Analysis I for ITET/RW

E. Kowalski

Contents

Preliminaries: logic, numbers, sets, maps

1.1. Logic

1.1.1. Introduction. Mathematics depends on being able to make very precise and unambiguous statements. Experience shows that human languages often do not suffice for this purpose; sentences expressed in German or English often allow different interpretations. For instance, the word "or" is sometimes meant to describe exclusive possibilities, and sometimes allows the possibility that both are true.

EXAMPLE 1.1.1. What does the sentence "Tomorrow, it will rain" mean?
- Tomorrow, it will rain *all day*.
- Tomorrow, it will rain *at some point*.

In this script, we will nevertheless mostly express mathematical results in ordinary language. However, when needed, we will be able to use the notation of *formal logic* that we now introduce. These are also relevant in other areas of science, including in computer science, in handling binary data and binary logic.

We consider mathematical statements A, possibly depending on one or more "variables", in which case we will use notation like $\mathsf{A}(x)$. These statements are required to have an unambiguous "truth value": for any given value of the variables, they are either True or False (but may be True for certain values, and False for others).

EXAMPLE 1.1.2. The assertion

$$\mathsf{E}(n): \text{ ``}n \text{ is an even natural number''}$$

is of this type. So is

$$\mathsf{S}(n): \text{ ``}n \text{ is the square of a natural number''}$$

1.1.2. Logical operations. A number of operations and notation allow us to construct more and more sophisticated mathematical assertions starting from very simple ones.
- (Logical negation): if A is a mathematical assertion, then

$$\neg \mathsf{A}$$

  is the mathematical assertion which is True if A is False, and False if A is True. It is read "not A".

  For instance, with notation as in Example 1.1.2, the following are True:

$$\neg\mathsf{E}(3), \qquad \neg\mathsf{S}(7),$$

  and the following are False:

$$\neg\mathsf{E}(4), \qquad \neg\mathsf{S}(9).$$

- (Logical "or"): if A and B are assertions then

$$\mathsf{A} \lor \mathsf{B}$$

is the mathematical assertion which is True if *either* A is true, *or* B is true, *or both*. It is read "*A* or *B*".

For instance, with notation as in Example 1.1.2, the following are True:

$$\mathsf{E}(2) \vee \mathsf{S}(2), \qquad \mathsf{E}(9) \vee \mathsf{S}(4), \qquad \mathsf{E}(16) \vee \mathsf{S}(9)$$

and the following is False:

$$\mathsf{E}(3) \vee \mathsf{S}(3).$$

- (Logical "and"): if A and B are assertions then

$$\mathsf{A} \wedge \mathsf{B}$$

is the mathematical assertion which is True if *both* A is true *and* B is true. It is read "*A* and *B*".

For instance, with notation as in Example 1.1.2, the following is True:

$$\mathsf{E}(6) \wedge \mathsf{S}(25).$$

and the following are False:

$$\mathsf{E}(2) \wedge \mathsf{S}(2), \qquad \mathsf{E}(9) \wedge \mathsf{S}(4), \qquad \mathsf{E}(7) \wedge \mathsf{S}(8).$$

- (universal quantifier; "forall"): if $\mathsf{A}(x)$ is an assertion depending on a variable x, then

$$\forall x, \ \mathsf{A}(x)$$

is the mathematical assertion which is True if $\mathsf{A}(x)$ is True *for all* possible choices of x. It is read "for all x, $\mathsf{A}(x)$."

For instance, with notation as in Example 1.1.2, the assertion

$$\forall x, \ \mathsf{E}(x)$$

is False.

- (existential quantifier; "there exists")): if $\mathsf{A}(x)$ is an assertion depending on a variable x, then

$$\exists x, \ \mathsf{A}(x)$$

is the mathematical assertion which is True if there is *at least one* choice of x such that $\mathsf{A}(x)$ is True; it does not imply that this value is unique. It is read "for all x, $\mathsf{A}(x)$."

For instance, with notation as in Example 1.1.2, the assertion

$$\exists x, \ \mathsf{E}(x)$$

is True.

- (Logical "implication"): if A and B are assertions then

$$\mathsf{A} \rightarrow \mathsf{B}$$

is the mathematical assertion which is True if A implies B. It is read "A implies B".

For instance, with notation as in Example 1.1.2, the following is True:

$$\mathsf{E}(x) \rightarrow \mathsf{E}(x+2),$$

and the following is False:

$$\mathsf{E}(x) \rightarrow \mathsf{S}(x).$$

Combining assertions (with proper use of parentheses to make it clear in which order they are considered) with this constructions leads easily to many mathematical statements that can express complicated properties.

EXAMPLE 1.1.3. (1) Since the construction of $\forall x,\ \mathsf{A}(x)$ places no restriction on what x is (any mathematical object would do, whether a number, a circle, a disc, a pyramid, a function, etc), it may seem that only very few such statements have a chance to be true.

However, once can use subsidiary assertions and implication to, in effect, "restrict" the variable.

For instance, consider in addition to Example 1.1.2 the assertion

$$\mathsf{N}(n)\ :\ \text{``n is a natural number''}.$$

Then the assertion

$$\forall n,\ (\mathsf{N}(n) \to \mathsf{E}(2n))$$

is True: indeed, once we know that n is a natural number (i.e., that $\mathsf{N}(n)$ is True), then it follows that $2n$ is an even natural number (so that $\mathsf{E}(n)$ is also true).

(2) The logical or $\mathsf{A} \vee \mathsf{B}$ holds if either of the assertions A or B is True. We can combine this with another assertion to express the "exclusive" or, which is true if *one and only one* of the two is True.

One possibility for this purpose is

$$(\mathsf{A} \vee \mathsf{B}) \wedge (\neg(\mathsf{A} \wedge \mathsf{B})).$$

Indeed, since this is an "and" of two parts, it s True if and only if both sides are True. The first part $\mathsf{A} \vee \mathsf{B}$ is true if either A or B is True. But the second part $\neg(\mathsf{A} \wedge \mathsf{B})$ is a negation, so it is True if and only if $\mathsf{A} \wedge \mathsf{B}$ is False, which means that it is not the case that *both* A and B are True. So one, and only one, of the two is True.

(3) Similarly, we can express as follows the variant of the existential quantifier $\exists x,\ \mathsf{A}(x)$ that claims that there is a *unique* x such that $\mathsf{A}(x)$ is True:

$$(\exists x,\ \mathsf{A}(x)) \wedge (\forall y \forall z,\ (\mathsf{A}(y) \wedge \mathsf{A}(z) \to (y = z))).$$

Indeed, this is an "and" statement, so it holds when both sides are True. The first part $\exists x,\ \mathsf{A}(x)$ tells us that there exists *some* x for which $\mathsf{A}(x)$ is True. The second expresses that there is no more than one, by requiring that whenever $\mathsf{A}(y)$ and $\mathsf{A}(z)$ are *both* true (logical "and"), it follows that $y = z$. So all possible y such that $\mathsf{A}(y)$ is True are equal.

1.1.3. Negation rules. We conclude this section with an explanation of the rules for expressing the negation $\neg\mathsf{A}$ of a mathematical statement expressed using the notation above. This is important when applying the method of proof *by contraposition*: to prove that a statement A implies B, it is equivalent to prove that the negation of B implies the negation of A.

We express the rules in a table which we will then explain.

Rule	Negation
$\neg\mathsf{A}$	A
$\mathsf{A} \vee \mathsf{B}$	$(\neg\mathsf{A}) \wedge (\neg\mathsf{B})$
$\mathsf{A} \wedge \mathsf{B}$	$(\neg\mathsf{A}) \vee (\neg\mathsf{B})$
$\forall x,\ \mathsf{A}(x)$	$\exists x,\ \neg\mathsf{A}(x)$
$\exists x,\ \mathsf{A}(x)$	$\forall x,\ \neg\mathsf{A}(x)$
$\mathsf{A} \to \mathsf{B}$	$\mathsf{A} \wedge (\neg\mathsf{B})$

We can easily check these by checking when any of the left-hand column statement is False, which is when the negation is True:

- (Negation) To say that $\neg \mathsf{A}$ is not True, means that A is true.
- (Or) If it is not the case that either A or B is True, then this means that A is False and B is False, and conversely.
- (And) If it is not the case that A and B are True, then this means that either A is False or B is False.
- (For all) If it is not true that $\mathsf{A}(x)$ is True for all x, that means that for *some x*, the statement $\mathsf{A}(x)$ is False.
- (There exists) If it is not true that there exists x for which $\mathsf{A}(x)$ is True, that means that the negation of $\mathsf{A}(x)$ is True for *all x*.
- (Implication) This is the only challenging interpretation, since it doesn't immediately intuitively capture the idea of a "causal" relationship that we feel should be there when we say that A implies B. However, one can argue that, in order to say that it is not true that A implies B, it must be the case that A is (or can be, when there is a variable) True, but B is not. This issue becomes usually clearer when working with examples.

EXAMPLE 1.1.4. We refer to the statements and notation of Example 1.1.3 and "compute" their negations. The method is elementary: isolate first if the statement is a Negation, and Or, an And, etc, then apply the corresponding rule to combine negations of the substatements which are the arguments of the Negation, Or, And, etc, and continue until no more operation can be done.

(1) The statement

$$\forall n, \ (\mathsf{N}(n) \to \mathsf{E}(2n))$$

expresses the fact that $2n$ is an even natural number when n is a natural number. Using the rules above, its negation (which is False, of course) is the statement

$$\exists n \, (\mathsf{N}(n) \land \neg\mathsf{E}(2n)).$$

Precisely, applying the method sketched previously, we have two steps before reaching the complete negation:

$$\neg(\forall n, \ (\mathsf{N}(n) \to \mathsf{E}(2n)))$$
$$\exists n, \ \neg(\mathsf{N}(n) \to \mathsf{E}(2n))$$
$$\exists n, \ (\mathsf{N}(n) \land \neg\mathsf{E}(2n))$$

(the first is the rule for $\exists n, \ \mathsf{A}(n)$, the second is the rule for $\neg(\mathsf{A} \to \mathsf{B})$).

This is indeed correct: it states that there is an n, which is a natural number (because $\mathsf{N}(n)$ must be True), and for which $2n$ is not an even integer (because $\neg\mathsf{E}(2n)$ must be True).

(2) The statement

$$(\mathsf{A} \lor \mathsf{B}) \land (\neg(\mathsf{A} \land \mathsf{B})).$$

expresses that one and only one of the two statements A and B is True. Its negation should express that either both are True, or both are False. By the rules above, this negation is

$$((\neg\mathsf{A}) \land (\neg\mathsf{B})) \lor (\mathsf{A} \land \mathsf{B}),$$

obtained in the following steps:

$$\neg(\ (\mathsf{A} \vee \mathsf{B}) \wedge (\neg(\mathsf{A} \wedge \mathsf{B})) \ )$$
$$(\neg(\mathsf{A} \vee \mathsf{B})) \vee \neg((\neg(\mathsf{A} \wedge \mathsf{B})))$$
$$((\neg\mathsf{A}) \wedge (\neg\mathsf{B})) \vee (\mathsf{A} \wedge \mathsf{B})$$

(first we have the negation of an And, then in parallel the negation of an Or and of a Negation).

Indeed, this has the expected meaning: the first part of this logical Or is True if and only if both statements are False, and the second is True if and only if both are True.

(3) The statement

$$(\exists x, \ \mathsf{A}(x)) \wedge (\forall y \forall z, \ (\mathsf{A}(y) \wedge \mathsf{A}(z) \rightarrow (y = z))).$$

expresses the fact that there is a unique x for which $\mathsf{A}(x)$ is True. We express its negation using the rules in the table: it becomes

$$(\forall x, \ \neg\mathsf{A}(x)) \vee (\exists y \exists z, \ (\mathsf{A}(y) \wedge \mathsf{A}(z)) \wedge \neg(y = z))).$$

In other words: either $\mathsf{A}(x)$ is always False, or there exist y and z such that $\mathsf{A}(y)$ and $\mathsf{A}(z)$ are both True, *and* in addition y is not equal to z – so the statement $\mathsf{A}(x)$ is True for at least two different values of x.

In this case, the successive steps are as follows;

$$\neg(\ (\exists x, \ \mathsf{A}(x)) \wedge (\forall y \forall z, \ (\mathsf{A}(y) \wedge \mathsf{A}(z) \rightarrow (y = z))) \ )$$
$$\neg(\exists x, \ \mathsf{A}(x)) \vee \neg(\ (\forall y \forall z, \ (\mathsf{A}(y) \wedge \mathsf{A}(z) \rightarrow (y = z))) \ )$$
$$(\forall x, \ \neg\mathsf{A}(x)) \vee (\exists y \exists z, \ \neg(\ (\mathsf{A}(y) \wedge \mathsf{A}(z) \rightarrow (y = z)) \ )$$
$$(\forall x, \ \neg\mathsf{A}(x)) \vee (\exists y \exists z, \ (\mathsf{A}(y) \wedge \mathsf{A}(z) \wedge \neg(y = z))$$

(negation of an And, then in parallel of a Forall and two Exists, then of an Implication).

1.2. Numbers and induction

1.2.1. Numbers. We denote

- by $\mathbf{N} = \{1, 2, \ldots\}$ the set of all natural numbers,[1]
- by $\mathbf{N}_0 = \{0, 1, 2, \ldots\}$ the natural numbers including 0,
- by $\mathbf{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$ the integers,[2]
- by $\mathbf{Q}$ the rational numbers,
- by $\mathbf{R}$ the real numbers,
- by $\mathbf{C}$ the complex numbers.

We mostly assume known the basic properties and structures of $\mathbf{N}$, $\mathbf{Z}$ and $\mathbf{Q}$ (addition, multiplication, division, comparison of integers and rational numbers). We will discuss in more details the real and complex numbers later.

1.2.2. Proof by induction. The natural numbers are often used to perform *proofs by induction.*[3*]

---

* Endnotes will give German translations of certain important mathematical terms.

THEOREM 1.2.1 (Induction principle). *Suppose that for each $n \in \mathbf{N}$ we have a mathematical statement $\mathsf{A}(n)$. Then $\mathsf{A}(n)$ is True for all $n \in \mathbf{N}$, provided the following conditions hold:*

(1) *The statement $\mathsf{A}(1)$ is True.*
(2') *For all $n \in \mathbf{N}$, the statement*

$$\mathsf{A}(n) \to \mathsf{A}(n+1)$$

is True, or in other words, $\mathsf{A}(n)$ implies $\mathsf{A}(n+1)$.

REMARK 1.2.2. If the statements are indexed by $n \in \mathbf{N}_0$ instead of $n \in \mathbf{N}$, we can replace the two conditions by

(1') The statement $\mathsf{A}(0)$ is True.
(2') For all $n \geqslant 0$, the statement

$$\mathsf{A}(n) \to \mathsf{A}(n+1)$$

is True.

In either (2) or (2'), one assumes that $\mathsf{A}(n)$ is true, in order to deduce the next case of the statement. One says that $\mathsf{A}(n)$ is the *induction hypothesis*. One can replace it by the stronger-looking assumption that all the statements $\mathsf{A}(0)$, $\mathsf{A}(1)$, ..., $\mathsf{A}(n)$ are true.

EXAMPLE 1.2.3. (1) The following formula is particularly important and should be remembered.

PROPOSITION 1.2.4. *Let x be a number not equal to 1. Let $n \in \mathbf{N}_0$. We have*

(1.1)
$$1 + x + \cdots + x^n = \frac{1 - x^{n+1}}{1 - x}.$$

PROOF. We proceed by induction on n: the statement $\mathsf{A}(n)$, for $n \in \mathbf{N}_0$, is that the formula (1.1) is true.

For $n = 0$, the formula is

$$1 = \frac{1 - x}{1 - x}$$

which is indeed true. Now we take $n \geqslant 0$ and we make the induction assumption that the formula holds for this value of the parameter. In order to deduce the formula for $n + 1$, we simply note that

$$1 + x + \cdots + x^n + x^{n+1} = (1 + x + \cdots + x^n) + x^{n+1} = \frac{1 - x^{n+1}}{1 - x} + x^{n+1},$$

where we used the *induction hypothesis* to replace the first sum by the result of applying (1.1).

Now we continue the computation easily:

$$1 + \cdots + x^{n+1} = \frac{1 - x^{n+1}}{1 - x} + x^{n+1} = \frac{1 - x^{n+1} - (1 - x)x^{n+1}}{1 - x}$$
$$= \frac{1 - x^{n+2}}{1 - x}.$$

which is (1.1) with n replaced by $n + 1$. $\square$

(2) Here is an another example, which is rather surprising at first sight. Define a sequence of integers F_n for $n \in \mathbf{N}$ by putting

$$F_1 = F_2 = 1, \qquad F_{n+2} = F_{n+1} + F_n \text{ for } n \geqslant 1$$

(these are known as Fibonacci numbers); for instance, we have $F_3 = 2$, $F_4 = 2 + 1 = 3$, $F_5 = 3 + 2 = 5$, etc.

It is in general almost impossible to find a simple expression for sequences obtained by such inductive definition (when new terms are defined in terms of previous ones). But here, we have

(1.2)
$$F_n = \frac{\alpha^n - \beta^n}{\alpha - \beta}$$

where

$$\alpha = \frac{1 + \sqrt{5}}{2}, \qquad \beta = \frac{1 - \sqrt{5}}{2}.$$

We will see later how one can understand where this formula comes from (see Example 4.3.8). But at least, once it is stated, we easily check that it is indeed true using induction.

Here, because the definition involves previous terms of the sequence, it is better to use the variant of induction where we check the formula for $n = 1$ and $n = 2$, and then assume (1.2) for n and $n + 1$ in order to deduce it for $n + 2$. (You should try to see why this is a valid form of induction.)

For $n = 1$, the formula (1.2) is

$$1 = F_1 = \frac{\alpha - \beta}{\alpha - \beta} = 1,$$

and for $n = 2$, it becomes

$$1 = F_2 = \frac{\alpha^2 - \beta^2}{\alpha - \beta} = \alpha + \beta = 1,$$

which is also true.

Now suppose for the purpose of induction that (1.2) is true for n and $n + 1$, and let's try to prove it for F_{n+2}. It is easy to see how to proceed: by definition, we have

$$F_{n+2} = F_n + F_{n+1},$$

so that the *two* induction assumptions lead to

$$F_{n+2} = \frac{1}{\alpha - \beta}(\alpha^n - \beta^n + \alpha^{n+1} - \beta^{n+1}).$$

We can simplify this: we have

$$\alpha^n - \beta^n + \alpha^{n+1} - \beta^{n+1} = \alpha^n(1 + \alpha) - \beta^n(1 + \beta),$$

and we note that

$$1 + \alpha = \frac{3 + \sqrt{5}}{2} = \left(\frac{1 + \sqrt{5}}{2}\right)^2 = \alpha^2, \quad 1 + \beta = \frac{3 - \sqrt{5}}{2} = \left(\frac{1 - \sqrt{5}}{2}\right)^2 = \beta^2,$$

which leads to

$$F_{n+2} = \frac{1}{\alpha - \beta}(\alpha^{n+2} - \beta^{n+2}).$$

This concludes the inductive step of the proof.

We see that in this proof, the key property of α and β is that they are the two roots of the algebraic equation

$$X^2 - X - 1 = 0.$$

1.2.3. The factorial. We conclude this section by defining the factorial[4] function $n!$ for integers $n \in \mathbf{N}_0$: we put

$$n! = 1 \cdot 2 \cdot \cdots \cdot n,$$

if $n \geqslant 1$ (the product of the integers from 1 to n); for $n = 0$, the correct convention is to define

$$0! = 1.$$

We can see then that we have the inductive property

$$(n+1)! = (n+1) \cdot n!$$

for all $n \in \mathbf{N}_0$, including for $n = 0$. The first few values of $n!$ are:

n	0	1	2	3	4	5	6	7	8
$n!$	1	1	2	6	24	120	720	5040	40320

1.3. Sets

1.3.1. Introduction. The goal of this section is to introduce notation about *sets*;[5] these are very useful to speak formally and consistently about many different types of mathematical objects. In fact, in a precise sense, all mathematical objects can be considered to be sets of a kind or another.

A (mathematical) set X is an unordered collection of mathematical objects, which is uniquely determined by the set of elements that it contains. These elements are arbitrary mathematical objects (in particular, they can be sets themselves).

We use the notation $a \in X$ to say that some object a belongs to a set X; when we want to state that a mathematical object a is not an element of a set X, we write $a \notin X$ (this is therefore the negation of the statement A that states that a is in X).

The fact that a set is determined by its elements means that two sets X and Y are equal if and only if the following statement is True:

$$\forall\, a, ((a \in X) \to (a \in Y)) \wedge ((a \in Y) \to (a \in X)),$$

which we also abbreviate in

$$\forall\, a, ((a \in X) \leftrightarrow (a \in Y))$$

(the double arrow is read "if and only if").

EXAMPLE 1.3.1. (1) The various sets of numbers are all sets:

$$\mathbf{N}_0, \quad \mathbf{N}, \quad \mathbf{Z}, \quad \mathbf{Q}, \quad \mathbf{R}. \quad \mathbf{C}.$$

(2) We sometimes define a set by listing within curly brackets its elements:

$$X = \{0, 1, 2, 3\}$$

is a set with 4 elements. The fact that sets are "unordered" means that we can write the elements of that list in any order. We can also repeat some elements multiple times without changing the set (since this doesn't change what the elements are).

So, for instance, we have

$$\{0, 1, 2, 3\} = \{0, 3, 1, 2\} = \{0, 0, 0, 1, 3, 2, 2\}.$$

(3) To illustrate that elements of a set can be arbitrary, define the set

$$X = \{0, 1, \{1, 2\}, \mathbf{N}\}.$$

This is a set with 4 elements: the integers 0 and 1, the set $\{1, 2\}$, and the set of natural numbers. So one can write $\mathbf{N} \in X$.

1.3.2. Defining sets. The most usual method to define a set is to have a statement $\mathsf{A}(x)$ depending on a variable x, and a given set X, and then to look at the set of all elements of X for which $\mathsf{A}(x)$ is True. This is denoted

$$Y = \{a \in X \mid \mathsf{A}(a) \text{ holds}\}, \text{ or simply } Y = \{a \in X \mid \mathsf{A}(a)\}.$$

The vertical bar is read "such that".

EXAMPLE 1.3.2. (1) Using notation from Example 1.1.2, the set of even natural numbers can be defined by

$$Y = \{n \in \mathbf{N} \mid \mathsf{E}(n)\}.$$

(2) Define

$$Y = \{n \in \mathbf{Z} \mid \text{there exist } a, b, c, d \text{ in } \mathbf{N}_0 \text{ such that } n = a^2 + b^2 + c^2 + d^2\}.$$

This set Y turns out to be equal to $\mathbf{N}_0$, but this is very far from obvious: it was first proved by Lagrange in the 18th Century.

(3) It is important, when defining sets in this manner, to specify the set X in which elements are taken. Otherwise, paradoxical results may follow. For instance, consider the "set"

$$P = \{x \mid x \notin x\}.$$

If this is a set, then we can ask whether $P \in P$ or not. But note that if $P \in P$, then by "definition" we would have $P \notin P$. And if $P \notin P$, then again, that would mean that $P \in P$.

(This construction is a mathematical version of a standard language paradox: in a town, the barber shaves precisely those people who do not shave themselves. Does the barber shave himself?)

1.3.3. Operations on sets. There are important operations with sets which we now describe.

- **Union:**[6] if X and Y are sets, the union $X \cup Y$ is the set such that $a \in X \cup Y$ if and only if $a \in X$ or $a \in Y$, or both. In other words, in logical sentences:

$$(a \in X \cup Y) \leftrightarrow ((a \in X) \vee (a \in Y)).$$

For instance, we can write

$$\mathbf{N}_0 = \mathbf{N} \cup \{0\}.$$

- **Intersection:**[7] if X and Y are sets, the intersection $X \cap Y$ is the set such that $a \in X \cap Y$ if and only if $a \in X$ and $a \in Y$. In other words, in logical sentences:

$$(a \in X \cap Y) \leftrightarrow ((a \in X) \wedge (a \in Y)).$$
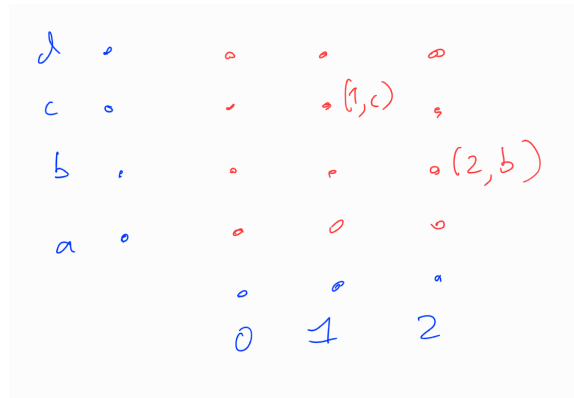
For instance, we can write

$$\mathbf{Z} \cap \{a \in \mathbf{Q} \mid 0 \leqslant a \leqslant 1\} = \{0, 1\}.$$

- **The empty set**[8]: if we consider instead of the last example the similar intersection

$$\mathbf{Z} \cap \{a \in \mathbf{Q} \mid 0 < a < 1\},$$

then we obtain an important set, namely the *empty set*, which is the only set with no elements at all. It is denoted $\emptyset$. So

$$\mathbf{Z} \cap \{a \in \mathbf{Q} \mid 0 < a < 1\} = \emptyset.$$

Many other intersection sets are empty, for instance

$$\{0, 1, 2\} \cap \{-2, 3, 4\} = \emptyset.$$

The empty set is also formally very useful when defining solution sets of equations, since it is important to allow the possibility than an equation has no solution.

- **Product**:[9] if X and Y are sets, then the product $X \times Y$ is the set of all *ordered pairs* (a, b) where $a \in X$ and $b \in Y$. The fact that these are ordered pairs (in contrast with sets) means that

$$(a, b) = (c, d)$$

is equivalent to $a = c$ and $b = d$, so that for instance $(1, 2) \neq (2, 1)$, whereas $\{1, 2\} = \{2, 1\}$.

For instance

$$\{0, 1, 2\} \times \{a, b, c, d\} =$$
$$\{(0, a), (0, b), (0, c), (0, d), (1, a), (1, b), (1, c), (1, d), (2, a), (2, b), (2, c), (2, d)\}.$$

A product can be visualized as points on a plane or a grid, as in the following picture:

Another good example of a product is $\mathbf{R} \times \mathbf{R}$, which is the usual euclidean plane, where (x, y) represents the point with first coordinate x and second coordinate y.

Note that $X \times \emptyset$ and $\emptyset \times X$ are both empty, whatever the choice of X.

1.3.4. Subsets. A set Y is a *subset*[10] of a set X if all elements of Y are also elements of X. We then say that Y is contained in X, and denote this property by

$$Y \subset X.$$

EXAMPLE 1.3.3. (1) We have

$$\mathbf{N} \subset \mathbf{N}_0 \subset \mathbf{Z} \subset \mathbf{Q} \subset \mathbf{R} \subset \mathbf{C}.$$

(2) The empty set is contained in all sets: $\emptyset \subset X$ is true, whatever X is.
(3) We always have $X \subset X$.
(4) For any sets X and Y, we have

$$X \subset X \cup Y, \quad Y \subset X \cup Y, \quad X \cap Y \subset X, \quad X \cap Y \subset Y.$$

(5) Suppose that $Z \subset X$ and that $W \subset Y$; then

$$Z \times W \subset X \times Y.$$

On the other hand, there are usually some subsets of $X \times Y$ which are not of the form $Z \times W$. For instance, in the plane $\mathbf{R} \times \mathbf{R}$, the line

$$\Delta = \{(x, y) \in \mathbf{R} \times \mathbf{R} \mid x = y\}$$

(the diagonal in the plane) is not of the form $Z \times W$.

1.3.5. Cardinality. If a set X is finite, which means that it has only a finite number of elements, then we call this number of elements the *cardinality* of X. This is denoted by $\mathrm{Card}(X)$, and is an element of $\mathbf{N}_0$.[11]

For instance, the empty set is finite with cardinality 0.

EXAMPLE 1.3.4. (1) The set

$$X = \{\emptyset, \{\emptyset\}\}$$

has two elements: the empty set, and the set $\{\emptyset\}$ (which itself has one element). Similarly

$$X = \{\mathbf{N}, \{\mathbf{N}_0\}\}$$

has two elements; one is the (infinite) set $\mathbf{N}$, and the other is the one-element set $\{\mathbf{N}_0\}$.

(2) If X and Y are finite sets, then so are $X \cup Y$, $X \cap Y$ and $X \times Y$. We have

$$\mathrm{Card}(X \times Y) = \mathrm{Card}(X)\,\mathrm{Card}(Y),$$

and

$$\mathrm{Card}(X \cup Y) + \mathrm{Card}(X \cap Y) = \mathrm{Card}(X) + \mathrm{Card}(Y).$$

(this is because $\mathrm{Card}(X) + \mathrm{Card}(Y)$ "counts twice" the elements of $X \cup Y$ which are in both sets, in other words those of $X \cap Y$).

1.4. Maps

1.4.1. Introduction. Many parts of mathematics, and many applications of mathematics, deal with the question of solving equations, and equations exist of completely different nature. These include:

- Algebraic equations, such as

(1.3)
$$x^3 + 3x - 1 = 0$$

   where the unknown x is a single number.
- Newton's equations of classical mechanics: the unknown here is for instance the trajectory of a particle, or of a 3-dimensional object. It involves therefore, for each time $t \geqslant 0$, a point in the space $\mathbf{R}^3$.
- Maxwell's equations, which are the fundamental equations of classical electromagnetism: there the unknown is, for each time $t \geqslant 0$ and each point x in space, the value of the electric and magnetic field at that point and at that time (these amount to two vectors in $\mathbf{R}^3$, so six coordinates).
- And there are the Schrödinger equations of quantum mechanics, the Einstein equations of general relativity, etc, *including equations which have not yet been written down by anyone!*

In order to speak uniformly of *all possible equations* (including those that no scientist has yet identified as useful), mathematicians have introduced a language that may look abstract and complicated but allows us to use the same words for certain fundamental properties of equations, independently of their nature. These properties include, for instance, the question of whether there is a solution, and if yes, whether it is unique.

We represent any equation in the form

$$f(x) = y$$

where

- x is the *unknown*, an element of a given set X;
- y is the "right hand side", an element of another given set Y;
- f, which describes the equation, is a *map*[12] from X to Y, which means that it is a well-defined rule that associates to *each* element of X a *unique* element $f(x)$ in Y. We denote by

$$f \colon X \to Y$$

the fact that f is a map from X to Y.

If $f \colon X \to Y$ is a map from X to Y, then we say that X is the *definition set* or *domain*[13] and that Y is the *target set*.[14] If Y is a set of numbers, one often says that f is a function[15]. For $x \in X$, one says that $f(x)$ is *the image of* x (by f).[16]

Two maps $f \colon X \to Y$ and $g \colon X' \to Y'$ are *equal* if and only if $X = X'$ and $Y = Y'$, and if furthermore $f(x) = g(x)$ for all $x \in X = X'$.

EXAMPLE 1.4.1. (1) For an algebraic equation, one usually takes $X = Y = \mathbf{C}$, or $X = Y = \mathbf{R}$, with $y = 0$, and f is the left-hand side of the equation, for instance

$$f(x) = x^3 + 3x - 1$$

for the equation (1.3).

Note that whether the equation $f(x) = 0$ has a solution may depend on the choice of X and Y; this explains why the equality of maps depends on the choice of the definition and target sets, and not simply on the fact that the "formula" defining f is the same. Indeed, it is natural to expect that equal functions should have the same solutions!

As an example, the functions

$$f_1 \colon \mathbf{R} \to \mathbf{R} \qquad f_2 \colon \mathbf{C} \to \mathbf{C}$$

defined by $f_1(x) = x^2 + 1$ and $f_2(x) = x^2 + 1$ are not equal; indeed, the equation $f_1(x) = 0$ has no solution, whereas the equation $f_2(x) = 0$ does have solutions.

(2) A map $f \colon X \to Y$ has the property that $f(x)$ is always defined, whatever $x \in X$ is taken. For instance, there is no function $f_1 \colon \mathbf{Z} \to \mathbf{Q}$ such that $f_1(n) = 1/n$, because this is not well-defined when $n = 0$. One can however define $f_2 \colon X \to \mathbf{Q}$ where

$$X = \{n \in \mathbf{Z} \mid n \neq 0\}$$

by $f_2(n) = 1/n$. Or one can define $f_3 \colon \mathbf{Z} \to \mathbf{Q}$ by defining $f_3(0) = 0$ and $f_3(n) = 1/n$ for $n \neq 0$. The functions f_2 and f_3 are not equal.

(3) The usual operations (addition, multiplication, etc) are maps of a special kind: for instance $f_4 \colon \mathbf{Q} \times \mathbf{Q} \to \mathbf{Q}$ defined by

$$f_4(x, y) = x + y$$

defines addition. Division can be defined as a map $f_5 \colon \mathbf{Q} \times \mathbf{Q}^* \to \mathbf{Q}$, where

$$\mathbf{Q}^* = \{x \in \mathbf{Q} \mid x \neq 0\},$$

by $f_5(x, y) = x/y$.

1.4.2. Injective, surjective, bijective. If we have a map $f \colon X \to Y$ that we use to write down equations

$$f(x) = y$$

it is very often the case that we want to look at equations where y is not simply a fixed element of Y – for instance, y could be related to "initial conditions" which may vary with different experiments.

The following are then very natural questions concerning the equation $f(x) = y$, with unknown x:

- Does it have a solution for *all* values of y in Y?
- Does it have *at most* one solution?
- Does it have *exactly one* solution?

A positive answer is a property, which may or may not be true, of the map f. Mathematicians use the following terminology[†] to avoid repeating these three sentences over and over again:

- If the equation $f(x) = y$ has *always at least one solution*, we say that f is *surjective*.[17]
- If the equation $f(x) = y$ has *never more than one solution*, we say that f is *injective*.[18]
- If the equation $f(x) = y$ has *always exactly one solution*, we say that f is *bijective*.[19]

REMARK 1.4.2. In all three definitions, the unknown of the equation is $x \in X$, and the "always" or "never" refers to the property being true for all possible $y \in Y$.

Note that by definition, to say that f is bijective is the same as saying that f is *both* injective and surjective.

EXAMPLE 1.4.3. (1) The *standard example* of a surjective map is the following: we consider sets A and Y, where A is not empty, and we denote $X = A \times Y$ and the map

$$p \colon X \to Y$$

defined by $p(a, y) = y$ for all pairs $(a, y) \in A \times Y = X$. (This is often called a *coordinate projection*.)

This map p is surjective because the corresponding equation $p(x) = y$ means precisely that the second coordinate of $x \in A \times Y$ is equal to y, and so we can fix any $a_0 \in A$, and $x = (a_0, y)$ is a solution.

(2) The *standard example* of an injective map is the following: we consider a subset X of a set Y, and the map

$$i \colon X \to Y$$

defined by $i(x) = x$ for all $x \in X$ (which makes sense because $X \subset Y$, so that $x \in Y$ also; this is called an *inclusion map*.)

This is injective, because for any $y \in Y$, the corresponding equation $i(x) = y$ means $x = y$, so there is at most one possible choice for x. But note that it may be that y was chosen to be an element of Y that is not in X, in which case the equation $i(x) = y$ has *no* solution.

[†] This terminology was only fixed relatively recently – before the 1950's, all kinds of words would be used for any of the three possibilities, making communication sometimes awkward.

(3) The *standard example* of a bijective map is the *identity map*[20] of a set X, which is the map

$$f \colon X \to X$$

defined by $f(x) = x$ for all $x \in X$. In this case, the only solution of the equation $f(x) = y$ is given by $x = y$; since here all values of y are elements of X (because the target set is X), this is always a solution.

We denote the identity map of X by Id_X.

We will not emphasize too much the use of this terminology when this is not essential, but we will be able to see how it allows us to speak uniformly of many different properties.

REMARK 1.4.4. (1) How does one prove that a given map $f \colon X \to Y$ is injective, or surjective?

For surjectivity: take an arbitrary $y \in Y$, and attempt to solve the equation $f(x) = y$; if this is always successful, then f is surjective.

For injectivity: the most convenient approach is usually to take elements x_1 and x_2 in X, and to assume that $f(x_1) = f(x_2)$, and then to deduce that $x_1 = x_2$. This is equivalent to f being injective:

- If f is injective and $f(x_1) = f(x_2)$, then putting $y = f(x_1)$, the equation $f(x) = y$ has (at least) the solutions x_1 and x_2; but for an injective map, there can be at most one solution, so that x_1 must be equal to x_2, which establishes the condition we stated.
- Conversely, assuming that this condition is true, we look at a given $y \in Y$ and at the equation $f(x) = y$; if it had at least two different solutions x_1 and x_2, that would mean that $f(x_1) = f(x_2)$ but $x_1 \neq x_2$, which is impossible.

(2) How does one prove that a given map $f \colon X \to Y$ is *not* injective, or *not* surjective?

To prove that f is not surjective, it is enough to find a single $y_0 \in Y$ such that $f(x) \neq y_0$ for all $x \in X$.

To prove that f is not injective, it is enough to find two elements $x_1 \neq x_2$ in X such that $f(x_1) = f(x_2)$.

We now consider a number of further examples.

EXAMPLE 1.4.5. (1) Squaring a number defines maps

$$s_1 \colon \mathbf{N} \to \mathbf{N}, \quad s_2 \colon \mathbf{Z} \to \mathbf{Z}, \quad s_3 \colon \mathbf{Q} \to \mathbf{Q}, \quad s_4 \colon \mathbf{R} \to \mathbf{R}, \quad s_5 \colon \mathbf{C} \to \mathbf{C},$$

(so $s_1(n) = n^2$, etc).

We then have the following facts:

- s_1 is injective but not surjective,
- s_2, s_3 and s_4 are not injective, and not surjective,
- s_5 is not injective, but is surjective.

Indeed:

- If two natural numbers have the same square, then they are equal (which means that s_1 is injective) but 3 is not the square of a natural number (which means that s_1 is not surjective).
- Since $(-1)^2 = 1^2$, we have two elements of $\mathbf{Z}$, or $\mathbf{Q}$, or $\mathbf{R}$ with the same square, and that means that none of the maps s_2, s_3 and s_4 is injective.

  Also $-1 \in \mathbf{Z}$ is not the square of any integer, or rational number, or real number, which means that the equation $x^2 = -1$ has no solution in any of these three sets, and therefore none of s_2, s_3, s_4 is surjective.

- Every complex number is the square of another complex number (which means that s_5 is surjective) but $(-1)^2 = 1^2$ shows that s_5 is not surjective.

(2) The map

$$f: \mathbf{Z} \times \mathbf{Z} \to \mathbf{Z}$$

defined by $f(a, b) = a + b$ is surjective, but not injective: to see that it is surjective, observe that $f(0, b) = b$ for any $b \in \mathbf{Z}$, and to see that it is not injective, note for instance that $f(1, -1) = f(2, -2) = 0$.

(3) The map

$$f: \mathbf{N} \times \mathbf{N} \times \mathbf{N} \to \mathbf{N}$$

defined by

$$f(n, m, p) = 2^n 3^m 5^p$$

is not surjective (for instance, there is no (n, m, p) with $f(n, m, p) = 7$). It is however injective, because of the Fundamental Theorem of Arithmetic: a natural number has a unique expression as a product of primes with various exponents, so that

$$2^n 3^m 5^p = 2^{n'} 3^{m'} 5^{p'}$$

can only happen if $(n, m, p) = (n', m', p')$. (For instance, if $n > n'$, this equality would imply that

$$2^{n-n'} 3^m 5^p = 3^{m'} 5^{p'}$$

which is impossible because the left-hand side is an even integer, and the right-hand side is not.)

(4) Let $\mathbf{R}_+$ be the set of non-negative real numbers (which means that $x \geqslant 0$). Then squaring map defines $s: \mathbf{R}_+ \to \mathbf{R}_+$ (because the square of a non-negative real number is also non-negative). This map is *bijective*: it is injective because $x^2 = y^2$ can only happen if $x = y$ or $x = -y$, and if x and y are in $\mathbf{R}_+$, then either $y < 0$, so $x = -y$ is not possible, or $y = 0$, and then $-y = y$. In

The map s is surjective because of the existence of square roots of non-negative real numbers (which we will recall later as a basic property of real numbers).

1.4.3. Composition. Maps between sets come with an extremely important operation, called *composition*.[21]

Suppose that we have sets X, Y, Z and maps $f: X \to Y$ and $g: Y \to Z$, where the important point is that the target set of the first map is the definition set of the second. Then we define a map

$$h: X \to Z$$

by putting

$$h(x) = g(f(x))$$

for all $x \in X$; this makes sense, since $f(x)$ belongs to Y, so that we can evaluate g at this value, and obtain an element of Z. This map is denoted $g \circ f$, and called the *composition* of g and f. To check that it exists, one often uses the diagram

$$X \xrightarrow{f} Y \xrightarrow{g} Z.$$

On can also compose more than two maps; given

$$X \xrightarrow{f} Y \xrightarrow{g} Z \xrightarrow{h} W$$

we can form either $h \circ (g \circ f)$ or $(h \circ g) \circ f$. Both of these are the same maps $X \to W$, because they have the same definition set X and the same target set W, and for any $x \in X$, we get

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

This common map is usually written simply $h \circ g \circ f$.

EXAMPLE 1.4.6. Consider the maps

$$\mathbf{R} \xrightarrow{\ f\ } \mathbf{R}^2 \xrightarrow{\ g\ } \mathbf{R}$$

such that $f(x) = (x, x)$ and $g(x, y) = xy$. Then $g \circ f$ is the squaring map $\mathbf{R} \to \mathbf{R}$ since

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

for all x.

(2) For any sets X and Y and any maps

$$f \colon X \to Y, \qquad g \colon Y \to X,$$

we have the formulas

$$f \circ \mathrm{Id}_X = f, \qquad \mathrm{Id}_X \circ g = g$$

where Id_X is the identity map of X.

1.4.4. Inverse of a bijective map. Suppose that $f \colon X \to Y$ is a *bijective map*. Then, for any y in Y, there is a unique element $x \in X$ such that $f(x) = y$ (a unique solution of the equation $f(x) = y$). Since x exists for every y and is unique, we can define a map $g \colon Y \to X$ such that $g(y) = x$ for all $y \in X$. (Note that this is often not defined by an easy formula, even if f is.)

DEFINITION 1.4.7. The map $g \colon Y \to X$ defined above for a bijective map $f \colon X \to Y$ is called the *inverse*[22] of f, and is often denoted $g = f^{-1}$.

PROPOSITION 1.4.8. *Let f be a bijective map $X \to Y$ and $g = f^{-1}$ its inverse.*
(1) *The map g is bijective and its inverse is f.*
(2) *We have*

(1.4) $$f(g(y)) = y, \qquad g(f(x)) = x$$

for every x and y, or in other words

$$f \circ g = \mathrm{Id}_Y, \qquad g \circ f = \mathrm{Id}_X.$$

PROOF. One can prove directly (1), but it is better to see that it follows from (2), which shows that formulas for compositions can be very useful. We will then prove (2).

So assume that (2) is true.

The map g is surjective: indeed, let $x \in X$; we must find $y \in Y$ with $g(y) = x$, and for this purpose, we define simply $y = f(x)$, since then

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

by the second formula in (1.4).

The map g is injective: indeed, let y_1 and y_2 be elements of Y such that $g(y_1) = g(y_1)$. We apply the map f to both sides of this equation and obtain

$$y_1 = f(g(y_1)) = f(g(y_2)) = y_2$$

by using (twice) the first formula in (1.4).

This shows that g is bijective. To find its inverse $g^{-1}\colon X \to X$, given an element x in X, we must find the unique $y = g^{-1}(x)$ such that $g(y) = x$. We have seen that $g(f(x)) = x$, and that means that this unique value of y is equal to $f(x)$. This means that $g^{-1} = f$.

Now we will prove (2). First, let $y \in Y$. By definition, $x = g(y)$ is the unique element of x such that $f(x) = y$, so that $f(g(y)) = y$. Secondly, let $x \in X$. Then $g(f(x)) = x'$ is the unique element of X such that $f(x') = f(x)$. Since x has this property, the uniqueness means that $x' = x$, so $g(f(x)) = x$. $\qquad\square$

EXAMPLE 1.4.9. For $\mathbf{R}_+$ the set of non-negative real numbers and f the bijective map $\mathbf{R}_+ \to \mathbf{R}_+$ defined by $f(x) = x^2$ (Example 1.4.5, (4)), the map f^{-1} is the squareroot function: $f^{-1}(x) = \sqrt{x}$. The properties above become the well-known formulas

$$\sqrt{x^2} = x, \qquad (\sqrt{y})^2 = y$$

for any $x \geqslant 0$ and $y \geqslant 0$.

1.4.5. Other interpretations of maps. We have motivated the use of maps to formalize very general equations. However, as often is the case, a useful scientific concept has more than one application or interpretation. In this case, we can also use maps to represent other kinds of date.

For instance, we mentioned that the solution to the Newton equations for the movement of a particle consists in giving the position of the article at all times $t \geqslant 0$. This means that this solution can be seen as a map

$$f\colon \mathbf{R}_+ \to \mathbf{R}^3,$$

where $f(t)$ is the point in space where the particle is located at time t.
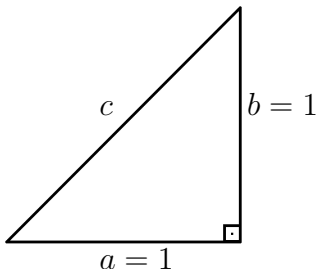
Similarly, we will see that many signals that occur in electrical engineering and similar engineering fields can be interpreted as maps of some kind.

1.5. The real numbers

Reference: [2, 2.1, 2.2, 2.3].

The most important set of numbers for analysis is the set $\mathbf{R}$ of real numbers. It is required because, although all algebraic operations (addition, multiplication, division) are defined for rational numbers, there are some equations which do not have rational solutions, although they "obviously" have some solution. The simplest is

$$x^2 = 2.$$

The geometric picture and the formula of Pythagoras $a^2 + b^2 =$

c^2 show that the length c has square equal to 2. But no rational number x satisfies $x^2 = 2$. (We recall the proof: if there were such a number, we could express it as a fraction $x = p/q$ with p and q in $\mathbf{N}$, and not both even; then we get $2q^2 = p^2$, so that p^2 is even; in this case, the number p itself must be even, so we can write $p = 2m$ for some integer $m \geqslant 1$, and then the last equation simplifies to $q^2 = 2m^2$, which means that q also is even, and that is impossible.)

Other "geometrically obvious" numbers, such as the length 2π of the perimeter of a circle of radius 1, are also not rational (but this is much harder to prove!)

In $\mathbf{R}$, however, all such "obvious" numbers have a meaning. However, it is not so easy to construct rigorously the set of real numbers to understand its properties – this was first done in 1858 by R. Dedekind, when he was teaching Analysis at ETH. Instead of presenting such a construction, we will list the properties of the real numbers – in a precise sense, one can show that they *characterize* the real numbers. In the remainder of the course, we will see that starting from these basic properties, we can deduce all the results of analysis (including give a rigorous definition of π), and also devise efficient methods for numerical computations.

The structures that are required to characterize $\mathbf{R}$ are:

- The addition, and the number 0,
- The multiplication, and the number 1,
- Division by non-zero numbers,
- The *order relation* $a \leqslant b$ for real numbers.

The difference between $\mathbf{R}$ and $\mathbf{Q}$ (where all these operations also exist) will be summarized in a single statement called *completeness*[23] of $\mathbf{R}$, which expresses a property of continuity (the fact that "there are no holes" in $\mathbf{R}$, in a certain sense).

We list the various properties of these operations in order. The first of them are all very well-known, and we will not discuss them very much – notice in particular that they are all true for rational numbers!

Properties of addition. The following are true for any real numbers:

$$(1.5) \qquad\qquad a + 0 = 0 + a = a$$

$$(1.6) \qquad\qquad a + b = b + a$$

$$(1.7) \qquad\qquad a + (b + c) = (a + b) + c$$

$$(1.8) \qquad \text{there is a unique real number } -a \text{ such that } a + (-a) = (-a) + a = 0.$$

Properties of multiplication. The following are true for any real numbers:

$$(1.9) \qquad\qquad a \cdot 1 = 1 \cdot a = a$$

$$(1.10) \qquad\qquad ab = ba$$
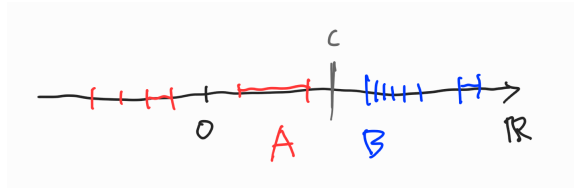
$$(1.11) \qquad\qquad a(bc) = (ab)c$$

$$(1.12) \qquad \text{if } a \neq 0, \text{ there is a unique real number } a^{-1} \text{ such that } a \cdot a^{-1} = a^{-1} \cdot a = 1.$$

Properties of addition and multiplication together. The following are true for any real numbers:

$$(1.13) \qquad\qquad a \cdot 0 = 0 \cdot a = 0$$

$$(1.14) \qquad\qquad a(b + c) = ab + ac$$

$$(1.15) \qquad\qquad (a + b)c = ac + bc.$$

Properties of the order relation. The following are true for any real numbers:

(1.16) $$a \leqslant a$$

(1.17) $$a \leqslant b \text{ and } b \leqslant a \text{ imply that } a = b$$

(1.18) $$a \leqslant b \text{ and } b \leqslant c \text{ imply that } a \leqslant c$$

(1.19) $$\text{either } a \leqslant b \text{ or } b \leqslant a$$

(1.20) $$a \leqslant b \text{ implies that } a + c \leqslant b + c$$

(1.21) $$a \leqslant b \text{ and } c \geqslant 0 \text{ imply that } ac \leqslant bc.$$

And finally comes the last property, which is only valid for $\mathbf{R}$ and not for $\mathbf{Q}$.

Completeness of R. Whenever we have two non-empty subsets A and B of real numbers with the property that any element a of A is smaller or equal to any element of B, then there is a real number c "between them": there exists $c \in \mathbf{R}$ such that

$$a \leqslant c \leqslant b$$

for all $a \in A$ and $b \in B$.

This is illustrated in the following picture:

To see that $\mathbf{Q}$ does not have this final property, we use it to prove:

THEOREM 1.5.1. *There exists a real number x such that $x^2 = 2$.*

PROOF. We will construct sets A and B which satisfy the property of the completeness statement, and for which the real number c will have to be a square root of 2.

The idea is quite easy: we define as before

$$\mathbf{R}_+ = \{x \in \mathbf{R} \mid x \geqslant 0\}$$

and

(1.22) $$A = \{a \in \mathbf{R}_+ \mid a^2 \leqslant 2\},$$

(1.23) $$B = \{b \in \mathbf{R}_+ \mid b^2 \geqslant 2\}.$$

These sets are not empty: for instance, $0 \in A$ and $2 \in B$. In addition, the condition $a \leqslant b$ is equivalent to $a^2 \leqslant b^2$ for elements of $\mathbf{R}_+$ (as can be checked using all the properties above), and therefore every element of A is smaller or equal than any element of B. By the completeness property, there exists therefore a real number c "between" A and B.

We now prove that $c^2 = 2$ by proving that we can derive a contradiction if we assume either that $c^2 > 2$, or that $c^2 < 2$. Then only the possibility $c^2 = 2$ remains.

We begin by assuming that $c^2 < 2$, and we will then show how to construct an element a of A such that $c < a$, which contradicts the fact that c is "to the right" of A. To do this, we first observe that $c \geqslant 1$, because $1 \in A$. The reason that this works is that the squaring operation is *continuous*: if we change the argument c by a very small amount e, then the value of the square $(c + e)^2$ changes also very little, and in particular the value $(c + e)^2$ will remain below 2.

To do this precisely, define $L = 2 - c^2$; this is a strictly positive real number. We consider a real number e such that $0 < e < c$, and define $a = c + e$, so that $c < a$. Under which condition can we ensure that $a \in A$? We compute

$$a^2 = (c + e)^2 = c^2 + 2ec + e^2 \leqslant c^2 + 3ec,$$

where the last inequality comes from the fact that $e < c$, so that $e^2 \leqslant ec$. If e is chosen so that $3ec < L$, for instance $e = L/(6c)$, then we get $3ec < L$ and

$$a^2 < c^2 + L = 2,$$

which implies that $a \in A$.

The other assumption $c^2 > 2$ is similar: we then construct an element b of B with $b < c$ – details are omitted. $\square$

REMARK 1.5.2. (1) Once we know the existence of $\sqrt{2}$, the sets A and B used above are just

(1.24) $$A = \{a \in \mathbf{R}_+ \mid a \leqslant \sqrt{2}\},$$

(1.25) $$B = \{b \in \mathbf{R}_+ \mid b \geqslant \sqrt{2}\},$$

in which case it is immediate that the only possible number c "in the midlle" is $\sqrt{2}$. The use of the square allowed us to replace this definition, which would not make sense without knowing the existence of $\sqrt{2}$, by one that is equivalent but involves only rational numbers.

(2) There are many different properties which, together with those of $\mathbf{Q}$, are equivalent to the completeness property. The version in [2] is more complicated to state, but easier to apply to construct (for instance) the square root of real numbers. However, we will soon have easier and convenient tools to solve such equations, so we have used the version that is also discussed in [1, 1.1].

(3) What this result really proves is that if there exists a set $\mathbf{R}$ with operations that satisfy all the rules that we listed above, then that set contains an element with $x^2 = 2$. We have *not* however proved the existence of the set of real numbers, which is more difficult.

(4) In the proof, we have used a number of simple properties that follow from the list of properties of real numbers, (for instance that $a^2 \geqslant 0$ for all $a \in \mathbf{R}$, or that $a \leqslant b$ is equivalent to $a^2 \leqslant b^2$ when a and b are in $\mathbf{R}_+$, or that $0 < a \leqslant b$ implies $0 < b^{-1} < a^{-1}$, etc). Proving these is a good exercise.

Although there are "more" real numbers than rational numbers, the following important fact shows that rational numbers can be bound between any two distinct real numbers (one says that they are "dense" in $\mathbf{R}$).

THEOREM 1.5.3. *Let $x < y$ be real numbers. There exists a rational number r such that*

$$x \leqslant r \leqslant y.$$

PROOF. We assume for simplicity that $0 \leqslant x < y$, the other possibilities being similar. Since $y - x > 0$, we find an integer $n \in \mathbf{N}$ such that

$$0 < \frac{1}{n} < y - x.$$

Let k be the element of $\mathbf{N}_0$ such that $k/n \leqslant x$ but $(k+1)/n > x$. Then because $(k+1)/n - k/n = 1/n < y - x$, it follows that

$$x < \frac{k+1}{n} \leqslant y$$

and so the rational number $(k+1)/n$ satisfies the condition we want. $\qquad\square$

1.6. Complex numbers

1.6.1. Definition and simple properties. Reference: [2, Kap. 3].

The set $\mathbf{C}$ of complex numbers completes the collection of number sets. Its additional property in comparison with $\mathbf{R}$ is that all polynomial equations of degree $n \geqslant 1$, of the form

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$$

with arbitrary coefficients $a_i \in \mathbf{C}$ and unknown $x \in \mathbf{C}$ have at least one solution.

REMARK 1.6.1. In other words, for any given coefficients $a_9, \ldots, a_{n-1}$, the map

$$f \colon \mathbf{C} \to \mathbf{C}$$

defined by

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

is surjective. In general it is not injective.

We recall a construction of $\mathbf{C}$ and some basic properties.

DEFINITION 1.6.2. We define $\mathbf{C} = \mathbf{R} \times \mathbf{R}$, with special elements $0_{\mathbf{C}} = (0,0)$ and $1_{\mathbf{C}} = (1,0)$ and $i = (0,1)$, and with operations

$$(a,b) + (c,d) = (a+c, b+d),$$
$$(a,b) \cdot (c,d) = (ac - bd, ad + bc).$$

We view an element (a,b) of $\mathbf{C}$ as a point in the plane $\mathbf{R} \times \mathbf{R}$. We have $(a,b) = (c,d)$ if and only if $a = c$ and $b = d$.

With these definitions, we see that

$$(a,b) = (a,0) + (0,b) = (a,0) + (0,1)(b,0) = (a,0) + i(b,0).$$

If we decide to identity $(a,0)$ with $a \in \mathbf{R}$ (as a point on the real axis in the plane), this becomes

$$(a,b) = a + ib.$$

Then the product rule follows by "the usual algebraic rules" together with the single relation $i^2 = (-1,0) = -1$. Indeed:

$$(a,b)(c,d) = (a+ib)(c+id) = ac + iad + ibc + i^2bd = (ac - bd) + i(ac + bd).$$

Moreover, when restricted to $\mathbf{R}$ viewed as a subset of $\mathbf{C}$, the addition and multiplication correspond to those of real numbers:

$$(a,0) + (b,0) = (a+b, 0)$$
$$(a,0) \cdot (b,0) = (ab, 0).$$

This justifies the view that $\mathbf{C}$ is an extension of $\mathbf{R}$.

With these operations, the properties of addition and multiplication of real numbers (except those that involve the order relation) are all true. We list them again for completeness:

Properties of addition. The following are true for any complex numbers:

(1.26) $$u + 0 = 0 + u = u$$

(1.27) $$u + v = v + u$$

(1.28) $$u + (v + w) = (u + v) + w$$

(1.29)   there is a unique complex number $-u$ such that $u + (-u) = (-u) + u = 0$.

Properties of multiplication. The following are true for any complex numbers:

(1.30) $$u \cdot 1 = 1 \cdot u = u$$

(1.31) $$uv = vu$$

(1.32) $$u(vw) = (uv)w$$

(1.33)
   if $u \neq 0$, there is a unique complex number u^{-1} such that $u \cdot u^{-1} = u^{-1} \cdot u = 1$.

Properties of addition and multiplication together. The following are true for any complex numbers:

(1.34) $$u \cdot 0 = 0 \cdot u = 0$$

(1.35) $$u(v + w) = uv + uw$$

(1.36) $$(u + v)w = uw + vw.$$

All of these are easy to check by direct computation, with the exception of the existence of the inverse u^{-1} of a non-zero complex number u. We check this by observing that if $u = (a, b) = a + ib$, then

$$(a + ib)(a - ib) = a^2 + b^2,$$

which is a strictly positive real number if $u \neq 0$ (since then either $a \neq 0$ or $b \neq 0$). So $uv = 1$ where

$$v = \frac{a}{a^2 + b^2} - i\frac{b}{a^2 + b^2}.$$

The fact that v is the unique solution of the equation $uv = 1$ is then elementary.
   We define two important functions that appeared in this small computation:

DEFINITION 1.6.3. Let $u = a + ib \in \mathbf{C}$.
(1) The *complex conjugate*, or simply conjugate,[24] of u is

$$\bar{u} = a - ib.$$

(2) The *modulus*[25] of u is

$$|u| = \sqrt{a^2 + b^2}.$$

Note that in defining $|u|$ we need the existence of the square root of any non-negative real number, which generalizes Theorem 1.5.1, and can be proved similarly; we will see how to do it more easily later, and for the moment take this existence for granted.

> PROPOSITION 1.6.4. *The following are true for all complex numbers:*
> (1) *We have* $|uv| = |u|\,|v|$.
> (2) *We have* $|u| = 0$ *if and only if* $u = 0$.
> (3) *We have* $|u + v| \leqslant |u| + |v|$.
> (4) *We have*
> $$\overline{u + v} = \bar{u} + \bar{v}, \quad \overline{uv} = \bar{u}\,\bar{v}$$
> (5) *We have* $u \in \mathbf{R}$ *if and only if* $\bar{u} = u$.
> (6) *We have* $u = a + ib$ *where* a *and* b *are given by*
> $$a = \frac{u + \bar{u}}{2}, \qquad b = \frac{u - \bar{u}}{2i}.$$

All of these properties are easy, except (3). It can however be understood geometrically: the modulus $|u|$ is the distance in the euclidean plane from the origine $0 = (0,0)$ to the point u, and more generally $|u - v|$ is the distance between the points u and v. Then (3) is the *triangle inequality*: in the triangle with vertices 0, u, $u + v$, the distance $|u + v|$ from 0 to $u + v$ is at most the sum of the distance $|u|$ from 0 to u and that from u to $u + v$, which is $|u + v - u| = |v|$.

REMARK 1.6.5. If $u \in \mathbf{R}$ then $|u| = \sqrt{u^2}$ is simply equal to u if $u \in \mathbf{R}_+$, and to $-u$ otherwise. The real number $|u|$ is also called the *absolute value* of u.

We now use these to check a special case of the fundamental theorem that we already mentioned:

> THEOREM 1.6.6. *Let* $n \geqslant 1$. *Let* $a_0, \ldots, a_n$ *be arbitrary complex numbers with* $a_n \neq 0$. *Then the equation*
> $$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0$$
> *with unknown* $x \in \mathbf{C}$ *has at least one solution.*

We prove this in the special case of the equation
$$x^2 = a_0,$$
namely we prove that every complex number is the square of another complex number.

Write $a_0 = a + ib$ and $x = c + id$ with $a, \ldots, d$ real. Then the equation $x^2 = a_0$ holds if and only if
$$\begin{cases} c^2 - d^2 = a, \\ 2cd = b. \end{cases}$$

We use a trick to see how to solve this easily: let $r_0 \geqslant 0$ be the modulus of a_0; note that any solution must satisfy $|x|^2 = |a_0|$, so that $c^2 + d^2 = r_0$. Combined with the first equation, we see that the only possibilities for c^2 and d^2 are

(1.37)
$$\begin{cases} c^2 = \frac{1}{2}(r_0 + a) \\ d^2 = \frac{1}{2}(r_0 - a). \end{cases}$$

Since $r_0^2 = a^2 + b^2$, we have $|a| \leqslant r_0$, which means that
$$-r_0 \leqslant a \leqslant r_0,$$
so that the numbers $a + r_0$ and $a - r_0$ are both non-negative. It follows that (by the extension of Theorem 1.5.1 to all non-negative real numbers) that real numbers c and d satisfying (1.37) exist. Then the equation $c^2 - d^2 = a$ holds, and moreover
$$2c^2d^2 = r_0^2 - a^2 = b^2,$$

and one needs simply to adjust the signs of c and d to ensure that $2cd = b$ (if $b \geqslant 0$, take c and d to be the square roots of $\frac{1}{2}(r_0 + a)$ and $\frac{1}{2}(r_0 - a)$, and if $b < 0$, take c to be the square root and d to be the opposite).

1.6.2. Sum and product notation. We will frequently need a notation for sums of many complex numbers, or products of them.

If I is a finite set, and we have complex numbers u_i defined for all $i \in I$ (which really means that we have a map $f \colon I \to \mathbf{C}$, where we use the notation u_i instead of $f(i)$), then we denote by

$$\sum_{i \in I} u_i$$

the sum of all u_i, and by

$$\prod_{i \in I} u_i$$

their product.

EXAMPLE 1.6.7. Let $I = \{1, 2, 3\}$. Then

$$\sum_{i \in I} u_i = u_1 + u_2 + u_3,$$

and

$$\prod_{i \in I} u_i = u_1 u_2 u_3.$$

By convention, if I is empty, the sum is equal to 0 and the product is equal to 1. This allows us to say that if I and J are finite sets without common element, then

$$\sum_{i \in I \cup J} u_i = \sum_{i \in I} u_i + \sum_{i \in J} u_i, \qquad \prod_{i \in I \cup J} u_i = \prod_{i \in I} u_i \cdot \prod_{i \in J} u_i.$$

Quite frequently, the set I is a subset of consecutive integers in $\mathbf{Z}$: for some $N \leqslant M$, we have

$$I = \{N, N+1, \dots, M\}.$$

In this case, we write

$$\sum_{i \in I} u_i = \sum_{n=N}^{M} u_n, \qquad \prod_{i \in I} u_i = \prod_{n=N}^{M} u_n.$$

1.6.3. Binomial coefficients and the binomial theorem. Reference: [2, 1.2].
The standard formula

$$(a + b)^2 = a^2 + 2ab + b^2$$

extends to complex numbers. Its generalization to other powers also does, and we recall the corresponding statement. This requires the binomial coefficients.[26]

DEFINITION 1.6.8 (Binomial coefficients). Let k and n be elements of $\mathbf{N}_0$. We denote by $\binom{n}{k}$ the binomial coefficient "n choose k",[27] which is the number of subsets of cardinality k in a set with n elements.

Note that this number does not depend on the set with n elements that is used (they are all "equivalent" for this purpose); we will usually use $I_n = \{1, \dots, n\}$.

EXAMPLE 1.6.9. We have $\binom{n}{k} = 0$, unless $0 \leqslant k \leqslant n$. Moreover

$$\binom{n}{0} = \binom{n}{n} = 1$$

(because the only subset with 0 elements is the empty set, and the only subset with n elements of I_n is the set I_n itself.

Moreover, we have

$$\binom{n}{1} = \binom{n}{n-1} = n,$$

because there are as many subsets of I_n with 1 elements as there are elements, and moreover, a subset with $n-1$ element is determined uniquely by the unique element of I_n that does not belong to it.

More generally, we get

$$\binom{n}{k} = \binom{n}{n-k}$$

because the subsets of I_n with k elements correspond to those with $n-k$ elements by replacing each subset S by its *complement*, the set of elements of I_n that are not in S.

We use the binomial coefficients to state and prove the binomial formula:[28]

THEOREM 1.6.10 (Binomial theorem). *Let $n \in \mathbf{N}_0$. For any complex numbers u and v, we have*

$$(u+v)^n = u^n + \binom{n}{1}u^{n-1}v + \cdots + \binom{n}{k}u^{n-k}v^k + \cdots + \binom{n}{n-1}uv^{n-1} + v^n$$

$$= \sum_{k=0}^{n} \binom{n}{k}u^{n-k}v^k.$$

PROOF. We first give an informal proof that explains why the binomial coefficients have to be there, but then also describe a proof by induction on n.

The n-th power of $u+v$ is a product of n terms each equal to $u+v$:

$$(u+v)^n = (u+v)\cdots(u+v).$$

We use the rule $(u+v)w = uw + vw$ repeatedly; one sees that this gives rise to 2^n terms, each of which is a product of n complex numbers z_i for $1 \leqslant i \leqslant n$, where z_i is either u or v. For instance

$$(u+v)(u+v) = u(u+v) + v(u+v) = uu + uv + vu + vv.$$

So each of the z_i is of the form $u^{n-k}v^k$ for some integer k, which is the number of times the factor v appears. This means that

$$(u+v)^n = c(n,0)u^n + c(n,1)u^{n-1}v + \cdots + c(n,k)u^{n-k}v^k$$
$$+ \cdots + c(n,n-1)uv^{n-1} + c(n,n)v^n,$$

where $c(n,k)$ is the number of terms z_i where we have picked k times the number v. This is equal to $\binom{n}{k}$, because each of these numbers z_i corresponds to the subset of $\{1,\ldots,n\}$ which indicates for which factor of the product we take the number v.

This explanation might look difficult to rigorously justify. We proceed then by induction on $n \in \mathbf{N}_0$. For $n = 0$, the formula is $1 = 1$, so it is correct. (If one prefers to start with the less obvious case $n = 1$, it is then $u + v = u + v$.)

Assume that the formula is true for $(u + v)^n$. Then

$$(u + v)^{n+1} = (u + v)(u + v)^n = (u + v) \sum_{k=0}^{n} \binom{n}{k} u^{n-k} v^k$$

$$= \sum_{k=0}^{n} \binom{n}{k} u^{n+1-k} v^k + \sum_{k=0}^{n} \binom{n}{k} u^{n-k} v^{k+1}.$$

In the second term, we use the new variable $\ell = k + 1$. The second term is then equal to

$$\sum_{\ell=1}^{n+1} \binom{n}{\ell - 1} u^{n+1-\ell} v^\ell.$$

We can rename the variable ℓ to k again and (separating the cases $k = 0$ and $k = n + 1$) we obtain

$$(u + v)^{n+1} = u^{n+1} + \sum_{k=1}^{n} \left(\binom{n}{k} + \binom{n}{k - 1} \right) u^{n+1-k} v^k + v^{n+1}.$$

Since $\binom{n+1}{0} = \binom{n+1}{n+1} = 1$, we will therefore conclude the proof as soon as we know that

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

We explain the proof of this below... □

LEMMA 1.6.11 (Pascal's Triangle). *For any n and k in $\mathbf{N}_0$, we have*

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

PROOF. The number $\binom{n+1}{k}$ of subsets of $I = \{1, \ldots, n+1\}$ of size k is equal to $A + B$, where A is the number of such subsets S where $n + 1 \notin S$ and B is the number of such subsets S where $n + 1 \in S$.

We have $A = \binom{n}{k}$, because A counts simply the k-element subsets of $\{1, \ldots, n\}$.

We have $B = \binom{n}{k-1}$, because in addition to $n + 1$, the subsets S must contain $k - 1$ elements of $\{1, \ldots, n\}$. □

EXAMPLE 1.6.12. The binomial coefficients are usually represented in a triangular form ("Pascal's Triangle", although it was known before Pascal), with the coefficients $\binom{n}{k}$ for a given n in each row. The lemma allows us to quickly compute a new row from the previous one, but putting in the k-th column the sum of the number above, and the one above and to the left (see below for that computation of $\binom{4}{2} = 6$).

n							
0	1						
1	1	1					
2	1	2		1			
3	1	3	+	3	1		
				‖			
4	1	4		6	4	1	
5	1	5		10	10	5	1
6	1	6		15	20	15	6	1

Another formula for binomial coefficients is the following:

PROPOSITION 1.6.13. *Let n and k be elements of $\mathbf{N}_0$ with $0 \leqslant k \leqslant n$. We have*

(1.38)
$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

For instance, we get
$$\binom{n}{2} = \frac{n!}{2(n-2)!} = \frac{n(n-1)}{2}.$$

PROOF. It is easiest here to use induction on $\mathbf{N}_0$, using Lemma 1.6.11. For $n = 0$ there is a single binomial coefficient equal to 1 and the right-hand side is
$$\frac{0!}{0!0!} = 1.$$

Now suppose that the formula (1.38) holds for a given n and all $0 \leqslant k \leqslant n$. Let then k be such that $0 \leqslant k \leqslant n+1$. Then using the formula $m! = m(m-1)!$ multiple times, we get

$$\frac{(n+1)!}{k!(n+1-k)!} = \frac{n!(n+1-k+k)}{k!(n+1-k)!} = \frac{n!(n+1-k)}{k!(n+1-k)!} + \frac{n!k}{k!(n+1-k)!}$$
$$= \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-(k+1))!}.$$

Using the induction hypothesis twice, we see that this is equal to

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

by Lemma 1.6.11. $\qquad\square$

CHAPTER 2

Constructing real numbers

In this chapter, we consider two general methods to construct new real numbers: using the *infimum* or the *supremum* of a suitable set of real numbers, or using infinite sequences or sums. The second are also extremely useful to *approximate* real numbers, usually with rational numbers. We will in particular explain how this leads to the usual decimal expansion of a real number.

2.1. Intervals

The most important sets of real numbers for analysis are *intervals*. These are sets defined as contain those real numbers that are either (1) between two bounds; (2) larger than some bound; (3) smaller than some bound, and where moreover we allow "between", "larger" and "smaller" to be defined either by strict inequalities or by non-strict inequalities.

More precisely, we introduce the following notation. For *bounded intervals*, those "between" two real numbers, say a and b, with $a \leqslant b$, we have four possibilities:

$$[a, b] = \{x \in \mathbf{R} \mid a \leqslant x \leqslant b\}$$
$$[a, b[= \{x \in \mathbf{R} \mid a \leqslant x < b\}$$
$$]a, b] = \{x \in \mathbf{R} \mid a < x \leqslant b\}$$
$$]a, b[= \{x \in \mathbf{R} \mid a < x < b\}.$$

Note how the use of brackets indicates whether the endpoint a or b is included or not.

We then have intervals "larger" than a bound, for which we use the symbol $+\infty$ to indicate the absence of upper-bound:

$$[a, +\infty[= \{x \in \mathbf{R} \mid a \leqslant x\}$$
$$]a, +\infty[= \{x \in \mathbf{R} \mid a <\},$$

and similarly with $-\infty$:

$$] -\infty, b] = \{x \in \mathbf{R} \mid x \leqslant b\}$$
$$] -\infty, b[= \{x \in \mathbf{R} \mid x < b\}.$$

Finally, we view $\mathbf{R}$ as an interval, with no restriction either above or below, and we sometimes write

$$] -\infty, +\infty[= \mathbf{R}.$$

Note that the empty set is also an interval, for instance $\emptyset =]0, 0[$.

DEFINITION 2.1.1. A *closed* interval is an interval of the form either $[a, b]$, or $[a, +\infty[$ or $] -\infty, a]$, which means that the endpoints are included (when they exist). We also say that $\emptyset$ and $\mathbf{R}$ are closed intervals.

An *open* interval is an interval of the form either $]a, b[$, or $]a, +\infty[$ or $] -\infty, a[$, which means that the endpoints are excluded. We also say that $\emptyset$ and $\mathbf{R}$ are open intervals.

An interval like $[a, b[$ is neither open nor closed. The set $\mathbf{R}$ is both open and closed (another illustration of the fact that mathematical and common languages can have very different interpretations...)

All intervals can be characterized by the following property, which informally states that they are the subsets of $\mathbf{R}$ "without holes".

> PROPOSITION 2.1.2. *A subset $I \subset \mathbf{R}$ is an interval if and only if, whenever $a \leqslant b$ are elements of I, all real numbers c such that $a \leqslant c \leqslant b$ are also in I.*

(In other words: if $a \leqslant b$ are elements of I, then the interval $[a, b]$ is contained in I.)

It is easy to prove, simply using the definition, that any interval has this property. We will not prove the converse here, but it should feel intuitively reasonable (see Exercise 2.3.4).

2.2. Upper and lower bounds, minimum and maximum

In the completeness property of $\mathbf{R}$, we see that condition that a real number c is larger or equal to all elements of a subset $A \subset \mathbf{R}$. This type of constraints occurs frequently.

> DEFINITION 2.2.1. Let A be a set of real numbers and $c \in \mathbf{R}$.
> The number c is an *upper-bound* for A if all elements of A are $\leqslant c$.
> The number c is a *lower-bound* for A if all elements of A are $\geqslant c$.

EXAMPLE 2.2.2. (1) Any real number $c \leqslant 1$ is a lower-bound of the set $\mathbf{N}$ of natural numbers. On the other hand, the set $\mathbf{N}$ has no upper-bound (because there is no real number larger than all integers).

(2) The interval $I = [0, 1]$ has many upper-bounds and many lower-bounds: any real number $c \geqslant 1$ is an upper-bound (and no other, because if $c < 1$, then $a = 1$ is an element of $[0, 1]$ with $c < a$, which contradicts the definition of upper-bound), and any real number $c < 0$ is a lower-bound.

(3) If $A' \subset A$, then any upper-bound of A is an upper-bound of A', and similarly for lower-bounds.

(4) The interval $J =]0, 1[$ has the same upper-bounds and lower-bounds as the interval $[0, 1]$. Indeed, any $c \geqslant 1$ is an upper-bound of J. If $c < 1$, on the other hand, although $1 \notin J$, there is a real number $a \in J$ such that $c < a$: if $c \leqslant 0$, we can take $a = 0$, and otherwise the middle-point

$$a = \frac{1 + c}{2},$$

of the segment from c to 1 is in J, and $c < a$.

Examples (2) and (4) show that for a set $A \subset \mathbf{R}$, it is possible that an upper-bound of A belongs to A (like for $[0, 1]$ and the upper-bound 1), but it is also possible that A has upper-bounds, but none belongs to A.

In the first case, we note that the upper-bound c in A must be unique: any other upper-bound $c' \in A$ satisfies $c \leqslant c'$ (because $c \in A$ and c' is an upper-bound of A) and $c' \leqslant c$ (conversely). This justifies the following definition:

> DEFINITION 2.2.3 (Maximum and minimum). Let A be a set of real numbers. The *maximum* of A is the unique upper-bound of A that belongs to A, if it exists; it is then denoted $\max(A)$.
> The *minimum* of A is the unique lower-bound of A that belongs to A, if it exists; it is then denoted $\min(A)$.

EXAMPLE 2.2.4. The interval $I = [a, b]$, for $a \leqslant b$, satisfies

$$\max(A) = b, \qquad \min(A) = a.$$

The interval $]a, b[$ has no maximum and no minimum. The interval $[a, b[$ has no maximum, but has the minimum a.

REMARK 2.2.5. We note that if A has a maximum, then $\max(A)$ is the smallest upper-bound of A: in fact, if $c < \max(A)$, then the fact that $\max(A) \in A$ shows that c is not an upper-bound of A.

2.3. Infimum, supremum and completeness

Many sets of real numbers have upper bounds or lower bounds, but among these, many do not have a maximum or a minimum, as in the example of an open interval $]a, b[$. Nevertheless, in that case, the number b is clearly "the best" upper-bound for $]a, b[$: although it is not an element of the interval, it is an upper-bound, and no smaller real number is an upper-bound. For this reason, b is called the *supremum* of $]a, b[$.

This is a very general fact:

THEOREM 2.3.1. *Let A be a non-empty subset of real numbers which has at least one upper-bound. Then the set of upper-bounds of A has a minimum, called the* supremum *of A, which is denoted* $\sup(A)$.

PROOF. This is an important application of the property of completeness. We denote by B the set of all real numbers $b \in \mathbf{R}$ which are upper-bounds for A. By assumption, B is not empty, and by definition of upper-bounds, we have

$$a \leqslant b$$

for all $a \in A$ and $b \in B$.

We can therefore apply the completeness property to the sets A and B. Let c be the element "in the middle" given by this property. Then we have

$$a \leqslant c \leqslant b$$

for all $a \in A$ and $b \in B$. The inequality $a \leqslant c$ for all $a \in A$ means that c is an upper-bound for A; that means that $c \in B$. The inequality $c \leqslant b$ for all $b \in B$ means that c is a lower-bound for B. So c is a lower-bound of B that belongs to B, so it is the minimum of B. $\qquad \square$

EXAMPLE 2.3.2. (1) We have

$$\sup(]0, 1[) = 1$$

according to Example 2.2.2, (4), since the set of upper-bounds of $]0, 1[$ is $[1, +\infty[$, and this has minimum equal to 1.

(2) If $A \neq \emptyset$ has an upper-bound, then the set of upper-bounds of A is the interval

$$[\sup(A), +\infty[.$$

Indeed, $\sup(A)$ is an upper-bound, and all larger or equal real numbers $b \geqslant \sup(A)$ are of course upper-bounds also.

(3) If A has a maximum, then this is also its supremum: $\sup(A) = \max(A)$ when $max(A)$ exists. Indeed, we have seen this in Remark 2.2.5.

(3) Let

$$A = \{x \in \mathbf{Q}_+ \mid x^2 \leqslant 2\}.$$

Then A is not empty and 2 is an upper-bound. So it has a supremum, and in fact
$$\sup(A) = \sqrt{2}.$$
(so this is another case where the supremum does not belong to A).

Indeed, let $c = \sup(A)$; it is a positive number, and we need to prove that $c^2 = 2$. Note first that if $x \in A$, then $x \leqslant c$, so that $x^2 \leqslant c^2$; since $x^2 \leqslant 2$ by definition of A, we must have $c^2 \geqslant 2$. It remains to check that $c^2 \leqslant 2$. Assume that we had $c^2 < 2$. Then as in the proof of Theorem 1.5.1, we can find $e > 0$ such that $(c+e)^2 < 2$. Now we can find a rational number x with $c < x < c + e$ (because, by Theorem 1.5.3, there is always a rational number between two real numbers $x < y$), and then $x^2 \leqslant (c+e)^2 < 2$, so that x would be an element of A larger than the upper-bound c, which is impossible.

REMARK 2.3.3. (1) The statement of the theorem (that all non-empty subsets of $\mathbf{R}$ bounded from above have a supremum) is equivalent to the property of completeness. To see this, assume that the theorem is known to be true, and let us try to prove the completeness property. So let A and B be non-empty sets of real numbers, with $a \leqslant b$ for all $a \in A$ and $b \in B$. Then A has an upper bound (any element of B, which is not empty), so we can define $c = \sup(A)$, since we assume that Theorem 2.3.1 is true. Since c is an upper-bound of A, we have $a \leqslant c$ for all $a \in A$. And since every element b of B is an upper-bound of A, we have $c \leqslant b$, since the supremum of A is the minimum of all upper-bounds. So we get
$$a \leqslant c \leqslant b,$$
which means that c is "in the middle", and this establishes the completeness property.

(2) The analogue of the theorem for lower-bounds is that if $A \subset \mathbf{R}$ is a non-empty set of real numbers which has a lower-bound, then the set of its lower-bounds has a maximum, called the *infimum* of A, which is denoted $\inf(A)$.

For instance, we have $\inf(]0, 1[) = 0$.

(3) When a non-empty set $A \subset \mathbf{R}$ has no upper-bound, it cannot have a supremum; we use the notation
$$\sup(A) = +\infty$$
to indicate this concisely. Similarly, if A has no lower-bound, we write
$$\inf(A) = -\infty.$$

(4) Let A be a non-empty subset of $\mathbf{R}$ and $B \subset \mathbf{R}$ such that $A \subset B$. Then the following elementary facts are often useful to compare the possible supremum and infimum of A and B:

- If B has an upper-bound, then so does A and
$$\sup(A) \leqslant \sup(B),$$

- If B has a lower-bound, then so does A and
$$\inf(B) \leqslant \inf(A).$$

This is because, in the first case for instance, the set of upper-bounds of B contains the set of upper-bounds of A.

EXERCISE 2.3.4. Let $I \subset \mathbf{R}$ be a subset of $\mathbf{R}$ such that whenever $a \leqslant b$ are in I, the interval $[a, b]$ is contained in A. Show that I is an interval.

(Hint: consider first the case when I is not empty and bounded from above and below; let $x = \inf(I)$ and $y = \sup(I)$; show that I is one of the intervals $[x, y]$, $]x, y]$, $[x, y[$ or $]x, y[$, depending on whether I has minimum or maximum, or both, or none.)

To summarize the basic idea of the supremum: suppose that we have a subset A of real numbers which is not empty and has an upper-bound, so that $\sup(A)$ exists. How does one prove that a given real number $c \in \mathbf{R}$ is equal to $\sup(A)$? Usually, the best method is to prove the two inequalities

(2.1) $$c \geqslant \sup(A)$$

(2.2) $$c \leqslant \sup(A)$$

separately. Of course, together they imply that $c = \sup(A)$ (but either can be proved separately, if we only want to show that c satisfies one of these).

To prove (2.1), it suffices to prove that c is an upper bound of A, or in other words that

$$a \leqslant c$$

for all elements $a \in A$.

To prove (2.2), we need to prove that any real number $d < c$ is not an upper bound of A. In other words, for all $d < c$, we need to find an element $a \in A$ such that $d < a$.

For many sets of real numbers, what we would really like to know is that they have a maximum. But to prove this is often difficult. However, it is usually much easier to prove that the set has an upper-bound; then the supremum always exists, and one can try, as a second step, to check if the supremum belongs to the set of interest. We will see various ways of ensuring that.

2.4. Sequences

We now have ways to construct many interesting real numbers (solutions of equations of various types in particular), using the supremum or infimum of sets of real numbers. However, we would like to also be able to *compute* concretely with these numbers. We can do this by using "sequences"

$$x_1, \ x_2, \ x_3, \ \ldots, \ x_n, \ \ldots$$

of rational numbers that get "closer and closer" to a given real number. The definition of sequences is simple, and we can extend it to complex numbers right away:

DEFINITION 2.4.1. A *sequence*[29] of complex numbers is a map $s \colon \mathbf{N} \to \mathbf{C}$. If all values of s are real numbers, we say that it is a *real sequence*.

In terms of notation, a sequence s with $s(n) = a_n$ is often denoted $s = (a_n)_{n \in \mathbf{N}}$, or simply $s = (a_n)$ or $s = (a_n)_n$. Two sequences (a_n) and (b_n) are equal if and only if $a_n = b_n$ for all $n \in \mathbf{N}$.

REMARK 2.4.2. (1) A sequence is *ordered*, and can have repetitions, and must not be confused with the *set of values* $\{a_n\}$. For instance, the sequence $s(n) = 2$ for all $n \in \mathbf{N}$ is an infinite sequence where all terms are equal to 2, and the sequences

$$(1, -1, 1, -1, 1, -1, \ldots)$$

and

$$(-1, 1, -1, 1, -1, 1, \ldots)$$

are distinct (although the sets of values of both sequences are the same set $\{-1, 1\}$).

In particular, to avoid confusion, we speak of the terms of the sequence, and say that a_n is the n-term of the sequence $(a_n)_{n \in \mathbf{N}}$.

(2) It is sometimes more convenient to order a sequence using $\mathbf{N}_0$, which amounts to considering maps $s\colon \mathbf{N}_0 \to \mathbf{C}$. In this case, one writes $s = (a_n)_{n\in\mathbf{N}_0}$, or simply (a_n) if it is clear that the first term is a_0.

EXAMPLE 2.4.3. (1) Constant sequences are defined by $s(n) = a$ for all n, where $a \in \mathbf{C}$ is a fixed number.

(2) [Arithmetic progressions][30] Let a and b be complex numbers. The sequence

$$(an + b)_{n\in\mathbf{N}_0} = (b, a + b, 2a + b, \ldots)$$

is called an arithmetic progression; the number a is called the *common difference*[31] (because two consecutive terms

$$an + b, \quad a(n + 1) + b$$

always differ by a).

(3) [Geometric progressions][32] Let a and b be complex numbers. The sequence

$$(ba^n)_{n\in\mathbf{N}_0} = (b, ab, a^2b, \ldots)$$

is called a geometric progression; the number a is called the *common ratio* (because, if a and b are non-zero, two consecutive terms

$$ba^n, \quad ba^{n+1}$$

have quotient equal to a).

(4) [Sequences defined by induction] Suppose that $f\colon \mathbf{C} \to \mathbf{C}$ (or $f\colon \mathbf{R} \to \mathbf{R}$) is a function. Given $a \in \mathbf{C}$, we can then define a sequence (a_n) by starting with a and applying f repeatedly: we define $a_1 = a$ and $a_{n+1} = f(a_n)$ for all $n \in \mathbf{N}$. (So, for instance, we have $a_3 = f(f(f(a)))$.) Such a sequence is said to be defined inductively (and the initial term could also be a_0).

For instance, arithmetic and geometric progressions are of this type: the sequence $(an + b)_{n\in\mathbf{N}_0}$ can also be described by

$$a_0 = b, \qquad a_{n+1} = a + a_n,$$

and the sequence $(ba^n)_{n\in\mathbf{N}_0}$ can be described by

$$a_0 = b, \qquad a_{n+1} = aa_n.$$

The Fibonacci numbers of Example 1.2.3, (2), are also defined inductively, but this time with two initial terms $a_1 = a_2 = 1$ and a rule of the form

$$a_{n+2} = f(a_n, a_{n+1}),$$

where $f(x, y) = x + y$. This can of course be generalized to other functions.

In general, it is however impossible to find a "simple" expression for the n-th term of a sequence which is defined by induction (a formula for a_n that does not mention previous terms of the sequence).

One can operate on sequences, by performing the operations for each value separately. More generally:

DEFINITION 2.4.4. Let X be an arbitrary set. For functions f_1 and f_2 from X to $\mathbf{C}$, we denote by $f_1 + f_2$ the sum of the functions, defined by

$$(f_1 + f_2)(x) = f_1(x) + f_2(x)$$

for all $x \in X$, and by $f_1 f_2$ the product, defined by

$$(f_1 f_2)(x) = f_1(x) f_2(x).$$

If f_2 is such that $f_2(x) \neq 0$ for all x, then f_1/f_2 is defined by

$$(f_1/f_2)(x) = f_1(x)/f_2(x).$$

So, for instance, the sum of the arithmetic progression $s_1 = (2n)_{n \in \mathbf{N}_0}$ and of the geometric progression $s_2 = (3^n)_{n \in \mathbf{N}_0}$ is the sequence

$$s_3 = (2n + 3^n)_{n \in \mathbf{N}_0} = (1, 5, 13, \ldots)$$

REMARK 2.4.5. In the language of linear algebra, the set E of complex-valued (resp. real-valued) functions from X to $\mathbf{C}$ (resp. from X to $\mathbf{R}$) is a complex (resp. real) vector-space; the zero vector is the constant function always equal to 0.

DEFINITION 2.4.6. A sequence $s = (a_n)$ is *bounded* if and only if there exists a real number $R \geqslant 0$ such that

$$|a_n| \leqslant R$$

for all $n \in \mathbf{N}$.

EXAMPLE 2.4.7. (1) An arithmetic progression $(an + b)_{n \in \mathbf{N}_0}$ is bounded if and only if $a = 0$. Indeed, if $a = 0$, then the sequence is constant; conversely, if $a \neq 0$, then using the lower bound

$$|an + b| \geqslant |a|n - |b|$$

we see that, for any real number $R \geqslant 0$, we have $|an + b| > R$ for any integer $n > (R + |b|)/|a|$.

(2) A geometric progression $(ba^n)_{n \in \mathbf{N}_0}$ with $b \neq 0$ is bounded if and only if $|a| \leqslant 1$. Indeed, if $|a| \leqslant 1$, then

$$|ba^n| = |b||a|^n \leqslant |b|$$

for all $n \in \mathbf{N}_0$. Conversely, if $|a| > 1$, then we have $|ba^n| = |b||a|^n$. If we write $|a| = 1 + e$ with $e > 0$, then by the binomial formula, we get

$$|ba^n| = |b|(1 + e)^n = |b|\left(1 + ne + \binom{n}{2}e^2 + \cdots\right) \geqslant |b|(1 + ne),$$

and this becomes $> R$ when $n > R/(e|b|)$.

2.5. Convergence of sequences of complex numbers

Reference: [2, 5.1, 5.2].

We are now ready to give the most important definition of analysis, that of *convergence* of a sequence of real or complex numbers.

The intuition is the following: a sequence (a_n) converges to a real number a if "the numbers a_n get closer and closer, and arbitrarily close, to a as n increases". This means that a_n, for n large enough, can be used to *approximate* a with *arbitrary precision*.

EXAMPLE 2.5.1. When we write

$$\pi = 3.14159265358979323846264338327950288842\ldots,$$

what we mean is that the sequence

$$(3,\ 3.1,\ 3.14,\ 3.141,\ 3.1415,\ \ldots)$$

(whose terms are rational numbers, in that case) gives better and better approximations of the real number π. This sequence *converges* to π.

In order to make this definition precise, we simply need to specify what "closer and closer", or "arbitrarily close", really means. Intuitively, this means that the difference, or rather the absolute value of the difference, is not too large, or becomes smaller as n increases.

DEFINITION 2.5.2. Let $\varepsilon > 0$ be a positive real number. Two complex numbers a and b are at distance less than ε if $|a - b| < \varepsilon$.

REMARK 2.5.3. If a and b are real numbers, this means that

$$a - \varepsilon < b < a + \varepsilon,$$

which one can also summarize as saying that b is an approximation of a with error at most ε.

Geometrically, if we view a as fixed, the set of complex numbers b at distance less than ε of a is the "interior" of the disc centered at a with radius ε (this excludes the circle, defined by $|a - b| = \varepsilon$).

This leads to the definition of convergence

DEFINITION 2.5.4. Let $s = (a_n)_{n \in \mathbf{N}}$ be a sequence of complex numbers and $a \in \mathbf{C}$. The sequence s *converges to a as n tends to infinity*, denoted

$$a_n \longrightarrow a, \qquad \text{or} \qquad \lim_{n \to +\infty} a_n = a,$$

if the following is true: for any positive real number ε, there exists an integer N, depending on ε, such that

$$|a_n - a| < \varepsilon$$

whenever $n \geqslant N$. We then also say that a is the *limit*[33] of the sequence (a_n).

As a matter of terminology, one sometimes also says that a_n tends to a as $n \to +\infty$.

REMARK 2.5.5. (1) The precise translation in a logical formula is that (a_n) converges to a if and only if

$$(2.3) \qquad \forall\, \varepsilon > 0,\ \exists\, N \in \mathbf{N},\ \forall\, n \geqslant N,\ |a_n - a| < \varepsilon$$

(2) It is intuitively clear that if a sequence (a_n) converges, then the limit a is unique: a sequence cannot converge to two different numbers. Let us check this precisely: suppose that (a_n) converges to a and b. Pick any $\varepsilon > 0$; there exist by assumption an integer N_1 large enough so that $|a_n - a| < \varepsilon/2$ for $n \geqslant N_1$, and an integer N_2 large enough so that $|a_n - b| < \varepsilon/2$ for $n \geqslant N_2$. If n is the larger of N_1 and N_2, then we get

$$|a - b| \leqslant |a - a_n| + |a_n - b| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

by the triangle inequality. So $|a - b|$ is smaller than any positive real number ε, which is only possible if $|a - b| = 0$ (otherwise, take $\varepsilon = |a - b|$ to get a contradiction), or in other words if $a = b$.

(3) It is important to remark that *many sequences do not converge*. In principle, before speaking of the value of the limit, and attempting to use it in computations, one must prove that it exists. Doing otherwise may lead to serious problems (as we will

illustrate on some examples below). For instance, the sequence defined by $a_n = (-1)^n$ does not converge to any real number (check this rigorously).

In general, unless stated differently, when we state a result in the form *we have*

$$\lim_{n \to +\infty} a_n = (\text{something})$$

for some sequence (a_n), the meaning is that we claim *first* that the limit exists, and *secondly* that it has the value on the right-hand side.

(4) A sequence of real numbers can only converge to a real limit (check this as an exercise).

EXAMPLE 2.5.6. (1) A constant sequence converges to the corresponding constant value.

(2) Suppose that (a_n) and (b_n) are sequences which coincide except for a finite number of values of n: there exists an integer M such that $a_n = b_n$ for all $n \geqslant M$. Then the sequence (a_n) converges if and only if (b_n) does, and when this is the case, their limits are the same.

To see this, suppose that (a_n) converges to a. Let's prove that (b_n) also does. Let $\varepsilon > 0$ be given; by assumption there exists $N \in \mathbf{N}$ such that $|a_n - a| < \varepsilon$ for $n \geqslant N$. Now let N' be the larger of N and M; for all $n \geqslant N'$ we have

$$|b_n - a| = |a_n - a| < \varepsilon.$$

This means that the sequence (b_n) converges to a. Exchanging the role of the two sequences gives the converse assertion.

We express this property by saying that *convergence is an asymptotic property of the sequence*; it doesn't depend on the first terms of the sequence, but only on what happens for n getting larger and larger; another asymptotic property is that of being bounded by *some $R \geqslant 0$*, although the precise bound may depend on the first terms.

(3) In some sense, in order to understand convergence, we only need to understand sequences of non-negative real numbers that converge to 0. Indeed, an arbitrary sequence $s = (a_n)$ of complex numbers converges to a if and only if the sequence (b_n) with $b_n = |a_n - a|$ converges to 0. This is simply because $|b_n - 0| = b_n = |a_n - a|$.

(4) In this respect, the following fact is then very useful: if (b_n) converges to 0, and if the sequence (c_n) is bounded, then $(b_n c_n)$ also converges to 0. Indeed, suppose that $|c_n| \leqslant R$ for all n, for some $R > 0$; then to have $|b_n c_n| < \varepsilon$, it suffices to have $|b_n| < \varepsilon/R$, which is true for all n large enough.

A very important fact, which we will interpret also later as a case of *continuity* is that we can operate easily on convergent sequences. Before stating this, we have two elementary but very useful lemmas.

LEMMA 2.5.7. *Let (a_n) be a convergent sequence. Then (a_n) is bounded.*

PROOF. Let a be the limit of the sequence. Take $\varepsilon = 1$ in the definition; we find $N \in \mathbf{N}$ such that $|a_n - a| < 1$ for all $n \geqslant N$. Then

$$|a_n| \leqslant |a| + |a_n - a| \leqslant |a| + 1$$

for all $n \geqslant N$. Let

$$R = \max(|a_1|, \ldots, |a_{N-1}|, |a| + 1).$$

Then we get $|a_n| \leqslant R$ for all n. $\qquad\square$

Another lemma is very convenient to check convergence properties.

LEMMA 2.5.8 (Convergence by comparison). *Let (a_n) be a sequence, $a \in \mathbf{C}$, and (b_n) a sequence of non-negative real-numbers that converges to 0. Assume that there exists $M \in \mathbf{N}$ such that for $n \geqslant M$, we have*

$$|a_n - a| \leqslant b_n.$$

Then (a_n) converges to a.

PROOF. This follows from the definition: given any $\varepsilon > 0$, there exists $N \in \mathbf{N}$ such that $|b_n| = b_n < \varepsilon$ for $n \geqslant N$; if n is at least the larger of M and N, then we get $|a_n - a| \leqslant b_n < \varepsilon$. $\qquad\square$

PROPOSITION 2.5.9. *Let $s_1 = (a_n)$ and $s_2 = (b_n)$ be sequences. Assume that s_1 converges to a and that s_2 converges to b.*
   (1) *The sequence $s_1 + s_2$ converges to $a + b$.*
   (2) *The sequence $s_1 s_2$ converges to ab.*
   (3) *If $b \neq 0$, then there exists $M \in \mathbf{N}$ such that $b_n \neq 0$ for all $n \geqslant M$; the sequence (c_n) defined by*

$$c_n = \begin{cases} 0 & \text{if } n < M \\ a_n/b_n & \text{if } n \geqslant M \end{cases}$$

converges to a/b.
   (4) *If s_1 and s_2 are real sequences and $a_n \geqslant b_n$ for all n, then $a \geqslant b$.*

PROOF. For the proof of (1), let $\varepsilon > 0$. There exist by assumption an integer N_1 so that $|a_n - a| < \varepsilon/2$ for $n \geqslant N_1$, and an integer N_2 so that $|b_n - b| < \varepsilon/2$ for $n \geqslant N_2$. If N is the larger of N_1 and N_2, then we get

$$|(a_n + b_n) - (a + b)| = |a_n - a + b_n - b| \leqslant |a_n - a| + |b_n - b| < \varepsilon$$

for $n \geqslant N$ by the triangle inequality. This proves that $(a_n + b_n)_{n \in \mathbf{N}}$ converges to $a + b$.

For the proof (2) and (3), we will use another presentation of the argument, which is often much more convenient: for (2), we attempt to bound from above the quantity $|a_n b_n - ab|$, without assumption on n at first; we then try to check that the resulting upper-bound tends to 0 as $n \to +\infty$, and then use Lemma 2.5.8.

The trick in (2), which is suggested by the fact that we know that $a_n - a$ and $b_n - b$ are small for large n, is to write

$$ab - a_n b_n = a(b - b_n) + ab_n - a_n b_n = a(b - b_n) + (a - a_n)b_n.$$

This expresses the sequence $(a_n b_n - ab)_{n \in \mathbf{N}}$ as the sum of the sequence $(a(b - b_n))_{n \in \mathbf{N}}$ and the sequence $((a - a_n)b_n)_{n \in \mathbf{N}}$. But we claim that both of these converge to 0; if this is true then (by part (1)), the sequence $(a_n b_n - ab)_{n \in \mathbf{N}}$ tends to 0.

But

$$|a(b - b_n)| = |a||b - b_n|, \qquad |b_n(a - a_n)| = |b_n||a - a_n|,$$

so the two sequences are expressed as the product of a bounded sequence (in the first case, a constant sequence) multiplied by a sequence converging to 0; they converge to 0 by Example 2.5.6 (4).

We now prove (3). First, we need to show that the terms of the sequence (b_n) are non-zero when n is large; this is because they become close enough to $b \neq 0$. Precisely, apply the definition of convergence with $\varepsilon = \frac{1}{2}|b|$: there exists M such that $|b_n - b| < \frac{1}{2}|b|$ for $n \geqslant M$, and then we get

(2.4) $$|b_n| \geqslant |b| - |b - b_n| \geqslant \tfrac{1}{2}|b| > 0$$

by the triangle inequality, which implies that $b_n \neq 0$ for $n \geqslant M$.

Now, we proceed a bit as in (2): we have

$$\frac{a_n}{b_n} - \frac{a}{b} = \frac{a_n b - a b_n}{b b_n}.$$

This gives

$$\left| \frac{a_n}{b_n} - \frac{a}{b} \right| = \frac{|a_n b - a b_n|}{|b b_n|} \leqslant \frac{2}{|b|^2} |a_n b - a b_n|$$

for $n \geqslant M$, according to (2.4). But note that the sequence $(a_n b - a b_n)_n$ converges to 0, by (1) applied to $(a_n b)$ and $(-a b_n)$ (which converge to ab and $-ab$, respectively, by (2)); then Example 2.5.6, (4), and Lemma 2.5.8, imply that $a_n / b_n \to a/b$.

Finally, to prove (4), we first assume that $a_n \geqslant 0$ for all n, and prove that $a \geqslant 0$. Indeed, suppose that $a < 0$; let $\varepsilon = |a|/2 > 0$; then for n large enough we have $|a_n - a| < \varepsilon$, and then

$$a_n < a + \varepsilon = a + \tfrac{1}{2}|a| = \tfrac{1}{2}a < 0,$$

which is a contradiction. $\qquad\square$

REMARK 2.5.10. A warning! Suppose that (a_n) and (b_n) are convergent sequences and that we know that $a_n > b_n$ for all $n \in \mathbf{N}$; we can then only conclude in general that $a \geqslant b$, *and not that $a > b$.* An example that illustrates this is $a_n = 1/n$ and $b_n = 0$; then $a_n > b_n$, but both sequences converge to 0.

We can also reduce in principle the study of limits of complex sequences to the case of real ones.

PROPOSITION 2.5.11. *Let (a_n) be a sequence of complex numbers, and write $a_n = x_n + i y_n$ with x_n and y_n real. Let $a = x + iy \in \mathbf{C}$ with x and y real.*
(1) *The sequence (a_n) converges to a if and only if $x_n \to x$ and $y_n \to y$.*
(2) *If (a_n) converges to a, then $(\bar{a}_n)$ converges to $\bar{a}$, and $(|a_n|)$ converges to $|a|$.*

PROOF. Note first that

$$|a_n - a| \leqslant |x_n - x| + |i(y_n - y)| = |x_n - x| + |y_n - y|$$

for all $n \in \mathbf{N}$, by the triangle inequality. If (x_n) converges to x and (y_n) converges to y, then by Proposition 2.5.9, (1), the right-hand sequence converges to 0; by Lemma 2.5.8, it follows that $a_n \to a$.

Conversely, we use the inequalities

$$|x_n - x| \leqslant \sqrt{|x_n - x|^2 + |y_n - y|^2} = |a_n - a|, \quad |y_n - y| \leqslant |a_n - a|$$

to conclude.

Since $\bar{a}_n = x_n - i y_n$, part (1) implies immediately the first part of (2). Then the triangle inequality implies that

$$||a_n| - |a|| \leqslant |a_n - a|,$$

because

$$|a| - |a_n - a| \leqslant |a_n| \leqslant |a| + |a_n - a|,$$

so that the second part follows. $\qquad\square$

2.6. Some basic limits

The next result gives some of the most fundamental examples of convergence; they should be kept in mind in particular as good tools for applications of Lemma 2.5.8.

PROPOSITION 2.6.1. *We have the following limits:*

$$(2.5) \qquad \lim_{n \to +\infty} \frac{1}{n^k} = 0 \qquad \text{for all } k > 0,$$

$$(2.6) \qquad \lim_{n \to +\infty} a^n = 0 \qquad \text{if } |a| < 1,$$

$$(2.7) \qquad \lim_{n \to +\infty} \frac{n^k}{b^n} = 0 \qquad \text{if } k \in \mathbf{R} \text{ and } |b| > 1,$$

$$(2.8) \qquad \lim_{n \to +\infty} \frac{a^n}{n!} = 0 \qquad \text{for all } a \in \mathbf{R}.$$

So for instance, we get

$$\frac{1}{\sqrt{n}} \to 0, \quad \frac{1}{2^n} \to 0, \quad \frac{n^3}{2^n} \to 0, \quad \frac{4^n}{n!} \to 0.$$

PROOF. The limit (2.5) is left as an exercise. The case of (2.6) follows from (2.7) applied with $k = 0$ and $b = 1/a$. The limit (2.7) is easiest to prove using some later results, so we omit the proof for the moment (see Example 2.8.4). Finally, we prove (2.8) by noting that

$$\left| \frac{a^n}{n!} \right| = \frac{|a| \cdots |a|}{1 \cdot 2 \cdots n}.$$

where the numerator has n factors. Let N be an integer larger than $2|a|$. Then for $n \geqslant N$, we get

$$\left| \frac{a^n}{n!} \right| = \frac{|a|^N}{N!} \frac{1}{2^{n-N}} = \frac{(2|a|)^N}{N!} \frac{1}{2^n},$$

which converges to 0 according to (2.6), since the first factor is a constant sequence. $\square$

EXAMPLE 2.6.2. We can now compute many other limits. For instance, let $k \leqslant l$ be positive integers, let $(a_0, \ldots, a_k)$ be complex numbers with $a_k \neq 0$, and let $(b_0, \ldots, b_l)$ be complex numbers with $b_l \neq 0$. Then

$$(2.9) \qquad \lim_{n \to +\infty} \frac{a_k n^k + \cdots + a_1 n + a_0}{b_l n^l + \cdots + b_1 n + b_0} = \begin{cases} \frac{a_k}{b_l} & \text{if } k = l \\ 0 & \text{if } k < l. \end{cases}$$

To see this, we use a useful trick: we write

$$a_k n^k + \cdots + a_1 n + a_0 = a_k n^k \left(1 + \frac{a_{k-1}}{a_k} \frac{1}{n} + \cdots + \frac{a_0}{a_k} \frac{1}{n^k} \right),$$

$$b_l n^l + \cdots + b_1 n + b_0 = b_l n^l \left(1 + \frac{b_{l-1}}{b_l} \frac{1}{n} + \cdots + \frac{b_0}{b_l} \frac{1}{n^l} \right)$$

This allows us to write

$$\frac{a_k n^k + \cdots + a_1 n + a_0}{b_l n^l + \cdots + b_1 n + b_0} = \frac{a_k}{b_l} \frac{1}{n^{l-k}} \frac{c_n}{d_n}$$

with

$$c_n = 1 + \frac{a_{k-1}}{a_k} \frac{1}{n} + \cdots + \frac{a_0}{a_k} \frac{1}{n^k}, \quad d_n = 1 + \frac{b_{l-1}}{b_l} \frac{1}{n} + \cdots + \frac{b_0}{b_l} \frac{1}{n^l}$$

By Proposition 2.5.9, (1) and (3), and (2.5), we see that

$$\lim_{n \to +\infty} \frac{c_n}{d_n} = 1.$$

By (2.5) again if $l > k$, and $n^0 = 1$ if $k = l$, we have

$$\lim_{n \to +\infty} \frac{a_k}{b_l} \frac{1}{n^{l-k}} = \begin{cases} 0 & \text{if } l > k \\ \frac{a_k}{b_l} & \text{if } l = k. \end{cases}$$

Since this limit exists, using once more Proposition 2.5.9, (2), we deduce that

$$\lim_{n \to +\infty} \frac{a_k n^k + \cdots + a_1 n + a_0}{b_l n^l + \cdots + b_1 n + b_0} = \lim_{n \to +\infty} \frac{a_k}{b_l} \frac{1}{n^{l-k}},$$

which is what we claimed.

We will see later that when $k > l$, the sequence in the left-hand side of (2.9) does not converge.

2.7. The decimal expansion of a real number

We will now explain how the decimal expansion of a real number may be constructed and interpreted as a "natural" sequence of rational numbers converging to a given real number.

Let

$$D = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}.$$

This is the finite set of (decimal) digits.

Let $a \in \mathbf{R}_+$. There exists a unique element of $\mathbf{N}_0$, denoted $\lfloor a \rfloor$, such that

$$\lfloor a \rfloor \leqslant a < \lfloor a \rfloor + 1.$$

Let $a' = a - \lfloor a \rfloor$, so that $0 \leqslant a' < 1$.

We now construct inductively sequences $(d_n)_{n \in \mathbf{N}}$ and $(a_n)_{n \in \mathbf{N}}$, with $d_n \in D$ and $a_n \in \mathbf{Q} \cap [0, 1[$, such that

(2.10) $$a_n \leqslant a' < a_n + 10^{-n}$$

for all $n \in \mathbf{N}$.

For $n = 1$, we let $d_1 \in D$ be the unique $d \in D$ such that

$$a' \in \left[\frac{d}{10}, \frac{d+1}{10} \right[,$$

(note that these intervals have no common points, and that their union is the interval $[0, 1[$), and we denote $a_1 = d_1/10$. We then have $a_1 \leqslant a' \leqslant a_1 + 1/10$, so the first step of the induction is done.

If we assume that d_n and a_n are defined and that (2.10) holds, we define d_{n+1} to be the unique $d \in D$ such that

$$a' \in \left[a_n + \frac{d}{10^{n+1}}, a_n + \frac{d+1}{10^{n+1}} \right[$$

(in other words, knowing that

$$a_n \leqslant a' < a_n + 10^{-n},$$

we split the interval of "uncertainty" in ten equal size sub-intervals, and look in which of these a' can be found). We put

(2.11) $$a_{n+1} = a_n + \frac{d_{n+1}}{10^{n+1}}.$$

By construction, we have then

$$a_{n+1} \leqslant a' < a_n + \frac{d_{n+1} + 1}{10^{n+1}} = a_{n+1} + \frac{1}{10^{n+1}},$$

veryfing (2.10) at the next step, and so the inductive construction is done.

The construction shows that we can express a_n in the more concise more

$$a_n = \frac{d_1}{10} + \cdots + \frac{d_n}{10^n},$$

or

$$a_n = 0.d_1 \cdots d_n$$

in decimal notation.

THEOREM 2.7.1 (Decimal expansion). *The sequence* $(\lfloor a \rfloor + a_n)_{n \in \mathbf{N}}$, *where* (a_n) *is defined above, converges to* a.

PROOF. According to the property (2.10) in the construction of the decimal expansion, we have

$$|\lfloor a \rfloor + a_n - a| = |a_n - a'| \leqslant \frac{1}{10^n}$$

for all $n \in \mathbf{N}$. Since the sequence (10^{-n}) tends to 0 (by (2.6)), we deduce the convergence from Lemma 2.5.8. $\square$

We will say that the *integral part* $\lfloor a \rfloor$ and the sequence $(d_n)_{n \in \mathbf{N}}$ define the *decimal expansion* of a non-negative real number a. These are uniquely determined; if $a \in \mathbf{R}$ is negative, then we also need to recall the minus sign.

EXAMPLE 2.7.2. Let $a = \sqrt{2}$. Since the function $f(x) = x^2$ on $\mathbf{R}_+$ is increasing, we have $1 \leqslant a < 2$. So $\lfloor a \rfloor = 1$ and $a' = \sqrt{2} - 1$.

By computing $(1 + d/10)^2$ for $d \in D$ and comparing it with 2, we see that

$$1 + \frac{4}{10} \leqslant a < 1 + \frac{5}{10},$$

and then computing $(1 + 4/10 + d/100)^2$, we get the next "digit"

$$1 + \frac{4}{10} + \frac{1}{100} \leqslant a < 1 + \frac{4}{10} + \frac{2}{100}$$

(because the square of 1.41 is $\leqslant 2$, but that of 1.42 is > 2).

This can be continued indefinitely, and with enough patience, one gets the first 28 digits of $\sqrt{2}$:

$$1.4142135623730950488016887242 \leqslant \sqrt{2} < 1.4142135623730950488016887242 + 10^{-29}.$$

This method of approximating $\sqrt{2}$ is relatively quick: as an iterative algorithm, each step adds one digit of precision, which means that the error is divided by 10 approximately. We will later see that there are even quicker algorithms (a form of Newton's method, although it was known much earlier than the general version), where (roughly speaking) each step multiplies the number of correct digits by 2; see Proposition 2.8.5 and Example 5.4.11.

In terms of the decimal expansion of real numbers, convergence of a real sequence (a_n) to a can be thought of as follows: for any given number $P \in \mathbf{N}$ of "digits of precision", all the numbers a_n have *the same* integral part $\lfloor a_n \rfloor$ and *the same* first P digits when n

is larger than some $N \in \mathbf{N}$. This is because the integral part and the first P digits of a_n "almost always" coincide with those of a if

$$|a_n - a| < 10^{-P-1},$$

which occurs for all n large enough. (This does not work all the time, because of examples like 0.999 and 1.001, which are very close, but have no common digit, but it is nevertheless a good intuition.)

2.8. Proving convergence without knowing the limit

If we want to use convergent sequences to construct interesting real numbers, a difficulty is that the definition of convergence requires to know what is the limit in advance.

There are two very important convergence conditions that do not require to know the limit. The first is very simple, but does not apply to all convergent sequences. The second, called the Cauchy Criterion, is a general condition that is equivalent to the convergence of a sequence of complex numbers.

The first one concerns *monotone* sequences.

DEFINITION 2.8.1. A sequence (a_n) of real numbers is non-decreasing if $a_n \leqslant a_{n+1}$ for all n. It is non-increasing if $a_{n+1} \leqslant a_n$ for all n. If it is either non-decreasing or non-increasing, we say that it is *monotone*.

REMARK 2.8.2. (1) Many sequences are not monotone – a simple example is $a_n = (-1)^n$.

(2) If (a_n) is non-decreasing, then we get by induction

$$a_n \leqslant a_m$$

for all $m \geqslant n$. If (a_n) is non-increasing, then similarly, we get

$$a_n \geqslant a_m$$

for $m \geqslant n$.

THEOREM 2.8.3. *A monotone sequence* $s = (a_n)$ *of real numbers converges if and only if it is bounded. In fact, if s is non-decreasing, then*

$$\lim_{n \to +\infty} a_n = \sup\{a_n \mid n \in \mathbf{N}\},$$

and if s is non-increasing, then

$$\lim_{n \to +\infty} a_n = \inf\{a_n \mid n \in \mathbf{N}\}.$$

PROOF. We prove the result for a non-decreasing sequence. We know that a convergent sequence is bounded. Conversely, assume that (a_n) is non-decreasing and bounded. The set of values $A = \{a_n\}$ is then bounded and non-empty, hence its supremum $a = \sup(A)$ exists by Theorem 2.3.1. We now prove that (a_n) converges to a. Let $\varepsilon > 0$ be given. Then $a - \varepsilon < a$, so that $a - \varepsilon$ is not an upper-bound of the set A; this means that there is some element of A, say $a_N \in A$, such that $a - \varepsilon < a_N \leqslant a$. Since (a_n) is non-decreasing, for any $n \geqslant N$, we get

$$a - \varepsilon < a_N \leqslant a_n \leqslant a$$

(the second inequality is because a is an upper-bound of A), so $|a - a_N| < \varepsilon$. $\qquad\square$

EXAMPLE 2.8.4. (1) Let $k \in \mathbf{R}$ and $b \in \mathbf{C}$ with $|b| > 1$. Let

$$a_n = \frac{n^k}{b^n}$$

for $n \in \mathbf{N}$. We want to prove that (a_n) converges to 0 (which establishes (2.7)). It is enough to prove that the sequence

$$|a_n| = \frac{n^k}{|b|^n}$$

converges to 0, which means that (by replacing b with $|b|$) we may assume that b is a real number > 1. We can also replace k by a larger number (which only makes $|a_n|$ larger) and therefore assume that $k \in \mathbf{N}$.

For $n \in \mathbf{N}$, we get

$$\frac{a_{n+1}}{a_n} = \frac{1}{b}\left(\frac{n+1}{n}\right)^k.$$

According to (2.9), we have then

$$\frac{a_{n+1}}{a_n} \to \frac{1}{b} < 1$$

as $n \to +\infty$. In particular, there exists $M \in \mathbf{N}$ and $c < 1$ such that $a_{n+1}/a_n \leqslant c$ for an $\geqslant M$, or equivalently $0 \leqslant a_{n+1} \leqslant c a_n \leqslant a_n$ for $n \geqslant M$. Consequently, the sequence (a_n) is non-increasing for $n \geqslant M$, and (because convergence is an asymptotic property) it converges to the infimum a of the values of a_n for $n \geqslant M$, by Theorem 2.8.3. We have $a \geqslant 0$ since the terms of the sequence are non-negative. The inequality $a_{n+1} \leqslant c a_n$ for $n \geqslant M$ implies then

$$a \leqslant c a_n$$

for all $n \geqslant M$. Since $c > 0$, this means that $c^{-1} a \leqslant a_n$ for all $n \geqslant M$, hence $c^{-1} a \leqslant a$, which is absurd unless $a = 0$ (because $c^{-1} > 1$).

(2) Let $A \subset \mathbf{R}$ be a non-empty set of real numbers which has an upper-bound and let $a = \sup(A)$. We can construct as follows a sequence (a_n), with values in A, converging to a.

First if $a \in A$ (so A has a maximum), then we can put $a_n = a$ for all n.

So assume that $a \notin A$. For $n \in \mathbf{N}$, the number $a - 1/n$ cannot be an upper-bound of A, so there exists $a_n \in A$ such that

$$a - \frac{1}{n} < a_n < a.$$

This implies that $|a_n - a| < 1/n$, so that (a_n) converges to a by Lemma 2.5.8 and (2.5).

(3) We show how to use Theorem 2.8.3 to construct the square root of any non-negative real number c. It suffices to do this when $c \geqslant 1$. Indeed, if we can do this, then for $c < 1$, we find a such that $a^2 = 1/c$, and then $(1/a)^2 = c$.

PROPOSITION 2.8.5. *Let $c \geqslant 1$ be a real number. Define a sequence (a_n) by $a_1 = c$ and*

$$a_{n+1} = \frac{1}{2}\left(a_n + \frac{c}{a_n}\right).$$

Then (a_n) is non-increasing and its limit a satisfies $a^2 = c$.

PROOF. We first prove that

$$1 \leqslant a_n \leqslant c$$

for all $n \in \mathbf{N}$ (in particular, this means that $a_n \neq 0$, so a_{n+1} can always be defined).

We proceed naturally by induction. We have $1 \leqslant a_1 \leqslant c$. Suppose that $1 \leqslant a_n \leqslant c$. Then

$$a_{n+1} = \frac{1}{2}\left(a_n + \frac{c}{a_n}\right) \leqslant \frac{1}{2}(c+c) = c$$

$$a_{n+1} = \frac{1}{2}\left(a_n + \frac{c}{a_n}\right) \geqslant \frac{1}{2}(1+1) = 1,$$

which concludes the proof by induction.

Next, we claim that $a_n^2 \geqslant c$ for all $n \in \mathbf{N}$. This is true for $n = 1$ since $a_1^2 = c^2 \geqslant c$ (because $c \geqslant 1$).

Now we can check that (a_n) is non-increasing. We have

$$a_{n+1} - a_n = \frac{1}{2}\left(\frac{c}{a_n} - a_n\right) = \frac{c - a_n^2}{2a_n} \leqslant 0$$

by what we just proved.

By Theorem 2.8.3, the sequence (a_n) converges. Let a be its limit. The sequence (a_{n+1}) also converges to a; but

$$a_{n+1} = \frac{1}{2}\left(a_n + \frac{c}{a_n}\right),$$

and the right-hand side, according to Proposition 2.5.9, (3) and (1), converges to $\frac{1}{2}(a + c/a)$. Since the limit of a convergent sequence is unique, this means that

$$a = \frac{1}{2}\left(a + \frac{c}{a}\right),$$

which translates to $a^2 = c$. $\qquad\qquad\square$

For instance, let $c = 2$. Then the first few steps of the sequence are $a_1 = 2$ and

$$a_2 = \frac{3}{2}$$
$$a_3 = \frac{17}{12}$$
$$a_4 = \frac{577}{408} = 1.41421568627450980392156862745098039 22\ldots$$
$$a_5 = \frac{665857}{470832} = 1.41421356237468991062629557889013491 01\ldots,$$

whereas

$$\sqrt{2} = 1.41421356237309504880168872420969809 786\ldots$$

Note that $|a_4 - \sqrt{2}| < 3 \cdot 10^{-6}$ and $|a_5 - \sqrt{2}| < 2 \cdot 10^{-12}$, which displays the remarkable efficiency of this algorithm: the number of correct digits is roughly multiplied by 2 when going from a_n to a_{n+1}.

The second criterion for convergence "without knowing the limit" is due to Cauchy. It can be motivated by the observation that if a sequence (a_n) of real numbers converges to a, then for n and m both large, the numbers a_n and a_m are very close: if $n \geqslant N$ and $m \geqslant N$, where N is such that $|a_n - a| < \varepsilon$ for $n \geqslant N$, then

$$|a_n - a_m| \leqslant |a_n - a| + |a - a_m| < 2\varepsilon.$$

The inequality $|a_n - a_m| < 2\varepsilon$ does not refer to a, which makes the following definition possible:

DEFINITION 2.8.6. A *Cauchy sequence* (a_n) is a sequence of complex numbers which has the above property: for any $\varepsilon > 0$, there exists $N \in \mathbf{N}$ such that if n and m are integers both larger or equal to N, we have

$$|a_n - a_m| < \varepsilon.$$

REMARK 2.8.7. In terms of logical formulas, this means that $(a_n)_{n \in \mathbf{N}}$ is a Cauchy sequence if and only if

$$\forall\, \varepsilon > 0,\ \exists\, N \in \mathbf{N},\ \forall\, n \geqslant N,\ \forall\, m \geqslant N,\ |a_n - a_m| < \varepsilon.$$

Up to replacing ε by $\varepsilon/2$ in the argument preceeding the definition, we see that it means that any convergent sequence is a Cauchy sequence. The remarkable fact (which is in fact also equivalent to completeness) is that the converse is true:

THEOREM 2.8.8 (Cauchy). *A sequence (a_n) of complex numbers is convergent if and only if it is a Cauchy sequence.*

We only give here an intuitive explanation of this important fact, in order to show that it is not so mysterious; a precise proof will be given in the next section.

We assume that all a_n are non-negative real numbers, and that they form a Cauchy sequence. We then look at the decimal expansions of the numbers a_n, and we observe that for an integer $P \in \mathbf{N}$, the condition

$$|a_n - a_m| < 10^{-P-1}$$

"almost implies" that the P first digits (and integral parts) of a_n and a_m are the same. The Cauchy condition implies that, whatever the choice of P, this will be the case whenever n and m are sufficiently large. But this means that we can "read" this common integral part, and these common digits, by looking at larger and larger P. Then we can construct the real number a with this integral part and these digits (this will be made precise later on); it shouldn't be surprising then that (a_n) converges to a.

Proving that a sequence (a_n) is a Cauchy sequence can seem difficult because there are so many parameters (ε, N, n and m) in the definition. In practice, the simplest approach is often to start with arbitrary integers m and n, and attempt to find a good upper-bound for $|a_m - a_n|$. For symmetry reasons, it is possible to assume that $m \geqslant n$, which can be helpful; one then attempts to find a bound of the form

$$|a_m - a_n| \leqslant b_n$$

where b_n is independent of m, and we know that $b_n \to 0$. Indeed, if this is the case, then for any $\varepsilon > 0$, we find $N \in \mathbf{N}$ such that $|b_n| = b_n < \varepsilon$ for $n \geqslant N$. Then if both m and n are integers $\geqslant N$, then either $m \geqslant n$, in which case we get $|a_m a_n| \leqslant b_n < \varepsilon$ by the above assumption, or $n \geqslant n$, in which case we exchange the role of m and n and get $|a_n - a_m| \leqslant b_m < \varepsilon$.

EXAMPLE 2.8.9. (1) The following is an important example of application of the Cauchy criterion:

PROPOSITION 2.8.10. *Let c be a real number with $0 \leqslant c < 1$. Suppose that (a_n) is a sequence of complex numbers such that*

$$|a_{n+2} - a_{n+1}| \leqslant c|a_{n+1} - a_n|$$

for all $n \in \mathbf{N}$. Then the sequence (a_n) is convergent.

PROOF. Let $n \in \mathbf{N}$ and $m \geqslant n$ be given. Since the information we have concerns the distance between consecutive terms of the sequence, we naturally split the difference

$a_m - a_n$ in intermediate steps $a_{k+1} - a_k$ with $n \leqslant k \leqslant m - 1$; then we can reduce these to $a_{n+1} - a_n$ by induction:

$$|a_{k+1} - a_k| \leqslant c|a_k - a_{k-1}| \leqslant \cdots \leqslant c^{k-n}|a_{n+1} - a_n|.$$

This leads to

$$a_m - a_n = (a_m - a_{m-1}) + \cdots + (a_{n+1} - a_n) = \sum_{k=n}^{m-1}(a_{k+1} - a_k),$$

and by the triangle inequality to

$$|a_m - a_n| \leqslant \sum_{k=n}^{m-1} c^{k-n}|a_{n+1} - a_n|$$

$$= |a_{n+1} - a_n|\left(1 + c + \cdots + c^{m-n-1}\right) = \frac{1 - c^{m-n}}{1 - c}|a_{n+1} - a_n|$$

by (1.1). Since $0 \leqslant c < 1$, we have $1 - c^{m-n} \leqslant 1$, so this means that

$$|a_m - a_n| \leqslant \frac{1}{1 - c}|a_{n+1} - a_n|$$

whenever $n \leqslant m$. Note that m does not appear anymore on the right-hand side. This means that we can obtain the Cauchy condition as soon as we prove that $a_{n+1} - a_n$ tends to 0 as $n \to +\infty$. We do this using induction again:

$$|a_{n+1} - a_n| \leqslant c|a_n - a_{n-1}| \leqslant c^{n-1}|a_2 - a_1|.$$

Since the sequence (c^n) tends to 0 by (2.7), we are done. $\qquad\square$

(2) The Cauchy criterion can be a useful way to prove that a sequence does not converge. One of the most important examples is the following. We define a_n for $n \in \mathbf{N}$ by

$$a_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n} = \sum_{k=1}^{n} \frac{1}{k}.$$

We claim that (a_n) diverges. To see this, we compute $a_{2n} - a_n$, and show that it is not very small, however large is. Precisely, we get

$$a_{2n} - a_n = \frac{1}{2n} + \cdots + \frac{1}{n+1} \geqslant n \cdot \frac{1}{2n} = \frac{1}{2}.$$

This is not compatible with the Cauchy criterion: indeed, with $\varepsilon = \frac{1}{2}$, this would imply that there exists $N \in \mathbf{N}$ such that, for all $n \geqslant N$ and $m \geqslant N$, and in particular for $n \geqslant N$ and $m = 2n \geqslant N$, we have

$$|a_m - a_n| = a_{2n} - a_n < \frac{1}{2},$$

which we just saw is not the case.

Since, we have $a_{n+1} \geqslant a_n$, this means by Theorem 2.8.3 that the sequence (a_n) is not bounded, so that for any $R > 1$, there exists $n \in \mathbf{N}$ such that $a_n > R$. We will later be able to approximate the smallest value of n for which this is true, as a function of R.

2.9. Subsequences

The sequence $(1, -1, 1, -1, \ldots)$ is an example of a bounded sequence of real numbers that does not converge. However, if we only look at the odd-numbered terms $(1, 1, 1, \ldots)$, we have a constant sequence, which is convergent. Remarkably, a version of this fact holds for *any* bounded sequence of complex numbers.

First we explain what is the most general form of "taking the odd-numbered terms".

DEFINITION 2.9.1 (Subsequence). Let $s = (a_n)_{n \in \mathbf{N}}$ be a sequence of complex numbers. A sequence $(b_k)_{k \in \mathbf{N}}$ is called a *subsequence*[34] of s if there exist integers

$$(2.12) \qquad\qquad n_1 < n_2 < \cdots < n_k < \cdots$$

such that

$$b_k = a_{n_k}$$

for all $k \in \mathbf{N}$.

If the sequence (b_k) converges, then its limit is called an *accumulation point*[35] or *limit point* of s.

We note (since it is sometimes useful) that the condition (2.12) implies (by induction on k) that $n_k \geqslant k$ for all $k \in \mathbf{N}$.

EXAMPLE 2.9.2. (1) Let $a_n = 2^n$. Among the subsequences of (a_n), we have

$$(2^{2k})_{k \in \mathbf{N}}, \qquad (2^{2k+1})_{k \in \mathbf{N}}, \qquad (2^{2^k})_{k \in \mathbf{N}}$$

with, respectively, $b_k = a_{2k}$, $b_k = a_{2k+1}$ and $b_k = a_{2^k}$.

On the other hand, the sequence

$$(b_k) = (2, 8, 4, 32, 16, \ldots)$$

is *not* a subsequence of (a_n): although each term b_k is a term of the original sequence, $b_k = a_{n_k}$, the integers n_k do not satisfy the requirement (2.12); for instance, $b_2 = a_3$ and $b_3 = a_2$, so $n_2 = 3$ and $n_3 = 2$.

(2) The sequence $(a_n)_{n \in \mathbf{N}} = ((-1)^n)_{n \in \mathbf{N}}$ has the constant sequences 1 and -1 as subsequences, in fact in many different ways: for instance $1 = a_{2k+1} = a_{4k+1}$, and $-1 = a_{2k} = a_{2^k}$ for all $k \in \mathbf{N}$. In particular, 1 and -1 are both accumulation points of (a_n) (one can check that they are the only ones).

(3) If a sequence $s = (a_n)$ converges to $a \in \mathbf{C}$, then a is the unique accumulation point of s. This should be intuitively clear: the terms of the sequence approach a, and cannot be made to approach a different limit even by taking only some of them. Rigorously, let (b_k) be a convergent subsequence of s, with $b_k = a_{n_k}$. Since $n_k \geqslant k$, we see that for any given $\varepsilon > 0$, if $N \in \mathbf{N}$ is such that $|a_n - a| < \varepsilon$ for all $n \geqslant N$, then we also get $|b_k - a| = |a_{n_k} - a| < \varepsilon$ for $k \geqslant N$, since $n_k \geqslant k \geqslant N$.

(Remarkably, the converse is true for *bounded* sequences: a bounded sequence which has only one accumulation point a converges to a; see Proposition 2.9.5.)

(4) A sequence (a_n) may well have infinitely many accumulation points. For instance, one can show that the set of accumulation points of the sequence $(\cos(n))_{n \in \mathbf{N}}$, which is bounded with values in the closed interval $I = [-1, 1]$, is equal to the whole interval I.

Now we state the important theorem concerning existence of accumulation points.

THEOREM 2.9.3 (Bolzano–Weierstrass). *Let $(a_n)_{n \in \mathbf{N}}$ be a bounded sequence of complex numbers. Then (a_n) has at least one accumulation point.*

PROOF. We begin with the case of a real sequence, which is more intuitive. Let $R \geqslant 0$ be such that $|a_n| \leqslant R$ for all $n \in \mathbf{N}$. The idea is to apply *successive dissections*: we first split the interval $[-R, R]$ in two equal subintervals, and note that *at least one* of the two subintervals must contain infinitely many terms of the sequence; then we split this subintervals into two even smaller ones, and iterate, finding smaller and smaller intervals that contain infinitely many terms of the sequence – this ends up with an accumulation point.

To be precise, let $I_1 = [\alpha_1, \beta_1]$ with $\alpha_1 = -R$ and $\beta_1 = R$, so a_n in in I_1 for all $n \in \mathbf{N}$. We construct by induction a sequence of intervals

$$I_k = [\alpha_k, \beta_k],$$

and a sequence of integers n_k, with

$$(2.13) \quad \alpha_{k-1} \leqslant \alpha_k \leqslant \beta_k \leqslant \beta_{k-1}, \quad \beta_k - \alpha_k = \frac{1}{2}(\beta_{k-1} - \alpha_{k-1}), \quad n_k > n_{k-1}, \quad a_{n_k} \in I_k$$

if $k \geqslant 2$, and with the property that there are infinitely many integers n with $a_n \in I_k$.

Taking $n_1 = 1$, we have already done this for $k = 1$; now assume that I_k has been constructed with the property (2.13). We have

$$I_k = \left[\alpha_k, \frac{\beta_k + \alpha_k}{2}\right] \cup \left[\frac{\beta_k + \alpha_k}{2}, \beta_k\right],$$

and among the integers n with $a_n \in I_k$, there must be infinitely many with

$$a_n \in \left[\alpha_k, \frac{\beta_k + \alpha_k}{2}\right],$$

or infinitely many with

$$a_n \in \left[\frac{\beta_k + \alpha_k}{2}, \beta_k\right]$$

(otherwise, there would only be finitely many n with $a_n \in I_k$). Let I_{k+1} be one of these two intervals, chosen so that it has this property; its endpoints α_{k+1} and β_{k+1} satisfy

$$\alpha_k \leqslant \alpha_{k+1} \leqslant \beta_{k+1} \leqslant \beta_k,$$

and the length $\beta_{k+1} - \alpha_{k+1}$ of I_{k+1} is half of the length of I_k. Finally, since infinitely many n exist with $a_n \in I_{k+1}$, we can choose n_{k+1} be be any integer larger than n_k with $a_{n_{k+1}} \in I_{k+1}$.

This concludes the inductive definition. Now let $b_k = a_{n_k}$ for $k \in \mathbf{N}$; the sequence (b_k) is a subsequence of (a_n) by construction, with $b_k \in I_k$ for all $k \in \mathbf{N}$.

By (2.13), the sequence (α_k) is non-decreasing and bounded by R:

$$-R = \alpha_1 \leqslant \alpha_2 \leqslant \cdots \leqslant \alpha_k \leqslant \beta_k \leqslant \beta_1 = R.$$

Hence the sequence (α_k) has a limit α by Theorem 2.8.3. Similarly the sequence (β_k) is non-increasing and bounded, so it has a limit β.

Since $\beta_1 - \alpha_1 = 2R$, we have $\beta_k - \alpha_k = 2^{2-k}R$ for $k \in \mathbf{N}$ by (2.12) and induction. So by Proposition 2.5.9 and (2.6), we get $\beta - \alpha = 0$. Finally, since $b_k \in I_k$, we have

$$|b_k - \alpha| \leqslant \beta_k - \alpha_k = 2^{2-k}R,$$

which implies that (b_k) also converges to α.

This concludes the proof of the theorem for real sequences. If (a_n) is a complex sequence, then we proceed as follows: since the sequence $(\mathrm{Re}(a_n))_n$ is a bounded sequence of real numbers, there is, by what we just saw, a convergent subsequence (b_k) with

$b_k = \operatorname{Re}(a_{n_k})$; denote its limit by b. Now the sequence $(\operatorname{Im}(a_{n_k}))_k$ is also a bounded sequence of real numbers; we find therefore a convergent subsequence $(c_l)_{l \in \mathbf{N}}$, with

$$c_l = b_{k_l}$$

for some integers k_l; let $c \in \mathbf{R}$ be its limit.

Now the sequence $(d_l) = (a_{n_{k_l}})_{l \in \mathbf{N}}$ is still a subsequence of (a_n); its real part is a subsequence of (b_k), hence also converges to b, while its imaginary part was chosen to converge to c. By Proposition 2.5.11, it follows that d_l converges to $a + ib$. This is therefore an accumulation point of the sequence (a_n). $\qquad\square$

REMARK 2.9.4. An unbounded sequence can have accumulation points: for instance, let

$$a_n = (1 + (-1)^n)2^n$$

for $n \in \mathbf{N}$, which defines sequence

$$(0, 8, 0, 32, 0, 128, 0, \ldots).$$

This sequence is not bounded (because it contains the terms 2^{2n+1}, which are not bounded), but clearly admits 0 as accumulation point.

As an application of the Bolzano–Weierstrass Theorem, we can now prove Theorem 2.8.8.

PROOF OF THEOREM 2.8.8. We need to prove that all Cauchy sequences converge. We will do this by combining two steps:

Step 1. Any Cauchy sequence (a_n) is bounded.

Step 2. If $a \in \mathbf{C}$ is an accumulation point of a Cauchy sequence (a_n), then (a_n) converges to a.

By Step 1, Theorem 2.9.3 implies that a Cauchy sequence (a_n) has an accumulation point a; then by Step 2, the sequence converges to a.

We now complete Step 1. Let (a_n) be a Cauchy sequence. Let $\varepsilon = 1$ and let $N \in \mathbf{N}$ be such that $|a_n - a_m| < 1$ for all $n \geqslant N$. Then for $n \geqslant N$, we get

$$|a_n| \leqslant |a_n - a_N| + |a_N| \leqslant |a_N| + 1,$$

by the triangle inequality; therefore, for all $n \in \mathbf{N}$, we have

$$|a_n| \leqslant \max(|a_1|, \ldots, |a_N|, |a_N| + 1).$$

Finally, we complete Step 2. Let (a_n) be a Cauchy sequence, and $a \in \mathbf{C}$ an accumulation point of (a_n). Let (b_k) with $b_k = a_{n_k}$ be a subsequence converging to a.

We now need to prove that (a_n) converges to a. Fix $\varepsilon > 0$. For any integer $n \in \mathbf{N}$, we use another parameter $k \in \mathbf{N}$ to write

$$|a_n - a| \leqslant |a_n - b_k| + |b_k - a|.$$

We need to select k carefully. First, since the sequence (a_n) is a Cauchy sequence, there exists $N \in \mathbf{N}$ such that $|a_n - a_m| < \varepsilon/2$ if n and m are $\geqslant N$. In particular $|a_n - b_k| < \varepsilon/2$ if $n \geqslant N$ and if $k \geqslant N$ (since then $n_k \geqslant N$).

Secondly, since (b_k) converges to a, there exists $K \in \mathbf{N}$ such that $|b_k - a| < \varepsilon/2$ if $k \geqslant K$. If we take for k the *fixed* value $k = \max(K, N)$, then we see that

$$|a_n - a| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for any $n \geqslant N$. This concludes the proof of convergence. $\qquad\square$

We finish this section by proving a fact we mentioned earlier in a remark.

> **PROPOSITION 2.9.5.** *Let $(a_n)_{n \in \mathbf{N}}$ be a bounded sequence of complex numbers. If the sequence has a unique limit point a, then*
> $$\lim_{n \to +\infty} a_n = a.$$

PROOF. We show that the negation of the statement leads to a contradiction. So we need to write the negation of the formula (2.3); using the rules in Section 1.1.3, we obtain

$$\exists\, \varepsilon > 0,\ \forall\, N \in \mathbf{N},\ \exists\, n \geqslant N,\ |a_n - a| \geqslant \varepsilon.$$

In words: for some fixed $\varepsilon > 0$, we can find for any N an integer larger than N such that a_n is at distance at least ε from a.

We deduce from this that there is a subsequence (b_k) of the sequence (a_n) which satisfies

$$|b_k - a| \geqslant \varepsilon$$

for all $k \in \mathbf{N}$ (by induction on k: applying the negation property for $N = 1$ first, then $N = n_1 + 1$, etc, one constructs the sequence n_k so that $b_k = a_{n_k}$ satisfies this property).

Since (a_n) is bounded, so is the subsequence (b_k); according to Theorem 2.9.3, the sequence (b_k) has a convergent subsequence; let b be its limit; it is also a limit point of the original sequence (a_n). But the inequality $|b_k - a| \geqslant \varepsilon$ implies (by replacing b_k by the terms of the convergent subsequence and using Proposition 2.5.11 and Proposition 2.5.9) that $|b - a| \geqslant \varepsilon > 0$. So b is an accumulation point of the original sequence which is not equal to a, which contradicts the assumption. $\qquad \square$

2.10. Series

Many sequences $(s_n)_{n \in \mathbf{N}}$ are defined in such a way that we have $s_1 = a_1$ and

$$s_{n+1} = s_n + a_{n+1} \qquad \text{for } n \geqslant 1,$$

where $(a_n)_{n \in \mathbf{N}}$ is another sequence of complex numbers. In such a case, we say that the sequence (s_n) is the *series*[36] with terms (or summands) (a_n), and we denote the sequence (s_n) by the expression

$$\sum_{n=1}^{+\infty} a_n.$$

By induction, we see that the n-th term of the sequence can be expressed in the form

$$s_n = a_1 + a_2 + \cdots + a_n = \sum_{k=1}^{n} a_k,$$

and the s_n are called the *partial sums* of the series.

If the sequence (s_n) converges, then its limit $s \in \mathbf{C}$ is called the *sum* of the series with terms (a_n), and we write

$$\sum_{n=1}^{+\infty} a_n = s.$$

As before, when we write such a formula, with a complex number in the right-hand side, we first mean that the series converges, and then that its sum is the number s. We may also have series with terms starting with a_0, which we write

$$\sum_{n=0}^{+\infty} a_n.$$

EXAMPLE 2.10.1. (1) The decimal expansion of a real number a is of this form, according to (2.11); Theorem 2.7.1 can be restated, for $a \geqslant 0$, by saying that

$$a = \lfloor a \rfloor + \sum_{i=1}^{+\infty} d_i 10^{-i}$$

where the sequence (d_i) is the sequence of digits defined in Section 2.7.

(2) An important example of convergent series is the *geometric series*

$$\sum_{n=0}^{+\infty} a^n$$

where $|a| < 1$. Indeed, by (1.1), the partial sums are

$$1 + a + \cdots + a^n = \frac{1 - a^{n+1}}{1 - a}$$

for $n \in \mathbf{N}$. Since $a^{n+1} \to 0$ by (2.6), we deduce that the series converges and that

$$\sum_{n=0}^{+\infty} a^n = \frac{1}{1 - a}.$$

REMARK 2.10.2. Conversely, we can interpret any sequence (a_n) also as a series, by writing

$$a_n = (a_n - a_{n-1}) + \cdots + (a_3 - a_2) + (a_2 - a_1) + a_1 = a_1 + \sum_{k=1}^{n} (a_k - a_{k-1}).$$

Translating directly the properties of convergence of sequences, we obtain immediately the following results:

PROPOSITION 2.10.3. *Let* $\sum a_n$ *and* $\sum b_n$ *be convergent series, with sums a and b, respectively. Then*

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

$$\sum_{n=1}^{+\infty} (a_n + b_n) = a + b$$

$$\sum_{n=1}^{+\infty} (z a_n) = za \ \textit{for } z \in \mathbf{C}$$

$$\textit{if } a_n \leqslant b_n \ \textit{for all } n \in \mathbf{N} \ \textit{then } a \leqslant b$$

$$\sum_{n=1}^{+\infty} \bar{a}_n = \bar{a}.$$

Moreover, if we write $x_n = \mathrm{Re}(a_n)$ *and* $y_n = \mathrm{Im}(a_n)$, *then*

$$\sum_{n=1}^{+\infty} x_n = \mathrm{Re}(a), \qquad \sum_{n=1}^{+\infty} y_n = \mathrm{Im}(a).$$

PROOF. For the first point, we note that in terms of the partial sums (s_n), we have

$$a_n = s_n - s_{n-1}.$$

Both sequences (s_n) and (s_{n-1}) converge to a, so by Proposition 2.5.9, (1), the sequence (a_n) converges to $a - a = 0$.

The second statement is also a consequence of the same part of Proposition 2.5.9, and the third corresponds to multiplying a convergent sequence with a constant. For the fourth, note that the assumption implies that

$$\sum_{k=1}^{n} a_k \leqslant \sum_{k=1}^{n} b_k$$

for all $n \in \mathbf{N}$, hence $a \leqslant b$ by Proposition 2.5.9, (4). And the last parts come from Proposition 2.5.11. $\square$

REMARK 2.10.4. (1) Note that there is no obvious representation of the product

$$ab = \Big(\sum_{n=1}^{+\infty} a_n\Big)\Big(\sum_{n=1}^{+\infty} b_n\Big)$$

as a series, so the second formula above is restricted to multiplying by a fixed number z. We will see later some cases where the product of two series has a convenient expression as a single series.

(2) It is important to remember that *the converse of the first property does not hold*: there are series $\sum a_n$, with $a_n \to 0$, which are not convergent. We already saw this in Example 2.8.9, (2), where $a_n = 1/n$: the series

$$\sum_{n=1}^{\infty} \frac{1}{n}$$

does not converge.

The analogue of Theorem 2.8.3 is very simple for series: since the difference of consecutive partial sums $s_{n+1} - s_n$ is just a_{n+1}, this is the sign of the terms that matters.

THEOREM 2.10.5. *Let (a_n) be a real sequence with $a_n \geqslant 0$ for all $n \geqslant 1$. Then the series $\sum a_n$ converges if and only if its partial sums are bounded, i.e, if there exists $R \geqslant 0$ such that*

$$a_1 + \cdots + a_n \leqslant R$$

for all $n \in \mathbf{N}$.

When $a_n \geqslant 0$ for all n, we will often write

$$\sum_{n=1}^{+\infty} a_n < +\infty, \qquad \sum_{n=1}^{+\infty} a_n = +\infty$$

to indicate that the series is convergent or not (in the second case, we say that it is *divergent*). So for instance, we write

$$\sum_{n=1}^{+\infty} \frac{1}{n} = +\infty.$$

In the application of Cauchy's Criterion to series, we need to consider differences $|s_n - s_m|$. It is here usually most convenient to assume that $n \geqslant m$; then

$$s_n - s_m = a_n + \cdots + a_{m+1} = \sum_{k=m+1}^{n} a_k.$$

Checking Cauchy's Criterion requires us to estimate the modulus of these numbers from above. And here there is a difference between series and general sequences, because there is a natural first upper-bound, based on applying directly the triangle inequality: we have

$$|s_n - s_m| \leqslant |a_n| + \cdots + |a_{m+1}| = \sum_{k=m+1}^{n} |a_k|.$$

The right-hand side is also what is needed to apply the Cauchy Criterion, but to the sequence of non-negative real numbers $(|a_n|)_n$. This means that we obtain from Theorem 2.8.8 and Theorem 2.10.5 the following very useful fact: if the terms (a_n) of a series are such that

$$\sum_{n=1}^{+\infty} |a_n| < +\infty,$$

then the series $\sum a_n$ converges. This is important enough that this condition has a name:

DEFINITION 2.10.6. A series $\sum a_n$ is *absolutely convergent*[37] if the series

$$\sum_{n=1}^{+\infty} |a_n|$$

converges.

So we have seen the first part of the following result:

THEOREM 2.10.7. *If the series $\sum a_n$ converges absolutely, then it converges, and moreover we then have*

$$\left| \sum_{n=1}^{+\infty} a_n \right| \leqslant \sum_{n=1}^{+\infty} |a_n|.$$

PROOF. We need only check the last inequality (which is fairly natural, as an extension of the triangle inequality). For any $n \geqslant 1$, the usual triangle inequality gives

$$\left| \sum_{k=1}^{n} a_k \right| \leqslant \sum_{k=1}^{n} |a_k|.$$

Since the series $\sum a_n$ converges, the left-hand side converges to

$$\left| \sum_{n=1}^{+\infty} a_n \right|$$

by Proposition 2.5.11; the second converges to

$$\sum_{n=1}^{+\infty} |a_n|,$$

so that we conclude by applying Proposition 2.5.9, (4). $\qquad \square$

REMARK 2.10.8. In practice, we often prove absolute convergence by proving an inequality

$$(2.14) \qquad\qquad\qquad |a_n| \leqslant b_n$$

where the series $\sum b_n$ is already known to be convergent. Indeed, this implies that

$$\sum_{k=1}^{n} |a_k| \leqslant \sum_{k=1}^{n} b_k \leqslant \sum_{k=1}^{+\infty} b_k$$

for all $n \in \mathbf{N}$, and we can apply Theorem 2.10.5. (We see here how the two methods of proving convergence without knowing the limit complement each other.)

Conversely, note that if we know that

(2.15)
$$|a_n| \geqslant b_n$$

and that $\sum b_n$ diverges, then the series $\sum a_n$ does not converge absolutely (but it may still converge, as we will see in an example below).

In both cases, as usual, it is enough if there exists $M \in \mathbf{N}$ such that the comparison bound (2.14) or (2.15) is true for all $n \geqslant M$.

EXAMPLE 2.10.9. (1) As an example, consider an arbitrary sequence $(d_n)_{n \in \mathbf{N}}$ of decimal digits, so $d_n \in \{0, \ldots, 9\}$. The series

$$\sum_{n=1}^{+\infty} d_n 10^{-n}$$

is then convergent since $|d_n 10^{-n}| \leqslant 9 \cdot 10^{-n}$, and we know that the series $\sum 10^{-n}$ is convergent (Example 2.10.1, (2)). This means that any infinite sequence of decimal digits defines a number (in Section 2.7, we showed that any real number has a decimal expansion, but it could have been the case that some sequences of digits are not possible).

There is however a small issue to remember when dealing with decimal expansions: if we start with an arbitrary sequence (d_n) of digits and define

$$a = \sum_{n=1}^{+\infty} \frac{d_n}{10^n},$$

then the decimal digits of a, as defined in Section 2.7, do not always coincide with the sequence (d_n). For instance, take $d_1 = 0$ and $d_n = 9$ for all $n \geqslant 2$, so that

$$a = 0.09999999\ldots$$

We have then

$$a = 9 \sum_{n=2}^{+\infty} \frac{1}{10^n} = \frac{9}{100} \sum_{n=0}^{+\infty} 10^{-n} = \frac{9}{100} \frac{1}{1 - 1/10} = \frac{9}{90} = \frac{1}{10},$$

and the sequence of digits for $1/10$, according to Section 2.7, is

$$0, \ 1, \ 0, \ 0, \ \ldots$$

(2) The series

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

is an example of a series which is convergent, but not absolutely convergent. To see this, note first that $|(-1)^n/n| = 1/n$, so the series does not converge absolutely. To prove the convergence, we consider the odd and even partial sums:

$$u_n = s_{2n} = 1 - \frac{1}{2} + \cdots + \frac{1}{2n}, \qquad v_n = s_{2n+1} = 1 - \frac{1}{2} + \cdots - \frac{1}{2n+1}.$$

We then observe that

$$v_n \leqslant v_{n+1} \leqslant u_{n+1} \leqslant u_n$$

holds for all $n \in \mathbf{N}$: indeed

$$v_{n+1} - v_n = \frac{1}{2n+2} - \frac{1}{2n+3} \geqslant 0, \qquad u_{n+1} - u_n = \frac{1}{2n+2} - \frac{1}{2n+1} \leqslant 0,$$

$$u_n - v_n = \frac{1}{2n+1} \geqslant 0.$$

The sequences (u_n) and (v_n) are therefore monotone; moreover

$$v_1 \leqslant v_n \leqslant u_n \leqslant u_1,$$

for all $n \in \mathbf{N}$, so these sequences are bounded, and converge to some limits u and v. Since $u_n - v_n \to 0$, it follows that $v = u$. Now, each partial sum of the series is either s_{2n} or s_{2n+1} for some n, and then it follows that the whole series converges to $u = v$. (One can show in this case that the sum of the series is $\log(2)$).

It is easy to see that the argument above is more general, and leads to the following useful result:

PROPOSITION 2.10.10. *Let (a_n) be a non-increasing sequence of positive real numbers which converges to 0. The series*

$$\sum_{n=1}^{+\infty} (-1)^{n-1} a_n$$

converges.

This applies for instance to

$$\sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{\sqrt{n}}, \qquad \sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{\log(2n)},$$

and many other series; such series are called *alternating series*.

REMARK 2.10.11. It is important to know that absolutely convergent series are "much better behaved" than series that are convergent, but not absolutely; for this reason, it is best to try to prove absolute convergence of a series, even if it might be easier to just prove convergence.

The following theorem gives a striking illustration of the difference of behavior of the two kinds of series, and indicates that manipulating infinite sums without being careful may be risky.

DEFINITION 2.10.12. Let (a_n) be a sequence of complex numbers. A *rearrangement* of the series $\sum a_n$ is a series $\sum b_k$ for which there exists a bijective map $f \colon \mathbf{N} \to \mathbf{N}$ such that

$$b_k = a_{f(k)}$$

for all $k \in \mathbf{N}$.

For instance, the series

$$a_2 + a_1 + a_4 + a_3 + a_6 + a_5 + \cdots$$

where we exchange successive pairs of terms $a_{2n+1} + a_{2n}$, or the series

$$a_4 + a_2 + a_3 + a_1 + a_5 + a_6 + a_7 + \cdots,$$

where we only permute the first four terms, are rearrangements of $\sum a_n$. On the other hand, the series

$$a_2 + a_2 + a_1 + a_3 + a_4 + \cdots,$$

is not a rearrangement, because we are repeating the term a_2 twice. Neither is

$$a_2 + a_3 + a_4 + a_5 + \cdots,$$

because we have omitted the term a_1.

If we have a sum with finitely many terms, then a rearrangement doesn't change its value (since $a + b = b + a$ for any complex numbers a and b), for instance

$$a_1 + a_2 + a_3 + a_4 = a_4 + a_2 + a_1 + a_3.$$

It is then maybe natural to expect that the same is true for "infinite sums", which are series.

However, this is only the case for absolutely convergent series. More precisely, one can prove the following:

THEOREM 2.10.13. (1) *If* $\sum a_n$ *is absolutely convergent, then are rearrangement* $\sum b_k$ *is also absolutely convergent, and*

$$\sum_{n=1}^{+\infty} a_n = \sum_{k=1}^{+\infty} b_k.$$

(2) *If* $a_n \in \mathbf{R}$ *for all* n *and the series* $\sum a_n$ *is convergent, but not absolutely convergent, then for all* $c \in \mathbf{R}$, *there exists a rearrangement* $\sum b_k$ *such that*

$$\sum_{k=1}^{+\infty} b_k = c.$$

There exist also rearrangements that are not convergent.

See [2, §6.3] for a discussion and a concrete example of a rearrangement of the series $\sum (-1)^{n-1}/n$ which has a different values as the series itself.

For every sequence (a_n) that is known to converge to 0, it is natural to ask whether the series $\sum a_n$ is in fact convergent. We now do this with the examples of Proposition 2.6.1.

PROPOSITION 2.10.14. (1) *The series*

$$\sum_{n=1}^{+\infty} \frac{1}{n^k}$$

converges for $k > 1$ *and diverges for* $k \leqslant 1$.
(2) *The series*

$$\sum_{n=0}^{+\infty} \frac{n^k}{b^n}$$

converges absolutely if $|b| > 1$ *and* $k \in \mathbf{R}$.
(3) *The series*

$$\sum_{n=0}^{+\infty} \frac{a^n}{n!}$$

converges absolutely for all $a \in \mathbf{C}$.

PROOF. (1) The terms $1/n^k$ are of course positive. Since we already know that $\sum 1/n$ diverges, and since $1/n^k \geqslant 1/n$ if $k \leqslant 1$, the comparison principle as in Remark 2.10.8 implies that $\sum 1/n^k$ diverges for $k \leqslant 1$.

Now we show how to prove the convergence for $k > 1$ by bounding the partial sums and using the convergence of the geometric series. Namely, for $m \geqslant 1$ such that $n \leqslant 2^m - 1$,

we write

$$1 + \frac{1}{2^k} + \cdots + \frac{1}{n^k} \leqslant \sum_{i=1}^{2^m-1} \frac{1}{i^k}$$

and we split the right-hand side into the sums

$$\sum_{i=2^j}^{2^{j+1}-1} \frac{1}{i^k}$$

for $0 \leqslant j \leqslant m - 1$; for instance, for $j = 2$, this is

$$\frac{1}{4^k} + \cdots + \frac{1}{7^k}.$$

In each sum the largest term is the first one, and there are 2^j terms, so that

$$\sum_{i=2^j}^{2^{j+1}-1} \frac{1}{i^k} \leqslant \frac{2^j}{2^{jk}} = b^j$$

with $b = 2^{1-k}$. Since $k > 1$ by assumption, we have $0 < b < 1$, and hence by (1.1), we get

$$\sum_{i=1}^{n} \frac{1}{i^k} \leqslant \sum_{j=0}^{m} b^j \leqslant \frac{1}{1-b}$$

for all n, which proves the convergence in that case.

(2) This is a generalization of the geometric series. Here we may observe that if $c > 1$ is such that $1 < c < |b|$ (for instance $c = (1 + |b|)/2$), then we have

$$\lim_{n \to +\infty} \frac{n^k}{(|b|/c)^n} = 0$$

by (2.7). So there exists some integer $N \in \mathbf{N}$ such that

$$\frac{|n|^k}{b^n} \leqslant \frac{1}{2} \frac{1}{c^n}$$

for all $n \geqslant N$. Since the geometric series $\sum (1/c)^n$ converges, we deduce by comparison that

$$\sum_{n=1}^{+\infty} \frac{|n|^k}{b^n} < +\infty,$$

which proves the absolute convergence of $\sum n^k/b^n$.

(3) We proceed in a similar way. By (2.8), we have

$$\lim_{n \to +\infty} \frac{(2|a|)^n}{n!} = 0,$$

so that there exists $N \in \mathbf{N}$ such that

$$\frac{|a|^n}{n!} \leqslant \frac{1}{2^n}$$

for all $n \geqslant N$. The series $\sum 2^{-n}$ converges, hence by comparison we deduce that the series $\sum a^n/n!$ converges absolutely. $\qquad \square$

One can of course ask if the sums of the series above are interesting or simple numbers. They are indeed all interesting; the function

$$\zeta(k) = \sum_{n=1}^{+\infty} \frac{1}{n^k}$$

for $k > 1$ is known as the *Riemann zeta function*, and remains very mysterious; one knows that

(2.16)
$$\zeta(2) = \sum_{n=1}^{+\infty} \frac{1}{n^2} = \frac{\pi^2}{6},$$

for example, and similar formulas for $\zeta(2k)$, when $k \geqslant 1$ is an integer, but it is not known if $\zeta(5)$ is a rational number or not.

Since it is a geometric series, we have

$$\sum_{n=0}^{+\infty} b^{-n} = \frac{1}{1 - 1/b},$$

for $|b| > 1$; the series $\sum n^k b^{-n}$ also have "nice" expressions for all $k \in \mathbf{N}_0$.

Finally, the series

$$\sum_{n=0}^{+\infty} \frac{a^n}{n!}$$

has the value $\exp(a) = e^a$, as we will see later (in fact, this will be the rigorous definition of the exponential function).

EXAMPLE 2.10.15. To see how useful these basic examples are, together with the comparison method, consider the series

$$\sum_{n=0}^{+\infty} \frac{\cos(3\exp(14\sqrt{\pi}n!)) + \sin(3\cos(n^3)))}{n!}.$$

Without knowing anything about the numerator, except that $|\cos(x)| \leqslant 1$ for all $x \in \mathbf{R}$, we can conclude that

$$\left| \frac{\cos(3\exp(14\sqrt{\pi}n!)) + \sin(3\cos(n^3)))}{n!} \right| \leqslant \frac{1}{n!}$$

hence the series converges absolutely, and its sum has absolute value $\leqslant \sum 1/n!$.

2.11. Convergence to infinity

Certain sequences (a_n) are not convergent but have a different type of very "regular" behavior: the terms a_n become larger and larger as n increases. This is the subject of the next definition:

DEFINITION 2.11.1. Let (a_n) be a sequence of real numbers.
(1) We say that the sequence a_n converges to $+\infty$, denoted

$$\lim_{n \to +\infty} a_n = +\infty,$$

if for all real numbers $T \in \mathbf{R}$, there exists $N \in \mathbf{N}$ such that $a_n > T$ for all $n \geqslant N$.
(2) We say that the sequence a_n converges to $-\infty$, denoted

$$\lim_{n \to +\infty} a_n = -\infty,$$

if for all real numbers $T \in \mathbf{R}$, there exists $N \in \mathbf{N}$ such that $a_n < T$ for all $n \geqslant N$.

EXAMPLE 2.11.2. We have
$$\lim_{n\to+\infty} n = +\infty, \qquad \lim_{n\to+\infty} -n^2 = -\infty.$$

Convergence to infinity can in fact be related to convergence to a real number.

PROPOSITION 2.11.3. *Let (a_n) and (b_n) be sequences of real numbers.*
(1) We have
$$\lim_{n\to+\infty} a_n = +\infty$$
if and only if $a_n > 0$ for all n large enough and
$$\lim_{n\to+\infty} \frac{1}{a_n} = 0.$$
(2) We have
$$\lim_{n\to+\infty} a_n = -\infty$$
if and only if $a_n < 0$ for all n large enough and
$$\lim_{n\to+\infty} \frac{1}{a_n} = 0.$$

PROOF. We prove (1), the second part being similar. If $a_n \to +\infty$, then the definition with $T = 0$ implies that $a_n > 0$ for n large enough; we then deduce that $1/a_n \to 0$ since, when $a_n > 0$, the condition $a_n > T$ is equivalent to $0 < |1/a_n| = 1/a_n < T^{-1}$. For any $\varepsilon > 0$, we take $T = \varepsilon^{-1}$, and obtain the definition of convergence to 0.

Conversely, if there exists $M \in \mathbf{N}$ such that $a_n > 0$ for $n \geqslant M$ and if $1/a_n \to 0$, then similarly, for a given $T > 0$, we take $\varepsilon = T^{-1}$ in the definition of convergence to 0, and obtain an $N \in \mathbf{N}$ such that
$$0 < a_n = |a_n| < T^{-1}$$
for $n \geqslant N$, so that $a_n > T$ for $n \geqslant N$. $\qquad\square$

REMARK 2.11.4. (1) It is possible that $1/a_n \to 0$ but that the sequence does not converge to either $+\infty$ or $-\infty$. For instance, for $a_n = (-1)^n/n$, we have $1/a_n = (-1)^n n$, which does not have a constant sign for n large.

(2) It is not true in general that a sequence that is not convergent must converge to either $+\infty$ or $-\infty$, as the example of $a_n = (-1)^n$ shows.

EXAMPLE 2.11.5. (1) According to Proposition 2.5.9, we have for instance
$$\lim_{n\to+\infty} \frac{a^n}{n^k} = +\infty, \text{ for } a > 1 \text{ and } k \in \mathbf{R},$$
$$\lim_{n\to+\infty} \frac{n!}{a^n} = +\infty, \text{ for } a > 0.$$

(2) We can now complete Example 2.6.2 in the missing case. Precisely, let $k \leqslant l$ be positive integers, let $(a_0, \dots, a_k)$ be complex numbers with $a_k \neq 0$, and let $(b_0, \dots, b_l)$ be complex numbers with $b_l \neq 0$. Assume that $k > l$. Then
$$\lim_{n\to+\infty} \frac{a_k n^k + \cdots + a_1 n + a_0}{b_l n^l + \cdots + b_1 n + b_0} = \begin{cases} +\infty & \text{if } a_k/b_l > 0 \\ -\infty & \text{if } a_k/b_l < 0. \end{cases}$$

Indeed, we have first
$$\lim \frac{b_l n^l + \cdots + b_1 n + b_0}{a_k n^k + \cdots + a_1 n + a_0} = 0$$

by Example 2.6.2. We then just need to check that, for n large enough, the sign of

$$a_k n^k + \cdots + a_1 n + a_0$$

is the same as the sign of a_k, and similarly for the other term and the sign of b_l. This is true because the other terms are smaller for n very large. Precisely, assume first that $a_k > 0$, the other case being similar. Since

$$\lim_{n \to +\infty} \frac{a_{k-1} n^{k-1} + \cdots + a_1 n + a_0}{a_k n^k} = 0,$$

there exists $N \in \mathbf{N}$ such that

$$|a_{k-1} n^{k-1} + \cdots + a_1 n + a_0| \leqslant \frac{a_k n^k}{2},$$

for all $n \geqslant N$, from which the triangle inequality gives

$$a_k n^k + \cdots + a_1 n + a_0 \geqslant a_k n^k - \frac{1}{2} a_k n^k > 0$$

for $n \geqslant N$.

(3) If (a_n) is a non-decreasing sequence, then we know by Theorem 2.8.3 that it converges if and only if it is bounded from above. If that is not the case, then (a_n) is unbounded, and we then have

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

Indeed, for any $T \in \mathbf{R}$, there exists some N such that $a_N > T$, since the sequence is unbounded from above; since it is non-decreasing, we obtain

$$a_n \geqslant a_N > T$$

for all $n \geqslant N$, which proves that $a_n \to +\infty$.

From Proposition 2.11.3, or directly, one can easily deduce the following result (where we consider mostly sequences converging to $+\infty$, leaving the analogues of convergence to $-\infty$ to the reader).

PROPOSITION 2.11.6. *Let (a_n) and (b_n) be sequences of real numbers.*
(1) If $a_n \to +\infty$ and (b_n) is bounded from below, then

$$\lim_{n \to +\infty} (a_n + b_n) = +\infty.$$

(2) If $a_n \to +\infty$ and there exists $\delta > 0$ such that $b_n \geqslant \delta$ for all n, then

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

PROOF. We establish (1) and leave (2) as an exercise (taking the inverse is there very useful).

If (b_n) is bounded from below, this means that there exists a real number R such that $b_n \geqslant R$. Then for any $T \in \mathbf{R}$, we find $N \in \mathbf{N}$ such that $a_n > T - R$ for all $n \geqslant N$, and then we get $a_n + b_n > T$ for all $n \geqslant N$. $\square$

REMARK 2.11.7. Case (1) applies, for instance, if the sequence (b_n) is convergent, or if $b_n \geqslant 0$ for all n. For instance, it follows that

$$\lim_{n \to +\infty} \left(\frac{n^3 + 2}{n^2 + 1} + 10000 \cdot (-1)^n \right) = +\infty.$$

Of course, there are variants of these facts for sequences converging to $-\infty$, which we leave to the reader to state if needed (they can be reduced to the previous case by replacing the terms of a sequence with their opposites).

CHAPTER 3

Continuous functions

In this chapter, we begin the study of *functions* defined on $\mathbf{R}$ or on a subset of $\mathbf{R}$, and with real values. These are fundamental in analysis, and the most common functions (such as square root, logarithm, exponential, trigonometric functions, etc) are of this type.

3.1. Functions and graphs

Let $I \subset \mathbf{R}$ be a set of real numbers. We will call *function* on I a map $f \colon I \to \mathbf{R}$ or $f \colon I \to \mathbf{C}$ (in the second case, we will usually speak of *complex-valued functions*). We will assume most of the time that I is an interval, but this is not always necessary.

EXAMPLE 3.1.1. (1) A *polynomial function*[38] (or sometimes simply *polynomial*) is a function $f \colon \mathbf{R} \to \mathbf{C}$ such that
$$f(x) = a_k x^k + \cdots + a_1 x + a_0$$
for some complex numbers $(a_0, \ldots, a_k)$, called the *coefficients* of the polynomial. If $a_k \neq 0$, then it is a *polynomial function* of degree k. If all coefficients are zero, then the function is always zero, and doesn't have a well-defined degree. If all coefficients are real numbers, then f defines a function $f \colon \mathbf{R} \to \mathbf{R}$.

It is important to note that the degree and the coefficients are determined uniquely by the function: it is not possible to have an equality
$$a_k x^k + \cdots + a_1 x + a_0 = b_l x^l + \cdots + b_1 x + b_0$$
for all x unless $k = l$ and $a_i = b_i$ for all i.

(2) Let f_2 be a non-zero polynomial and let I be the set of real numbers where $f_2(x) \neq 0$ (this only excludes finitely many values of x). If f_1 is a polynomial, then we can define a *rational function* $f \colon I \to \mathbf{R}$ by $f(x) = f_1(x)/f_2(x)$.

(3) Let I be any subset of $\mathbf{R}$. The function defined by
$$f(x) = \begin{cases} 1 & \text{if } x \in I \\ 0 & \text{if } x \notin I \end{cases}$$
is called the *characteristic function* of the set I.

As in Definition 2.4.4, we can add and multiply functions defined on the same set I, and we can divide a function f by a function g such that $g(x) \neq 0$ for all $x \in I$. For instance, polynomials are obtained by sums and products from the constant functions and the identity function $f(x) = x$.

Real-valued functions defined on subsets of $\mathbf{R}$ can be graphically represented in a way that makes it often possible to understand intuitively their properties.

DEFINITION 3.1.2. Let $I \subset \mathbf{R}$ be a set of real numbers. Let $f \colon I \to \mathbf{R}$ be a function defined on I. The *graph* of f is the subset G_f of the plane $\mathbf{R} \times \mathbf{R}$ defined by
$$G_f = \{(x, y) \in I \times \mathbf{R} \mid y = f(x)\}.$$
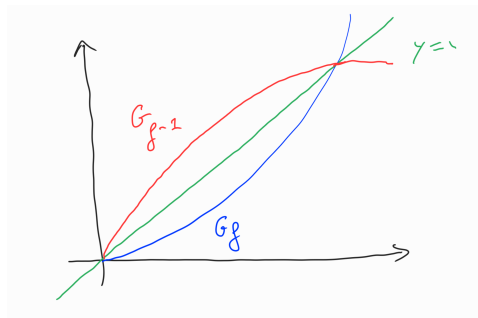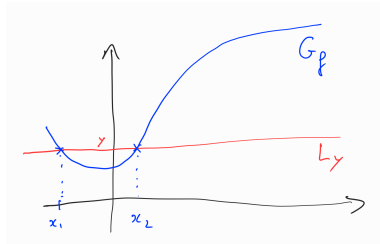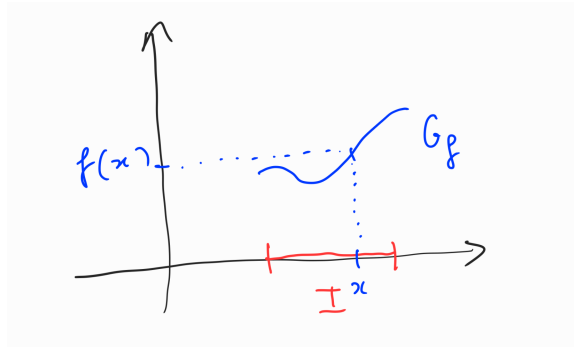
REMARK 3.1.3. Not all subsets G of the plane can be the graph of a function. To see whether this is the case, one must check that for every $x \in \mathbf{R}$, there is at most one value of y for which $(x, y) \in G$; the set of x where there is exactly one value of y is then the definition set I, and the function is defined by mapping x to this unique value of y with $(x, y) \in G$.

EXAMPLE 3.1.4. (1) Let $f \colon I \to \mathbf{R}$ be a function. For any $y_0 \in \mathbf{R}$, the intersection of the graph of f with the horizontal line L_{y_0} with equation $y = y_0$ is the set of points (x, y) such that, on the one hand, we have $y = f(x)$ (because $(x, y) \in G_f$) and on the other hand $y = y_0$ (because $(x, y) \in L_{y_0}$). This means that these are the points (x, y_0) where x is a solution of the equation $f(x) = y_0$.

From this, we see that

- To say that f is injective means that whatever the value of $y_0 \in \mathbf{R}$, the horizontal line L_{y_0} with equation $y = y_0$ intersects the graph G_f in *at most* one point;
- To say that f is *surjective* is to say that whatever the value of $y_0 \in \mathbf{R}$, the horizontal line L_{y_0} with equation $y = y_0$ intersects the graph G_f in at least one point.

(2) Let I and J be subsets of $\mathbf{R}$ and $f \colon I \to J$ a *bijective function*. The graph of the inverse bijection f^{-1} (see Definition 1.4.7) is obtained by taking the symmetric of the graph G_f with respect to the diagonal (which means that (x, y) is in G_f if and only if $(y, x) \in G_{f^{-1}}$).

DEFINITION 3.1.5. Let $I \subset \mathbf{R}$ be an interval (Section 2.1). A function $f \colon I \to \mathbf{R}$ is said to be *non-decreasing* (resp. *strictly increasing*) if we have $f(x) \leqslant f(y)$ whenever $x \leqslant y$ (resp. if $f(x) < f(y)$ if $x < y$).

A function $f \colon I \to \mathbf{R}$ is said to be *non-increasing* (resp. *strictly decreasing*) if we have $f(x) \geqslant f(y)$ whenever $x \leqslant y$ (resp. if $f(x) > f(y)$ if $x < y$).

A function that is either non-decreasing or non-increasing is called *monotone*; if it is either strictly increasing or strictly decreasing, it is called *strictly monotone*.

EXAMPLE 3.1.6. (1) The function defined by $f(x) = x^2$ on $\mathbf{R}_+$ is strictly increasing. The function defined by $g(x) = x^2$ on $\mathbf{R}_- =]-\infty, 0]$ is strictly decreasing.

(2) The function defined by $f(x) = x^3$ on $\mathbf{R}$ is strictly increasing.

(3) A constant function $f(x) = a$ for $x \in I$ is non-decreasing, but not strictly increasing (unless I is reduced to a single point).

Note that any strictly monotone function $f \colon I \to \mathbf{R}$ is *injective*: indeed, if $x \neq y$, then either $x < y$, in which case $f(x) < f(y)$, or $y < x$, in which case $f(y) < f(x)$; in both cases, we get $f(x) \neq f(y)$.

3.2. Continuous functions

Reference: [2, 7.1, 7.2, 7.4, 7.5].

Suppose that I is a set of real numbers and $f \colon I \to \mathbf{R}$ a function. How can we hope to compute its value for a real number $x_0 \in I$ which is only determined or represented by an approximation x_1? Even if we can compute $f(x_1)$, this can only be useful if f has the property that *the values of f are close when evaluated at nearby values of the variable x, like x_0 and x_1.*

This property of a function is called *continuity*.[39] It is defined precisely as follows:

DEFINITION 3.2.1. Let I be a set of real numbers and $f \colon I \to \mathbf{C}$ a function.

Let $x_0 \in I$. We say that f is continuous[40] at x_0 if for every $\varepsilon > 0$, there exists $\delta > 0$ such that, whenever $x \in I$ satisfies $|x - x_0| < \delta$, we have $|f(x) - f(x_0)| < \varepsilon$.

If f is continuous at all $x \in I$, then we say simply that f is continuous on I.

REMARK 3.2.2. (1) The logical formula for continuity of f at x_0 is

$$\forall \varepsilon > 0, \; \exists \delta > 0, \; \forall x \in I, \; (|x - x_0| < \delta \longrightarrow |f(x) - f(x_0)| < \varepsilon),$$

and for continuity on I, it becomes

$$\forall x_0 \in I, \; \forall \varepsilon > 0, \; \exists \delta > 0, \; \forall x \in I, \; (|x - x_0| < \delta \longrightarrow |f(x) - f(x_0)| < \varepsilon).$$

Intuitively, this means that if we want to know what is $f(x_0)$ with a certain precision (determined by ε), it suffices to know x_0 with the precision δ.

Geometrically, one can say (at least if I is an interval) that f is continuous if "one can draw the graph of f in the plane without lifting the pen". However, a continuous function can be quite complicated. Below is a fairly typical example.

(2) There are many analogies between properties of convergence of sequences and continuity. For instance, the analogue of the fact that convergence is an asymptotic property of a sequence (Example 2.5.6, (2)) is the fact that continuity at x_0 is a *local property* of the function f for values of x close to x_0. Precisely, suppose that g is a function on I and that there exists $\alpha > 0$ such that

$$f(x) = g(x)$$
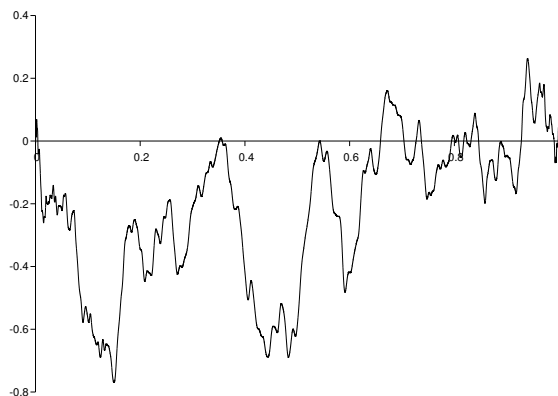
whenever $x \in I$ satisfies $|x - x_0| < \alpha$ (so the graphs of f and g coincide at least over the interval defined by $|x - x_0| < \alpha$). Then f is continuous at x_0 if and only if g is. (The reason is that for a given $\varepsilon > 0$, we can replace the value of δ in the logical formula for the continuity of f by $\min(\alpha, \delta)$.)

(3) If $f \colon I \to \mathbf{C}$ is continuous on I, then it is also continuous on any subset $J \subset I$.

EXAMPLE 3.2.3. (1) Constant functions are continuous, and so is the identity function $I \to \mathbf{R}$ (which maps x to x); more generally, for any complex number c, the function defined by $f(x) = cx$ on I is continuous on I (exercise).

(2) Let $I = [a, b]$ for some real numbers $a < b$. The characteristic function f of I, viewed as a function from $\mathbf{R}$ to $\mathbf{R}$, is continuous at all $x_0 \in \mathbf{R}$ except for $x_0 = a$ and $x_0 = b$.

Indeed, if $x_0 \notin \{a, b\}$, then there exists $\alpha > 0$ such that f is constant when $|x - x_0| < \alpha$, and one can use the fact that continuity is a local property (for instance, if $a < x_0 < b$, one can take $\alpha = \min(\frac{1}{2}(x_0 - a), \frac{1}{2}(b - x_0))$; and then $f(x) = 1$ for $|x - x_0| < \alpha$).

If $x_0 = a$, on the other hand, for any $\delta > 0$, the interval $]a - \delta, a + \delta[$ contains real numbers x that are not in I, and then $f(x) = 0$, so $|f(x) - f(x_0)| = 1$, which means that the formula defining continuity is not satisfied when $\varepsilon < 1$.

(3) Although most of the functions that we encounter in applications are continuous, or at worse continuous outside of a few points, it is important to know that there are functions that are not continuous at *any* point. For instance, let $I = [0, 1]$ and define

$$f(x) = \begin{cases} 0 & \text{if } x \notin \mathbf{Q} \\ b & \text{if } x = a/b \text{ with } a \in \mathbf{N}_0,\ b \in \mathbf{N}, \text{ without common factor.} \end{cases}$$

So for instance, $f(0) = 1$, $f(1/2) = 2$, $f(3/5) = 5$, but $f(1/\sqrt{2}) = 0$ because $1/\sqrt{2} \notin \mathbf{Q}$.

We claim that, for any $x_0 \in [0, 1]$, the function f is not continuous at x_0. To see this, consider first $x_0 \notin \mathbf{Q}$ and take $\varepsilon = 1/2$ in the definition of continuity. If f were continuous at x_0, there would exist $\delta > 0$ such that $|f(x)| = f(x) < 1/2$ when $|x - x_0| < \delta$. But in the interval $[0, 1] \cap]x_0 - \delta, x_0 + \delta[$, there always exists a rational number x, for which we have $f(x) \geqslant 1$ (Theorem 1.5.3).

Now, on the other hand, if $x \in \mathbf{Q}$, and $x = a/b$ without common factor, so that $f(x) = b \geqslant 1$, then let again $\varepsilon = 1/2$. If f were continuous at x_0, there would exist $\delta > 0$ such that $|f(x) - b| = f(x) < 1/2$ when $|x - x_0| < \delta$, which means that $f(x) = b$ since $f(x)$ is a non-negative integer. But in the interval $[0, 1] \cap]x_0 - \delta, x_0 + \delta[$, there always exists an irrational number x for which $f(x) = 0$ (for instance $x = x_0 + \frac{1}{n\sqrt{2}}$ or $x = x_0 - \frac{1}{n\sqrt{2}}$, for $n > \delta^{-1}$).

In order to prove continuity, we can also use a comparison approach that is easier to handle than the definition.

> LEMMA 3.2.4. *Let I be a set of real numbers and let f, g be functions $I \to \mathbf{C}$. Suppose that g is continuous on I, and that we have*
>
> $$|f(x) - f(y)| \leqslant |g(x) - g(y)|$$
>
> *for all x and y in I. Then f is continuous on I.*

PROOF. Indeed, given $\varepsilon > 0$, if we take $\delta > 0$ such that $|g(x) - g(y)| < \varepsilon$ whenever $|x - y| < \varepsilon$, then we also obtain

$$|f(x) - f(y)| \leqslant |g(x) - g(y)| < \varepsilon$$

for $|x - y| < \delta$. □

One example of comparison is particularly helpful and important.

> DEFINITION 3.2.5. *Let I be a set of real numbers and $f\colon I \to \mathbf{C}$ a function. Let $c \in \mathbf{R}_+$. One says that f is* Lipschitz-continuous *if*
>
> $$|f(x) - f(y)| \leqslant c|x - y|$$
>
> *for all x and y in I.*
> *We also say that c is a Lipschitz constant for f (it is not unique).*

Any Lipschitz-continuous function is continuous, by applying the lemma to the function $g(x) = cx$.

REMARK 3.2.6. Intuitively, to say that f is Lipschitz is to say that, in order to compute $f(x_0)$ with N digits of precision, we need to know x_0 with $N+n$ digits approximately, where n is the number of digits of c.

As in the case of convergent sequences, we can operate with continuous functions, and continuity is preserved. In addition, we can also use composition, which provides a powerful tool to prove the continuity of almost all functions defined "using elementary functions".

We state the results for continuity over the whole interval, but they are also true for continuity at a single point x_0.

We begin with a useful lemma. A continuous function is not necessarily bounded (meaning that there might not exist $R \in \mathbf{R}_+$ such that $|f(x)| \leqslant R$ for all $x \in I$; an example is the identity function $f(x) = x$ on $I = \mathbf{R}$). However, we have the following:

> LEMMA 3.2.7. *Let I be a set of real numbers and $f\colon I \to \mathbf{C}$ a function. Let $x_0 \in I$ such that f is continuous at x_0.*
>
> *(1) If f is continuous at x_0, then it is* locally bounded *around x_0, which means that there exists $\alpha > 0$ such that f is bounded for $x \in I$ such that $|x - x_0| < \alpha$.*
>
> *(2) If f is real-valued and $f(x_0) > 0$ then there exists $\alpha > 0$ such that for all $x \in I$ such that $|x - x_0| < \alpha$, we have*
>
> $$f(x) \geqslant \tfrac{1}{2} f(x_0),$$
>
> *and in particular $f(x) > 0$.*
> *Similarly if $f(x_0) < 0$, then there exists $\alpha > 0$ such that $f(x) \leqslant \tfrac{1}{2} f(x_0) < 0$ for all $x \in I$ such that $|x - x_0| < \alpha$.*

PROOF. (1) Applying continuity with $\varepsilon = 1$, and then the triangle inequality, we find $\alpha > 0$ such that

$$|f(x)| \leqslant |f(x_0)| + |f(x) - f(x_0)| \leqslant |f(x_0)| + 1$$

for $x \in I$ with $|x - x_0| < \alpha$.

(2) Applying continuity with $\varepsilon = \frac{1}{2} f(x_0) > 0$, and then the triangle inequality, we find $\alpha > 0$ such that

$$f(x) \geqslant f(x_0) - |f(x) - f(x_0)| \geqslant f(x_0) - \tfrac{1}{2} f(x_0) \geqslant \tfrac{1}{2} f(x_0) > 0$$

when $x \in I$ is such that $|x - x_0| < \alpha$. $\qquad\square$

> PROPOSITION 3.2.8. *Let I be a set of real numbers and let f, g be continuous functions $I \to \mathbf{C}$.*
>
> (1) *The functions $f + g$ and fg are continuous on I, and if $g(x) \neq 0$ for all $x \in I$, then f/g is continuous.*
>
> (2) *Let $J \subset \mathbf{R}$ be such that $f(x) \in J$ for all $x \in I$. Let $h \colon J \to \mathbf{C}$ be a function on J. The composition*
> $$h \circ f \colon I \to \mathbf{C},$$
> *such that $(h \circ f)(x) = h(f(x))$ for all $x \in I$, is continuous.*

PROOF. We leave (1) as an exercise (but we will also show how to deduce this later from another continuity condition).

(2) Let $x_0 \in I$ be given. Let $\varepsilon > 0$. Since h is continuous at $f(x_0)$, there exists $\delta_1 > 0$ such that $|h(y) - h(f(x_0))| < \varepsilon$ if $|y - f(x_0)| < \delta_1$. And since f is continuous at x_0, there exists $\delta > 0$ such that $|f(x) - f(x_0)| < \delta_1$ if $|x - x_0| < \delta$. Hence, whenever $|x - x_0| < \delta$, we get $|h(f(x)) - h(f(x_0))| < \varepsilon$ by applying the first inequality to $y = f(x)$. $\qquad\square$

With this proposition, it follows that essentially any function constructed using finitely many "elementary" operations or functions is continuous where it is defined.

EXAMPLE 3.2.9. (1) Any polynomial function is continuous on $\mathbf{R}$ (or on any subset of $\mathbf{R}$). Similarly, if f is a polynomial function and I does not contain any zero of f, then for any polynomial g, the rational function defined by $h(x) = f(x)/g(x)$ for $x \in I$ is continuous on I.

(2) Since the squaring function is continuous, for any continuous function $f \colon I \to \mathbf{R}$, the function f^2 is also continuous.

(3) The function on $\mathbf{R}$ defined by $v(x) = |x|$ is continuous. This follows for instance from the formula

$$\big| |x| - |y| \big| \leqslant |x - y|$$

which proves that it is even Lipschitz-continuous. Using composition, it follows that for any continuous function f, the function $|f|$ is also continuous.

There is another approach to proving continuity, which uses sequences, and which is also a powerful way to prove that sequences converge.

> PROPOSITION 3.2.10. *Let I be a set of real numbers. Let $f \colon I \to \mathbf{C}$ be a function. Then f is continuous at a point $x_0 \in I$ if and only if, for any sequence (a_n) with values in I that converges to x_0, we have*
> $$\lim_{n \to +\infty} f(a_n) = f(x_0).$$

PROOF. Assume first that f is continuous at x_0. Let (a_n) be a convergent sequence with limit x_0 and with $a_n \in I$ for all n. Let $\varepsilon > 0$ and let δ be such that $|f(x) - f(x_0)| < \varepsilon$ when $|x - x_0| < \delta$. Since $a_n \to x_0$, there exists $N \in \mathbf{N}$ such that $|a_n - x_0| < \delta$ for $n \geqslant N$; for all such n, we then get $|f(a_n) - f(x_0)| < \varepsilon$, which means that the sequence $(f(a_n))$ converges, with limit $f(x_0)$.

Conversely, assume that the convergence condition is satisfied. To prove that f is continuous at x_0, we assume the opposite, and deduce a contradiction.

The negation of convergence at x_0 is the formula

$$\exists \varepsilon > 0, \ \forall \delta > 0, \ \exists x \in I, \ (|x - x_0| < \delta \wedge |f(x) - f(x_0)| \geqslant \varepsilon)$$

(in other words: for some fixed $\varepsilon > 0$, however small δ is, we can find $x \in I$ at distance less than δ for which $f(x)$ is at least at distance ε from $f(x_0)$).

We fix the ε given by this statement. We then apply it to $\delta = 1/n$, where $n \in \mathbf{N}$ is an arbitrary integer; we denote by a_n a value in I with $|a_n - x_0| < 1/n$ and $|f(a_n) - f(x_0)| > \varepsilon$. This last condition, since it holds for all n, implies that $f(a_n)$ does *not* converge to $f(x_0)$. But on the other hand, we have

$$|a_n - x_0| < \frac{1}{n}$$

by construction, which implies that $a_n \to x_0$ (since $1/n \to 0$). So there is at least one sequence converging to x_0 whose image by f does not converge to $f(x_0)$. $\qquad\square$

EXAMPLE 3.2.11. (1) Proposition 3.2.10 leads to a proof of Proposition 3.2.8, based on the known statements for convergence of sequences. For instance, for the statements in part (1) of Proposition 3.2.8, note that if f and g are defined at x_0, then for any sequence (a_n) in I that converges to x_0, we have $(f + g)(a_n) \to f(x_0) + g(x_0)$ and $(fg)(a_n) \to f(x_0)g(x_0)$ by Proposition 2.5.9 and Proposition 3.2.10.

(2) Suppose that $f \colon \mathbf{R} \to \mathbf{R}$ is a continuous function. Let $a_1 \in \mathbf{R}$ and defined a sequence (a_n) inductively by

$$a_{n+1} = f(a_n)$$

for $n \in \mathbf{N}$. Suppose that we know that (a_n) is convergent (that might not always be the case!). Then its limit a satisfies $f(a) = a$.

Indeed, in the equation $a_{n+1} = f(a_n)$, the left-hand side converges to a, while the right-hand side converges to $f(a)$, so these numbers must be equal.

Graphically, this means that we can find all possible limits of such inductive sequences by looking at intersection points of the graph of f with the diagonal

$$\Delta = \{(x, x) \in \mathbf{R} \times \mathbf{R} \mid x \in \mathbf{R}\}$$

(since these intersection points are precisely the points (x, y) with $x = y = f(x)$).

3.3. Global properties of continuous functions

We have discussed now *local* properties of continuity. Even when stated over the whole definition set, the statements that were proved could apply to continuity at a single x_0.

Howver, functions which are continuous *on certain sets of real numbers*, satisfy two extremely important *global* properties, which we now state.

THEOREM 3.3.1 (Intermediate value theorem). *Let I be an interval of real numbers and let $a < b$ be elements of I. If $f \colon I \to \mathbf{R}$ is a continuous function and if*

$$f(a) < f(b) \quad (\text{resp. } f(a) > f(b))$$

then for any $c \in [f(a), f(b)]$ (resp. for any $c \in [f(b), f(a)]$), there exists $x \in [a, b]$ such that $f(x) = c$.

Another way to state this theorem is to say that the image of an interval by a continuous function is an interval.
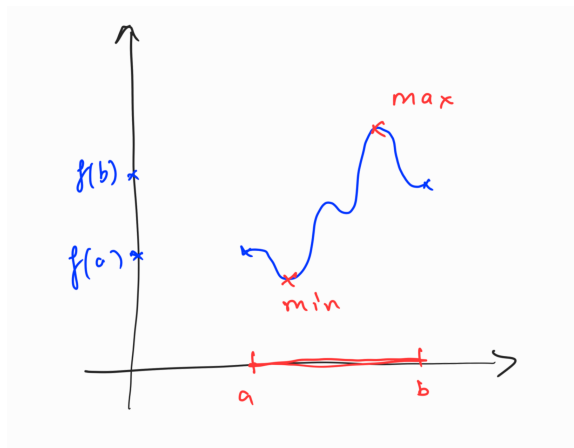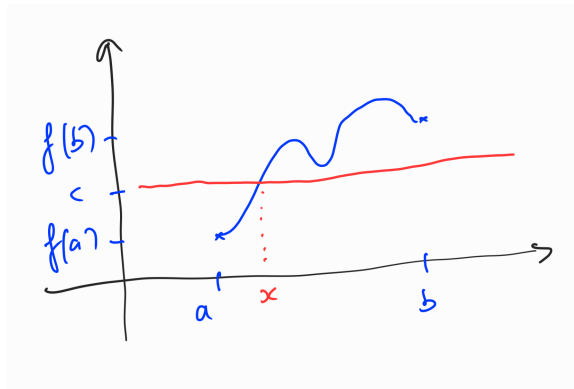
THEOREM 3.3.2 (Extremum Theorem). *Let $a < b$ be real numbers. Let $f\colon [a,b] \to \mathbf{R}$ be a continuous function. Then the set of values*

$$f([a,b]) = \{f(x) \mid x \in [a,b]\} \subset \mathbf{R}.$$

has a maximum and a minimum.

This statement implies in particular that f is bounded on an interval of the form $[a,b]$. If we combine it with the intermediate value theorem, we can see that for any $a \leqslant b$, there exist $c \leqslant d$ in $\mathbf{R}$ such that

$$f([a,b]) = [c,d].$$

Here c is the minimum value of f on $[a,b]$, and d the maximum; to say that they are the minimum and the maximum is to say in particular that there exist x_0 and x_1 in $[a,b]$ such that

$$f(x_0) = c = \min_{x \in [a,b]} f(x), \qquad f(x_1) = d = \max_{x \in [a,b]} f(x).$$

REMARK 3.3.3. (1) The extremum theorem only holds for intervals of the form $[a,b]$ (bounded, and containing both endpoints). This can be seen with the following examples: if $I = \,]a,b]$ or $]a,b[$, take $f(x) = 1/(x-a)$, which is not bounded; if $I = [a,b[$, take $f(x) = 1/(x-b)$; if I is not bounded (either in the positive or negative direction), take $f(x) = x^2$.

For this reason, such intervals have a special name: they are called *compact intervals*.[41]

(2) If f is monotone on $[a,b]$, but not necessarily continuous, then it is also true that f has a maximum and a minimum, which are in fact the end points. For instance, if f is

non-decreasing, then $\max f(x) = f(b)$ and $\min f(x) = f(a)$, simply because we have

$$f(a) \leqslant f(x) \leqslant f(b)$$

for all $x \in I$. On the other hand, the image of f is then not always equal to the whole interval $[f(a), f(b)]$, as shown by the example of the function

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

Before we prove these theorems, let us illustrate one important application.

EXAMPLE 3.3.4. (1) Let $k \in \mathbf{N}$ be an *odd* integer and $a_0, \ldots, a_k$ real numbers, where $a_k \neq 0$. Then we claim that there exists (at least) one real number $x \in \mathbf{R}$ such that

$$a_k x^k + \cdots + a_1 x + a_0 = 0.$$

(So it follows that the function $f \colon \mathbf{R} \to \mathbf{R}$ defined by $f(x) = a_k x^k + \cdots + a_1 x + a_0$ is surjective.)

To see this, note that we may assume that $a_k > 0$. Now define the continuous function $f \colon \mathbf{R} \to \mathbf{R}$ by

$$f(x) = a_k x^k + \cdots + a_1 x + a_0.$$

We claim that there exists $a < 0$ such that $f(a) < 0$ and $b > 0$ such that $f(b) > 0$. Then the intermediate value theorem shows that there exists $x \in [a, b]$ with $f(x) = 0$, which was our goal.

To check the claim, we note that from Example 2.11.5, we have

$$\lim_{n \to +\infty} f(n) = +\infty$$

so that there certainly exists some integer $b \in \mathbf{N}$ with $f(b) > 0$. Moreover, since the degree k of the polynomial is odd, we have

$$f(-n) = -a_k n^k + \cdots - a_1 n + a_0,$$

and hence for the same reason, we have

$$\lim_{n \to -\infty} f(-n) = -\infty,$$

which gives an integer $a \in \mathbf{N}$ such that $f(-a) < 0$.

(2) Let $k \in \mathbf{N}$ be an integer. Define $f(x) = x^k$ for $x \in \mathbf{R}_+$; this is a continuous function from $\mathbf{R}_+$ to $\mathbf{R}_+$. We claim that it is surjective, so that for any $y \geqslant 0$, there exists some real number $x \geqslant 0$ such that $x^k = y$ (which is unique because the function f is strictly monotone; it is denoted $\sqrt[k]{y}$ or $y^{1/k}$).

Indeed, let $y \in \mathbf{R}_+$. Since

$$\lim_{n \to +\infty} n^k = +\infty,$$

there exists $n \in \mathbf{N}$ such that $n^k > y$, and then we have

$$0 = f(0) < y < f(n)$$

and by the Intermediate Value Theorem, there must exist $x \in [0, n]$ such that $x^n = y$.

PROOF OF THEOREM 3.3.1. The idea is to use a subdivision argument, somewhat similar to the construction of a convergent subsequence, although it will be seen to be more effective (in the sense that it can actually be used to approximate the solution of the intermediate value problem).

We consider the case where $f(a) < f(b)$ and $f(a) < c < f(b)$. First, by replacing f with g defined by $g(x) = f(x) - c$, we are in the situation where $g(a) < 0$, $g(b) > 0$, and we want to find x such that $g(x) = 0$.

We construct by induction on $n \in \mathbf{N}_0$ a sequence of intervals

$$I_n = [a_n, b_n]$$

such that

$$(3.1) \qquad\qquad a_n \leqslant a_{n+1}, \qquad b_{n+1} \leqslant b_n,$$

$$(3.2) \qquad\qquad g(a_n) < 0, \qquad g(b_n) \geqslant 0, \qquad b_n - a_n = \frac{1}{2^n}(b - a)$$

for all n. For $n = 0$, we can take $a_0 = a$ and $b_0 = b$. If I_n has been constructed, we consider $c = (a_n + b_n)/2$, and define I_{n+1} depending on the sign of $g(c)$. Precisely, we put

$$I_{n+1} = [a_n, c] = [a_n, (a_n + b_n)/2]$$

if $g(c) \geqslant 0$, and

$$I_{n+1} = [c, b_n] = [(a_n + b_n)/2, b_n]$$

if $g(c) < 0$. Then the conditions (3.1) and (3.2) are satisfied.

Now we conclude in a classical way: from the construction, we have

$$a = a_0 \leqslant \cdots \leqslant a_n \leqslant b_n \leqslant \cdots \leqslant b_0 = b$$

for all n, so the sequence (a_n) is bounded and non-decreasing and converges to some number x_0, while (b_n) is bounded and non-increasing and converges to x_1. Since

$$b_n - a_n = \frac{1}{2^n}(b - a),$$

we have $x_0 = x_1$. Let us denote by x this common value; we claim that $g(x) = 0$, which will conclude the proof.

From $x = \lim a_n$ and the continuity of g, we know that $g(x) = \lim g(a_n)$ by Proposition 3.2.10; since $g(a_n) \leqslant 0$ for all n, it therefore follows that $g(x) \leqslant 0$ (by Proposition 2.5.9, (4)). Similarly, $x = \lim b_n$ implies that $g(x) = \lim g(b_n)$, and from $g(b_n) \geqslant 0$, it follows that $g(x) \geqslant 0$. Both combined give $g(x) = 0$. $\qquad\square$

Proof of Theorem 3.3.2. We prove the existence of the maximum, the case of the minimum being similar. We first assume that f is bounded, and will check that this is indeed the case afterwards. In this is the case, then the set $f([a, b])$ is a non-empty bounded set of real numbers (it contains $f(a)$). Let M be its supremum (which exists by Theorem 2.3.1). We need to prove that M is the value $g(x)$ for some $x \in [a, b]$.

We know at least that for any $n \in \mathbf{N}$, the number $M - 1/n$ is not an upper-bound for $f([a, b])$. This means that there exists $a_n \in [a, b]$ such that

$$M - \frac{1}{n} < f(a_n) \leqslant M.$$

Now we want to use Proposition 3.2.10, but the sequence (a_n) may not converge. However, we have $a \leqslant a_n \leqslant b$ for all n, so the Bolzano–Weierstrass Theorem (Theorem 2.9.3) implies that there exists a convergent subsequence (b_k) with $b_k = a_{n_k}$ for some $n_k \in \mathbf{N}$. Let $x \in [a, b]$ be its limit. We have

$$M - \frac{1}{n_k} < f(b_k) \leqslant M$$

for all $k \in \mathbf{N}$. Since $n_k \geqslant k$, we have $1/n_k \to 0$, and then since $f(b_k)$ converges to $f(x)$ by continuity, we deduce that $f(x) = M$.
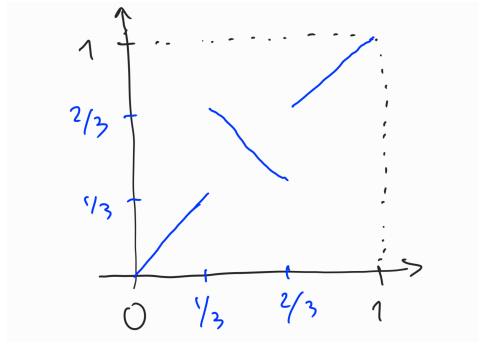
There remains to check that f is indeed bounded. The argument is similar: if f is not bounded, then for any $n \in \mathbf{N}$, we can find $a_n \in [a, b]$ such that $|f(a_n)| \geqslant n$. We find a convergent subsequence (b_k), with limit x, by Theorem 2.9.3; the bounds

$$|f(a_n)| \geqslant n$$

for all $n \in \mathbf{N}$ imply that $|f(a_n)| \to +\infty$, which contradicts Proposition 3.2.10, since $f(a_n)$ should converge to the real number $f(x)$. $\qquad\square$

3.4. Injective continuous functions

We have already noted that a strictly monotone function on an interval I is necessarily injective. In general, the converse is not true: for instance, with $I = [0, 1]$, defining

$$f(x) = \begin{cases} x & \text{if } 0 \leqslant x < 1/3, \\ -x + 1 & \text{if } 1/3 \leqslant x < 2/3 \\ x & \text{if } 2/3 \leqslant x \leqslant 1, \end{cases}$$

we have an injective function that is not monotone (since it is increasing on $[0, 1/3[$ and decreasing on $[1/3, 2/3[$).

However, the converse does hold for continuous functions.

PROPOSITION 3.4.1. *Let I be an interval. A continuous function $f \colon I \to \mathbf{R}$ is injective if and only if it is strictly monotone.*

PROOF. We first suppose that I is of the form $I = [a, b]$ with $a < b$. Suppose that f is injective and continuous on I. Either $f(a) < f(b)$ or $f(a) > f(b)$, since f is injective; we assume that the first inequality holds, and then we prove that f is strictly increasing (the other case is similar).

Let c and d be real numbers such that

$$a \leqslant c < d \leqslant b.$$

We then use the Intermediate Value Theorem as follows:

- We have $f(d) > f(a)$; indeed, otherwise we have

$$f(d) < f(a)$$

  (we have $f(d) \neq f(a)$ since f is injective) and the Intermediate Value Theorem gives some $x \in [d, b]$ such that $f(x) = f(a)$, contradicting the injectivity of f.
- We have $f(c) < f(d)$; indeed, we can assume that $a < c$ (the previous case handles the situation when $a = c$), and then we would get otherwise the inequalities

$$f(a) < f(d) < f(c)$$

71

and the Intermediate Value Theorem would give some x with $x \in [a, c]$ such that $f(x) = f(d)$, which again contradicts the injectivity of f.

The case of a general interval can be reduced to that case by noting that there are always sequences $a_n < b_n$ such that

$$I = \bigcup_{n \in \mathbf{N}} [a_n, b_n]$$

and a_n is decreasing and b_n is increasing (for instance, if I is the interval $I = \,]a, +\infty[$, we can take $a_n = a + 1/n$ and $b_n = a + n$). By the above, f is strictly increasing or strictly decreasing on $[a_1, b_1]$. Suppose that it is strictly increasing; then for any n, the function f is strictly monotone on $[a_n, b_n]$, and it must be strictly increasing on $[a_n, b_n]$, which contains $[a_1, b_1]$. $\qquad\square$

Let $f \colon I \to \mathbf{R}$ be continuous and strictly monotone. Let $J = f(I)$ be the image of f; this is also an interval by the Intermediate Value Theorem. Then if we view f as a function $f \colon I \to J$, we have a bijective function (since f is injective, and changing the target set to J makes it surjective).

The inverse bijection f^{-1} is also strictly monotone, of the same type as f: for instance, if f is increasing, and if $u < v$ are elements of J, then the elements $a = f^{-1}(u)$ and $b = f^{-1}(v)$ such that $u = f(a)$ and $v = f(b)$ satisfy $a < b$, because otherwise from $b < a$ we would get $f(b) < f(a)$.

On the other hand, it is not obvious that f^{-1} should be continuous. This is however true:

> **PROPOSITION 3.4.2.** *Let $f \colon I \to \mathbf{R}$ be continuous and strictly monotone and let $J = f(I)$ be the image of f. Then the inverse $f^{-1} \colon J \to I$ of the bijection $f \colon I \to J$ is continuous.*

PROOF. We assume again first that $I = [a, b]$ for some $a < b$, and we consider the case when f is increasing. Then $J = [f(a), f(b)]$. Let $y \in J$. We prove continuity of f^{-1} at y using Proposition 3.2.10. Thus let (b_n) be a sequence in J that converges to y. We have unique elements $a_n = f^{-1}(b_n)$ and $x = f^{-1}(y)$ in I such that $f(a_n) = b_n$ and $f(x) = y$.

The sequence (a_n) is bounded, since its terms belong to $I = [a, b]$. By the Bolzano–Weierstrass Theorem 2.9.3, there exists a convergent subsequence (a_{n_k}); its limit x satisfies $a \leqslant x \leqslant b$ since all terms do. For all k, we have

$$f(a_{n_k}) = b_{n_k},$$

so the sequence $f(a_{n_k})$ converges to $y = \lim b_{n_k}$. On the other hand, by continuity of f, the same sequence converges to $f(x)$, so that $y = f(x)$.

We are therefore in the situation where a bounded sequence has a unique limit point; applying Proposition 2.9.5, it follows that $a_n \to x$, which means that $f^{-1}(b_n) \to f^{-1}(y)$. Proposition 3.2.10 then shows that f is continuous at y.

In the general case, we use the fact that continuity is a local property: we can find $c < d$ such that $a < y < b$ and

$$[c, d] \subset J$$

and we can argue with the restriction of f to $[f^{-1}(c), f^{-1}(d)]$, which is continuous and (by monotony and the intermediate value theorem) has image $[c, d]$. (To be fully precise, such c and d exist unless a or b is an endpoint of the interval J; when that is the case, say $d = \max J$, we find $c < y \leqslant b$ with $[c, b] \subset J$ and we use this interval instead.) $\qquad\square$

EXAMPLE 3.4.3. It follows for instance that, for any integer $k \in \mathbf{N}$, the function defined by $f(x) = x^{1/k}$ on $\mathbf{R}_+$ is continuous, since it is the inverse bijection of the function $g(x) = x^k$ from $\mathbf{R}_+$ to $\mathbf{R}_+$.

3.5. Other limits of functions

Let $f \colon I \to \mathbf{C}$ be a function defined on an interval $I \subset \mathbf{R}$. If a is an "endpoint" of the interval, even if $a \notin I$, it is possible to approach a by elements of I, and the behavior of $f(x)$ is then sometimes also regular, in the sense that the values approach a fixed real number (or maybe infinity, in the same way as sequences can converge to plus or minus infinity). This leads to various definition of limits of functions. We summarize (most of) them quickly:

- (Finite limit at a point $a \in I$): suppose that $]a - \eta, a + \eta[\subset I$ for some $\eta > 0$; then we say that f has a limit $y \in \mathbf{C}$ when x tends to a, denoted

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

  if for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $x \in I$ satisfies $|x - a| < \delta$, then $|f(x) - y| < \varepsilon$.
- (Finite limit on the left at a point a): suppose that $]a, b] \subset I$ for some $b \in I$; then we say that f has a limit y when x tends to a, denoted

$$\lim_{\substack{x \to a \\ x > a}} f(x) = y$$

  if for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $x \in I$ satisfies $|x - a| < \delta$, then $|f(x) - y| < \varepsilon$.
- (Finite limit on the right at a point b): suppose that $[a, b[\subset I$ for some $a \in I$; then we say that f has a limit y when x tends to b, denoted

$$\lim_{\substack{x \to b \\ x < b}} f(x) = y$$

  if for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $x \in I$ satisfies $|x - b| < \delta$, then $|f(x) - y| < \varepsilon$.
- (Finite limit at plus infinity): suppose that $[a, +\infty[\subset I$ for some $a \in I$; then we say that f has a limit y when x tends to $+\infty$, denoted

$$\lim_{x \to +\infty} f(x) = y$$

  if for all $\varepsilon > 0$, there exists $T > a$ such that if $x > T$, then $|f(x) - y| < \varepsilon$.
- (Finite limit at minus infinity): suppose that $] - \infty, b] \subset I$ for some $b \in I$; then we say that f has a limit y when x tends to $-\infty$, denoted

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

  if for all $\varepsilon > 0$, there exists $T < b$ such that if $x < T$, then $|f(x) - y| < \varepsilon$.
- (Infinite limit on the right at a point a): suppose that $]a, b] \subset I$ for some $b \in I$; then we say that f tends to $+\infty$ (resp. to $-\infty$) when x tends to a, denoted

$$\lim_{\substack{x \to a \\ x > a}} f(x) = +\infty, \ \text{resp.} \ \lim_{\substack{x \to a \\ x > a}} f(x) = -\infty$$

  if for all $M > 0$, there exists $\delta > 0$ such that if $x \in I$ satisfies $|x - a| < \delta$, then $f(x) > M$ (resp $f(x) < -M$).

- (Infinite limit at plus infinity): suppose that $[a, +\infty[\subset I$ for some $a \in I$; then we say that f has limit $+\infty$ (resp. $-\infty$) when x tends to $+\infty$, denoted

$$\lim_{x \to +\infty} f(x) = +\infty, \quad \text{resp.} \quad \lim_{x \to +\infty} f(x) = -\infty$$

  if for all $M > 0$, there exists $T > a$ such that if $x > T$, then $f(x) > M$ (resp. $f(x) < -M$).

- (Infinite limit at minus infinity): suppose that $]-\infty, b] \subset I$ for some $b \in I$; then we say that f has limit $+\infty$ (resp. $-\infty$) when x tends to $-\infty$, denoted

$$\lim_{x \to -\infty} f(x) = +\infty, \quad \text{resp.} \quad \lim_{x \to -\infty} f(x) = -\infty$$

  if for all $M > 0$, there exists $T < b$ such that if $x < T$, then $f(x) > M$ (resp. $f(x) < -M$).

We have not included all the possible variants....

However, all of these can be proved or remembered in a uniform manner by using the analogue of Proposition 3.2.10, and the definition of sequences converging to plus or minus infinity:

PROPOSITION 3.5.1. *In all of the above situations, the limit holds if and only if, for any sequence (a_n) in I converging to the "limit point", whether finite or infinite, the sequence $(f(a_n))$ converges to the stated limit, whether finite or infinite.*

EXAMPLE 3.5.2. (1) We have the following limits:

$$\lim_{\substack{x \to 0 \\ x > 0}} \frac{\sin(x)}{x} = 1, \qquad \lim_{x \to +\infty} \frac{1}{x^2 + 1} = 0$$

$$\lim_{\substack{x \to 0 \\ x < 0}} \frac{1}{x} = -\infty, \qquad \lim_{x \to -\infty} e^{-2x} = +\infty.$$

For instance, we check the second using sequences: for any sequence (a_n) that converges to $+\infty$, we have $a_n^2 + 1 \to +\infty$, hence $1/(a_n^2 + 1) \to 0$.

(2) A function $f \colon [a, b] \to \mathbf{C}$ is continuous at x_0 if and only if

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

(3) Limits can sometimes be used to prove the existence of a maximum or a minimum even for functions defined on other types of intervals using other properties. For instance, let $I = \mathbf{R}$ and assume that f is a continuous function such that $f(x) > 0$ for all $x \in \mathbf{R}$ and

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

Then we claim that f has a maximum (but not necessarily a minimum; an example is $f(x) = 1/(1+x^2)$, where the maximum is $1 = f(0)$, and where there is no minimum since $f(x) > 0$ for all $x \in \mathbf{R}$).

Indeed, since $f(0) > 0$, the assumption implies that for there exists $R > 0$ such that we have $0 < f(x) \leqslant f(0)$ when $|x| > R$. On the compact interval $[-R, R]$, the function f has a maximum M_0 at some point $x_0 \in [-R, R]$, and we have $M \geqslant f(0)$. For $|x| > R$, we get $f(x) \leqslant f(0) \leqslant M$, so M is a maximum for all values of $f(x)$.

(4) Generalizing Examples 2.6.2 and 2.11.5, if $k \in \mathbf{N}_0$ and $l \in \mathbf{N}_0$ are integers and

$$f(x) = a_k x^k + \cdots + a_1 x + a_0, \qquad g(x) = b_l x^l + \cdots + b_1 x + b_0$$

are polynomials with $a_k \neq 0$ and $b_l \neq 0$, then

$$\lim_{x \to +\infty} \frac{f(x)}{g(x)} = \begin{cases} \frac{a_k}{b_l} & \text{if } k = l \\ +\infty & \text{if } k > l \text{ and } a_k/b_l > 0 \\ -\infty & \text{if } k > l \text{ and } a_k/b_l < 0 \\ 0 & \text{if } l > k. \end{cases}$$

3.6. Continuous functions defined on subsets of C

We have defined and studied functions on subsets of real numbers, because only in such cases can be prove the intermediate value theorem or the extremum theorem. However, the definition 3.2.1 can be extended to subsets of $\mathbf{C}$: for $I \subset \mathbf{C}$ and $f \colon I \to \mathbf{C}$, one says that f is continuous if for all $x \in I$, and for every $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ for $y \in I$ such that $|x - y| < \delta$.

In other words, the intervals $]x_0 - \varepsilon, x_0 + \varepsilon[$ around a real number x_0 which describe numbers "close to x_0" are replaced by discs centered at x_0 (which are the sets of complex numbers defined by inequalities $|x - x_0| < \varepsilon$).

The "local" results of Section 3.2 extend to continuous functions defined on subsets of $\mathbf{C}$, and we will use (for instance) Proposition 3.2.8 and Proposition 3.2.10 for functions defined on subsets of $\mathbf{C}$.

CHAPTER 4

Sequences and series of functions and elementary functions

4.1. Uniform convergence

Reference: [2, 15.1].

We have seen that any finite number of "usual" operations, including addition, multiplication, division (when it makes sense) and composition, transforms continuous functions into continuous functions. We now discuss what happens when we attempt to perform some of these operations infinitely many times. In other words, we look at the possibility of defining the values of a function $f(x)$ for $x \in I$ as the limit of sequences $(f_n(x))$, which depend on $x \in I$.

EXAMPLE 4.1.1. (1) The sum of the geometric series with parameter x such that $x \in I =\ ]-1, 1[$ is of this form: we have

$$\frac{1}{1-x} = \lim_{n \to +\infty} f_n(x)$$

for any $x \in I$, where

$$f_n(x) = 1 + \cdots + x^n.$$

In this case, the functions $f_n \colon I \to \mathbf{R}$ are continuous (they are polynomials), and the limit is also continuous (since $1 - x \neq 0$).

(2) However the following example shows that even if all f_n are continuous functions, and if $(f_n(x))$ converges for all $x \in I$, the values of the limit $\lim f_n(x) = f(x)$ might not define a continuous function.

Let $I = [0, 1]$. Define $f_n(x) = x^n$ for $x \in I$; this is a continuous function. But note that for $0 \leqslant x < 1$ we have

$$\lim_{n \to +\infty} x^n = 0$$

(by (2.6)), while

$$\lim_{n \to +\infty} 1^n = 1.$$

Therefore $f_n(x)$ converges for all x to

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

This function f is not continuous at $x_0 = 1$.

We will use the same notation (f_n) for a sequence of functions as for a sequence of numbers; note that we can add sequences of functions, multiply them, etc

The first thing to do is to find a condition on a sequence of functions that ensures that, if they converge and are continuous, then so is the limit. It is not obvious that a simple condition of this type should exist, but it turns out that there is one, which intuitively is that the "speed of convergence" of $f_n(x)$ to the limit should be "the same" for all x.
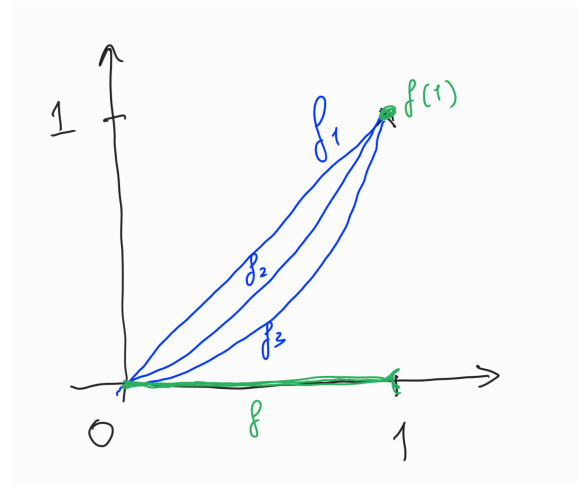
DEFINITION 4.1.2 (Uniform convergence). Let $I \subset \mathbf{C}$ be a set of complex numbers. For $n \in \mathbf{N}$, let $f_n \colon I \to \mathbf{C}$ be an arbitrary function. Let $f \colon I \to \mathbf{C}$ be a function. One says that *the sequence (f_n) converges uniformly*[42] *to f on I* if, for all $\varepsilon > 0$, there exists $N \in \mathbf{N}$, such that for all $x \in I$, we have $|f(x) - f_n(x)| < \varepsilon$ for $n \geqslant N$.

REMARK 4.1.3. (1) The logical formula for uniform convergence is:

$$\forall \varepsilon > 0, \ \exists N \in \mathbf{N}, \ \forall x \in I, \ (n \geqslant N \longrightarrow |f(x) - f_n(x)| < \varepsilon).$$

We should compare this to the formula that states that $(f_n(x))$ converges for all $x \in I$, which can be written

$$\forall \varepsilon > 0, \ \forall x \in I, \ \exists N \in \mathbf{N}, \ (n \geqslant N \longrightarrow |f(x) - f_n(x)| < \varepsilon).$$

One sees that the only difference is the exchange of the location of $\exists N \in \mathbf{N}$ and $\forall x \in I$: this has the effect that N *only depends on ε*, and not on the point x.

In particular, uniform convergence implies that $f_n(x)$ converges to $f(x)$ for all x, but it is a stronger condition.

(2) In practice, one attempts to prove that (f_n) converges uniformly to f by comparison: one tries to prove an inequality of the form

$$|f(x) - f_n(x)| \leqslant b_n$$

for all $x \in I$, where (b_n) is a *fixed* sequence of non-negative real numbers (independent of the choice of x) that converges to 0. Then we get

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

for all $x \in I$ as soon as $b_n < \varepsilon$.

THEOREM 4.1.4. *Let $I \subset \mathbf{C}$. Let (f_n) be a sequence of* continuous *functions defined on I and f a function $f \colon I \to \mathbf{C}$. Suppose that (f_n) converges* uniformly *to f on I. Then f is continuous on I.*

PROOF. Let $x_0 \in I$. We check continuity at x_0. Let $x \in I$. We need to bound $|f(x) - f(x_0)|$ when x is close to x_0. Since we only know that $f(x)$ is approached by $f_n(x)$ and $f(x_0)$ by $f_n(x_0)$ for n large enough, we use the triangle inequality to write

$$|f(x) - f(x_0)| \leqslant |f(x) - f_n(x)| + |f_n(x) - f_n(x_0)| + |f_n(x_0) - f(x_0)|$$

where $n \in \mathbf{N}$ can be chosen arbitrarily. Note that all three terms are small under suitable conditions: the first and third, when n is large enough, because $f_n(x) \to f(x)$ and

$f_n(x_0) \to f(x_0)$, and the second when x is close enough to x_0, because f_n is continuous at x_0 for a fixed n. However, we need all three to be *simultaneously* small, which is where uniform convergence is needed.

Precisely, let $\varepsilon > 0$ be given. Because of *uniform* convergence, we can find a *single* integer $n \in \mathbf{N}$ such that

$$|f(x) - f_n(x)| < \frac{\varepsilon}{3}, \qquad |f(x_0) - f_n(x_0)| < \frac{\varepsilon}{3},$$

whatever the choice of $x \in I$ is. The inequality above implies for this value of n that

$$|f(x) - f(x_0)| \leqslant \frac{2\varepsilon}{3} + |f_n(x) - f_n(x_0)|,$$

for all $x \in I$. Now, since the particular function f_n is continuous, we obtain $\delta > 0$ such that $|f_n(x) - f_n(x_0)| < \varepsilon/3$ whenever $x \in I$ satisfies $|x - x_0| < \delta$. For all such x, we get

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

and this proves the continuity of f at x_0. □

EXAMPLE 4.1.5. (1) We go back to the sequence (f_n) defined on $[0, 1]$ by $f_n(x) = x^n$. According to the theorem, this cannot converge uniformly to the function f equal to 0 for $x \neq 1$ and to 1 for $x = 1$. Let us see this concretely.

We have

$$|f(x) - f_n(x)| = \begin{cases} 0 & \text{if } x = 1 \\ x^n & \text{if } 0 \leqslant x < 1. \end{cases}$$

If we want this to be (say) $< \frac{1}{2}$ for a given $x < 1$, we need $x^n < \frac{1}{2}$, or $n > 2/\log(1/x)$, and this value of n increases when x is closer and closer to 1. This means that the convergence is *not* uniform.

(2) Could we have deduced the continuity of the sum of the geometric series from Theorem 4.1.4?

Here the functions f_n are given for $x \in]-1, 1[$ by

$$f_n(x) = 1 + \cdots + x^n = \frac{1 - x^{n+1}}{1 - x},$$

so that

$$|f(x) - f_n(x)| = \frac{|x|^{n+1}}{|1 - x|}.$$

For the same reason as in (1), we see that there is no uniform convergence for $x \in]-1, 1[$. However, we can be more clever and still obtain the result. Let $-1 < a < b < 1$ be two real numbers, and suppose that we consider only $x \in]a, b[$. Then we obtain

$$|f(x) - f_n(x)| = \frac{\max(|a|, |b|)^{n+1}}{\min(|1 - a|, |1 - b|)}.$$

for $x \in]a, b[$. The right-hand side converges to 0 as $n \to +\infty$ (by (2.6)) so we deduce that (f_n) converges uniformly to $f(x) = 1/(1 - x)$ on $]a, b[$. In particular, by Theorem 4.1.4, the limit f is a continuous function on $]a, b[$. Since any $x \in]-1, 1[$ belongs to some interval $]a, b[$ with $-1 < a < b < 1$, and since continuity is a local property, this means that the limit f is in fact continuous on all of I.

This example computation is very important, because it is very frequent that one doesn't prove uniform convergence over the whole definition set, but only over suitable smaller subsets; this shows that this can be enough to deduce continuity everywhere.

There is also an analogue of the Cauchy Criterion for uniform convergence.

PROPOSITION 4.1.6. *Let $I \subset \mathbf{C}$ and let (f_n) be a sequence of functions on I.*
Suppose that for every $\varepsilon > 0$, there exists $N \in \mathbf{N}$, such that for all integers $n \geqslant N$ and $m \geqslant N$, and for all $x \in I$, we have

$$|f_n(x) - f_m(x)| < \varepsilon.$$

Then (f_n) converges uniformly to some function f.

PROOF. For any given $x \in I$, the usual Cauchy Criterion is satisfied for the sequence $(f_n(x))$, so there exists a limit

$$f(x) = \lim_{n \to +\infty} f_n(x)$$

for all $x \in I$ (Theorem 2.8.8). We now prove that (f_n) converges uniformly to f.

Let $\varepsilon > 0$ and let $x \in I$. Let $N \in \mathbf{N}$ such that

$$|f_n(x) - f_m(x)| < \frac{\varepsilon}{2}$$

for all $x \in I$ when n and m are $\geqslant N$. We keep $n \geqslant N$ fixed, and view this inequality as a property of the sequence $(|f_n(x) - f_m(x)|)_{m \in \mathbf{N}}$. This sequence converges to $|f_n(x) - f(x)|$, so these inequalities give*

$$|f_n(x) - f(x)| \leqslant \frac{\varepsilon}{2} < \varepsilon$$

for all $n \geqslant N$ and all $x \in I$, which gives the uniform convergence. $\qquad\square$

REMARK 4.1.7. As usual, we can prove the Cauchy Criterion for uniform convergence of (f_n) by proving, for $m \geqslant n$, an inequality

$$|f_n(x) - f_m(x)| \leqslant b_n$$

for all $x \in I$, where (b_n) is a fixed sequence converging to 0 (which is independent of x).

4.2. Normal convergence

Reference: [2, 7.3].

Consider now a *series* of functions, which means a sequence, denoted

$$\sum_{n=1}^{+\infty} f_n$$

where each $f_n \colon I \to \mathbf{C}$ is a function, corresponding to the partial sums

$$s_n(x) = \sum_{k=1}^{n} f_k(x).$$

Supppose that each f_n is bounded on I, say $|f_n(x)| \leqslant b_n$ for all $x \in I$, where $b_n \in \mathbf{R}_+$ is independent of x. Then for $m \geqslant n$, we obtain

$$|s_n(x) - s_m(x)| = \left| \sum_{k=n+1}^{m} s_k(x) \right| \leqslant \sum_{k=n+1}^{m} b_k$$

by the triangle inequality. If we assume that the series $\sum b_n$ is convergent, then the Cauchy Criterion for this series implies that the sequence (s_n) converges uniformly on I (by Remark 4.1.7).

DEFINITION 4.2.1. A series of functions $\sum f_n$ *converges normally*[43] if each $|f_n|$ is bounded on I by some $b_n \in \mathbf{R}_+$ such that $\sum b_n$ converges.

---

* By Remark 2.5.10, we cannot deduce $|f_n(x) - f(x)| < \varepsilon/2$.

We have therefore checked:

THEOREM 4.2.2. *Suppose that the series $\sum f_n$ converges normally on I. Then it converges uniformly on I.*

REMARK 4.2.3. (1) This concept is particularly useful when $I = [a, b]$ for some real numbers $a < b$, and if f_n is continuous for each n, since then the Extremum Theorem 3.3.2 shows that each f_n is indeed bounded.

(2) Normal convergence implies uniform convergence for series, but is not equivalent to it.

Combined with Theorem 4.1.4, normal convergence leads to:

COROLLARY 4.2.4. *Let (f_n) be a sequence of continuous functions on I such that $\sum f_n$ converges normally on I. Then the series converges uniformly to a continuous function f on I.*

EXAMPLE 4.2.5. The series
$$\sum_{n=1}^{+\infty} \frac{\cos(nx)}{n^2}$$
for $x \in \mathbf{R}$ converges normally, since the series $\sum n^{-2}$ is convergent. On the other hand, the series
$$\sum_{n=1}^{+\infty} \frac{\cos(nx)}{n}$$
doesn't converge normally, since for $x = 0$ we have $\cos(nx) = 1$, so we need to take $b_n = 1$ and $\sum n^{-1}$ does not converge.

4.3. Power series

Reference: [2, 6.4, 7.3].

We now consider the simplest type of series of functions, which can be used to define and study the most important elementary functions. These are the *power series*[44], of the form
$$\sum_{n=0}^{+\infty} a_n x^n = a_0 + a_1 x + \cdots + a_n x^n + \cdots$$
(note that these can be interpreted as a natural attempt to generalize the notion of polynomials when we allow infinitely many coefficients, and in particular that any polynomial function is a power series, where the coefficients a_n vanish for n large enough).

The power series $\sum a_n x^n$ certainly converges for $x = 0$ (where the sum is equal to a_0). At other points, convergence is controlled by the following proposition:

PROPOSITION 4.3.1. *Let (a_n) be a sequence of complex numbers. Let $x_0 \in \mathbf{C}$ be a non-zero complex number such that the series*
$$\sum_{n=0}^{+\infty} a_n x_0^n$$
converges.

(1) For any $x \in \mathbf{C}$ such that $|x| < |x_0|$, the series $\sum a_n x^n$ converges absolutely.

(2) For any $r < |x_0|$, the series $\sum a_n x^n$ converges normally on the closed disc D centered at 0 with radius r, in other words on
$$D_r = \{x \in \mathbf{C} \mid |x| \leqslant r\}.$$
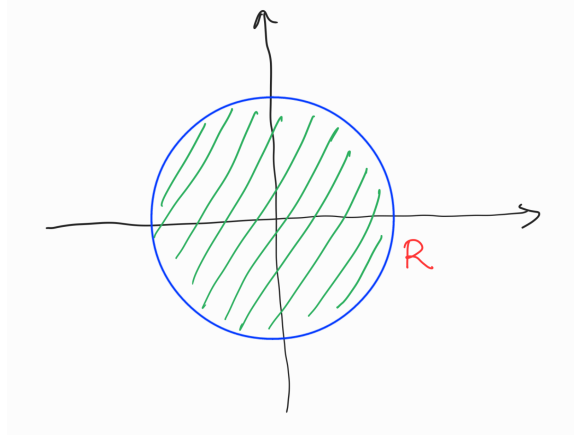
PROOF. The convergence of the series $\sum a_n x_0^n$ implies that $a_n x_0^n \to 0$ (see Proposition 2.10.3), and in particular the sequence $(a_n x_0^n)$ is bounded (Lemma 2.5.7). Let $M \in \mathbf{R}_+$ be such that $|a_n x_0^n| \leqslant M$ for all $n \in \mathbf{N}_0$.

Let $x \in \mathbf{C}$ be such that $|x| < |x_0|$. Then for all $n \in \mathbf{N}_0$, we have

$$|a_n x^n| = |(x/x_0)|^n |a_n x_0|^n \leqslant M|(x/x_0)|^n,$$

and since $|x/x_0| < 1$, the series $\sum a_n x^n$ converges absolutely by comparison with the geometric series.

Moreover, if $|x| \leqslant r < |x_0|$, then the inequality above gives

$$|a_n x^n| = |(x/x_0)|^n |a_n x_0|^n \leqslant M(r/|x_0|)^n,$$

and since the right-hand side is now a sequence independent of $x \in D$, which defines a convergent series, this means that the power series converges normally on D. $\qquad\square$

This suggests the following definition:

DEFINITION 4.3.2. Let (a_n) be a sequence of complex numbers. The *radius of convergence*[45] of the series $\sum a_n x^n$ is defined to be $R = +\infty$ if the series converges for all $x \in \mathbf{C}$, and otherwise

$$R = \sup\{r \in \mathbf{R}_+ \mid \sum a_n x_0^n \text{ converges for some } x_0 \text{ with } |x_0| = r\}.$$

The previous proposition leads then to the following result:

COROLLARY 4.3.3. *Let (a_n) be a sequence of complex numbers and R the convergence radius of the power series $\sum a_n x^n$.*

(1) For any r such that $0 \leqslant r < R$, the series $\sum a_n x^n$ converges normally on

$$D_r = \{x \in \mathbf{C} \mid |x| \leqslant r\}.$$

(2) For any $x \in \mathbf{C}$ such that $|x| < R$, the series $\sum a_n x^n$ converges absolutely, and its sum is a continuous function on the open disc

$$D'_R = \{x \in \mathbf{C} \mid |x| < R\}.$$

(3) For any $x \in \mathbf{C}$ with $|x| > R$, the series $\sum a_n x^n$ diverges.

REMARK 4.3.4. If $R = +\infty$, then the condition $0 \leqslant r < R$ is always satisfied.

PROOF. (1) If $0 \leqslant r < R$, then r is not an upper-bound of the (non-empty) set

$$E = \{r \in \mathbf{R}_+ \mid \sum a_n x_0^n \text{ converges for some } x_0 \text{ with } |x_0| = r\}.$$

So there exists $s \in E$ such that $r < s$, and by definition there exists $x_0 \in \mathbf{C}$ with $|x_0| = s$ for which the series converges; the proposition then implies that the power series converges normally on D_r.

(2) If $|x| < R$, then putting $r = |x|$ in (1), we get the absolute convergence of the power series at x. Moreover, since the functions $f_n(x) = a_n x^n$ are continuous, Corollary 4.2.4 implies that the sum of the series is continuous on D_r. Since continuity is a local property, and any $x \in D'_R$ belongs to D_r for some $r < R$ (namely, $x \in D_{|x|}$), this means that the sum of the series is continuous on D'_R.

(3) Suppose that $|x| > R$. Then $\sum a_n x^n$ must diverge, since otherwise, Proposition 4.3.1 would imply that $\sum a_n y^n$ converges for $R \leqslant |y| < |x|$, which is not the case by definition. $\qquad \square$

REMARK 4.3.5. (1) We will see later that power series have much stronger regularity properties than only continuity.

(2) This corollary gives *no information* about the convergence of the series $\sum a_n x^n$ on the circle with radius equal to the radius of convergence (when R is finite and positive). This is because there is no general result here: the series might converge on the whole circle, at no point of the circle, or just at some points – we will see examples of all these possibilities.

EXAMPLE 4.3.6. (1) It can well happen that the radius of convergence is equal to 0, so that the power series only converges when $x = 0$. For instance consider $a_n = n!$, so that the power series is

$$\sum_{n=0}^{+\infty} n! x^n.$$

For any non-zero complex number x, we have

$$\lim_{n \to +\infty} |n! x^n| = +\infty$$

by Example 2.11.5, (1), so that the series $\sum n! x^n$ cannot be convergent.

(2) The most important power series with infinite radius of convergence is

$$\sum_{n=0}^{+\infty} \frac{x^n}{n!}.$$

Indeed, this series converges for all $x \in \mathbf{C}$ by Proposition 2.10.14, (3).

(3) The geometric series

$$\sum_{n=0}^{+\infty} x^n = \frac{1}{1-x}$$

has radius of convergence equal to 1. In this case, if $|x| = 1$, we have $|x^n| = 1$ for all $n \in \mathbf{N}_0$, so that the series diverges at all points of the circle with radius 1.

(4) The series

$$\sum_{n=1}^{+\infty} \frac{x^n}{n}$$

has radius of convergence equal to 1. Indeed, since $|x^n/n| \leqslant |x^n|$, it converges absolutely where the geometric series converges, and since

$$\lim_{n \to +\infty} \frac{x^n}{n} = +\infty$$

for $|x| > 1$, it cannot converge at any $x \in \mathbf{C}$ with $|x| > 1$.

Note here that the series diverges for $x = 1$ (it is the series $\sum 1/n$, which diverges by Example 2.8.9, (2)), but converges (not absolutely) for $x = -1$ (it is then the alternating series of Example 2.10.9, (2)).

(5) The series

$$\sum_{n=1}^{+\infty} \frac{x^n}{n^2}$$

has radius of convergence equal to 1, for similar reasons as in the previous example, but the series converges absolutely (even normally) whenever $|x| = 1$ since then

$$\left|\frac{x^n}{n^2}\right| \leqslant \frac{1}{n^2}$$

and the series $\sum 1/n^2$ is convergent (Proposition 2.10.14, (1)).

To determine, or at least estimate, the radius of convergence of a power series $\sum a_n x^n$, one can as usual attempt a comparison: if (b_n) is a sequence of non-negative real numbers such that

$$|a_n| \leqslant b_n$$

for all $n \in \mathbf{N}_0$, then the radius of convergence for $\sum a_n x^n$ is at least that of $\sum b_n x_n$.

Moreover, we have the following useful fact:

> LEMMA 4.3.7. *Suppose that $\sum a_n x^n$ has radius of convergence equal to R. Then for any $k \in \mathbf{R}$, the power series*
>
> $$\sum_{n=1}^{+\infty} a_n n^k x^n$$
>
> *has the same radius of convergence.*

We start the series here at $n = 1$ because k may be negative.

PROOF. It suffices to check that if

$$\sum a_n x_0^n$$

converges then for $|x| < |x_0|$ and $|y| > |x_0|$, the series

$$\sum a_n n^k x^n \qquad \text{and} \qquad \sum a_n n^k y^n$$

are respectively convergent and divergent. The first property is proved as in Proposition 4.3.1, using Proposition 2.10.14, (2). The second is obtained by an argument like that in Corollary 4.3.3, (3). $\qquad\square$

EXAMPLE 4.3.8. We give here an interesting example of application of power series, which is one of the simplest cases of the idea of *generating functions*, first invented by Euler.

Let $(F_n)_{n \in \mathbf{N}_0}$ be the sequence defined inductively by $F_0 = F_1 = 1$ and $F_{n+2} = F_{n+1} + F_n$ for all $n \in \mathbf{N}_0$ (this is just the Fibonacci sequence of Example 1.2.3, (2), renumbered to start at F_0). We consider the power series

$$\sum_{n=0}^{+\infty} F_n x^n.$$

First we check that this series has a positive radius of convergence R. Indeed, by induction on n, we see that $0 \leqslant F_n \leqslant 2^n$ for all $n \in \mathbf{N}_0$. Therefore the radius of convergence R is at least that of $\sum 2^n x^n$, which is equal to $1/2$, since this series converges for $|2x| < 1$, and diverges for $|2x| > 1$. Let $f(x)$ be the sum of the power series for $|x| < R$.

We will now use the inductive definition of the sequence to *compute* the function f. This will allow us to find a formula for the sequence.

We observe that

$$x^2 f(x) = \sum_{n=0}^{+\infty} F_n x^{n+2} = \sum_{n=2}^{+\infty} F_{n-2} x^n, \qquad x f(x) = \sum_{n=1}^{+\infty} F_{n-1} x^n.$$

Therefore we get

$$(x^2 + x) f(x) = \sum_{n=2}^{+\infty} F_{n-2} x^n + \sum_{n=1}^{+\infty} F_{n-1} x^n = F_0 x + \sum_{n=2}^{+\infty} (F_{n-2} + F_{n-1}) x^n$$

$$= F_0 x + \sum_{n=2}^{+\infty} F_n x^n$$

$$= f(x) + (F_0 - F_1)x - F_0 = f(x) - 1.$$

It follows that

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

In order to deduce a formula for F_n from this expression, we need a different way to express it as a power series. For this, we observe that

$$1 - x - x^2 = (1 - x/\alpha)(1 - x/\beta)$$

where

$$\alpha = \frac{-1 + \sqrt{5}}{2}, \qquad \beta = \frac{-1 - \sqrt{5}}{2}$$

(these are the two roots of the equation $x^2 + x - 1 = 0$). Then we find real numbers a and b such that

$$\frac{1}{1 - x - x^2} = \frac{a}{1 - x/\alpha} + \frac{b}{1 - x/\beta}.$$

Indeed, this is true with

$$a = \frac{\beta}{\alpha - \beta}, \qquad b = -\frac{\alpha}{\alpha - \beta}.$$

Then by the geometric series, we deduce that for $|x| < \min(|\alpha|, |\beta|)$, we have

$$\sum_{n=0}^{+\infty} F_n x^n = f(x) = a \sum_{n=0}^{+\infty} \frac{x^n}{\alpha^n} + b \sum_{n=0}^{\infty} \frac{x^n}{\beta^n}.$$

We will see later that two power series can only be equal as functions if all their coefficients agree, and we conclude that

$$F_n = \frac{a}{\alpha^n} + \frac{b}{\beta^n}$$

for all $n \in \mathbf{N}_0$. (Once the formula is known, it can be checked by induction, as in Example 1.2.3, (2), but the method that we have just described can lead to the *discovery* of the formula.)
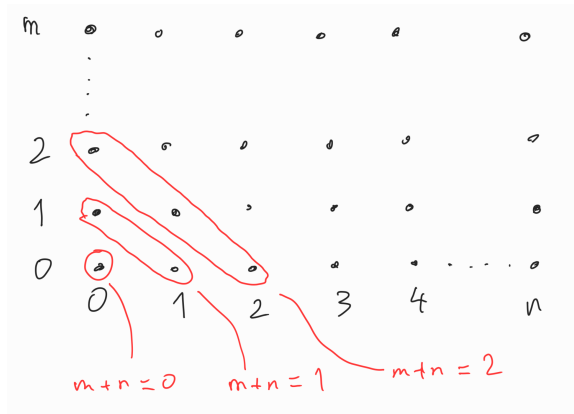
4.4. The elementary functions, I: the exponential

In the next two sections, we will use power series to define the basic elementary functions: the exponential function, the trigonometric functions, and we will establish *from scratch* their basic properties, including the periodicity of the trigonometric functions (thus defining the number π for which the smallest positive period is 2π). We can also define the logarithm as reciprocal bijection of the exponential (and later the inverse trigonometric functions).

> DEFINITION 4.4.1 (Exponential, cosine and sine). The (complex) exponential function $\exp\colon \mathbf{C} \to \mathbf{C}$ is the function defined by
> $$\exp(x) = \sum_{n=0}^{+\infty} \frac{x^n}{n!}.$$

We already know that the exponential power series has infinite radius of convergence, so the function is indeed defined for all $x \in \mathbf{C}$, and it is a continuous function on $\mathbf{C}$.

Note that although we defined $\exp(x)$ for every $x \in \mathbf{C}$, the exponential of a real number is a real number. By looking at the constant term, we see that $\exp(0) = 1$.

The next result is the most important property of the exponential function.

> THEOREM 4.4.2. *For any x and y in $\mathbf{C}$, we have*
> $$(4.1) \qquad\qquad \exp(x)\exp(y) = \exp(x+y).$$

PROOF. This is an example of a general fact about multiplication of two absolutely convergent series. The idea is to multiply

$$\sum_n a_n, \quad \text{and} \quad \sum_n b_n$$

by taking all the products $a_n b_m$ and combining them according to the value of $n + m$.

If we assume that this "works" (in the sense that if we denote by c_k the sum of the $a_n b_m$ where $n + m = k$, the series $\sum_k c_k$ converges and that its sum is the product of the two series), then we get the series

$$\sum_{k=0}^{+\infty}\Big(\sum_{n=0}^{k} \frac{x^n y^{k-n}}{n!(k-n)!}\Big),$$

85
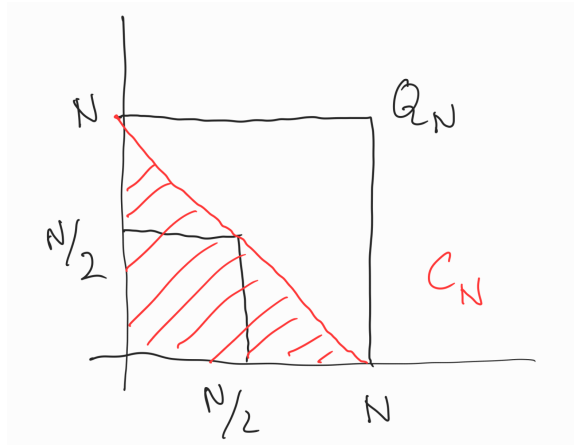
in the case of $\exp(x)\exp(y)$. we recognize the terms of this series are

$$\sum_{n=0}^{k}\frac{x^n y^{k-n}}{n!(k-n)!}=\frac{1}{k!}\sum_{n=0}^{k}\frac{k!}{n!(k-n)!}x^n y^{k-n}=\frac{1}{k!}\sum_{n=0}^{k}\binom{k}{n}x^n y^{k-n}=\frac{(x+y)^k}{k!}$$

by (1.38) and the binomial theorem (Theorem 1.6.10), so that the series is

$$\sum_{k=0}^{+\infty}\frac{(x+y)^k}{k!}=\exp(x+y).$$

This explains why the statement of the theorem is to be expected; the proof that the product of the series has the required property is explained below. $\qquad\square$

PROPOSITION 4.4.3. *Let $\sum a_n$ and $\sum b_n$ be absolutely convergent series with sums a and b respectively. For $k\in\mathbf{N}_0$, let*

$$c_k=\sum_{k=0}^{n}a_k b_{n-k}=\sum_{n+m=k}a_n b_m.$$

Then the series $\sum c_k$ converges absolutely and its sum is ab.

PROOF. We first assume that $a_n\geqslant 0$ and $b_m\geqslant 0$ for all n and m. Then the point is that the partial sum

$$\sum_{k=0}^{N}c_k$$

can be interpreted as the sum of the products $a_n b_m$ over the points $(n,m)\in\mathbf{N}_0\times\mathbf{N}_0$ in the set C_N defined by the condition $n+m\leqslant k$. On the other hand, the product of partial sums

$$\Big(\sum_{n=0}^{N}a_n\Big)\Big(\sum_{m=0}^{N}b_m\Big)$$

is the sum of the products $a_n b_m$ over the set Q_N of all points (n,m) with $0\leqslant n\leqslant N$ and $0\leqslant n\leqslant N$. Now we observe that

$$Q_{N/2}\subset C_N\subset Q_N,$$

which because all coefficients $a_n b_m$ are non-negative implies that

$$\left(\sum_{n=0}^{N/2} a_n\right)\left(\sum_{m=0}^{N/2} b_m\right) \leqslant \sum_{k=0}^{N} c_k \leqslant \left(\sum_{n=0}^{N} a_n\right)\left(\sum_{m=0}^{N} b_m\right).$$

Because the right-hand side is bounded by ab, it follows that the series $\sum_k c_k$ has bounded partial sums, and therefore converges; then since both left and right-hand sides converge to ab, we deduce that the sum of the series $\sum c_k$ is also ab.

We now come back to the general case where a_n and b_m are complex numbers. Note that

$$|c_k| \leqslant \sum_{n=0}^{k} |a_n|\,|b_{k-n}|$$

by the triangle inequality. Since we assumed that the series are absolutely convergent, this implies that

$$\sum_k |c_k| \leqslant \left(\sum_n |a_n|\right)\left(\sum_n |b_n|\right),$$

which shows that the series $\sum c_k$ is absolutely convergent. There only remains to check that its sum c is equal to ab. For this, we write

$$\left|\sum_{k=0}^{N} c_k - \left(\sum_{n=0}^{N} a_n\right)\left(\sum_{m=0}^{N} b_m\right)\right| \leqslant \left(\sum_{n>N/2} |a_n|\right)\left(\sum_{mn>N/2} |b_m|\right),$$

where the right-hand side is the product of two sequences that tend to 0 as $N \to +\infty$. Since the left-hand side converges to $|c - ab|$, we get $c = ab$ as claimed. $\qquad\square$

This formula leads to many important consequences.

COROLLARY 4.4.4. (1) *For any $x \in \mathbf{C}$, we have $\exp(x) \neq 0$ and $\exp(x)^{-1} = \exp(-x)$.*
(2) *For any $x \in \mathbf{R}$, we have $\exp(x) > 0$, and the exponential on $\mathbf{R}$ is strictly increasing with*

$$\lim_{x \to -\infty} \exp(x) = 0, \qquad \lim_{x \to +\infty} \exp(x) = +\infty.$$

(3) *For any $x \in \mathbf{C}$, we have $\exp(\bar{x}) = \overline{\exp(x)}$.*

PROOF. Since $\exp(x)\exp(-x) = \exp(0) = 1$, we get the first property.

Then we note that $\exp(x) \geqslant 1$ for all $x \in \mathbf{R}_+$, because then all terms of the series are $\geqslant 0$ and the first term is equal to 1. Since $\exp(-x) = 1/\exp(x)$, it follows that $\exp(x) > 0$ for all real x. If we then take $x < y$ in $\mathbf{R}$, then we get

$$\exp(y) - \exp(x) = \exp(x)(\exp(y - x) - 1) > 0$$

since $\exp(y - x) \geqslant 1 + (y - x) > 1$ for $y - x > 0$. Similarly, for $x \geqslant 0$, we have $\exp(x) \geqslant 1 + x$, and therefore $\exp(x) \to +\infty$ when $x \to +\infty$. For $x < 0$, we have

$$\exp(x) = \frac{1}{\exp(-x)} \to 0$$

since $-x \to +\infty$ when $x \to -\infty$. This proves (2).

For (3), since the conjugation function is continuous and since $\overline{x^n n} = \bar{x}^n$ for all $n \in \mathbf{N}_0$, we get

$$\overline{\exp(x)} = \overline{\sum_{n=0}^{+\infty} \frac{x^n}{n!}} = \sum_{n=0}^{+\infty} \frac{\bar{x}^n}{n!} = \exp(\bar{x}).$$

$\qquad\square$

This allows us to define the logarithm function on $]0, +\infty[$, and the general power functions. Indeed, part (2) of the corollary implies that the continuous function exp is injective on $\mathbf{R}$, and that its image is the interval $]0, +\infty[$.

DEFINITION 4.4.5 (Logarithm). The logarithm function is the function $\log \colon]0, +\infty[\to \mathbf{R}$ which is the reciprocal bijection of the exponential.

According to Proposition 3.4.2, the logarithm is continuous and is a strictly increasing bijection, with image $\mathbf{R}$, so that

$$\lim_{x \to +\infty} \log(x) = +\infty, \qquad \lim_{\substack{x \to 0 \\ x > 0}} \log(x) = -\infty.$$

We have $\log(1) = 0$ since $\exp(0) = 1$, and the formulas

$$\exp(x + y) = \exp(x)\exp(y), \qquad \exp(-x) = 1/\exp(x)$$

implies that

$$\log(ab) = \log(a) + \log(b), \qquad \log(1/a) = -\log(a)$$

for all $a > 0$ and $b > 0$ (because, for instance, the exponential of the left-hand side is ab, and is equal to the exponential $\exp(\log(a) + \log(b)) = \exp(\log(a))\exp(\log(b))$ of the right-hand side).

DEFINITION 4.4.6 (Power functions). Let $x > 0$ and let $a \in \mathbf{C}$. We define

$$x^a = \exp(a\log(x)).$$

It follows that $\log(x^a) = a\log(x)$.

This new definition is compatible with the usual definition of x^n if $n \in \mathbf{Z}$, and moreover leads to the usual notation e^x for the exponential.

PROPOSITION 4.4.7. Let $x > 0$, $y > 0$ and $a \in \mathbf{R}$, $b \in \mathbf{R}$.
(1) If $a \in \mathbf{Z}$ then x^a as defined above corresponds to the usual definition as integral power or inverse of an integral power; if $k \in \mathbf{N}$, then $x^{1/k}$ is the unique positive real number such that $(x^{1/k})^k = x$, in particular $x^{1/2} = \sqrt{x}$.
(2) We have
$$(xy)^a = x^a y^a.$$
(3) We have
$$x^{a+b} = x^a x^b, \qquad (x^a)^b = x^{ab}.$$
(4) Let $e = \exp(1) \in \mathbf{R}_+$. For any $x \in \mathbf{C}$, we have $e^x = \exp(x)$.

PROOF. (1) We note that according to the exponential definition, we have

$$x^1 = \exp(1\log(x)) = \exp(\log(x)) = x,$$

and

$$x^{-1} = \exp(-\log(x)) = \exp(\log(1/x)) = \frac{1}{x}.$$

The fact that x^a has the usual meaning for $a \in \mathbf{Z}$ follows by induction using the property (3) that we will prove soon. And if $k \in \mathbf{N}$, then we get

$$(x^{1/k})^k = x^1 = x.$$

(2) We have

$$(xy)^a = \exp(a\log(xy)) = \exp(a(\log(x) + \log(y)))$$
$$= \exp(a\log(x))\exp(a\log(y)) = x^a y^a.$$
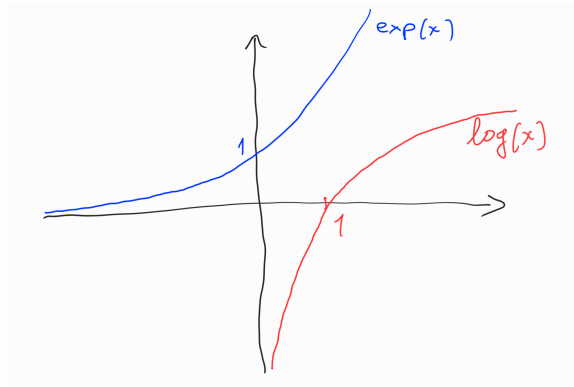
(3) We have

$$x^{a+b} = \exp((a+b)\log(x)) = \exp(a\log(x) + b\log(x))$$
$$= \exp(a\log(x))\exp(b\log(x)) = x^a x^b,$$

and

$$(x^a)^b = \exp(b\log(x^a)) = \exp(ab\log(x)) = x^{ab}.$$

(4) We have $\log(e) = 1$, so that

$$\exp(x) = \exp(x\log(e)) = e^x.$$

$\square$

REMARK 4.4.8. The number $e = \exp(1)$ is by definition the real number given by

$$e = 1 + \frac{1}{2} + \cdots + \frac{1}{n!} + \cdots$$

Note that this series converges very quickly; in fact the remainder after $n \geqslant 1$ terms is

$$\frac{1}{n!} + \frac{1}{(n+1)!} + \cdots = \frac{1}{n!}\left(1 + \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \cdots\right).$$

For any integer $j \geqslant 1$, we have

$$(n+1)\cdots(n+j) \geqslant 2^j,$$

so we obtain the upper-bound

$$0 \leqslant e - \sum_{k=0}^{n-1} \frac{1}{k!} \leqslant \frac{1}{n!}\sum_{j=0}^{+\infty} \frac{1}{2^j} = \frac{2}{n!}.$$

Computing a few terms, we find the decimal approximation

$$e = 2.718281828459045235360287471352\ldots$$

PROPOSITION 4.4.9. *For any real number r and any $a > 0$, we have*

$$\lim_{x \to +\infty} x^r e^{ax} = +\infty, \qquad \lim_{x \to +\infty} x^r e^{-ax} = 0,$$

and if $r > 0$ and $a \geqslant 0$, then

$$\lim_{\substack{x \to 0 \\ x > 0}} x^r \log(x)^a = 0, \qquad \lim_{x \to +\infty} x^r \log(x)^a = +\infty.$$

One remembers this by saying that e^x "grows much faster" than any power of x for $x \to +\infty$, and "goes to zero much faster than any power of x grows", and conversely that $\log(x)$ grows slower than any power of x at infinity.

PROOF. We have
$$x^r e^{ax} = (x^{r/a} e^x)^a,$$
so that it suffices to consider the case $a = 1$ of the first limit. Let $m \in \mathbf{N}$ be such that $r \geqslant -m$; note that for $x \geqslant 0$, we get
$$e^x \geqslant \frac{x^{m+1}}{(m+1)!},$$
hence
$$x^r e^x \geqslant \frac{x}{(m+1)!},$$
which tends to $+\infty$ when $x \to +\infty$. For the second limit, note that
$$x^r e^{-ax} = (x^{-r} e^{ax})^{-1},$$
which converges to 0 since $x^{-r} e^{ax} \to +\infty$ as $x \to +\infty$.

For the other limits, write $y = \log(x)$ for $x > 0$; then $x = e^y$ and $x^r \log(x)^a = e^{ry} y^a$; since $y \to +\infty$ if $x \to +\infty$, we get
$$x^r \log(x)^a \to +\infty \text{ as } x \to +\infty,$$
and since $y \to -\infty$ if $x \to 0$ (with $x > 0$), we get
$$x^r \log(x)^a \to 0 \text{ as } x \to 0, \ x > 0.$$

$\square$

4.5. The elementary functions, II: trigonometry

DEFINITION 4.5.1 (Cosine and sine). The (complex) sine and cosine functions are defined by

$$(4.2) \qquad \cos(x) = \frac{e^{ix} + e^{-ix}}{2}, \qquad \sin(x) = \frac{e^{ix} - e^{-ix}}{2i}.$$

By addition and composition, the cosine and sine functions are continuous on $\mathbf{C}$. Since $e^0 = 1$, we have
$$\cos(0) = 1, \qquad \sin(0) = 0.$$
Moreover, we see that
$$\cos(-x) = \cos(x), \qquad \sin(-x) = -\sin(x)$$
for $x \in \mathbf{C}$.

By definition, we obtain the other key formula
$$(4.3) \qquad\qquad\qquad e^{ix} = \cos(x) + i\sin(x).$$

By taking the n-th power of (4.3) for $n \in \mathbf{Z}$ and using (4.2), we also get
$$(4.4) \qquad \cos(nx) + i\sin(nx) = e^{inx} = (e^{ix})^n = (\cos(x) + i\sin(x))^n.$$

for any $x \in \mathbf{R}$.

Before checking that these are the usual trigonometric functions, we first compute the power series for these functions.

PROPOSITION 4.5.2. *For any $x \in \mathbf{C}$, we have*

$$\sin(x) = \sum_{n=0}^{+\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{6} + \frac{x^5}{120} + \cdots,$$

$$\cos(x) = \sum_{n=0}^{+\infty} \frac{(-1)^n}{(2n)!} x^{2n} = 1 - \frac{x^2}{2} + \frac{x^4}{24} + \cdots,$$

and both power series have infinite radius of convergence. In particular, the cosine and sine of a real number are both real, so that

$$\cos(x) = \mathrm{Re}(e^{ix}), \qquad \sin(x) = \mathrm{Im}(e^{ix}).$$

PROOF. By definition of the exponential, we have

$$e^{ix} = \sum_{n=0}^{+\infty} \frac{(ix)^n}{n!} = \sum_{n=0}^{+\infty} \frac{i^n x^n}{n!},$$

and since (i^n) is the repeating sequence

$$(1, i, -1, -i, 1, i, -1, -i, \ldots),$$

this becomes

$$e^{ix} = \sum_{n=0}^{+\infty} \frac{x^{4n}}{(4n)!} + i \sum_{n=0}^{+\infty} \frac{x^{4n+1}}{(4n+1)!)} - \sum_{n=0}^{+\infty} \frac{x^{4n+2}}{(4n+2)!)} - i \sum_{n=0}^{+\infty} \frac{x^{4n+3}}{(4n+3)!)}.$$

Similarly, replacing x by $-ix$, we get

$$e^{-ix} = \sum_{n=0}^{+\infty} \frac{x^{4n}}{(4n)!} - i \sum_{n=0}^{+\infty} \frac{x^{4n+1}}{(4n+1)!)} - \sum_{n=0}^{+\infty} \frac{x^{4n+2}}{(4n+2)!)} + i \sum_{n=0}^{+\infty} \frac{x^{4n+3}}{(4n+3)!)}.$$

Adding or subtracting e^{ix} and e^{-ix}, we get the result.

Since the power series converge for all $x \in \mathbf{C}$, we deduce that the radius of convergence is $+\infty$ (which one can also check directly). And the final part of the proposition is clear since the coefficients of the power series are real numbers. □

PROPOSITION 4.5.3. *For any $x \in \mathbf{R}$, we have $|e^{ix}| = 1$, and therefore*

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

In particular, for $x \in \mathbf{R}$, we have

$$|\sin(x)| \leqslant 1, \qquad |\cos(x)| \leqslant 1.$$

PROOF. For $x \in \mathbf{R}$, we have

$$|e^{ix}|^2 = e^{ix} \overline{e^{ix}} = e^{ix} e^{-ix} = 1$$

by Corollary 4.4.4, and then

$$\cos(x)^2 + \sin(x)^2 = \mathrm{Re}(e^{ix})^2 + \mathrm{Im}(e^{ix})^2 = |e^{ix}|^2 = 1.$$

□

REMARK 4.5.4. Let $x = a + ib$ be a complex number with $(a, b) \in \mathbf{R}^2$. We have

$$e^x = e^{a+ib} = e^a(\cos(b) + i \sin(b))$$

and in particular

$$|e^x| = e^a = e^{\mathrm{Re}(x)}.$$

EXAMPLE 4.5.5. The property (4.1) and the relations (4.3) and (4.2) lead to a systematic way of proving the usual trigonometric identities, such that that relating $\cos(x + y)$ to $\cos(x)$, $\cos(y)$, $\sin(x)$ and $\sin(y)$, or the formula for $\sin(kx)$ in terms of powers of $\cos(x)$ and $\sin(x)$, or the converse expressions for $\cos(x)^k$ in terms of $\cos(mx)$ and $\sin(nx)$. We illustrate the method with some basic examples of each – one should know the method, but not try to remember the formulas themselves!

(1) *How to compute* $\cos(x + y)$ *and* $\sin(x + y)$?

We can deal with both simultaneously (maybe faster than doing each individually) using (4.3) followed by (4.1):

$$\cos(x + y) + i\sin(x + y) = e^{i(x+y)} = e^{ix}e^{iy},$$

and then we use (4.3) again and multiply the complex numbers:

$$
\begin{aligned}
e^{ix}e^{iy} &= (\cos(x) + i\sin(x))(\cos(y) + i\sin(y)) \\
&= (\cos(x)\cos(y) - \sin(x)\sin(y)) + i(\sin(x)\cos(y) + \cos(x)\sin(y)).
\end{aligned}
$$

Looking at the real and imaginary parts, we get

$$(4.5) \qquad\qquad \cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y)$$
$$(4.6) \qquad\qquad \sin(x + y) = \sin(x)\cos(y) + \cos(x)\sin(y).$$

(2) *What is* $\sin(x)^4$ *in terms of cosine or sines of multiples of* x?

Here we use (4.2), following by the binomial expansion for $(a + b)^4$ and various applications of (4.1):

$$
\begin{aligned}
\sin(x)^4 &= \left(\frac{e^{ix} - e^{-ix}}{2i}\right)^4 \\
&= \frac{1}{16}\left(e^{4ix} - 4e^{2ix} + 6 - 4e^{-2ix} + e^{-4ix}\right).
\end{aligned}
$$

We then recombine and recognize that

$$\sin(x)^4 = \frac{1}{8}(\cos(4x) - 4\cos(2x) + 3).$$

(3) *What are* $\cos(5x)$ *and* $\sin(5x)$ *in terms of powers of* $\cos(x)$ *and* $\sin(x)$?

Here the basic formula is (4.4); for $k = 5$, using the binomial formula leads to

$$
\begin{aligned}
\cos(5x) + i\sin(5x) &= (\cos(x) + i\sin(x))^5 = \cos(x)^5 + 5i\cos(x)^4\sin(x) \\
&- 10\cos(x)^3\sin(x)^2 - 10i\cos(x)^2\sin(x)^3 + 5\cos(x)\sin(x)^4 + i\sin(x)^5.
\end{aligned}
$$

Identifying again the real and imaginary parts, we deduce that

$$
\begin{aligned}
\cos(5x) &= \cos(x)^5 - 10\cos(x)^3\sin(x)^2 + 5\cos(x)\sin(x)^4 \\
\sin(5x) &= 5\cos(x)^4\sin(x) - 10\cos(x)^2\sin(x)^3 + \sin(x)^5.
\end{aligned}
$$

Note that here also it is in fact faster to compute both $\cos(5x)$ and $\sin(5x)$ at the same time.

The next step in order to understand the trigonometric functions is to determine their exact image, and to prove that they are periodic (which leads to the definition of the number π).

> **LEMMA 4.5.6.** *There exists a real number $\pi \in [0, 4]$ such that $\cos(\pi/2) = 0$, and such that $\cos(x) > 0$ for $0 \leqslant x < \pi/2$. Moreover, we have $\sin(\pi/2) = 1$ and $\sin(x) > 0$ for $0 < x \leqslant \pi/2$.*

PROOF. We have $\cos(0) = 1 > 0$. We claim that $\cos(2) < 0$; using the intermediate value theorem, this implies the existence of at least one $x \in]0, 2]$ such that $\cos(x) = 0$. We then define

$$\pi = 2 \inf\{x \in [0, 2] \mid \cos(x) = 0\}.$$

This has the required property: because the infimum of a set of real numbers is the limit of a sequence (x_n) in that set (see Example 2.8.4, (2)), the continuity of the cosine leads to

$$\cos(x) = \lim_{n \to +\infty} \cos(x_n) = 0,$$

and there can be no smaller solution to the equation $\cos(x) = 0$.

To prove the claim, we use the fact that for $0 \leqslant x \leqslant 4$, we can write

$$\cos(x) - \left(1 - \frac{x^2}{2}\right) = \frac{x^4}{24} - \frac{x^6}{760} + \cdots$$

where the series on the right-hand side is alternating, as in Proposition 2.10.10, since (for $0 \leqslant x \leqslant 4$) the ratio between the absolute value of consecutive terms is $x^2/((2n+2)(2n+1))$, with $n \geqslant 2$, which is $\leqslant 16/30 < 1$.

It is then the case that the sum of the series is always located between the odd-index partial sums and the even-index partial sums. In particular, we get

$$1 - \frac{x^2}{2} \leqslant \cos(x) \leqslant 1 - \frac{x^2}{2} + \frac{x^4}{24}$$

for $x \geqslant 0$. For $x = 2$, this gives

$$\cos(x) \leqslant 1 - 2 + \frac{16}{24} = -\frac{1}{3} < 0.$$

We end by checking that $\sin(\pi/2) = 1$. Since $\cos(\pi/2)^2 + \sin(\pi/2)^2 = 1$, the only possibilities are that $\sin(\pi/2) = 1$ or $\sin(\pi/2) = -1$. But since we can also write

$$\sin(x) - \left(x - \frac{x^3}{6}\right) = \frac{x^5}{120} - \frac{x^7}{7!} + \cdots,$$

we have in the same way as above the inequality

$$\sin(x) \geqslant x - \frac{x^3}{6}$$

for $0 \leqslant x \leqslant 4$, which implies $\sin(x) > 0$ for $x^2 < 6$, in particular for all $x \leqslant 2$, including $x = \pi/2$. $\qquad \square$

> **THEOREM 4.5.7.** *For any $x \in \mathbf{R}$, we have*
> $$\cos(x + \pi) = -\cos(x), \qquad \sin(x + \pi) = -\sin(x)$$
> $$\cos(x + 2\pi) = \cos(x), \qquad \sin(x + 2\pi) = \sin(x)$$
> $$\cos(x + \tfrac{1}{2}\pi) = -\sin(x), \qquad \sin(x + \tfrac{1}{2}\pi) = \cos(x).$$

PROOF. By (4.5), we get

$$\cos(x + \tfrac{1}{2}\pi) = \cos(x)\cos(\tfrac{1}{2}\pi) - \sin(x)\sin(\tfrac{1}{2}\pi) = -\sin(x),$$

and

$$\sin(x + \tfrac{1}{2}\pi) = \sin(x)\cos(\tfrac{1}{2}\pi) + \cos(x)\sin(\tfrac{1}{2}\pi) = \cos(x).$$

For $x = \pi/2$, we deduce that $\cos(\pi) = -1$ and $\sin(\pi) = 0$. Then by the same formula we get
$$\cos(x + \pi) = \cos(x)\cos(\pi) - \sin(x)\sin(\pi) = -\cos(x),$$
and
$$\sin(x + \pi) = \sin(x)\cos(\pi) + \cos(x)\sin(\pi) = -\sin(x).$$
With $x = \pi$, this gives $\cos(2\pi) = 1$ and $\sin(2\pi) = 0$, and then once more, we get
$$\cos(x + 2\pi) = \cos(x)\cos(2\pi) - \sin(x)\sin(2\pi) = \cos(x),$$
and
$$\sin(x + 2\pi) = \sin(x)\cos(2\pi) + \cos(x)\sin(2\pi) = \sin(x).$$
$\square$

REMARK 4.5.8. We repeat the important formulas for cosine and sine of special numbers:
$$\cos(\tfrac{1}{2}\pi) = 0, \qquad \sin(\tfrac{1}{2}\pi) = 1$$
$$\cos(\pi) = -1, \qquad \sin(\pi) = 0$$
$$\cos(2\pi) = 1, \qquad \sin(2\pi) = 0.$$
Because of the periodicity, we get more generally for $k \in \mathbf{Z}$
$$\cos(k\pi) = (-1)^k, \qquad \sin(k\pi) = 0.$$
These formulas can also be remembered in their exponential form
$$e^{i\pi/2} = i, \qquad e^{i\pi} = -1, \qquad e^{2i\pi} = 1.$$

LEMMA 4.5.9. *For $\alpha \in \,]0, 2\pi[$, we have $e^{i\alpha} \neq 1$.*

PROOF. Consider the set X of $\alpha \in \,]0, 2\pi[$ such that $e^{i\alpha} = 1$. We want to prove that X is empty. We assume that it isn't and we will get a contradiction.

If X is not empty, then it has an infimum β, since X is bounded. We have $\beta > 0$: indeed, for $\alpha \in \,]0, \pi/2]$, we have $\sin(\alpha) = \text{Im}(e^{i\alpha}) > 0$ hence $\alpha \notin X$. Now β is the limit of a sequence $\alpha_n \in X$, and since the exponential is continuous, we get $e^{i\beta} = \lim e^{i\alpha_n} = 1$.

Note that $e^{i\beta/4}$ is equal to either 1, -1, i or $-i$ (since its fourth power is equal to 1, so the square is 1 or -1). In the first two cases, either $\beta/4$ or $\beta/2$ are elements of X smaller than β, which is impossible. If $e^{i\beta/4} = i$ or -1, then we get $\cos(\beta/4) = 0$, which contradicts the definition of π, since $\beta/4 < \pi/2$. This ends the proof of (1). $\square$

COROLLARY 4.5.10. *(1) For any pair $(a, b) \in \mathbf{R}^2$ such that $a^2 + b^2 = 1$, there exists a unique $\theta \in [0, 2\pi[$ such that*
$$e^{i\theta} = a + ib, \quad \text{or equivalently} \quad a = \cos(\theta) \text{ and } b = \sin(\theta).$$
In other words, the function from $[0, 2\pi[$ to $\mathbf{C}$ which maps θ to $e^{i\theta}$ is injective and its image is equal to the circle with center 0 and radius 1.
(2) The function cosine is injective and strictly decreasing on $[0, \pi]$, and the function sine is injective and strictly increasing on $[-\frac{1}{2}\pi, \frac{1}{2}\pi]$.

PROOF. We first show the existence of the real number θ. We may assume that $a \neq 1$ (in that case we can take $\theta = 0$). We have then $-1 \leqslant a < 1$. Because $\cos(0) = 1$ and $\cos(\pi) = -1$, there exists $\theta \in \,]0, \pi]$ such that $\cos(\theta) = a$. Then
$$\sin(\theta)^2 = 1 - a^2 = b^2,$$
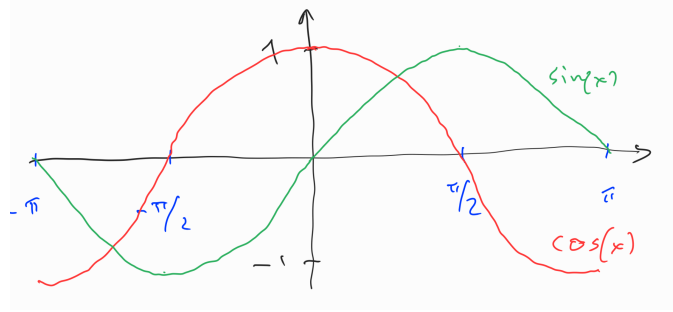
so either $\sin(\theta) = b$, in which case we are done, or $\sin(\theta) = -b$. In this second situation, we observe that

$$\cos(2\pi - \theta) = \cos(\theta) = a, \qquad \sin(2\pi - \theta) = -\sin(\theta) = b,$$

so that $2\pi - \theta \in [\pi, 2\pi[$ is then a solution.

We now prove that the solution is unique. Suppose that a pair of numbers (θ_1, θ_2) with $0 \leqslant \theta_1 \leqslant \theta_2 < 2\pi$ satisfy

$$e^{i\theta_1} = e^{i\theta_2}.$$

Then we get $e^{i(\theta_2 - \theta_1)} = 1$, with $\theta_2 - \theta_1 \in [0, 2\pi[$; by the previous lemma, this is only possible if $\theta_1 = \theta_2$.

We now prove (2). Suppose that x and y in $[0, \pi]$ satisfy $\cos(x) = \cos(y)$. Since $\cos(\pi - x) = -\cos(x)$ and $\cos(x) \geqslant 0$ for $0 \leqslant x \leqslant \pi/2$, either x and y are both $\leqslant \pi/2$, or both larger. We assume the first case (the second is similar). Then we have

$$\sin(x) \geqslant 0, \qquad \sin(y) \geqslant 0$$

by the last part of Lemma 4.5.6. This implies that

$$\sin(x) = \sqrt{1 - \cos(x)^2}, \qquad \sin(y) = \sqrt{1 - \cos(y)^2},$$

so that we get $e^{ix} = e^{iy}$, hence $x = y$ by (1).

So we have shown that cosine is injective on $[0, \pi]$; it must be strictly monotone by Proposition 3.4.1, and since $\cos(0) = 1$ and $\cos(\pi) = -1$, it must be strictly decreasing.

Using the formula $\sin(x) = -\cos(x + \pi/2)$ (see Theorem 4.5.7), we deduce that sine is strictly increasing on $[-\pi/2, \pi/2]$. $\qquad \square$

REMARK 4.5.11. The graphs of cosine and sine have therefore the following shape: Note however that to determine that the "slopes" of these curves at various points are roughly correct, we will have to wait for the definition of the derivative, which will allow us to know that $\sin(x)$ is "very close to x" when $|x|$ is small.

---

COROLLARY 4.5.12. *Let x be a complex number.*
(1) *There exists a real number $\theta \in [0, 2\pi[$ such that*

$$x = |x|e^{i\theta},$$

and θ is unique if x is not zero.
(2) *The image of the exponential function* $\exp\colon \mathbf{C} \to \mathbf{C}$ *is the set of non-zero complex numbers, and $e^x = e^y$ if and only if there exists $k \in \mathbf{Z}$ such that $x - y = 2k\pi$.*

PROOF. (1) If $x = 0$, we can take any θ. Otherwise, we note that $|x/|x|| = 1$ so the previous corollary implies that we can find a unique $\theta \in [0, 2\pi[$ with $\frac{x}{|x|} = e^{i\theta}$.
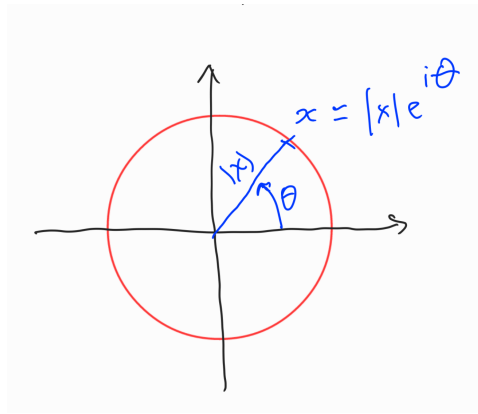
(2) If $x \neq 0$, then we get $x = |x|e^{i\theta} = e^{\log(|x|)}e^{i\theta} = e^{\log(|x|)+i\theta}$, so that x is in the image of the exponential. Since $e^x \neq 0$ for all x, the set of non-zero complex numbers is indeed the image of the exponential.

If $e^x = e^y$, then we get $e^{x-y} = 1$. Taking the modulus, we deduce that $e^{\mathrm{Re}(x-y)} = 1$, so that $\mathrm{Re}(x - y) = 0$. Hence there exists $\alpha \in \mathbf{R}$ such that $x - y = i\alpha$. If we subtract from α a suitable multiple $2k\pi$ of 2π, with $k \in \mathbf{Z}$, we can ensure that

$$0 \leqslant \alpha - 2k\pi < 2\pi,$$

and then $e^{i(\alpha-2k\pi)} = 1$ implies that $\alpha = 0$ by Lemma 4.5.9. $\qquad\square$

REMARK 4.5.13. (1) Geometrically, the number θ is the angle between the positive real-axis and the segment joining the origin to x. It is unique in $[0, 2\pi[$ (if $x \neq 0$), but it is not unique as a real number: $\theta + 2k\pi$ has the same property for all $k \in \mathbf{Z}$. We say that θ is the *argument* of x.

(2) The pair $(|x|, \theta)$ is called the *polar coordinates* of x. Note that it is not unique! This representation is convenient for multiplication: if x and y are complex numbers, with arguments θ and φ, then

$$xy = |xy|e^{i(\theta+\varphi)},$$

so the argument of the product is the sum of the argument of the factors (provided one allows the argument to exceed 2π).

Here is an important application.

PROPOSITION 4.5.14. *Let $k \in \mathbf{N}$. Let $y \in \mathbf{C}$ be a non-zero complex number. There are k roots x in $\mathbf{C}$ of the equation $x^k = y$, given by*

$$x_j = |y|^{1/k}e^{i\theta/k+2\pi ij/k}, \qquad 0 \leqslant j < k,$$

where $\theta \in [0, 2\pi[$ is the argument of y.

PROOF. For x_j as above, we have $x_j^k = |y|e^{i\theta+2\pi ij} = |y|e^{i\theta} = y$, so these numbers are solutions of the equations. They are distinct, because if we assume that $x_j = x_l$, with $j \leqslant l$, then we get
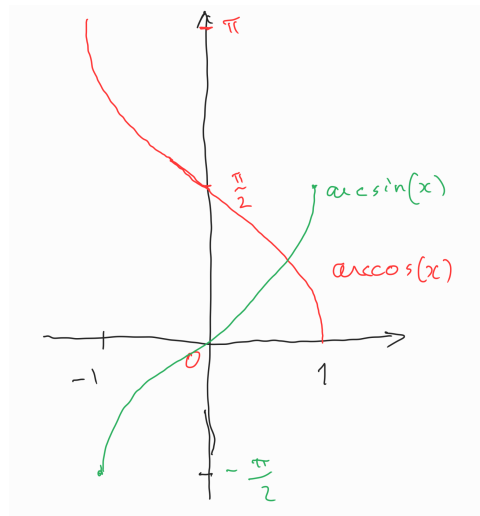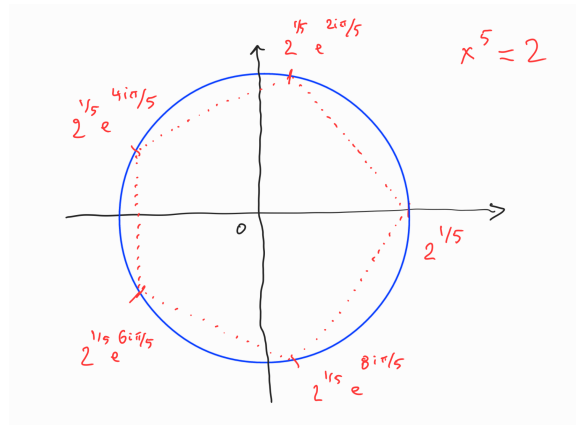
$$e^{i\theta/k+2\pi ij/k} = e^{i\theta/k+2\pi il/k}$$

or

$$e^{2\pi i(j-l)/k} = 1,$$

which cannot happen unless $j = l$, since $0 \leqslant 2\pi(j-l)/k < 2\pi$.

So we have found k distinct solutions of an algebraic equation of degree k, and these must be all the solutions (which is also easy to check directly in this case). $\qquad\square$

REMARK 4.5.15. Geometrically, the solutions x_j in the plane form the vertices of a regular polygon with k sides (e.g., a square if $k = 4$).

We end this chapter by defining the inverse trigonometric functions using Corollary 4.5.10.

> DEFINITION 4.5.16. The *inverse cosine function* is the reciprocal bijection arccos: $[-1, 1] \to [0, \pi]$ of the cosine function restricted to $[0, \pi]$.
>
> The *inverse sine function* is the reciprocal bijection arcsin: $[-1, 1] \to [-\pi/2, \pi/2]$ of the sine function restricted to $[-\pi/2, \pi/2]$.

These are indeed defined since cosine (resp. sine) is strictly decreasing (resp. strictly increasing) on $[0, \pi]$ (resp. on $[-\pi/2, \pi/2]$) with $\cos(0) = 1$ and $\cos(\pi) = -1$ (resp. $\sin(-\pi/2) = -1$ and $\sin(\pi/2) = 1$). The inverse cosine is strictly decreasing, and the inverse sine is strictly increasing.

EXAMPLE 4.5.17. (1) We have $\cos(\arccos(x)) = x$ and $\sin(\arcsin(x)) = x$ for all $x \in [-1, 1]$, but be careful that that $\arccos(\cos(x)) = x$ only holds if $x \in [0, \pi]$, although cosine is defined for all $x \in \mathbf{R}$.

(2) There is a close relationship between arcsin and arccos, as between cosine and sine. In fact, we have

$$\arccos(x) = \frac{\pi}{2} - \arcsin(x),$$

for all $x \in [-1, 1]$, since the number $\pi/2 - \arcsin(x)$ belongs to the interval $[0, \pi]$ for $x \in]-\pi/2, \pi/2[$ and the right-hand side satisfies

$$\cos(\tfrac{1}{2}\pi - \arcsin(x)) = \sin(\arcsin(x)) = x.$$

Moreover, we also have

$$\sin(\arccos(x)) = \sqrt{1 - x^2}, \qquad \cos(\arcsin(x)) = \sqrt{1 - x^2}$$

for all $x \in [-1, 1]$. Indeed, we have first

$$1 = \cos(\arccos(x))^2 + \sin(\arccos(x))^2 = x^2 + \sin(\arccos(x))^2,$$

so that $\sin(\arccos(x))$ can only be equal to $\sqrt{1 - x^2}$ or $-\sqrt{1 - x^2}$. Since $\arccos(x) \in [0, \pi]$, and since $\sin(y) \geqslant 0$ for $y \in [0, \pi]$ (by Lemma 4.5.6 for $0 \leqslant y \leqslant \pi/2$, and then since $\sin(\pi - y) = -\sin(y - \pi) = \sin(y + \pi) = \sin(y))$, this holds for $\pi/2 \leqslant y \leqslant \pi$), the only possibility is $\sqrt{1 - x^2}$.

Similarly, from

$$1 = \cos(\arcsin(x))^2 + \sin(\arcsin(x))^2 = \cos(\arcsin(x))^2 + x^2,$$

and from the fact that $\arcsin(x) \in [-\pi/2, \pi/2]$, where cosine is non-negative, we deduce that $\cos(\arcsin(x)) = \sqrt{1 - x^2}$.

CHAPTER 5

Differentiable functions

This chapter considers "better" regularity conditions on functions than their continuity, corresponding to the fact that, close a given point x_0, we can approximate very well a function f by the simplest non-constant function, that is, by a linear function $g(x) = ax + b$. We will only consider here functions defined on intervals in $\mathbf{R}$. The parameter a that appears is the *derivative* of f at the point x_0.

5.1. Definition and algebraic properties

Let I be an interval in $\mathbf{R}$ and $f\colon I \to \mathbf{R}$ a function. Let $x_0 \in I$. If we try to approximate "as best as possible" the function f by $g(x) = ax + b$, then it is natural to take b such that $g(x_0) = f(x_0)$. This leads to

$$g(x) = ax + f(x_0) - ax_0 = a(x - x_0) + f(x_0).$$

In order for $g(x)$ to be "close to" $f(x)$ when x is close to x_0, we need (with $\approx$ to denote a good approximation)

$$f(x) \approx f(x_0) + a(x - x_0),$$

which suggests that the constant a should be "close to"

$$\frac{f(x) - f(x_0)}{x - x_0}.$$

This suggests to take the limit of this last expression, if it exists, as defining a. This is geometrically quite natural, since this ratio is simply the slope of the line in the plane joining the two points

$$(x_0, f(x_0)), \qquad (x, f(x))$$

of the graph of f, so that to look at its limit means to look at points very close on the graph, and asking if the slope tends to a fixed number.
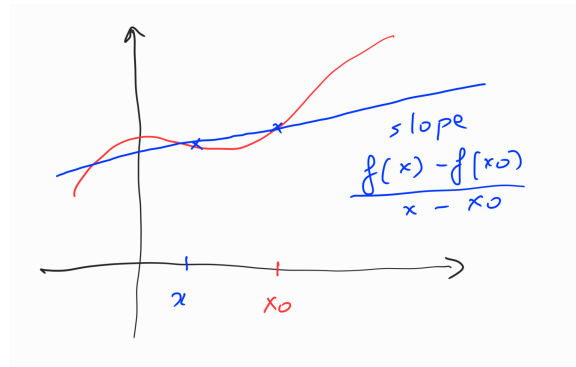
DEFINITION 5.1.1. Let $I \subset \mathbf{R}$ be an interval and $f \colon I \to \mathbf{R}$ a function. Let x_0 be an element of I.

If the limit

(5.1)
$$\lim_{\substack{x \to x_0 \\ x \neq x_0}} \frac{f(x) - f(x_0)}{x - x_0}$$

exists, then we say that f is *differentiable* at x_0 with *derivative* equal to the limit, which is denoted $f'(x_0)$.

If f is differentiable at all $x_0 \in I$, it is said to be differentiable on I, and the function f' on I that maps $x \in I$ to $f'(x)$ is called the *derivative* of f.

REMARK 5.1.2. (1) If x_0 is an endpoint of the interval I (say $I = [a, b]$ and $x_0 = a$), then the limit in the definition is understood to have x restricted to lie in I. Because this turns out to imply that the properties of the derivative at such a point are not the same as "interior" points, we sometimes emphasize this by saying that when x_0 is the minimum (resp. maximum) of I, the limit (when it exists) is the *right-derivative* of f at x_0, denoted $f'_r(x_0)$ (resp. the *left-derivative*, denoted $f'_l(x_0)$).

(2) If the limit (5.1) is $+\infty$ or $-\infty$, then one sometimes writes $f'(x_0) = +\infty$ or $f'(x_0) = -\infty$.

(3) It is often useful to write the limit (5.1) in the form

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f(x_0 + h) - f(x_0)}{h},$$

because the limit is then always as $h \to 0$.

The geometric interpretation is that the derivative (if it exists) is the slope of the *tangent line* of f at a point x (defined, intuitively, as a line in the plane that "just touches" the graph of f at the point $(x, f(x))$).

Note that since the definition is with a limit as $x \to x_0$, the existence (and the value) of the derivative are local properties of the function f, that only depend on its values on a small interval containing x_0.

DEFINITION 5.1.3. Let $I \subset \mathbf{R}$ be an interval and $f \colon I \to \mathbf{R}$ a function. Let x_0 be an element of I such that f is differentiable at x_0. The line with equation

$$y - y_0 = f'(x_0)(x - x_0),$$

which has slope $f'(x_0)$ and passes through $(x_0, f(x_0))$, is called the *tangent line to the graph of f* at $(x_0, f(x_0))$.

If $f'(x_0) = +\infty$ or $f'(x_0) = -\infty$, then the vertical line with equation

$$x = x_0$$

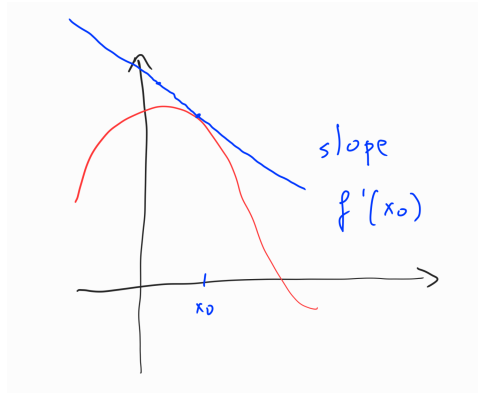is called the *tangent line to the graph of f* at $(x_0, f(x_0))$.

EXAMPLE 5.1.4. (1) Consider $I = \mathbf{R}$ and $f(x) = x$. Let $x_0 \in \mathbf{R}$ and $x \neq x_0$. Then we get

$$\frac{f(x) - f(x_0)}{x - x_0} = 1,$$

so that f is differentiable at x_0 with derivative constant equal to 1.

(2) Consider $I = \mathbf{R}$ and $f(x) = x^2$. Let $x_0 \in \mathbf{R}$ and $x \neq x_0$. Then we get

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{(x - x_0)(x + x_0)}{x - x_0} = x + x_0,$$

hence the limit as $x \to x_0$ exists and is equal to $2x_0$. Therefore the square function is differentiable on $\mathbf{R}$ with derivative $f'(x) = 2x$ for all x.

(3) Consider $I = \mathbf{R}_+$ and $g(x) = \sqrt{x}$. For $x_0 = 0$. Then for $x > 0$, we get

$$\frac{g(x) - g(x_0)}{x - x_0} = \frac{\sqrt{x}}{x} = \frac{1}{\sqrt{x}},$$

and therefore the derivative on the right of g is equal to $+\infty$. This corresponds to the fact that the slope of the graph of g around 0 is vertical.

(4) Altough the most commonly used functions are differentiable, this is not always the case. For instance, let $I = \mathbf{R}$ and $f(x) = |x|$. Then if we take $x_0 = 0$, then the ratio

$$\frac{f(x) - f(0)}{x} = \frac{|x|}{x}$$

is equal to 1 if $x < 0$ and to -1 if $x < 0$, which means that there is no derivative at $x_0 = 0$.

In fact, one can show that in a certain precise sense "almost all" continuous functions are *not* differentiable at any point (this is a consequence of the theory of *Brownian motion*). A concrete example (not obvious; it was proposed by Weierstrass in 1872, but only proved to be suitable by Hardy in 1916) is the function defined by

$$f(x) = \sum_{n=1}^{+\infty} \frac{\sin(3^n x)}{3^n}.$$

The following fact is often useful, even as a way to prove continuity.

PROPOSITION 5.1.5. *Let $I \subset \mathbf{R}$ be an interval and let $f \colon I \to \mathbf{R}$ be a function differentiable on I. Then f is continuous on I.*

PROOF. Let $x_0 \in I$. Define the function

$$g(x) = \frac{f(x) - f(x_0)}{x - x_0}$$

for $x \neq x_0$. From the existence of the limit (5.1), we deduce that there exists $\delta > 0$ such that

$$|g(x)| \leqslant |f'(x_0)| + 1$$

when $|x - x_0| < \delta$ and $x \neq x_0$. Define $g(x_0) = 1$, so that the inequality is also true for $x = x_0$. Since

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

for all $x \in I$, including $x = x_0$, we deduce that for $|x - x_0| < \delta$, we have

$$|f(x) - f(x_0)| \leqslant M|x - x_0|$$

101

where $M = 1 + |f'(x_0)|$. Using Lemma 3.2.4 (or arguing directly), we deduce that f is continuous at x_0. $\qquad\square$

As is the case of continuity, the property of differentiability is preserved under the usual algebraic operations, and moreover there are easy formulas to compute the corresponding derivatives.

PROPOSITION 5.1.6. *Let $I \subset \mathbf{R}$ be an interval and let f, $g\colon I \to \mathbf{R}$ be functions differentiable on I.*

(1) The function $f + g$ is differentiable on I and $(f + g)' = f' + g'$.

(2)(Leibniz rule) The function fg is differentiable on I and $(fg)' = f'g + g'f$. In particular, if $f = a$ is a constant function, then $(ag)' = ag'$.

(3) If $g(x) \neq 0$ for all $x \in I$, then the function f/g is differentiable on I and $(f/g)' = (f'g - fg')/g^2$.

(4) (Chain rule) If J is an interval containing the image of f and f_1 is a differentiable function $J \to \mathbf{R}$, the the composition $f_1 \circ f$ is differentiable on I and

$$(f_1 \circ f)' = f'(f_1' \circ f),$$

or in other words

$$(f_1 \circ f)'(x) = f'(x)f_1'(f(x))$$

for all $x \in I$.

(5) (Reciprocal) If f is injective and J is its image, and if f' is non-zero on I, then the reciprocal bijection function $f^{-1}\colon J \to I$ is differentiable with

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

PROOF. (1) We have

$$\frac{(f + g)(x) - (f + g)(x_0)}{x - x_0} = \frac{f(x) - f(x_0)}{x - x_0} + \frac{g(x) - g(x_0)}{x - x_0},$$

and the right-hand side converges to $f'(x_0) + g'(x_0)$ as $x \to x_0$.

(2) We have

$$\frac{(fg)(x) - (fg)(x_0)}{x - x_0} = g(x)\frac{f(x) - f(x_0)}{x - x_0} + f(x_0)\frac{g(x) - g(x_0)}{x - x_0},$$

and the right-hand side converges to $g(x_0)f'(x_0) + f(x_0)g'(x_0)$ as $x \to x_0$, since g is continuous at x_0, so that $g(x) \to g(x_0)$ as $x \to x_0$ (Proposition 5.1.5). If $f = a$ is constant, then we get $(af)'(x_0) = af'(x_0)$.

(3) Using (2), it suffices to prove that the derivative of $1/g$ is $-g'/g^2$. We have

$$\frac{1/g(x) - 1/g(x_0)}{x - x_0} = \frac{1}{g(x)g(x_0)}\frac{g(x_0) - g(x)}{x - x_0},$$

which converges to $-g'(x_0)/g(x_0)^2$ (again because g is continuous).

(4) Formally, we write

$$\frac{f_1 \circ f(x) - f_1 \circ f(x_0)}{x - x_0} = \frac{f_1 \circ f(x) - f_1 \circ f(x_0)}{f(x) - f(x_0)}\frac{f(x) - f(x_0)}{x - x_0}.$$

Since f is continuous, we have $f(x) \to f(x_0)$ as $x \to x_0$, which implies that

$$\frac{f_1 \circ f(x) - f_1 \circ f(x_0)}{f(x) - f(x_0)} \to f_1'(f(x_0)).$$

The second factor converges to $f'(x_0)$, which gives the result. (This is not quite a rigorous proof, because it could be that $f(x) = f(x_0)$ for many values of x when it approaches x_0, so that the expression above is not well-defined; however, one can use note that we always have

$$\frac{f_1 \circ f(x) - f_1 \circ f(x_0)}{x - x_0} = h(x)\frac{f(x) - f(x_0)}{x - x_0}$$

where we put $h(x) = f_1'(f(x_0))$ if $f(x) = f(x_0)$, and

$$h(x) = \frac{f_1 \circ f(x) - f_1 \circ f(x_0)}{f(x) - f(x_0)}$$

otherwise; it is then easy to deduce, using sequences, that $h(x) \to f_1'(f(x_0))$ as $x \to x_0$.)

(5) Let $y_0 \in J$, For all $y \in J$ different from y_0, we define have

$$\frac{f^{-1}(y) - f^{-1}(y_0)}{y - y_0} = \frac{f^{-1}(y) - f^{-1}(y_0)}{f(f^{-1}(y)) - f(f^{-1}(y_0))}.$$

As $y \to y_0$, we have $f^{-1}(y) \to f^{-1}(y_0)$ since f^{-1} is continuous, hence this converges to $1/f'(f^{-1}(y))$. $\qquad\square$

We now prove that almost all elementary functions are differentiable, which implies that all finite combinations of such functions are differentiable where they are defined (and shows that their derivatives can be efficiently computed).

PROPOSITION 5.1.7. (1) *Any polymomial*

$$p(x) = a_n x^n + \cdots + a_1 x + a_0$$

is differentiable on $\mathbf{R}$ *with*

$$p'(x) = na_n x^{n-1} + (n-1)a_{n-1} x^{n-2} + \cdots + a_1.$$

(2) *The exponential and trigonometric functions are differentiable on* $\mathbf{R}$ *with*

$$\exp' = \exp, \qquad \cos' = -\sin, \qquad \sin' = \cos,$$

(3) *The logarithm is differentiable on* $]0, +\infty[$ *with*

$$\log'(x) = \frac{1}{x}$$

for all $x > 0$.

(4) *For any* $a \in \mathbf{R}$, *the function* $f(x) = x^a$ *is differentiable on* $]0, +\infty[$ *with* $f'(x) = ax^{a-1}$.

(5) *The functions* arccos *and* arcsin *are differentiable on* $]-1, 1[$ *with derivatives*

$$\arccos'(x) = -\frac{1}{\sqrt{1-x^2}}, \qquad \arcsin'(x) = \frac{1}{\sqrt{1-x^2}}.$$

PROOF. (1) Using Proposition 5.1.6, it is enough to consider the case of $p_n(x) = x^n$ for $n \in \mathbf{N}_0$, for which we need to prove that

$$p_n'(x) = nx^{n-1} \text{ for } n \in \mathbf{N}, \qquad p_0'(x) = 0.$$

This can be done by induction on n. If $n = 0$, then $p_0 = 1$ is constant and has derivative equal to 0. If $n = 1$, then $p_1(x) = x$ has derivative 1 by Example 5.1.4, (1). Now assume that $n \in \mathbf{N}$ and that the result holds for p_n and consider p_{n+1}. We have $p_{n+1} = p_n p_1$, so that by the Leibniz rule, the function p_{n+1} is differentiable and

$$p_{n+1}'(x) = (p_n p_1)'(x) = p_n'(x)p_1(x) + p_n(x)p_1'(x) = nx^{n-1} \times x + x^n \times 1 = (n+1)x^n,$$

which concludes the proof.

(2) The derivatives of exp, sin and cos can be deduced from their power series expansions; for example, treating the power series as if it were a polynomial and using (1), we would get

$$\exp'(x) = 1 + 2\frac{x}{2} + 3\frac{x^2}{6} + \cdots + n\frac{x^{n-1}}{n!} + \cdots = \exp(x).$$

This can be justified, as we will explain later, but we can also prove this in a more elementary manner using the fundamental property (4.1). Precisely, let $x_0 \in \mathbf{R}$, and for $h \in \mathbf{R}$, write first

$$\frac{e^{x_0+h} - e^{x_0}}{h} = e^{x_0}\frac{e^h - 1}{h}.$$

As $h \to 0$, this converges if and only if $(e^h - 1)/h$ converges, and the limit is then e^{x_0} multiplied by that limit (this reduces the question to the differentiability at 0).

For $h \neq 0$, we get using the power series the formula

$$\frac{e^h - 1}{h} = 1 + hg(h),$$

where

$$g(h) = \frac{1}{2} + \cdots + \frac{h^n}{(n+2)!} + \cdots$$

Observe that the coefficients of this series satisfy $|h^n/(n+2)!| \leqslant h^n$ for $n \in \mathbf{N}_0$, so that if $-1 < h < 1$, we have

$$|g(h)| \leqslant \sum_{n=0}^{+\infty} |h|^n = \frac{1}{1 - |h|}.$$

It follows that

$$\left|\frac{e^h - 1}{h} - 1\right| \leqslant \frac{|h|}{1 - |h|},$$

and since the right-hand side tends to 0 as $h \to 0$, we conclude that $(e^h - 1)/h \to 1$ as $h \to 0$, which leads to the formula for the derivative of the exponential.

For cosine and sine, we use the power series, which will be justified later on (see Example 5.2.3): differentiating term by term from

$$\cos(x) = 1 - \frac{x^2}{2} + \cdots + \frac{(-1)^n x^{2n}}{(2n)!} + \cdots$$

as for a polynomial, we get

$$\cos'(x) = -x + \cdots + (-1)^n\frac{x^{2n-1}}{(2n-1)!} + \cdots = -\sin(x),$$

and similarly with the sine.

(3) By definition, the logarithm is the reciprocal bijection of the exponential function on $\mathbf{R}$; since $\exp' = \exp$ which is never zero, so we get

$$\log'(x) = \frac{1}{\exp'(\log(x))} = \frac{1}{\exp(\log(x))} = \frac{1}{x}$$

for all $x > 0$ by Proposition 5.1.6, (5).

(4) We have by definition

$$x^a = e^{a\log(x)} = f(g(x))$$

where $f(x) = \exp(x)$ and $g(x) = a\log(x)$. By the Chain Rule, it follows that the power function is differentiable, and that the value of its derivative at $x_0 > 0$ is

$$g'(x_0)f'(g(x_0)) = \frac{a}{x_0}e^{a\log(x_0)} = ax_0^{a-1}$$

by Proposition 4.4.7, (3).

(5) The function arccos is the reciprocal of the function $\cos\colon [0,\pi] \to [-1,1]$, which has derivative $-\sin(x)$. We have $\sin(x) > 0$ on $]0,\pi[$. By Proposition 5.1.6, (5) again, we get

$$\arccos'(x) = -\frac{1}{\sin(\arccos(x))}.$$

for $x \in]-1,1[$ (which is the image of cos when restricted to $]0,\pi[$), and by Example 4.5.17, (2), we get

$$\arccos'(x) = -\frac{1}{\sqrt{1-x^2}}.$$

The case of arcsin is similar . $\qquad\square$

EXAMPLE 5.1.8. Let $I = [1,+\infty[$ and

$$f(x) = \log(x + \sqrt{x^2-1})$$

for $x \in I$. By composition and addition of differentiable functions (and the fact that $x^2 - 1 \geqslant 0$ and $x + \sqrt{x^2-1} > 0$ for $x \in I$), it follows that f is differentiable. We compute its derivative using the chain rule: putting $g(x) = x + \sqrt{x^2-1}$, which satisfies

$$g'(x) = 1 + \frac{2x}{2\sqrt{x^2-1}} = 1 + \frac{x}{\sqrt{x^2-1}},$$

we get

$$f'(x) = \frac{g'(x)}{x+\sqrt{x^2-1}} = \left(1 + \frac{x}{\sqrt{x^2-1}}\right)\frac{1}{x+\sqrt{x^2-1}}.$$

Since derivatives are defined as limits, they can often be used themselves to compute various limits that can be represented in terms of limits of $(f(x) - f(x_0))/(x - x_0)$. Here is a standard example:

PROPOSITION 5.1.9 (L'Hospital's Rule). *Let f and g be functions defined and differentiable on $[a,b]$ with $a < b$. Suppose that $f(a) = g(a) = 0$. If $g'(a) \neq 0$, then*

$$\lim_{\substack{x\to a\\x<a}} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}.$$

Similarly if $f(b) = g(b) = 0$ and $g'(b) \neq 0$, we have

$$\lim_{\substack{x\to b\\x<b}} \frac{f(x)}{g(x)} = \frac{f'(b)}{g'(b)}.$$

PROOF. For $x > a$, we note that

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

where the first factor converges to $f'(a)$ and the second to $1/g'(a)$. $\qquad\square$

EXAMPLE 5.1.10. Consider
$$\lim_{x \to 0} \frac{\cos(x) - 1}{\sin(x)}.$$
We can apply the L'Hospital Rule to compute both the limit as $x \to 0$ with $x > 0$ or $x \to 0$ with $x < 0$. We find
$$\lim_{\substack{x \to 0 \\ x > 0}} \frac{\cos(x) - 1}{\sin(x)} = \frac{-\sin(0)}{\cos(0)} = 0,$$
and the same value for the limit with $x < 0$. We conlude that the limit exists and is equal to 0.

There exist generalizations of the L'Hospital Rule, but in many cases it is better to use Taylor polynomials to compute such limits (see Example 5.7.10).

5.2. Derivative of functions defined as limits

Let $I \subset \mathbf{R}$ be an interval of real numbers. If we have a sequence of functions $f_n \colon I \to \mathbf{R}$ that converges to a function f, and if all functions f_n are differentiable, it is natural to ask whether the limit f is also differentiable (and if Yes, one might hope that the derivative of f is also the limit of the derivatives f_n'). Clearly, we need to assume at least that the convergence is uniform, since otherwise the limit f might not even be continuous, which is a necessary condition to be differentiable. But it is possible to find examples where the uniform convergence is not sufficient; in fact, more generally, it is a result of Weierstrass that *any* continuous function $f \colon [0,1] \to \mathbf{R}$, possibly nowhere differentiable, can be expressed as the uniform limit of a sequence of polynomials (f_n), which are of course all differentiable functions.

REMARK 5.2.1. Bernstein found a particulary simple example of such a sequence of polynomials: for any continuous function $f \colon [0,1] \to \mathbf{R}$, he proved that the functions
$$f_n(x) = \sum_{k=0}^n \binom{n}{k} x^k (1-x)^{n-k} f\left(\frac{k}{n}\right),$$
which are polynomials of degree at most n, converge uniformly to f on $[0,1]$ (this example is actually very important in applications, in the theory of *splines* for computer-aided design for instance).

However, we can obtain the result if we assume also that the sequence of derivatives f_n' converges itself uniformly.

THEOREM 5.2.2. *Let I be an interval in $\mathbf{R}$ and for $n \in \mathbf{N}$, let $f_n \colon I \to \mathbf{R}$ be a differentiable function. Let $f \colon I \to \mathbf{R}$ be a function. Assume that f_n' is continuous for all n, and moreover that (f_n) converges to f uniformly on I, and that (f_n') converges uniformly to some function g.*
  Then the function f is differentiable on I, and $f' = g$, which is a continuous function.

We will prove this theorem in the next chapter, since it is easiest to do it using the properties of the integral of functions.

EXAMPLE 5.2.3. (1) Let $(a_n)_{n \in \mathbf{N}_0}$ be a sequence of real numbers such that the power series $\sum a_n x^n$ has a positive radius of convergence $R > 0$. Let f be the continuous function on $]-R, R[$ defined by the sum of the series.

We claim that we can apply Theorem 5.2.2 to the sequence of partial sums, and conclude that the sum f of the power series is differentiable on $]-R, R[$ and satisfies

$$f'(x) = a_1 + 2a_2 x + \cdots + na_n x^{n-1} + \cdots = \sum_{n=0}^{+\infty} (n+1)a_{n+1}x^n.$$

Precisely, we need as usual to first restrict the domain. The partial sums are the polynomials

$$s_n(x) = a_0 + a_1 x + \cdots + a_n x^n,$$

which are therefore differentiable with a continuous derivative, given by the polynomials

$$s_n'(x) = a_1 + \cdots + na_n x^{n-1}.$$

These are the partial sums of the power series

$$\sum_{n=0}^{+\infty} (n+1)a_{n+1}x^n.$$

This power series has the same radius of convergence as the original series, by Lemma 4.3.7 applied to $k = 1$, and therefore it converges uniformly to some function g on any interval $]-r, r[$ where $r < R$. By the theorem, we deduce that f is differentiable on $]-r, r[$ with

$$f'(x) = \sum_{n=0}^{+\infty} (n+1)a_{n+1}x^n.$$

for $-r < x < r$. Taking a sequence of r that tends to R (for instance $r_n = R - 1/n$), we conclude (because the differentiability is a local property) that for any x such that $|x| < R$, the function f has a derivative at x given be the sume of the series above.

(2) If we apply (1) to cases where we have a concrete expression for the sum of the power series, then we obtain new identities. For instance, from the geometric series expansion

$$\sum_{n=0}^{+\infty} x^n = \frac{1}{1-x},$$

we obtain

$$\sum_{n=0}^{+\infty} (n+1)x^n = \frac{1}{(1-x)^2},$$

which implies also (by multiplying by x on both sides and renumbering) that

$$x + 2x^2 + 3x^3 + \cdots = \sum_{n=1}^{+\infty} nx^n = \frac{x}{(1-x)^2}.$$

If we differentiate again, we get

$$1 + 4x + 9x^2 + \cdots = \sum_{n=0}^{+\infty} (n+1)^2 x^n = \frac{1+x}{(1-x)^3},$$

since the derivative of $x/(1-x)^2$ is equal to

$$\frac{(1-x)^2 - x(2x-2)}{(1-x)^4} = \frac{1-x^2}{(1-x)^4} = \frac{1+x}{(1-x)^3}.$$

Differentiating further, one can in principle get such formulas for

$$\sum_{n=0}^{+\infty} n^k x^n$$

for any $k \in \mathbf{N}_0$.

5.3. Derivatives of complex-valued functions

We add a few remarks on the generalization of the notion of derivative to functions $f \colon I \to \mathbf{C}$, where I is still an interval in $\mathbf{R}$.

This can be done by using the same definition:

$$f'(x_0) = \lim_{\substack{x \to x_0 \\ x \neq x_0}} \frac{f(x) - f(x_0)}{x - x_0}$$

where the right-hand side is a limit of quantities in $\mathbf{C}$. Alternatively, one can write

$$f = f_1 + if_2$$

where $f_1(x) = \mathrm{Re}(f(x))$ and $f_2(x) = \mathrm{Im}(f(x))$, and then (using Proposition 2.5.11 for instance) it follows that f is differentiable at x_0 with derivative $f'(x_0)$ if and only if both f_1 and f_2 are differentiable at x_0 with

$$f'(x_0) = f_1'(x_0) + if_2'(x_0).$$

All the properties we have seen (Proposition 5.1.6 and Theorem 5.2.2, in particular) hold for functions from I to $\mathbf{C}$.

EXAMPLE 5.3.1. Let $f(x) = e^{ax}$ for $x \in \mathbf{R}$, where $a \in \mathbf{C}$. If a is not a real number (for instance if $a = i$), then f is really a function from $\mathbf{R}$ to $\mathbf{C}$. It is differentiable on $\mathbf{R}$ with

$$f'(x) = ae^{ax}$$

for all $x \in \mathbf{R}$.

We can check this using the power series expansion and Theorem 5.2.2, or using the second interpretation with real and imaginary parts. Indeed, write $a = \alpha + i\beta$ with $\alpha = \mathrm{Re}(a)$ and $\beta = \mathrm{Im}(a)$. We have

$$f(x) = e^{\alpha x} \cos(\beta x) + ie^{\alpha x} \sin(\beta x).$$

The real and imaginary parts f_1 and $f-2$ are differentiable as products of differentiable functions, so the function f is differentiable. Using the Leibniz rule and Proposition 5.1.7, we compute

$$f_1'(x) = \alpha e^{\alpha x} \cos(\beta x) - \beta e^{\alpha x} \sin(\beta x)$$
$$f_2'(x) = \alpha e^{\alpha x} \sin(\beta x) + \beta e^{\alpha x} \cos(\beta x).$$

Hence

$$\begin{aligned}
f'(x) &= f_1'(x) + if_2'(x) \\
&= \Big(\alpha e^{\alpha x} \cos(\beta x) - \beta e^{\alpha x} \sin(\beta x)\Big) + i\Big(\alpha e^{\alpha x} \sin(\beta x) + \beta e^{\alpha x} \cos(\beta x)\Big) \\
&= \alpha e^{\alpha x}(\cos(\beta x) + i\sin(\beta x)) + i\beta e^{\alpha x}(\cos(\beta x) + i\sin(\beta x)) = ae^{ax},
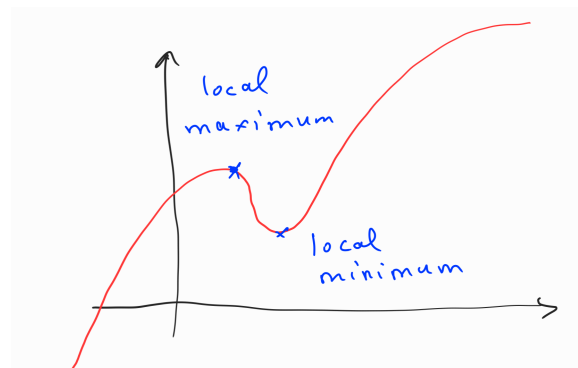\end{aligned}$$

as claimed.

FIGURE 5.1. Local Maximum and Local Minimum

REMARK 5.3.2. One can also ask whether it makes sense to speak of the derivative of a function $f\colon \mathbf{C} \to \mathbf{C}$, since the ratios

$$\frac{f(x) - f(x_0)}{x - x_0}$$

make sense as complex numbers for $x \neq x_0$ (even if there is no interpretation as the slope of a line).

Such a definition does make sense, but it turns out that, except for very simple formal properties, the functions $f\colon \mathbf{C} \to \mathbf{C}$ that are differentiable have very different properties as functions of a real variable. The most striking is probably the following: any function $f\colon \mathbf{C} \to \mathbf{C}$ that is differentiable everywhere in the complex sense can be represented as a *power series with infinite radius of convergence.*

5.4. Global properties of differentiable functions

As was the case for continuity, the power of having differentiable functions is revealed in the global properties of the derivative. The two important statements are somewhat related to the intermediate value theorem and the extremum theorem. Before we state them, we need to generalize the notion of the maximum or minimum of a function.

DEFINITION 5.4.1 (Local extremum). Let $I \subset \mathbf{R}$ be an interval and $f\colon I \to \mathbf{R}$ a function. Let $x_0 \in I$.
(1) The function f has a *local maximum* at x_0 if there exists $\delta > 0$ such that
$$f(x) \leqslant f(x_0)$$
for $x \in I$ such that $|x - x_0| < \delta$.
(2) The function f has a *local minimum* at x_0 if there exists $\delta > 0$ such that
$$f(x) \geqslant f(x_0)$$
for $x \in I$ such that $|x - x_0| < \delta$.
(3) If f has either a local maximum or a local minimum at x_0, then we say that f has a *local extremum* at x_0.

REMARK 5.4.2. In other words, in comparison with a maximum or a minimum of all values of a function (which we often call a *global* maximum or minimum, to avoid any ambiguity), we only ask that, at a local maximum x_0 for instance, the graph of the function is below the horizontal line $y = f(x_0)$ when x is not too far from x_0. Clearly, if f has a maximum at x_0, then it has a local maximum. An example of a local minimum
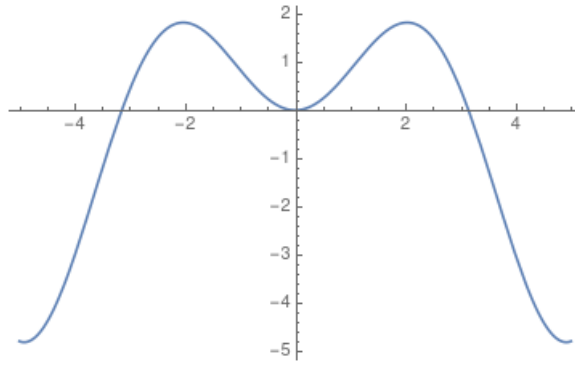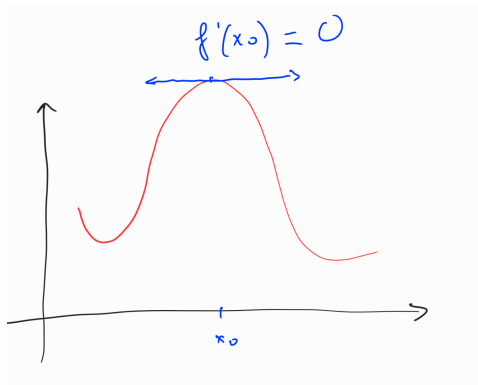
FIGURE 5.2. Graph of $f(x) = x\sin(x)$

FIGURE 5.3. Local Extremum Theorem

that is not a global minimum is given by the function

$$f(x) = x\sin(x)$$

at $x_0 = 0$. Indeed, since $\sin(-x) = -\sin(x)$ and $\sin(x) \geqslant 0$ for $0 \leqslant x \leqslant \pi/2$, we have

$$f(x) \geqslant 0 = f(0)$$

for $-\pi/2 \leqslant x \leqslant \pi/2$, which shows that 0 is a local minimum. It is not a global minimum because f is sometimes negative, for instance $f(3\pi/2) = -3\pi/2$ since $\sin(3\pi/2) = -1$.

THEOREM 5.4.3 (Local Extremum Theorem). *Let $I \subset \mathbf{R}$ be an interval and $f \colon I \to \mathbf{R}$ a differentiable function. Let $x_0 \in I$ be such that f has a local extremum at x_0, and such that x_0 is not either the minimum or maximum of I. Then $f'(x_0) = 0$.*
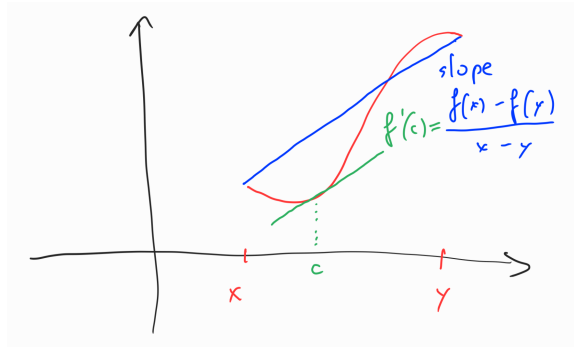
PROOF. Since x_0 is not the minimum or maximum of I, there exists $\delta > 0$ such that the interval $]x_0 - \delta, x_0 + \delta[$ is contained in I.

Assume that x_0 is a local maximum, the other case being similar. We may then assume that δ is small enough to satisfy

$$f(x) \leqslant f(x_0)$$

for $x \in]x_0 - \delta, x_0 + \delta[$. It follows that

$$\frac{f(x_0 + h) - f(x_0)}{h} \leqslant 0$$

for $0 < h < \delta$ and
$$\frac{f(x_0 + h) - f(x_0)}{h} \geqslant 0$$
for $-\delta < h < 0$. From the first, letting $h \to 0$, we deduce that $f'(x_0) \leqslant 0$, and from the second, that $f'(x_0) \geqslant 0$. Hence $f'(x_0) = 0$. $\qquad\square$

REMARK 5.4.4. (1) The statement means that, at a local extremum that is not an endpoint of I, the tangent line to the graph of f is horizontal, which is something that is intuitively clear.

(2) The theorem does not always hold for the endpoints. For instance, if $I = [0, 1]$ and $f(x) = x$, then $x_0 = 0$ is a local minimum (even global minimum), but $f'(0) = 1$.

(3) The condition that $f'(x_0) = 0$ is necessary, but *not sufficient*, to ensure that x_0 is a local extremum. For instance, let $I = \mathbf{R}$ and $f(x) = x^3$. Then $f'(x) = 3x^2$, so that $f'(0) = 0$, but $x_0 = 0$ is not a local extremum since $f(x) < 0$ for $x < 0$ and $f(x) > 0$ for $x > 0$. In the next section, we will see how one can often determine if a zero of the derivative really corresponds to a local extremum.

(4) Let $f \colon [a, b] \to \mathbf{R}$ be continuous. We know, from the Extremum Theorem, that there exists points x_0 and x_1 such that
$$f(x_0) \leqslant f(x) \leqslant f(x_1)$$
for all $x \in [a, b]$. The Local Extremum Theorem provides an approach to finding the possible values of x_0 and x_1, if the function f is in addition differentiable on $[a, b]$:

- Find the set X of all solutions x of the equation $f'(x) = 0$.
- Evaluate f at the endpoints (a and b) and at all points of X; then the possible values of x_0 (resp. x_1) are those $x \in X \cup \{a, b\}$ where $f(x_0)$ is the smallest (resp. the largest).

This approach is often successful, especially when the set X is finite.

THEOREM 5.4.5 (Mean-value Theorem). *Let $I \subset \mathbf{R}$ be an interval and $f \colon I \to \mathbf{R}$ a differentiable function. Let $a < b$ be elements of I. There exists a real number $c \in\,]a, b[$ such that*
$$\frac{f(b) - f(a)}{b - a} = f'(c).$$

PROOF. Let
$$g(x) = \frac{f(b) - f(a)}{b - a}(x - a) + f(a),$$
which is the linear function such that the line in the plane joining $(a, f(a))$ to $(b, f(b))$ has equation $y = g(x)$ (because $g(a) = f(a)$ and $g(b) = f(b)$).

Consider the differentiable function $h(x) = f(x) - g(x)$. It satisfies $h(a) = h(b) = 0$. If the function h is everywhere equal to 0, then $f'(x) = g'(x) = (f(b) - f(a))/(b - a)$ for all $x \in]a, b[$, and we are done.

If h is not the zero function, then either its maximum or its minimum (which exist because h is continuous) is non-zero. Assume that $\max h(x) \neq 0$; let then $c \in [a, b]$ be such that

$$h(c) = \max_{x \in [a,b]} h(x).$$

Since $h(c) \neq 0$ by assumption, we deduce that $c \notin \{a, b\}$, and by the Local Extremum Theorem, we conclude that $h'(c) = 0$, which leads to

$$f'(c) - \frac{f(b) - f(a)}{b - a} = 0.$$

The case of a non-zero minimum is similar. $\qquad\square$

REMARK 5.4.6. (1) Sometimes we apply the mean-value theorem to two elements a and b of the interval I without knowing if $a < b$ or $a > b$. The conclusion is however unchanged, if one states that c is between a and b. Indeed, if $b < a$, we have

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

(the slope of the line between two points does not depend on the order of the two points).

(2) It is essential for the Mean Value Theorem that f be differentiable everywhere. Even a single point where it is not can lead to failure! For instance, let $I = \mathbf{R}$ and $f(x) = |x|$, which is differentiable everywhere except at $x = 0$. If we take $a = -1$ and $b = 1$, then the slope $(f(b) - f(a))/(b - a)$ is zero, which is not the value of $f'(x)$ wherever it exists (since $f'(x) = -1$ if $x < 0$ and $f'(x) = 1$ if $x > 0$).

COROLLARY 5.4.7. *Let $I \subset \mathbf{R}$ be an interval and $f \colon I \to \mathbf{R}$ a differentiable function.*
(1) *The function f is non-decreasing if and only if $f'(x) \geqslant 0$ for all $x \in I$.*
(2) *If $f'(x) > 0$ for all $x \in I$, then f is strictly increasing.*
(3) *The function f is non-increasing if and only if $f'(x) \leqslant 0$ for all $x \in I$.*
(4) *If $f'(x) < 0$ for all $x \in I$, then f is strictly decreasing.*

PROOF. (1) If f is non-decreasing on I, then for any x and $x_0 \in I$, with $x \neq x_0$, we get

$$\frac{f(x) - f(x_0)}{x - x_0} \geqslant 0,$$

and letting $x \to x_0$, we deduce that $f'(x_0) \geqslant 0$ for all $x_0 \in I$.

Conversely, suppose that $f'(x_0) \geqslant 0$ for all $x_0 \in I$. Then if $x < y$ are elements of I, the Mean-Value Theorem implies that there exists $x_0 \in]x, y[$ such that

$$\frac{f(y) - f(x)}{y - x} = f'(x_0) \geqslant 0,$$

so that $f(y) \geqslant f(x)$. We also see that if $f'(x_0) > 0$ for all x_0, then f is in fact strictly increasing, proving (2).

The proofs of (3) and (4) are similar. $\qquad\square$

REMARK 5.4.8. The example of $I = \mathbf{R}$ and $f(x) = x^3$, with $f'(0) = 0$, shows that it is not necessary to have $f'(x) > 0$ for all x in order for a function to be strictly increasing.

COROLLARY 5.4.9. *Let $I = [a, b]$ be an interval and $f \colon I \to \mathbf{R}$ a differentiable function such that f' is continuous. Then f is Lipschitz-continuous, and in fact*

$$|f(x) - f(y)| \leqslant M|x - y|$$

for all x and y in $[a, b]$, where $M = \max |f'(x)|$.

PROOF. For all $x \neq y$ in I, we find by the Mean-Value Theorem a real numbers x_0 in I such that

$$\frac{f(x) - f(y)}{x - y} = f'(x_0),$$

so that

$$|f(x) - f(y)| = |f'(x_0)||x - y| \leqslant M|x - y|.$$

$\square$

COROLLARY 5.4.10. *Let $I \subset \mathbf{R}$ be an interval and $f \colon I \to \mathbf{R}$ a differentiable function. The function f is constant on I if and only if $f' = 0$.*

PROOF. If f is constant then its derivative is everywhere 0. Conversely, suppose $f'(x) = 0$ for all $x \in I$. Let $x_0 \in I$ be fixed. For any $x \neq x_0$, we find $c \in I$ such that

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(c) = 0,$$

so that $f(x) = f(x_0)$ for all x, and hence f is constant. $\square$

EXAMPLE 5.4.11. (1) Corollary 5.4.7, together with the computation of the derivatives of cosine and sine, Lemma 4.5.6 and the symmetry properties of trigonometric functions, allows us to sketch the graphs of these two functions. For instance, we recover the fact that on the interval $[0, \pi/2]$, the cosine function is decreasing, since $\cos' = -\sin \leqslant 0$, and in fact strictly so since $\sin(x) > 0$ outside of the endpoints. Moreover, $\cos'(0) = -\sin(0) = 0$, so that the tangent line to the graph of the cosine at $x = 0$ is horizontal, and $\cos'(\pi/2) = -\sin(\pi/2) = -1$ shows that the tangent line at $x = \pi/2$ has slope -1.

(2) Among many applications, we illustrate one use of the derivative to obtain a very fast algorithm (due to Newton) to find the solutions of many equations involving differentiable functions.

Let $I \subset \mathbf{R}$ be an interval of real numbers, and let $f \colon I \to \mathbf{R}$ be a differentiable function with continuous derivative, which we assume to satisfy $f'(x) > 0$ for all x (in particular, it is strictly increasing). If we know two values $a < b$ such that

$$f(a) < 0 < f(b),$$

then the Intermediate Value Theorem, and the fact that f is injective, imply that there exists a unique $x_0 \in]a, b[$ such that $f(x_0) = 0$.

Newton's Algorithm attempts to construct x_0 as the limit of the sequence $(x_n)_{n \in \mathbf{N}}$ defined inductively by choosing some value of x_1, and defining

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad \text{for } n \geqslant 1.$$

Why this formula? The idea is that, given an approximation x_n of the root x, we define x_{n+1} to be the intersection point of the horizontal axis with the tangent line to the graph at the point $(x_n, f(x_n))$ (see Figure 5.4.11). This (usually) "leads in the right direction", so that should give a better approximation. Since the tangent has equation
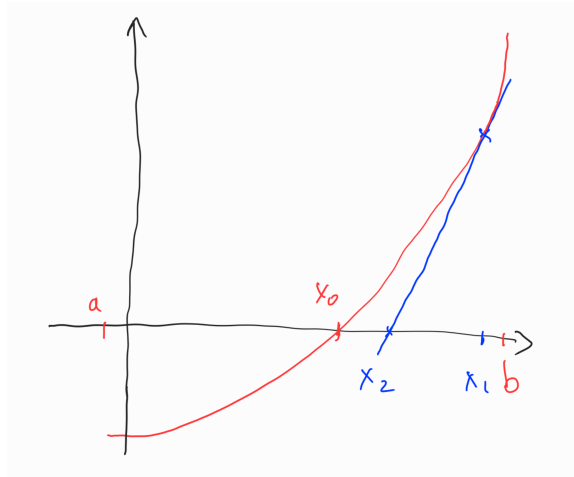
$$y - f(x_n) = f'(x_n)(x - x_n),$$

FIGURE 5.4. Newton's Algorithm

we see that the value x_{n+1} of x where $y = 0$ is given precisely by the expression above.

Under our assumptions, the inductive definition can be expressed in the form $x_{n+1} = g(x_n)$ where $g(x) = x - f(x)/f'(x)$ is continuous, so that the limit x of the sequence (x_n), if it exists satisfies $x = g(x)$, which is equivalent to

$$x = x - \frac{f(x)}{f'(x)}, \quad \text{or } f(x) = 0.$$

In practice, the sequence does not always converge (for instance, x_{n+1} might not belong to the interval of definition of f anymore), but it often does very fast.

Consider for instance a real number $c > 1$ and put $f(x) = x^2 - c$ on $[1, c]$, so that the zero of f in $[1, c]$ is $\sqrt{c}$. Then if we take the starting point $x_1 = c$, the inductive definition becomes

$$x_1 = c, \qquad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - c}{2x_n} = \frac{1}{2}\left(x_n - \frac{c}{x_n}\right),$$

which we recognize as the sequence that was used in Proposition 2.8.5 to construct the square-root of c.

(3) With the help of the derivative and Corollary 5.4.7, it is possible to investigate the basic properties of many simple differentiable function (e.g., where are they monotone, the location of local extrema, etc).

For instance, define the tangent function by

$$\tan(x) = \frac{\sin(x)}{\cos(x)}$$

for all x where this makes sense, namely for $x \in D$, the set of real numbers x such that x is not of the form $\frac{1}{2}\pi + k\pi$ for some $k \in \mathbf{Z}$ (for which the cosine would be zero). Note that this set is not an interval, but a union of infinitely many intervals of the form $I_k =]\frac{1}{2}\pi + k\pi, \frac{1}{2}\pi + (k+1)\pi[$, in particular on $I = I_{-1} =]-\pi/2, \pi/2[$. It is enough in fact to study the function on that single interval because

$$\tan(x + \pi) = \frac{\sin(x + \pi)}{\cos(x + \pi)} = \frac{-\sin(x)}{-\cos(x)} = \tan(x),$$

so the tangent function has period π. Morevoer, we have $\tan(-x) = -\tan(x)$, so in principle we can understand the function from its behavior on the interval $[0, \pi/2[$.

114
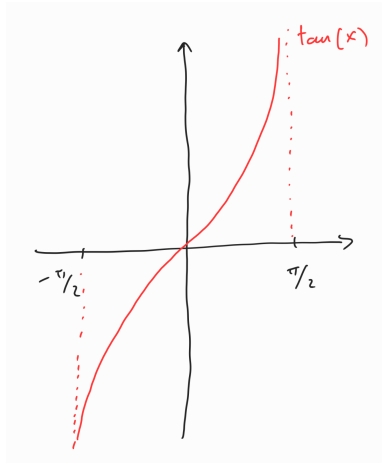
As a ratio of two functions that are differentiable, the tangent function is differentiable on any of those intervals, with derivative given by

$$\tan'(x) = \frac{\sin'(x)\cos(x) - \sin(x)\cos'(x)}{\cos(x)^2} = \frac{\cos(x)^2 + \sin(x)^2}{\cos(x)^2} = \frac{1}{\cos(x)^2}.$$

In particular, the tangent function is strictly increasing on each of the intervals I_k. It is therefore injective. To compute its image on I, we study the limit as $x \to -\pi/2$ and $x \to +\pi/2$. Since $\sin(\pi/2) = 1$ and $\sin(-\pi/2) = -1$, we get

$$\lim_{\substack{x\to-\pi/2 \\ x>-\pi/2}} \tan(x) = -\infty, \qquad \lim_{\substack{x\to\pi/2 \\ x<\pi/2}} \tan(x) = +\infty.$$

The image of tan is therefore an interval in $\mathbf{R}$ with no upper bound or lower bound, which is only possible if the image is $\mathbf{R}$. So the restriction of the tangent function is a bijection

$$\tan\colon\,]-\pi/2, \pi/2[\to \mathbf{R}.$$

The reciprocal bijection is denoted

$$\arctan\colon \mathbf{R} \to\,]-\pi/2, \pi/2[.$$

It is strictly increasing and satisfies

$$\lim_{x\to-\infty} \tan(x) = -\frac{\pi}{2}, \qquad \lim_{x\to+\infty} \tan(x) = \frac{\pi}{2},$$

as well as $\arctan(-x) = -\arctan(x)$. It is also differentiable since $\tan'(x) \neq 0$ on I, and

$$\arctan'(x) = \frac{1}{\tan'(\arctan(x))}$$

for all $x \in \mathbf{R}$. To simplify this, we use a different expression for the derivative of the tangent, which is

$$\tan'(x) = \frac{1}{\cos(x)^2} = 1 + \frac{\sin(x)^2}{\cos(x)^2} = 1 + \tan(x)^2.$$

This leads to

$$\tan'(\arctan(x)) = 1 + \tan(\arctan(x))^2 = 1 + x^2$$

for all $x \in \mathbf{R}$. In other words, we have the simple expression

(5.2) $$\arctan'(x) = \frac{1}{1 + x^2}$$

for all $x \in \mathbf{R}$.

The function arctan can be used sometimes to compute the argument of a complex number. Indeed, if $e^{i\theta} = x + iy$ with $-\pi/2 < \theta < \pi/2$ (which means that $x > 0$) then we have $y/x = \sin(\theta)/\cos(\theta) = \tan(\theta)$, hence

(5.3) $$\theta = \arctan(y/x).$$

5.5. Higher derivatives

If a function $f \colon I \to \mathbf{R}$ is differentiable on I, then its derivative f' is another function $f' \colon I \to \mathbf{R}$, and one can ask whether it is also differentiable. This leads to the definition of higher derivatives.

DEFINITION 5.5.1. Let $I \subset \mathbf{R}$ be an interval and $f \colon I \to \mathbf{R}$ a function. Let $k \in \mathbf{N}$. The k-th derivative $f^{(k)}$, if it exists, is the function $f^{(k)} \colon I \to \mathbf{R}$ defined inductively as follows:

(1) If f is differentiable, then $f^{(1)}$ is the derivative f' on I.
(2) If $f^{(k-1)}$ exists and is differentiable on I, then $f^{(k)} = (f^{(k-1)})'$.

If $f^{(k)}$ exists, then we say also that f is k-times differentiable on I. We also denote $f'' = f^{(2)}$ when it exists, and sometimes $f''' = f^{(3)}$.

REMARK 5.5.2. Note that $f^{(k)}$ exists, then so do the derivatives $f^{(j)}$ for $1 \leqslant j \leqslant k$. Moreover, if $1 \leqslant j < k$, then $f^{(j)}$ is $(k-j)$-times differentiable, and we have

$$(f^{(j)})^{(l)} = f^{(j+l)}$$

for $1 \leqslant l \leqslant k - j$.

If f and g are k-times differentiable on I, then for any real numbers a and b, the function $af + bg$ is also k-times differentiable, and

$$(af + bg)^{(j)} = af^{(j)} + bg^{(j)}$$

for $1 \leqslant j \leqslant k$.

We have seen that it is often also useful to know that the derivative of a function is continuous. We give a name to the set of functions with such properties (as well as for higher derivative).

DEFINITION 5.5.3. Let $I \subset \mathbf{R}$ be an interval. Let $k \in \mathbf{N}$. The set $C^k(I)$ is the set of all functions $f \colon I \to \mathbf{R}$ such that f is k-times differentiable on I and moreover $f^{(k)}$ is continuous on I. An element of $C^k(I)$ is called *a function of class C^k on I*.

If $f \in C^k(I)$ for all $k \in \mathbf{N}$, then we say that f is indefinitely differentiable on I; the set of such functions is denoted $C^\infty(I)$.

We denote also simply $C^0(I)$ or $C(I)$ the set of continuous functions on I, and for $f \in C^0(I)$, we write $f^{(0)} = f$.

REMARK 5.5.4. (1) We have $C^k(I) \subset C^{k-1}(I)$ for all $k \in \mathbf{N}$, and $C^\infty(I)$ is the intersection of all the spaces $C^k(I)$ for $k \in \mathbf{N}$.

(2) If f and g are in $C^k(I)$, with $k \in \mathbf{N}$ or $k = \infty$, then for any real numbers a and b we have $af + bg \in C^k(I)$, by Remark 5.5.2. So the set $C^k(I)$ is a vector space. Moreover, by Proposition 5.1.6, we see that the product of two functions f and g in $C^k(I)$ is also in $C^k(I)$, as well as f/g if $g(x) \neq 0$ for all $x \in I$. Also, the composite of two functions of class C^k is of class C^k by the Chain Rule.

EXAMPLE 5.5.5. The following is an example of a differentiable function f that is not in $C^1(I)$, or in other words, such that the derivative f' is not continuous. A standard example is the function $f : \mathbf{R} \to \mathbf{R}$ defined by $f(0) = 0$ and

$$f(x) = x^2 \sin(1/x)$$

for $x \neq 0$. It is easy to see that f is differentiable on the negative and positive numbers, with

$$f'(x) = 2x \sin(1/x) - \cos(1/x)$$

for $x \neq 0$. Using the definition, one also sees that

$$\frac{f(x) - f(0)}{x} = x \sin(1/x) \to 0$$

as $x \to 0$, so that f is also differentiable at $x = 0$ with $f'(0) = 0$. However, the derivative f' is *not* continuous at $x = 0$, because this would imply that $f'(x) \to 0$ as $x \to 0$, and this would require the function $g(x) = \cos(1/x)$ to have a limit as $x \to 0$, which is not the case.

Concerning the product, we have a precise formula for the k-th derivative, generalizing the Leibniz rule for the first derivative.

LEMMA 5.5.6 (General Leibniz formula). *Let $k \in \mathbf{N}$. Let $I \subset \mathbf{R}$ be an interval and f, $g \colon I \to \mathbf{R}$ functions that are k-times differentiable on I. Then fg is k-times differentiable and we have*

$$(fg)^{(k)} = \sum_{j=0}^{k} \binom{k}{j} f^{(j)} g^{(k-j)}.$$

PROOF. We proceed by induction on $k \in \mathbf{N}$. For $k = 1$, the formula is

$$(fg)' = fg' + f'g,$$

which is the Leibniz Rule of Proposition 5.1.6. If we assume that $k \in \mathbf{N}$ and that the formula holds for functions that are k-times differentiable, then for f and g assumed to be $(k+1)$-times differentiable on I, we write

$$(fg)^{(k+1)} = ((fg)^{(k)})'.$$

The induction hypothesis means that

$$(fg)^{(k)} = \sum_{j=0}^{k} \binom{k}{j} f^{(j)} g^{(k-j)}.$$

We compute its derivative using additivity and the Leibniz Rule for each term, and we obtain

$$(fg)^{(k+1)} = \sum_{j=0}^{k} \binom{k}{j} \left(f^{(j+1)} g^{(k-j)} + f^{(j)} g^{(k-j+1)} \right).$$

This is equal to

$$\sum_{j=1}^{k+1} \binom{k}{j-1} f^{(j)} g^{(k+1-j)} + \sum_{j=0}^{k} \binom{k}{j} f^{(j)} g^{(k+1-j)} =$$

$$f^{(k+1)} + \sum_{j=1}^{k} \left(\binom{k}{j-1} + \binom{k}{j} \right) f^{(j)} g^{(k+1-j)} + g^{(k+1)},$$

and using the formula from Lemma 1.6.11, we obtain the formula for the $(k+1)$-st derivative. $\square$

EXAMPLE 5.5.7. (1) From Proposition 5.1.7, we see by induction that

(1) Any polynomial belongs to $C^\infty(\mathbf{R})$;
(2) The exponential and cosine and sine belong to $C^\infty(\mathbf{R})$;
(3) The function $f(x) = \log(x)$ and, for any $a \in \mathbf{R}$, the function $g(x) = x^a$ belong to $C^\infty(]0, +\infty[)$.
(4) The functions arccos and arcsin belong to $C^\infty(]-1, 1[)$.

In some cases, we can also compute all derivatives. For instance, we have

$$\exp^{(k)} = \exp$$

for all k,x since $\exp' = \exp$. For cosine and sine, we have a periodicity of order 4:

$$\cos' = -\sin, \quad \cos'' = -\cos, \quad \cos^{(3)} = \sin, \quad \cos^{(4)} = \cos,$$
$$\sin' = \cos, \quad \sin'' = -\sin, \quad \sin^{(3)} = -\cos, \quad \sin^{(4)} = \sin,$$

and then the pattern repeats.

For the logarithm, we get $f'(x) = 1/x$, so $f''(x) = -1/x^2$, and then by induction

$$\log^{(k)}(x) = (-1)^{k-1}\frac{(k-1)!}{x^k}.$$

For the power function $f(x) = x^a$, we get

$$f^{(k)}(x) = a(a-1)\cdots(a-k+1)x^{a-k}$$

(noting that if $a \in \mathbf{N}$, then the first factor will be zero for $k \geqslant a+1$, in which case $f^{(k)}$ is always zero).

(2) We can also sometimes use the higher derivatives to prove inequalities. For instance, consider the function

$$f(x) = \sin(x) - \left(x - \frac{x^3}{6}\right).$$

Using properties of alternating series, we could check earlier that $f(x) \geqslant 0$ for $0 \leqslant x \leqslant 4$. In fact, the inequality holds for all $x \in \mathbf{R}_+$.

To check this, we notice that $f(0) = 0$ and that

$$f'(x) = \cos(x) - \left(1 - \frac{x^2}{2}\right).$$

We still have $f'(0) = 0$, but the sign of the derivative is not yet obvious. However, we can differentiate further and get

$$f''(x) = -\sin(x) + x, \qquad f^{(3)}(x) = -\cos(x) + 1.$$

It follows that $f^{(3)}(x) \geqslant 0$ for all x, so that f'' is non-decreasing, and since $f''(0) = 0$, we get $f'(x) \geqslant 0$ for $x \geqslant 0$, hence f is non-decreasing on $\mathbf{R}_+$, and since $f(0) = 0$, we get finally

$$f(x) \geqslant 0$$

for all $x \geqslant 0$.

PROPOSITION 5.5.8. *Let (a_n) be a sequence of real numbers such that the power series $\sum a_n x^n$ has positive radius of convergence R, and let $f \colon]-R, R[\to \mathbf{C}$ be the sum of the power series.*

We have $f \in C^\infty(]-R, R[)$, and

$$f^{(k)}(x) = \sum_{n=0}^{+\infty} (n+1) \cdots (n+k) a_{n+k} x^n$$

for $k \in \mathbf{N}_0$ and $|x| < R$. In particular, we have

$$a_k = \frac{f^{(k)}(0)}{k!}$$

for all $k \in \mathbf{N}_0$.

PROOF. The first part follows quickly by induction on k from the differentiability of power series in Example 5.2.3 (note that it is valid for $k = 0$, being just the power series expansion of f). Looking at the constant term, we see that

$$f^{(k)}(0) = (1 \cdot 2 \cdots k) a_k = k! a_k,$$

hence the last formula. □

EXAMPLE 5.5.9. Let $p(x) = a_n x^n + \cdots + a_1 x + a_0$ be a polynomial. Then it is also a power series, and we get

$$p(x) = \sum_{k=0}^{n} \frac{p^{(k)}(0)}{k!} x^k$$

for all $x \in \mathbf{R}$.

COROLLARY 5.5.10. *Let (a_n) and (b_n) be sequences of complex numbers such that the power series $\sum a_n x^n$ and $\sum b_n x^n$ have positive radius of convergence R_1 and R_2, respectively. If there exists $R \leqslant \min(R_1, R_2)$ such that the sums of the two series are equal for $|x| < R$, then $a_n = b_n$ for all $n \in \mathbf{N}_0$.*

PROOF. let f be the common sum of the two series for $|x| < r$. Since the derivatives are determined locally, we get

$$a_n = \frac{f^{(n)}(0)}{n!} = b_n$$

by the previous proposition. □

5.6. Convex functions

There are two further important applications of higher-derivatives. One is the geometric interpretation of the sign of the second derivative, and the other is the definition of polynomial approximations of higher-degree to a function that is k-times differentiable. We begin with the first topic, and the second will be handled in the next section.

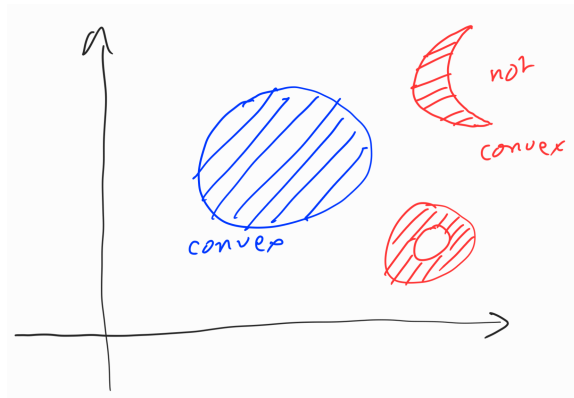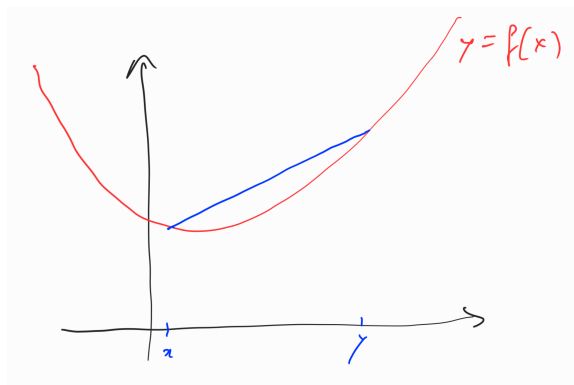The second derivative is closely connected to *convexity*.

FIGURE 5.5. Example of convex and non-convex sets

DEFINITION 5.6.1. (1) A subset $A \subset \mathbf{R}^2$ is *convex* if it contains the line segment joining any two of its points. In other words, if x_1 and x_2 are points of A, then all the points of the form

$$tx_1 + (1-t)x_2, \qquad 0 \leqslant t \leqslant 1$$

belong to A.

(2) Let $I \subset \mathbf{R}$ be an interval. A function $f \colon I \to \mathbf{R}$ is *convex* if the set

$$A_f = \{(x,y) \in \mathbf{R}^2 \mid y \geqslant f(x)\} \subset \mathbf{R}^2$$

is convex.

In practice, we use the convexity of a function through the following other definition, which implies a very useful type of inequalities:

LEMMA 5.6.2. *Let $I \subset \mathbf{R}$ be an interval.*

(1) A function $f \colon I \to \mathbf{R}$ is convex if and only if, for all $x \neq y$ in I and for all $t \in [0,1]$, we have

(5.4) $$f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y).$$

(2) If $f \colon I \to \mathbf{R}$ is convex, then for any integer $k \in \mathbf{N}$, any distinct elements $(x_1, \ldots, x_k)$ of I and any non-negative real numbers $(p_1, \ldots, p_k)$ such that

$$p_1 + \cdots + p_k = 1,$$

we have

(5.5) $$f(p_1 x + \cdots p_k x_k) \leqslant p_1 f(x_1) + \cdots + p_k f(x_k).$$

120

PROOF. (1) Suppose first that f is convex. Since the points $(x, f(x))$ and $(y, f(y))$ belong to the set A_f of the definition, so do the points in the segment joining them, which means that $(tx + (1-t)y, tf(x)(1-t)f(y)) \in A_f$, which translates to

$$tf(x) + (1-t)f(y) \geqslant f(tx + (1-t)y).$$

Conversely, assume the inequality (5.4) is valid, and let (x_1, y_1) and $(x_2, y_2) \in A_f$. We have $y_1 \geqslant f(x_1)$ and $y_2 \geqslant f(x_2)$; for $0 \leqslant t \leqslant 1$, we get

$$t(x_1, y_1) + (1-t)(x_2, y_2) = (tx_1 + (1-t)x_2, ty_1 + (1-t)y_2),$$

with

$$ty_1 + (1-t)y_2 \geqslant tf(x_1) + (1-t)f(x_2) \geqslant f(tx_1 + (1-t)y_1),$$

so that $t(x_1, y_1) + (1-t)(x_2, y_2) \in A_f$.

(2) For $k = 1$, we must have $p_1 = 1$ and the inequality is an equality. For $k = 2$, if we put $t = p_1$, then $p_2 = 1 - t$, so the inequality is simply (5.4). So we argue by induction on $k \geqslant 2$. We assume that (5.5) holds for an integer k, and that we have $(x_1, \ldots, x_{k+1})$ and $(p_1, \ldots, p_{k+1})$ in $\mathbf{R}_+$ with sum equal to 1. If $p_{k+1} = 1$, then all other p_i are zero, and we are done. Otherwise, let

$$y = \frac{p_1}{1 - p_{k+1}} x_1 + \cdots + \frac{p_k}{1 - p_{p+1}} x_k,$$

and $t = 1 - p_{k+1}$. Then the left-hand side of (5.5) is

$$f(ty + (1-t)x_{k+1}) \leqslant tf(y) + (1-t)f(x_{k+1}) = tf(y) + p_{k+1}f(x_{k+1}).$$

Moreover, since

$$\frac{p_1}{1 - p_{k+1}} + \cdots + \frac{p_k}{1 - p_{p+1}} = \frac{1 - p_{k+1}}{1 - p_{k+1}} = 1,$$

the induction assumption gives

$$tf(y) = (1 - p_{k+1})f\left(\sum_{i=1}^{k} \frac{p_i}{1 - p_{k+1}} x_i\right) \leqslant \sum_{i=1}^{k} p_i f(x_i),$$

and the result follows. □

EXAMPLE 5.6.3. (1) The set A_f is the set of points in the plane above the graph of the function f. Looking at simple cases, we see that the function $f(x) = |x|$ is convex on $\mathbf{R}$, as is the function $f(x) = x^2$. The function $f(x) = x^3$ is convex on $\mathbf{R}_+$, but not on $\mathbf{R}$ (for instance the segment joining $(-1, f(-1)) = (-1, -1)$ and $(0, f(0)) = (0, 0)$ is not contained in A_f).

(2) If f and g are convex and if a and b are non-negative real numbers, then the function $af + bg$ is convex: this follows easily by checking (5.4). Moreover, the function $\max(f, g)$ is convex: here we can deduce this from

$$A_{\max(f,g)} = A_f \cap A_g,$$

and the fact that the intersection of two convex sets is always convex.

(3) If (f_n) is a sequence of functions on I, and if $f_n(x)$ converges for all x to a limit $f(x)$, then the limit function f is convex on I: this follows from (5.4).

The link with higher-derivatives is the following statement:

THEOREM 5.6.4. *Let $I \subset \mathbf{R}$ be an interval. Let $f \in C^2(I)$. The function f is convex if and only if $f''(x) \geqslant 0$ for all $x \in I$, or in other words if and only if f' is non-decreasing on I.*
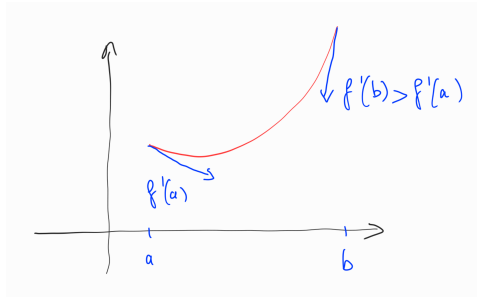
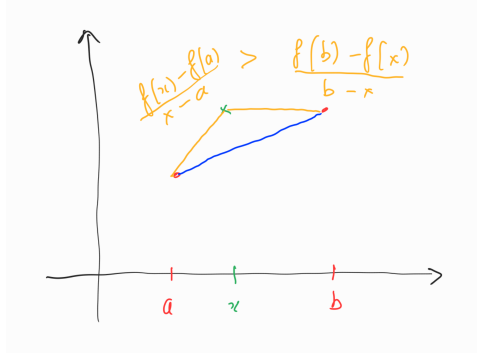FIGURE 5.6. First step of the proof of Theorem 5.6.4

FIGURE 5.7. Second step of the proof of Theorem 5.6.4

PROOF. Suppose first that f is convex. Let $a < b$ be elements of I. We consider again the function

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

describing the line from $(a, f(a))$ to $(b, f(b))$, and $h = f - g$; this is a function of class C^2. Since f is convex, it follows that $h(x) \leqslant 0$ for all $x \in [a, b]$. But moreover $h(a) = h(b) = 0$. This implies that $h'(a) \leqslant 0 \leqslant h'(b)$ (for instance because if $h'(a) > 0$, then by continuity we have $h'(x) > 0$ for x close to a, and then h would be strictly increasing close to a (by Corollary 5.4.7, (2)), contradicting $h(x) \leqslant 0 = h(a)$), which translates into

$$f'(a) - \frac{f(b) - f(a)}{b - a} \leqslant 0 \leqslant f'(b) - \frac{f(b) - f(a)}{b - a}.$$

Therefore the derivative f' is non-decreasing on $[a, b]$, and hence $f'' \geqslant 0$ on $[a, b]$ (Corollary 5.4.7, (1)). Since a and b are arbitrary, this means that $f'' \geqslant 0$ on I.

Conversely, we now assume that $f'' \geqslant 0$ on I, and therefore that f' is non-decreasing on I.

Let $a < b$ be elements of I, and construct the function $h = f - g$ as in (1). We need to check that $h \leqslant 0$ on $[a, b]$. If this is not the case, then there exists $x \in]a, b[$ such that $h(x) > 0$. Hence

$$f(x) > g(x) = \frac{f(b) - f(a)}{b - a}(x - a) + f(a), \quad \text{or} \quad \frac{f(x) - f(a)}{x - a} > \frac{f(b) - f(a)}{b - a}.$$

Moreover, writing $x - a = x - b + a - b$, we see that we can also write

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

122

for $x \in I$, so $h(x) > 0$ also implies

$$f(x) > \frac{f(b) - f(a)}{b - a}(x - b) + f(b), \quad \text{or} \quad \frac{f(b) - f(a)}{b - a} > \frac{f(b) - f(x)}{b - x}.$$

By the Mean-Value Theorem, we find $c \in]a, x[$ and $d \in]x, b[$ such that

$$\frac{f(x) - f(a)}{x - a} = f'(c), \qquad \frac{f(b) - f(x)}{b - x} = f'(d),$$

and deduce that $f'(d) < f'(c)$. Since $c < d$, this contradicts the fact that f' is non-decreasing. □

EXAMPLE 5.6.5. This criterion allows to check very easily that various functions are convex (or not), and from (5.4), we can deduce a number of very useful (and non-obvious) inequalities.

(1) First we note that we easily derive another consequence of the convexity for a C^2 function $f \colon I \to \mathbf{R}$: the graph of f is *above* the tangent at any point x_0. More precisely, since the equation of the tangent at x_0 is

$$y = f'(x_0)(x - x_0) + f(x_0),$$

the distance between the point $(x, f(x))$ and the tangent is

$$g(x) = f(x) - (f'(x_0)(x - x_0) + f(x_0)).$$

Our claim is that $g(x) \geqslant 0$ for all $x \in I$.

To see this, note that $g(x_0) = 0$, and that g is differentiable with

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

for $x \in I$. If f is convex, then $f'' \geqslant 0$, which means that f' is increasing. So:

- If $x \geqslant x_0$, then $g'(x) \geqslant 0$, so g is non-decreasing for $x \geqslant x_0$, which implies that indeed $g(x) \geqslant g(x_0) = 0$ for $x \geqslant x_0$.
- If $x \leqslant x_0$, then $g'(x) \leqslant 0$, so g is non-increasing for $x \leqslant x_0$, hence $g(x) \geqslant g(x_0) = 0$ also for $x \leqslant x_0$.

(2) Consider $f(x) = e^x$. This satisfies $f'' = f \geqslant 0$, so that f is convex on $\mathbf{R}$. We deduce that for any $x < y$ and $0 \leqslant t \leqslant 1$, we have

$$e^{tx + (1-t)y} \leqslant te^x + (1 - t)e^y.$$

(3) Let $r \in \mathbf{R}$. Define $f(x) = x^r$ on $I =]0, +\infty[$. Then f is in $C^2(I)$ and $f''(x) = r(r - 1)x^{r-2}$. It is therefore non-negative on I if either $r \geqslant 1$ or if $r \leqslant 0$, and f is convex in either of these cases, but not for $0 < r < 1$.

Assume that $r \geqslant 1$. We deduce that

$$(tx + (1 - t)y)^r \leqslant tx^r + (1 - t)y^r$$

(4) Let $p > 1$ and let $q > 1$ be the real number such that

$$\frac{1}{p} + \frac{1}{q} = 1$$

(for instance $q = 2$ if $p = 2$, and $q = 4/3$ if $p = 4$). The function $f(x) = -\log(x)$ is convex on $]0, +\infty[$, since

$$f'(x) = -\frac{1}{x}, \qquad f''(x) = \frac{1}{x^2},$$

123

hence for $x > 0$ and $y > 0$, we have

$$-\log\left(\frac{x}{p} + \frac{y}{q}\right) \leqslant -\frac{1}{p}\log(x) - \frac{1}{q}\log(y).$$

We take the opposite (changing the direction of the inequality) and then compute the exponential on both sides. This leads to

$$x^{1/p}y^{1/q} \leqslant \frac{x}{p} + \frac{y}{q}.$$

Replacing x by x^p and y by y^q, we obtain *Young's inequality*

$$xy \leqslant \frac{x^p}{p} + \frac{y^q}{q}.$$

COROLLARY 5.6.6 (Hölder's inequality). *Let $p > 1$ and let $q > 1$ be the real number such that*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

For any $k \in \mathbf{N}$, and for any complex numbers $(x_1, \dots, x_k)$ and $(y_1, \dots, y_k)$, we have

$$\left|\sum_{i=}^{k} x_i y_i\right| \leqslant \left(\sum_{i=1}^{k} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{k} |y_i|^q\right)^{1/q}.$$

For instance, when $p = q = 2$, we get the *Cauchy–Schwarz inequality*

$$\left|\sum_{i=}^{k} x_i y_i\right| \leqslant \left(\sum_{i=1}^{k} |x_i|^2\right)^{1/2} \left(\sum_{i=1}^{k} |y_i|^2\right)^{1/2}.$$

PROOF. Using the triangle inequality first, we see that it suffices the prove the inequality when $x_i \geqslant 0$ and $y_i \geqslant 0$. Removing any i where $x_i y_i = 0$, we can even assume that $x_i > 0$ and $y_i > 0$. Let then

$$N = \left(\sum_{i=1}^{k} x_i^p\right)^{1/p}, \quad M = \left(\sum_{i=1}^{k} y_i^q\right)^{1/q}.$$

These are positive real numbers; let

$$\widetilde{x}_i = \frac{x_i}{N}, \qquad \widetilde{y}_i = \frac{y_i}{M}.$$

We apply Young's inequality to $\widetilde{x}_i$ and $\widetilde{y}_i$, for each i, getting

$$\widetilde{x}_i \widetilde{y}_i \leqslant \frac{\widetilde{x}_i^p}{p} + \frac{\widetilde{y}_i^q}{q}.$$

Now summing over i, we deduce that

$$\frac{1}{NM}\sum_{i=1}^{k} x_i y_i \leqslant \frac{1}{p}\sum_{i=1}^{k} \widetilde{x}_i^p + \frac{1}{q}\sum_{i=1}^{k} \widetilde{y}_i^q.$$

But note that

$$\sum_{i=1}^{k} \widetilde{x}_i^p = \frac{1}{N^p}\sum_{i=1}^{k} x_i^p = 1, \quad \sum_{i=1}^{k} \widetilde{y}_i^p = \frac{1}{M^q}\sum_{i=1}^{k} y_i^q = 1,$$

so the inequality becoms

$$\sum_{i=1}^{k} x_i y_i \leqslant NM,$$

which is precisely what we wanted to prove. □

COROLLARY 5.6.7 (Minkowski's Inequality). *Let $p > 1$. For any $k \in \mathbf{N}$, and for any complex numbers $(x_1, \ldots, x_k)$ and $(y_1, \ldots, y_k)$, we have*

$$\left(\sum_{i=}^{k} |x_i + y_i|^p\right)^{1/p} \leqslant \left(\sum_{i=1}^{k} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{k} |y_i|^p\right)^{1/p}.$$

PROOF. We note first that

$$\sum_{i=}^{k} |x_i + y_i|^p = \sum_{i=1}^{k} |x_i + y_i|\ |x_i + y_i|^{p-1} \leqslant \sum_{i=1}^{k} |x_i|\ |x_i + y_i|^{p-1} + \sum_{i=1}^{k} |y_i|\ |x_i + y_i|^{p-1}.$$

We apply Hölder's inequality to both terms in this sum: we have

$$\sum_{i=1}^{k} |x_i|\ |x_i + y_i|^{p-1} \leqslant \left(\sum_{i=1}^{k} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{k} |x_i + y_i|^{q(p-1)}\right)^{1/q}$$

and similarly for (y_i). Note that $q(p-1) = p$, so that by adding the two terms we get

$$\sum_{i=}^{k} |x_i + y_i|^p \leqslant \left(\sum_{i=}^{k} |x_i + y_i|^p\right)^{1/q} \left(\left(\sum_{i=1}^{k} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{k} |y_i|^p\right)^{1/p}\right).$$

Again since $1 - 1/q = 1/p$, we obtain the result (if the left-hand side is non-zero, but the inequality is clear if it is equal to 0). □

5.7. Taylor polynomials

Reference: [2, 14.1].

We used the idea of linear approximation of a function to motivate the definition of the derivative. For functions which are k-times differentiable, we can get approximations by polynomials of higher degree. Which polynomials should be used is suggested by the formula

$$p(x) = \sum_{n=0}^{k} \frac{p^{(n)}(0)}{n!} x^n$$

valid for all x if p is a polynomial of degree k. More precisely:

DEFINITION 5.7.1 (Taylor polynomials). Let $k \in \mathbf{N}$. Let $I \subset \mathbf{R}$ be an interval and $f \colon I \to \mathbf{R}$ a function that is k-times differentiable on I. Let $x_0 \in I$. The polynomial

$$T_k f(x; x_0) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k = \sum_{n=0}^{k} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

is called the *k-th Taylor polynomial of f at x_0*.

REMARK 5.7.2. (1) The function $T_k f(x; x_0)$ is a polynomial of degree at most k; it can be of smaller degree (if $f^{(k)}(x_0) = 0$).

(2) If f is a polynomial of degree $d \in \mathbf{N}$, then we have

$$T_k f(x; x_0) = f(x)$$

for all x if $k \geqslant d$. (However, if $k < d$, then the k-th Taylor polynomial is not equal to f, since not all derivatives are used.)

(3) The Taylor polynomials are constructed so that the relation $(T_k f)' = T_{k-1}(f')$ holds (we omit the point x_0, which is the same on both sides).
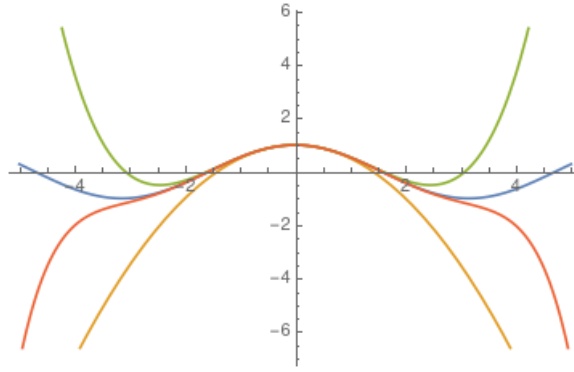
FIGURE 5.8. Taylor polynomials of cosine

EXAMPLE 5.7.3. (1) Let $f(x) = e^x$; then $f^{(k)} = f$ for all k. Taking $x_0 = 0$, we get the Taylor polynomials

$$T_k \exp(x; 0) = 1 + x + \cdots + \frac{x^k}{k!}.$$

These are the partial sums of the power series defining the exponential; in particular, in that case, they converge to the function f.

Take now instead $x_0 = 1$. Then $f^{(k)}(1) = e^1 = e$ for all k, so that

$$T_k \exp(x; 1) = e + e(x - 1) + \cdots + \frac{e(x - 1)^k}{k!}.$$

Here we get also

$$\lim_{k \to +\infty} T_k \exp(x; 1) = e \sum_{n=0}^{+\infty} \frac{(x - 1)^k}{k!} = e \cdot e^{x-1} = e^x.$$

(2) If $f(x) = 1/(1 - x)$ for $x \in\,]-2, 1[$, then for $x_0 = 1$, we easily obtain

$$T_k f(x; 0) = 1 + x + \cdots + x^k.$$

Note that, in this case, we do not have $T_k f(x; 0) \to f(x)$ for all x (for instance for $x = -1$).

(3) We present in Figure 5.7.3 a graph of $\cos(x)$ (in blue) and of the first Taylor polynomials for $x_0 = 0$, evaluated for $-5 \leqslant x \leqslant 5$.

How good the Taylor polynomials approximates f is described by the next theorem (another version of which, with a better estimate of the error, will be proved later).

THEOREM 5.7.4. *Let $k \in \mathbf{N}_0$. Let $I \subset \mathbf{R}$ be an interval and let $f \colon I \to \mathbf{R}$ be a function that is $(k + 1)$-times differentiable on I. Let $x_0 \in I$. For any $x \in I$, there exists $c \in I$ between x and x_0 such that*

$$f(x) = T_k f(x; x_0) + \frac{f^{(k+1)}(c)}{(k + 1)!}(x - x_0)^{k+1}$$

or in other words

$$f(x) = \sum_{n=0}^{k} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(k+1)}(c)}{(k + 1)!}(x - x_0)^{k+1}.$$

EXAMPLE 5.7.5. For $k = 1$ and $x_0 = 0$, this means that

$$f(x) = f(0) + x f'(0) + \frac{x^2}{2} f''(c).$$

PROOF. We prove the result under the assumption that $f^{(k+1)}$ is continuous, although that is not necessary for the validity of the theorem; we will obtain in the next chapter a different, and more powerful, form of this expression.

We note first that if $x = x_0$, then we can take $c = x_0$ indeed. Assuming that $x \neq x_0$, there exists a unique real number a such that

$$f(x) = \sum_{n=0}^{k} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{a}{(k+1)!}(x - x_0)^{k+1}.$$

We define a function g on I by

$$g(y) = f(y) - \sum_{n=0}^{k} \frac{f^{(n)}(x_0)}{n!}(y - x_0)^n - \frac{a}{(k+1)!}(y - x_0)^{k+1}.$$

for $y \in I$, so that the condition on a becomes $g(x) = 0$.

The function g is $(k+1)$-times differentiable on I. Computing its j-th derivative for $0 \leqslant j \leqslant k$, we obtain

$$g^{(j)}(y) = f^{(j)}(y) - \sum_{n=j}^{k} \frac{f^{(n)}(x_0)}{(n-j)!}(y - x_0)^{n-j} - \frac{a}{(k+1)!}(y - x_0)^{k+1},$$

for $y \in I$ and

$$g^{(k+1)}(y) = f^{(k+1)}(y) - a.$$

In particular, this leads to

$$g(x_0) = \cdots = g^{(k)}(x_0) = 0, \qquad g^{(k+1)}(x_0) = f^{(k+1)}(x_0) - a.$$

Now suppose first that $g^{(k+1)}(x_0) > 0$. We claim that there must exist some c between x and x_0 such that $g^{(k+1)}(c) = 0$. Indeed, otherwise, by continuity of $g^{(k+1)}$, we get $g^{(k+1)}(y) > 0$ between x and x_0, so that $g^{(k)}$ is strictly increasing, hence $g^{(k)}(y) > g^{(k)}(x_0) = 0$ between x_0 and x; by induction we would deduce that g is strictly increasing between x_0 and x which contradicts the fact that $g(x_0) = g(x) = 0$.

Now the number c satisfies

$$0 = g^{(k+1)}(c) = f^{(k+1)}(c) - a,$$

and the relation $g(x) = 0$ leads to the conclusion. We argue similarly if $g^{(k+1)}(x_0) < 0$, and if $g^{(k+1)}(x_0) = 0$, then we are also done. $\qquad \square$

COROLLARY 5.7.6. *Let* $k \in \mathbf{N}$. *Let* $I \subset \mathbf{R}$ *be an interval and let* $f \colon I \to \mathbf{R}$ *be a function that is* $(k+1)$*-times differentiable on* I. *Let* $x_0 \in I$. *Assume that* $m \leqslant M$ *are real numbers such that*

$$m \leqslant |f^{(k+1)}(x)| \leqslant M$$

for all $x \in I$. *Then we have*

$$\frac{m}{(k+1)!}|x - x_0|^{k+1} \leqslant \left| f(x) - T_k f(x; x_0) \right| \leqslant \frac{M}{(k+1)!}|x - x_0|^{k+1}$$

for all $x \in I$.

EXAMPLE 5.7.7. Let $f(x) = \log(1 + x)$ on $I =]\frac{1}{2}, 1[$. We will use the Taylor formula to prove that

$$f(x) = \log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots = \sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n} x^n.$$

In particular, the logarithm has a power series expansion on this interval (we will later see that the formula also holds for $-1 < x \leqslant 0$, but that's not obvious from the theorem). Note that, in any case, the radius of convergence of this series is equal to 1 (like the geometric series), so that one cannot use the power series to compute (for instance) $\log(1 + 2) = \log(3)$.

We first compute the derivatives of f, with is of class C^∞ on I. We have

$$f'(x) = \frac{1}{1+x}, \quad f''(x) = -\frac{1}{(1+x)^2}, \quad f^{(3)}(x) = \frac{2}{(1+x)^3},$$

and more generally by induction

(5.6)
$$f^{(n)}(x) = (-1)^{n-1} \frac{(n-1)!}{(1+x)^n}.$$

Since $f(0) = 0$, it follows that the Taylor polynomial of f of degree n is

$$f'(0)x + \frac{f''(0)}{2} x^2 + \cdots + \frac{f^{(n)}(0)}{n!} x^n = x - \frac{x^2}{2} + \cdots + (-1)^{n-1} \frac{x^n}{n}.$$

There exists therefore c between 0 and x such that

$$\left| \log(x) - \left(x - \frac{x^2}{2} + \cdots + (-1)^{n-1} \frac{x^n}{n} \right) \right| = \frac{n!}{(n+1)!} \frac{|x|^n}{(1+c)^n}.$$

If $0 \leqslant x < 1$, then the right-hand side is $\leqslant |x|^n/(n+1)$, which converges to 0 as $n \to +\infty$. If $-1/2 < x \leqslant 0$, then we have $x \leqslant c \leqslant 0$, so that $|x/(1+c)| \leqslant |x/(1+x)| < 1$, and we get the same conclusion.

The next corollary explains how the Taylor polynomials can provide better and better approximation of a function.

COROLLARY 5.7.8. *Let $k \in \mathbf{N}$. Let $I \subset \mathbf{R}$ be an interval and let $f \colon I \to \mathbf{R}$ be a function of class C^{k+1}. Let $x_0 \in I$. We have*

$$f(x) = T_k f(x; x_0) + (x - x_0)^k r(x)$$

where

$$\lim_{x \to x_0} r(x) = 0,$$

or in other words

$$\lim_{x \to x_0} \frac{1}{(x - x_0)^k} \left(f(x) - T_k f(x; x_0) \right) = 0.$$

PROOF. Since $f^{(k+1)}$ is continuous, there exists $M \in \mathbf{R}_+$ such that $|f^{(k+1)}(x)| \leqslant M$ for all $x \in I$ such that $|x - x_0| \leqslant 1$. Then by the theorem, for any $x \neq x_0$ in I with $|x - x_0| \leqslant 1$, there exists c between x and x_0 such that

$$\left| \frac{1}{(x - x_0)^k} \left(f(x) - T_k f(x; x_0) \right) \right| = \frac{|f^{(k+1)}(c)|}{(k+1)!} |x - x_0| \leqslant \frac{M}{(k+1)!} |x - x_0|,$$

which tends to 0 as $x \to x_0$. $\qquad\qquad \square$

REMARK 5.7.9. (1) Concretely, assume that x is an approximation of x_0 with $m \geqslant 1$ digits of precision. Then the limit above shows that for m large enough, the difference

$$f(x) - T_k f(x; x_0)$$

is "much smaller" than $|x - x_0|^k$, which means that the approximation of $f(x)$ by $T_k f(x; x_0)$ has roughly km digits of precision.

(2) Given an interval I and a function $f \in C^\infty(I)$, we can consider the power series

$$\sum_{n=0}^{+\infty} \frac{f^{(n)}(0)}{n!} x^n.$$

One might expect that there is some relation between this power series and the function f. However, this is not at all the case in general! Almost the only thing that can be said is that, if this power series has a positive radius of convergence, then its sum g satisfies

$$g^{(n)}(0) = f^{(n)}(0)$$

for all $n \in \mathbf{N}_0$.

More precisely, the following can happen:

- It may be that f is defined on $\mathbf{R}$, but the power series has zero radius of convergence (for instance, there is a function $f \in C^\infty(\mathbf{R})$ with $f^{(n)}(0) = (n!)^2$ for all n, which implies that the power series is the power series

$$\sum_{n=0}^{+\infty} n! x^n,$$

  whose radius of convergence is zero.
- It may be that the radius of convergence of the power series is $+\infty$, and yet its sum g only satisfies $f(x) = g(x)$ for $x = 0$. An example is the function f defined by $f(0) = 0$ and $f(x) = \exp(1/x^2)$ for $x \neq 0$. It needs to be checked using the definition that f is indeed in $C^\infty(\mathbf{R})$ (the issue is at 0, of course), but one can do this and prove that

$$f^{(n)}(0) = 0$$

  for all $n \in \mathbf{N}_0$. So the power series has all coefficients 0, hence converge at all points to 0, whereas $f(x) = 0$ is only true for $x = 0$.

Another basic application of Taylor polynomials is as a way to study certain limits. The idea is to reduce complicated limits as $x \to x_0$ of ratios $f(x)/g(x)$ to that of ratios of polynomials, by showing that one can replace the numerator and denominator with Taylor polynomials for f and g of suitable order. We illustrate this with an example.

EXAMPLE 5.7.10. Does the limit

$$\lim_{x \to 0} \frac{\cos(x) - 1 + x^2/2}{\sin(x^4)}$$

exist? If Yes, what is its value? Since numerator and denominator tend to 0 as $x \to 0$, we can be tempted to use L'Hospital's Rule (Proposition 5.1.9). But the derivative of the denominator is $4x^3 \cos(x^4)$, which also vanishes at 0, so we would have to do it iteratively.

It is simpler to use the Taylor polynomials around 0. Using the polynomial of degree 4 for $\cos(x)$, we find that the denominator is of the form

$$\cos(x) - 1 + \frac{x^2}{2} = \frac{x^4}{24} + x^4 r_1(x)$$

where $r_1(x) \to 0$ as $x \to 0$ by Corollary 5.7.8. Similarly, (but using simply the Taylor polynomial of degree 1 of $\sin(y)$, and replacing $y = x^4$) we have

$$\sin(x^4) = x^4 + x^4 r_2(x^4)$$

with $r_2(y) \to 0$ as $y \to 0$. So
$$\frac{\cos(x) - 1 + x^2/2}{\sin(x^4)} = \frac{x^4/24 + x^4 r_1(x)}{x^4 + x^4 r_2(x^4)} = \frac{1/24 + r_1(x)}{1 + r_2(x^4)} \to \frac{1}{24}$$
as $x \to 0$.

As a final application of higher derivatives, we can find a criterion to determine whether a point where $f'(x_0) = 0$ is a local extremum, or not.

THEOREM 5.7.11. *Let $k \in \mathbf{N}$. Let I be an interval and let $f \colon I \to \mathbf{R}$ be of class C^k on I. Let $x_0 \in I$, neither the maximum nor the minimum of I, if these exist. Suppose that $f'(x_0) = 0$, and that there exists $j \leqslant k$ such that*
$$f'(x_0) = f''(x_0) = \cdots = f^{(j-1)}(x_0) = 0,$$
and $f^{(j)}(x_0) \neq 0$.
  (1) *If j is odd, then x_0 is not a local extremum of f.*
  (1) *If j is even, then the point x_0 is a local minimum of f if and only if $f^{(j)}(x_0) > 0$.*
  (2) *If j is even, then the point x_0 is a local maximum of f if and only if $f^{(j)}(x_0) < 0$.*

PROOF. Note first that since $f^{(j)}(x_0) \neq 0$, and f is of class C^j, so that the j-th derivative is continuous, we know that there exists $\delta > 0$ such that $f^{(j)}(x)$ is of the same sign as $f^{(j)}(x_0)$ if $|x - x_0| < \delta$, by Lemma 3.2.7, (2).

Under the given assumptions, the $(j-1)$-st Taylor polynomial of f at x_0 is the constant $f(x_0)$. So Theorem 5.7.4 states that for any $x \neq x_0$ in I, there exists c between x and x_0 such that
$$f(x) = f(x_0) + \frac{(x - x_0)^j}{j!} f^{(j)}(c), \text{ or } f(x) - f(x_0) = \frac{(x - x_0)^j}{j!} f^{(j)}(c).$$
If $|x - x_0| < \delta$, then we also have $|c - x_0| < \delta$, since c is between x and x_0, so the sign of $f^{(j)}(c)$ is the same as the sign of $f^{(j)}(x_0)$. So this formula allows us to see what is the sign of $f(x) - f(x_0)$ when $|x - x_0| < \delta$, and all the statements follow.

For instance, if j is odd, then $(x - x_0)^j$ changes sign when x moves from being $< x_0$ to being $> x_0$, so the sign of $f(x) - f(x_0)$ is not constant on any small interval around x_0, which means that f does not have a local extremum at x_0.

On the other hand, if j is even, then $(x - x_0)^j \geqslant 0$ for all x, so that the sign of $f(x) - f(x_0)$ is the same as that of $f^{(j)}(x_0)$, which gives the last two statements. $\qquad \square$

The most important special case is the following:

COROLLARY 5.7.12. *Let I be an interval and let $f \colon I \to \mathbf{R}$ be of class C^2 on I. Let $x_0 \in I$, neither the maximum nor the minimum of I, if these exist. Suppose that $f'(x_0) = 0$, and that $f''(x_0) \neq 0$.*
  (1) *The point x_0 is a local minimum of f if and only if $f''(x_0) > 0$.*
  (2) *The point x_0 is a local maximum of f if and only if $f''(x_0) < 0$.*

Note that, as usual, when using these results to locate the local extrema, one has to consider separately the possible maximum or minimum of the interval I.

EXAMPLE 5.7.13. (1) Let $n \in \mathbf{N}$ and $f(x) = x^n e^{-x}$ on $\mathbf{R}_+$. We attempt to find if f has a maximum or a minimum on $\mathbf{R}_+$. For the minimum, the answer is easy to see: we have $f(x) \geqslant 0$ for all x, and $f(0) = 0$, so the minimum of f is 0, achieved at 0.

Now to find local extrema, we compute
$$f'(x) = n x^{n-1} e^{-x} - x^n e^{-x} = x^{n-1} e^{-x} (n - x).$$
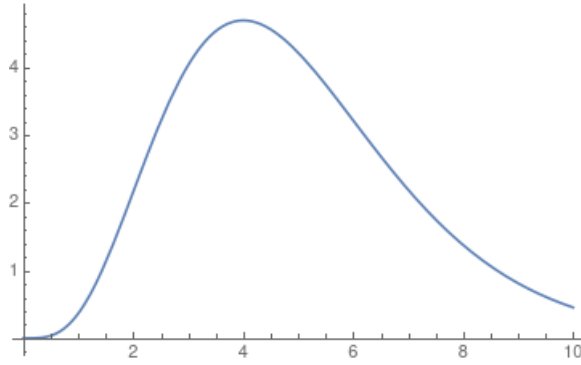
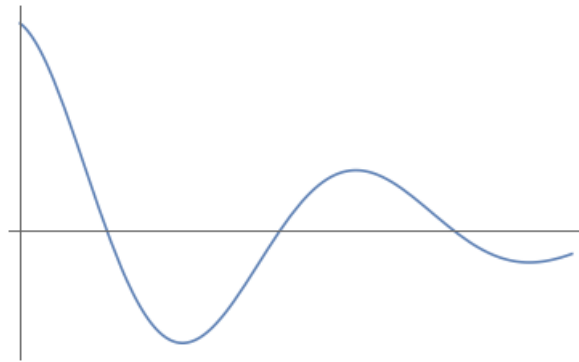FIGURE 5.9. Graph of $f(x) = x^4 e^{-x}$

FIGURE 5.10. Graph of $f(x) = \cos(x)e^{-x}$

We have $f'_d(0) = 0$ if $n \geqslant 2$, and otherwise the only possible local extremum is $x = n$.

We now check first that $x = n$ is a local maximum. Indeed, differentiating a second time f', or using the Leibniz formula for the second derivative of a product with x^n and e^{-x} (see Lemma 5.5.6), we get

$$f''(x) = n(n-1)x^{n-2}e^{-x} - 2nx^{n-1}e^{-x} + x^n e^{-x} = x^{n-2}e^{-x}(n(n-1) - 2nx + x^2).$$

In particular, taking $x = n$, we get

$$f''(n) = n^{n-2}e^{-n}(n^2 - n - 2n^2 + n^2) = -n^{n-1}e^{-n} < 0.$$

According to the corollary, this means that f has a *local maximum* at $x = n$.

We now check that $x = n$ is a global maximum. For this purpose, observe that $f'(x) < 0$ for $x > n$, by the formula above, so that f is strictly decreasing for $x > n$, in particular $f(x) \leqslant f(n)$ for $x \geqslant n$. Moreover $f'(x) > 0$ for $0 \leqslant x < n$, so the function is strictly increasing on $[0, n]$, and $f(x) \leqslant f(n)$ for $0 \leqslant x \leqslant n$ also.

We conclude that $f(n) = (n/e)^n$ is the maximum of f for $x \geqslant 0$.

(2) Let $f(x) = \cos(x)e^{-x}$ on $\mathbf{R}$. What are the local extrema of f (if any)?

We compute first the derivative

$$f'(x) = -\sin(x)e^{-x} - \cos(x)e^{-x} = -e^{-x}(\cos(x) + \sin(x)).$$

So we need to find the solutions of the equation

$$\cos(x) + \sin(x) = 0.$$

For such a value of x, we get $1 = \cos(x)^2 + \sin(x)^2 = 2\cos(x)^2$, so that the possible values of $(\cos(x), \sin(x))$ are

$$\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right), \quad \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).$$

These correspond to the points

$$e^{-i\pi/4}, \quad e^{3i\pi/4}$$

on the unit circle. The corresponding values of x are $3\pi/4$ or $-\pi/4$, but only up to addition of a multiple of 2π. These can be summarized with the formula

$$x_k = \frac{3\pi}{4} + k\pi,$$

where $k \in \mathbf{Z}$ (because $x_{-1} = -\pi/4$, we recover all the values this way).

Now to identify whether these are really local extrema, we compute

$$f''(x) = -\cos(x)e^{-x} + 2\sin(x)e^{-x} + \cos(x)e^{-x} = 2\sin(x)e^{-x},$$

hence

$$f''(x_k) = 2e^{-x_k}\sin\left(\frac{3\pi}{4} + k\pi\right).$$

If k is even, this gives

$$f''(x_k) = 2e^{-x_k}\sin\left(\frac{3\pi}{4}\right) = \sqrt{2}e^{-x_k} > 0,$$

and if k is odd then

$$f''(x_k) = 2e^{-x_k}\sin\left(-\frac{\pi}{4}\right) = -\sqrt{2}e^{-x_k} < 0.$$

So for x_{2k}, we have a local minimum, and for x_{2k+1}, we have a local maximum.

Since $e^{-x} \to +\infty$ as $x \to -\infty$, and $\cos(x)$ can be either 1 or -1 with x more and more negative, it follows from the continuity of f that its image is equal to $\mathbf{R}$. This means that none of the local extrema that we have found is a global maximum or minimum.

CHAPTER 6

Integration

In this chapter, we define another essential process of analysis, that of *integration*. This has many extremely varied applications, among which:

- It "reverses" the process of differentiation: given a continuous function g, it gives a function f with $f' = g$.
- It can be used to compute, and even to define rigorously, the area of subsets of $\mathbf{R}^2$ (such as a disc).
- It can be used to compute the length of a curve in the plane, for instance the perimeter of a circle.
- It can be used to define completely new functions, some of which are very important in analysis (for instance, the Gamma function of Euler, which "extends" the factorial to all real numbers).
- It provides the tool to find the decomposition of a periodic signal in a sum of "pure" waves (in other words, it gives the values of the numbers a_n when a function f is defined as a sum, possibly infinite, of the form

$$f(x) = \sum_n a_n \cos(nx), \text{ or } f(x) = \sum_n a_n \sin(nx).$$

- It leads to the correct mathematical concepts of *probability* and to the whole theory of probability and all its applications.
- It can be used to define "Hilbert spaces", which are the natural infinite-dimensional analogues of the euclidean plane and 3-dimensional space; these spaces provide the right setting for Quantum Mechanics and are therefore essential to much of modern technology.

6.1. Primitives

We begin with the simplest approach, by attempting to reverse the differentiation process. This is not theoretically satisfactory, because (in contrast to computing the derivative) there is no good algorithm to do this in general.

DEFINITION 6.1.1. Let I be an interval of $\mathbf{R}$ and $g\colon I \subset \mathbf{R}$ an arbitrary function. A *primitive*[46] of g is a function $f\colon I \to \mathbf{R}$ such that f is differentiable on I and $f' = g$ on I.

PROPOSITION 6.1.2. *Let I be an interval of $\mathbf{R}$ and $g\colon I \subset \mathbf{R}$ an arbitrary function. If there exists a primitive f of g, then:*

(1) All primitives of g are of the form $f + c$ for some constant $c \in \mathbf{R}$, and all such functions are primitives of g.

(2) For any $x_0 \in I$, there exists a unique primitive $\widetilde{f}$ of g such that $\widetilde{f}(x_0) = 0$.

When g has a primitive, we will use the notation

$$\int_{x_0}^{x} g(t)dt$$

for the value at x of the unique primitive of g that is zero at x_0. This is called "the integral of g between x_0 and x".

We sometimes write

$$\int g(t)dt$$

for an *arbitrary* primitive of g (this is therefore not a well-defined function).

PROOF. (1) Since $(f + c)' = f' = g$, adding a constant to a primitive f of g gives another primitive. Conversely, if $f_1 \colon I \to \mathbf{R}$ is a primitive of g, then $f' = f_1'$, which means that $(f_1 - f)' = 0$. By Corollary 5.4.10, this means that $f_1 - f$ is a constant, say equal to c, so that $f_1 = f + c$.

(2) Let $\widetilde{f} = f - f(x_0)$; then $\widetilde{f}$ is a primitive of g with $\widetilde{f}(x_0) = f(x_0) - f(x_0) = 0$. It is the only one, since a primitive f_1 of g with this property must be of the form $f_1 = f + c$, and evaluating at x_0, we get $0 = f_1(x_0) = f(x_0) + c$, so that $c = -f(x_0)$. $\square$

This proposition is not very useful unless one has a way of knowing that one primitive exists. For the moment, we can only "observe" that some functions do, but in Section 6.2, we will show that any continuous function as a primitive.

EXAMPLE 6.1.3. (1) Just by looking at the derivatives of any known function, we get a list of primitives. For instance, we get

$$\int_0^x t^n dt = \frac{1}{n+1} t^{n+1} \text{ if } n \in \mathbf{N}_0, \qquad \int_1^x \frac{1}{t} dt = \log(x),$$

$$\int_0^x e^t dt = e^x - 1, \qquad \int_0^x \frac{1}{1+t^2} dt = \arctan(x).$$

Note that e^x is a primitive of e^x that does not vanish at any point, so it is not of the form $\int_{x_0}^x e^t dt$.

(2) If f is differentiable on I, then f' has f as primitive, and we get

$$\int_{x_0}^x f'(t)dt = f(x) - f(x_0)$$

for any x_0 and any x in I. Indeed, the right-hand, as a function of x, is a primitive of f' and takes the value 0 at $x = x_0$.

Similarly, any rule for differentiating functions leads to a rule for computing primitives. But these are, except for addition, rather more complicated than for the derivatives.

PROPOSITION 6.1.4. *Let $I \subset \mathbf{R}$ be an interval, and let g_1, g_2 be real-valued functions on I.*

(1) If g_1 and g_2 have primitives, then for any real numbers a and b, the function $ag_1 + bg_2$ has primitives. Moreover, for any $x_0 \in I$, we have

(6.1) $$\int_{x_0}^x (ag_1(t) + bg_2(t))dt = a \int_{x_0}^x g_1(t)dt + b \int_{x_0}^x g_2(t)dt,$$

for all $x \in I$.

(2) If g_1 and g_2 are differentiable on I, then $g_1 g_2'$ has a primitive if and only if $g_1' g_2$ does, and

(6.2) $$\int_{x_0}^x g_1'(t)g_2(t)dt = g_1(x)g_2(x) - g_1(x_0)g_2(x_0) - \int_{x_0}^x g_1(t)g_2'(t)dt$$

for any $x_0 \in I$ and $x \in I$.

The formula (6.2) is called *integration by parts.*[47] It should be noted that it does *not* compute directly any primitive, but it reduces the computation for one function to that for another, which might be simpler.

PROOF. (1) This is very simple since, if $f_1' = g_1$ and $f_2' = g_2$, then the derivative of $af_1 + bf_2$ is $af_1' + bf_2' = ag_1 + bg_2$, so that $af_1 + bf_2$ is a primitive of $ag_1 + bg_2$; then (6.1) is valid because both sides are primitives that take the value 0 at x_0.

(2) Here we have a form of the Leibniz formula, although this might not be clear: the function $g_1 g_2$ has derivative $g_1' g_2 + g_1 g_2'$ by Proposition 5.1.6, (2). Looking at the value at x_0, this means that

$$\int_{x_0}^x (g_1'(t)g_2(t) + g_1(t)g_2'(t))dt = g_1(x)g_2(x) - g_1(x_0)g_2(x_0).$$

Using the additivity property from (1), we obtain (6.2) by subtracting the second part. $\square$

REMARK 6.1.5. If $g\colon [a, b] \to \mathbf{R}$ is an arbitrary function, one sometimes writes

$$[g(t)]_a^b = g(b) - g(a).$$

So for instance the formula (6.2) can be written

$$\int_{x_0}^x g_1'(t)g_2(t)dt = [g_1(t)g_2(t)]_{x_0}^x - \int_{x_0}^x g_1(t)g_2'(t)dt.$$

EXAMPLE 6.1.6. (1) Since any power function $f(x) = x^n$ with $n \in \mathbf{N}_0$ has a primitive, the first property implies that any polynomial

$$p(x) = a_n x^n + \cdots + a_1 x + a_0$$

has a primitive, and that, for instance

$$\int_0^x p(t)dt = \frac{1}{n+1}a_n x^{n+1} + \frac{1}{n}a_{n-1}x^n + \cdots + \frac{1}{2}a_1 x^2 + a_0 x.$$

(2) The function $f(x) = xe^x$ on $\mathbf{R}$ has a primitive. Indeed, we can write $f = g_1 g_2'$ with $g_1(x) = x$ and $g_2(x) = e^x$. Then $g_1' g_2 = \exp$ has the primitive $\exp$, so that (6.2) implies that f has a primitive, and that (for instance)

$$\int_0^x te^t dt = xe^x - \int_0^x e^t dt = xe^x - e^x.$$

(3) The function $f(x) = \log(x)$ on $]0, +\infty[$ has a primitive. Here, we use a trick and write $f(x) = g_1' g_2$ with $g_1(x) = x$ and $g_2(x) = \log(x)$. Then $g_1(x)g_2'(x) = x \cdot x^{-1} = 1$, which has a primitive, equal to x, so we can deduce by (6.2) that

$$(6.3) \qquad \int_1^x \log(t)dt = x\log(x) - \int_1^x dt = x\log(x) - (x - 1).$$

We then have the application of the chain rule to primitives.

PROPOSITION 6.1.7. *Let I and J be intervals of $\mathbf{R}$ and let $h\colon I \to J$ and $g\colon J \to \mathbf{R}$ be functions such that h is differentiable. If g has a primitive, then the function $h' \cdot (g \circ h)$ has a primitive on I, and for any $x_0 \in I$ and $x \in I$, we have*

$$(6.4) \qquad \int_{x_0}^x h'(t)g(h(t))dt = \int_{h(x_0)}^{h(x)} g(t)dt.$$

PROOF. Let f be a primitive of g. Then we have by the Chain Rule

$$(f \circ h)' = h' \ (f' \circ h) = h' \ (g \circ h),$$

so that $h' \ (g \circ h)$ has a primitive, namely $f \circ h$. This means also that

$$\int_{x_0}^{x} h'(t)g(h(t))dt = f(h(x)) - f(h(x_0)) = \int_{h(x_0)}^{h(x)} g(t)dt.$$

$\square$

EXAMPLE 6.1.8. (1) Let $f(x) = xe^{x^2}$ for $x \in \mathbf{R}$. Then f has a primitive, because we can write $f(x) = \frac{1}{2}h'(x)g(h(x))$, where $h(x) = x^2$ and $g(x) = e^x$; since g has the primitive exp, then we get by (6.4) the formula

$$\int_0^x te^{t^2}dt = \frac{1}{2}\int_{0^2}^{x^2} e^t dt = \frac{1}{2}(e^{x^2} - 1).$$

(2) The formula (6.4) is called "change of variable formula"[48]. It is usually applied by starting from the left-hand side

$$\int_{x_0}^x h'(t)g(h(t))dt,$$

and saying "we make the change of variable $u = h(t)$, with $du = h'(t)dt$" (which is a formal way to remember the procedure), and then

$$\int_{x_0}^x h'(t)g(h(t))dt = \int_{h(x_0)}^{h(x)} g(u)du.$$

The factor $h'(t)$ might not be obvious, and one might need to perform some computations to see it.

The formula is often applied "starting from the right-hand side", namely from

$$\int_{y_0}^y g(t)dt$$

for some y_0 and y. If we put $t = h(u)$ where h is bijective, then

$$\int_{y_0}^y g(t)dt = \int_{x_0}^x g(h(u))h'(u)du$$

with x_0 such that $h(x_0) = y_0$ and $h(x) = y$.

(3) Consider

$$\int_{x_0}^x \frac{1}{t(\log(t))}dt$$

with $1 < x_0 < x$. We put $u = \log(t)$, do that $du = t^{-1}dt$, and the integral becomes

$$\int_{x_0}^x \frac{1}{t(\log(t))}dt = \int_{\log(x_0)}^{\log(x)} \frac{1}{u}du = \log(\log(x)) - \log(\log(x_0)).$$

(4) Consider

$$\int_{x_0}^x \sqrt{1 - t^2}dt$$

136

where $-1 \leqslant x_0 \leqslant x \leqslant 1$. It is natural to want to see t as the cosine of some angle θ. So we put $t = \cos(\theta)$ with $0 \leqslant \theta \leqslant \pi$; then $\sqrt{1 - t^2} = \sqrt{1 - \cos(\theta)^2} = \sin(\theta)$ and $dt = -\sin(\theta)d\theta$, which gives

$$\int_{x_0}^x \sqrt{1 - t^2}dt = -\int_{\arccos(x_0)}^{\arccos(x)} \sin(\theta)^2 d\theta.$$

The primitive of $\sin(\theta)^2$ can be computed by reducing to cosine and sine of multiples of θ, as in Example 4.5.5, (2). We find

$$\sin(\theta)^2 = \frac{1 - \cos(2\theta)}{2},$$

which has primitive $\theta/2 - \sin(2\theta)/4$, so that

$$\int_{x_0}^x \sqrt{1 - t^2}dt = -\frac{1}{2}\int_{\arccos(x_0)}^{\arccos(x)} (1 - 2\cos(2\theta))d\theta = -\left[\frac{\theta}{2} - \frac{\sin(2\theta)}{4}\right]_{\arccos(x_0)}^{\arccos(x)}.$$

(5) An important case of the change of variable formula is

(6.5) $$\int_{x_0}^x g(at + b)dt = \frac{1}{a}\int_{ax_0+b}^{ax+b} g(t)dt$$

for $a \neq 0$ and $b \in \mathbf{R}$; here $u = at + b$ with $du = adt$.

Among the properties of primitives, we highlight one in particular:

PROPOSITION 6.1.9. *Let I be an interval and $g \colon I \to \mathbf{R}$ a bounded function which has a primitive on I. Then for any x_0 and x in I, we have*

$$\left|\int_{x_0}^x g(t)dt\right| \leqslant M|x - x_0|$$

where M is such that $|g(t)| \leqslant M$ for all t between x_0 and x.

PROOF. We can assume that $x \neq x_0$, since otherwise both sides of the inequality are equal to 0. Let f be a primitive of g. Then

$$\int_{x_0}^x g(t)dt = f(x) - f(x_0).$$

Since f is differentiable on I, the Mean-Value Theorem implies that there exists c between x and x_0 such that

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(c) = g(c),$$

and therefore

$$\left|\int_{x_0}^x g(t)dt\right| = |f(x) - f(x_0)| = |g(c)| \, |x - x_0| \leqslant M|x - x_0|.$$

$\square$

6.2. The Riemann Integral

Reference: [2, 11.1, 11.2, 11.3, 11.4, 11.5].

In Section 6.1, we have formally defined the integral or primitives, but we can only compute these using fairly constrained rules. It turns out, by experience, that for many functions, even simple ones like $f(x) = e^{x^2}$ or $f(x) = e^{-x^2}$, one never seems to be able to find a primitive of f. This is not a sign of a lack of imagination: one can actually prove that a primitive of $f(x) = e^{x^2}$ cannot be expressed as an "elementary" function.
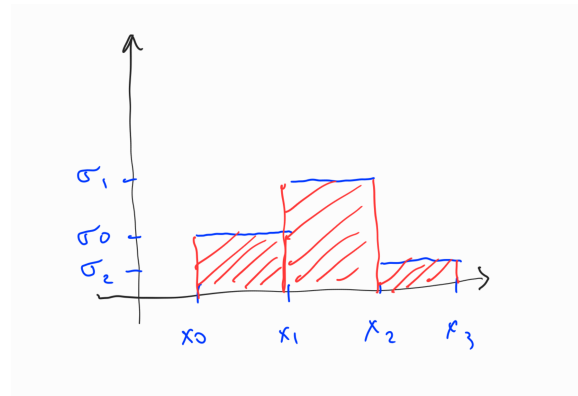
FIGURE 6.1. A step-function and its integral

Nevertheless, this function, and in fact any continuous function, has a primitive. This must however be constructed in a completely different manner than by applying simple rules, similar to those for computing derivatives.

One motivation for the construction can be described as follows: one can deduce from Proposition 6.1.9 that if we have a sequence of functions (g_n) which converge uniformly on I to a function g, and which all have primitives f_n, then the limit function g also has a primitive, and moreover the f_n (normalized to take value 0 at some x_0) converge uniformly to the primitive of g with value 0 at x_0. So it suffices to find "sufficiently many" functions with primitives to approximate arbitrarily a given function in order to obtain its primitives.

One could work with polynomials, which have primitives and can approach any continuous function on a compact interval $[a, b]$ (as we mentioned before see Remark 5.2.1), but it is more customary to use even simpler function to define the Riemann integral.

DEFINITION 6.2.1 (Step-function). Let $I = [a, b]$ be a compact interval with $a < b$. A function $s \colon I \to \mathbf{R}$ is a *step-function*[49] on I if there exist $k \in \mathbf{N}$ and numbers
$$a = x_0 < x_1 < \cdots < x_k = b$$
such that s is constant, equal to some $\sigma_i \in \mathbf{R}$, on $]x_i, x_{i+1}[$ for all i.

The *integral* of s over $[a, b]$ is the real number
$$\int_a^b s(t)dt = \sum_{i=0}^{k-1} \sigma_i(x_{i+1} - x_i).$$

EXAMPLE 6.2.2. (1) If $s \colon I \to \mathbf{R}$ is constant, equal to $c \in \mathbf{R}$, then it is a step-function and
$$\int_a^b s(t)dt = c(b - a).$$
Note that this is compatible with the previous notation since s has the function $f(x) = cx$ as primitive. Note also, however, that a step-function is in general *not* continuous on $[a, b]$.

(2) One should check that the value of the integral does not depend on the choice of the points x_i used to see that s is a step-function. However, this is elementary, and can be seen in the case of (1), where the computation does not require knowing that $x_0 = a$ and $x_1 = b$: for any
$$a = x_0 < x_1 < \cdots < x_k = b$$

138

we have $s(x) = c$ on $]x_i, x_{i+1}[$ and

$$\sum_{i=0}^{k-1} s_i(x_{i+1} - x_i) = c(x_1 - x_0 + x_2 - x_1 + \cdots + x_k - x_{k-1}) = c(x_k - x_0) = c(b - a).$$

In the general case, one can also observe that the set X of values of a step function s is finite. For each $\sigma \in X$, the set of $x \in I$ where $s(x) = \sigma$ is a finite union of intervals. If we denote by L_σ the sum of the lengths of these intervals, then

$$\int_a^b s(t)dt = \sum_{\sigma \in X} \sigma L_\sigma,$$

and since the right-hand side does not depend on the points (x_i), the same is true for the left-hand side.

REMARK 6.2.3. A very important remark for the applications of the integral is that if s is a *non-negative* step-function on I, then the integral defined above has a geometric interpretation: it is the *area* of the subset of the plane defined by

$$C_s = \{(x, y) \in \mathbf{R}^2 \mid a \leqslant x \leqslant b, \quad 0 \leqslant y \leqslant s(x)\},$$

which is the part of the plane between the x-axis and the graph of s. Indeed, the graph of s is a sequence of horizontal segments, and the set C_s is a union of rectangles with the horizontal side of length $x_{i+1} - x_i$ and the vertical side of length σ_i. So $\sigma_i(x_{i+1} - x_i)$ is the area of ach individual rectangle, and the sum is the total area of C_s.

PROPOSITION 6.2.4. *Let $I = [a, b]$ with $a < b$.*
(1) If s_1 and s_2 are step-functions on I then so is $s_1 + s_2$ and

$$\int_a^b (s_1(t) + s_2(t))dt = \int_a^b s_1(t)dt + \int_a^b s_2(t)dt.$$

(2) If s is a step-function on I, then it is bounded, the function $|s|$ is a step-function and

$$\left| \int_a^b s(t)dt \right| \leqslant \int_a^b |s(t)|dt \leqslant M(b - a),$$

where M is such that $|s(t)| \leqslant M$ for all $t \in I$.
(3) Let $c \in I$. For any step-function s on I, the restriction of s to $[a, c]$ and $[c, b]$ are step-functions on these intervals and

(6.6)
$$\int_a^b s(t)dt = \int_a^c s(t)dt + \int_c^b s(t)dt.$$

PROOF. (1) is clear if s_1 and s_2 are constant on the same intervals, and we can reduce to this case by taking the subdivision of $[a, b]$ given by ordering the union of the subdivision points for both functions.
(2) With the notation of the definition of step-functions, we get first

$$\left| \int_a^b s(t)dt \right| = \left| \sum_{i=0}^{k-1} \sigma_i(x_{i+1} - x_i) \right| \leqslant \sum_{i=0}^{k-1} |\sigma_i|(x_{i+1} - x_i) = \int_a^b |s(t)|dt,$$

and then

$$\int_a^b |s(t)|dt = \sum_{i=0}^{k-1} |\sigma_i|(x_{i+1} - x_i) \leqslant M \sum_{i=0}^{k-1} (x_{i+1} - x_i) = M(b - a).$$

(3) This is again straightforward, because we can always use a decomposition (x_i) of $[a, b]$ adapted to s such that $c = x_j$ for some j (by adding this point in any given decomposition). □

As we will show, and as we suggested, we can then obtain the integral of any function that can be uniformly approached by a sequence of step-functions. We first name such functions to avoid repeating frequently this condition.

DEFINITION 6.2.5 (Ruled functions). Let $I = [a, b]$ with $a < b$. A function $g : I \to \mathbf{R}$ is called a *ruled function*[50] on I if there exists a sequence (s_n) of step-functions on I that converges uniformly to g on I.

In order to see that this is a useful notion, we immediately show that continuous functions are ruled functions (but not conversely); this will imply that all results below that assume that a function is ruled apply to continuous functions in particular.

THEOREM 6.2.6. *Let $I = [a, b]$ with $a < b$. Let $g : I \to \mathbf{R}$ be a continuous function. For any $x \in [a, b]$, the restriction of g to $[a, x]$ is a uniform limit of step-functions.*

In order to prove this result, we need a property of continuous functions on compact intervals that is called *uniform* continuity: it states that in Definition 3.2.1 of a continuous function, we can take the δ that ensures that $|f(x) - f(y)| < \varepsilon$ when $|x - y| < \delta$ to be *independent* of x.

THEOREM 6.2.7 (Uniform continuity). *Let $I = [a, b]$ with $a < b$. Let $g : I \to \mathbf{C}$ be a continuous function. Then g is uniformly continuous, in the sense that for any $\varepsilon > 0$, there exists $\delta > 0$ such that $|x - y| < \delta$ implies $|g(x) - g(y)| < \varepsilon$.*

PROOF. We will use an argument by contradiction. So suppose that g is *not* uniformly continuous. This means that there exists a fixed $\varepsilon > 0$ such that, for any $\delta > 0$, we can find two elements x and y of I such that $|x - y| < \delta$ but $|g(x) - g(y)| \geqslant \varepsilon$. We apply this with $\delta = 1/n$ for $n \in \mathbf{N}$, and denote by x_n and y_n two elements of I that satisfy these conditions.

Since (x_n) is bounded, there exists a subsequence $(x_{n_k})_k$ that converges to some $x \in I$ (Theorem 2.9.3). The inequalities $|x_n - y_n| < 1/n$ imply that $y_{n_k} \to x$. Since g is continuous, we deduce that

$$|g(x_{n_k}) - g(y_{n_k})| \to |f(x) - f(x)| = 0.$$

But this contradicts the lower bounds

$$|g(x_{n_k}) - g(y_{n_k})| \geqslant \varepsilon > 0$$

for all k. □

PROOF OF THEOREM 6.2.6. We can assume that $x = b$. We construct the sequence (s_n) as follows.

According to the uniform continuity (Theorem 6.2.7 applied with $\varepsilon = 1/n$, and taking m so that $1/m < \delta$), for any $n \in \mathbf{N}$, there exists $m \in \mathbf{N}$ such that

$$|g(x) - g(y)| < \frac{1}{n}$$

if $|x - y| < 1/m$. We efine a step-function s_n using the subdivision

$$x_0 = a < x_1 = a + \frac{b-a}{k} < \cdots < x_{k-1} = a + (k-1)\frac{b-a}{k} < x_k = b,$$
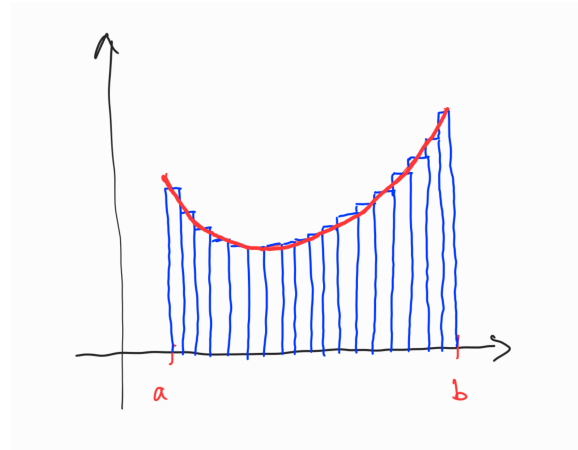
FIGURE 6.2. Approximation of a continuous function

and the values
$$\sigma_i = g\left(\frac{x_i + x_{i+1}}{2}\right),$$
and we specify furthermore that $s_n(x_i) = g(x_i)$.

Then by construction, we get for $x \in [x_i, x_{i+1}]$ that
$$|g(x) - s_n(x)| \leqslant \left|g(x) - g\left(\frac{x_i + x_{i+1}}{2}\right)\right| < \frac{1}{n},$$
and this implies that (s_n) converges uniformly to g on I. □

Now we construct the integral of ruled functions.

COROLLARY 6.2.8. *Let* $I = [a, b]$ *with* $a < b$. *Let* $g: I \to \mathbf{R}$ *be a ruled function, and* (s_n) *a sequence of step-functions on* I *that converges uniformly to* g.
(1) *The sequence* (x_n) *defined by*
$$x_n = \int_a^b s_n(t)dt$$
converges to a real number x.
(2) *The real number* x *does not depend on the choice of the sequence* (s_n) *of step-functions that converges uniformly to* g *on* I.
(3) *If* $|g(x)| \leqslant M$ *for all* x *in* $[a, b]$, *then the limit satisfies*
$$x \leqslant M(b - a).$$

PROOF. (1) We check that the sequence (x_n) is a Cauchy sequence. Indeed, for n and m in $\mathbf{N}$, we have
$$|x_n - x_m| = \left|\int_a^b (s_n(t) - s_m(t))dt\right| \leqslant (b - a)M_{n,m}$$
where $M_{n,m}$ is the supremum of $|s_n - s_m|$ on I. But the definition of uniform convergence, and the Cauchy Crtierion for uniform convergence precisely imply that $M_{n,m}$ is smaller than any given $\varepsilon > 0$ when n and m are both suitably large. So the sequence (x_n) is a Cauchy sequence of real numbers, which therefore converges.

(2) We use a common trick: if (t_n) is another sequence of step-functions that converges uniformly to g, then the sequence (u_n) defined by
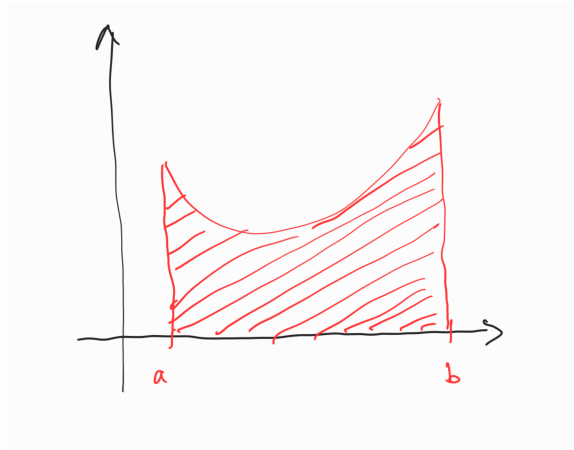$$u_{2n} = s_n, \qquad u_{2n+1} = t_n,$$

141

FIGURE 6.3. Definition of the area

is a third such sequence. According to (1), the corresponding integrals converge to some $x \in \mathbf{R}$; but we have subsequences

$$\int_a^b u_{2n}(t)dt = \int_a^b s_n(t)dt, \qquad \int_a^b u_{2n+1}(t)dt = \int_a^b t_n(t)dt,$$

and both must converge also to the same limit x, which implies the result.

(3) Let $k \in \mathbf{N}$. Since (s_n) converges to g uniformly on I, there exists $N \in \mathbf{N}$ such that $|s_n(t) - g(t)| \leqslant 1/k$ for all $t \in I$ and all $n \geqslant N$. Then we get

$$|s_n(t)| \leqslant M + \frac{1}{k}$$

by the triangle inequality, so that by Proposition 6.2.4, we get

$$\left| \int_a^b s_n(t)dt \right| \leqslant \int_a^b |s_n(t)|dt \leqslant M(b-a) + \frac{b-a}{k}$$

for all $n \geqslant N$. Since the left-hand side converges to $|x|$ as $n \to +\infty$, this implies that $|x| \leqslant M(b-a) + (b-a)/k$. But then, since k is arbitrary, letting $k \to +\infty$, we deduce that $|x| \leqslant M$. $\square$

DEFINITION 6.2.9. Let $I = [a,b]$ with $a < b$. Let $g \colon I \to \mathbf{R}$ be a ruled function. The number x defined in the corollary is called the *integral of g from a to b*, and denoted

$$\int_a^b g(t)dt.$$

With this notation, the third part of the corollary becomes the fundamental inequality

(6.7) $$\left| \int_a^b g(t)dt \right| \leqslant M(b-a)$$

for any ruled function g on $[a,b]$ such that $|g| \leqslant M$ on $[a,b]$.

Remark 6.2.3 gives the motivation for the next definition:

DEFINITION 6.2.10 (Area). Let $I = [a,b]$ with $a < b$ and $g \colon I \to \mathbf{R}$ a non-negative ruled function. The *area* of the set

$$C_g = \{(x,y) \in \mathbf{R}^2 \mid a \leqslant x \leqslant b, \quad 0 \leqslant y \leqslant g(x)\},$$

is *defined* to be equal to the integral of g over I.

Moreover, we also obtain the linearity properties of the integral:

PROPOSITION 6.2.11. *Let $I = [a, b]$ with $a < b$.*
(1) If $g_1, g_2 \colon I \to \mathbf{R}$ are ruled functions, then for any real numbers c and d, the function $cg_1 + dg_2$ is ruled, and we have

$$\int_a^b (cg_1(t) + dg_2(t))dt = c \int_a^b g_1(t)dt + d \int_a^b g_2(t)dt.$$

(2) If $c \in [a, b]$ and if g is a ruled function on I, then the restrictions of g to $[a, c]$ and $[c, b]$ are ruled and

(6.8)
$$\int_a^b g(t)dt = \int_a^c g(t)dt + \int_c^b g(t)dt.$$

PROOF. If g_1 is the limit of the sequence $(s_{1,n})$ of step-functions, and s_2 is the limit of the sequence $(s_{2,n})$ of step-functions, then the function $cg_1 + dg_2$ is the uniform limit of the step-functions $cs_{1,n} + ds_{2,n}$, so it is a ruled function. Using Proposition 6.2.4, we know that

$$\int_a^b (cs_{1,n}(t) + ds_{2,n}(t))dt = c \int_a^b s_{1,n}(t)dt + d \int_a^b s_{2,n}(t)dt$$

for all n; then the left-hand side converges to

$$\int_a^b (cg_1(t) + dg_2(t))dt$$

and the right-hand side to

$$c \int_a^b g_1(t)dt + d \int_a^b g_2(t)dt.$$

The result follows.

To prove (2), note that if (s_n) is a sequence of step0functions which converges uniformly to g on $[a, b]$, then the sequence of the restrictions of s_n to $[a, c]$ and $[c, b]$ converge uniformly to g on these intervals, and the formula

$$\int_a^b s_n(t)dt = \int_a^c s_n(t)dt + \int_c^b s_n(t)dt$$

for any n (see (6.6)) leads to the statement. $\qquad\square$

The crucial theorem is the following:

THEOREM 6.2.12 (Fundamental theorem of calculus). *Let $I = [a, b]$ with $a < b$. Let $g \colon I \to \mathbf{R}$ be a continuous function. The function f defined on $[a, b]$ by*

$$f(x) = \int_a^x g(t)dt$$

is differentiable on $[a, b]$ and is a primitive of g with $f(a) = 0$.

This theorem shows that the integral notation from Definition 6.2.9 is compatible with that of Section 6.1 whenever we have a function g with a known primitive.

PROOF. This is simpler than it looks. For $x_0 < x$ in I, we get by (6.8) the relation

$$f(x) - f(x_0) = \int_{x_0}^x g(t)dt = \int_{x_0}^x g(x_0)dt + \int_{x_0}^x (g(t) - g(x_0))dt.$$

The first term is equal to $(x - x_0)g(x_0)$. The second is small when x is close to x_0, because g is continuous. Precisely, for $\varepsilon > 0$, let $\delta > 0$ be such that

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

when $|x - x_0| < \delta$. Then for all such x, we get

$$\left| \int_{x_0}^{x} (g(t) - g(x_0))dt \right| \leqslant \varepsilon |x - x_0|$$

by (6.7), and therefore

$$\left| \frac{f(x) - f(x_0)}{x - x_0} - g(x_0) \right| = \left| \frac{1}{x - x_0} \int_{x_0}^{x} (g(t) - g(x_0))dt \right| \leqslant \varepsilon.$$

This shows that f has right-derivative $g(x_0)$, and a similar argument proves that the left-derivative exists and is also equal to $g(x_0)$. $\qquad \square$

The proof has the following corollaries:

> COROLLARY 6.2.13. *Let $I = [a, b]$ with $a < b$. Let $g \colon I \to \mathbf{R}$ be continuous. We have*
>
> $$\int_a^b g(t)dt = \lim_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} g\left(a + k\frac{b-a}{n}\right) = \lim_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n} g\left(a + k\frac{b-a}{n}\right).$$

The sums of the type

$$\frac{1}{n} \sum_{k=0}^{n-1} g\left(a + k\frac{b-a}{n}\right)$$

are called *Riemann sums* for the function g. This corollary shows that they provide an approximation to the integral of g.

PROOF. Since

$$\frac{1}{n} \sum_{k=0}^{n} g\left(a + k\frac{b-a}{n}\right) = \frac{1}{n} \sum_{k=0}^{n-1} g\left(a + k\frac{b-a}{n}\right) + \frac{g(b)}{n},$$

and $g(b)/n \to 0$ as $n \to +\infty$, the existence of either limit implies that both exist and are equal. So it suffices to check that

$$\int_a^b g(t)dt = \lim_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} g\left(a + k\frac{b-a}{n}\right).$$

But, for a given $n \in \mathbf{N}$, the right-hand side is the integral of a step-function s_n such that

$$s_n(x) = g\left(a + k\frac{b-a}{n}\right)$$

for $a + k(b-a)/n \leqslant x < a + (k+1)(b-a)/n$, where $0 \leqslant k \leqslant n - 1$. We then see that for any $x \in I$, we have

$$|g(x) - s_n(x)| \leqslant b_n,$$

where

$$b_n = \sup\{|f(s) - f(t)| \mid |s - t| \leqslant 1/n\}.$$

Theorem 6.2.7 (the uniform continuity of g) implies that $b_n \to 0$ as $n \to +\infty$, which means that (s_n) converges uniformly to g on I. By Corollary 6.2.8, (2), we have

$$\lim_{n \to +\infty} \int_a^b s_n(t)dt = \int_a^b g(t)dt$$

which gives the result. □

EXAMPLE 6.2.14. Consider the sequence defined by
$$a_n = \frac{1}{n}\Big(\cos(0) + \cos(\pi/n) + \cdots + \cos((n-1)\pi/n)\Big)$$
for $n \in \mathbf{N}$. By the corollary applied to $f(x) = \cos(\pi x)$ on $[0,1]$, we have
$$\lim_{n\to+\infty} a_n = \int_0^1 \cos(\pi t)dt = \frac{1}{\pi}(\sin(\pi) - \sin(0)) = 0.$$

COROLLARY 6.2.15. *Let $I = [a,b]$ with $a < b$. Let (g_n) be a sequence of ruled functions on I that converges uniformly to g, and such that g is ruled. Then*
$$\int_a^b g(t)dt = \lim_{n\to+\infty} \int_a^b g_n(t)dt.$$

PROOF. We have, by (6.7) and linearity, the bound
$$\left|\int_a^b g(t)dt - \int_a^b g_n(t)dt\right| = \left|\int_a^b (g(t) - g_n(t))dt\right| leq b_n(b-a)$$
where b_n is the supremum of the numbers $|g(t) - g_n(t)|$ for $t \in [a,b]$. The uniform convergence means that $b_n \to 0$ as $n \to +\infty$, and the conclusion follows. □

EXAMPLE 6.2.16. Let (a_n) be a sequence of real numbers such that the power series $\sum a_n x^n$ has positive radius of convergence R. Then for $x \in]-R, R[$, we have
$$\int_0^x \Big(\sum_{n=0}^{+\infty} a_n t^n\Big)dt = \sum_{n=0}^{+\infty} \frac{a_n}{n+1}x^{n+1}.$$

(1) For instance, consider the geometric series
$$\frac{1}{1-x} = \sum_{n=0}^{+\infty} x^n$$
for $|x| < 1$. We deduce that for $|x| < 1$, we have
$$\int_0^x \frac{1}{1-t}dt = \sum_{n=0}^{+\infty} \frac{x^{n+1}}{n+1}.$$
But the left-hand side, by the change of variable $u = 1 - t$ (see (6.5)), or by computing the derivative, is equal to $-\log(1-x)$. So we have
$$\log(1-x) = -\sum_{n=0}^{+\infty} \frac{x^{n+1}}{n+1} = -\sum_{n=1}^{+\infty} \frac{x^n}{n}$$
for $|x| < 1$. This is more precise than the result previously obtained in Example 5.7.7 using Taylor polynomials.

(2) As another example, from the geometric series applied to $-x^2$ with $|x| < 1$, we get
$$\frac{1}{1+x^2} = 1 - x^2 + x^4 + \cdots + (-1)^n x^{2n} + \cdots$$
for $|x| < 1$. Integrating, and using (5.2) and $\arctan(0) = 0$, we conclude that
$$\arctan(x) = x - \frac{x^3}{3} + \cdots + \frac{(-1)^n x^{2n+1}}{2n+1} -$$

for $|x| < 1$. In fact, one can prove that this formally is still valid for $x = 1$, where the series converges (it is an alternating series). From $\arctan(1) = \pi/4$, we deduce a series representation for π:

$$\pi = 4\Big(1 - \frac{1}{3} + \frac{1}{5} - \cdots\Big) = 4\sum_{n=0}^{+\infty} \frac{(-1)^n}{2n+1}.$$

Note however that this series converges very slowly; for instance the sum of the 10000 first terms is

$$3.14149265359004323845918383748153781\ldots$$

which is only correct up to about 10^{-4}. But one can get many better formulas; for instance, $\arctan(1/\sqrt{3}) = \pi/6$ (because $\sin(\pi/6) = 1/2$ and $\cos(\pi/2) = \sqrt{3}/2$), and therefore

$$\frac{\pi}{6} = \sum_{n=0}^{+\infty} (-1)^n \frac{3^{-(2n+1)/2}}{2n+1} = \frac{1}{\sqrt{3}} \sum_{n=0}^{+\infty} \frac{(-1)^n}{3^n(2n+1)}.$$

The series converges now much faster since its terms go to 0 faster than $1/3^n$. Just the first ten terms give the approximation

$$3.141590510938080099642754229944255043 7\ldots$$

of π, which is correct up to approximately $2 \cdot 10^{-6}$.

One can find even better formulas (involved smaller numbers that $1/\sqrt{3}$, so that the power series has even smaller coefficients). For instance, Machin noticed in 1706 that

$$\frac{\pi}{4} = 4\arctan\Big(\frac{1}{5}\Big) - \arctan\Big(\frac{1}{239}\Big),$$

which he used to compute (by hand) the first hundred decimal digits of π. The formula itself can be obtained from the computation

$$(5+i)^4(-239+i) = -4 \cdot 13^4 \cdot (1+i),$$

by observing that using (5.3) and $\arctan(-x) = -\arctan(x)$, we have

$$5+i = (26)^{1/2}\, e^{i\arctan(1/5)}, \quad -239+i = (239^2+1)^{1/2}\, e^{-i\arctan(1/239)}$$

$$1+i = 2^{1/2}\, e^{i\pi/4}.$$

As another application Corollary 6.2.15, we now prove Theorem 5.2.2 about the differentiability of functions obtained as limits of functions.

PROOF OF THEOREM 5.2.2. The assumption is that we have an interval I in $\mathbf{R}$ and a sequence (f_n) of functions of class C^1 on I which converge uniformly to a function f, and for which (f_n') converges uniformly to a function g. We want to deduce that f is differentiable with derivative $f' = g$.

Because this is a local question, we can assume that $I = [a, b]$ for some $a < b$ in I. Then we note that since (f_n') converges uniformly to g, Corollary 6.2.15 shows that

$$\int_a^x f_n'(t)dt = f_n(x) - f_n(a)$$

converges uniformly on I to

$$\int_a^x g(t)dt.$$

On the other hand, $f_n(x) - f_n(a)$ converges uniformly, by assumption, to $f(x) - f(a)$. So we get

$$f(x) - f(a) = \int_a^x g(t)dt$$

for $x \in [a, b]$. The right-hand side function is differentiable on I with derivative g by the Fundamental Theorem of Calculus, and the result follows. $\quad\square$

Finally, it is convenient to define as follows the integral from a to b even if $a > b$.

DEFINITION 6.2.17. Let I be an interval and let $g \colon I \to \mathbf{R}$ be a function. Let a and b be elements of I with $a > b$. If g is ruled on $[b, a]$, then we define

$$\int_a^b g(t)dt = -\int_b^a g(t)dt.$$

We also define

$$\int_a^a g(t)dt = 0.$$

With this definition one sees that

$$\int_a^b g(t)dt = \int_a^c g(t)dt + \int_c^b g(t)dt$$

holds for all a, b, c in I, if g is ruled on the corresponding intervals (for instance, if g is continuous on I). And if g is continuous and f is a primitive of g, then we have

$$\int_a^b g(t)dt = f(b) - f(a)$$

in all cases.

6.3. Properties and applications of the integral

First, Theorem 6.2.12 means that in Propositions 6.1.2, 6.1.4 and 6.1.7, whenever there is an assumption that a function g has a primitive, we can apply the result if g is continuous. We state the formulas for integration by parts and change of variable for ease of reference:

- (Integration by parts) If g_1 and g_2 are in $C^1(I)$, then

(6.9) $$\int_{x_0}^x g_1'(t)g_2(t)dt = g_1(x)g_2(x) - g_1(x_0)g_2(x_0) - \int_{x_0}^x g_1(t)g_2'(t)dt$$

  for any $x_0 \in I$ and $x \in I$. (The assumption implies that both integrals exist, as integrals of continuous functions.)
- (Change of variable) If g is continuous and $h \in C^1(I)$, then for any $x_0 \in I$ and $x \in I$, we have

(6.10) $$\int_{x_0}^x h'(t)g(h(t))dt = \int_{h(x_0)}^{h(x)} g(t)dt.$$

EXAMPLE 6.3.1. We compute now the area $A(R)$ of a disc D_R of radius $R > 0$ centered at 0 in the plane. This is twice the area of the half-disc D_R^+ with the same radius and center, which contains the points of the disc with non-negative y-coordinate. Then D_R^+ is also the set

$$D_R^+ = \{(x, y) \in \mathbf{R}^2 \mid y \geqslant 0 \text{ and } \sqrt{x^2 + y^2} \leqslant R\},$$

which can be described also as

$$D_R^+ = \{(x, y) \in \mathbf{R}^2 \mid 0 \leqslant y \leqslant g(x)\}$$

where $g \colon [-R, R] \to \mathbf{R}$ is defined by $g(x) = \sqrt{R^2 - x^2}$. By the Definition 6.2.10, this means that

$$A(R) = 2 \int_{-R}^{R} \sqrt{R^2 - t^2} dt.$$

We simplify the integral by making the change of variable $t = Ru$, so that $dt = Rdu$, and

$$A(R) = 2R^2 \int_{-1}^{1} \sqrt{1 - u^2} du = R^2 A(1).$$

Applying Example 6.1.8, (4), with $\arccos(-1) = \pi$ and $\arccos(0) = 1$, we get

$$A(R) = R^2 A(1) = \pi R^2.$$

We next obtain some consequences of the definition that result easily from the corresponding properties for step-functions.

PROPOSITION 6.3.2. *Let $I = [a, b]$ with $a < b$.*
(1) If $g_1, g_2 \colon I \to \mathbf{R}$ are continuous functions on I and $g_1 \leqslant g_2$, then

(6.11)
$$\int_a^b g_1(t) dt \leqslant \int_a^b g_2(t) dt$$

(2) If $g \geqslant 0$ is continuous on I, then for $a \leqslant c \leqslant d \leqslant b$, we have

$$\int_c^d g(t) dt \leqslant \int_a^b g(t) dt$$

(3) If $g \geqslant 0$ is continuous on I, then

$$\int_a^b g(t) dt \geqslant 0,$$

with equality if and only if $g(x) = 0$ for all x.

PROOF. (1) is a consequence, for instance, of Corollary 6.2.13. It implies the first part of (3) by taking $g_1 = 0$, with integral 0, and $g_2 = g$.

To prove (2), we note that

$$\int_a^b g(t) dt = \int_a^c g(t) dt + \int_c^d g(t) dt + \int_d^b g(t) dt$$

by (6.8), and the first and third terms are non-negative by (1).

To prove the last part of (3), we assume that there exists $x_0 \in [a, b]$ such that $g(x_0) > 0$. Then by Lemma 3.2.7, (2), there exists an interval $[c, d]$ with $c < d$ containing x_0 such that $g(x) \geqslant \frac{1}{2} g(x_0)$ for $x \in [c, d]$. Then by (2), and then by (1) applied with the constant function $\frac{1}{2} g(x_0)$ on $[c, d]$, we get

$$\int_a^b g(t) dt \geqslant \int_c^d g(t) dt \geqslant (d - c) \frac{g(x_0)}{2} > 0.$$

$\square$

A common application of integrals is to obtain estimates for series, or their partial sums, as the following examples illustrates.

$\int_1^5 g(t)dt \le a_1 + a_2 + a_3 + a_4$

$\int_1^5 g(t)dt \ge a_2 + a_3 + a_4 + a_5$
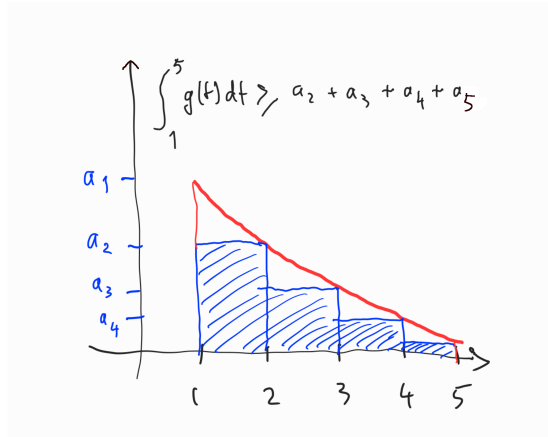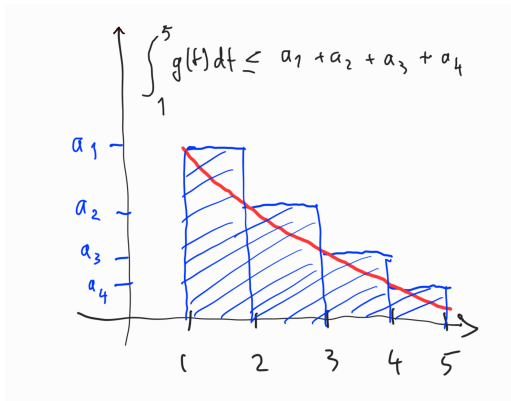
EXAMPLE 6.3.3. (1) Let $(a_n)_{n\in\mathbf{N}}$ be a sequence of real numbers. Suppose that there exists a continuous function $g\colon [1,+\infty[\to \mathbf{R}$ such that $a_n = g(n)$ for $n \in \mathbf{N}$, and suppose that g is non-increasing and non-negative (for instance, this applies to $a_n = 1/n^k$ where $k > 0$).

For $k \in \mathbf{N}$ and $t \in [k, k+1]$, we have then

$$a_{k+1} = g(k+1) \leqslant g(t) \leqslant a_k = g(k).$$

Integrating from k to $k+1$ (using (6.11)), this gives

$$\int_k^{k+1} a_{k+1}dt = a_{k+1} \leqslant \int_k^{k+1} g(t)dt \leqslant a_k.$$

Let $n \in \mathbf{N}$. Summing these inequalities from $k = 1$ to $k = n$, and using (6.6), we get

$$a_2 + \cdots + a_{n+1} \leqslant \int_1^{n+1} g(t)dt \leqslant a_1 + \cdots + a_n.$$

Take for instance $g(x) = 1/n^c$ where $c > 1$. We obtain

$$1 + \frac{1}{2^c} + \cdots + \frac{1}{n^c} \leqslant 1 + \int_1^{n+1} \frac{1}{t^c}dt$$

and

$$1 + \frac{1}{2^c} + \cdots + \frac{1}{n^c} \geqslant \int_1^{n+1} \frac{1}{t^c}dt.$$

Since $f(x) = x^{1-c}/(1-c)$ is a primitive of g, this leads to the inequalities

$$\frac{1}{c-1}\left(1 - \frac{1}{n^{c-1}}\right) \leqslant 1 + \frac{1}{2^c} + \cdots + \frac{1}{n^c} \leqslant 1 + \frac{1}{c-1}\left(1 - \frac{1}{n^{c-1}}\right)$$

for any $n \in \mathbf{N}$.

(2) Suppose now, in the opposite direction, that $a_n = g(n)$ where $g\colon [1,+\infty[\to \mathbf{R}_+$ is continuous and non-decreasing. Then we can still estimate sums like

$$s_n = a_1 + \cdots + a_n.$$

Indeed, for $k \in \mathbf{N}$ and $t \in [k, k+1]$, we now have

$$a_k = g(k) \leqslant g(t) \leqslant a_{k+1},$$

so

$$a_k \leqslant \int_k^{k+1} g(t)dt \leqslant a_{k+1},$$

and therefore, summing for k from 1 to n, we deduce that

$$s_n \leqslant \int_1^{n+1} g(t)dt \leqslant s_{n+1} - a_1.$$

As an illustration, let $g(x) = \log(x)$. A primitive is computed in (6.3), and since $a_n = \log(n)$ implies that $s_n = \log(n!)$, we get

$$\log(n!) \leqslant \int_1^{n+1} \log(t)dt = (n+1)\log(n+1) - n \leqslant \log((n+1)!),$$

or equivalently

$$n\log(n) - n + 1 \leqslant \log(n!) \leqslant (n+1)\log(n+1) - n.$$

This is quite a decent approximation since it is not very difficult to check that

$$\lim_{n \to +\infty} \frac{(n+1)\log(n+1) - n}{n\log(n) - n + 1} = 1.$$

Another application of the integral is a different form of the Taylor formula of Theorem 5.7.4.

> **THEOREM 6.3.4.** *Let $k \in \mathbf{N}_0$. Let $I \subset \mathbf{R}$ be an interval and let $f \colon I \to \mathbf{R}$ be a function that is in $C^{k+1}(I)$. Let $x_0 \in I$. For any $x \in I$, we have*
> $$f(x) = T_k f(x; x_0) + \frac{1}{k!}\int_{x_0}^x f^{(k+1)}(t)(x-t)^k dt.$$

Note that the integral exists since the function $f^{(k+1)}$ is continuous.

PROOF. For $k = 0$, this formula becomes

$$f(x) = f(x_0) + \int_{x_0}^x f'(t)dt,$$

which is correct since the integral is equal to $f(x) - f(x_0)$.

In order to finish the proof, we can therefore proceed by induction. We assume that the result holds for the $(k-1)$-st Taylor polynomial, so that

$$f(x) = T_{k-1}f(x; x_0) + \frac{1}{(k-1)!}\int_{x_0}^x f^{(k)}(t)(x-t)^{k-1}dt.$$

We evaluate the integral using (6.9): we write

$$f^{(k)}(t)(x-t)^{k-1} = g_1(t)g_2'(t)$$

where $g_1 = f^{(k)}$ and $g_2 = -(x-t)^k/k$, so that

$$\frac{1}{(k-1)!}\int_{x_0}^x f^{(k)}(t)(x-t)^{k-1}dt = \frac{1}{k!}f^{(k)}(x_0)(x-x_0)^k + \frac{1}{k!}\int_{x_0}^x f^{(k+1)}(t)(x-t)^k dt,$$

since $g_1'(x)g_2(x) = 0$. Since

$$T_{k-1}f(x; x_0) + \frac{1}{k!}f^{(k)}(x_0)(x-x_0)^k = T_k f(x; x_0),$$

we obtain the stated formula. $\square$

EXAMPLE 6.3.5. Combined with inequalities like (6.7), this form of the Taylor formula is often more useful than Theorem 5.7.4. For instance, consider $f(x) = \log(1 + x)$ as in Example 5.7.7. Taking $x_0 = 0$, we get

$$\log(1 + x) = x - \frac{x^2}{2} + \cdots + (-1)^{k-1}\frac{x^k}{k} + \frac{(-1)^k}{k!}\int_0^x \frac{k!}{(1+t)^{k+1}}(x - t)^k dt$$

for any $k \in \mathbf{N}$, using the formula (5.6). We consider only $x = 1$; then the remainder is

$$(-1)^k \int_0^1 \frac{(1 - t)^k}{(1 + t)^{k+1}} dt.$$

Since $|1 - t| \leqslant 1$ for $0 \leqslant t \leqslant 1$, and moreover $-(1 + t)^{-k}/k$ is a primitive of $(1 + t)^{-k-1}$, we deduce that the size of the remainder is at most $1/k$. Consequently, we have

$$f(1) = \log(2) = \sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$$

6.4. Some standard integrals

We summarize some basic integrals that are often useful. When we give a primitive, we just write $\int g(t)dt$ for a particular primitive of g.

1. An important change of variable We have

$$\int_{x_0}^x f(at + b)dt = \frac{1}{a}\int_{ax_0+b}^{ax+b} f(u)du$$

for $a \neq 0$ and $b \in \mathbf{R}$. This is used very often to simplify certain computations (for instance, to replace an e^{at} by e^t, where the primitive is more obvious).

2. Elementary functions. We have

$$\int e^t dt = e^t, \qquad \int t^a dt = \frac{1}{1 + a}t^{a+1} \quad (a \neq -1), \qquad \int \frac{1}{t}dt = \log(t)$$

$$\int \cos(t)dt = \sin(t), \qquad \int \sin(t)dt = -\cos(t).$$

3. Reciprocal functions. We have

$$\int \frac{1}{\sqrt{1 - t^2}}dt = \arcsin(t), \qquad -\int \frac{1}{\sqrt{1 - t^2}}dt = \arccos(t),$$

$$\int \frac{1}{1 + t^2}dt = \arctan(t).$$

We also explain a few standard integration techniques.

4. Powers times exponential or trigonometric functions. Integrals of the form

$$\int_a^b t^k \cos(ct)dt, \qquad \int_a^b t^k \sin(ct)dt, \qquad \int_a^b t^k e^{ct}dt,$$

with $k \in \mathbf{N}_0$ and $c \in \mathbf{R}$ can be computed by induction using integration by parts, where one differentiates t^k and integrates the trigonometric or exponential function. This needs to be repeated until the integral only involves the exponential or trigonometric function, at which point we know a primitive.

EXAMPLE 6.4.1. To compute
$$\int_0^x t^2 e^{3t} dt,$$
we integrate by parts twice:
$$\int_0^x t^2 e^{3t} dt = \frac{1}{3} x^2 e^{3x} - \frac{2}{3} \int_0^x t e^{3t} dt$$
$$= \frac{1}{3} x^2 e^{3x} - \frac{2}{3} \left(\frac{1}{3} x e^{3x} - \frac{1}{3} \int_0^x e^{3t} dt \right)$$
$$= e^{3x} \left(\frac{x^2}{3} - \frac{2x}{9} + \frac{2}{27} \right).$$

5. Trigonometric and exponential functions. Integrals of the form
$$\int_a^b \cos(rt) e^{st} dt, \qquad \int_a^b \sin(rt) e^{st} dt$$
with r and $s \in \mathbf{R}$ can be computed by doing two integration by parts to obtain a linear equation for the integral.

EXAMPLE 6.4.2. Let $s \neq 0$ and
$$I = \int_0^x \cos(rt) e^{st} dt.$$
By integrating by parts one, differentiating the cosine, we get
$$I = \frac{\cos(rx) e^{sx} - 1}{s} + \frac{r}{s} \int_0^x \sin(rt) e^{st} dt.$$
We integrate by parts the second integral, differentiating the sine, and get
$$I = \frac{\cos(rx) e^{sx} - 1}{s} + \frac{r \sin(rx) e^{sx}}{s^2} - \frac{r^2}{s^2} \int_0^x \cos(rt) e^{st} dt.$$
The integral is the same as I, so that
$$\left(1 + \frac{r^2}{s^2} \right) I = \frac{\cos(rx) e^{sx} - 1}{s} + \frac{r \sin(rx) e^{sx}}{s^2}.$$

6. Products of trigonometric functions. Integrals of the form
$$\int_a^b \cos(rt) \cos(st) dt, \qquad \int_a^b \cos(rt) \sin(st) dt, \qquad \int_a^b \sin(rt) \sin(st) dt,$$
with r and s in $\mathbf{R}$ can also be computed by two integration by parts.

EXAMPLE 6.4.3. Let $s \neq 0$ and
$$I = \int_0^x \cos(rt) \sin(st) dt.$$
By integrating by parts one, differentiating the cosine, we get
$$I = \frac{-\cos(rx) \cos(sx) + 1}{s} - \frac{r}{s} \int_0^x \sin(rt) \cos(st) dt.$$
We integrate by parts the second integral, differentiating the sine, and get
$$I = \frac{-\cos(rx) \cos(sx) + 1}{s} - \frac{r \sin(rx) \sin(sx)}{s^2} + \frac{r^2}{s^2} \int_0^x \cos(rt) \sin(st) dt,$$

so
$$\left(1 - \frac{r^2}{s^2}\right)I = \frac{1 - \cos(rx)\cos(sx)}{s} - \frac{r\sin(rx)\sin(sx)}{s^2}.$$

7. Powers of trigonometric functions. Integrals of the form
$$\int_a^b \cos(rt)^k dt, \qquad \int_a^b \sin(rt)^k dt,$$
with $k \in \mathbf{N}_0$ and $r \in \mathbf{R}$ can be computed expressing $\cos(x)^k$ as linear combination of $\cos(mx)$ and $\sin(nx)$ for suitable integers m and n (see Example 4.5.5, (2) for the basic method).

EXAMPLE 6.4.4. We compute $\int_0^x \sin(t)^3 dt$ by writing
$$\sin(t)^3 = \left(\frac{e^{it} - e^{-it}}{2i}\right)^3 = -\frac{1}{4}(\sin(3t) - 3\sin(t)),$$
getting
$$\int_0^x \sin(t)^3 dt = \frac{1}{12}(\cos(3x) - 1) - \frac{3}{4}(\cos(x) - 1).$$

8. Orthogonality relations. Let n and m be in $\mathbf{Z}$. We can then compute using the previous examples that
$$\int_0^{2\pi} \cos(nt)\sin(mt)dt = 0,$$
$$\int_0^{2\pi} \cos(nt)\cos(mt)dt = 0 \text{ if } n \neq m,$$
$$\int_0^{2\pi} \sin(nt)\sin(mt)dt = 0 \text{ if } n \neq m,$$
$$\int_0^{2\pi} \cos(nt)^2 dt = \pi \text{ if } n \neq 0,$$
$$\int_0^{2\pi} \sin(nt)^2 dt = \pi \text{ if } n \neq 0.$$

These formulas are fundamental in *Fourier analysis*.

9. Rational functions. If
$$g(x) = \frac{p(x)}{q(x)}$$
where p and q are polynomials, then one can always compute a primitive of g in terms of rational functions, polynomials, logarithms of polynomials and arctan of polynomials. Note that conversely each of these functions have indeed derivatives that are rational functions; for instance
$$\arctan(x^3 + 2)' = \frac{3x^2 + 1}{1 + (x^3 + 2)^2}, \qquad \log(x^4 + 2x + 1) = \frac{4x^3 + 2}{x^4 + 2x + 1}.$$

The method uses two steps:

- (Reduction to simple terms) One expresses the rational function as a sum of simpler functions of the following forms
$$\text{(polynomials)}, \qquad \frac{ax + b}{(\alpha x^2 + \beta x + \gamma)^k}, \qquad \frac{a}{(\alpha x + \beta)^k},$$

where $k \in \mathbf{N}$ and, in the first case, $\beta^2 - 4\alpha\gamma < 0$, and $\alpha \neq 0$ in the second case.

The fact that this reduction is always possible is really part of linear algebra. In simple cases, one can find the corresponding simple expressions by trial and error using "unknown coefficients".

- (Integration of simple terms) There remains to integrate each of the special rational functions. For polynomials, this is of course elementary. For

$$\int \frac{1}{(\alpha t + \beta)^k} dt$$

one performs the change of variable $u = \alpha t + \beta$ to reduce to $\int u^{-k} du$, which is known.

For the second, we have two basic cases

$$\int \frac{t}{(\alpha t^2 + \beta t + \gamma)^k} dt, \qquad \int \frac{1}{(\alpha t^2 + \beta t + \gamma)^k} dt.$$

Since

$$\alpha x^2 + \beta x + \gamma = \alpha \left(\left(x + \frac{\beta}{2\alpha} \right)^2 + \frac{\gamma}{\alpha} - \frac{\beta^2}{4\alpha^2} \right)$$

we can use substitutions to reduce to

$$\int \frac{at}{(t^2 + 1)^k} dx, \qquad \int \frac{b}{(t^2 + 1)^k} dt.$$

The first of these can be computed by the substitution $u = t^2 + 1$, $du = 2t\, dt$:

$$\int_a^b \frac{t}{(t^2 + 1)^k} dt = \frac{1}{2} \int_{a^2}^{b^2} \frac{1}{u^k} du = \begin{cases} \frac{1}{2} \log(b^2/a^2) & \text{if } k = 1 \\ \frac{1}{2(k+1)} (b^{-2(k+1)} - a^{-2(k+1)}) & \text{if } k \neq 1. \end{cases}$$

There remains to deal with

$$I_k = \int_a^b \frac{1}{(t^2 + 1)^k} dt.$$

Since we can compute $I_1 = \arctan(b) - \arctan(a)$, it suffices to find a relation between I_k and I_{k+1}. Using integration by parts, we get

$$I_k = \frac{b}{(b^2 + 1)^k} - \frac{a}{(a^2 + 1)^k} + 2k \int_a^b \frac{t^2}{(t^2 + 1)^{k+1}} dt.$$

Since

$$\int_a^b \frac{t^2}{(t^2 + 1)^{k+1}} dt = I_k - I_{k+1},$$

we get

$$I_k = \frac{b}{(b^2 + 1)^k} - \frac{a}{(a^2 + 1)^k} + 2k I_k - 2k I_{k+1},$$

or

$$I_{k+1} = \left(1 - \frac{1}{2k} \right) I_k + \frac{b}{(b^2 + 1)^k} - \frac{a}{(a^2 + 1)^k}.$$

Here is an example done in detail to see these steps implemented.

We want to compute

$$f(x) = \int_1^x \frac{t^2 + t}{6t^3 - t^2 + t - 1} dt.$$

To determine which values of x are allowed, and to implement the first step, we need to factor the denominator to see where it vanishes. It turns out that

$$6t^3 - t^2 + t - 1 = (2t - 1)(3t^2 + t + 1),$$

and since the discriminant of the quadratic polynomial $3t^2 + t + 1$ is $1 - 4 \cdot 3 = -11$, it has no real roots. So the rational function

$$g(x) = \frac{x^2 + x}{6x^3 - x^2 + x - 1}$$

is continuous on the intervals $]1/2, +\infty[$ and $] -\infty, 1/2[$. We can then define f by the integral above for $x \geqslant 1$ for instance.

Next, the decomposition of g should be of the form

$$g(x) = \frac{\alpha}{2x - 1} + \frac{\beta x + \gamma}{3x^2 + x + 1}.$$

We can find the coefficent α easily by multiplying both sides by $2x - 1$ and taking the limit as $x \to 1/2$: the right-hand side converges to α, and so

$$\alpha = \lim_{x \to 1/2} \frac{(2x - 1)(x^2 + x)}{6x^3 - x^2 + x - 1} = \lim_{x \to 1/2} \frac{x^2 + x}{3x^2 + x + 1} = \frac{3/4}{9/4} = \frac{1}{3}.$$

We can find γ for instance by putting $x = 0$: we get

$$0 = g(0) = -\alpha + \gamma \quad \text{so} \quad \gamma = \alpha = 1/3.$$

To compute β, we could evaluate the value at (say) $x = 1$, for we can compute the limit of $xg(x)$ as $x \to +\infty$: we find

$$\frac{1}{6} = \lim_{x \to +\infty} xg(x) = \lim_{x \to +\infty} \left(\frac{\alpha x}{2x - 1} + \frac{\beta x^2 + \gamma x}{3x^2 + x + 1} \right) = \frac{\alpha}{2} + \frac{\beta}{3} = \frac{1}{6} + \frac{\gamma}{3},$$

so that $\gamma = 0$. We therefore get

$$g(x) = \frac{1}{3} \left(\frac{1}{2x - 1} + \frac{1}{3x^2 + x + 1} \right)$$

(which can be checked to be correct by reducing the same denominator), so that

$$f(x) = \frac{1}{3} \int_1^x \frac{1}{2t - 1} dt + \frac{1}{3} \int_1^x \frac{1}{3t^2 + t + 1} dt$$

for $x \geqslant 1$.

In the first integral, we make the substitution $u = 2t - 1$, so $t = (u + 1)/2$ and $dt = \frac{1}{2} du$, and (since $u = 1$ for $t = 1$) we get

$$\frac{1}{3} \int_1^x \frac{1}{2t - 1} dt = \frac{1}{6} \int_1^{2x-1} \frac{1}{u} du = \frac{1}{6} \log(2x - 1).$$

In the second integral, we know that we want to reduce to an arctangent integral. We "complete the square" in the denominator:

$$3t^2 + t + 1 = 3\left(t^2 + \frac{1}{3}t + \frac{1}{3} \right) = 3\left(\left(t + \frac{1}{6} \right)^2 + \frac{1}{3} - \frac{1}{6^2} \right) = 3\left(\left(t + \frac{1}{6} \right)^2 + \frac{11}{36} \right).$$

If we make first the substitution $u = t + 1/6$, do $du = dt$, we get

$$\frac{1}{3} \int_1^x \frac{1}{3t^2 + t + 1} dt = \frac{1}{9} \int_{7/6}^{x+1/6} \frac{1}{u^2 + 11/36} du.$$

155

In order to reach finally the proper form to get an arctangent, we write the denominator in the form

$$u^2 + \frac{11}{36} = \frac{11}{36}\left(\frac{36u^2}{11} + 1\right) = \frac{11}{36}\left(\left(\frac{6u}{\sqrt{11}}\right)^2 + 1\right).$$

So if we make the subsitution $v = 6cu$ with $c = 1/\sqrt{11}$, so that $du = \frac{1}{6c}dv$, we get

$$\frac{1}{9}\int_{7/6}^{x+1/6} \frac{1}{u^2 + 11/36}du = \frac{1}{9} \times \frac{36}{11} \times \frac{1}{6c}\int_{7c}^{(6x+1)c} \frac{1}{v^2+1}dv$$

$$= \frac{2}{3\sqrt{11}}\left(\arctan(\tfrac{6}{\sqrt{11}}x + \tfrac{1}{\sqrt{11}}) - \arctan(\tfrac{7}{\sqrt{11}})\right).$$

Finally, combining everything, we deduce that

$$f(x) = \frac{1}{6}\log(2x - 1) + \frac{2}{3\sqrt{11}}\left(\arctan(\tfrac{6}{\sqrt{11}}x + \tfrac{1}{\sqrt{11}}) - \arctan(\tfrac{7}{\sqrt{11}})\right)$$

for $x \geqslant 1$.

6.5. Improper integrals

In many applications, one is interested in a generalization of the integral where the interval of integration is not necessarily compact. Examples are

$$\int_0^{+\infty} e^{-t}t^n dt, \qquad \int_{-1}^1 \frac{1}{\sqrt{1-t^2}}dt$$

where, in the first case, the interval is unbounded, and in the second, it is really $]-1, 1[$ that is involved since the function that one integrates is not defined at -1 and 1.

Such integrals are called *improper integrals*.[51] They are defined, as one might expect, using limits.

DEFINITION 6.5.1. Let $a \in \mathbf{R}$ and let $I = [a, +\infty[$. Let $g\colon I \to \mathbf{R}$ be a continuous function. We say that g has an (improper) integral over I if the limit

$$\lim_{x\to+\infty}\int_a^x g(t)dt$$

exists. Its value is called the integal of g over I, and is denoted

$$\int_a^{+\infty} g(t)dt.$$

A similar definition applies to improper integrals on $]-\infty, b]$, namely

$$\int_{-\infty}^b g(t)dt = \lim_{x\to-\infty}\int_x^b g(t)dt,$$

and, if g is continuous on $]a, b]$ or $[a, b[$, but not defined at a or b, we define

$$\int_a^b g(t)dt = \lim_{\substack{x\to a \\ x>a}}\int_x^b g(t)dt, \qquad \int_a^b g(t)dt = \lim_{\substack{x\to b \\ x<b}}\int_a^x g(t)dt.$$

On the other hand, for improper integrals over $\mathbf{R}$, or with g undefined at both endpoints, one must be careful. The correct definition is

$$\int_{-\infty}^{+\infty} g(t)dt = \int_{-\infty}^0 g(t)dt + \int_0^{+\infty} g(t)dt,$$

or in other words, the improper integral over $\mathbf{R}$ exists if and only if the improper integrals over $]-\infty, 0]$ and over $[0, \infty[$ exist, and is then the sum of these two integrals. Similarly, if g is defined on $]a, b[$, but not at the endpoints, we pick any $x_0 \in]a, b[$ and define

$$\int_a^b g(t)dt = \int_a^{x_0} g(t)dt + \int_{x_0}^b g(t)dt,$$

when both integrals on the right-hand side exist (the value of the sum does not depend on the choice of x_0).

When they exist, improper integrals satisfy some of the basic properties of usual integrals. For instance, if the improper integrals of g_1 and g_2 exist, and if c, d are real numbers, then we have the linearity property

$$\int_a^{+\infty} (cg_1(t) + dg_2(t))dt = c \int_a^{+\infty} g_1(t)dt + d \int_a^{+\infty} g_2(t)dt,$$

and in particular the left-hand side integral also exists. This is also the case for the other types of improper integrals.

Similarly, for any $b \geqslant a$, we have

$$\int_a^{+\infty} g(t)dt = \int_a^b g(t)dt + \int_b^{+\infty} g(t)dt,$$

if either of the improper integrals exists (then the other does).

In order to prove the existence of improper integrals, the use of comparison principles is the most useful.

PROPOSITION 6.5.2. *Let* $g\colon [a, b[\to \mathbf{R}$ *be a continuous function, where* $b = +\infty$ *is allowed.*

(1) *If there exists* $h\colon [a, b[\to \mathbf{R}$ *such that*

$$|g(x)| \leqslant h(x)$$

for all $x \in [a, b[$ *and such that*

$$\int_a^b h(t)dt$$

exists, then the improper integral of g *on* $[a, b[$ *exists, and*

(6.12) $$\left| \int_a^b g(t)dt \right| \leqslant \int_a^b h(t)dt.$$

(2) *If* $g \geqslant 0$, *then the improper integral of* g *over* $[a, b[$ *exists if and only if there exists* M *such that*

$$\int_a^x g(t)dt \leqslant M$$

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

(3) *If there exists* $h\colon [a, b[\to \mathbf{R}$ *such that*

$$g(x) \geqslant h(x)$$

for all $x \geqslant a$ *and such that*

$$\int_a^b h(t)dt$$

does not exists, then the improper integral of g *on* $[a, b[$ *does not exist.*

Note that the first part is very similar to the fact that an absolutely convergent series is convergent, whereas the second corresponds to the convergence of monotone bounded sequence. Indeed, the proofs will use the same ideas.

PROOF. (1) We write the argument for $b = +\infty$. First we show that, for any sequence (x_n) tending to $+\infty$, the sequence

$$I_n = \int_a^{x_n} g(t)dt$$

is convergent. We use the Cauchy Criterion: let $n \in \mathbf{N}$ and $m \geqslant n$; then

$$|I_m - I_n| = \left| \int_{x_n}^{x_m} g(t)dt \right| \leqslant \int_{x_n}^{x_m} |g(t)|dt \leqslant \int_{x_n}^{x_m} h(t)dt = J_m - J_n$$

where

$$J_n = \int_a^{x_n} h(t)dt.$$

By assumption, the sequence (J_n) converges to the improper integral of h, hence by the Cauchy Criterion for (J_n), we deduce that (I_n) is also a Cauchy sequence, and therefore converges.

It now remains to prove that the limit of the sequence (I_n) does not depend on the choice of the sequence (x_n). This is done with the same trick as in the proof of Corollary 6.2.8: if (y_n) is another sequence tending to $+\infty$, we define a third sequence (z_n) by

$$z_{2n} = x_n, \qquad z_{2n+1} = y_n,$$

and the fact that (z_n) converges, by the first step, implies that (x_n) and (y_n) have the same limit.

Finally, from

$$\left| \int_a^x g(t)dt \right| \leqslant \int_a^x h(t)dt$$

for $x \geqslant a$, we deduce that (6.12) holds.

(2) Again we consider $b = +\infty$. Define

$$f(x) = \int_a^x g(t)dt$$

for $x \geqslant a$. Since $g \geqslant 0$, the function f is non-decreasing; by definition, the improper integral of f over $[a, +\infty[$ exists if and only if the function f has a limit as $x \to +\infty$.

So, if we assume first that the improper integral of g exists, then the function f has some limit M as $x \to +\infty$, and therefore $f(x) \leqslant M$ for all x.

Conversely, assume that f is bounded by some real number M. Let I be the supremum of the set X of values of $f(x)$ for $x \geqslant a$, which exists since X is non-empty and bounded from above. For any $\varepsilon > 0$, since $I - \varepsilon$ is not an upper-bound of X, there exists $x_0 \geqslant a$ such that $f(x_0) \geqslant I - \varepsilon$. We then get

$$I - \varepsilon \leqslant f(x_0) \leqslant f(x) \leqslant I$$

for all $x \geqslant x_0$. This implies that the function f converges to I as $x \to +\infty$.

(3) This is easy using (2): again writing the proof of $b = +\infty$, we have

$$\int_a^x g(t)dt \geqslant \int_a^x h(t)dt$$

for all $x \geqslant a$; the assumption implies that the right-hand side tends to $+\infty$ when $x \rightarrow +\infty$, and therefore so does the left-hand side, and hence the improper integral doesn't exist. □

REMARK 6.5.3. It might seem more natural to define

$$(6.13) \qquad \int_{-\infty}^{+\infty} g(t)dt = \lim_{x \to +\infty} \int_{-x}^{x} g(t)dt,$$

but this leads to paradoxical results. For instance, we may expect that

$$\int_{-\infty}^{+\infty} g(t)dt = \int_{-\infty}^{0} g(t)dt + \int_{0}^{+\infty} g(t)dt$$

holds for improper integrals, and (6.13) does not satisfy this natural condition. For instance, if g is any *odd*, which means that $g(-t) = -g(t)$ for all t, then we get by the substitution $u = -t$ the relation

$$\int_{-x}^{x} g(t)dt = \int_{x}^{-x} g(u)du = - \int_{-x}^{x} g(t)dt,$$

which means that the limit of the integral over $[-x, x]$ is always equal to 0. So, if we take $g(x) = x$, then (6.13) would lead to

$$\int_{-\infty}^{+\infty} tdt = 0,$$

whereas the improper integrals of g over $[0, +\infty[$ or $]-\infty, 0]$ do not exist.

EXAMPLE 6.5.4. (1) For any $a > 0$, the improper integral

$$\int_{0}^{+\infty} e^{-at}dt$$

exists, and is equal to $1/a$. Indeed, for $x > 0$, we have

$$\int_{0}^{x} e^{-at}dt = -\frac{1}{a}e^{-ax} + \frac{1}{a},$$

which converges to $1/a$ as $x \to +\infty$.

So if g is any function such that $|g(t)| \leqslant e^{-at}$ for some $a > 0$ and $t \geqslant 0$, then the improper integral of g also exists.

(2) Let $c \in \mathbf{R}$. For $c > 1$, the improper integral

$$\int_{1}^{+\infty} \frac{1}{t^c}dt$$

exists and is equal to $1/(c-1)$, while if $c \leqslant 1$, then the improper integral does not exist. Indeed, we have

$$\int_{1}^{x} \frac{1}{t^c}dt = \frac{1}{1-c}\left(\frac{1}{x^{c-1}} - 1\right) \qquad \text{if } c \neq -1,$$

$$\int_{1}^{x} \frac{1}{t}dt = \log(x),$$

for $x \geqslant 1$, and this converges if and only if $c \geqslant 1$. Hence, for instance, the improper integral

$$\int_{1}^{+\infty} \frac{1}{t^c}dt$$

159

exists, but

$$\int_1^{+\infty} \frac{2+\cos(t)}{t} dt$$

doesn't exist.

Simiarly, we consider the improper integral

(6.14)
$$\int_0^1 \frac{1}{t^c} dt$$

for $c > 0$ (the integral is not an improper integral when $c \leqslant 0$). For $y > 0$, we get

$$\int_y^1 \frac{1}{t^c} dt = \frac{1}{1-c}(1 - y^{1-c})$$

if $c \neq 1$ and

$$\int_y^1 \frac{1}{t} dt = -\log(y).$$

This means that the improper integral (6.14) exists if and only if $c < 1$; in that case, we have

$$\int_0^1 \frac{1}{t^c} dt = \frac{1}{1-c}.$$

In particular, putting both parts of this example together, the improper integral

$$\int_0^{+\infty} \frac{1}{t^c} dt$$

does not exist for *any* $c \geqslant 0$.

(3) There are many analogies between improper integrals and series, but there are also significant differences. For instance, it is not necessary that $g(t) \to 0$ as $t \to +\infty$ for the improper integral

$$\int_a^{+\infty} g(t) dt$$

to exist. An example is given by $g(t) = \cos(t^2)$: although there are infinitely many $t \to +\infty$ with $g(t) = 1$ (namely $t = \sqrt{2k\pi}$ with $k \in \mathbf{N}$), the improper integral of g over $[1, +\infty[$ (for instance) exists.

We can see this as follows: for $x \geqslant 1$, we first use the change of variable $u = t^2$ to get

$$\int_1^x \cos(t^2) dt = 2 \int_1^{x^2} \frac{\cos(u)}{\sqrt{u}} du.$$

Now we integrate by parts:

$$\int_1^{x^2} \frac{\cos(u)}{\sqrt{u}} du = \frac{\sin(x^2)}{x} + \frac{1}{2} \int_1^{x^2} \frac{\sin(u)}{u^{3/2}} du.$$

The first term tends to 0 as $x \to +\infty$, while for the second, we note that

$$\left| \frac{\sin(u)}{u^{3/2}} \right| \leqslant \frac{1}{u^{3/2}}$$

for all $u \geqslant 1$, and since the improper integral of $u^{-3/2}$ exists on $[1, +\infty[$ (see (2) above), we conclude by comparison that the improper integral of $\sin(u)/u^{3/2}$ also exists.

(4) For $c > 0$, we consider the existence of the improper integral

$$\Gamma(c) = \int_0^{+\infty} t^{c-1} e^{-t} dt.$$

There are two cases. If $c \geqslant 1$, then the function

$$g(x) = x^{c-1}e^{-x}$$

is continuous on $[0, +\infty[$, so the improper integral only involves the limit "at infinity". We use comparison. The function

$$h(x) = e^{-x/2}$$

is also continuous on $\mathbf{R}_+$, and since

$$\lim_{x \to +\infty} \frac{g(x)}{h(x)} = \lim_{x \to +\infty} x^{c-1}e^{-x/2} = 0,$$

there exists $M \geqslant 0$ such that

$$0 \leqslant g(x) \leqslant Mh(x) = Me^{-x/2}$$

for all $x \in \mathbf{R}_+$. By comparison and Example (1), we conclude that the improper integral $\Gamma(c)$ exists for $c \geqslant 1$.

If $0 \leqslant c < 1$, then the function g is only continuous on $]0, +\infty[$, so that the improper integral $\Gamma(c)$ involves two limits. We split the integral at $t = 1$ in order to understand what happens.

Arguing exactly as before by comparing with $e^{-x/2}$, we can see that the improper integral

$$\int_1^{+\infty} g(t)dt$$

does exist. On the other hand, for $0 < t \leqslant 1$, we have

$$g(t) = t^{c-1}e^{-t} \leqslant t^{c-1}.$$

Again by comparison and the second part of Example (2), the improper integral from 0 to 1 also exists.

We conclude that $\Gamma(c)$ is well-defined for all $c > 0$.

We now claim that the following properties are true: first, $\Gamma(1) = 1$, and $\Gamma(c + 1) = c\Gamma(c)$ for all $c > 0$. Once this is done we can use induction to conclude that for $n \in \mathbf{N}$, we have

$$\Gamma(n) = (n - 1)!.$$

First, we get

$$\Gamma(1) = \int_0^{+\infty} e^{-t}dt = 1.$$

Next, fix $c > 0$. For $x > 1$ and $0 < y < 1$, we observe that integration by parts leads to

$$\int_1^x t^c e^{-t}dt = x^c e^{-x} - e^{-1} + c\int_1^x t^{c-1}e^{-t}dt$$

$$\int_y^1 t^c e^{-t}dt = e^{-1} - y^c e^{-y} + c\int_y^1 t^{c-1}e^{-t}dt.$$

Letting $x \to +\infty$ in the first formula and $y \to 0$ in the second, we deduce that

$$\int_1^{+\infty} t^c e^{-t}dt = -e^{-1} + c\int_1^{+\infty} t^{c-1}e^{-t}dt$$

$$\int_0^1 t^c e^{-t}dt = e^{-1} + c\int_0^1 t^{c-1}e^{-t}dt.$$

Adding up, this means that
$$\Gamma(c+1) = c\Gamma(c).$$

6.6. A short introduction to Fourier series

We conclude with a very brief discussion of one of the fundamental application of analysis, especially of differential and integral calculus: the theory of Fourier series.

The motivation is the following: we are given a periodic signal (which could be a sound wave, a light signal, etc), that is represented by a function $f\colon \mathbf{R} \to \mathbf{R}$, which satisfies
$$f(x + 2\pi) = f(x)$$
for all $x \in \mathbf{R}$.

REMARK 6.6.1. The choice of the period 2π is not essential, but is convenient because this means that the functions sine and cosine are examples. If we have instead $f(x+T) = f(x)$, for some other fixed number T, then the basic examples are $f(x) = \cos(2\pi x/T)$ and $f(x) = \sin(2\pi x/T)$, since for instance
$$\cos\left(\frac{2\pi}{T}(x + T)\right) = \cos\left(\frac{2\pi x}{T} + 2\pi\right) = \cos\left(\frac{2\pi x}{T}\right).$$

Further examples of 2π-periodic functions are
$$c_k(x) = \cos(kx), \qquad s_k(x) = \sin(kx)$$
for all $k \in \mathbf{N}_0$. Note that c_0 is the constant function 1, whereas s_0 is the constant function zero.

The graphs of these functions are sometimes called "pure waves"; their graphs are similar, except that c_k has k-complete oscillations in the basic period interval $[0, 2\pi]$.

Fourier's idea and insight, that is now one of the most important in both pure and applied mathematics, is that *any* periodic signal (with maybe some regularity condition), should be represented as a *superposition* of these basic "pure" waves, or in other words, that $f(x)$ should be sum of a series
$$f(x) = a_0 + \sum_{k=1}^{+\infty}(a_k c_k(x) + b_k s_k(x)),$$
or equivalently
$$f(x) = a_0 + \sum_{k=1}^{+\infty}(a_k \cos(kx) + b_k \sin(kx)),$$
for suitable coefficients $(a_k)_{k \in \mathbf{N}_0}$ and $(b_k)_{k \in \mathbf{N}}$, now called *Fourier coefficients*.

EXAMPLE 6.6.2. A finite sum of the type
$$a_0 + \sum_{k=1}^{K}(a_k \cos(kx) + b_k \sin(kx))$$
for $K \in \mathbf{N}$ is called a *trigonometric polynomial*.

To illustrate that even finite sums of pure waves quickly achieve complicated shapes, Figure 6.4 is the graph over $[0, 2\pi]$ of the trigonometric polynomial
$$f(x) = \cos(x) - \sin(x) + \frac{1}{2}\cos(2x) + \frac{3}{5}\sin(2x) - 3\cos(3x) + \sin(9x).$$

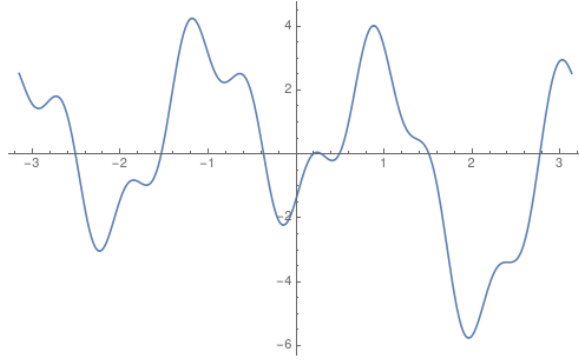Trigonometric polynomials have good properties. For instance:

FIGURE 6.4. A trigonometric polynomial

- The product of two trigonometric polynomials is also a trigonometric polynomial. This results from formulas such as

$$\cos(m_1 x) \sin(m_2 x) = \frac{1}{2}(\sin((m_1 + m_2)x) + \sin((m_2 - m_1)x))$$

for m_1 and $m_2 \in \mathbf{N}_0$, which can be checked from (4.5), or proved as follows:

$$\begin{aligned}
\cos(m_1 x) \sin(m_2 x) &= \frac{1}{4i}(e^{im_1 x} + e^{-im_1 x})(e^{im_2 x} - e^{-im_2 x}) \\
&= \frac{1}{4i}\left(e^{i(m_1 + m_2)x} - e^{-i(m_1 + m_2)x} + e^{i(m_2 - m_1)x} - e^{i(m_1 - m_2)x}\right) \\
&= \frac{1}{2}(\sin((m_1 + m_2)x) + \sin((m_2 - m_1)x)).
\end{aligned}$$

- If f is a trigonometric polynomial, then for any $x_0 \in \mathbf{R}$, the function $g(x) = f(x + x_0)$ is a trigonometric polynomial. This follows from the formulas (4.5), which imply that the functions $f(x) = \cos(k(x + x_0))$ or $f(x) = \sin(k(x + x_0))$ are trigonometric polynomials for all $k \in \mathbf{N}_0$ and $x_0 \in \mathbf{R}$.

Fourier realized that, if f is a superposition of pure waves, then (at least if f is continuous) the coefficients can be computed quite simply:

THEOREM 6.6.3. *Let $f \colon \mathbf{R} \to \mathbf{R}$ be a 2π-periodic continuous function which has a representation*

$$f(x) = a_0 + \sum_{k=1}^{+\infty}(a_k \cos(kx) + b_k \sin(kx)),$$

where the series on the right converges uniformly on $[0, 2\pi]$. Then we have

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(t)dt,$$

$$a_m = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos(mt)dt, \quad b_m = \frac{1}{\pi}\int_0^{2\pi} f(t)\sin(mt)dt, \qquad \text{for } m \in \mathbf{N}.$$

PROOF. Let $m \in \mathbf{N}$. By assumption, the trigonometric series converges uniformly to f on $[0, 2\pi]$, and it follows that

$$a_0 \cos(mx) + \sum_{k=1}^{+\infty}(a_k \cos(kx)\cos(mx) + b_k \sin(kx)\cos(mx)),$$

163

converges uniformly to $f(x)\cos(mx)$, simply because the partial sums satisfy

$$\left| f(x)\cos(mx) - \left(a_0\cos(mx) + \sum_{k=1}^{K}(a_k\cos(kx)\cos(mx) + b_k\sin(kx)\cos(mx)) \right) \right| \leqslant$$

$$\left| f(x) - \left(a_0 + \sum_{k=1}^{K}(a_k\cos(kx) + b_k\sin(kx)) \right) \right|,$$

so the uniform convergence follows from the assumption.

Consequently, by Corollary 6.2.15, we have

$$\int_0^{2\pi} f(t)\cos(mt)dt = a_0\int_0^{2\pi}\cos(mt)dt +$$

$$\sum_{k=1}^{+\infty}\left(a_k\int_0^{2\pi}\cos(kt)\cos(mt)dt + b_k\int_0^{2\pi}\sin(kt)\cos(mt)dt \right).$$

But from the *orthogonality relations* (Section 6.4, example 8), we have

$$\int_0^{2\pi}\cos(mt)dt = 0, \quad \int_0^{2\pi}\sin(kt)\cos(mt)dt = 0$$

and

$$\int_0^{2\pi}\cos(kt)\cos(mt)dt = 0$$

if $k \neq m$. So the result is that

$$\int_0^{2\pi} f(t)\cos(mt)dt = a_m\int_0^{2\pi}\cos(mt)^2 dt.$$

By the last of the orthogonality relations, the last integral is equal to π, and hence

$$\frac{1}{\pi}\int_0^{2\pi} f(t)\cos(mt)dt = a_m,$$

as claimed.

The other formulas are proved similarly. $\qquad\qquad\square$

These formulas are quite remarkable since they give a direct unique formula for the representation of f as a trigonometric series. But they do not tell us if, conversely, the series formed with these coefficients converges, or if it does, if its sum is equal to f.

Among many statements in this direction, the simplest is the following:

> THEOREM 6.6.4. *Suppose that $f \in C^2(\mathbf{R})$ has period 2π. Define*
>
> $$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(t)dt,$$
>
> $$a_m = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos(mt)dt, \quad b_m = \frac{1}{\pi}\int_0^{2\pi} f(t)\sin(mt)dt, \qquad \text{for } m \in \mathbf{N}.$$
>
> *Then the series*
>
> $$a_0 + \sum_{k=1}^{+\infty}(a_k\cos(kx) + b_k\sin(kx))$$
>
> *converges uniformly on $\mathbf{R}$, and its sum is equal to f.*

PROOF. **Step 1**. We will prove that there exists a real number $c \geqslant 0$ such that

$$|a_m| \leqslant \frac{c}{m^2}, \qquad |b_m| \leqslant \frac{c}{m^2}$$

for all $m \in \mathbf{N}$. Since $|\cos(mx)| \leqslant 1$ and $|\sin(mx)| \leqslant 1$, it follows that

$$|a_m \cos(mx) + b_m \sin(mx)| \leqslant \frac{2c}{m^2}$$

for all $x \in \mathbf{R}$ and all $m \in \mathbf{N}$. The trigonometric series is therefore normally convergent, and in particular uniformly convergent (Theorem 4.2.2).

Let $m \in \mathbf{N}$. To prove the inequality for a_m (the one for b_m is similar), we use the assumption that $f \in C^2(\mathbf{R})$ to integrate by parts twice, integrating the cosine term: we get

$$\pi a_m = \int_0^{2\pi} f(t) \cos(mt) dt$$

$$= f(2\pi) \cos(2\pi m) - f(0) \cos(0) - \frac{1}{m} \int_0^{2\pi} f'(t) \sin(mt) dt$$

$$= -\frac{1}{m} \left(f'(2\pi) \sin(2\pi m) - f'(0) \sin(0) + \frac{1}{m} \int_0^{2\pi} f''(t) \cos(mt) dt \right)$$

$$= -\frac{1}{m^2} \int_0^{2\pi} f''(t) \cos(mt) dt,$$

where we have also used the periodicity of f and f', and of cosine and sine to see that the first part of the integration by parts are zero.

Since f'' is continuous, there exists M such that $|f''(t)| \leqslant M$ for all $t \in [0, 2\pi]$, and using the triangle inequality, we conclude that

$$|a_m| \leqslant \frac{2\pi M}{\pi m^2} = \frac{2M}{m^2}.$$

Step 2. Let g be the sum of the series

$$g(x) = a_0 + \sum_{k=1}^{+\infty} (a_k \cos(kx) - b_k \sin(kx))$$

which exists and is continuous by Step 1. By Theorem 6.6.3, its Fourier coefficients are

$$a_k(g) = a_k, \qquad b_k(g) = b_k$$

for $k \in \mathbf{N}_0$.

Let $\varphi = f - g$; this is a continuous periodic function and its Fourier coefficients are

$$a_k(\varphi) = \frac{1}{\pi} \int_0^{2\pi} \varphi(t) dt = a_k - a_k(g) = 0, \qquad b_k(\varphi) = b_k - b_k(g) = 0.$$

So we have to show that a continuous function whose Fourier coefficients are all zero is everywhere zero. We use an idea of Lebesgue to do this.

If φ is not everywhere zero, then up to changing its sign and multiplying φ by a sufficiently large number, we can find x_0 such that $\varphi(x_0) > 1$. It follows by continuity that there exists $\delta > 0$ such that $\varphi(x) > 1$ for x in the interval $I = [x_0 - \delta, x_0 + \delta]$. We can assume that $0 < x_0 < 2\pi$, and that $\delta < 1/2$. We denote then $J = [x_0 - \delta/2, x_0 + \delta/2]$, a slightly smaller subinterval.
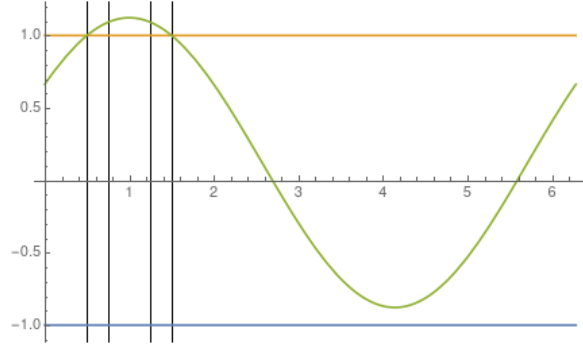
FIGURE 6.5. The function k_1 with $\delta = 1/2$ and $x_0 = 1$

Step 3. We construct a special trigonometric polynomial that will "test" the vanishing of the Fourier coefficients. Let $n \in \mathbf{N}$ and

$$k_n(x) = (1 - \cos(\delta) + \cos(x - x_0))^n.$$

By Example 6.6.2, the function $k_1(x)$ is a trigonometric polynomials, and therefore the function k_n is also one, as a product of finitely many trigonometric polynomials. Now writing k_n as a trigonometric polynomial, we observe that the linearity of the integral combined with the vanishing of the Fourier coefficients of φ imply that

$$(6.15) \qquad \int_0^{2\pi} \varphi(t) k_n(t) dt = 0.$$

We will obtain a contradiction by proving that, in fact, this integral tends to $+\infty$ as $n \to +\infty$.

Step 4. Observe first that

$$(6.16) \qquad 1 - \cos(\delta) + \cos(x - x_0) \geqslant 1 \text{ if } x \in I,$$

$$(6.17) \qquad 1 - \cos(\delta) + \cos(x - x_0) \geqslant 1 + \cos(\delta/2) - \cos(\delta) > 1 \text{ if } x \in J,$$

$$(6.18) \qquad |1 - \cos(\delta) + \cos(x - x_0)| \leqslant 1 \text{ if } x \notin I.$$

These properties are illustrated in Figure 6.5.

The first and second inequalities come from the fact that cosine is strictly decreasing on $[-\delta, \delta]$; for the third, we use

$$1 - \cos(\delta) + \cos(x - x_0) \geqslant -\cos(\delta) \geqslant -1,$$

for all x, and

$$\cos(x - x_0) \leqslant \cos(\delta)$$

if either $\delta \leqslant x - x_0 \leqslant \pi$ or $-\pi \leqslant x - x_0 \leqslant -\delta$ (which is easier to verify on a graph).

We now write

$$\int_0^{2\pi} \varphi(t) k_n(t) dt = \int_I \varphi(t) k_n(t) dt + \int_0^{x_0-\delta} \varphi(t) k_n(t) dt + \int_{x_0+\delta}^{2\pi} \varphi(t) k_n(t) dt$$

$$\geqslant \int_J \varphi(t) k_n(t) dt + \int_0^{x_0-\delta} \varphi(t) k_n(t) dt + \int_{x_0+\delta}^{2\pi} \varphi(t) k_n(t) dt,$$

because $k_n \geqslant 0$ on I (see (6.16)).

By (6.17), we have

$$\int_J \varphi(t) k_n(t) dt \geqslant \delta (1 + \cos(\delta/2) - \cos(\delta))^n,$$

166

which tends to $+\infty$. On the other hand, because of (6.18) and the triangle inequality, we get

$$\left|\int_0^{x_0-\delta} \varphi(t)k_n(t)dt\right| \leqslant (x_0 - \delta) \leqslant 2\pi$$

$$\left|\int_0^{x_0-\delta} \varphi(t)k_n(t)dt\right| \leqslant (2\pi - (x_0 + \delta)) \leqslant 2\pi$$

for all $n \in \mathbf{N}$.

So we get

$$\int_0^{2\pi} \varphi(t)k_n(t)dt \geqslant \delta(1 + \cos(\delta/2) - \cos(\delta))^n - 4\pi,$$

which tends to $+\infty$ as $n \to +\infty$, and therefore contradicts (6.15). $\qquad\square$

REMARK 6.6.5. (1) We mentioned that this is not best possible, and for instance Dirichlet proved the uniform convergence of Fourier series for periodic functions in $C^1(\mathbf{R})$. However, it cannot be extended to all continuous functions: there exist continuous periodic functions f and $x_0 \in \mathbf{R}$ such that the Fourier series for f does not converge at x_0.

(2) A remarkable "numerical" consequence of Theorem 6.6.4, and of its generalizations, is the *Parseval formula*

$$2a_0^2 + \sum_{k=1}^{+\infty}(a_k^2 + b_k^2) = \frac{1}{\pi}\int_0^{2\pi} |f(t)|^2 dt,$$

valid for any 2π-periodic function $f \in C^2(\mathbf{R})$, and in fact in much greater generality: it holds for *any* continuous function, for instance. Applied to suitable functions, this leads for instance to a proof of the formula

$$\sum_{k=1}^{+\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$

(mentioned in (2.16)).

Greek

We list here the greek letters often used in mathematics with their names.

α	alpha
β	beta
γ, Γ	gamma, capital gamma
δ	delta
η	eta
ε	epsilon
φ, Φ	phi, capital phi
κ	kappa
χ	chi
π, Π	pi, capital pi
ψ, Ψ	psi, capital psi
ϱ	rho
σ, Σ	sigma, capital sigma
θ	theta
ξ	xi
ζ	zeta
ω, Ω	omega, capital omega

Dictionary

1. Natural numbers=Natürliche Zahlen

2. Integers=ganze Zahlen

3. Induction principle=Induktionsprinzip

4. Factorial=Fakultät

5. Set=Menge

6. Union=Vereinigung

7. Intersection=Durchschnitt

8. Empty set=Leermenge

9. Product=Produkt

10. Subset=Teilmenge

11. Cardinality=Kardinalität

12. Map=Abbildung

13. Definition set=Definitionsmenge

14. Target set=Zielmenge

15. Function=Funktion

16. Image of x=Bild von x

17. Surjective=Surjektiv

18. Injective=Injektiv

19. Bijective=Bijektiv

20. Identity map=Identitätsabbildung

21. Composition=Verknüpfung

22. Inverse of f=Umkehrabbildung von f

23. Completeness=Vollständigkeit

24. Conjugate=Konjugierte Zahl

25. Modulus=Absolutbetrag

26. Binomial coefficients=Binomialkoeffizienten

27. n choose k=k aus n

28. Binomial formula=Binomialentwicklung

29. Sequence=Folge

30. Arithmetic progression=arithmetische Folge

31. Common difference=Differenz

32. Geometric progression=geometrische Folge

33. Limit=Grenzwert=Limes

34. Subsequence=Teilfolge

35. Accumulation point=Häufungswert

36. Series=Reihe

37. Converges absolutely=konvergiert absolut

38. Polynomial=Polynom

39. Continuity=Stetigkeit

40. Continuous=Stetig

41. Compact interval=Kompaktes Intervall

42. Converges Uniformly=konvergiert gleichmässig

43. Converges normally=konvergiert normal

44. Power series=Potenzreihe

45. Radius of convergence=Konvergenzradius

46. Primitive=Stammfunktion

47. Integration by parts=Partielle Integration

48. Change of variable=Substitutionsregel

49. Step-function=Treppenfunktion

50. Ruled function=Regelfunktion

51. Improper integrals=Uneigentliche Integrale

Bibliography

[1] M. Burger: *Skript Analysis I für INFK*.

[2] K. Königsberger: *Analysis I*, Springer.