

If x and y are not both zero then

$$\left(\frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2} \right)$$

is a multiplicative inverse of (x, y). In fact,

$$(x, y)(\frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2}) = (\frac{x^2}{x^2 + y^2} + \frac{y^2}{x^2 + y^2}, \frac{-xy}{x^2 + y^2} + \frac{yx}{x^2 + y^2}) = (1, 0).$$

This motivates the following definition.

A complex number is a vector (x, y) with $x, y \in \mathbb{R}$. The sum of two complex numbers $(x, y) + (u, v)$ is defined as the vector $(x + u, y + v)$. The product $(x, y)(u, v)$ is defined as the vector $(xu - yv, xv + yu)$. The set of complex numbers is denoted by $\mathbb{C}$.

Usually we denote a complex number by just one letter, e.g., $z = (x, y)$, $w = (u, v)$.

All the axioms for multiplication and addition of the rational or real numbers hold for the complex numbers. Any set that satisfies those axioms is called a *field*. Thus the sets of rational, real and complex numbers are fields.

1. $z + w = w + z$ for all $z, w \in \mathbb{C}$.
2. $(z + w) + s = z + (w + s)$ for any $z, w, s \in \mathbb{C}$.
3. Denote the complex number $(0, 0)$ by 0. Then $0 + z = z + 0 = 0$, i.e. 0 is the neutral element for the addition of complex numbers.
4. For any complex number $z = (x, y)$ there is a complex number $w = (-x, -y)$ such that $z + w = 0$. We write $w = -z$.
5. $z \cdot 0 = 0 \cdot z = 0$ for any $z \in \mathbb{C}$.
6. $zw = wz$ for any $z, w \in \mathbb{C}$.
7. $(zw)s = z(ws)$ for any $z, w, s \in \mathbb{C}$.
8. Denote the complex number $(1, 0)$ by 1. Then $1 \cdot z = z \cdot 1 = z$, i.e. 1 is the neutral element for the multiplication of complex numbers.
9. For any complex number $z = (x, y)$ with $z \neq 0$ there is a complex number w such that $zw = 1$. We write $w = \frac{1}{z}$ or $w = z^{-1}$.
10. $(z + w)s = (zs + ws)$ for any $z, w, s \in \mathbb{C}$.

The real numbers can be included in $\mathbb{C}$ as a subset, $\mathbb{R}$ can be identified with the complex numbers of the form $(x, 0)$. For such numbers we just write x instead of $(x, 0)$. The multiplication rule simplifies to

$$x(u, x) = (x, 0)(u, v) = (xu, xv).$$

It is common to denote the complex number $(0, 1)$ by i. It is called the *imaginary unit*. It has the surprising property

$$\mathrm{i}^2 = (0, 1)^2 = (-1, 0) = -1.$$

The square of the imaginary unit is a negative real number! We will find out later that any complex number has a square root.

Now any complex number can be written as

$$z = (x, y) = (x, 0) + (0, y) = (x, 0) + (y, 0)(0, 1) = x + y\,\mathrm{i}\,.$$

This is the standard notation for complex numbers. The component x is the *real part* and y is the *imaginary part* of z. Notice that both real and imaginary part are real numbers. We write $\operatorname{Re} z = x$ for the real part of z and $\operatorname{Im} z = y$ or the imaginary part of z

♦Example.

(a) $(1+3i)+(3+i) = 4+4i = 4(1+i)$

(b) $(1+3i)-(3+i) = -2+2i = 2(-1+i)$

(c) $(\pi+i)-(1+\sqrt{2}i) = (\pi-1)+(1-\sqrt{2})i$

(d) $\left(\frac{1}{2}+\frac{1}{3}i\right)+\left(\frac{1}{4}-\frac{1}{6}i\right) = \frac{3}{4}+\frac{1}{6}i$

The multiplication rule for complex numbers looks rather difficult at the first glance. However it is easy to multiply complex numbers by expanding the expression

$$(x+\mathrm{i}\,y)(u+\mathrm{i}\,v) = xu+\mathrm{i}\,xv+\mathrm{i}\,yu+\mathrm{i}^2\,yv.$$

Then remember that $\mathrm{i}^2 = -1$ and extract the real and the imaginary parts. This yields

$$(x+\mathrm{i}\,y)(u+\mathrm{i}\,v) = xu-yv+\mathrm{i}(xv+yu).$$

♦Example.

(a)
$$\begin{aligned}(3+4i)(6+i) &= 3\cdot(6+i)+4i\cdot 6+i)\\ &= 18+3i+24i+4i^2\\ &= 18+3i+24i-4\\ &= 14+27i.\end{aligned}$$

(b)
$$\begin{aligned}(2-7i)(3-2i) &= 2(3-2i)-7i(3-2i)\\ &= 6-4i-21i-14\\ &= -8-25i.\end{aligned}$$

(c)
$$\begin{aligned}(\sqrt{2}+i\sqrt{3})(1-i) &= \sqrt{2}(1-i)+i\sqrt{3}(1-i)\\ &= \sqrt{2}-i\sqrt{2}+i\sqrt{3}+\sqrt{3}\\ &= (\sqrt{2}+\sqrt{3})+i(\sqrt{3}-\sqrt{2}).\end{aligned}$$

(d)
$$\begin{aligned}(\sqrt{2}-i)^2 &= (\sqrt{2})^2-2\sqrt{2}i+i^2\\ &= 2-2\sqrt{2}i-1\\ &= 1-2\sqrt{2}i.\end{aligned}$$

Equality of Complex Numbers

To specify a complex number we must give two real numbers, the real and imaginary parts. So two complex numbers are equal if and only if their real and imaginary parts are equal (respectively).

For complex numbers $z_1 = a_1 + ib_1$ *and* $z_2 = a_2 + ib_2$ *we have* $z_1 = z_2$ *if and only if* $a_1 = a_2$ *and* $b_1 = b_2$.

♦Example. Find all complex numbers for which

$$z^2 = -3 + 4i.$$

Solution We write $z = x + iy$, with x and y real. Substituting into the equation we have

$$\begin{aligned} z^2 = (x+iy)^2 &= -3 + 4i \\ \text{i.e. } x^2 - y^2 + i2xy &= -3 + 4i. \end{aligned}$$

Now equate real and imaginary parts – remember the complex number on the left can only equal that on the right if and only if their real and imaginary parts are (respectively) equal. We get

$$x^2 - y^2 = -3 \text{ and } 2xy = 4.$$

From the second of these equations we have

$$y = \frac{2}{x},$$

which we substitute into the first equation. This gives

$$x^2 - \left(\frac{2}{x}\right)^2 = -3.$$

Multiplying this equation through by x^2 gives as

$$\begin{aligned} x^4 - 4 &= -3x^2 \\ \text{i.e. } x^4 + 3x^2 - 4 &= 0. \end{aligned}$$

This is a quadratic in x^2, we factorise

$$(x^2 + 4)(x^2 - 1) = 0,$$

so that $x^2 = 1$ or $x^2 = -4$. But x must be real, so we cannot have $x^2 = -4$. So we conclude $x^2 = 1$, which gives $x = \pm 1$. We found earlier that $y = 2/x$, so we have two possible solutions

$$(x, y) = (1, 2) \text{ or } (-1, -2).$$

Giving two possible complex numbers z,

$$z = 1 + 2i \quad \text{or} \quad z = -1 - 2i.$$

□

The geometry of Complex numbers

Similarly to the visualisation of real numbers as points at a line, we can represent complex numbers as points in the plane. A number $z = x + \mathrm{i}\, y$ corresponds to the point with coordinates (x, y). Below we introduce some functions of complex numbers that have a clear geometric meaning.

Conjugate of a Complex Number

For any complex number $z = x + \mathrm{i}\, y$ we define the conjugate

$$\bar{z} = x - \mathrm{i}\, y.$$

Notice that all we have to do to get the complex conjugate of a complex number is to replace the imaginary part by its negative.

♦Example.

(a) If $z = 3 + 2\,\mathrm{i}$ then $\overline{z} = 3 - 2\,\mathrm{i}$.

(b) If $z = 27 - 5\,\mathrm{i}$ then $\overline{z} = 27 + 5\,\mathrm{i}$.

(c) If $z = 5$ then $\overline{z} = 5$.

(d) If $z = 6\,\mathrm{i}$ then $\overline{z} = -6\,\mathrm{i}$.

Taking the conjugate is a simple but very important function $\mathbb{C} \to \mathbb{C}$. It has the following properties:

1. $\bar{\bar{z}} = z$
2. $\overline{z + w} = \bar{z} + \bar{w}$
3. $\overline{zw} = \bar{z}\bar{w}$.

4. $z\bar{z} = (\operatorname{Re} z)^2 + (\operatorname{Im} z)^2$.

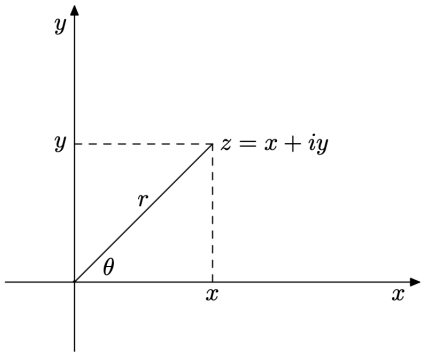
5. $\operatorname{Re} z = \frac{1}{2}(z + \bar{z})$.

6. $\operatorname{Im} z = \frac{1}{2\mathrm{i}}(z - \bar{z})$.

The conjugate $\bar{z}$ of the number z considered as a point in the plane is the reflection of z with respect to the x-axis.

Absolute value and argument

A point $x \neq 0$ in the plane can be determined by its Cartesian coordinates (x, y) but also by its distance from the origin and the angle between the x-axis and the unique line through 0 and z. (See the picture below.)



The distance between z and 0 is the length of the hypotenuse of a right triangle with catheti[11] $\operatorname{Re} z$ and $\operatorname{Im} z$. According to Pythagoras' theorem this distance equals

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

We define the *absolute value* (or *modulus*) function $\mathbb{C} \to \mathbb{R}^+$ by $z \mapsto |z| = \sqrt{z\bar{z}}$. The absolute value of a complex numbers is non-negative and it equals zero if and only if $z = 0$.

Notice that

$$|x| = |\operatorname{Re} z| \leq |z| \text{ and } |y| = |\operatorname{Im} z| \leq |z|.$$

♦Example.

(a) For $z = 2 + 3\,\mathrm{i}$ we have $z\overline{z} = 2^2 + 3^2 = 13$

(b) If $z = 2$ then $z\overline{z} = 2^2 = 4$

[11] Cathetus (pl. catheti) is the name for the two short sides of a right triangle.

(c) If $z = -3\,\mathrm{i}$ then $z\overline{z} = (-3)^2 = 9$.

If θ denotes the angle (measured in radians) between the x-axis and the line through 0 and z then

$$\begin{aligned} \operatorname{Re} z = x &= |z| \cos\theta \\ \operatorname{Im} z = y &= |z| \sin\theta. \end{aligned}$$

The angle θ is called the *argument*[12] of z. We write

$$\theta = \arg z.$$

If z is in the right half plane (i.e. if $\operatorname{Re} z > 0$) then $\arg z$ can be found from

$$\tan\theta = \frac{\operatorname{Im} z}{\operatorname{Re} z},$$

hence $\arg z = \arctan \frac{\operatorname{Im} z}{\operatorname{Re} z}$. If $\operatorname{Re} z < 0$ we have the modified formula $\arg z = \arctan \frac{\operatorname{Im} z}{\operatorname{Re} z} + \pi$. If $\operatorname{Re} z = 0$ the argument of z is $\frac{\pi}{2}$ (if $\operatorname{Im} z > 0$) or $-\frac{\pi}{2}$ (if $\operatorname{Im} z < 0$) or undefined (if $\operatorname{Im} z = 0$).

Notice that the argument of a complex number is only defined up to a summand of $2k\pi$ where k is an integer. We could resolve this ambiguity by restricting the arguments to the interval

$$0 \le \theta < 2\pi,$$

(or, better, to $-\pi < \theta \le \pi$. However the resulting arg function would be discontinuous: If we approach $z = 1$ from a region below the x-axis $\arg z$ would tend to 2π. This is different from $\arg 1 = 0$, which is the limit for z approaching 1 from a region above the x-axis.

The coordinates $(r, \theta) = (|z|, \arg z)$ are called *polar coordinates*. They are very well adapted to the multiplication of complex numbers.

Theorem 27. *If z, w are non-zero complex numbers then*

$$\begin{aligned} |zw| &= |z||w| \\ \arg zw &= \arg z + \arg w. \end{aligned}$$

Notice that the formula $\arg zw = \arg z + \arg w$ would not be true if we took the arguments only between 0 and 2π because the sum of two such argument needs not to stay within that interval.

[12]Notice that the word argument can have different meanings in mathematics. One meaning is as defined here. The other meaning is the input of a function and has been used before.

Proof. $|zw| = |z||w|$ is equivalent to $|zw|^2 = |z|^2|w|^2$. Here the LHS equals

$$zw\overline{zw} = zw\bar{z}\bar{w} = z\bar{z}w\bar{w} = |z|^2|w|^2,$$

which is the required RHS.

The second statement is more tricky. Recall the addition formulae for sin and cos.

$$\begin{aligned}\sin(\theta+\phi) &= \sin\theta\cos\phi + \cos\theta\sin\phi\\ \cos(\theta+\phi) &= \cos\theta\cos\phi - \sin\theta\sin\phi.\end{aligned}$$

Now let $\theta = \arg z$ and $\phi = \arg w$. Then

$$\begin{aligned}zw &= |z|(\cos\theta + \mathrm{i}\sin\theta)|w|(\cos\phi + \mathrm{i}\sin\phi)\\ &= |z||w|[(\cos\theta\cos\phi - \sin\theta\sin\phi) + \mathrm{i}(\sin\theta\cos\phi + \cos\theta\sin\phi)]\\ &= |z||w|(\cos(\theta+\phi) + \mathrm{i}\sin(\theta+\phi)).\end{aligned}$$

This implies that (up to a summand of $2k\pi$) the argument of the product is the sum of the arguments of the factors. □

It readily follows that division of complex numbers in polar coordinates can be carried out by dividing the respective absolute values and subtracting the arguments. In particular, we have

$$\arg\frac{1}{z} = -\arg z, \quad \arg\bar{z} = \arg\frac{|z^2|}{z} = \arg\frac{1}{z} = -\arg z.$$

A shorter notation[13] for $\cos\theta + \mathrm{i}\sin\theta$ is $\mathrm{e}^{\mathrm{i}\theta}$. The formula

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$$

is called *Euler's formula*. For the time being we just take it as a definition of the exponential function applied on imaginary numbers. Notice that this definition is compatible with the usual rule for the exponential function.

$$\mathrm{e}^{\mathrm{i}\theta}\,\mathrm{e}^{\mathrm{i}\phi} = \mathrm{e}^{\mathrm{i}(\theta+\phi)}.$$

Moreover, we can define the exponential function for an arbitrary complex input $z = x + \mathrm{i}y$ as

$$\mathrm{e}^{x+\mathrm{i}y} = \mathrm{e}^x \cdot \mathrm{e}^{\mathrm{i}y} = \mathrm{e}^x(\cos y + \mathrm{i}\sin y).$$

We have the remarkable Eulers's identity

$$\mathrm{e}^{\mathrm{i}\pi} = \cos\pi + \mathrm{i}\sin\pi = -1$$

[13] In school mathematics sometimes $\operatorname{cis}\theta$ is used instead of $\mathrm{e}^{\mathrm{i}\theta}$.

which relates the numbers $1, \mathrm{e}, \mathrm{i}, \pi$ to each other.

The multiplication formula has an immediate consequence for powers of complex numbers:

$$z^n = (|z|\,\mathrm{e}^{\mathrm{i}\arg z})^n = |z|^n\,\mathrm{e}^{\mathrm{i}\,n\arg z}\,.$$

This is de Moivre's formula. Again $n\arg z$ needs not to be in the interval $[0, 2\pi)$ even if $\arg z$ was.

For $z = \cos\theta + \mathrm{i}\sin\theta$ this becomes

$$(\cos\theta + \mathrm{i}\sin\theta)^n = \cos n\theta + \mathrm{i}\sin n\theta,$$

which gives the formulae for sin and cos of $n\theta$:

$$\cos n\theta = \mathrm{Re}(\cos\theta + \mathrm{i}\sin\theta)^n$$
$$\sin n\theta = \mathrm{Im}(\cos\theta + \mathrm{i}\sin\theta)^n.$$

Roots

An n-th root of a (complex) number a is defined as a number z such that

$$z^n = a.$$

From de Moivre's formula we find

$$|z|^n\,\mathrm{e}^{\mathrm{i}\,n\arg z} = |a|\,\mathrm{e}^{\mathrm{i}\arg a}\,.$$

This implies that $|z|$ must be the n-th root of the non-negative number $|a|$ in the usual real sense. To determine the argument of z is more subtle due to the ambiguity. We have

$$n\arg z = \arg a + 2k\pi,$$

where k can be any integer. On dividing by n we get

$$\arg z = \frac{1}{n}\arg a + \frac{2k\pi}{n}.$$

The additional summand $\frac{2k\pi}{n}$ is, in general, not an integer multiple of 2π. Therefore we obtain n different roots corresponding to $k = 0, 1, \ldots, n-1$, namely

$$|a|^{\frac{1}{n}}\,\mathrm{e}^{\frac{\mathrm{i}\arg a}{n}},\ |a|^{\frac{1}{n}}\,\mathrm{e}^{\frac{\mathrm{i}\arg a}{n}+\frac{2\pi\,\mathrm{i}}{n}},\ |a|^{\frac{1}{n}}\,\mathrm{e}^{\frac{\mathrm{i}\arg a}{n}+\frac{4\pi\,\mathrm{i}}{n}},\ \ldots,\ |a|^{\frac{1}{n}}\,\mathrm{e}^{\frac{\mathrm{i}\arg a}{n}+\frac{2(n-1)\pi\,\mathrm{i}}{n}}\,.$$

Notice that $|a|^{\frac{1}{n}}\,\mathrm{e}^{\frac{\mathrm{i}\arg a}{n}+\frac{2n\pi\,\mathrm{i}}{n}}, |a|^{\frac{1}{n}}\,\mathrm{e}^{\frac{\mathrm{i}\arg a}{n}+\frac{2(n+1)\pi\,\mathrm{i}}{n}}, \ldots$ do not give new solutions because their arguments differ by 2π from the arguments of $|a|^{\frac{1}{n}}\,\mathrm{e}^{\frac{\mathrm{i}\arg a}{n}},\ |a|^{\frac{1}{n}}\,\mathrm{e}^{\frac{\mathrm{i}\arg a}{n}+\frac{2\pi\,\mathrm{i}}{n}}, \ldots$.

The n-th root of a complex number z is not a function in the usual sense because it does not assign to z a unique output. When we write $\sqrt[n]{z}$ for complex z we mean

the set of all complex numbers w such that $w^n = z$. This is in contrast to the definition of the n-th root of a real number.

♦Example. Compute $\sqrt{\mathrm{i}}$.

We have $|\,\mathrm{i}\,| = 1$ and $\arg \mathrm{i} = \frac{\pi}{2}$. We find the two roots

$$\mathrm{e}^{\frac{\pi}{2}\,\mathrm{i}} = \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}\,\mathrm{i} \text{ and } \mathrm{e}^{(\frac{\pi}{2}+\frac{2\pi}{2})\,\mathrm{i}} = -\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}\,\mathrm{i}\,.$$

♦Example. Compute all n-th roots of 1.

We have $|1| = 1$ and $\arg 1 = 0$. Therefore the roots are

$$\mathrm{e}^{0\,\mathrm{i}} = 1, \quad \mathrm{e}^{\frac{2\pi\,\mathrm{i}}{n}}, \quad \mathrm{e}^{\frac{4\pi\,\mathrm{i}}{n}}, \ldots, \mathrm{e}^{\frac{2(n-1)\pi\,\mathrm{i}}{n}}\,.$$

If we denote $\mathrm{e}^{\frac{2\pi\,\mathrm{i}}{n}} = \varepsilon$ then all roots can be expressed as

$$\varepsilon, \varepsilon^2, \ldots, \varepsilon^n (= 1).$$

♠ *Exercises 46.* Show that $\varepsilon + \varepsilon^2 + \cdots + \varepsilon^n = 0$.
(Hint. Use the formula $\sum_{k=0}^{n} q^n = \frac{q^{n+1}-1}{q-1}$.)

Geometrically, these complex numbers represent n points in the plan, which have distance 1 from the origin. The sectors cut out by two adjacent roots open at an angle $\frac{2\pi}{n}$, so that the n roots form a regular n-gon inscribed in the unit circle.

The triangle inequality

For real numbers we have the inequality

$$|a + b| \leq |a| + |b|.$$

The analogous inequality is also true for complex numbers. An equivalent form is

$$|a - b| \leq |a| + |b|$$

(just replace b by $-b$ and use $|-b| = |b|$.) Now we look at the triangle through the origin and a, b interpreted as points in the plane. Then $|a|$ and $|b|$ are the length of the sides $0a$ and $0b$ respectively, whereas $|a - b|$ is the length of the side ab. The geometric meaning of the inequality above is that in any such triangle the length of the third side cannot be bigger than the sum of the lengths of the other two sides. This is called *triangle inequality*. A formal proof is given below.

Theorem 28. *If $z, w \in \mathbb{C}$ then*

$$|z + w| \leq |z| + |w|.$$

Here equality holds if and only if $\arg z = \arg w$ *(or one of z, w is zero).*

Proof. We have

$$|z+w|^2 = |z|^2+|w|^2+2\operatorname{Re} z\bar{w} \leq |z|^2+|w|^2+2|\operatorname{Re} z\bar{w}| \leq |z|^2+|w|^2+2|z||w| = |z|^2+|w|^2,$$

since

$$\operatorname{Re} z\bar{w} \leq |\operatorname{Re} z\bar{w}| \leq |z\bar{w}| = |z||w|.$$

Equality occurs if $\operatorname{Re} z\bar{w} = |z\bar{w}|$, i.e.

$$|z||w|\,\mathrm{e}^{\mathrm{i}(\arg z - \arg w)} = |z||w|,$$

which requires that $z = 0$ or $w = 0$ or $\arg z = \arg w$. □

A similar argument to the proof above can be used to prove the law of cosines:

We compute

$$\begin{aligned} |a-b|^2 &= (a-b)(\bar{a}-\bar{b}) = |a|^2 + |b|^2 - a\bar{b} - b\bar{a} \\ &= |a|^2 + |b|^2 - 2\operatorname{Re} a\bar{b} \\ &= |a|^2 + |b|^2 - 2\operatorname{Re}|a||b|(\cos(\arg a - \arg b) + \mathrm{i}\sin(\arg a - \arg b)) \\ &= |a|^2 + |b|^2 - 2|a||b|\cos(\arg a - \arg b). \end{aligned}$$

Here, we used $\arg \bar{b} = -\arg b$. In our proof we have implicitly used the addition theorem for sine and cosine.

Final remarks

The methods developed for solving systems of linear equations have been based on the arithmetic properties of the real numbers. We have seen that the complex numbers satisfy the same properties. Therefore the whole theory (including the notion of determinants) carries over to linear equations with complex coefficients. In such case the solutions will be complex as well.

For solving polynomial equations, the set of complex numbers is even more suitable than the set of real numbers. We know that the quadratic equation

$$x^2 + 1 = 0$$

cannot have a real solution. This follows from $x^2 \geq 0$, which implies $x^2 + 1 \geq 1 > 0$. We have seen that this equation has complex solutions, namely, $x = \pm\,\mathrm{i}$.

In fact, any polynomial (of order at least 1) with complex coefficients has complex roots. This is the statements of the Fundamental Theorem of Algebra:

Theorem 29. *Any polynomial equation*

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 = 0$$

where $a_0, \ldots, a_n$ *are arbitrary complex coefficients with* $a_n \neq 0$ *and* $n > 0$ *has at least one complex solution.*

The Fundamental Theorem of Algebra is usually attributed to C.F. Gauss, although the first proof is due to J.-R. Argand. The proof requires techniques beyond this unit. The units on Complex analysis and Topology both feature such proofs.

In the case of quadratic equations the well-known solution formula remains true and delivers a solution even if the discriminant is negative. For example we can solve

$$z^2 + z + 1 = 0$$

by applying the usual quadratic formula

$$\begin{aligned} z &= \frac{-1 + \sqrt{1^2 - 4}}{2} \\ &= \frac{-1 + \sqrt{-3}}{2}. \end{aligned}$$

Clearly the solutions are complex, we need to write them in the standard $a + \mathrm{i}\, b$ format. We note that

$$\sqrt{-3} = \{\sqrt{3}\, \mathrm{e}^{\frac{\pi \mathrm{i}}{2}}, \sqrt{3}\, \mathrm{e}^{\frac{3\pi \mathrm{i}}{2}}\} = \pm \mathrm{i}\, \sqrt{3}.$$

So the solutions to the quadratic are

$$z = \frac{-1 + \sqrt{-3}}{2} = -\frac{1}{2} \pm \mathrm{i}\, \frac{\sqrt{3}}{2}.$$

♦Example. Solve the quadratic equation $z^2 - (2 - 2\,\mathrm{i})z - 1 - 2\,\mathrm{i} = 0$.

We have

$$z_{1,2} = \frac{2 - 2\,\mathrm{i} + \sqrt{4}}{2} = 1 - \mathrm{i} \pm 1 = \{-\mathrm{i}, 2 - \mathrm{i}\}.$$

♠ *Exercises 47.*

1. Express each of the following complex numbers in the form $x + \mathrm{i}\, y$.

(a)	$(2 - i)(3 + 2i)$	(b)	$(6 + 5i)(2 + 7i)$
(c)	$(3 - 2i)^2$	(d)	i^3
(e)	$\dfrac{2 - i}{1 + i}$	(f)	$\dfrac{2}{3 + i} - \dfrac{1 + i}{1 - i}$
(g)	$\dfrac{1 - i}{(2 + i)^2}$	(h)	i^7

2. Show that $\dfrac{1+\sin\theta+i\cos\theta}{1+\sin\theta-i\cos\theta} = \sin\theta + i\cos\theta$.

3. Solve the following equations for z, writing your solution in the form $a+ib$

   (a) $(-1+2i)z - 1 = 3i$

   (b) $z^2 + 2i + 5 = 0$

   (c) $5z^2 - 4z + 1 = 0$.

4. Find all solutions of the equation

$$z^2 = 6 - 8i.$$

5. For each of the following complex numbers write down the complex conjugate and modulus

   (a) $6 + 2i$

   (b) $1 - 3i$

   (c) $\frac{1+i}{\sqrt{2}}$

   (d) $\frac{1}{1+i}$

   (e) $\frac{2-3i}{1-i}$

   (f) i.

6. Let $z_1, z_2 \in \mathbb{C}$. Show that

$$|z_1 + z_2|^2 + |z_1 - z_2|^2 = 2|z_1|^2 + 2|z_2|^2.$$

7. For $z_1, z_2 \in \mathbb{C}$ prove that $\overline{\left(\dfrac{z_1}{z_2}\right)} = \dfrac{\overline{z}_1}{\overline{z}_2}$, for $z_2 \neq 0$.

8. Show that $\arg(-z) = \arg(z) + \pi$.

9. Find a formula for $\sin 3\theta$. (Hint. Use de Moivre's theorem.)

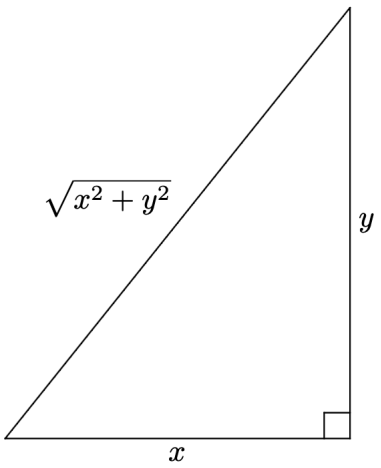
28 The Inner or dot Product

The content of this and the next lecture will be taught in the second year unit Pmth212.

From school mathematics we are used to measuring lengths to line segments and angles to pairs of intersecting lines. How are we to do this in our vector space setting?

In general, lengths and angles between vectors are defined using what is known as an inner product. The inner product is a mapping, which associates to each pair of vectors a scalar. We will not pursue things in such generality here. The interested student will meet inner products in the units Pmth212 and Pmth213.

What we require here is an inner product which leads naturally to the *Euclidean* distance measure via the Pythagoras theorem.



In fact we just about have such an inner product at hand. Take two vectors $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ in $\mathbb{R}^n$, then we can define a map $\mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ as follows

$$\boxed{\boldsymbol{x} \cdot \boldsymbol{y} = x_1y_1 + x_2y_2 + \ldots + x_ny_n.}$$

This *inner product* (in $\mathbb{R}^3$ often called the dot product) of two vectors is easy to remember, it is just the sum of the products of the components of the two vectors.

The inner product has the following obvious properties:

1. It is symmetric, i.e.

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{y} \cdot \boldsymbol{x}$$

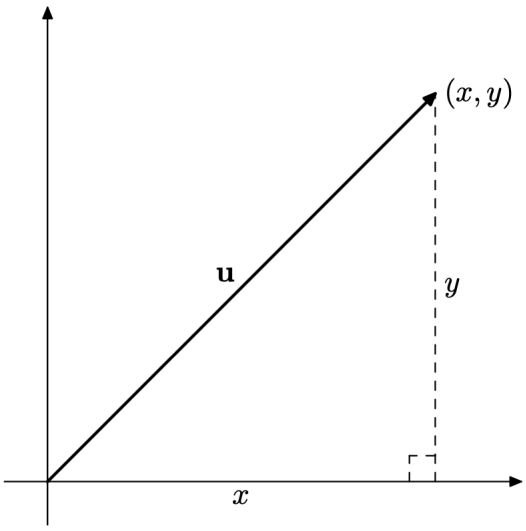
for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

2. It is bilinear, i.e.
$$(\boldsymbol{x}_1 + \boldsymbol{x}_2) \cdot \boldsymbol{y} = \boldsymbol{x}_1 \cdot \boldsymbol{y} + \boldsymbol{x}_2 \cdot \boldsymbol{y},$$
for any $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y} \in \mathbb{R}^n$, and
$$(\alpha \boldsymbol{x}) \cdot \boldsymbol{y} = \alpha(\boldsymbol{x} \cdot \boldsymbol{y})$$
for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. (The analogous linearity property with respect to the second factor holds due to the symmetry.)

3. It is positive definite, i.e.
$$\boldsymbol{x} \cdot \boldsymbol{x} = (x_1)^2 + (x_2)^2 + \cdots + (x_n)^2 \geq 0.$$
and can be equal to 0 only if all $x_1 = x_2 = \cdots = x_n = 0$, i.e. $\boldsymbol{x} = 0$.

The length of a vector is now just given as
$$\begin{aligned} \|\boldsymbol{x}\| &= \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}} \\ &= \sqrt{(x_1)^2 + (x_2)^2 + \ldots + (x_n)^2}. \end{aligned}$$

This is the usual length given by Pythagoras' theorem. For example in $\mathbb{R}^2$



The length of $\boldsymbol{u} = (x, y)$ is just $\|\boldsymbol{u}\| = \sqrt{x^2 + y^2}$.

♦Example. Let $\boldsymbol{i} = (1, 0, 0)$, $\boldsymbol{j} = (0, 1, 0)$ and $\boldsymbol{k} = (0, 0, 1)$ the so-called standard vectors in $\mathbb{R}^3$. Calculate the following

(a) $\boldsymbol{i} \cdot \boldsymbol{i}$ (b) $\boldsymbol{j} \cdot \boldsymbol{j}$ (c) $\boldsymbol{k} \cdot \boldsymbol{k}$ (d) $\boldsymbol{i} \cdot \boldsymbol{j}$
(e) $\boldsymbol{i} \cdot \boldsymbol{k}$ (f) $\boldsymbol{j} \cdot \boldsymbol{k}$

Solution

(a) $\boldsymbol{i} \cdot \boldsymbol{i} = 1^2 + 0^2 + 0^2 = 1$

The length of $\boldsymbol{i}$ is 1.

(b) $\boldsymbol{j} \cdot \boldsymbol{j} = 0^2 + 1^2 + 0^2 = 1$

(c) $\boldsymbol{k} \cdot \boldsymbol{k} = 0^2 + 0^2 + 1^2 = 1$

(d) $\boldsymbol{i} \cdot \boldsymbol{j} = 1 \times 0 + 0 \times 1 + 0 \times 0 = 0$

(e) $\boldsymbol{i} \cdot \boldsymbol{k} = 1 \times 0 + 0 \times 0 + 0 \times 1 = 0$

(f) $\boldsymbol{i} \cdot \boldsymbol{k} = 0 \times 0 + 1 \times 0 + 0 \times 1 = 0$

□

♦Example. Calculate the following

(a) $\boldsymbol{a} \cdot \boldsymbol{b}$ where $\boldsymbol{a} = (1, -2), \boldsymbol{b} = (3, 4)$

(b) $(\boldsymbol{i} + \boldsymbol{j} - \boldsymbol{k}) \cdot (2\boldsymbol{i} + \boldsymbol{j} - \boldsymbol{k})$

Solution

(a)

$$\begin{aligned} \boldsymbol{a} \cdot \boldsymbol{b} &= 1 \times 3 + (-2) \times 4 \\ &= 3 - 8 \\ &= -5 \end{aligned}$$

(b) Either use the earlier example after expanding brackets or think of the vectors in row vector form.

$$\begin{aligned} (\boldsymbol{i} + \boldsymbol{j} - \boldsymbol{k}) \cdot (2\boldsymbol{i} + \boldsymbol{j} - \boldsymbol{k}) &= 1 \times 2 + 1 \times 1 + (-1) \times (-1) \\ &= 2 + 1 + 1 \\ &= 4. \end{aligned}$$

□

We mentioned earlier that inner products also have something to say about the angle between two vectors. We first look at $\mathbb{R}^2$, where the inner product can be

expressed with complex numbers. The inner product of $z = (x, y)$ and $w = (u, v)$ is $z \cdot w = xu + yv = \operatorname{Re} z\bar{w}$. It follows

$$z \cdot w = |z||w| \operatorname{Re} \mathrm{e}^{\mathrm{i} \arg z - \arg w} = |z||w| \cos(\arg z - \arg w) = |z||w| \cos\theta,$$

where $\theta = \arg z - \arg w$ is the angle between the vectors z and w.

The following theorem shows explicitly how the dot product gives you information on the angle between a pair of vectors in any $\mathbb{R}^n$.

Theorem 30. *Let θ be the acute angle between two vectors $\boldsymbol{z}$ and $\boldsymbol{w}$ in $\mathbb{R}^n$. Then*

$$\boldsymbol{z} \cdot \boldsymbol{w} = \|\boldsymbol{z}\| \|\boldsymbol{w}\| \cos\theta. \tag{19}$$

Proof. Consider the triangle OPQ where O is the origin and P, Q are the tips of the vectors $\boldsymbol{z}, \boldsymbol{w}$ with tails placed at O. Then the lengths of the sides are

$$|OP| = \sqrt{\boldsymbol{z} \cdot \boldsymbol{z}}, \quad |OQ| = \sqrt{\boldsymbol{w} \cdot \boldsymbol{w}}, \quad |PQ| = \sqrt{(\boldsymbol{z} - \boldsymbol{w}) \cdot (\boldsymbol{z} - \boldsymbol{w})}.$$

The laws of cosines states

$$|PQ|^2 = |OP|^2 + |OQ|^2 - 2|OP||OQ| \cos\theta.$$

In terms of the inner product this can be rewritten as

$$(\boldsymbol{z} - \boldsymbol{w}) \cdot (\boldsymbol{z} - \boldsymbol{w}) = \boldsymbol{z} \cdot \boldsymbol{z} + \boldsymbol{w} \cdot \boldsymbol{w} - 2\|\boldsymbol{z}\| \|\boldsymbol{w}\| \cos\theta.$$

This yields

$$\begin{aligned} \|\boldsymbol{z}\| \|\boldsymbol{w}\| \cos\theta &= \frac{1}{2}(\boldsymbol{z} \cdot \boldsymbol{z} + \boldsymbol{w} \cdot \boldsymbol{w} - (\boldsymbol{z} - \boldsymbol{w}) \cdot (\boldsymbol{z} - \boldsymbol{w})) \\ &= \frac{1}{2}(\boldsymbol{z} \cdot \boldsymbol{z} + \boldsymbol{w} \cdot \boldsymbol{w} - \boldsymbol{z} \cdot \boldsymbol{z} - \boldsymbol{w} \cdot \boldsymbol{w} + \boldsymbol{z} \cdot \boldsymbol{w} + \boldsymbol{w} \cdot \boldsymbol{z}) \\ &= \frac{1}{2}(\boldsymbol{z} \cdot \boldsymbol{w} + \boldsymbol{w} \cdot \boldsymbol{z}) \\ &= \boldsymbol{z} \cdot \boldsymbol{w} \end{aligned}$$

as required. □

A consequence of the relation (19) is the important *Cauchy-Schwarz inequality*

$$|\boldsymbol{x} \cdot \boldsymbol{y}| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|,$$

which follows from $|\cos\theta| \leq 1$.

Our theorem gives also a nice criterion for determining when two vectors are *orthogonal* (i.e. perpendicular).

Corollary *Two non-zero vectors* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *are orthogonal if and only if* $\boldsymbol{u} \cdot \boldsymbol{v} = 0$.

Proof. The proof is a very simple consequence of the earlier theorem. Note that it is an 'if and only if' proof. Firstly, if $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal then the angle between them, θ, is $\frac{\pi}{2}$ so

$$\begin{aligned} \boldsymbol{u} \cdot \boldsymbol{v} &= \|\boldsymbol{u}\|\|\boldsymbol{v}\| \cos\theta \\ &= \|\boldsymbol{u}\|\|\boldsymbol{v}\| \cos\frac{\pi}{2} = 0. \end{aligned}$$

On the other hand if $\boldsymbol{u} \cdot \boldsymbol{v} = 0$ then, as $\|\boldsymbol{u}\| \neq 0$ and $\|\boldsymbol{v}\| \neq 0$, we have $\cos\theta = 0$. As θ is the acute angle between $\boldsymbol{u}$ and $\boldsymbol{v}, \theta = \frac{\pi}{2}$. The vectors are orthogonal. □

♦Example. Find the angle between the following two lines

$$\begin{aligned} OP: &\quad \text{joining 0 to } (1,1,2) \\ OQ: &\quad \text{joining 0 to } (0,1,1,). \end{aligned}$$

Solution. We have

$$\begin{aligned} \overrightarrow{OP} &= (1,1,2)(= \boldsymbol{i} + \boldsymbol{j} + 2\boldsymbol{k}) \\ \overrightarrow{OQ} &= (0,1,1)(= \boldsymbol{j} + \boldsymbol{k}). \end{aligned}$$

So $\overrightarrow{OP} \cdot \overrightarrow{OQ} = 1 \times 0 + 1 \times 1 + 2 \times 1 = 3$.

$$\begin{aligned} \text{Also,} \qquad |\overrightarrow{OP}| &= \sqrt{1^2 + 1^2 + 2^2} = \sqrt{6} \text{ and} \\ |\overrightarrow{OQ}| &= \sqrt{0^2 + 1^2 + 1^2} = \sqrt{2}. \end{aligned}$$

If θ is the angle between $\overrightarrow{OP}$ and $\overrightarrow{OQ}$ then we have

$$\overrightarrow{OP} \cdot \overrightarrow{OQ} = |\overrightarrow{OP}||\overrightarrow{OQ}| \cos\theta.$$

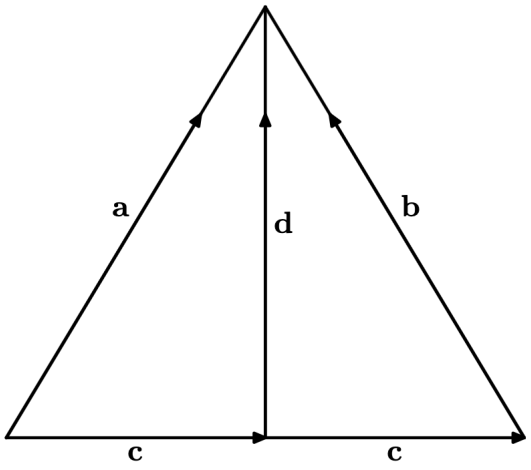
So,

$$\begin{aligned} 3 &= \sqrt{6} \cdot \sqrt{2} \cos\theta \\ \text{i.e.} \quad \cos\theta &= \frac{3}{\sqrt{6}\sqrt{2}} = \frac{3}{\sqrt{12}} \\ &= \frac{3}{2\sqrt{3}} \\ &= \frac{\sqrt{3}}{2} \end{aligned}$$

Hence the angle θ is $\frac{\pi}{6}$ or 30°. □

♦Example. Use vectors to prove that the median drawn from the vertex made by the equal sides of an isosceles triangle is perpendicular to the third side of the triangle.

Solution. Let $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ and $\boldsymbol{d}$ be as shown



Notice that the median $\boldsymbol{d}$ bisects the base of the isosceles triangle represented by $2\boldsymbol{c}$. We use the vector rule of addition

$$\begin{aligned} \boldsymbol{a} &= \boldsymbol{c} + \boldsymbol{d} \text{ and} \\ \boldsymbol{d} &= \boldsymbol{c} + \boldsymbol{b}. \end{aligned}$$

From this pair of equations we deduce that

$$\begin{aligned} \boldsymbol{d} &= \frac{1}{2}(\boldsymbol{a} + \boldsymbol{b}) \text{ and} \\ \boldsymbol{c} &= \frac{1}{2}(\boldsymbol{a} - \boldsymbol{b}). \end{aligned}$$

$$\begin{aligned} \text{So that } \boldsymbol{d} \cdot \boldsymbol{c} &= \frac{1}{4}(\boldsymbol{a} \cdot \boldsymbol{a} + \boldsymbol{a} \cdot \boldsymbol{b} - \boldsymbol{b} \cdot \boldsymbol{a} - \boldsymbol{b} \cdot \boldsymbol{b}) \\ &= \frac{1}{4}(\|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2), \end{aligned}$$

since $\|\boldsymbol{u}\| = \sqrt{\boldsymbol{u} \cdot \boldsymbol{u}}$ for any vector $\boldsymbol{u}$. However, as the triangle is isosceles $\|\boldsymbol{a}\| = \|\boldsymbol{b}\|$ — the sides given by $\boldsymbol{a}$ and $\boldsymbol{b}$ have equal length. Thus,

$$\boldsymbol{d} \cdot \boldsymbol{c} = 0.$$

We conclude that the median (represented by $\boldsymbol{d}$) is perpendicular to the base (represented by $2\boldsymbol{c}$). □

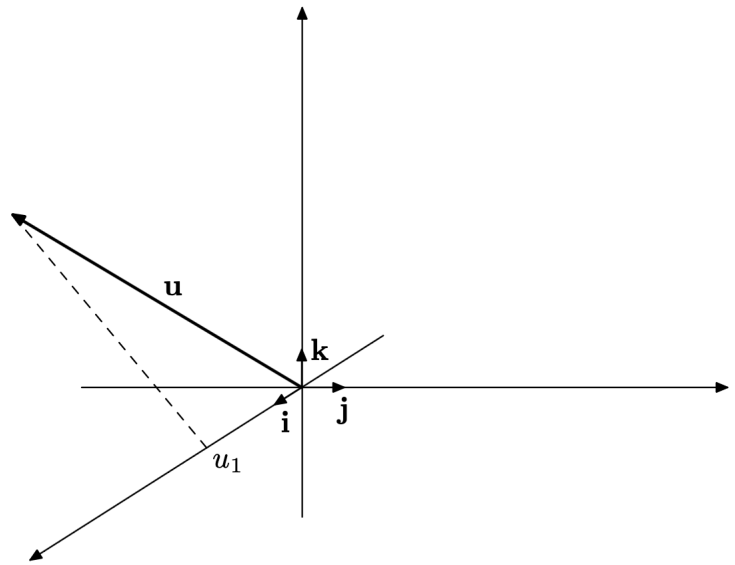
Orthogonal Projection

In $\mathbb{R}^3$ our basis vectors $\boldsymbol{i}, \boldsymbol{j}$ and $\boldsymbol{k}$ are mutually orthogonal (each one is perpendicular to the other two), unit vectors (they all have length 1).

A general vector $\boldsymbol{u}$ in $\mathbb{R}^3$ can be written as

$$\boldsymbol{u} = u_1\boldsymbol{i} + u_2\boldsymbol{j} + u_3\boldsymbol{k},$$

where the u_i are the components of $\boldsymbol{u}$ with respect to the basis $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$. We can think of u_1 as the component of the projection onto $\boldsymbol{i}$ of $\boldsymbol{u}$ — in fact it is the perpendicular or orthogonal projection.



In the same sense u_2 and u_3 are the projections onto $\boldsymbol{j}$ and $\boldsymbol{k}$ respectively.

We now want to use our inner product to characterise such projections. We note that

$$\begin{aligned} u_1 &= \boldsymbol{i} \cdot \boldsymbol{u} \\ u_2 &= \boldsymbol{j} \cdot \boldsymbol{u} \\ \text{and} \quad u_3 &= \boldsymbol{k} \cdot \boldsymbol{u}. \end{aligned}$$

So we find the component u_1 of the projection of $\boldsymbol{u}$ onto $\boldsymbol{i}$ by simply taking the dot product.

Let's generalise. Let $\boldsymbol{e}$ be any vector, suppose we want to find the component of the projection of $\boldsymbol{u}$ onto $\boldsymbol{e}$. First, we need to make $\boldsymbol{e}$ into a **unit vector**, i.e a vector of length 1. We are interested only in the component of $\boldsymbol{u}$ in the direction of $\boldsymbol{e}$. The unit vector in the $\boldsymbol{e}$ direction is

$$\hat{\boldsymbol{e}} = \frac{\boldsymbol{e}}{\|\boldsymbol{e}\|}.$$

Note that, $\hat{\boldsymbol{e}} \cdot \hat{\boldsymbol{e}} = \frac{\boldsymbol{e}.\boldsymbol{e}}{\|\boldsymbol{e}\|^2} = \frac{\|\boldsymbol{e}\|^2}{\|\boldsymbol{e}\|^2} = 1$, so $\hat{\boldsymbol{e}}$ has indeed got unit length.

The required component of projection is now simply

$$\hat{\boldsymbol{e}} \cdot \boldsymbol{u}.$$

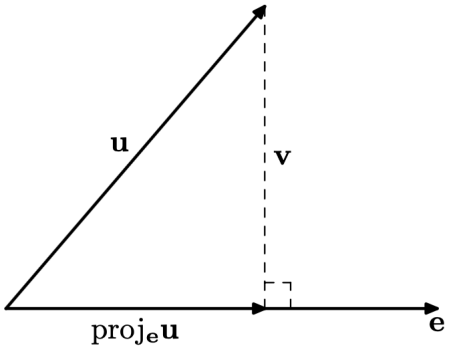
The projection of the vector $\boldsymbol{u}$ onto $\boldsymbol{e}$ is then the vector of length $\hat{\boldsymbol{e}} \cdot \boldsymbol{u}$ in the $\boldsymbol{e}$ i.e. $\hat{\boldsymbol{e}}$, direction.

The orthogonal projection of $\boldsymbol{u}$ *onto a nonzero vector* $\boldsymbol{e}$ *is* $$\mathit{proj}_{\boldsymbol{e}}\boldsymbol{u} = (\hat{\boldsymbol{e}} \cdot \boldsymbol{u})\hat{\boldsymbol{e}},$$ *a vector of length* $(\hat{\boldsymbol{e}} \cdot \boldsymbol{u})$ *in the* $\hat{\boldsymbol{e}}$ *direction.*

As $\hat{\boldsymbol{e}} = \frac{\boldsymbol{e}}{\|\boldsymbol{e}\|}$ this can also be written as

$$\text{proj}_{\boldsymbol{e}}\boldsymbol{u} = \frac{\boldsymbol{e} \cdot \boldsymbol{u}}{\|\boldsymbol{e}\|^2}\boldsymbol{e}.$$

Note $|\text{proj}_{\boldsymbol{e}}\boldsymbol{u}| = |\hat{\boldsymbol{e}} \cdot \boldsymbol{u}| = \frac{|\boldsymbol{e} \cdot \boldsymbol{u}|}{\|\boldsymbol{e}\|}$.



Notice that the vector labelled $\boldsymbol{v}$ is orthogonal to $\boldsymbol{e}$ (and $\text{proj}_{\boldsymbol{e}}\boldsymbol{u}$). It is known as the *component of* $\boldsymbol{u}$ *orthogonal to* $\boldsymbol{e}$. In fact, using the vector addition rule

$$\boldsymbol{v} = \boldsymbol{u} - \text{ proj}_{\boldsymbol{e}}\boldsymbol{v}.$$

♦Example. Find the orthogonal projection and component orthogonal to it for

$$\boldsymbol{u} = \boldsymbol{i} + \boldsymbol{j} + \boldsymbol{k}$$

in the direction of $\boldsymbol{e} = \boldsymbol{i} + \boldsymbol{j}$.

Solution. Unit vector in $\boldsymbol{e}$ direction,

$$\hat{\boldsymbol{e}} = \frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} = \frac{\boldsymbol{i} + \boldsymbol{j}}{\sqrt{1^2 + 1^2}} = \frac{1}{\sqrt{2}}(\boldsymbol{i} + \boldsymbol{j}).$$

$$\begin{aligned} \text{Then } \ \hat{\boldsymbol{e}} \cdot \boldsymbol{u} &= \frac{1}{\sqrt{2}}(1+1) = \sqrt{2}, \ \text{ so that} \\ \text{proj}_{\boldsymbol{e}}\boldsymbol{u} &= \frac{\sqrt{2}}{(\sqrt{2})^2}(\boldsymbol{i}+\boldsymbol{j}) \\ &= \frac{1}{\sqrt{2}}(\boldsymbol{i}+\boldsymbol{j}). \end{aligned}$$

The vector orthogonal to $\text{proj}_{\boldsymbol{e}}\boldsymbol{u}$ is

$$\boldsymbol{u} - \ \text{proj}_{\boldsymbol{e}}\boldsymbol{u} = (1 - \frac{1}{\sqrt{2}})\boldsymbol{i} + (1 - \frac{1}{\sqrt{2}})\boldsymbol{j} + \boldsymbol{k}.$$

□

♠ *Exercises 48.*

1. In each part find the inner product of the vectors and the cosine of the angle between them.

   (a) $\boldsymbol{u} = \boldsymbol{i} + \boldsymbol{j}, \boldsymbol{v} = \boldsymbol{i} - \boldsymbol{j}$

   (b) $\boldsymbol{u} = (1, -1), \boldsymbol{v} = (2, -3)$

   (c) $\boldsymbol{u} = 2\boldsymbol{i} - \boldsymbol{j} + \boldsymbol{k}, \boldsymbol{v} = -\boldsymbol{i} + 3\boldsymbol{j} + \boldsymbol{k}.$

   (d) $\boldsymbol{u} = \boldsymbol{i} + \boldsymbol{j} - \boldsymbol{k}, \boldsymbol{v} = 3\boldsymbol{i} - \boldsymbol{k}.$

2. Use vectors to show that $A(2, -1, 1), B(3, 2, -1)$ and $C(7, 0, -2)$ are vertices of a right angled triangle.

3. In each part find the orthogonal projection on $\boldsymbol{e} = \boldsymbol{i} + \boldsymbol{j} - \boldsymbol{k}$ and also the vector component orthogonal to $\boldsymbol{e}$.

   (a) $\boldsymbol{u} = 4\boldsymbol{i} - \boldsymbol{j} + 7\boldsymbol{k}$ (b) $\boldsymbol{u} = \boldsymbol{i} + \boldsymbol{j} + \boldsymbol{k}$

   (c) $\boldsymbol{u} = \boldsymbol{i} - 2\boldsymbol{j}$ (d) $\boldsymbol{u} = -\boldsymbol{i} + \boldsymbol{j}.$

*4. Use vectors to prove that the angle inscribed in a semi-circle is a right angle.

Oriented area

Let $\boldsymbol{z}$ and $\boldsymbol{w}$ be vectors in the two-dimensional plane, which can again be expressed through complex numbers $z = x + \mathrm{i}\, y, w = u + \mathrm{i}\, v$. These vectors span a triangle with vertices $0, z, w$. The area of this triangle is

$$\frac{1}{2}|z||w| \sin\theta$$

where $\theta \in [0, \pi]$ is the angle between the vectors $\boldsymbol{z}$ and $\boldsymbol{w}$. We have

$$|z||w| \sin\theta = -\operatorname{Im} z\bar{w} = xv - yu = \begin{vmatrix} x & u \\ y & v \end{vmatrix}.$$

The expression

$$\frac{1}{2}|z||w|\sin\theta = \frac{1}{2}\begin{vmatrix} x & u \\ y & v \end{vmatrix}$$

can become positive, zero or negative if θ is interpreted as the angle between $\boldsymbol{z}$ and $\boldsymbol{w}$ in counterclockwise orientation. The sign changes if we swap the two vectors. Thus the expression includes two pieces of information: the area (as absolute value) and the orientation (as sign). We call the number

$$\frac{1}{2}\begin{vmatrix} x & u \\ y & v \end{vmatrix}$$

the oriented area of the triangle spanned by the two vectors $\boldsymbol{z}$ and $\boldsymbol{w}$. The traditional area is then just the absolute value of that expression.

The expression

$$\begin{vmatrix} x & u \\ y & v \end{vmatrix}$$

can be interpreted as a scalar-valued product of the vectors $\boldsymbol{z} = (x, y)$ and $\boldsymbol{w} = (u, v)$ in $\mathbb{R}^2$, similar to the dot product. We denote this product by $\boldsymbol{z} \wedge \boldsymbol{w}$. This wedge product is only defined for vectors in $\mathbb{R}^2$. It is bilinear, like the dot product, however it is antisymmetric

$$\boldsymbol{z} \wedge \boldsymbol{w} = -\boldsymbol{w} \wedge \boldsymbol{z}.$$

The wedge product gives the oriented area of the parallelogram spanned by the vectors $\boldsymbol{z}, \boldsymbol{w}$.

Remark. Felix Klein, a prominent mathematician of the 19th and early 20th century, initiated a revision of the school curriculum in mathematics in his time. Many of his thoughts are summarised in the book "Elementary mathematics from an Advanced Standpoint". This book was meant as a contribution to the training of maths teachers. One of the first topics in geometry is the oriented area. He shows that many statements in geometry become more natural when area is replaced by oriented area. E.g., it is a standard procedure to compute the area of a convex polygon by dissecting it into triangles with one common vertex inside the polygon and the other vertices being adjacent vertices of the polygon. This procedure will also work for non-convex polygons and an arbitrary common vertex if oriented areas are used. This idea has far-reaching consequences for computing the area of a curvilinear shape. This will be studied in MTHS130.

♦Example. Compute the area of the pentagon with vertices $(-1,-1)$, $(1,-1)$, $(0,0)$, $(0,1)$, $(-1,0)$.

This pentagon is not convex but using orientated areas we can express the area

as

$$\begin{aligned} A &= \frac{1}{2}\left(\begin{vmatrix} -1 & 1 \\ -1 & -1 \end{vmatrix} + \begin{vmatrix} 1 & 0 \\ -1 & 0 \end{vmatrix} + \begin{vmatrix} 0 & 0 \\ 0 & 1 \end{vmatrix} + \begin{vmatrix} 0 & -1 \\ 1 & 0 \end{vmatrix} + \begin{vmatrix} -1 & -1 \\ 0 & -1 \end{vmatrix}\right) \\ &= \frac{1}{2}(2+0+0+1+1) = 2 \end{aligned}$$

Another application of the wedge product $\boldsymbol{u} \wedge \boldsymbol{v}$ is a test for linear dependence. We have

$$\boldsymbol{u} \wedge \boldsymbol{v} = 0$$

if and only if $\sin\theta = 0$ or $\boldsymbol{u} = \mathbf{0}$ or $\boldsymbol{v} = \mathbf{0}$. This corresponds exactly to the situation when $\boldsymbol{u}$ and $\boldsymbol{v}$ are linearly dependent.

29 The Cross Product

For our final lecture on vectors and vector spaces we want to examine a notion which is very specific to vectors in $\mathbb{R}^3$. This is the *cross product*. Although there are generalisations of the vector product to higher dimensional vector spaces they require more technical machinery, only in $\mathbb{R}^3$ does the vector product have a natural definition within the vector space itself.

What we want to do is to define a "product" of two (non-parallel) vectors which produces a new vector orthogonal (perpendicular) to the original pair. Here is our definition.

If $\boldsymbol{u} = u_1\boldsymbol{i} + u_2\boldsymbol{j} + u_3\boldsymbol{k}$ *and* $\boldsymbol{v} = v_1\boldsymbol{i} + v_2\boldsymbol{j} + v_3\boldsymbol{k}$ *are two vectors in* $\mathbb{R}^3$ *then the cross product* $\boldsymbol{u} \times \boldsymbol{v}$ *is the vector defined by*

$$\boldsymbol{u} \times \boldsymbol{v} = (u_2v_3 - u_3v_2)\boldsymbol{i} - (u_1v_3 - u_3v_1)\boldsymbol{j} + (u_1v_2 - u_2v_1)\boldsymbol{k}.$$

There are in fact deeper mathematical reasons why we would choose such a bizarre looking definition. We'll just have to accept it for the time being. At least until you have done some more mathematics. What we want to do is explore some of the consequences of the definition. The cross product became popular initially because of its great utility in applications to fluid mechanics and electromagnetism.

Our definition of the cross product is, as it stands, difficult to use and remember. However, if you look at the three components of $\boldsymbol{u} \times \boldsymbol{v}$, i.e. $(u_2v_3 - u_3v_2)$, $-(u_1v_3 - u_3v_1)$ and $(u_1v_2 - u_2v_1)$, you should be reminded of the determinant! You can verify for yourself the following formula.

$$\boldsymbol{u} \times \boldsymbol{v} = \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}.$$

In practice this is how one remembers the cross product definition.

♦Example. Calculate $\boldsymbol{u} \times \boldsymbol{v}$ where $\boldsymbol{u} = \boldsymbol{i} - \boldsymbol{j} + \boldsymbol{k}$ and $\boldsymbol{v} = 2\boldsymbol{i} + 3\boldsymbol{j} - \boldsymbol{k}$. Verify that $\boldsymbol{u} \times \boldsymbol{v}$ is orthogonal to $\boldsymbol{u}$ and $\boldsymbol{v}$.

Solution.

$$\begin{aligned} \boldsymbol{u}\times\boldsymbol{v} &= \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ 1 & -1 & 1 \\ 2 & 3 & -1 \end{vmatrix} \\ &= [(-1)\times(-1) - 3\times 1]\boldsymbol{i} - [1\times(-1) - 2\times 1]\boldsymbol{j} + [1\times 3 - 2\times(-1)]\boldsymbol{k} \\ \text{i.e. } \boldsymbol{u}\times\boldsymbol{v} &= -2\boldsymbol{i} + 3\boldsymbol{j} + 5\boldsymbol{k}. \end{aligned}$$

To check the orthogonality of $\boldsymbol{u}$ and $\boldsymbol{v}$ with $\boldsymbol{u}\times\boldsymbol{v}$ we need to calculate the angle between $\boldsymbol{u}$ and $\boldsymbol{u}\times\boldsymbol{v}$; and, $\boldsymbol{v}$ and $\boldsymbol{u}\times\boldsymbol{v}$. We use the inner product formula. Let θ be the angle between $\boldsymbol{u}$ and $\boldsymbol{u}\times\boldsymbol{v}$. Then,

$$\begin{aligned} \cos\theta &= \frac{\boldsymbol{u}\cdot(\boldsymbol{u}\times\boldsymbol{v})}{\|\boldsymbol{u}\|\|\boldsymbol{u}\times\boldsymbol{v}\|} \\ &= \frac{1\times(-2) + (-1)\times 3 + 1\times 5}{\sqrt{1^2+(-1)^2+1^2}\sqrt{(-2)^2+3^2+5^2}} \\ &= 0. \end{aligned}$$

So $\cos\theta_1 = 0$ and $\theta_1 = \frac{\pi}{2}$, $\boldsymbol{u}$ is orthogonal to $\boldsymbol{u}\times\boldsymbol{v}$. □

It is worth noting at this point the differences between the scalar and cross products.

- The inner (or dot) product is defined on any $\mathbb{R}^n$. The cross product is defined only in $\mathbb{R}^3$.
- The inner product produces a scalar, i.e. $\boldsymbol{u}\cdot\boldsymbol{v}$ is a scalar. The cross product produces a vector, i.e. $\boldsymbol{u}\times\boldsymbol{v}$ is a vector.

Properties of the Cross Product

We summarise the main properties of the cross product in the following theorem.

Theorem 31. *If $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w}$ are any vectors in $\mathbb{R}^3$ and λ is any scalar, then*

1. $\boldsymbol{u}\times\boldsymbol{v} = -(\boldsymbol{v}\times\boldsymbol{u})$ *(anti-commutativity)*
2. $\boldsymbol{u}\times\boldsymbol{u} = \boldsymbol{0}$ *(this is actually a consequence of 1.)*
3. $\lambda(\boldsymbol{u}\times\boldsymbol{v}) = (\lambda\boldsymbol{u})\times\boldsymbol{v} = \boldsymbol{u}\times(\lambda\boldsymbol{v})$
4. $\boldsymbol{w}\times(\boldsymbol{u}+\boldsymbol{v}) = \boldsymbol{w}\times\boldsymbol{u} + \boldsymbol{w}\times\boldsymbol{v}$ *(Properties 3. and 4. are called bilinearity.)*

5. $\boldsymbol{u} \cdot (\boldsymbol{u} \times \boldsymbol{v}) = 0$ *and* $\boldsymbol{v} \cdot (\boldsymbol{u} \times \boldsymbol{v}) = 0$.

6. $\boldsymbol{u} \times (\boldsymbol{v} \times \boldsymbol{w}) - (\boldsymbol{u} \times \boldsymbol{v}) \times \boldsymbol{w} = \boldsymbol{v} \times (\boldsymbol{u} \times \boldsymbol{w})$. *This property is called Jacobi identity. It shows that the cross product is not associative.*

7. $\boldsymbol{u} \times (\boldsymbol{v} \times \boldsymbol{w}) = \boldsymbol{v}(\boldsymbol{u} \cdot \boldsymbol{w}) - \boldsymbol{w}(\boldsymbol{u} \cdot \boldsymbol{v})$.

Proof.

1. follows from the determinantal formula for $\boldsymbol{u} \times \boldsymbol{v}$ — interchange the rows of the determinant to create $\boldsymbol{v} \times \boldsymbol{u}$, but interchanging rows of a determinant multiplies the determinant by -1.

2, 3, and 4 also follow easily from the determinant formula. They are left as an exercise.

5. says that both $\boldsymbol{u}$ and $\boldsymbol{v}$ are perpendicular to $\boldsymbol{u} \times \boldsymbol{v}$. The proof is easy, following from the general formulae for the dot and cross products.

6. A direct verification is possible but rather tedious. Using bilinearity one can reduce the problem to verifying the identity just for combinations of the basis vectors $\boldsymbol{i}$, $\boldsymbol{j}$ and $\boldsymbol{k}$. Moreover, if one basis vector appears twice, one of the products vanish and the other two become identical, thus the identity holds. Using the symmetry of the identity it is enough to prove it for $\boldsymbol{u} = \boldsymbol{i}$, $\boldsymbol{v} = \boldsymbol{j}$ and $\boldsymbol{w} = \boldsymbol{k}$. In this case all three products vanish.

7. We leave this as a (challenging) exercise. □

Note that the cross product anti–commutes, i.e. $\boldsymbol{u} \times \boldsymbol{v} = -\boldsymbol{v} \times \boldsymbol{u}$, and is not associative. This is quite unlike ordinary multiplication and the inner product.

♦Example. The vectors $\boldsymbol{i}, j$ and $\boldsymbol{k}$ are mutually orthogonal unit vectors show that

$$\boldsymbol{i} \times \boldsymbol{j} = \boldsymbol{k}, \quad \boldsymbol{j} \times \boldsymbol{k} = \boldsymbol{i} \text{ and } \boldsymbol{k} \times \boldsymbol{i} = \boldsymbol{j}$$

Solution.

We will show $\boldsymbol{i} \times \boldsymbol{j} = \boldsymbol{k}$. Note

$$\begin{aligned} \boldsymbol{j} &= 1\boldsymbol{i} + 0\boldsymbol{j} + 0\boldsymbol{k} \\ \text{and} \quad \boldsymbol{j} &= 0\boldsymbol{i} + 1\boldsymbol{j} + 0\boldsymbol{k}, \\ \boldsymbol{i} \times \boldsymbol{j} &= \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix} \\ &= \begin{vmatrix} 0 & 0 \\ 1 & 0 \end{vmatrix} \boldsymbol{i} - \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} \boldsymbol{j} + \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} \boldsymbol{k} \\ &= \boldsymbol{k}. \end{aligned}$$

The other formulae follow in a similar manner. □

You will recall that we were able to calculate the scalar product in terms of the lengths of the vectors and the angle between them. Is a similar type of formula valid for the cross product? The following theorem provides the answer.

Theorem 32. *Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be vectors in $\mathbb{R}^3$ with θ being the smaller angle between them. Then*

$$\|\boldsymbol{u} \times \boldsymbol{v}\| = \|\boldsymbol{u}\|\|\boldsymbol{v}\| \sin\theta.$$

Proof. We have

$$\begin{aligned} \cos\theta &= \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}, \\ \text{so} \qquad \sin\theta &= \sqrt{1 - \cos^2\theta} \\ &= \sqrt{1 - \left(\frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}\right)^2}. \end{aligned}$$

Then,

$$\begin{aligned} \|\boldsymbol{u}\|\|\boldsymbol{v}\| \sin\theta &= \|\boldsymbol{u}\|\|\boldsymbol{v}\| \sqrt{1 - \frac{(\boldsymbol{u} \cdot \boldsymbol{v})^2}{\|\boldsymbol{u}\|^2\|\boldsymbol{u}\|^2}} \\ &= \sqrt{\|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2 - (\boldsymbol{u} \cdot \boldsymbol{v})^2} \\ &= \sqrt{(u_1^2 + u_2^2 + u_3^2)(v_1^2 + v_2^2 + v_3^2) - (u_1v_1 + u_2u_2 + u_3v_3)^2} \\ &= \sqrt{(u_2v_3 - u_3v_2)^2 + (u_1v_3 - u_3v_1)^2 + (u_1v_2 - u_2v_1)^2} \\ &= \|\boldsymbol{u} \times \boldsymbol{v}\|. \qquad \square \end{aligned}$$

Notice in proving our formula we derived the following interesting formula

$$\|\boldsymbol{u} \times \boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2 - (\boldsymbol{u} \cdot \boldsymbol{v})^2.$$

We have the following easy corollary.

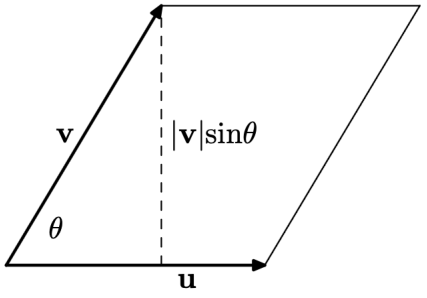
Corollary. *Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be two non-zero vectors in $\mathbb{R}^3$. Then*

(a) The area of the parallelogram with sides $\boldsymbol{u}, \boldsymbol{v}$ *is* $\|\boldsymbol{u} \times \boldsymbol{v}\|$.

(b) $\boldsymbol{u} \times \boldsymbol{v} = \boldsymbol{0}$ *if and only if* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *are parallel.*

Proof.

(a)



$$\begin{aligned} \text{Area of the parallelogram} &= \text{(base)} \times \text{(perpendicular height)} \\ &= \|\boldsymbol{u}\|\|\boldsymbol{v}\| \sin\theta \\ &= \|\boldsymbol{u} \times \boldsymbol{v}\|. \end{aligned}$$

(b) $\boldsymbol{u}$ and $\boldsymbol{v}$ are assumed nonzero so $\|\boldsymbol{u}\| \neq 0$ and $\|\boldsymbol{v}\| \neq 0$. So we have

$$\boldsymbol{u} \times \boldsymbol{v} = \boldsymbol{0} \text{ if and only if } \sin\theta = 0.$$

This is true if and only if $\theta = 0$ or $\theta = \pi$. So $\boldsymbol{u} \times \boldsymbol{v} = \boldsymbol{0}$ if and only if $\boldsymbol{u}$ and $\boldsymbol{v}$ are parallel (or anti-parallel).

□

♦Example. Find the area of the triangle whose vertices are

$$P_1(1,1,1), \quad P_2(-1,1,0) \text{ and } P_3(0,2,1).$$

Solution.

The area of the triangle A, say, is half the area of the parallelogram determined by vectors

$$\begin{aligned} \overrightarrow{P_1P_2} &= (-1-1, 1-1, 0-1) \\ &= -2\boldsymbol{i} - \boldsymbol{k} \\ \text{and} \quad \overrightarrow{P_3P_2} &= (-1-0, 1-2, 0-1) \\ &= -\boldsymbol{i} - \boldsymbol{j} - \boldsymbol{k}. \end{aligned}$$

So

$$A = \frac{1}{2}\left|\overrightarrow{P_1P_2} \times \overrightarrow{P_3P_2}\right|.$$

Now,

$$\begin{aligned}
\overrightarrow{P_1P_2} \times \overrightarrow{P_3P_2} &= \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ -2 & 0 & -1 \\ -1 & -1 & -1 \end{vmatrix} \\
&= -\boldsymbol{i} - \boldsymbol{j} + 2\boldsymbol{k}. \\
\text{Then } \ A = \frac{1}{2}|-\boldsymbol{i} - \boldsymbol{j} + 2\boldsymbol{k}| &= \frac{1}{2}\sqrt{(-1)^2 + (-1)^2 + 2^2} \\
&= \frac{\sqrt{6}}{2} = \sqrt{\frac{3}{2}}.
\end{aligned}$$

□

The Mixed Triple Product

The fact that the cross product produces a vector means that we can define a product of three vectors using the cross product and the inner product.

If $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w}$ are vectors in $\mathbb{R}^3$ we define the mixed triple product of $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w}$ as the scalar

$$\boldsymbol{u} \cdot (\boldsymbol{v} \times \boldsymbol{w}).$$

We can give a rather nice formula for the triple product in terms of a determinant.

$$\begin{aligned}
\boldsymbol{v} \times \boldsymbol{w} &= \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \\
&= \begin{vmatrix} v_2 & v_3 \\ w_2 & w_3 \end{vmatrix}\boldsymbol{i} - \begin{vmatrix} v_1 & v_3 \\ w_1 & w_3 \end{vmatrix}\boldsymbol{j} + \begin{vmatrix} v_1 & v_2 \\ w_1 & w_2 \end{vmatrix}\boldsymbol{k}
\end{aligned}$$

So,

$$\begin{aligned}
\boldsymbol{u} \cdot (\boldsymbol{v} \times \boldsymbol{w}) &= u_1\begin{vmatrix} v_2 & v_3 \\ w_2 & w_3 \end{vmatrix} - u_2\begin{vmatrix} v_1 & v_3 \\ w_1 & w_3 \end{vmatrix} + u_3\begin{vmatrix} v_1 & v_2 \\ w_1 & w_2 \end{vmatrix} \\
&= \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}.
\end{aligned}$$

$$\boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w}) = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}.$$

♦Example. Calculate the triple product $\boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w})$ if $\boldsymbol{u} = \boldsymbol{i}+\boldsymbol{j}-\boldsymbol{k}, \boldsymbol{v} = 2\boldsymbol{i}-\boldsymbol{j}$ and $\boldsymbol{w} = -\boldsymbol{i}+3\boldsymbol{k}$.

Solution.

$$\begin{aligned} \boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w}) &= \begin{vmatrix} 1 & 1 & -1 \\ 2 & -1 & 0 \\ -1 & 0 & 3 \end{vmatrix} \\ &= -8. \end{aligned}$$

□

The scalar triple product has a geometrical interpretation as the oriented volume of the parallelepiped defined by the three vectors $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w}$. This can be seen as follows,

$$\begin{aligned} \text{Volume of parallelepiped, } V &= \text{(Area of base) } \times \text{ (perpendicular height)} \\ &= |\boldsymbol{v}\times\boldsymbol{w}|h. \end{aligned}$$

Where h is the perpendicular height,

$$\begin{aligned} h = \left\|\text{proj}_{\boldsymbol{v}\times\boldsymbol{w}}\boldsymbol{u}\right\| &= \frac{|\boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w})|}{\|\boldsymbol{v}\times\boldsymbol{w}\|^2}\|\boldsymbol{v}\times\boldsymbol{w}\| \\ &= \frac{|\boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w})|}{\|\boldsymbol{v}\times\boldsymbol{w}\|}. \end{aligned}$$

So we have

$$\begin{aligned} V &= |\boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w})|, \\ \text{or} \qquad V &= \pm\boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w}). \end{aligned}$$

As in the case of the oriented area, the triple product gives additional information of a + or − sign. The + indicated that the vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ follow the right hand rule, e.g. the thumb of the right hand points in $\boldsymbol{w}$ direction when your fingers move $\boldsymbol{u}$ into $\boldsymbol{v}$.

The oriented area of a parallelogram can be derived from the triple product in the following way: Let $\boldsymbol{u} = (u_1, u_2, 0)$ and $\boldsymbol{v} = (v_1, v_2, 0)$ be two vectors in the x, y-plane. Then the oriented area of the parallelogram formed by $\boldsymbol{u}, \boldsymbol{v}$ is equal to

the oriented volume of the parallelepiped formed by $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w} = (0, 0, 1)$. Notice that $\boldsymbol{w}$ has length 1 and is perpendicular to the x, y-plane. Now

$$\boldsymbol{u} \cdot (\boldsymbol{v} \times \boldsymbol{w}) = \begin{vmatrix} u_1 & u_2 & 0 \\ v_1 & v_2 & 0 \\ 0 & 0 & 1 \end{vmatrix} = \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix}.$$

♦Example. Verify the parallelepiped volume formula by calculating the volume of the unit cube with sides $\boldsymbol{i}, \boldsymbol{j}$ and $\boldsymbol{k}$.

Solution

$$\begin{aligned} \text{Volume} &= |\boldsymbol{i} \cdot (\boldsymbol{j} \times \boldsymbol{k})| \\ &= |\boldsymbol{i} \cdot \boldsymbol{i}| \\ &= 1. \end{aligned}$$

□

The triple product can also be used as a test of linear dependence of three vectors in $\mathbb{R}^3$. The vectors

$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}, \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

are linearly dependent if an only if they lie in the same plane, that is the volume of the parallelepiped spanned is zero. Therefore, the three vectors are linearly dependent if and only if the determinant

$$\begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = 0.$$

♠ *Exercises 49.*

1. If $\boldsymbol{u} = \boldsymbol{i} + 2\boldsymbol{j} - \boldsymbol{k}, \boldsymbol{v} = -4\boldsymbol{i} + \boldsymbol{j} + 2\boldsymbol{k}$ calculate the following
   (a) $\boldsymbol{u} \times \boldsymbol{v}$
   (b) $\boldsymbol{u} \times (\boldsymbol{u} + \boldsymbol{v})$
   (c) the area of the triangle with $\boldsymbol{u}$ and $\boldsymbol{v}$ as two of its sides.
2. Prove property 5 of cross products:
   $$\boldsymbol{u} \cdot (\boldsymbol{u} \times \boldsymbol{v}) = 0 \text{ and } \boldsymbol{v} \cdot (\boldsymbol{u} \times \boldsymbol{v}) = 0.$$
3. Let $\boldsymbol{u} = \boldsymbol{i} - \boldsymbol{j}, \boldsymbol{v} = 2\boldsymbol{i} - \boldsymbol{j} + 2\boldsymbol{k}$ and $\boldsymbol{w} = 2\boldsymbol{j} - 3\boldsymbol{k}$. Calculate

(a) $(\boldsymbol{u} \times \boldsymbol{v}) \times \boldsymbol{w}$ (b) $\boldsymbol{u} \times (\boldsymbol{v} \times \boldsymbol{w})$
(c) $\boldsymbol{u} \cdot (\boldsymbol{v} \times \boldsymbol{w})$ (d) $\boldsymbol{v} \times (\boldsymbol{w} \times \boldsymbol{u})$

4. Let P_1, P_2, P_3 and P_4 be the following four points in $\mathbb{R}^3$, $P_1(-1,0,0)$, $P_2(0,1,-1)$, $P_3(1,0,1)$, $P_4(0,0,1)$. Calculate

   (a) the area of the triangle formed by P_1, P_2 and P_3,

   (b) the volume of the parallelepiped with sides given by the three vectors $\overrightarrow{P_1P_2}, \overrightarrow{P_1P_3}$ and $\overrightarrow{P_1P_4}$.

5*. Let d be the perpendicular distance from a point P to the line through two points Q and R. Show that

$$d = \frac{|\overrightarrow{PQ} \times \overrightarrow{QR}|}{|\overrightarrow{QR}|}.$$

30 Appendix: Archimedean axiom

It is easy to show that for any two positive rational numbers $x = \frac{p}{q}$ and $y = \frac{s}{t}$ there is a natural number n such that

$$xn > y.$$

Indeed, just take $n = 2qs$. Then

$$xn = 2ps > s \geq \frac{s}{t} = y.$$

For real numbers we stipulate this as an additional axiom that makes sure that there aren't "too many" real numbers. This is the *Archimedean axiom*:

For any two positive real numbers x, y there is a natural number n such that $xn > y$.

The Archimedean axiom essentially says that we can make a positive number arbitrarily large by adding sufficiently many copies of it. It is clearly equivalent to the statement:

For any two positive real numbers x, y there is a natural number n such that $\frac{y}{n} < x$.

This axiom is needed to prove seemingly obvious statements like "The set $\mathbb{N}$ is not bounded" or "For any two real numbers $a < b$ there exists a rational number c such that $a < c < b$.

Proposition 11. *The set of natural numbers $\mathbb{N}$ is unbounded above.*

Proof. We show that the assumption that $\mathbb{N}$ is bounded above contradicts the Archimedean axiom. Assume that there is an upper bound $K \in \mathbb{R}$ that is

$$n = n \cdot 1 \leq K$$

for all $n \in \mathbb{N}$. This means that the Archimedean axiom does not hold for $x = 1$ and $y = K$. □

The following Corollary is just a reformulation of the Proposition above:

Corollary 3.

$$\forall \quad K \in \mathbb{R} \quad \exists n \in \mathbb{N} \textit{ such that } n > K.$$

It turns out that the unboundedness of the set of natural numbers $\mathbb{N}$ in $\mathbb{R}$ is equivalent to the Archimedean axiom. The converse of Proposition 11 is also true:

Proposition 12. *If the Archimedean axiom is not satisfied then* $\mathbb{N}$ *is bounded.*

Proof. If the Archimedean axiom is not satisfied then there exist positive real numbers x, y such that $nx \leq y$ for any $n \in N$. This means that $\frac{y}{x}$ is an upper bound for $\mathbb{N}$. □

We will need the following

Lemma 1. *Let* $a > 0$ *be a real number. Then there exists a unique natural number* s *such that* $s \leq a$ *and* $s + 1 > a$.

Proof. We give a proof by contradiction and induction. Assume that such number does not exist and consider the set S of all natural numbers s with $s \leq a$. The set S is bounded above by a. Clearly, $0 \in S$. Our assumption means that for any s that belongs to S, $s + 1$ does belong to S as well. By induction, $S = \mathbb{N}$ which contradicts Proposition 11. The contradiction proves that the desired number s exists.

We prove uniqueness. Let s and t be two such integers. Then

$$s \leq a < s + 1$$
$$t \leq a < t + 1$$

implies

$$t < s + 1$$
$$s < t + 1$$

hence

$$0 \leq |t - s| < 1.$$

Since $t - s$ is an integer it follows that $t - s = 0$, thus $s = t$. □

♠ *Exercises 50.* Show that for any real number a there exists an integer s such that $s \leq a$ and $s + 1 > a$.

We are now ready to prove a stronger version of the density property for rational numbers:

Theorem 33. *For any two real numbers $a < b$ there exists $x \in \mathbb{Q}$ such that $a < x < b$.*

Proof. First, we find a natural number n such that $n(b-a) > 1$. Then we choose another natural number $m > -na$. Then the interval $(an + m, bn + m)$ has length greater than 1 and $an + m > 0$. Our aim is to show that this interval contains an integer. Indeed, by Lemma 1 there exists a natural number s such that $s \leq an + m$ and $s + 1 > an + m$. On the other hand $s + 1 \leq an + m + 1 < bn + m$ and therefore

$$an + m < s + 1 < bn + m$$

that is

$$a < \frac{s + 1 - m}{n} < b.$$

Therefore the rational number $\frac{s+1-m}{n}$ has the desired property. □

♠ *Exercises 51.* Show that for any two real numbers $a < b$ there are infinitely many rational numbers x such that $a < x < b$.

Another consequence of Lemma 1 is:

Theorem 34. *Any real number a is the limit of a sequence of rational numbers.*

Proof. Given a real number a. For any positive integer n there is an integer s_n such that

$$s_n \leq an < s_n + 1$$

and hence

$$\frac{s_n}{n} \leq a < \frac{s_n + 1}{n}.$$

Now,

$$\lim_{n\to\infty} \frac{s_n}{n} = a$$

because

$$\forall \varepsilon > 0 \quad \exists N = \left\lceil \frac{1}{\varepsilon} \right\rceil \quad \forall n > N: \quad \left| a - \frac{s_n}{n} \right| < \frac{s_n + 1}{n} - \frac{s_n}{n} = \frac{1}{n} < \frac{1}{N} < \varepsilon. \quad \square$$

Finally, let us show that any real number can be expressed in a unique way as an infinite decimal fraction. For a non-negative real number we can write

$$x = s_0.s_1 s_2 s_3 \cdots = \lim_{n \to \infty} \sum_{j=0}^{n} s_j 10^{-j},$$

where $s_j \in \mathbb{N}$ and $0 \leq s_j \leq 9$ for $j > 0$. For negative x we can find the decimal representation for

$$-x = s_0.s_1 s_2 s_3 \ldots$$

and get

$$x = -s_0.s_1 s_2 s_3 \ldots$$

We show by induction that there exist unique s_j such that

$$0 \leq x - x_n < 10^{-n}$$

with

$$x_n = \sum_{j=0}^{n} s_j 10^{-j}.$$

From Lemma 1 we get the unique $s_0 \in \mathbb{N}$ such that

$$s_0 \leq x < s_0 + 1,$$

which is equivalent to

$$0 \leq x - x_0 = x - s_0 < 1 = 10^0.$$

This starts the induction. For the induction step assume that unique $s_0, s_1, \ldots, s_n$ exist such that

$$0 \leq x - x_n < 10^{-n},$$

which is equivalent to

$$0 \leq 10^{n+1}(x - x_n) < 10.$$

To satisfy the inequalities

$$0 \leq x - x_n - s_{n+1} 10^{-n-1} < 10^{-n-1}$$

we need

$$s_{n+1} 10^{-n-1} \leq x - x_n < (s_{n+1} + 1) 10^{-n-1},$$

that is

$$s_{n+1} \leq (x - x_n)10^{n+1} < (s_{n+1} + 1).$$

Again, by Lemma 1, there exists a unique integer that satisfies the inequality above. Since

$$(x - x_n)10^{n+1} < 10$$

we have $s_{n+1} \leq 9$. By construction

$$0 \leq x - x_n - s_{n+1}10^{-n-1} < 10^{-n-1}.$$

Index