

Foundations of Mathematical Analysis

Fabio Bagagiolo

Dipartimento di Matematica, Università di Trento

email:fabio.bagagiolo@unitn.it

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 3 |
| 2 | Basic concepts in mathematical analysis and some complements | 6 |
| 2.1 | Limit, supremum, and monotonicity | 6 |
| 2.2 | The Bolzano-Weierstrass theorem and the Cauchy criterium | 9 |
| 2.3 | Superior and inferior limit and semicontinuous functions | 12 |
| 2.4 | Infinite number series | 16 |
| 2.5 | Rearrangements | 21 |
| 2.6 | Sequences of functions | 25 |
| 2.7 | Series of functions | 30 |
| 2.8 | Power series | 32 |
| 2.9 | Fourier series | 35 |
| 2.10 | Historical notes | 38 |
| 3 | Real numbers and ordered fields | 43 |
| 3.1 | Ordering and Archimedean properties | 44 |
| 3.2 | Isomorphisms and complete ordered fields | 50 |
| 3.3 | Choose your axiom, but choose one! | 57 |
| 3.4 | Historical notes and complements | 60 |
| 4 | Cardinality | 66 |
| 4.1 | The power of a set | 66 |
| 4.2 | Historical notes and comments | 74 |
| 5 | Topologies | 76 |
| 5.1 | On the topology in \mathbb{R} | 76 |
| 5.2 | Metric spaces | 77 |
| 5.3 | Completeness | 79 |
| 5.4 | Compactness in metric spaces | 84 |
| 5.5 | The importance of being compact | 87 |

| | | |
|----------|---|------------|
| 5.6 | Topological spaces | 91 |
| 5.7 | Convergence and continuity in topological spaces | 97 |
| 5.8 | Compactness in topological spaces and sequential compactness | 100 |
| 5.9 | Fast excursion on density, separability, Ascoli-Arzelà, Baire and Stone-Weierstrass | 105 |
| 5.10 | Topological equivalence and metrizability | 113 |
| 5.11 | Historical notes | 115 |
| 6 | Construction of some special functions | 117 |
| 6.1 | A continuous nowhere derivable function | 117 |
| 6.2 | Abundance of smooth functions in \mathbb{R}^n | 119 |

1 Introduction

These are the notes of the course *Foundations of Analysis* delivered for the “laurea magistrale” in mathematics at the University of Trento. The natural audience of such a course (and hence of such notes) is given by students who have already followed a three years curriculum in mathematics, or, at least, who are supposed to be already familiar with the notions of rational numbers, real numbers, ordering, sequence, series, functions, limit, continuity, differentiation and so on. Roughly speaking, these notes are not for beginners. Actually, many of the elementary concepts of the mathematical analysis will be recalled along the notes, but this will be always done just thinking that the reader already knows such concepts and moreover has already worked¹ with it.

There are two main concepts that a student faces when she starts to study mathematical analysis. They are the concept of limit and the concept of supremum also said superior extremum (as well as of infimum or inferior extremum). Such two concepts are those which are the basic bricks for constructing all the real analysis² as we study, learn, know, investigate and apply nowadays. We can say that “limit” and “supremum” form the “core” of mathematical analysis, from which all is generated. In Figure 1 many of the generated things are reported; the ones inside the ellipse are somehow touched by this notes, in particular in Sections 2 and 6. Indeed, this is one of the main goal of the notes: starting from the basic concepts of limit and supremum, show how all the rest comes out.

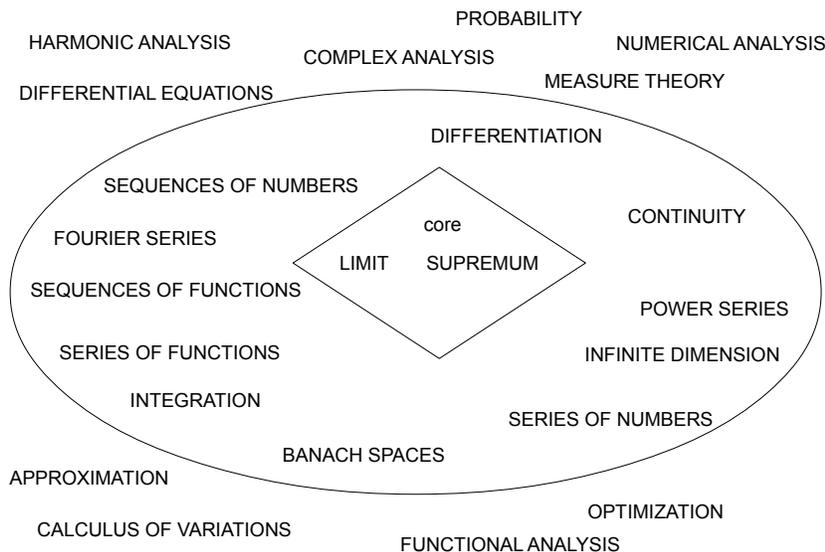


Figure 1: The core of the Mathematical Analysis

The other goal of these notes is to break the “core”, separately analyzing and gener-

¹Made exercises.

²And in some sense also the complex analysis.

alizing the two concepts of limit and supremum. This will lead us to consider, from one side, metrics and topologies in general sets, and, from the other side, ordered fields with the archimedean and completeness property. Figure 2 shows such an approach, which will be developed in Sections 3, 4, and 5. One of the main results here reported is the fact that the set of the real numbers is the unique³ complete ordered field. In some sense, we can say that, if we want to make analysis, we have to make it on the real numbers: no alternatives are given.

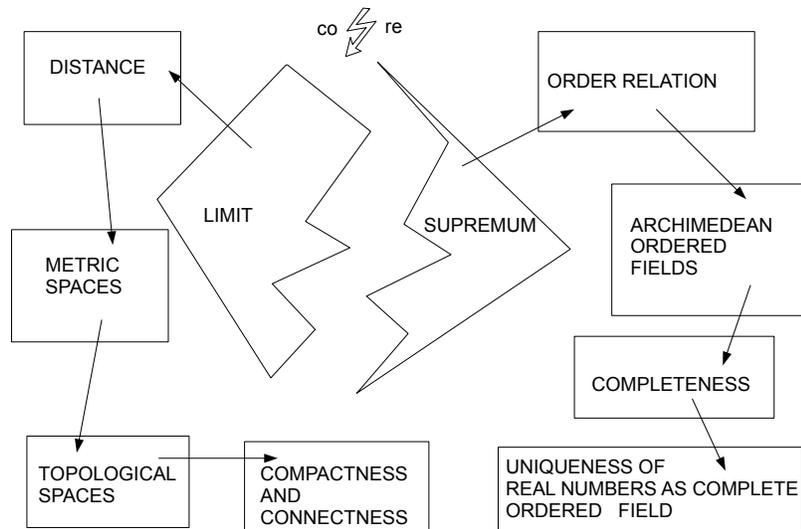


Figure 2: Breaking the core

The references reported as last section were all consulted and of inspiration for writing these notes. I would also like to thank some colleagues of mine with whom, during the writing of the notes, I have discussed about the subject and asked for some clarifications. They are Stefano Baratella, Gabriele Greco, Valter Moretti, Francesco Serra Cassano, Andrea Pugliese, Marco Sabatini, Raul Serapioni, and our late lamented Mimmo Luminati.

The mathematical notations in these notes are the standard ones. In particular \mathbb{R} , \mathbb{N} , \mathbb{Q} and \mathbb{C} respectively stand for the sets of real, natural, rational and complex numbers. Moreover, if $n \in \mathbb{N} \setminus \{0\}$, then \mathbb{R}^n is, as usual, the n -dimensional space $\mathbb{R} \times \dots \times \mathbb{R}$ i.e. the cartesian product of \mathbb{R} n -times. With the notation $[a, b]$ we will mean the interval of real numbers $\{x \in \mathbb{R} | a \leq x \leq b\}$, that is the closed interval containing its extreme points. In the same way $]a, b[$ will denote the open interval without extreme points $\{x \in \mathbb{R} | a < x < b\}$ ⁴, $]a, b]$ the semi-open interval $\{x \in \mathbb{R} | a < x \leq b\}$ and $[a, b[$ the

³In the sense that they are all isomorphic.

⁴Here, we of course permit $a = -\infty$ as well as $b = +\infty$.

semi-open interval $\{x \in \mathbb{R} | a \leq x < b\}$ ⁵

In these notes the formulas will be enumerated by $(x.y)$ where x is the number of the section (independently from the number of the subsection) and y is the running number of the formula inside the section. Moreover, the statements will be labeled by “S $x.y$ ” where “S” is the type of the statement (Theorem, Proposition, Lemma, Corollary, Definition, Remark, Example), x is the number of the section (independently from the number of the subsection), and y is the running number of the statement inside the section (independently from the type of the statement).

The symbol “□” will mean the end of a proof.

Every comment, suggestion and mistake report will be welcome, both concerning mathematics and English.

⁵In the last two cases we, respectively, permit $a = -\infty$ and $b = +\infty$.

2 Basic concepts in mathematical analysis and some complements

2.1 Limit, supremum, and monotonicity

There is a rather simple but fundamental result which links together the concepts of limit and the one of supremum. It is the result concerning the limit of monotone sequences.

Definition 2.1 A sequence of real numbers is a function from the set of natural numbers \mathbb{N} to the set of real numbers \mathbb{R} . That is, for every natural numbers $n = 0, 1, 2, 3, \dots$ we choose a real number $a_0, a_1, a_2, a_3, \dots$. We indicate the sequence by the notation $\{a_n\}_{n \in \mathbb{N}}$.⁶

We say that the sequence is monotone increasing if

$$a_n \leq a_{n+1} \quad \forall n \in \mathbb{N}.$$

We say that it is monotone decreasing if

$$a_n \geq a_{n+1} \quad \forall n \in \mathbb{N}.$$

If ℓ is a real number, we say that the sequence converges to ℓ (or that ℓ is the limit of the sequence) if

$$\forall \varepsilon > 0 \exists \bar{n} \in \mathbb{N} \text{ such that } n \geq \bar{n} \implies |a_n - \ell| \leq \varepsilon.$$

Similarly, we say that the sequence converges (often also said diverges) to $+\infty$ (or, respectively, to $-\infty$) if

$$\begin{aligned} \forall M > 0 \exists \bar{n} \in \mathbb{N} \text{ such that } n \geq \bar{n} &\implies a_n \geq M \text{ (or, respectively,} \\ \forall M < 0 \exists \bar{n} \in \mathbb{N} \text{ such that } n \geq \bar{n} &\implies a_n \leq M). \end{aligned}$$

Definition 2.2 Given a nonempty subset $A \subseteq \mathbb{R}$ and a real number m , we say that m is a majorant or an upper bound (respectively: a minorant or a lower bound) of A if

$$a \leq m \quad \forall a \in A \quad (\text{respectively: } m \leq a \quad \forall a \in A).$$

The set A is said bounded from above if it has a majorant, it is said bounded from below if it has a minorant, it is said bounded if it has both majorants and minorants.

A real number $\ell \in \mathbb{R}$ is said to be the supremum of A , and we write $\ell = \sup A$, if it is the minimum of the majorants of A , that is if it is a majorant of A and any other real number strictly smaller than ℓ cannot be a majorant. In other words

⁶Actually, this is the image of the function from \mathbb{N} to \mathbb{R} , and indeed we will often identify the sequence with its image, which is the subset of \mathbb{R} $\{a_0, a_1, a_2, a_3, \dots\}$ whose a more compact notation is just $\{a_n\}_{n \in \mathbb{N}}$. Often, with an abuse of notation, we will indicate a sequence just by $\{a_n\}$ or even by a_n .

⁷Note that such a definition of increasingness, as well as the definition of decreasingness, takes also account of the constant case $a_n = a_{n+1}$ for some (or even every) n .

$$a \leq \ell \forall a \in A, \text{ and } (\ell' \in \mathbb{R}, a \leq \ell' \forall a \in A \implies \ell \leq \ell').$$

We say that ℓ is the infimum of A , writing $\ell = \inf A$, if it is the maximum of the minorants of A , that is if it is a minorant and any other real number strictly larger than ℓ cannot be a minorant. In other words

$$a \geq \ell \forall a \in A, \text{ and } (\ell' \in \mathbb{R}, a \geq \ell' \forall a \in A \implies \ell \geq \ell').$$

We say that the supremum of A is $+\infty$, and we write $\sup A = +\infty$, (respectively: the infimum is $-\infty$, and we write $\inf A = -\infty$) if A has no majorants (respectively: A has no minorants), that is

$$\begin{aligned} &\forall M > 0 \exists a \in A \text{ such that } a > M \text{ (respectively,} \\ &\forall M < 0 \exists a \in A \text{ such that } a < M). \end{aligned}$$

Note that when the supremum is a real number ℓ then, the fact that it is the minimum of the majorants, can be stated as

$$a \leq \ell \forall a \in A \text{ and } (\forall \varepsilon > 0 \exists a \in A \text{ such that } a > \ell - \varepsilon),$$

and similarly for the maximum of the minorants

$$a \geq \ell \forall a \in A \text{ and } (\forall \varepsilon > 0 \exists a \in A \text{ such that } a < \ell + \varepsilon).$$

The following result is of fundamental importance for the whole building of the mathematical analysis, and indeed it will be the subject of some of the next sections⁸.

Theorem 2.3 *Let $A \subseteq \mathbb{R}$ be a non empty set. Then, its supremum exists (possibly equal to $+\infty$) and it is unique; in particular, if A is bounded from above, then its supremum is a finite real number. Similarly, its infimum exists (possibly equal to $-\infty$) and it is unique; in particular, if A is bounded from below, then its infimum is a finite real number.*

Remark 2.4 *Let us note that the importance of Theorem 2.3 is especially given by the existence result for bounded sets (the uniqueness being almost obvious by the definition). Indeed, if we look for the supremum of a set inside another universe-set, then the existence is not more guaranteed. Think for instance to the bounded set*

$$A = \left\{ q \in \mathbb{Q} \mid q^2 \leq 2 \right\} \subset \mathbb{Q},$$

and look for its supremum “inside \mathbb{Q} ”, that is look for a rational number \bar{q} which is a majorant of A and such that any other different rational majorant of A is strictly larger.

⁸Is it really a theorem or is it an assumption, an axiom? We are going to investigate such a question in a following part of these notes.

It is well known that such a number \bar{q} does not exist. On the other side, the supremum of A as subset of \mathbb{R} exists and it is equal to $\sqrt{2}$ (which is not a rational number, of course). This fact is also naively referred as “ \mathbb{Q} has holes, whereas \mathbb{R} has not holes”.

We also recall the difference between the concept of supremum and of maximum of a set $A \subseteq \mathbb{R}$. The maximum is a majorant which belongs to the set A , whereas the supremum is not required to belong to the set. In particular, if the set A has a maximum, then such a maximum coincides with the supremum; if the supremum belongs to the set, then it is the maximum; if the supremum does not belong to the set, then the set has no maximum. Similar considerations hold for minimum and infimum. For instance the interval $[0, 1[\subset \mathbb{R}$ has infimum and minimum both coincident with 0, has supremum equal to 1, but has no maximum.

Now we are ready to state and prove the result of main interest for this section

Theorem 2.5 (*Limit of monotone sequences*). Let $\{a_n\}_{n \in \mathbb{N}}$ be an increasing monotone sequence of real numbers, and let $\ell \in]-\infty, +\infty]$ be its supremum, that is

$$\ell = \sup\{a_0, a_1, a_2, a_3, \dots\} \in]-\infty, +\infty],$$

which exists as stated in Theorem 2.3. Then the sequence converges to ℓ . In other words, any increasing monotone sequence converges to its supremum (possibly equal to $+\infty$). Similarly, if the sequence is decreasing monotone, then it converges to its infimum (possibly equal to $-\infty$).

Proof. We prove only the case of increasing monotone sequence and finite supremum $\ell \in \mathbb{R}$.⁹ We have to prove that the sequence converges to ℓ , that is, for every $\varepsilon > 0$ we have to find a natural number \bar{n} (depending on ε) such that, for every larger natural number n , we have $|a_n - \ell| \leq \varepsilon$. In doing that we have to use the monotonicity and the property of the supremum. Since ℓ is a majorant, we have

$$a_n \leq \ell \quad \forall n \in \mathbb{N}.$$

Let us fix $\varepsilon > 0$, since ℓ is the minimum of the majorants we find $\bar{n} \in \mathbb{N}$ such that

$$\ell - \varepsilon < a_{\bar{n}}.$$

Using the monotonicity, we then get

$$n \geq \bar{n} \implies \ell - \varepsilon < a_{\bar{n}} \leq a_n \leq \ell < \ell + \varepsilon \implies |a_n - \ell| \leq \varepsilon.$$

□

Another way to state Theorem 2.5 is just to say “if a sequence is increasing monotone and bounded from above, then it converges to its (finite) supremum; if it is monotone

⁹The reader is invited to prove the other statements, that is: increasing monotone sequence and $\ell = +\infty$, decreasing monotone sequence and $\ell \in \mathbb{R}$, and decreasing monotone sequence and $\ell = -\infty$.

increasing and not bounded from above, then it diverges to $+\infty$ ". Note that a bounded sequence does not necessarily converge. Take for instance the sequence $a_n = (-1)^n$ which, of course, is not monotone.

Theorem 2.5 has great importance and it is widely used in analysis. In particular it also enlightens the importance of the property of monotonicity. In the sequel we present some important consequences of such a theorem.

2.2 The Bolzano-Weierstrass theorem and the Cauchy criterium

Definition 2.6 Given a sequence of real numbers $\{a_n\}_{n \in \mathbb{N}}$, a subsequence of it is a sequence of real numbers of the form $\{a_{n_k}\}_{k \in \mathbb{N}}$ where n_k stays for a strictly increasing function

$$\mathbb{N} \rightarrow \mathbb{N}, \quad k \mapsto n_k.$$

In other words, $k \mapsto n_k$ is a strictly increasing selection of indices and so a subsequence is a sequence given by a selection of infinitely many elements of the ordinary sequence a_n , which are labeled respecting the same ordinary order.

Example 2.7 Given the sequence $a_n = n^2 - n - 1$, that is

$$-1, -1, 1, 5, 11, 19, 29, 41, 55, 71, 89, \dots$$

defining $n_k = 2k + 1$, we get the subsequence a_{n_k} as

$$-1, 5, 19, 41, 71, \dots$$

Proposition 2.8 A sequence $\{a_n\}$ converges to $\ell \in [-\infty, +\infty]$ if and only if every strict¹⁰ subsequence of it converges to the same value ℓ .

Proof. We prove only the necessity in the case $\ell \in \mathbb{R}$. Let us suppose that all strict subsequences converge to ℓ and, by absurd, let us suppose that a_n does not converge to ℓ . This means that it is not true that

$$\forall \varepsilon > 0 \exists \bar{n} \in \mathbb{N} \text{ such that } n \geq \bar{n} \implies |a_n - \ell| \leq \varepsilon.$$

That is, there exists $\bar{\varepsilon} > 0$ such that it is not true that

$$\exists \bar{n} \in \mathbb{N} \text{ such that } n \geq \bar{n} \implies |a_n - \ell| \leq \bar{\varepsilon}.$$

This means that, for every $k \in \mathbb{N}$, we may find $n_k \geq k$ such that $n_k > n_{k-1} + 1$ and that

$$|a_{n_k} - \ell| > \bar{\varepsilon}.$$

Hence, the strict subsequence $\{a_{n_k}\}_k$ does not converge to ℓ and this is a contradiction.

□

The celebrated Bolzano-Weierstrass theorem is the following,

¹⁰i.e. a subsequence which is not coincident with the whole sequence itself.

Theorem 2.9 Every bounded¹¹ sequence of real numbers admits a convergent subsequence.

To prove the theorem we need the following bisection lemma.

Lemma 2.10 Let $I_0 = [a, b]$ be a bounded closed interval of \mathbb{R} . We divide it in two closed half-parts by sectioning it by its medium point $(a + b)/2$ and we call I_1 one of those parts, for instance the first one:

$$I_1 = \left[a, \frac{a + b}{2} \right].$$

Hence we divide I_1 in two closed half-parts by sectioning it by its medium point $(a + (a + b)/2)/2$ and we call I_2 one of those two parts, for instance the second one:

$$I_2 = \left[\frac{3a + b}{4}, \frac{a + b}{2} \right].$$

We proceed in this way, at every step we bisect the closed interval and we choose one of the two closed half-parts, hence, at every step, I_{n+1} is one of the two half-parts of I_n .

Then, the intersection of all (nested) closed intervals I_n is not empty and contains just one point only. That is there exists $c \in [a, b]$ such that

$$\bigcap_{n \in \mathbb{N}} I_n = I_0 \cap I_1 \cap I_2 \cap \cdots \cap I_n \cap \cdots = \{c\}.$$

Proof. Let us denote $I_n = [a_n, b_n]$ for every n and consider the two sequences $\{a_n\}, \{b_n\}$. First of all note that, since the intervals are (closed) and nested $a_n, b_n \in I_m \subseteq I_0$ for all $0 \leq m \leq n$, and hence the sequences are bounded. It is easy to prove that $\{a_n\}$ is increasing and $\{b_n\}$ is decreasing. Hence, by Theorem 2.5 they both converge to $c', c'' \in I_0$ respectively, with $a_n \leq c' \leq c'' \leq b_n$ for all n (this is obvious since $a_n \leq b_n$ for all n , a_n is increasing and b_n decreasing). By construction, it is also obvious that

$$0 \leq b_n - a_n \leq \frac{b - a}{2^n},$$

and this would imply $c' = c'' = c$. Since a generic x belongs to $\in \bigcap I_n$ if and only if

$$a_n \leq x \leq b_n \quad \forall n \in \mathbb{N},$$

we get the conclusion. □

Note that if we perform the same bisection procedure but we take the open half-parts of the intervals instead of the closed ones, then the conclusion is not more true in general. Consider, for example, the sequence of open intervals $]0, 1/2^n[$ whose intersection is empty.

¹¹A sequence a_n is bounded if there exists $M > 0$ such that $|a_n| \leq M$ for all $n \in \mathbb{N}$, that is if the subset $\{a_0, a_1, \dots\} \subset \mathbb{R}$ is bounded.

Proof of Theorem 2.9. Let $\{a_n\}$ be the sequence. Since it is bounded, there exists a closed bounded interval $I = [\alpha, \beta]$ which contains a_n for all n . We are going to apply the bisection procedure to I . Let us denote by I_0 one of the two closed half-parts of I which contains infinitely many a_n (at least one of such half-parts exists). Hence we define

$$n_0 = \min\{n \in \mathbb{N} \mid a_n \in I_0\}.$$

Now, let us denote by I_1 one of the two closed half-parts of I_0 which contains infinitely many a_n (again, at least one of such half-parts exists). We define

$$n_1 = \min\{n > n_0 \mid a_n \in I_1\}.$$

We proceed in this way by induction: we define I_{k+1} as one of the two closed half-parts of I_k which contains infinitely many a_n and define

$$n_{k+1} = \min\{n > n_k \mid a_n \in I_{k+1}\}.$$

By construction, the subsequence $\{a_{n_k}\}_k$ satisfies

$$a_{n_k} \in I_k = \bigcap_{j \leq k} I_j, \quad \forall k \in \mathbb{N}.$$

By Lemma 2.10, and by the definition of limit, we get that there exists $c \in I$ such that $a_{n_k} \rightarrow c$ as $k \rightarrow +\infty$. \square

An immediate consequence of Theorem 2.9 is the so-called Cauchy criterium for the convergence of a sequence of real numbers. Such a criterium gives a sufficient and necessary condition for a sequence to be convergent, and this without passing through the computation of the limit, which is usually a more difficult problem.

Proposition 2.11 *A real sequence $\{a_n\}$ converges if and only if*

$$\forall \varepsilon > 0 \exists m \in \mathbb{N} \text{ such that } n', n'' \geq m \implies |a_{n'} - a_{n''}| \leq \varepsilon. \quad (2.1)$$

The condition (2.1) is called Cauchy condition and a sequence satisfying it is called a Cauchy sequence.

Proof. We leave the proof of the necessary of the Cauchy condition to the reader as an exercise. Let us prove the sufficiency. We first prove that (2.1) implies the boundedness of the sequence. Indeed, let us fix $\varepsilon > 0$ and let m be as in (2.1), then we have

$$\begin{aligned} n \geq m &\implies |a_n| \leq |a_m| + |a_n - a_m| \leq |a_m| + \varepsilon =: M', \\ n < m &\implies |a_n| \leq \max\{|a_0|, |a_1|, \dots, |a_m|\} =: M''. \end{aligned}$$

Hence we get $|a_n| \leq \max\{M', M''\}$ for all n , that is the boundedness of the sequence. Hence, by Theorem 2.9, there exists a subsequence $\{a_{n_k}\}_k$ converging to a real number

c. If we prove that any other subsequence is convergent to the same c , then we are done. Let $\{a_{n_j}\}_j$ be another subsequence, let us fix $\varepsilon > 0$, and let m be as in (2.1). Let $\bar{k}, \bar{j} \in \mathbb{N}$ be such that

$$n_{\bar{k}} \geq m, \quad |a_{n_{\bar{k}}} - c| \leq \varepsilon, \quad j \geq \bar{j} \implies n^j \geq m.$$

Hence we get, for every $j \geq \bar{j}$,

$$|a_{n^j} - c| \leq |a_{n^j} - a_{n_{\bar{k}}}| + |a_{n_{\bar{k}}} - c| \leq 2\varepsilon,$$

from which the convergence $a_{n^j} \rightarrow c$ as $j \rightarrow +\infty$. □

Remark 2.12 *What does the Cauchy condition (2.1) mean? Roughly speaking, it means that, when n increases, the terms a_n of the sequence “accumulate” themselves. If we think to a sequence as a discrete evolution of a material point that at time $t = 0$ occupies the position a_0 on the real line, at time $t = 1$ occupies the position a_1 and so on, then (2.1) says that, when time goes to infinity, the material point brakes: it occupies positions which become closer and closer, even if it continues to move. Since the real numbers has the supremum property, which means that it has no holes, then the braking material point has to finish its running somewhere, that is it has to stop in a point, which exactly is the limit of the sequence.*

As it is obvious, if the same Cauchy property (2.1) holds for a sequence of rational numbers, then such a sequence does not necessarily converge to a rational number, and this is because the field of the rational numbers \mathbb{Q} “has many holes”. Take for instance the fundamental sequence $(1 + 1/n)^n$ which, as it is well known, converges to the irrational (even transcendental) Napier number e . Being convergent in \mathbb{R} it is then a Cauchy sequence of rational numbers, but it is not convergent to any point of \mathbb{Q} .

2.3 Superior and inferior limit and semicontinuous functions

Definition 2.13 *Given a sequence $\{a_n\}_n$, its inferior and superior limits are, respectively¹²*

$$\underline{a} = \liminf_{n \rightarrow +\infty} a_n := \lim_{n \rightarrow +\infty} \left(\inf_{m \geq n} a_m \right), \quad \bar{a} = \limsup_{n \rightarrow +\infty} a_n := \lim_{n \rightarrow +\infty} \left(\sup_{m \geq n} a_m \right).$$

Example 2.14 i) The sequence $\{(-1)^n\}_n$ has inferior limit equal to -1 and superior limit equal to 1 .

ii) The sequence

¹²It is intended that, if $\inf_{m \geq n} a_m = -\infty$ for all n , then, formally, $\lim_{n \rightarrow +\infty} (-\infty) = -\infty$. Note that this happens if and only if the sequence is not bounded from below. Also note that if $\inf_{m \geq \bar{n}} a_m > -\infty$ for some \bar{n} , then $\inf_{m \geq n} a_m > -\infty$ for all n . Similar considerations hold for the superior limit.

$$a_n = \begin{cases} n & \text{if } n \text{ is odd,} \\ 1 - \frac{1}{n} & \text{if } n \text{ is even, but not divisible by 4,} \\ \log\left(\frac{1}{n}\right) & \text{if } n \text{ is divisible by 4,} \end{cases}$$

has inferior limit equal to $-\infty$ and superior limit equal to $+\infty$.

iii) The sequence

$$a_n = \begin{cases} e^{\frac{27n}{n^2-97}} & \text{if } n \text{ is even,} \\ 1 - \frac{1}{\log(n+1)} & \text{if } n \text{ is odd,} \end{cases}$$

has inferior and superior limit both equal to 1^{13} .

Proposition 2.15 i) *The inferior and superior limits \underline{a}, \bar{a} of a sequence always exist in $[-\infty, +\infty]$;*

ii) *it always holds that $\underline{a} \leq \bar{a}$;*

iii) *an extended number $a \in [-\infty, +\infty]$ is the inferior limit \underline{a} (respectively: the superior limit \bar{a}) of the sequence $\{a_n\}$ if and only if the following two facts hold: for any subsequence $\{a_{n_k}\}$ it is $a \leq \liminf_{k \rightarrow +\infty} a_{n_k}$ (respectively: $a \geq \limsup_{k \rightarrow +\infty} a_{n_k}$) and there exists a subsequence $\{a_{n_k}\}$ such that $a = \lim_{k \rightarrow +\infty} a_{n_k}$ (respectively: $a = \lim_{k \rightarrow +\infty} a_{n_k}$);*

iv) *\underline{a} and \bar{a} are both equal to the same extended number $a \in [-\infty, +\infty]$ if and only if the whole sequence a_n converges (or diverges) to a .*

Proof. We only prove i) for the inferior limit, iii) for the inferior limit with $\underline{a} \in \mathbb{R}$ and only concerning the necessity of the existence of the subsequence $\lim_k a_{n_k} = \underline{a}$, and iv) with $\underline{a} \in \mathbb{R}$. The other points and cases are left as exercise. i) For every n , we define $\underline{a}_n = \inf_{m \geq n} a_m$. It is evident that $\underline{a} = \lim \underline{a}_n$ and that the sequence $\{\underline{a}_n\}$ is monotone increasing. Hence we conclude by Theorem 2.5. iii) By definition of infimum, for every integer $k > 0$ there exists $n_k \geq k$ such that

$$a_{n_k} - \frac{1}{k} \leq \underline{a}_k = \inf_{m \geq k} a_m \leq a_{n_k} \leq a_{n_k} + \frac{1}{k}.$$

Now, let us fix $\varepsilon > 0$ and take $k' \in \mathbb{N}$ such that $1/k' \leq \varepsilon/2$ and that $|\underline{a}_n - \underline{a}| \leq \varepsilon/2$ for $n \geq k'$. Hence we conclude by

$$k \geq k' \implies |a_{n_k} - \underline{a}| \leq |a_{n_k} - \underline{a}_k| + |\underline{a}_k - \underline{a}| \leq \frac{1}{k} + \frac{\varepsilon}{2} \leq \varepsilon.$$

iv) If $\lim_n a_n = a \in \mathbb{R}$ then, by definition of limit, for every $\varepsilon > 0$ there exists n' such that $|a_n - a| \leq \varepsilon$ for any $n \geq n'$. But this of course implies that $|\underline{a}_n - a|, |\bar{a}_n - a| \leq \varepsilon$ from

¹³Actually the whole series converges to 1.

which we conclude that $a = \bar{a} = \underline{a}$. Vice versa, if $\underline{a} = \bar{a} = a$, by absurd we suppose that a_n does not converge to a . But then, there exists a subsequence a_{n_k} whose inferior or superior limit is respectively strictly lower or strictly greater than a , which is a contradiction to point iii). Indeed, since a_n does not converge to a , we can find $\varepsilon > 0$ such that for every integer $k > 0$ there exists $n_k \geq k$ with $|a_{n_k} - a| \geq \varepsilon$, and this means that the above claimed subsequence exists. \square

Definition 2.16 Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a function and $x_0 \in \mathbb{R}$ be a fixed point. The inferior and the superior limits of f at x_0 are respectively¹⁴

$$\liminf_{x \rightarrow x_0} f(x) := \lim_{r \rightarrow 0^+} \left(\inf_{x \in [x_0 - r, x_0 + r] \setminus \{x_0\}} f(x) \right), \quad \limsup_{x \rightarrow x_0} f(x) := \lim_{r \rightarrow 0^+} \left(\sup_{x \in [x_0 - r, x_0 + r] \setminus \{x_0\}} f(x) \right).$$

The function f is said to be lower semicontinuous (l.s.c.) at x_0 (respectively said upper semicontinuous (u.s.c.) at x_0) if

$$f(x_0) \leq \liminf_{x \rightarrow x_0} f(x) \quad (\text{respectively } \limsup_{x \rightarrow x_0} f(x) \leq f(x_0)).$$

We just say that f is lower semicontinuous (respectively, upper semicontinuous), without referring to any point, if it is lower semicontinuous (respectively, upper semicontinuous) at x for every x of its domain.

We recall here the well known definitions of limit and continuity.

Definition 2.17 Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a function, $x_0 \in \mathbb{R}$ be a point and $\ell \in [-\infty, +\infty]$. We say that the limit of f at x_0 is the value ℓ , and we write $\lim_{x \rightarrow x_0} f(x) = \ell$ if (supposing $\ell \in \mathbb{R}$ ¹⁵): for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|x - x_0| \leq \delta \implies |f(x) - \ell| \leq \varepsilon.$$

If $\lim_{x \rightarrow x_0} f(x) = \ell \in \mathbb{R}$ and $f(x_0) = \ell$, then we say that f is continuous at x_0 .

Proposition 2.18 Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a function and $x_0 \in \mathbb{R}$ be a point, and let us denote $\underline{\ell} = \liminf_{x \rightarrow x_0} f(x)$, $\bar{\ell} = \limsup_{x \rightarrow x_0} f(x)$.

- i) The inferior and the superior limit of f at x_0 always exist in $[-\infty, +\infty]$ and $\underline{\ell} \leq \bar{\ell}$;
- ii) there exists $\ell \in [-\infty, +\infty]$ such that $\underline{\ell} = \bar{\ell} = \ell$ if and only if $\lim_{x \rightarrow x_0} f(x) = \ell$;
- iii) f is simultaneously lower and upper semicontinuous at x_0 if and only if it is continuous at x_0 ;

¹⁴It is intended that, if $\inf_{x \in [x_0 - r, x_0 + r] \setminus \{x_0\}} f(x) = -\infty$ for all $r \rightarrow 0^+$, then, formally, $\lim_{r \rightarrow 0^+} (-\infty) = -\infty$. Note that this happens if and only if the function is not bounded from below in any neighborhood of x_0 . Also note that if $\inf_{x \in [x_0 - \bar{r}, x_0 + \bar{r}] \setminus \{x_0\}} f(x) > -\infty$ for some \bar{r} , then $\inf_{x \in [x_0 - r, x_0 + r] \setminus \{x_0\}} f(x) > -\infty$ for all $0 < r \leq \bar{r}$. Similar considerations hold for the superior limit.

¹⁵The reader is suggested to write down the definition of limit in the case $\ell = \pm\infty$ and in the case $x_0 = \pm\infty$

iv) f is lower semicontinuous (respectively, upper semicontinuous) at x' if and only if, for every sequence x_n converging to x' we have

$$f(x') \leq \liminf_{n \rightarrow +\infty} f(x_n) \quad (\text{respectively, } \limsup_{n \rightarrow +\infty} f(x_n) \leq f(x'));$$

v) there exist two sequences $\{x_n\}, \{z_n\}$ such that, respectively, $\lim_{n \rightarrow +\infty} f(x_n) = \underline{\ell}$, $\lim_{n \rightarrow +\infty} f(z_n) = \bar{\ell}$.

Proof. Exercise. □

Remark 2.19 Very naively speaking, we can say that a function is lower semicontinuous (respectively, upper semicontinuous) if all its discontinuities are downward (respectively, upward) jumps.

Example 2.20 i) A very simple example is the following: let $a \in \mathbb{R}$ and consider the function

$$f(x) = \begin{cases} -1 & \text{if } x < 0, \\ a & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Then, f is lower semicontinuous if and only if $a \leq -1$, it is upper semicontinuous if and only if $a \geq 1$, it is neither lower nor upper semicontinuous if and only if $-1 < a < 1$.

ii) Consider the function

$$f(x) = \begin{cases} \sin\left(\frac{1}{x}\right) & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ x \sin\left(\frac{1}{x}\right) & \text{if } x < 0. \end{cases}$$

Then

$$\liminf_{x \rightarrow 0} f(x) = -1, \quad \limsup_{x \rightarrow 0} f(x) = 1$$

If, with obvious definition, we consider the inferior and superior limits at $x = 0$ from left and from right, we respectively obtain

$$\liminf_{x \rightarrow 0^-} f(x) = 0, \quad \limsup_{x \rightarrow 0^-} f(x) = 0, \quad \liminf_{x \rightarrow 0^+} f(x) = -1, \quad \limsup_{x \rightarrow 0^+} f(x) = 1.$$

Semicontinuity plays an important role in the existence of minima and maxima of functions¹⁶. Indeed, the well known Weierstrass theorem says that a continuous function

¹⁶This is one of the most important subject in mathematical analysis, as well as in the applied sciences. Just think to the physical principle which says that the equilibrium positions of a physical system are given by minima of the total energy of the system.

on a compact set reaches its maximum and minimum values. But, in the Weierstrass theorem, besides the compactness, we require the continuity of the function and this fact simultaneously gives the existence of maximum and minimum. However, sometimes we may be interested in minima only (for instance in the case of total energy) or in maxima only. Hence, requiring the continuity of the function seems to be redundant. Indeed, the semicontinuity is enough. The following result just says, for instance, that the lower semicontinuity brings those properties of continuity which are enough to guarantee the existence of minima.

Theorem 2.21 *Let $[a, b] \subset \mathbb{R}$ be a compact interval¹⁷, and let $f : [a, b] \rightarrow \mathbb{R}$ be a lower (respectively, upper) semicontinuous function. Hence, there exists $\underline{x} \in [a, b]$ (respectively $\bar{x} \in [a, b]$) such that $f(\underline{x})$ is the minimum of f on $[a, b]$ (respectively, $f(\bar{x})$ is the maximum of f on $[a, b]$).*

Proof. We prove only the case of lower semicontinuous functions, the other case being left as an exercise. Let $m \in [-\infty, +\infty[$ be the infimum of f :

$$m := \inf_{x \in [a, b]} f(x).$$

By definition of infimum, there exists a sequence of points $x_n \in [a, b]$ such that

$$\lim_{n \rightarrow +\infty} f(x_n) = m.$$

Since $[a, b]$ is compact, there exists a subsequence $\{x_{n_k}\}_k$ and a point $\underline{x} \in [a, b]$ such that

$$x_{n_k} \rightarrow \underline{x}, \text{ as } k \rightarrow +\infty.$$

By the lower semicontinuity of f and by Proposition 2.18, we have

$$m = \lim_{k \rightarrow +\infty} f(x_{n_k}) = \liminf_{k \rightarrow +\infty} f(x_{n_k}) \geq f(\underline{x}),$$

from which we conclude since, by the definition of infimum, this implies $m = f(\underline{x})$. \square

2.4 Infinite number series

Definition 2.22 *Given a sequence $\{a_n\}$, the associated series is the series*

$$\sum_{n=0}^{+\infty} a_n = a_0 + a_1 + a_2 + a_3 + \cdots \tag{2.2}$$

and the sequence is called the general term of the series.

¹⁷i.e. closed and bounded.

The notation in (2.2) is obviously conventional. What is the meaning of the right-hand side? It should have the meaning of the “sum of an infinite quantity of addenda”. But a series tells also us the order of adding the addenda: first a_1 to a_0 , then a_2 to the previously found sum, and so on. Hence it is natural to consider the following definition.

Definition 2.23 *Given a real number $s \in \mathbb{R}$, we say that the series $\sum_{n=0}^{+\infty} a_n$ converges to s or that its sum is s , if the sequences of the k -partial summation*

$$s_k = a_0 + a_1 + a_2 + \cdots + a_k, \quad k \in \mathbb{N},$$

converges to s . In a similar way we define the convergence (or the divergence) of the series to $+\infty$ as well as to $-\infty$.

Of course, there are series which are not convergent, neither to a finite sum nor to an infinite sum; in this case we say that the series oscillates. Think for instance to the series $\sum_{n=0}^{+\infty} (-1)^n = 1 - 1 + 1 - 1 + 1 - 1 + \cdots$, for which $s_k = 0$ if k is odd and $s_k = 1$ if k is even.

By the definition of convergence of a series as convergence of the sequence of its partial summations and by Proposition 2.11, we immediately get the following Cauchy criterium for the series.

Proposition 2.24 *A series $\sum_{n=0}^{+\infty} a_n$ of real numbers converges to a finite sum if and only if*

$$\forall \varepsilon > 0 \exists m \in \mathbb{N} \text{ such that } m \leq n' \leq n'' \implies \left| \sum_{n=n'}^{n''} a_n \right| \leq \varepsilon.$$

From the Cauchy criterium we immediately get the following necessary condition for the convergence of the series.

Proposition 2.25 *If the series $\sum a_n$ converges to a finite sum, then its general term is infinitesimal, that is*

$$\lim_{n \rightarrow +\infty} a_n = 0.$$

Proof. If, by absurd, the general term is not infinitesimal, then there exist $\varepsilon > 0$ and a subsequence a_{n_k} such that $|a_{n_k}| > \varepsilon$ for every k . Now, take m as in the Cauchy criterium and k such that $n_k \geq m$. Hence we have the contradiction

$$|a_{n_k}| = \left| \sum_{n=n_k}^{n_k} a_n \right| \leq \varepsilon.$$

□

Example 2.26 The harmonic series

$$\sum_{n=1}^{+\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \dots$$

is divergent to $+\infty$. Indeed, for any $k \in \mathbb{N}$ we have

$$s_{2k} - s_k = \sum_{n=k+1}^{2k} \frac{1}{n} \geq \sum_{n=k+1}^{2k} \frac{1}{2k} = \frac{1}{2}.$$

By induction we then deduce

$$s_{2k} \geq \frac{k}{2} + s_1 \rightarrow +\infty, \text{ as } k \rightarrow +\infty,$$

and hence $s_n \rightarrow +\infty$ because it is monotone.

This example also shows that the condition $a_n \rightarrow 0$ in Proposition 2.25 is only necessary for the convergence of the series and not sufficient: the harmonic series diverges but its general term is infinitesimal.

Example 2.27 The geometric series of reason $c \in \mathbb{R} \setminus \{0\}$ is the series

$$\sum_{n=0}^{+\infty} c^n = 1 + c + c^2 + c^3 + \dots$$

Very simple calculations yield, for any k and for $c \neq 1$ (for which the series is obviously divergent: $s_k = k$),

$$\left. \begin{array}{l} s_{k+1} = s_k + c^{k+1} \\ s_{k+1} = 1 + cs_k \end{array} \right\} \implies s_k = \frac{1 - c^{k+1}}{1 - c}.$$

We then conclude that the geometric series is convergent to a finite sum if and only if $-1 < c < 1$ (and the sum is $1/(1 - c)$), it is divergent to $+\infty$ if $c \geq 1$ and finally it oscillates if $c \leq -1$.

Theorem 2.28 (*Series with positive terms*) *If the general term of the series is made by numbers which are non negative ($a_n \geq 0 \forall n$), then the series cannot oscillate. In particular, if the sequences of the partial summations s_k is bounded, then the series converges to a finite sum, and, if s_k is unbounded, then it diverges to $+\infty$. Similar considerations hold for the series with non positive terms.*

Proof. Since $a_n \geq 0$, then the sequence s_k is increasing monotone, and so the proof is a straightforward consequence of Theorem 2.5. \square

Theorem 2.28 gives a characterization of convergence to a finite sum for series with positive terms: a positive terms series converges to a finite sum if and only if the sequence of its partial summations is bounded¹⁸. Such a characterization is the main ingredient of several convergence criteria for positive terms series.

¹⁸Since a convergent sequence is necessarily bounded, then the the boundedness of the sequence of partial summations is also necessary.

Proposition 2.29 Let $\sum_{n=0}^{+\infty} a_n, \sum_{n=0}^{+\infty} b_n$ be two positive terms series.

i) (Comparison criterium) If $\sum b_n$ is convergent to a finite sum, and if $0 \leq a_n \leq b_n$ for all n , then $\sum a_n$ is also convergent to a finite sum¹⁹.

ii) (Comparison criterium) If $\sum b_n$ diverges to $+\infty$, and if $0 \leq b_n \leq a_n$ for all n , then $\sum a_n$ is also divergent to $+\infty$.

iii) (Ratio criterium) If $a_n > 0$ for all n , let $\ell = \lim_{n \rightarrow +\infty} a_{n+1}/a_n \in [0, +\infty[$ exist. If $0 \leq \ell < 1$, then the series converges to a finite sum; if instead $\ell > 1$, then the series is divergent to $+\infty$; if $\ell = 1$, then everything is still possible²⁰.

iv) (n -th root criterium) If $a_n \geq 0$ for all n and if $\limsup_{n \rightarrow +\infty} (a_n)^{1/n} < 1$, then the series converges to a finite sum; if instead $\limsup_{n \rightarrow +\infty} (a_n)^{1/n} > 1$, then the series is divergent to $+\infty$; if that superior limit is equal to 1, then everything is still possible²¹.

Proof. We only prove i) and the first two assertions of iii) and iv).

i) Let s_k and σ_k be respectively the partial summations of $\sum a_n$ and of $\sum b_n$. Since the latter is convergent to a finite sum, then the sequence of partial summation, being convergent, is bounded by a constant M . But then by our comparison hypothesis, the sequence of the partial summations of $\sum a_n$ is bounded by the same constant too. We then conclude by Theorem 2.28.

iii) If $0 \leq \ell < 1$, then there exist $0 \leq c < 1$ and a natural number n_c such that

$$n \geq n_c \implies 0 < \frac{a_{n+1}}{a_n} \leq c \implies 0 < a_{n+1} \leq ca_n \leq c^{n-n_c} a_{n_c},$$

where the last inequality is obtained by induction. Let σ_h be the h -th partial summation of the geometric series of reason c , which is convergent and hence has bounded partial summations. Then, if s_k is the k -th partial summation of $\sum a_n$, we have, for every $k > n_c$,

$$0 < s_k = s_{n_c} + a_{n_c+1} + \cdots + a_k \leq s_{n_c} + a_{n_c} \sigma_{k-n_c} \leq M.$$

Hence, also $\sum a_n$ has bounded partial summation and hence, being a positive terms series, it converges to a finite sum.

If instead $\ell > 1$, then there exists \bar{n} such that

$$n \geq \bar{n} \implies \frac{a_{n+1}}{a_n} > 1,$$

and hence, for every $n \geq \bar{n}$

$$a_n > a_{n-1} > \cdots > a_{\bar{n}} > 0$$

which means that the general term of the series is not infinitesimal, and so the series is not convergent to a finite sum.

¹⁹Not necessarily the same sum. The property $a_n \leq b_n$ may be also satisfied only for all n sufficiently large, i.e. $n \geq \bar{n}$ for a suitable \bar{n} .

²⁰We need further investigation in order to understand the behavior of the series. This means that there are series for which $\ell = 1$ which are convergent as well as series which are divergent

²¹Same footnote as above.

iv) If the superior limit is less than 1, then there exist $0 \leq c < 1$ and \bar{n} such that

$$n \geq \bar{n} \implies (a_n)^{\frac{1}{n}} < c \implies a_n < c^n.$$

Hence, the series is dominated by the geometrical series of reason $0 \leq c < 1$, and so it converges.

If instead the superior limit is strictly larger than 1, then there exists a subsequence a_{n_k} such that

$$(a_{n_k})^{\frac{1}{n_k}} > 1 \implies a_{n_k} > 1 \forall k.$$

Hence the general term is not infinitesimal and the series is not convergent to a finite sum. \square

Remark 2.30 *Note that, concerning the ratio criterium, we can substitute the limit with the superior limit only in the case “less than 1”. Indeed, even if the superior limit is strictly larger than 1, the series may converge. Take for instance the series with general term given by, for $n \geq 1$,*

$$a_n = \begin{cases} \frac{1}{n^2} & \text{if } n \text{ is even,} \\ \frac{3}{n^2} & \text{if } n \text{ is odd} \end{cases}$$

(compute the superior limit of the ratio, which is equal to 3, and compare the series with the convergent one $\sum 3/(n^2)$). Obviously, in this example the limit of the ratio does not exist, being the inferior limit smaller than 1. However, for the case “larger than 1”, the limit may be substituted by the inferior limit.

It is clear which are the advantages of Proposition 2.29: 1) if we have a sufficiently large family of prototypes positive terms series, then, using the points i) and ii), we can compare them with other positive series and infer their convergence as well as their divergence; 2) the limits in iii) and iv) are sometimes not so hard to calculate. However its power is not confined to the study of positive terms series. Indeed, by virtue of the following result, Proposition 2.29 is also useful for obtaining convergence results for series with not equi-signed general term.

Proposition 2.31 *Let $\sum a_n$ be a series. We say that it is absolutely convergent to a finite sum if so is the associated series of the absolute values: $\sum |a_n|$.*

If a series is absolutely convergent to a finite sum, then it is also simply convergent to a finite sum²², that is it is convergent by itself, without the absolute values.

²²Not necessarily the same sum, of course.

Proof. It immediately follows from the inequality²³

$$\left| \sum_{n=n'}^{n''} a_n \right| \leq \sum_{n=n'}^{n''} |a_n|,$$

from the convergence of $\sum |a_n|$ and from the Cauchy criterium. \square

Of course, Proposition 2.31 is only a criterium for simple convergence, that is it does not give a necessary condition, but only a sufficient one. For example, the alternate harmonic series

$$\sum_{n=1}^{+\infty} \frac{(-1)^{n+1}}{n}$$

converges²⁴ to the finite sum $\log 2$, but it is not absolutely convergent since its series of absolute values is just the harmonic series.

Proposition 2.29 may be then also viewed as a criterium for absolute convergence, which may be useful for simple convergence too.

2.5 Rearrangements

Definition 2.23 may lead to incorrectly think that the sum of an infinite number series well behaves as the sum of a finite quantity of real numbers, and this because it links the sum of a series to the limit of the k -partial summations, which indeed are finite sum. Unfortunately, this is not correct. What does it mean that the finite sum of real numbers well behaves? It just means that such an operation (the finite summation) satisfies the well-known properties: in particular, the associative and the commutative ones. Actually, many of these properties also hold for the convergent series.

Proposition 2.32 *Let $\sum_{n=0}^{+\infty} a_n$, $\sum_{n=0}^{+\infty} b_n$ be two convergent series with sum $a, b \in \mathbb{R}$ respectively, and let c a real number. Then*

- i) the series $\sum_{n=0}^{+\infty} (a_n + b_n)$ is also convergent, with sum $a + b$;*
- ii) the series $\sum_{n=0}^{+\infty} (ca_n)$ is also convergent, with sum ca ;*
- iii) defining, for every $n \in \mathbb{N}$, $\alpha_n = a_{n/2}$ if n is even, $\alpha_n = 0$ if n is odd, then the series $\sum_{n=0}^{+\infty} \alpha_n$ is also convergent, with sum a ;*
- iv) given a strictly increasing subsequence of natural numbers $\{n_j\}_{j \in \mathbb{N}}$ with $n_0 = 0$, and defined $\alpha_j = a_{n_j} + a_{n_j+1} + \dots + a_{n_{j+1}-1}$, then the sequence $\sum_{j=0}^{+\infty} \alpha_j$ is also convergent, with sum a .*

In particular, the property ii) means that the distributive property holds for the infinite number series; property iii) (together with some of its obvious generalizations) means that, in a convergent series, between two terms, we can insert any finite quantity of zeroes without changing the sum of the series itself; property iv) means that the associative property holds.

²³Recall that the absolute value of a finite sum is less than or equal to the sum of the absolute values.

²⁴As the reader certainly well knows.

Proof. The easy proof is left as an exercise. □

What about the most popular property, the commutative one? In general it is not satisfied by a convergent series, as we are going to show.

Definition 2.33 Given a series $\sum_{n=0}^{+\infty} a_n$ a rearrangement of it is a series of the form $\sum_{n=0}^{+\infty} a_{\sigma(n)}$, where $\sigma : \mathbb{N} \rightarrow \mathbb{N}$ is a bijective function.

Example 2.34 The series

$$a_1 + a_0 + a_3 + a_2 + a_5 + a_4 + \dots$$

is a rearrangement of $\sum_{n=0}^{+\infty} a_n$ with $\sigma(n) = n + 1$ if n is even, $\sigma(n) = n - 1$ if n is odd.

It is evident that the holding of the commutative property for the series would mean that every rearrangement of a converging series is still converging to the same sum. Unfortunately such a last sentence is not true, as the following example shows.

Example 2.35 We know that the alternating harmonic series is convergent with sum $\log 2$:

$$\sum_{n=1}^{+\infty} (-1)^{n+1} \frac{1}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \dots = \log 2.$$

Hence, invoking points i), ii) and iii) of Proposition 2.32, we get (dividing by 2 and inserting zeroes)

$$\begin{aligned} \log 2 &= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \dots \\ &+ \\ \frac{\log 2}{2} &= 0 + \frac{1}{2} + 0 - \frac{1}{4} + 0 + \frac{1}{6} + 0 - \frac{1}{8} + \dots \\ &= \\ \frac{3}{2} \log 2 &= 1 + 0 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + 0 + \frac{1}{7} - \frac{1}{4} + \dots \end{aligned}$$

and, letting drop the zeroes, the last row is exactly a rearrangement of the alternating harmonic series with $\sigma : \mathbb{N} \setminus \{0\} \rightarrow \mathbb{N} \setminus \{0\}$ given by

$$\sigma(n) = \begin{cases} 1 & \text{if } n = 1, \\ n + \frac{n}{2} & \text{if } n \text{ is even,} \\ n - m - 1 & \text{if } n > 1, n = 4m + 3 \text{ or } n = 4m + 5, m \in \mathbb{N}. \end{cases}$$

Theorem 2.36 Let $\sum_{n=0}^{+\infty} a_n$ be a converging series but not absolutely convergent. Then for every (extended) real numbers

$$-\infty \leq \alpha \leq \beta \leq +\infty,$$

there exists a rearrangement of the series such that, denoting by s'_k its k -th partial summation, we have

$$\liminf_{k \rightarrow +\infty} s'_k = \alpha, \quad \limsup_{k \rightarrow +\infty} s'_k = \beta.$$

Remark 2.37 *Theorem 2.36 just says that, if the series is simply but not absolutely convergent (as the alternate harmonic series actually is), then we can always find a rearrangement which diverges to $-\infty$ (just taking $\alpha = \beta = -\infty$), a rearrangement which converges to any a-priori fixed finite sum S (just taking $\alpha = \beta = S$), a rearrangement which oscillates (just taking $\alpha \neq \beta$), and a rearrangement which diverges to $+\infty$ (just taking $\alpha = \beta = +\infty$).*

We are going to see in Theorem 2.38 that, if instead the series is absolutely convergent, than the “commutative property” holds. In some sense we can say that the “right” extension of the concept of finite sum to the series is the absolute convergence.

Proof of Theorem 2.36. For every $n \in \mathbb{N}$, we define

$$p_n = (a_n)^+, \quad q_n = (a_n)^-,$$

where $(a_n)^+ = \max\{0, a_n\}$ is the positive part of a_n and $(a_n)^- = \max\{0, -a_n\}$ is the negative part of a_n . Note that $p_n - q_n = a_n$, $p_n + q_n = |a_n|$, $p_n \geq 0$, $q_n \geq 0$. The series $\sum p_n$ and $\sum q_n$ either converge to a finite nonnegative sum or diverge to $+\infty$. By hypothesis, the series $\sum(p_n + q_n) = \sum |a_n|$ is divergent to $+\infty$, hence, the series $\sum p_n$, $\sum q_n$ cannot be both convergent (otherwise their sum must be convergent). By absurd, let us suppose that $\sum p_n$ is convergent (and so $\sum q_n$ divergent). Hence the series $\sum a_n = \sum(p_n - q_n)$ should be divergent, which is a contradiction. A similar conclusion holds if we suppose $\sum q_n$ convergent. Hence $\sum p_n$, $\sum q_n$ are both divergent to $+\infty$.

For every $j \in \mathbb{N}$ let us denote by P_j the j -th nonnegative term of $\{a_n\}$, and by Q_j the absolute value of the j -th negative term of $\{a_n\}$. Note that the sequences²⁵ P_j and Q_j are both infinitesimal as $j \rightarrow +\infty$ since they are subsequences of the sequence a_n which is the general term of a converging series; moreover, the series $\sum P_j$ and $\sum Q_j$ are both divergent since they respectively differ from $\sum p_n$ and $\sum q_n$ only for some zero terms (when $a_n \geq 0$ then $p_n = a_n$ and $q_n = 0$, when $a_n < 0$ then $p_n = 0$ and $q_n = -a_n$). Now, let us take two sequences $\{\alpha_n\}$, $\{\beta_n\}$ such that they are respectively strictly increasing and decreasing and that

$$\alpha_n \rightarrow \alpha, \quad \beta_n \rightarrow \beta \quad \text{as } n \rightarrow +\infty, \quad \alpha_n < \beta_n \quad \forall n.$$

Let m_1 be the first positive integer such that

$$P_1 + \cdots + P_{m_1} > \beta_1,$$

²⁵They are really sequences, that is we can really make j go to $+\infty$. Indeed, if for instance our construction of P_j gives only a finite number of js , then the series $\sum a_n$ is definitely given by negative terms and so its convergence would also imply its absolute convergence. Which is a contradiction.

and let be k_1 be the first integer such that

$$P_1 + \cdots + P_{m_1} - Q_1 - \cdots - Q_{k_1} < \alpha_1.$$

Again, let m_2 and k_2 be the first integers such that

$$\begin{aligned} P_1 + \cdots + P_{m_1} - Q_1 - \cdots - Q_{k_1} + P_{m_1+1} + \cdots + P_{m_2} &> \beta_2, \\ P_1 + \cdots + P_{m_1} - Q_1 - \cdots - Q_{k_1} + P_{m_1+1} + \cdots + P_{m_2} - Q_{k_1+1} - \cdots - Q_{k_2} &< \alpha_2. \end{aligned}$$

Since the series $\sum P_j$ and $\sum Q_j$ are both divergent to $+\infty$, those integers m_n and k_n always exists and hence we can repeat this procedure infinitely many times, and finally we get two subsequences of indices $\{m_n\}_n, \{k_n\}_n$. We then consider the series

$$\begin{aligned} P_1 + \cdots + P_{m_1} - Q_1 - \cdots - Q_{k_1} + P_{m_1+1} + \cdots + P_{m_2} - Q_{k_1+1} - \cdots - Q_{k_2} + \\ P_{m_2+1} + \cdots + P_{m_3} - Q_{k_2+1} - \cdots - Q_{k_3} + \cdots \end{aligned} \quad (2.3)$$

which is obviously a rearrangement of the series $\sum a_n$. Let x_n and y_n indicate the subsequence of partial summations of the series (2.3) whose last terms are P_{m_n} and Q_{k_n} , respectively. By our construction of m_n and k_n , it is²⁶

$$0 \leq x_n - \beta_n \leq P_{m_n}, \quad 0 \leq \alpha_n - y_n \leq Q_{k_n}.$$

Hence, since P_{m_n} and Q_{k_n} are infinitesimal, we get $x_n \rightarrow \beta$ and $y_n \rightarrow \alpha$ as $n \rightarrow +\infty$. Finally, by construction, α is the inferior limit and β is the superior limit of the sequence of partial summation of the series (2.3). Indeed, if ζ'_j is another subsequence of partial summations, its last term is of the form $P_{m_{n_j}+s_{n_j}}$ or $Q_{k_{n_j}+r_{n_j}}$ (it falls inside a packet of P 's or Q 's). In the first and in the second case we respectively have

$$y_{k_{n_j}} \leq \zeta'_j \leq x_{m_{n_j+1}}, \quad y_{k_{n_j+1}} \leq \zeta'_j \leq x_{m_{n_j}},$$

and we conclude by point iii) of Proposition 2.15. \square

Theorem 2.38 *Let $a \in \mathbb{R}$ be the (simple) finite sum of an absolutely convergent series ($a = \sum a_n$). Then every rearrangement of it is still absolutely convergent, and it simply converges to the same sum a .*

For proving Theorem 2.38 we first need the following lemma.

Lemma 2.39 *A series $\sum_{n=0}^{+\infty} a_n$ absolutely converges if and only if the following positive terms series are convergent*

$$\sum_{n=0}^{+\infty} (a_n)^+, \quad \sum_{n=0}^{+\infty} (a_n)^-.$$

²⁶If, for example, it is $x_n - \beta_n > P_{m_n}$, then we would have $P_1 + \cdots + P_{m_1} - Q_1 - \cdots - Q_{k_1} + \cdots + P_{m_n-1} > \beta_n$, which is a contradiction to the fact that m_n is the first integer such that...

Moreover, if the series absolutely converges, we have

$$\sum_{n=0}^{+\infty} a_n = \sum_{n=0}^{+\infty} (a_n)^+ - \sum_{n=0}^{+\infty} (a_n)^-, \quad \sum_{n=0}^{+\infty} |a_n| = \sum_{n=0}^{+\infty} (a_n)^+ + \sum_{n=0}^{+\infty} (a_n)^-. \quad (2.4)$$

Proof. Let us suppose $\sum a_n$ absolutely convergent. Hence, by the inequalities

$$(a_n)^+, (a_n)^- \leq |a_n| \quad \forall n,$$

and by comparison, we immediately get the convergence of the series $\sum (a_n)^+$ and $\sum (a_n)^-$. Vice versa, if those two series are convergent, then their sum is also convergent, and we conclude by the equality $|a_n| = (a_n)^+ + (a_n)^-$, which also gives the second equality in (2.4). The other one follows from the equality $a_n = (a_n)^+ - (a_n)^-$. \square

Proof of Theorem 2.38. Let us suppose that the series is given by nonnegative terms. Let $\sigma : \mathbb{N} \rightarrow \mathbb{N}$ be a bijective function as in the definition of rearrangement Definition 2.33, and let us consider the rearrangement given by the general term $b_n = a_{\sigma(n)}$. Let α_n and β_n be the sequences of the partial summations of the series $\sum_{n=0}^{+\infty} a_n$ and $\sum_{n=0}^{+\infty} b_n$ respectively. For every $n \in \mathbb{N}$ we define $m_n = \max\{\sigma(k) | k = 0, 1, \dots, n\}$. Since the series have nonnegative terms, we get

$$\beta_n = \sum_{k=0}^n b_k = \sum_{k=0}^n a_{\sigma(k)} \leq \sum_{j=0}^{m_n} a_j = \alpha_{m_n} \leq a.$$

This means that the partial summations β_n are bounded by a and so the rearranged series $\sum_{n=0}^{+\infty} b_n$ converges to a finite sum $b \leq a$.

Since, just regarding $\sum a_n$ as a rearrangement of $\sum b_n$, we can change the role between the two series, we then also get $a \leq b$ and so $a = b$.

If instead the series $\sum a_n$ has general term with non constant sign, then we still get the conclusion using Lemma 2.39 and observing that $\sum (b_n)^+$ and $\sum (b_n)^-$ are rearrangements of $\sum (a_n)^+$ and $\sum (a_n)^-$ respectively. \square

2.6 Sequences of functions

A natural question that may arise after the study of the sequences of real numbers is about the behavior of sequences of functions. If we have a numerable family of functions labeled by the natural numbers, $\{f_n\}_{n \in \mathbb{N}}$, can we say something about the changing of f_n (for instance of its graph) when n goes to $+\infty$? in particular, is there a “limit function” f to which f_n “tends” when $n \rightarrow +\infty$? and if all the f_n have the same property (convexity, continuity, derivability...), does such a property pass to the limit function f ? Of course, before answering to such questions, it is necessary to exactly define what does it mean that “ f_n tends to f ”. This is a crucial point, since we can give several definitions of convergence for functions, each one of them related to some particular properties to pass to the limit. In this subsection we focus only to the well known pointwise and uniform convergences.

Definition 2.40 Let $\{f_n\}_n, f$ be respectively a sequence of real-valued functions and a real-valued function, all defined on the same subset $A \subseteq \mathbb{R}$. We say that the sequence $\{f_n\}_n$ pointwise converges to f on A if, for every $x \in A$, the sequence of real numbers $\{f_n(x)\}_n$ converges to the number $f(x)$, in other words if

$$\lim_{n \rightarrow +\infty} f_n(x) = f(x) \quad \forall x \in A, \quad (2.5)$$

or, equivalently²⁷,

$$\forall \varepsilon > 0, \forall x \in A, \exists \bar{n} \in \mathbb{N} \text{ such that } n \geq \bar{n} \implies |f_n(x) - f(x)| \leq \varepsilon. \quad (2.6)$$

This is a first natural definition of convergence but, as we are going to see, rather poor.

Example 2.41 ²⁸ i) The sequence of function $f_n : [0, 1] \rightarrow \mathbb{R}$, for $n \geq 1$, defined by

$$f_n(x) = \begin{cases} nx & \text{if } 0 \leq x \leq \frac{1}{n}, \\ -nx + 2 & \text{if } \frac{1}{n} \leq x \leq \frac{2}{n}, \\ 0 & \text{if } x \geq \frac{2}{n} \end{cases}$$

is pointwise converging to $f \equiv 0$ in $[0, 1]$.

ii) The sequence $g_n : [-1, 1] \rightarrow \mathbb{R}$ defined for $n \geq 1$ by

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

is pointwise converging in $[-1, 1]$ to the function

$$g(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

iii) The sequence of functions $u_n : [0, 1] \rightarrow \mathbb{R}$ defined for $n \geq 1$ by

$$u_n(x) = \begin{cases} x & \text{if } 0 \leq x \leq \frac{1}{n}, \\ -x + \frac{2}{n} & \text{if } \frac{1}{n} \leq x \leq \frac{2}{n}, \\ 0 & \text{if } x \geq \frac{2}{n} \end{cases}$$

is pointwise converging to the function $u \equiv 0$.

iv) The sequence of functions $\varphi_n : [0, 1] \rightarrow \mathbb{R}$ defined for $n \geq 1$ by

$$\varphi_n(x) = \begin{cases} n^2x & \text{if } 0 \leq x \leq \frac{1}{n}, \\ -n^2x + 2n & \text{if } \frac{1}{n} \leq x \leq \frac{2}{n}, \\ 0 & \text{if } x \geq \frac{2}{n} \end{cases}$$

is pointwise converging in $[0, 1]$ to the function $\varphi \equiv 0$.

²⁷The reader is invited to prove the equivalence of the definitions.

²⁸The reader is invited to draw the graphs of the functions.

Let us make some considerations on the various cases reported in Example 2.41. i) The functions f_n are continuous and the limit f is continuous; for every $n \geq 1$, $\int_0^1 f_n = 1/n$ which converges (as $n \rightarrow +\infty$) to 0 which is the integral of the limit function. ii) The functions g_n are continuous, but the limit function g is not; for every $n \geq 1$, $\int_{-1}^1 g_n = 0$ which converges (as $n \rightarrow +\infty$) to 0 which is the integral of the limit function²⁹. iii) The functions u_n are continuous and the limit u is continuous; for every $n \geq 1$, $\int_0^1 u_n = 1/(n^2)$ which converges (as $n \rightarrow +\infty$) to 0 which is the integral of the limit function. iv) The functions φ_n are continuous and the limit φ is continuous; for every $n \geq 1$, $\int_0^1 \varphi_n = 1$ which converges (as $n \rightarrow +\infty$) to 1 which is not the integral of the limit function (which is equal to 0).

Hence, we deduce that, in general, the pointwise convergence is not sufficient for having the continuity of the limit (whenever the sequence is made by continuous functions) and also for passing to the limit inside the integral:

$$\lim_{n \rightarrow +\infty} \int f_n = \int f \left(= \int \left(\lim_{n \rightarrow +\infty} f_n \right) \right).$$

Anyway, as we can see from cases i) and iii) of Example 2.41, it may happen that the limit is continuous and that the integrals converge to the integral of the limit function. However, between i) and iii) there is a big difference: in i) the maximum of f_n is 1 which does not converge to the maximum of the limit function which is 0; in iii) the maximum of u_n is $1/n$ which converges to the maximum of the limit function which is 0. In some sense, we can say that in iii) the graphs of u_n “converge” to the graph of u , whereas, in i) the graphs of f_n do not converge to the graph of f ³⁰.

The convergence of maxima (as well as of minima) is of course an important property, which unfortunately is not in general guaranteed by the pointwise convergence. Hence, we need a stronger kind of convergence which, in some sense, takes account of the “convergence of the graphs”. This is the so-called uniform convergence.

Definition 2.42 *Let $A \subseteq \mathbb{R}$ be a subset, $\{f_n\}_n$ and f be, respectively, a sequence of real-valued functions and a real-valued function all defined on A . We say that the sequence $\{f_n\}$ uniformly converges to f in A if*

$$\lim_{n \rightarrow +\infty} \sup_{x \in A} |f_n(x) - f(x)| = 0 \tag{2.7}$$

or, equivalently³¹,

$$\forall \varepsilon > 0 \exists \bar{n} \in \mathbb{N} \text{ such that } n \geq \bar{n} \implies |f_n(x) - f(x)| \leq \varepsilon \forall x \in A. \tag{2.8}$$

²⁹Note that g is discontinuous but nevertheless integrable. Moreover, also the integrals of the absolute values converge: for every $n \geq 1$, $\int_{-1}^1 |g_n| = (2n - 1)/n$ which converges (as $n \rightarrow +\infty$) to 2 which is the integral of the limit function $|g|$.

³⁰In the graph of f_n there is always a pick at height 1 which stays well-distant from the graph of f which is the constant 0.

³¹The reader is invited to prove the equivalence of the definitions.

It is evident the difference between the pointwise convergence of Definition 2.40 and the uniform convergence of Definition 2.42. The former tests the convergence for every fixed point x (in other words: point-by-point), the latter tests the convergence in a “uniform” way, looking to the whole set A with its all points (this is pointed out by the presence of the supremum in (2.7) which is not present in (2.5), or, equivalently, by the fact that in (2.8) the integer \bar{n} does not depend on $x \in A$, as instead happens in (2.6), but it is chosen depending on ε only, that is in a uniform way with respect to $x \in A$).

Still referring to the “convergence of graphs”, we naively say that the uniform convergence implies the convergence of graphs since for every ε -strip around the graph of the limit function f

$$N_\varepsilon = \left\{ (x, y) \in A \times \mathbb{R} \mid |y - f(x)| \leq \varepsilon \right\},$$

for n sufficiently large, all the graphs of the f_n functions stay inside the ε -strip:

$$\exists \bar{n} \in \mathbb{N} \text{ such that } \Gamma_{f_n} \subset N_\varepsilon \quad \forall n \geq \bar{n},$$

where Γ_{f_n} is the graph of f_n over A :

$$\Gamma_{f_n} = \left\{ (x, y) \in A \times \mathbb{R} \mid y = f_n(x) \right\}.$$

Remark 2.43 *In example 2.41 the only uniformly convergent sequence is the sequence u_n of iii).*

From the Cauchy criterium Proposition 2.11, we have the following convergence criteria.

- i) $\{f_n\}$ pointwise converges to a real-valued function in A if and only if $\forall \varepsilon > 0 \forall x \in A \exists m_x \in \mathbb{N}$ such that $m_x \leq n' \leq n'' \implies |f_{n'}(x) - f_{n''}(x)| \leq \varepsilon$;*
- ii) $\{f_n\}$ uniformly converges to a real-valued function in A if and only if $\forall \varepsilon > 0 \exists m \in \mathbb{N}$ such that $m \leq n' \leq n'' \implies |f_{n'}(x) - f_{n''}(x)| \leq \varepsilon \forall x \in A$.*

Proposition 2.44 *If the sequence f_n uniformly converges to f in $A \subseteq \mathbb{R}$, then the following facts hold:*

- i) the sequence f_n pointwise converges to f in A ;*
- ii) f_n continuous for all $n \implies f$ continuous;*
- iii) if A is a bounded interval and f_n integrable on A for all n , then f is integrable on A and $\int_A f_n \rightarrow \int_A f$ as $n \rightarrow +\infty$;*
- iv) If f_n are derivable in A and if the derivatives f'_n uniformly converge on A to a function g , then f is derivable and $f' = g$ ³²;*
- v) f_n bounded for all $n \implies \sup_A f_n \rightarrow \sup_A f \in] - \infty, +\infty]$ ³³.*

³²Actually, it is sufficient the uniform convergence of the derivatives and the convergence of the functions in a fixed point. Also note that the only uniform convergence of derivable functions is not sufficient for the derivability of the limit function: think for example to a uniform approximation of the absolute value by smooth functions.

³³Similarly for what concerns the infimum. Here, the boundedness is required to give a meaning to the convergence of suprema as real (finite) numbers.

Proof. We prove only ii), iv) and v), the other ones being left as exercise. ii) Let us take $\varepsilon > 0$ and \bar{n} as in the definition of uniform convergence. Let us fix $x_0 \in A$ and take $\delta > 0$ such that

$$x \in A, |x - x_0| \leq \delta \implies |f_{\bar{n}}(x) - f_{\bar{n}}(x_0)| \leq \varepsilon.$$

Hence we get, for $x \in A$ and $|x - x_0| \leq \delta$

$$|f(x) - f(x_0)| \leq |f(x) - f_{\bar{n}}(x)| + |f_{\bar{n}}(x) - f_{\bar{n}}(x_0)| + |f_{\bar{n}}(x_0) - f(x_0)| \leq 3\varepsilon,$$

which proves the continuity of f in x_0 , which is arbitrary³⁴.

iv) Let us fix a point x of A (which here we suppose to be an open interval), and for every n let us consider the functions of $h > 0$

$$g_n(h) = \frac{f_n(x+h) - f_n(x)}{h}.$$

Since the functions f_n are derivable, by the Lagrange theorem applied to the function $f_n - f_m$ we get, for every n, m and h ,

$$g_n(h) - g_m(h) = f'_n(\xi) - f'_m(\xi) \text{ with } \xi \in]x, x+h[.$$

From this, by the uniform convergence of the functions f'_n and by the Cauchy criterium, we obtain the uniform convergence of g_n as $n \rightarrow +\infty$, with obvious limit function $\gamma : h \mapsto (f(x+h) - f(x))/h$ ³⁵. Hence, there exists an infinitesimal quantity with respect to $n \rightarrow +\infty$, $\mathcal{O}(n)$, (in particular independent from $h > 0$) such that

$$\left| \frac{f(x+h) - f(x)}{h} - \frac{f_n(x+h) - f_n(x)}{h} \right| \leq \mathcal{O}(n) \quad \forall h > 0.$$

From this we get (the first line of equalities is justified by the fact that the third limit exists)

$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \rightarrow 0} \left(\frac{f_n(x+h) - f_n(x)}{h} + \mathcal{O}(n) \right) \quad \forall n \\ \implies f'(x) &= f'_n(x) + \mathcal{O}(n) \quad \forall n \implies f'(x) = \lim_{n \rightarrow +\infty} f'_n(x), \end{aligned}$$

and the proof is concluded.

v) We prove the case where $\sup_A f \in]0, +\infty[$ ³⁶. By absurd, let us suppose the convergence of suprema not true. Hence, there exists $\bar{\varepsilon} > 0$ and a subsequence f_{n_k} such that

³⁴We need the uniform convergence in order to uniformly estimate $|f_{\bar{n}} - f|$ in different points x_0 and x .

³⁵For $h \in]0, \bar{h}[$, with $\bar{h} > 0$ such that $]x, x + \bar{h}[\subseteq A$.

³⁶We leave to the reader the other cases.

$$|\sup_A f_{n_k} - \sup_A f| > \bar{\varepsilon} \quad \forall k \in \mathbb{N}.$$

But this is obviously a contradiction to the uniform convergence. Indeed there exists \bar{k} such that $|f_{n_k}(x) - f(x)| \leq \bar{\varepsilon}/4$ for all $k \geq \bar{k}$ and for all $x \in A$ ³⁷, and hence, denoting by $\bar{x} \in A$ a point such that $f(\bar{x}) \geq \sup_A f - \bar{\varepsilon}/4$ and, for every k , $x_k \in A$ such that $f_{n_k}(x_k) \geq \sup_A f_{n_k} - \bar{\varepsilon}/4$, we get

$$\begin{aligned} \sup_A f - \frac{\bar{\varepsilon}}{4} &\leq f(\bar{x}) \leq f_{n_k}(\bar{x}) + \frac{\bar{\varepsilon}}{4} \leq \sup_A f_{n_k} + \frac{\bar{\varepsilon}}{4} \\ &\leq f_{n_k}(x_k) + \frac{\bar{\varepsilon}}{2} \leq f(x_k) + \frac{3\bar{\varepsilon}}{4} \leq \sup_A f + \frac{3\bar{\varepsilon}}{4}, \end{aligned}$$

which implies the contradiction

$$|\sup_A f_{n_k} - \sup_A f| \leq \frac{\bar{\varepsilon}}{2} \quad \forall k \geq \bar{k}.$$

□

2.7 Series of functions

Given a sequence of real valued functions $\{f_n\}_n$ defined on a set $A \subseteq \mathbb{R}$, we can consider the associated series $\sum_{n=0}^{+\infty} f_n$, and the natural question is to seek for a function $f : A \rightarrow \mathbb{R}$ which possibly represents the sum of the series, that is such that

$$\sum_{n=0}^{+\infty} f_n(x) = f(x) \quad \forall x \in A,$$

where the left-hand side is the series of real numbers $f_n(x)$.

Definition 2.45 *Given a series of real-valued functions on $A \subseteq \mathbb{R}$, $\sum_{n=0}^{+\infty} f_n$, and a function $f : A \rightarrow \mathbb{R}$, we say that the series pointwise converges to f on A if the sequence of functions given by the partial summations*

$$s_k : A \rightarrow \mathbb{R}, \quad x \mapsto s_k(x) := \sum_{n=0}^k f_n(x) \in \mathbb{R}$$

pointwise converges to f in A .

We say that the series uniformly converges to f in A if the sequence of partial summation $\{s_k\}$ uniformly converges to f in A .

³⁷If the sequence f_n uniformly converges to f then any subsequence f_{n_k} also uniformly converges to f .

Remark 2.46 From the Cauchy criterium for the numerical series Proposition 2.24, we get the following convergence criteria:

i) $\sum_{n=0}^{+\infty} f_n$ pointwise converges to a real-valued function in A if and only if

$$\forall \varepsilon > 0 \forall x \in A \exists m_x \in \mathbb{N} \text{ such that } m_x \leq n' \leq n'' \implies \left| \sum_{n=n'}^{n''} f_n(x) \right| \leq \varepsilon;$$

ii) $\sum_{n=0}^{+\infty} f_n$ uniformly converges to a real-valued function in A if and only if

$$\forall \varepsilon > 0 \exists m \in \mathbb{N} \text{ such that } m \leq n' \leq n'' \implies \left| \sum_{n=n'}^{n''} f_n(x) \right| \leq \varepsilon \forall x \in A.$$

The following criterium is a sufficient condition for uniform convergence. It is called the Weierstrass criterium, and a series which satisfies it is sometimes called totally convergent.

Proposition 2.47 Let $\sum_{n=0}^{+\infty} f_n$ be a series of real-valued functions on A , and let $\sum_{n=0}^{+\infty} M_n$ be a series of positive real numbers which is convergent to a finite sum. If

$$\sup_{x \in A} |f_n(x)| \leq M_n \forall n \in \mathbb{N},$$

then, the series of functions is uniformly convergent in A .

Proof. Since $\sum M_n$ is convergent, by the Cauchy criterium, for very $\varepsilon > 0$ there exists $m \in \mathbb{N}$ such that, for $m \leq n' \leq n''$ it is $\sum_{n=n'}^{n''} M_n \leq \varepsilon$. Hence, by our hypothesis, we get

$$\left| \sum_{n=n'}^{n''} f_n(x) \right| \leq \sum_{n=n'}^{n''} |f_n(x)| \leq \sum_{n=n'}^{n''} M_n \leq \varepsilon \forall x \in A,$$

which concludes the proof. □

Remark 2.48 The Weierstrass criterium is only a sufficient condition for the uniform convergence, that is there exist uniformly convergent series which do not fit the hypotheses of the criterium (they are not totally convergent). A very simple example is the series of constant functions for $n \geq 1$, $f_n : \mathbb{R} \rightarrow \mathbb{R}$, $f_n \equiv (-1)^n/n$. It is uniformly convergent on \mathbb{R} to the constant function $f \equiv \log 2$, but it is not totally convergent³⁸.

Remark 2.49 By Proposition 2.44 and by the definition of convergence of a series as convergence of the partial summations (which are finite sums), we immediately get the following: if $\sum_0^{+\infty} f_n$ uniformly converge to f in A then

i) f_n continuous for all $n \implies f$ continuous;

³⁸The reader is invited to prove such a sentence.

ii) (integration by series) if A is a bounded interval and f_n is integrable on A for all n , then f is integrable on A and

$$\int_A f = \sum_{n=0}^{+\infty} \int_A f_n;$$

iii) (derivation by series) if f_n are derivable in A and the series of derivatives $\sum_{n=0}^{+\infty} f'_n$ uniformly converge on A to a function g , then f is derivable and

$$f' = g = \sum_{n=0}^{+\infty} f'_n.$$

2.8 Power series

The first natural sequence of functions which are of interest are the so-called power series. They can be viewed as the natural extension of the polynomials. A polynomial p in the real variable $x \in \mathbb{R}$ is a real valued function given by a finite sum of powers of x with real coefficients:

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m,$$

where $a_i \in \mathbb{R}$, $i = 0, 1, \dots, m$, are fixed coefficients, and m is the degree of the polynomial³⁹. For instance the polynomial

$$p(x) = -1 + x^2 + 4x^3 - x^6,$$

is of degree 6 and $a_0 = -1, a_1 = 0, a_2 = 1, a_3 = 4, a_5 = 0, a_6 = -1$.

Polynomials are the functions which are better manageable, for what concerns evaluation, differentiation, integration and other elementary operations. It is then natural to extend the notion of polynomials to infinite sums of powers of the variable x , that is expressions of the form

$$\sum_{n=0}^{+\infty} a_n x^n = a_0 + a_1x + a_2x^2 + \cdots + a_n x^n + \cdots \quad (2.9)$$

where $\{a_n\}$ is a given sequence of real numbers. The expression (2.9) is called a power series and the sequence $\{a_n\}$ is the sequence of its coefficients⁴⁰. It is evident that a power series is a series of functions $\sum_{n=0}^{+\infty} f_n$ where f_0 is the constant a_0 and, for $n \geq 1$, f_n is the monomial function

$$f_n(x) = a_n x^n.$$

³⁹The highest power that occurs in its expression with non-zero coefficient.

⁴⁰Actually, the power series (2.9) is “centered” in 0. A more general expression is a power series centered in a point $x_0 \in \mathbb{R}$, which is an expression of the form $\sum_{n=0}^{+\infty} a_n (x-x_0)^n = a_0 + a_1(x-x_0) + a_2(x-x_0)^2 + \cdots$. However, in the sequel by “power series” we will always refer to expressions of the form (2.9), and all results for them can be transferred to power series centered in a general point x_0 .

Proposition 2.50 Given a power series $\sum a_n x^n$, there exists a convergence ray $\rho \in [0, +\infty]$ such that the series pointwise converges in $] - \rho, \rho[$ ⁴¹ and does not converge for any x such that $|x| > \rho$.

Moreover, if $\rho > 0$, the power series is uniformly convergent in any compact set $K \subset] - \rho, \rho[$.

Proof. Let us define

$$\rho = \begin{cases} +\infty & \text{if } \limsup_{n \rightarrow +\infty} |a_n|^{\frac{1}{n}} = 0, \\ \frac{1}{\limsup_{n \rightarrow +\infty} |a_n|^{\frac{1}{n}}} & \text{if } \limsup_{n \rightarrow +\infty} |a_n|^{\frac{1}{n}} \in]0, +\infty[, \\ 0 & \text{if } \limsup_{n \rightarrow +\infty} |a_n|^{\frac{1}{n}} = +\infty, \end{cases}$$

By the root criterium of Proposition 2.29, we immediately get that the series absolutely converges for every x such that $|x| < \rho$:

$$\limsup_{n \rightarrow +\infty} (|a_n x^n|)^{\frac{1}{n}} = |x| \limsup_{n \rightarrow +\infty} |a_n|^{\frac{1}{n}} < 1.$$

If instead $|x| > \rho \geq 0$, then, by contradiction, let us suppose that the series is convergent in x . Let us take $\xi \in \mathbb{R}$ such that $\rho < |\xi| < |x|$. Since the series converges in x , then the term $a_n x^n$ must be infinitesimal. Let us take \bar{n} such that $0 \leq |a_n x^n| < 1$ for all $n \geq \bar{n}$. Hence we get, for $n \geq \bar{n}$,

$$|a_n \xi^n| = |a_n x^n| \left| \frac{\xi}{x} \right|^n < \left| \frac{\xi}{x} \right|^n =: t^n.$$

Since $0 < t < 1$, the geometric series $\sum t^n$ is convergent and hence, by comparison, the series $\sum a_n \xi^n$ is absolutely convergent, which is a contradiction to the definition of ρ and to the root criterium⁴².

Now, we prove that the power series is uniformly convergent in $[-r, r]$ for all $0 < r < \rho$. This is immediate by the Weierstrass criterium since, we have

$$|a_n x^n| \leq |a_n| r^n \quad \forall x \in [-r, r],$$

and $\sum |a_n| r^n$ is convergent for the definition of ρ and the fact that $0 < r < \rho$. □

Proposition 2.51 Given a power series $\sum_{n=0}^{+\infty} a_n x^n$, we can consider the power series of derivatives

$$\sum_{n=1}^{+\infty} n a_n x^{n-1} = a_1 + 2a_2 x + 3a_3 x^2 + \dots,$$

⁴¹If $\rho = 0$, we take $\{0\}$. Note that, every power series is convergent in $x = 0$, and convergent to the constant a_0 (actually, any partial summation is equal to a_0).

⁴² $\limsup_{n \rightarrow +\infty} |a_n \xi^n|^{1/n} = |\xi|/\rho > 1$.

and the power series of primitives

$$\sum_{n=0}^{+\infty} \frac{a_n}{n+1} x^{n+1} = a_0 x + \frac{a_1}{2} x^2 + \frac{a_2}{3} x^3 + \frac{a_3}{4} x^4 + \dots$$

The series of derivatives and the series of primitives have the same ray of convergence as the originary one $\sum a_n x^n$.

Proof. Just note that, for $x \neq 0$ we have that

$$\begin{aligned} \sum_{n=1}^{+\infty} n a_n x^{n-1} \text{ converges if and only if } \sum_{n=1}^{+\infty} n a_n x^n = x \sum_{n=1}^{+\infty} n a_n x^{n-1} \text{ converges,} \\ \sum_{n=0}^{+\infty} \frac{a_n}{n+1} x^{n+1} \text{ converges if and only if } \sum_{n=0}^{+\infty} \frac{a_n}{n+1} x^n = \frac{1}{x} \sum_{n=0}^{+\infty} \frac{a_n}{n+1} x^{n+1} \text{ converges.} \end{aligned}$$

We then get the conclusion since

$$\limsup_{n \rightarrow +\infty} (n|a_n|)^{\frac{1}{n}} = \limsup_{n \rightarrow +\infty} (|a_n|)^{\frac{1}{n}} = \limsup_{n \rightarrow +\infty} \left(\frac{|a_n|}{n+1} \right)^{\frac{1}{n}}.$$

□

Remark 2.52 By the previous Proposition 2.51 and by Remark 2.49, we immediately get the following facts. Let $\sum_{n=0}^{+\infty} a_n x^n$ be a power series with convergence ray $\rho > 0$, and let $f :]-\rho, \rho[\rightarrow \mathbb{R}$ be its sum. Then f is continuous, derivable and integrable⁴³, in particular, for every $x \in]-\rho, \rho[$,

$$f'(x) = \sum_{n=0}^{+\infty} n a_n x^{n-1}, \quad \int_0^x f(s) ds = \sum_{n=0}^{+\infty} \frac{a_n}{n+1} x^{n+1}.$$

Moreover, f is also a C^∞ function⁴⁴ and, for every $k \in \mathbb{N}$, its k -th derivative is given by the series of the k -th derivatives.

One of the most important results about power series is the following well-known one, which we do not prove here.

Theorem 2.53 (Taylor series). Let $f :]-a, a[\rightarrow \mathbb{R}$ be a C^∞ function satisfying⁴⁵

$$\exists A > 0 \text{ such that } |f^{(n)}(x)| \leq A n! a^{-n} \quad \forall x \in]-a, a[, \quad \forall n \in \mathbb{N}. \quad (2.10)$$

Then, the power series

⁴³On every compact subinterval.

⁴⁴Derivable infinitely many times, and then all derivatives are of course continuous.

⁴⁵Here $f^{(n)}$ stays for the n -th derivative of f and, when $n = 0$, it is just f itself.

$$\sum_{n=0}^{+\infty} \frac{f^{(n)}(0)}{n!} x^n \quad (2.11)$$

pointwise converges to $f(x)$ for all $x \in]-a, a[$ (and hence uniformly in every compact subset).

In such a case, f is said an analytical function on $] - a, a[$ and the power series (2.11) is said to be the Taylor series (or Taylor expansion) of f around 0⁴⁶. Moreover, if a power series uniformly converges to f on an open interval containing 0, then it must be the power series (2.11).

The Taylor expansion can be also used for calculating the sum of several numerical series. For instance, also suitably using Theorem 2.53, it can be proved that

$$\log(1+x) = \sum_{n=1}^{+\infty} \frac{(-1)^{n+1}}{n} x^n \quad \forall x \in]-1, 1[, \quad (2.12)$$

from which, taking for example $x = -1/2$, we deduce:

$$-\log 2 = \log\left(\frac{1}{2}\right) = \log\left(1 + \left(-\frac{1}{2}\right)\right) = -\sum_{n=1}^{+\infty} \frac{1}{2^n n}$$

Actually, the equality in (2.12) also holds for $x = 1$ ⁴⁷. Hence, we also have

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

2.9 Fourier series

Theorem 2.53 gives a way for (usually locally) representing a function f by a power series. However, this is not always the best way for trying to represent a function by a series of functions. In particular, if $f : \mathbb{R} \rightarrow \mathbb{R}$ is periodic, that is there exists $T > 0$, called period, such that

$$f(x+T) = f(x) \quad \forall x \in \mathbb{R},$$

⁴⁶Actually, the condition (2.10) is only a sufficient condition for the convergence of the Taylor series to f : other conditions may be imposed to the derivatives of f , possibly restricting the set of convergence. However, let us point out that the only fact that $f \in C^\infty(]-a, a[)$ is not sufficient for being analytical, that is for the convergence to f of the Taylor series (take the function $f(x) = e^{-x^{-2}}$ if $x \neq 0$ and $f(0) = 0$ which is C^∞ on \mathbb{R} but not analytical: write down its Taylor series and look for its convergence).

Moreover, we have stated the theorem as expansion around 0, but similar statements hold for expansion around other points x_0 , changing the Taylor series in $\sum_{n=0}^{+\infty} (f^{(n)}(x_0))/(n!)(x-x_0)^n$. Indeed, the true definition of analyticity in an open interval I is that, for any point $x_0 \in I$, there exists $r > 0$ such that f is expandable in $]x_0 - r, x_0 + r[$ as Taylor series centered in x_0 .

⁴⁷This can be proved using the Leibniz criterium and the Abel theorem, which are not reported in this notes.

then, trying to represent f in the whole domain as a power series is probably not a good thing, since powers are not periodic. Moreover, as we will see, the concept of Fourier series that we are going to introduce is sometimes more "manageable" because it refers to coefficients given by integration rather than derivation, and hence it is also suitable for treating discontinuous (but integrable) functions.

Definition 2.54 Let $\{a_n\}_{n \in \mathbb{N}}$ and $\{b_n\}_{n > 0}$ be two sequences of real numbers. A Fourier series is an expression of the form

$$\frac{a_0}{2} + \sum_{n=1}^{+\infty} a_n \cos nx + \sum_{n=1}^{+\infty} b_n \sin nx, \quad (2.13)$$

and the sequences a_n, b_n are said the coefficients of the Fourier series.

A Fourier series is then a series of functions $\sum_{n=0}^{+\infty} f_n$, where $f_n : \mathbb{R} \rightarrow \mathbb{R}$ is given by

$$f_n(x) = \begin{cases} \frac{a_0}{2} & \text{if } n = 0, \\ a_n \cos nx + b_n \sin nx & \text{if } n > 0. \end{cases}$$

Proposition 2.55 If the Fourier series converges to a function f , then f is periodic with period 2π .

If the series $\sum a_n$ and $\sum b_n$ are both absolutely convergent, then the Fourier series is uniformly convergent in the whole \mathbb{R} .

Proof. The first assertion is obvious by the periodicity of \cos and \sin . The second one follows by the Weierstrass criterium Proposition 2.47 since

$$|a_n \cos nx + b_n \sin nx| \leq |a_n| + |b_n| \quad \forall x \in \mathbb{R}.$$

□

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a periodic function with period 2π , and suppose that it is integrable on $] -\pi, \pi[$. Then all the following integrals exist

$$\frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx =: a_n \quad \forall n \in \mathbb{N}, \quad \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx =: b_n \quad \forall n \in \mathbb{N}, \quad n > 0, \quad (2.14)$$

and they are called the Fourier coefficients of f . If a_n and b_n are the Fourier coefficients of f , then the series (2.13) is called the Fourier series of f .

The fact that we can calculate the Fourier coefficients of a function f and hence we can write its Fourier series does not absolutely mean that such a series converges and moreover that converges to f .

Here we report, without proof, two results concerning the pointwise and the uniform convergence of a Fourier series. However, we point out that in the theory of the Fourier series, the most suitable type of convergence is the convergence in $L^2(-\pi, \pi)$ that is the

one given by the convergence to zero of the integrals of the squared difference of the functions: f_n converges to f in $L^2(-\pi, \pi)$ if

$$\int_{-\pi}^{\pi} (f_n - f)^2 dx \rightarrow 0.$$

We do not treat such a type of convergence, however, let us note that the uniform convergence implies the L^2 convergence.

We first give a definition

Definition 2.56 *Given a function $f : \mathbb{R} \rightarrow \mathbb{R}$ and a point $x_0 \in \mathbb{R}$, we say that the function f satisfies the Dirichlet condition in x_0 if at least one of the following facts hold*

- i) f admits derivative in x_0 ;*
- ii) f is continuous in x_0 and admits right derivative and left derivative in x_0 , respectively:*

$$f'_+(x_0) = \lim_{h \rightarrow 0^+} \frac{f(x_0 + h) - f(x_0)}{h} \in \mathbb{R}, \quad f'_-(x_0) = \lim_{h \rightarrow 0^-} \frac{f(x_0 + h) - f(x_0)}{h} \in \mathbb{R};$$

- iii) f has a first-kind discontinuity in x_0 , that is*

$$\mathbb{R} \ni f^+(x_0) = \lim_{x \rightarrow x_0^+} f(x) \neq f^-(x_0) = \lim_{x \rightarrow x_0^-} f(x) \in \mathbb{R},$$

and the following limits exists in \mathbb{R}

$$\lim_{h \rightarrow 0^+} \frac{f(x_0 + h) - f^+(x_0)}{h}, \quad \lim_{h \rightarrow 0^-} \frac{f(x_0 + h) - f^-(x_0)}{h}.$$

Theorem 2.57 *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a periodic function with period 2π , which is integrable on $] -\pi, \pi[$. Moreover, let us suppose that f is piecewise continuous, that is we can split the interval $] -\pi, \pi[$ into a finite partition of subintervals $]a_i, b_i[$ such that f is continuous in every $]a_i, b_i[$ and the right and left limits $f^+(a_i)$, $f^-(b_i)$ exist in \mathbb{R} . Then we have the following.*

- i) The Fourier series of f converges in every point $x \in \mathbb{R}$ where the Dirichlet condition is satisfied, and it converges to the value*

$$s(x) = \frac{f(x)^+ + f(x)^-}{2}.$$

In particular note that, if f is continuous in x , then $s(x) = f(x)$ and so the Fourier series converges to $f(x)$.

- ii) If f is continuous, with piecewise continuous derivative, and if the Dirichlet condition holds everywhere, then the Fourier series of f uniformly converges to f in \mathbb{R} . In particular, if $f \in C^1(\mathbb{R})$, then the Fourier series uniformly converges to f .*

Remark 2.58 *It is obvious that the choice of 2π as period of the functions in this subsection is not relevant. All the same theory holds for function with different period $\tau > 0$. It is sufficient to replace (2.13) by*

$$\frac{a_0}{2} + \sum_{n=1}^{+\infty} a_n \cos\left(\frac{2\pi n}{\tau}x\right) + \sum_{n=1}^{+\infty} b_n \sin\left(\frac{2\pi n}{\tau}x\right),$$

and (2.14) by

$$a_n = \frac{2}{\tau} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(x) \cos\left(\frac{2\pi n}{\tau}x\right) dx, \quad b_n = \frac{2}{\tau} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} f(x) \sin\left(\frac{2\pi n}{\tau}x\right) dx.$$

As for the Taylor series, also the Fourier series may be used for calculating the sum of several numerical series. For instance, let us consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$ periodic of period 2π and such that $f(x) = x^2$ for $x \in [-\pi, \pi]$. If we calculate its Fourier coefficients we find⁴⁸

$$a_0 = \frac{2\pi^2}{3}, \quad a_n = \frac{4(-1)^n}{n^2}, \quad b_n = 0 \quad \forall n \in \mathbb{N} \setminus \{0\}.$$

The function f is continuous and satisfies the Dirichlet condition in all points $x \in \mathbb{R}$ and so

$$f(x) = \frac{\pi^2}{3} + 4 \sum_{n=1}^{+\infty} \frac{(-1)^n}{n^2} \cos(nx) \quad \forall x \in \mathbb{R}.$$

Hence we have

$$\begin{aligned} \pi^2 = f(\pi) &= \frac{\pi^2}{3} + 4 \sum_{n=1}^{+\infty} \frac{(-1)^n}{n^2} \cos(n\pi) = \frac{\pi^2}{3} + 4 \sum_{n=1}^{+\infty} \frac{1}{n^2} \implies \sum_{n=1}^{+\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \\ 0 = f(0) &= \frac{\pi^2}{3} + 4 \sum_{n=1}^{+\infty} \frac{(-1)^n}{n^2} \cos(0) = \frac{\pi^2}{3} + 4 \sum_{n=1}^{+\infty} \frac{(-1)^n}{n^2} \implies \sum_{n=1}^{+\infty} \frac{(-1)^{n+1}}{n^2} = \frac{\pi^2}{12}. \end{aligned}$$

2.10 Historical notes

Since the time of the Greeks, scientists faced the problem of working with something similar to a summation of infinite terms. The most important of such scientists was Archimedes of Syracuse (287-212. B.C.). In several works Archimedes rigorously exploited the so-called method of exhaustion, which however goes back to Eudoxus of Cnidus (408-355 B.C.). The method of exhaustion may be seen as the first attempt of calculating the areas of regions of the plane which are delimited by some curves⁴⁹, and hence it is the

⁴⁸The function f is even, that is $f(-x) = f(x)$ for all x , and so it is obvious that the coefficients b_n must be all equal to zero, since they are the coefficients of the “odd part” of f : the “sinus part”.

⁴⁹i.e. not necessarily segments.

precursor of the integral calculus. Of course our modern concept of integrals which, very naively speaking, may be seen as a generalization of the concept of summation to a set of more than infinitely countable terms⁵⁰, at that time was just replaced by the sum of a larger and larger number of addenda: the areas of some suitably inscribed figures.

The most famous application of the method of exhaustion by Archimedes was the quadrature of a parabolic segment. Using modern notations and describing only a particular case⁵¹, the problem is the following: to calculate the area of the bounded plane region between the x -axis and the parabola of equation $y = 1 - x^2$. Nowadays we immediately answer:

$$Area = \int_{-1}^1 (1 - x^2) dx = \frac{4}{3}.$$

Let us sketch the argumentation of Archimedes. Consider the triangle of vertices $(-1, 0)$, $(1, 0)$ and $(0, 1)$: it has area equal to 1 and it is inscribed in our region. To such a triangle, add two more triangles with vertices, respectively: $(-1, 0)$, $(0, 1)$, $(-1/2, 3/4)$ and $(1, 0)$, $(1/2, 3/4)$, $(0, 1)$. The total area of those new triangles is $1/4$. Hence the new polygonal figure, which is still inscribed in our parabolic region, has area equal to $1 + 1/4$. Now we can add four more triangles and this is done adding the vertices: $(\pm 1/4, 15/16)$ and $(\pm 3/4, 7/16)$. Such four triangles has a total area equal to $1/16$. Hence the new inscribed polygonal figure has area equal to

$$1 + \frac{1}{4} + \frac{1}{16}.$$

By his great ability in calculating, Archimedes showed that, at every steps, we can add to the polygonal figure, a number of triangles with total area equal to one-quarter of the added area at the previous step. Hence after n steps, the area of the inscribed polygonal figure is⁵²

$$1 + \frac{1}{4} + \frac{1}{16} + \cdots + \frac{1}{4^n} = \sum_{k=0}^n \frac{1}{4^k},$$

and he also proved that such a finite (but numerous) summation is always strictly less than $4/3$, but, on the other side, at very step the sum becomes closer to $4/3$.

The method of “exhaustion” just stays to indicate that we intend to exhaust the area of the parabolic region by inserting more and more triangles. If we really exhaust the area, then nowadays we immediately get

$$Area = \sum_{k=0}^{\infty} \frac{1}{4^k} = \frac{1}{1 - \frac{1}{4}} = \frac{4}{3}, \quad (2.15)$$

⁵⁰One per every point of the integration interval.

⁵¹Archimedes actually solved the problem for a more general situation.

⁵²In a modern notation.

which is the right answer. Actually, Archimedes did not compute the area as the sum of infinite addenda, since, at that time and also for more that 2000 years after, the infinite processes were not taken as possible⁵³: there were many troubles with the concept of “infinity”. However, Archimedes did prove that the area is equal to $4/3$, and in what he did we can now see the basic concept of modern definition of limit. Indeed, by a “reductio ad absurdum” he proved that the area cannot be larger as well as cannot be smaller than $4/3$: he showed that if the area were larger of $4/3$ then he could inscribe triangles until the total area was more than $4/3$ and this is a contradiction to the fact that every finite sum is smaller than $4/3$; on the other side, if the area were smaller than $4/3$, then he could find k such that, at the $k - th$ step the inscribed area is larger: another contradiction.

In this procedure we really recover our definition of limit: when we take an arbitrary $\varepsilon > 0$ and we prove that, for sufficiently large n , the partial summations satisfies $|s_n - 4/3| \leq \varepsilon$, we are actually saying that the sum of the series cannot be larger as well as cannot be smaller than $4/3$.

As we already said, Archimedes and the other Greeks, did not make limits, as we understand nowadays, since the right definition of “infinity” was far from coming. What Archimedes did was to conjecture the real value of the area and then to show that it cannot be any different value. In doing that he used the “series” in the sense that, adding more and more terms, he arrived to a contradiction.

However, Archimedes arrived very close to make a limit, and so to write the second equality in (2.15). Despite to this fact, in the following of the human history, mathematicians continued to have some troubles with the infinite series. From the middle age, they certainly knew that some infinite series have a finite sum and some other ones have an infinite sum. For instance, Nicole Oresme (France, 1323-1382) was probably the first to prove that the harmonic series has a sum equal to $+\infty$. Pietro Mengoli (Bologna, 1626 - 1686) showed that the alternating harmonic series has a finite sum equal to $\log 2$. Leonhard Euler (Basel 1707 - St. Petersburg 1783) first proved that the harmonic series of power 2, $(\sum 1/n^2)$, has a finite sum equal to $\pi^2/6$. Brook Taylor (England, 1685 - 1731) made his study on expansion of functions and discovered the (after him) called Taylor series. However, a right and rigorous definition of what should be the “convergence” of a series was yet to come until the XIX century. The lacking of a right definition of “convergence” may be recognized in the fact that, mathematicians retained that all the series should have a sum (finite or infinite) and so the treating of what nowadays we call “oscillating series” was avoided or ambiguous and mistaken. For instance, one problem pointed by Daniel Bernoulli (The Netherlands 1700 - Switzerland 1782) was the following: the alternating series of 1’s and -1 ’s can be obtained by setting $x = 1$ in the following power series:

⁵³Or at least as rigorous

$$\begin{aligned}
1 - x + x^2 - x^3 + x^4 - x^5 + \dots &= \frac{1}{1+x}, \\
1 - x + x^3 - x^4 + x^6 - x^7 + \dots &= (1-x)(1+x^3+x^6+\dots) = \frac{1-x}{1-x^3} = \frac{1}{1+x+x^2}, \\
1 - x^2 + x^3 - x^5 + x^6 - x^8 + \dots &= (1-x^2)(1+x^3+x^6+\dots) = \frac{1-x^2}{1-x^3} = \frac{1+x}{1+x+x^2},
\end{aligned}$$

and we respectively obtain, as sum of the alternating series, $1/2$, $1/3$, $2/3$ which, for Bernoulli and his colleagues (but for us today, too) was very unsatisfactory. What was mistaken in the reasoning is to assume that the series has a sum. A similar erratum reasoning is the following one⁵⁴: let $S = 1 + 1 - 1 + 1 - 1 + \dots$ be the sum of the series. Then we have

$$S = 1 - 1 + 1 - 1 + 1 - 1 + \dots = 1 - (1 - 1 + 1 - 1 + \dots) = 1 - S \implies S = \frac{1}{2}.$$

We now know that such a series has not a sum, however note that if it had finite sum, then it would be $1/2$: the mean value between 0 and 1 which are the values of the alternating sequence of finite summations.

Series without sum were a serious problem. However, also the series with finite sum were still a little bit obscure. Indeed, in the XVIII century, mathematicians were aware of the distinction between sum with a large number of addenda and infinite series. Moreover, they also knew that it is not always possible to treat infinite series as true summations, since infinite series do not well behave as standard summation, but they also knew that, when that possibility is given, then a lot powerful results are at disposal, such as the fact that the integral of a power series can be found integrating term by term, just as in a finite summation. They developed a sense for what was and was not legitimate, and they paid attention to not overcome such a limit. However, at the beginning of the XIX century this was not more sufficient, that limit must be overcome, and in particular, clarified.

One of the events that contributed to such a breakthrough was the work of Jean Baptiste Joseph Fourier (France, 1768 - 1830). In 1807, Fourier published his study on the propagation of heat in solid bodies and showed that, if the distribution of the heat at a part of the boundary of the body is given by a sum of trigonometric functions then also the solution is given by a sum of trigonometric functions. Anyway, the constant function $f(x) \equiv 1$ cannot be written as sum of trigonometric functions, but, on the other side, a constant temperature may be applied to any part of the boundary. To overcome this difficult, Fourier showed that, for $-1 < x < 1$ it is possible to write the constant function $f(x) \equiv 1$ as infinite sum of trigonometric functions

$$1 = \frac{4}{\pi} \sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{2n-1} \cos\left(\frac{(2n-1)\pi x}{2}\right) \quad (2.16)$$

⁵⁴Which seems to be made even by Euler.

and that the solution is then an infinite series of trigonometric functions. Moreover, he also gave a method for finding the (Fourier) coefficients for any function. The contemporaries of Fourier were very suspicious about such a solution, in particular about the infinite sum of trigonometric functions. Such a series were already appeared in the past years, but now Fourier bring them to a mandatory attention by the mathematicians. One of the most evident problems with such series was about the meaning of the function they represent. If we look to (2.16) we see that, if we take $1 < x < 2$ then the series has sum equal to -1 . In practice, the function represented by the series should be equal to the constant 1, but it is also periodic oscillating between 1 and -1 (and also 0 in the integer points). This fact was not acceptable. At that moment “functions” only means polynomials, powers, logarithms, trigonometric functions and any multiplication and linear combination of them and their inverse. Moreover the functions should have a “continuous” graph. Finally, “functions” are something differentiable infinitely many times and, knowing its derivative in a point, means knowing the function everywhere (Taylor series). This is not obviously true for the function in (2.16). This is not admissible. Moreover, in finding his solution, Fourier assumed that you can integrate and differentiate the trigonometric series term by term. However, the solution proposed by Fourier was actually modeling a real physical problem, and so it cannot be rejected without trying to understand why it seems to work. The answer for rejecting it is then that there are troubles with the convergence of series of trigonometric functions.

Actually the trouble was with the “convergence” in general. Another french mathematician, Augustin Luis Cauchy (1789 - 1857) started such a revolutionary process in re-founding what do we mean by infinite series, convergence and limit. He indeed, in some sense, went back to Archimedes’ ideas, in particular to the fact that (what now we call the limit) must be definitely neither larger nor smaller than the approximating sequence, that is our modern “ $\varepsilon - \delta$ ” definition. However, Cauchy bypassed Archimedes and the Greeks, whose vision was more of less static (they already knew the value of the possible limit and proved that it cannot be otherwise) since he introduced a “dynamic” concept of convergence and limit⁵⁵. The work of Cauchy was completed by other mathematicians, firstly by the german ones Georg Friedrich Bernhard Riemann (1826-1866) and Karl Theodor Wilhelm Weierstrass (1815-1897).

⁵⁵However, in his formulation, Cauchy avoided terms as “time” and “velocity” which were already used by some previous mathematicians and which rely on a “physical” point of view of functions, limits and derivatives.

3 Real numbers and ordered fields

In this section we are going to point out the properties of the set \mathbb{R} of the real numbers, both from an algebraic and from an analytical point of view. At this stage, we suppose that the reader already knows what the real numbers are.

The first properties of the real numbers are of course the algebraic properties. Let us start from these ones.

Definition 3.1 1) Given a nonempty set A , an operation on it is a function from the cartesian product $A \times A$ to A .

2) A nonempty set G is said to be a commutative (or abelian) group if there is an operation $\varphi : (a, b) \mapsto \varphi(a, b) =: a + b$ on it which satisfies the following properties:

$$\begin{aligned}
 2i) \text{ (associative property)} & \quad a + (b + c) = (a + b) + c \quad \forall a, b, c \in G, \\
 2ii) \text{ (neutral element)} & \quad \exists G \ni b =: 0 \text{ such that } a + 0 = a \quad \forall a \in G, \\
 2iii) \text{ (opposite element)} & \quad \forall a \in G \exists G \ni b =: (-a) \text{ such that } 0 = a + (-a) =: a - a, \\
 2iv) \text{ (commutative property)} & \quad a + b = b + a \quad \forall a, b \in G.
 \end{aligned} \tag{3.1}$$

3) A nonempty set F is said a field if there are two operations on it: $\varphi_1(a, b) =: a + b$, $\varphi_2(a, b) =: ab$ such that F is an abelian group with respect to φ_1 , $F \setminus \{0\}$ ⁵⁶ is an abelian group with respect to φ_2 , and if the following compatibility condition between the two operations is satisfied:

$$\text{(distributive property)} \quad c(a + b) = ca + cb \quad \forall a, b, c \in F \tag{3.2}$$

By the associative and commutative properties it follows that the neutral element of a group is unique as well as the opposite element. Indeed if $0, 0'$ are two neutral elements, we get

$$0 = 0 + 0' = 0' + 0 = 0',$$

and for any a, b, c :

$$a + c = b + c \implies a + c - c = b + c - c \implies a = b,$$

which implies the uniqueness of the opposite element since, if b, b' are two opposite elements of a , we have

$$a + b = 0 = a + b' \implies b = b'.$$

⁵⁶Here, 0 is the neutral element of φ_1 . The group $F \setminus \{0\}$ has a multiplicative representation, that is we indicate by 1 its unique neutral element and by a^{-1} or even by $1/a$ the opposite element. Also note that, by definition, it follows that the restriction of φ_2 to $(F \setminus \{0\}) \times (F \setminus \{0\})$ is an operation on $F \setminus \{0\}$; in particular, if $a, b \in F \setminus \{0\}$ then $ab \neq 0$ and $a^{-1} \neq 0$. This immediately implies that $ab = 0 \implies a = 0$ or $b = 0$.

Also note that, in the case of a field F , it is, for any $a \in F$,

$$0 \cdot a = a \cdot 0 = 0.$$

Indeed $a \cdot 0 = a(0 + 0) = a \cdot 0 + a \cdot 0 \implies a \cdot 0 = 0$ and similarly for $0 \cdot a$.

Proposition 3.2 *If F is a field, then, for every $a, b \in F$ the following holds: i) $(-1)a = -a$ ⁵⁷; ii) $(-a)^{-1} = -a^{-1}$;*

Proof. i) $a + (-1)a = a(1 - 1) = a \cdot 0 = 0$; ii) $(-a)(-a^{-1}) = (-1)(-1)(aa^{-1}) = 1$.

□

With the usual operations of sum and multiplication, the set of the real numbers \mathbb{R} is a field. Other well-known fields are the set of rational numbers \mathbb{Q} ⁵⁸ and the set of complex numbers \mathbb{C} ⁵⁹. The set of integers \mathbb{Z} is not a field⁶⁰, but the quotient sets $\mathbb{Z}/p\mathbb{Z}$, with $p \in \mathbb{N}$ a prime number, are all fields⁶¹. There is a main difference between $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ from one side and $\mathbb{Z}/p\mathbb{Z}$ from the other side: the first ones are infinite fields, that is with infinitely many elements, the second one are finite fields, that is with finitely many elements: just p elements. Mathematical analysis is mainly devoted to the study of the infinite fields. As we are going to see in this notes, there are also big differences between \mathbb{Q}, \mathbb{R} and \mathbb{C} : the first is ordered but not complete, the second is ordered and complete, the third is complete⁶² but not ordered. The (real) mathematical analysis is mainly devoted to the study of \mathbb{R} .

3.1 Ordering and Archimedean properties

As anticipated at the end of the previous subsection, \mathbb{R} is ordered.

Definition 3.3 *An order relation (or an ordering) on a nonempty set A (which is then said to be an ordered set) is a relation, denoted by “ \leq ”, between its elements⁶³ such that, for every $a, b, c \in A$,*

$$\begin{aligned} \text{(transitive property)} \quad & a \leq b, b \leq c \implies a \leq c, \\ \text{(reflexive property)} \quad & a \leq a, \\ \text{(anti-symmetric property)} \quad & a \leq b, b \leq a \implies a = b. \end{aligned}$$

The order relation is said to be total if, for every couple of elements $a, b \in A$, it is always true that at least one of the following relations hold: $a \leq b$ or $b \leq a$. If the ordering

⁵⁷From which $(-1)(-1) = 1$.

⁵⁸With the same operations as in \mathbb{R} .

⁵⁹With the known operations which extend the ones in \mathbb{R} .

⁶⁰ $\mathbb{Z} \setminus \{0\}$ is not a group with respect to the multiplication: there is no opposite element.

⁶¹With natural extension of sum and multiplication from elements of \mathbb{Z} to the elements of $\mathbb{Z}/p\mathbb{Z}$, which are equivalence classes.

⁶²In the sense of metric space.

⁶³A “relation” is often defined as a subset R of the cartesian product $A \times A$, so that a is in relation with b if and only if $(a, b) \in R$.

is total, then A is said to be totally ordered, if instead the ordering is not total⁶⁴ then A is said partially ordered.

Given an order relation on A we can always define a strict order relation on A , denoted by “ $<$ ”, as

$$a < b \iff a \leq b \text{ and } a \neq b.$$

Such a strict order relation satisfies the transitive property only.

If F is a field and it is also a totally ordered set, we say that F is an ordered field if the following compatibility conditions between ordering and operations hold

$$\begin{aligned} \text{i)} \quad a \leq b &\implies a + c \leq b + c \quad \forall c \in F, \\ \text{ii)} \quad a \leq b &\implies ac \leq bc \quad \forall 0 \leq c \in F. \end{aligned} \tag{3.3}$$

Remark 3.4 Instead of writing $a \leq b$, we will often say “ a is smaller than or equal to b ”, as well as “ b is larger than or equal to a ”, and also, in a more ambiguous manner, we will sometimes say “ a is smaller than b ” as well as “ b is larger than a ” for indicating both $a \leq b$ and $a < b$. Finally, $a \geq b$ (as well as $a > b$) will mean $b \leq a$ (as well as $b < a$).

Example 3.5 We give a simple (but important for the sequel of these notes) example of a partially, but not totally, ordered set. Let $x_0 \in \mathbb{R}$ and define the set

$$\mathcal{A} = \left\{ A \subseteq \mathbb{R} \mid \exists r > 0 \text{ such that }]x_0 - r, x_0 + r[\subseteq A \right\},$$

so that the elements of \mathcal{A} are subsets of \mathbb{R} containing open intervals centered in x_0 ⁶⁵. We define the following relation in \mathcal{A} :

$$\forall A_1, A_2 \in \mathcal{A}, \quad A_1 \leq A_2 \iff A_2 \subseteq A_1 : \tag{3.4}$$

that is the “inverse inclusion order”. The reader is invited to prove that (3.4) defines a partial, but not total, ordering on \mathcal{A} .

With the usual ordering, both \mathbb{Q} and \mathbb{R} are ordered fields. A first natural total ordering in the complex field \mathbb{C} is the lexicographical one: given $z_1 = a_1 + ib_1, z_2 = a_2 + ib_2 \in \mathbb{C}$

$$z_1 \leq z_2 \iff a_1 < a_2 \text{ or } (a_1 = a_2 \text{ and } b_1 \leq b_2).$$

It is immediate to see that this is a total ordering on \mathbb{C} , but also that, with such an ordering, \mathbb{C} is not an ordered field: the property ii) of (3.3) does not hold⁶⁶.

⁶⁴That is there exist two elements $a, b \in A$ such neither $a \leq b$ nor $b \leq a$ holds true.

⁶⁵Neighborhoods of x_0 ?

⁶⁶ $0 \leq i$, but $i^2 = i \cdot i = -1 < 0 = 0 \cdot i$.

Proposition 3.6 *If F is an ordered field, then, the implications in (3.3) are indeed equivalence, and moreover, for every $a, b \in F$, we have: i) $a \leq b \iff -b \leq -a$ ⁶⁷, ii) $a^2 := aa \geq 0$ ⁶⁸, iii) $a > 0 \iff a^{-1} > 0$, iv) $0 < a < b \iff 0 < b^{-1} < a^{-1}$, v) $a \leq b \iff a - b \leq 0$, vi) $a < 0, b > 0 \implies ab < 0$.*

Proof. For the first sentence, just take in (3.3-i.ii) $c = 0$ and $c = 1$ respectively. i) $a \leq b \implies a + (-a - b) \leq b + (-a - b) \implies -b \leq -a$; the opposite implication similarly comes starting from $-b \leq -a$; ii) If $a \geq 0$ then it is obvious by point ii) of (3.3), if instead $a < 0$ then $-a > 0$ and so, being $a = -(-a)$, $-a^2 = (-1)(-1)(-1)(-a)^2 = -(-a)^2 \leq 0 \implies a^2 \geq 0$; iii) if $a^{-1} < 0$ then $1 = a(a^{-1}) < 0$ which is absurd; the opposite implication similarly comes from the equality $(a^{-1})^{-1} = a$; iv) if by absurd hypothesis we have $0 < a^{-1} < b^{-1}$, then multiplying by $ab > 0$ we obtain a contradiction, v) $a \leq b \implies 0 = a - a \leq b - a$, vice versa $0 \leq b - a \implies a = 0 + a \leq (b - a) + a = b$; vi) if it was $ab > 0$ then $a = abb^{-1} > 0$, which is absurd. \square

Remark 3.7 *From the point ii) of Proposition 3.6 we immediately get the inequality $x^2 + 1 > 0$ for all $x \in F$. This means that, if F is an ordered field, then the equation $x^2 + 1 = 0$ has no solutions in F . In particular, this implies that there is not an ordering on the complex field \mathbb{C} which makes \mathbb{C} an ordered field.*

Many of the usual definitions which we already know and use for the real numbers can be suitably transferred to any ordered field (for instance all the definitions in Definition 2.2 (majorants, minorants, bounded subsets, superior extremum, inferior extremum)⁶⁹), moreover we may also define:

$$\begin{aligned}
 & \text{(non-negative values (resp: positive values))}: & x \in F \text{ such that } 0 \leq x \text{ (resp. } 0 < x), \\
 & \text{(non-positive values (resp: negative values))}: & x \in F \text{ such that } 0 \geq x \text{ (resp. } 0 > x), \\
 & \text{(absolute value)} & \forall x \in F, |x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}
 \end{aligned} \tag{3.5}$$

It can be also easily proved that the absolute value satisfies the following expected properties for every $x, y \in F$:

$$\begin{aligned}
 & |x| \geq 0, \quad |x| = 0 \iff x = 0, \\
 & |x + y| \leq |x| + |y|, \quad \left| |x| - |y| \right| \leq |x - y|, \\
 & |xy| = |x||y|, \quad \text{if } y \neq 0: \quad \left| \frac{x}{y} \right| = \frac{|x|}{|y|}.
 \end{aligned}$$

Proposition 3.8 *If F is an ordered field, then it contains infinitely many elements. Moreover, F is dense, that is*

$$\forall x, y \in F, x < y \implies \exists z \in F \text{ such that } x < z < y. \tag{3.6}$$

⁶⁷Which in particular implies: either $a < 0$ or $-a < 0$ for all $a \in F \setminus \{0\}$.

⁶⁸This of course implies $1 = 1^2 > 0$, which is of great importance.

⁶⁹Obviously, the existence of infimum and of supremum is not necessarily guaranteed.

Proof. The first sentence easily comes from the following inequalities⁷⁰

$$0 < 1 < 1 + 1 < 1 + 1 + 1 < 1 + 1 + 1 + 1 < \dots$$

Let us prove the second sentence. First note that

$$0 < a < 1 \implies 0 < ax < x \quad \forall x > 0.$$

Indeed, if it is not the case we would have $0 < x \leq ax$ which implies, multiplying by x^{-1} , $a \geq 1$. Now, let us take $x < y$. By point iv) of Proposition 3.6 we have $0 < (1 + 1)^{-1} < 1$. Hence, if $x = 0$ we immediately get

$$x = 0 < z = (1 + 1)^{-1}y < y,$$

and similarly for the case $y = 0$. If instead $0 < x < y$, let us first note the following

$$(1 + 1)^{-1} = 1 - (1 + 1)^{-1},$$

indeed: $(1 + 1)(1 - (1 + 1)^{-1}) = (1 + 1) - 1 = 1$. Hence $z = (1 + 1)^{-1}(x + y)$ satisfies the inequalities. Indeed, for instance, by absurd

$$x > (1 + 1)^{-1}(x + y) \implies (1 + 1)^{-1}y < x - (1 + 1)^{-1}x = (1 - (1 + 1)^{-1})x = (1 + 1)^{-1}x,$$

which is a contradiction to $0 < x < y$ ⁷¹. Similarly for the case $x < y < 0$. The case $x < 0 < y$ is obvious. □

Remark 3.9 *If F is a finite field⁷², then there is not an ordering which makes F an ordered field.*

Definition 3.10 *As suggested by the proof of Proposition 3.8, if F is an ordered field, we can consider the following infinite subset*

$$\{0, 1, 1 + 1, 1 + 1 + 1, 1 + 1 + 1 + 1, \dots\}$$

which we obviously call the subset of the natural elements of F and, from now on, we will denote

$$2 := 1 + 1, \quad 3 := 1 + 1 + 1 = 2 + 1, \quad 4 := 1 + 1 + 1 + 1 = 3 + 1 \quad \dots$$

We then say that $x \in F$ is an integer element if x or $-x$ is natural; we say that x is a rational element if there exist two integers m, n with $n \neq 0$ such that $x = m/n := mn^{-1}$.

⁷⁰By point i) of (3.3) and by the inequality $0 < 1$ we get $1 = 0 + 1 < 1 + 1$ and so on.

⁷¹Recall that $(1 + 1)^{-1}$ is positive.

⁷²In particular $\mathbb{Z}/p\mathbb{Z}$ with p prime.

We denote by $\mathbb{N}_F, \mathbb{Z}_F, \mathbb{Q}_F$ the set of naturals, of integers and of rationals of F , respectively. We then have: $\mathbb{N}_F \subset \mathbb{Q}_F \subset \mathbb{Q}_F$

Also note that the sum and the product between rational elements follow the usual rules as for the rational numbers:

$$\begin{aligned} \frac{m_1}{n_1} + \frac{m_2}{n_2} &= m_1 n_1^{-1} + m_2 n_2^{-1} = (n_1 n_2)(n_1 n_2)^{-1} (m_1 n_1^{-1} + m_2 n_2^{-1}) \\ &= (n_1 n_2)^{-1} (m_1 n_2 + m_2 n_1) = \frac{m_1 n_2 + m_2 n_1}{n_1 n_2}, \\ \frac{m_1}{n_1} \frac{m_2}{n_2} &= (m_1 n_1^{-1}) (m_2 n_2^{-1}) = n_1^{-1} n_2^{-1} m_1 m_2 = \frac{m_1 m_2}{n_1 n_2}. \end{aligned}$$

Remark 3.11 Another possible, equivalent, definitions of the natural elements are the following. We first give the definition of inductive subset $A \subseteq F$: A is inductive if $x + 1 \in A$ whenever $x \in A$. We say that $x \in F$ is natural if it belongs to all inductive subsets $A \subseteq F$ which contain the neutral element 0 . Hence, we can say that \mathbb{N}_F is the smallest inductive subset of F containing 0 .

Also note that, for every $x \in F$ and $n \in \mathbb{N}_F$, the element $nx \in F$ is exactly given by the sum of x with itself for n -times, where this second n is the familiar natural number $n = 1 + 1 + \dots + 1 \in \mathbb{N}$. So, when we write nx we can think to n as element of \mathbb{N}_F as well as element of the natural numbers \mathbb{N} , that is, in this particular context, we can identify \mathbb{N}_F with \mathbb{N} . Actually, as we are going to see in the sequel, such an identification is always possible via some suitable isomorphisms, and of course it extends to the identifications of \mathbb{Z}_F with the integer numbers \mathbb{Z} and of \mathbb{Q}_F with the rational numbers \mathbb{Q} .

Proposition 3.12 If F is an ordered field, then \mathbb{N}_F is well-ordered, that is every non-empty subset $A \subseteq \mathbb{N}_F$ has a minimum element:

$$\emptyset \neq A \subseteq \mathbb{N}_F \implies \exists \bar{n} \in A \text{ such that } \bar{n} \leq n \forall n \in A.$$

Proof. Let us suppose that $\emptyset \neq A \subseteq \mathbb{N}_F$ has no minimum. Since $0 = \min \mathbb{N}_F$ ⁷³, this implies that $0 \notin A$ ⁷⁴. Let us define

$$P = \left\{ n \in \mathbb{N}_F \mid m \notin A \forall m = 0, 1, \dots, n \right\}.$$

The set P is not empty since $0 \in P$ and moreover it is inductive. Indeed, if $n \in P$ then also $n + 1 \in P$, otherwise $n + 1$ would be the minimum of A . Hence, we must have $P = \mathbb{N}_F$ since \mathbb{N}_F is the smallest non-empty inductive set in F . This implies $A = \emptyset$ which is a contradiction. \square

Definition 3.13 An ordered field F is said to be an archimedean field if \mathbb{N}_F is not bounded from above. Equivalently if:

$$a, b \in F, b > 0 \implies \exists n \in \mathbb{N} \text{ such that } nb > a. \quad (3.7)$$

⁷³ \mathbb{N}_F has only non-negative elements, because $\mathbb{N}_F \setminus \{x < 0\}$ is an inductive set containing 0 and so $\mathbb{N}_F \subseteq \mathbb{N}_F \setminus \{x < 0\}$.

⁷⁴Otherwise 0 would also be the minimum of A .

The equivalence between the two definitions is easy

$$\begin{aligned} \implies) \quad & nb < a \forall n \in \mathbb{N} \implies n < ab^{-1} \forall n \in \mathbb{N}_F, \text{ contradiction;} \\ \impliedby) \quad & n < M \forall n \in \mathbb{N}_F \implies n \cdot 1 < M \forall n \in \mathbb{N}, \text{ contradiction.} \end{aligned}$$

The reader is invited to prove the following two facts: if F is an archimedean field then

$$\text{i) } \inf \left\{ \frac{1}{k} = k^{-1} \mid k \in \mathbb{N}_F \setminus \{0\} \right\} = 0; \quad (3.8)$$

ii) denoting by \mathbb{A} one of the subsets $\mathbb{N}_F, \mathbb{Z}_F, \mathbb{Q}_F$ or F , we say that $A \subseteq \mathbb{A}$ is bounded from above (respectively, bounded from below, bounded) in \mathbb{A} if there exists $x \in \mathbb{A}$ such that $a \leq x$ (respectively, $x \leq a$, $a \leq |x|$) for all $a \in A$; then, for every $A \subseteq \mathbb{A}$,

A is bounded from above (respectively: bounded from below, bounded) in $\mathbb{A} \iff$
 A is bounded from above (respectively: bounded from below, bounded) in F .

The field of rational numbers \mathbb{Q} and the field of real numbers \mathbb{R} are archimedean fields.

Proposition 3.14 *If F is an ordered field, then \mathbb{Q}_F is an archimedean field.*

Proof. The fact that \mathbb{Q}_F is an ordered field is obvious⁷⁵. Let us take $\bar{m}, \bar{n} \in \mathbb{N}_F, \bar{n} \neq 0$. Since $\bar{n} \geq 1$ we then immediately get

$$\mathbb{Q}_F \ni \frac{\bar{m}}{\bar{n}} \leq \bar{m} < \bar{m} + 1 \in \mathbb{N}_F,$$

which implies that $\mathbb{N}_{\mathbb{Q}_F} = \mathbb{N}_F$ is not bounded from above in \mathbb{Q}_F (by the arbitrariness of \bar{m}/\bar{n}). \square

Proposition 3.15 *If F is an archimedean field, then \mathbb{Q}_F is dense in F , that is*

$$\forall a, b \in F \text{ with } a < b, \exists q \in \mathbb{Q}_F \text{ such that } a < q < b.$$

Proof. It is not restrictive to assume $a > 0$ ⁷⁶. Since $b - a > 0$, by (3.8) there exists $n \in \mathbb{N}_F$ such that

$$0 < \frac{1}{n} < b - a.$$

Let m be the smallest element of \mathbb{N}_F such that

$$\frac{m}{n} > a.$$

⁷⁵Note that it is also obvious that $\mathbb{N}_{\mathbb{Q}_F} = \mathbb{N}_F$.

⁷⁶If $a = 0$ then we conclude by (3.8), if $a < 0 < b$ then just take $q = 0$, if $a < b \leq 0$ then just invert signs and inequalities.

Note that such m exists by the well-ordering property, and that it is certainly larger than 0. Hence, by the definition of m , we have

$$a < \frac{m}{n} = \frac{m-1}{n} + \frac{1}{n} < a + (b-a) = b,$$

and we conclude taking $q = m/n$. \square

Proposition 3.16 *Let F be an ordered field. A section of F is a couple of subsets of F , (A, B) , such that $A, B \neq \emptyset, F, A \cup B = F$ and $x \in A, y \in B \implies x \leq y$.*

If F is archimedean and (A, B) is a section of F , then there exist two sequences $\{a_n\}_{n \in \mathbb{N}}, \{b_n\}_{n \in \mathbb{N}}$ in F such that

$$\begin{aligned} & a_n \in A, b_n \in B \quad \forall n \in \mathbb{N}, \quad \{a_n\} \text{ is increasing, } \{b_n\} \text{ is decreasing,} \\ & \forall c \in F, c > 0 \exists n \in \mathbb{N} \text{ such that } b_n - a_n < c. \end{aligned}$$

Proof. This is very similar to the proof of the Bisection Lemma 2.10. Let us take $a_0 \in A$ and $b_0 \in B$ and, for every $n \in \mathbb{N}$, define

$$c_n = \frac{a_n + b_n}{2}, \quad a_{n+1} = \begin{cases} c_n & \text{if } c_n \in A, \\ a_n & \text{otherwise} \end{cases}, \quad b_{n+1} = \begin{cases} b_n & \text{if } c_n \in A, \\ c_n & \text{otherwise} \end{cases}.$$

The first properties of the sequences are obvious. For the last one, we can certainly assume that $a_n \neq b_n$ for all n , otherwise we must definitely have $a_n = b_n = c_n$ and the statement becomes obvious. Hence, we have

$$b_n - a_n = (b_0 - a_0) (2^n)^{-1},$$

where 2^n means the element $2 \in F$ multiplied n -times by itself. Since $2^n \geq n$ for every n ⁷⁷ (where the second n is the “ n ” of \mathbb{N}_F), then, by (3.8) we have

$$\inf \left\{ (b_0 - a_0)(2^n)^{-1} \mid n \in \mathbb{N} \right\} = (b_0 - a_0) \inf \left\{ (2^n)^{-1} \mid n \in \mathbb{N} \right\} = 0,$$

which concludes the proof. \square

3.2 Isomorphisms and complete ordered fields

Definition 3.17 *An ordered operative structure is a non-empty set A endowed by a number of operations $\varphi_A^1, \dots, \varphi_A^{n_A}$ and of an ordered relation \leq_A .*

We say that two ordered operative structures A, B , are isomorphic if $n_A = n_B$ and there exists a bijective function $\psi : A \rightarrow B$ such that

$$\begin{aligned} & a \leq_A b \iff \psi(a) \leq_B \psi(b) \quad \forall a, b \in A, \\ & \psi(\varphi_A^i(a, b)) = \varphi_B^i(\psi(a), \psi(b)) \quad \forall a, b \in A \quad \forall i = 1, \dots, n_A = n_B. \end{aligned}$$

Such a function ψ is called an isomorphism between the structures A and B .

⁷⁷Let $A \subseteq F$ be the sets of all positive natural elements such that $2^n < n$ and suppose that is is not empty. Since $2^1 > 1$, then the minimum n_0 of A , which exists because of the well-ordering property, satisfies $n_0 \geq 2$. Hence, $2^{n_0} < n_0$ and $2^{n_0-1} \geq n_0 - 1$ and we get a contradiction.

Remark 3.18 *It is quite obvious that, if F and G are ordered fields, then \mathbb{N}_F and \mathbb{N}_G are isomorphic, and then \mathbb{Q}_F and \mathbb{Q}_G are isomorphic.*

Definition 3.19 *An ordered field F is said to be a complete ordered field if every non-empty subset $A \subset F$ which is bounded from above, admits superior extremum in F , that is there exists $\bar{a} \in F$ such that*

$$a \leq \bar{a} \quad \forall a \in A; \quad x < \bar{a} \implies \exists a \in A \text{ such that } x < a.$$

From the definition it easily follows that \bar{a} is unique, moreover the “supremum property” of Definition 3.19 is equivalent to the obvious dual “inferior property”. We will also write $\sup A$ for the supremum and $\inf A$ for the infimum.

Proposition 3.20 *If an ordered field F is complete then it is also archimedean. If an ordered field F strictly contains a complete ordered field G ⁷⁸, then F is not archimedean.*

Proof. Let us prove the first sentence. By absurd, let us suppose that F is not archimedean, and so \mathbb{N}_F is bounded from above. Then, since \mathbb{N}_F is not empty, by completeness, there exists $M = \sup \mathbb{N}_F \in F$. Since $M - 1 < M$ there exists $n \in \mathbb{N}_F$ such that $M - 1 < n \leq M$. But \mathbb{N}_F is an inductive subset and so $n + 1 \in \mathbb{N}_F$. Since $n + 1 > M$ we get a contradiction.

Let us prove the second sentence. First of all note that $\mathbb{N}_F = \mathbb{N}_G$. Let us take $\xi \in F \setminus G$, and suppose that $\xi > 0$, which is not restrictive. If there is not a natural element $\bar{n} \in \mathbb{N}_F$ such that $\xi < \bar{n}$, then \mathbb{N}_F is bounded from above and F is not archimedean. Otherwise, the set

$$A = \left\{ x \in G \mid x < \xi \right\}$$

is not empty (it contains 0) and is bounded from above in G ⁷⁹. Since G is complete there exists $\bar{x} = \sup A \in G$ ⁸⁰. For any $n \in \mathbb{N}_F \setminus \{0\}$, we then have⁸¹

$$\bar{x} - \frac{1}{n} < \xi < \bar{x} + \frac{1}{n} \implies 0 < |\xi - \bar{x}| < \frac{1}{n},$$

which implies that (3.8) does not hold in F , and hence that F is not archimedean. \square

Proposition 3.21 *Let F be an archimedean ordered field. Then there exists a complete ordered field \tilde{F} which contains a subfield isomorphic to F .*

⁷⁸Of course with the same operations and ordering.

⁷⁹There exists $\bar{n} \in \mathbb{N}_F = \mathbb{N}_G$ such that $\xi < \bar{n}$; hence if $x \in A$ it cannot be $x > \bar{n}$.

⁸⁰Here the supremum “is in G ”, that is $\bar{x} \in G$ is the minimum among all elements $w \in G$ such that $x \leq w$ for all $x \in A$.

⁸¹Since $\bar{x} = \sup A$ in G , for every n there exists $\bar{\xi} \in A$ such that $\bar{x} - 1/n < \bar{\xi} < \xi$. On the other side, by contradiction, if it exists n such that $\xi \geq \bar{x} + 1/n$, then we would have $\xi > \bar{x} + 1/(n+1) \in G$ from which $\bar{x} + 1/(n+1) \in A$ which is a contradiction to the fact that \bar{x} is the supremum of A . Note that we need all these considerations because $\xi \notin G$ and so we may have both $\bar{x} > \xi$ and $\bar{x} < \xi$. But, what we certainly have is that whenever we subtract or add $1/n$ to \bar{x} we get under and get over ξ respectively.

Proof.

We define a *half-section* of F as any subset $L \subset F$ such that

$$L \neq \emptyset; L \neq F; L \text{ has not maximum}; x, y \in F, x \in L, y < x \implies y \in L.$$

An example of a half-section is the left-half-line

$$L = \{x \in F \mid x < r\},$$

where $r \in F$ is fixed⁸². We define

$$\tilde{F} := \{L \mid L \text{ is a half-section of } F\}.$$

Order relation.

$$\forall L, M \in \tilde{F}, L \leq M \iff L \subseteq M.$$

It is easy to see that this is a total relation on \tilde{F} .

Existence of supremum. Let $A \subseteq \tilde{F}$ be a non-empty bounded above subset. We define the subset of F

$$H = \bigcup_{L \in A} L \subseteq F, \text{ that is } x \in H \iff \exists L \in A \text{ such that } x \in L.$$

Let us prove that H is a half-section. Obviously, it is not empty. Since A is bounded from above, then there exists \bar{L} such that $L \subseteq \bar{L}$ for every $L \in A$, and so H cannot be the whole F . It does not have maximum since, for every $x \in H$ it is $x \in L$ for some $L \in A$, and L has not maximum, hence there exists $x' \in L \subseteq H$ such that $x < x'$. Finally, let us take $x \in H$ and $y < x$, hence $x \in L \subseteq H$ and so $y \in L \subseteq H$. Hence H is a half-section, that is $H \in \tilde{F}$. Now we prove that $H = \sup A$. By definition, it is obvious that $L \leq H$ for every $L \in A$. Moreover, if L' is a majorant of A , then $L \subseteq L'$ for all $L \in A$, and so, by definition, $H \subseteq L'$. Hence, H is the supremum.

Sum. We define a sum in \tilde{F} by

$$L + M = \left\{ z \in F \mid \exists x \in L, y \in M \text{ such that } z = x + y \right\} \quad \forall L, M \in \tilde{F}.$$

First, we have to prove that $L + M$ is a half-section of F , that is an element of \tilde{F} . Let us prove only the fourth condition. If $z \in L + M$, then by definition there exist $x \in L$ and $y \in M$ such that $z = x + y$. Hence, if $w < z$, then $w - z < 0$ and so $x + w - z < x$, which implies $x + w - z \in L$, from which $w = x + w - z + y \in L + M$.

⁸²However, the half-sections are not all of this kind, otherwise, in some sense, we would replicate F itself. Indeed, think for instance to the case $F = \mathbb{Q}$, we also consider the half-section $\{q \mid q \leq 0 \text{ or } (q > 0 \text{ and } q^2 < 2)\}$, which is not of the previous kind.

Now, we sketch the proof that such a sum in \tilde{F} makes \tilde{F} an abelian group⁸³. It can be proved that the sum is commutative and associative, and also that the half-section

$$\tilde{0} = \left\{ x \in F \mid a < 0 \right\},$$

is the neutral element. Let us show the construction of the opposites with more details. If $L \in \tilde{F}$, then we define the sets

$$\hat{L} = \left\{ x \in F \mid -x \notin L \right\}, \quad L^* = \begin{cases} \hat{L} & \text{if } \hat{L} \text{ has no maximum,} \\ \hat{L} \setminus \{\max \hat{L}\} & \text{otherwise} \end{cases}$$

It can be proved that L^* is a half-section of F and that $L + L^* \subseteq \tilde{0}$. Let us prove the opposite inclusion. Note that $(L, F \setminus L)$ is a section of F , and take two sequences $\{a_n\}, \{b_n\}$ as in Proposition 3.16⁸⁴. Hence we have, for every $n > 0$:

$$a_n \in L, \quad -b_n \in \hat{L}, \quad -b_n - \frac{1}{n} \in L^*, \quad (3.9)$$

and, by the property of the sequences a_n and b_n and by the archimedean property, for every $z < 0$, there exists n such that

$$0 < (b_n - a_n) + \frac{1}{n} < -z \implies z < -(b_n - a_n) - \frac{1}{n} \in L + L^* \implies z \in L + L^*, \quad (3.10)$$

and so $\tilde{0} \subseteq L + L^*$.

Product. We define the following product in \tilde{F} , for only positive elements $L, M > 0$:

$$L \cdot M := \tilde{0} \cup \{0\} \cup \left\{ z \in F \mid \exists 0 < x \in L, 0 < y \in M \text{ such that } z = xy \right\}$$

Again, it can be proved that it is well defined (the product is a half-section), that it is associative commutative, distributive, and that the neutral element is

$$\tilde{1} = \left\{ x \in F \mid x < 1 \right\}.$$

Moreover, similarly as above, in particular using the archimedean property, for every $0 < L \in \tilde{F}$ we can define

$$\hat{L} = \tilde{0} \cup \{0\} \cup \left\{ 0 < x \in F \mid x^{-1} \notin L \right\}, \quad L^* = \begin{cases} \hat{L} & \text{if } \hat{L} \text{ has no maximum,} \\ \hat{L} \setminus \{\max \hat{L}\} & \text{otherwise} \end{cases},$$

and prove that $L^* = L^{-1}$.

⁸³Most of such a proof is tedious and long. We only sketch the passages where the archimedean property of F is crucial.

⁸⁴Here, the archimedean property of F plays an essential role.

Once we have defined a product between positive elements, then we can extend it to negative elements by defining, for instance,

$$L \cdot M := -(L \cdot (-M)) \quad \forall L > 0, M < 0.$$

Compatibility with the ordering. This is almost immediate. For example, let us prove the compatibility between product of positive terms and ordering. Let $0 < L \leq M$ and $0 < C$ be chosen, and take $z \in L \cdot C$. If $z \leq 0$, then obviously $z \in M \cdot C$ too. Let us suppose $z > 0$. Hence, there exist $0 < a \in L$ and $0 < b \in C$ such that $z = ab$. Since $L \subseteq M$ then we also have $a \in M$ and so $z = ab \in M \cdot C$ too.

Hence, we have proved that \tilde{F} is a complete ordered field. It remains to prove that it contains a subfield which is isomorphic to F . But the isomorphism is immediate, it is the restriction of

$$\psi : F \rightarrow \tilde{F} \quad x \mapsto \psi(x) = \left\{ z \in F \mid z < x \right\}.$$

to its image as codomain⁸⁵.

□

Let us collect some of the results till now obtained and also some other obvious facts⁸⁶:

- i) F, G ordered fields $\implies \mathbb{Q}_F, \mathbb{Q}_G$ archimedean and isomorphic
 - ii) F, G isomorphic $\implies (F \text{ archimedean/complete} \iff G \text{ archimedean/complete})$
 - iii) F, G complete, $G \subseteq F \implies F = G$,
 - iv) F, G archimedean and isomorphic $\implies \tilde{F}, \tilde{G}$ isomorphic
 - v) F complete $\implies F, \tilde{F}$ isomorphic
- (3.11)

Proposition 3.22 *If F is an archimedean field, then \tilde{F} is isomorphic to $\tilde{\mathbb{Q}}_F$.*

Proof. Let us define the following function

$$\psi : \tilde{\mathbb{Q}}_F \rightarrow \tilde{F}, \quad L \mapsto L' = \left\{ x \in F \mid \exists q \in L, x < q \right\}. \quad (3.12)$$

It is easy to prove that L' is indeed an half-section of F and that ψ is injective. For instance, about the injectivity: if $L, M \in \tilde{\mathbb{Q}}_F$ and $L < M$, then let us take $q_1, q_2 \in M \setminus L$ with $q_1 < q_2$, which exist since, at least one $q \in M \setminus L$ exists, and M has no maximum. Hence, being F dense, there exists $x \in F$ such that $q_1 < x < q_2$ which of course implies $x \in M' \setminus L'$, that is $L' \neq M'$.

⁸⁵What is immediate is that such a restriction of ψ is a good candidate to be isomorphism. To prove that it really is an isomorphism requires, again, some patience.

⁸⁶iii) holds because if G is not the whole F , then F strictly contains a complete field and so it cannot be complete; iv): if $\psi : F \rightarrow G$ is isomorphism, then $\tilde{\psi} : \tilde{F} \rightarrow \tilde{G}$, $\tilde{\psi}(L) = \{\psi(x) : x \in L\}$ is isomorphism; v) is true since \tilde{F} contains a subfield isomorphic to F which is complete, and hence it must coincide with that subfield.

The fact that ψ maintains the orderings is easy. About the operations, let us prove as example the compatibility with the product of positive elements. Let $L, M \in \tilde{\mathbb{Q}}_F$, $L, M > 0$, and prove that $(LM)' = L'M'$. It is sufficient to work with positive elements $0 < x \in F$. If $x \in (LM)'$ then there exist $0 < q_L \in L, 0 < q_M \in M$ such that $x < q_L q_M$. From this we get that $x/q_L < q_M$ and hence $x/q_L \in M'$. Since we obviously have $q_L \in L'^{87}$, we finally get $x = (x/q_L)q_L \in L'M'$. On the other side, if $x \in L'M'$ then there exist $x_L \in L', x_M \in M'$ such that $x = x_L x_M$. Moreover, there exist $q_L \in L, q_M \in M$ such that $x_L < q_L, x_M < q_M$. Since $q_L q_M \in LM$ we conclude by $x = x_L x_M < q_L q_M$.

Hence, $\tilde{\mathbb{Q}}_F$ is isomorphic to $\psi\left(\tilde{\mathbb{Q}}_F\right)$ which is a complete subfield of \tilde{F} . Since \tilde{F} is complete too, then $\tilde{F} = \psi\left(\tilde{\mathbb{Q}}_F\right)$, and hence \tilde{F} is isomorphic to $\tilde{\mathbb{Q}}_F$. \square

Now, we can add a new point vi) to the collection (3.11)

$$vi) \quad F \text{ archimedean} \implies \tilde{F} \text{ isomorphic to } \tilde{\mathbb{Q}}_F. \quad (3.13)$$

Here is the main result of this section.

Theorem 3.23 *If F and G are two complete ordered fields, then they are isomorphic.*

Proof. Just using the six points in (3.11)–(3.13), denoting by “ \simeq ” the relation of being isomorphic, and noting the obvious fact that such a relation is transitive, we immediately get

$$F \simeq \tilde{F} \simeq \tilde{\mathbb{Q}}_F \simeq \tilde{\mathbb{Q}}_G \simeq \tilde{G} \simeq G.$$

\square

Remark 3.24 *By Theorem 3.23 we can say that “there exists only one complete ordered field”, where “uniqueness” must be intended as “they are all isomorphic”. In some sense, two different complete fields are the same objects with the same rules just only painted with different colors. We call such a unique field “the field of the real numbers” and we denote it by \mathbb{R} .*

Moreover, as conclusion of this section, we also deduce that: 1) \mathbb{R} is the unique complete ordered field; 2) \mathbb{Q} is the smallest ordered field: indeed every ordered field contains \mathbb{Q} , and it is also archimedean (and so the smallest archimedean); 3) every archimedean ordered field F is contained between \mathbb{Q} and \mathbb{R} : $\mathbb{Q} \subseteq F \subseteq \mathbb{R}^{88}$; 4) if an ordered field is contained between \mathbb{Q} and \mathbb{R} , then it is archimedean (because so is \mathbb{R}); 5) if an ordered field strictly contains \mathbb{R} then, it is not archimedean.

We end this section reporting two examples of an archimedean field strictly contained between \mathbb{Q} and \mathbb{R} and of a non-archimedean field, respectively.

⁸⁷This is because L has no maximum.

⁸⁸because $F \subseteq \tilde{F} = \mathbb{R}$.

Example 3.25 We consider the set

$$\mathbb{Q}[\sqrt{2}] = \left\{ x \in \mathbb{R} \mid \exists a, b \in \mathbb{Q} \text{ such that } x = a + \sqrt{2}b \right\}.$$

It is evident that $\mathbb{Q} \subset \mathbb{Q}[\sqrt{2}] \subset \mathbb{R}$ with strict inclusions. Moreover, $\mathbb{Q}[\sqrt{2}]$ is an archimedean ordered field. Since it is a subset of the archimedean ordered field \mathbb{R} , it is sufficient to prove that it is a field, that is closed for the sum and the product and their inversions. However, first of all note that, since $a, b \in \mathbb{Q}$, $a + \sqrt{2}b = 0$ if and only if $a = b = 0$, and the same also holds for $a^2 - 2b^2 = (a + \sqrt{2}b)(a - \sqrt{2}b)$.

$$(a + \sqrt{2}b) + (\alpha + \sqrt{2}\beta) = (a + \alpha) + \sqrt{2}(b + \beta) \in \mathbb{Q}[\sqrt{2}];$$

$$-(a + \sqrt{2}b) = (-a) + \sqrt{2}(-b) \in \mathbb{Q}[\sqrt{2}];$$

$$(a + \sqrt{2}b)(\alpha + \sqrt{2}\beta) = (a\alpha + 2b\beta) + \sqrt{2}(b\alpha + a\beta) \in \mathbb{Q}[\sqrt{2}];$$

$$\text{if } a \neq 0 \text{ or } b \neq 0: \frac{1}{a + \sqrt{2}b} = \frac{1}{a + \sqrt{2}b} \frac{a - \sqrt{2}b}{a - \sqrt{2}b} = \frac{a}{a^2 - 2b^2} - \sqrt{2} \frac{b}{a^2 - 2b^2} \in \mathbb{Q}[\sqrt{2}]$$

Example 3.26 Let us consider the following set of rational functions

$$F = \left\{ \frac{P(x)}{Q(x)} \mid P, Q \text{ polynomials in the variable } x \in \mathbb{R}, Q \neq 0 \right\}$$

Note that, for every rational function P/Q in F there exists a neighborhood of $+\infty$, that is a half-line $[m, +\infty[$, such that P/Q is defined on it (i.e. $Q(x) \neq 0$ for all $x \in [m, +\infty[$). Obviously F is endowed of a sum and of a multiplication which make it a field:

$$\frac{P_1}{Q_1} + \frac{P_2}{Q_2} = \frac{P_1Q_2 + P_2Q_1}{Q_1Q_2}, \quad \frac{P_1}{Q_1} \frac{P_2}{Q_2} = \frac{P_1P_2}{Q_1Q_2},$$

the neutral element for the sum is the null function (which corresponds to $P \equiv 0$ with every Q) and the unity is the constant function 1 (which corresponds to $P = Q \equiv 1$).

It is not restrictive to suppose that all the elements of F have the denominator Q such that its leading coefficient (the one of the maximum power) is strictly positive⁸⁹. We introduce the following total ordering in F :

$$\frac{P_1}{Q_1} \leq \frac{P_2}{Q_2} \iff \frac{P_2}{Q_2} - \frac{P_1}{Q_1} \text{ has the numerator with non-negative leading coefficient.}$$

Let us prove that it is an order relation. First of all note that a polynomial P has null leading coefficient if and only if it is the null polynomial (the null constant function), and this fact immediately gives the reflexivity and the anti-symmetry. Just a calculation gives that the sum of non-negative elements gives a non-negative element. Hence we get the transitivity:

$$\frac{P_1}{Q_1} \leq \frac{P_2}{Q_2}, \frac{P_2}{Q_2} \leq \frac{P_3}{Q_3} \implies \frac{P_3}{Q_3} - \frac{P_1}{Q_1} = \left(\frac{P_3}{Q_3} - \frac{P_2}{Q_2} \right) + \left(\frac{P_2}{Q_2} - \frac{P_1}{Q_1} \right) \geq 0$$

⁸⁹If not, just multiply P and Q by -1 .

Checking the compatibility of the ordering with the operations is also easy.

Hence F is an ordered field. It is obvious that \mathbb{N}_F is given by the constant functions $f \equiv n$ with $n \in \mathbb{N}$ (corresponding to $P \equiv n$, $Q \equiv 1$). Let us note that

$$1 > 0, \quad \frac{1}{x} > 0, \quad \text{and} \quad 1 - \frac{1}{x} = \frac{x-1}{x} > 0 \implies 0 < \frac{1}{x} < 1.$$

But we also have $n \cdot 1/x < 1$ for every $n \in \mathbb{N}$, indeed

$$1 - \frac{n}{x} = \frac{x-n}{x} > 0.$$

Hence, F is not archimedean (otherwise $n \cdot 1/x \geq 1$ for some n).

Remark 3.27 *We know that the archimedean fields are contained between \mathbb{Q} and \mathbb{R} . We also know that if an ordered field strictly contains \mathbb{R} , then it is not archimedean. The field in Example 3.26 indeed strictly contains an isomorphic copy of \mathbb{R} : the constant functions $f \equiv \alpha \in \mathbb{R}^{90}$. However, it is not necessary for a non-archimedean field to contain the real numbers. Just take the field of rational functions given by polynomials with rational coefficients: it is a non-archimedean field but it does not contain the real numbers⁹¹.*

3.3 Choose your axiom, but choose one!

In the previous sections we have stated and proved several results about ordered field, and in particular, we arrived to prove that every two complete fields are isomorphic, which, in some sense, means that there exists at most only one complete ordered field, which we call the real numbers field. The natural question that must be done at this stage is whether the real numbers exist or, in other words, whether a complete ordered field exists or not. We already know that, if an archimedean ordered field exists then a complete ordered field exists too. Hence the question may be shifted to a "lower level": does an archimedean ordered field exist?, which is in some sense equivalent to ask whether the field of rational numbers exists or not. However, the rational numbers can be naturally constructed starting from the natural numbers. Hence the question may be posed in a further "lower level": do the natural numbers exist?

⁹⁰By the way, note the little bit surprising fact, but coherent with an already done observation in a previous footnote, $\inf_{\mathbb{R}}\{\alpha \in \mathbb{R} \mid \alpha \geq (1/x)\} = 0 < (1/x)$, where $\inf_{\mathbb{R}}$ means the infimum inside the subfield \mathbb{R} .

⁹¹This last example may well explain why we need the archimedean property for the construction of the complete field \tilde{F} as done using the half-sections. It is almost clear that the constant function $\sqrt{2}$ is a "hole" of the field of the rational functions with rational coefficients. Hence, consider the half-section L given by the sets of the elements smaller than $\sqrt{2}$. Its opposite with respect to the sum should be the half-section L^* of elements smaller than $-\sqrt{2}$. But the "hole $\sqrt{2}$ " also contains all the elements which differ from $\sqrt{2}$ by an infinitesimal quantity (i.e. by a non-zero quantities whose absolute value is less than every $1/n$): for example, the function $\sqrt{2} - 1/x$. But then, it is not more true that $L + L^* = \tilde{0}$. Indeed, we cannot obtain $-1/x$ as sum of two elements of L and L^* , because we cannot "approximate" $\sqrt{2}$ more than any non-negative element as in the archimedean case: we always remain far from $\sqrt{2}$ of at least the quantity $1/x > 0$ (compare with (3.9),(3.10)).

At the end of the XIX century, an Italian mathematician, together with his school, Giuseppe Peano (Cuneo 1858- Torino 1932) intensively studied the problem of formalizing all the structure of mathematics, in particular the set of numbers and their properties. His goal was to develop a formal language able to contain the mathematical logic and all the results of the most important sectors of mathematics. In doing that he also introduced most of the symbols we are still using nowadays for formulas, such as \in , \subset , \cup , \cap , \forall , \exists . This was one of the power of his method: to avoid as more as possible any "metaphysical" language and to rely all the theory on rigorous symbolism, together with going to the essential properties of the structure under study. In particular, for what concerns the construction of natural numbers, he identified the basic characteristic properties of the set of natural numbers which are now worldwide recognized as the Peano axioms. Such axioms are the following; there exists a non empty set \mathbb{N} such that:

- i) (*successor function*) there exist an element $0 \in \mathbb{N}$ and an injective function $s : \mathbb{N} \rightarrow \mathbb{N} \setminus \{0\}$ (and $s(n)$ is said the successor of n),
- ii) (*Principle of Induction*) if a collection of predicates $P(n)$, one per every $n \in \mathbb{N}$, satisfies both

- a) $P(0)$ is true,
- b) $P(n)$ true $\implies P(s(n))$ true,

then $P(n)$ is true for all $n \in \mathbb{N}$.

If \mathbb{N} is a set satisfying the Peano axioms, then it can be endowed, in a natural manner, of a sum, a product and an order with the obvious compatibility conditions between them. Indeed, let us define $1 := s(0)$, then we can define, for all $n, m \in \mathbb{N}$

$$\begin{aligned} n + 1 &:= s(n), & n + (m + 1) &:= (n + m) + 1, \\ n \cdot 1 &:= n, & n(m + 1) &:= nm + n, \\ 2 &:= 1 + 1 = s(1) = s(s(0)) > 1 > 0. \end{aligned}$$

It can be shown that, using the Principle of Induction, these formulas well-define a sum, a product and a total ordering on \mathbb{N} .

If \mathbb{N}, \mathbb{N}' are two sets satisfying the Peano axioms, then there is an isomorphism between them, again defined using the Principle of Induction,

$$\psi : \mathbb{N} \rightarrow \mathbb{N}', \quad 0 \mapsto \psi(0) = 0', \quad s(0) \mapsto \psi(s(0)) = s'(0'), \quad s(n) \mapsto s'(\psi(n)),$$

where s, s' are the two successor functions. Hence, we can say that, if a set satisfying the Peano axioms exists, then it is unique: they are all isomorphic.

We now prove that, if F is an ordered field, then \mathbb{N}_F satisfies the Peano axioms. Of course there is an element $0 \in \mathbb{N}_F$ and the injective function s is $s : x \rightarrow x + 1$. The Principle of Induction holds for the fact that \mathbb{N}_F is well-ordered, that is every non empty subset of \mathbb{N}_F has a minimum element (see Proposition 3.12). Indeed, let us consider the set $A \subseteq \mathbb{N}_F$ such that $n \in A$ if and only if $P(n)$ is false, and, by absurd, suppose that A is non-empty. Then it must have a minimum element \bar{n} , which of course is not 0 since $P(0)$ is true. Hence, \bar{n} is not the minimum of \mathbb{N}_F and so it has a predecessor, which means

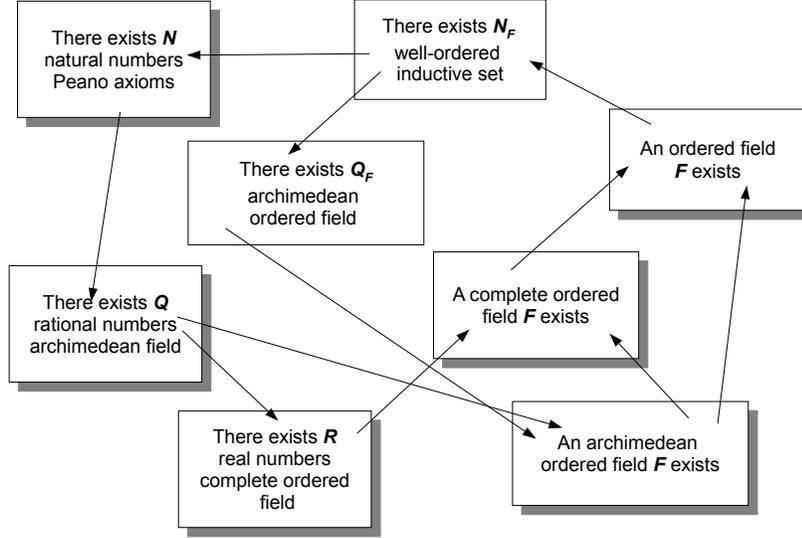


Figure 3: The sentences in the shaded rectangles may be chosen as axioms. Please, choose one.

that $\bar{n} - 1 \in \mathbb{N}_F$ and, of course, $\bar{n} - 1 \notin A$ for the minimality of \bar{n} . Hence $P(\bar{n} - 1)$ is true and, by the inductive assumption, we also get $P(\bar{n})$ true, which is a contradiction since $\bar{n} \in A$. Hence it must be $A = \emptyset$, that is $P(n)$ true for all $n \in \mathbb{N}_F$.

When we assume the Peano axioms, in particular the existence of such a set, the natural numbers, then it is easy to arrive to the construction of the set of rational numbers which turns out to be an archimedean field. From the field of rational numbers, by Proposition 3.21, we get the existence of a complete field which, by Theorem 3.23, turns out to be the unique complete field: the real numbers. On the other way, we can assume that the real numbers, as we imagine and know, exist (in particular that they satisfy the superior extremum property, that is that they form a complete field), and we immediately arrive to the existence of the natural numbers (i.e. $\mathbb{N}_{\mathbb{R}}$), satisfying the Peano axioms. Moreover, as explained above, we can also assume the existence of a generic ordered field F and we obtain the existence of a set satisfying the Peano axioms (\mathbb{N}_F) from which we obtain the existence of an archimedean field (\mathbb{Q}_F), and then obtain the existence of a complete ordered field ($\tilde{F} = \tilde{\mathbb{Q}}_F$). In similar way, we can assume the existence of an archimedean ordered field (typically the rational numbers \mathbb{Q}) and then we get the existence of the natural numbers satisfying the Peano axioms and the existence of a complete field (the real numbers, by uniqueness). In conclusion, we cannot avoid to assume an axiom between the following ones: a) the natural numbers exists satisfying the Peano axioms, b) the rational numbers exist forming an archimedean field, c) the real numbers exist forming a complete field, d) an ordered field exists, e) an archimedean field exists, f) a complete

field exists. Whichever axiom we have chosen, all the other ones become theorems⁹² and, in any case, we arrive to the real numbers as unique complete field. What is mandatory is to make such a choice: since nothing comes from nothing, one axiom must be chosen and all the rest will come. Figure 3 tries to explain this fact.

3.4 Historical notes and complements

As we have seen in the previous sections, the archimedean property of the rational (and so of the real) numbers plays an important role. In particular the consequence (3.8) is of great importance also in practice, because, for instance, it concerns the calculation of limits. Another (similar) important consequence is the density of the rationals in any archimedean field.

From a geometric point of view, the archimedean property, in its formulation as in (3.7), is equivalent to the geometric fact that, given two segments with positive lengths, it is always possible to cover one of the two by the union of a suitable finite number of subsequent copies of the other one. This absolutely intuitive property of the real line is at the basis of the method of exhaustion of Eudoxus and Archimedes. Indeed, as reported in the books of Elements by Euclid of Alexandria (325 B.C.-265 B.C.), the method of exhaustion relies on the sentence: “If from any quantity⁹³ we subtract not less than its half-part, and if from the remaining we subtract again not less than its half-part, and if we proceed in this way as many times as we want, then, at the end, we will remain with a quantity which is smaller than any other quantity could have been a-priori fixed”⁹⁴. Such a sentence comes from the archimedean property just by a “reductio ad absurdum”. In our modern language, this formulation of the method of exhaustion is of course equivalent to

$$0 < r < 1 \implies \lim_{n \rightarrow +\infty} r^n = 0,$$

where $1 - r$ is the fraction of the quantity to be subtracted. But, as already said, Greeks did not make limits.

The archimedean property is in some sense close to the completeness, because every complete field is archimedean and every archimedean field can be completed in a complete field. Indeed, the right use of the method of exhaustion should require the completeness, otherwise, said in a modern language, we do not know where we are converging to. For instance, in his studies about the area of the circle, Archimedes proved the fact that a circle is equivalent to the triangle with the circumference as basis and the radius as height, and in proving that he approximated the circle with inscribed and circumscribed regular polygons. For being rigorous, this procedure requires to suppose that any Cauchy

⁹²That is statements which are proved to be true.

⁹³Here, “quantity” means a geometric quantity as a length or a surface and so on.

⁹⁴Of course, the sentence continues to be true if we replace the half-part with any other fractional part: one third, one fourth, and so on. Moreover, as already said in the previous historical section, this has not to be understood as if Greeks made limits, but as the fact that repeating such a process a high number of times we can make the quantity smaller than an a-piori fixed quantity.

sequence of rational numbers⁹⁵ is converging in \mathbb{R} , and we know that this fact is in some sense equivalent to the completeness of the real numbers (see Proposition 2.11 and Remark 2.12).

The completeness of the real numbers was always assumed (tacitly or not, consciously or not) to be true as an axiom in the history of mathematics. Even if the concept of “real number” is rather modern, however the representation of the numbers on the line is rather natural, and the completeness of the lines of the geometry is rather obvious to be assumed. Such a principle of completeness can be explained in the following “nested interval principle” which is only another version of the Bolzano-Weierstrass Theorem 2.9 (or of the Bisection Lemma 2.10):

Nested interval principle. Let $\{a_n\}$ and $\{b_n\}$ be, respectively, an increasing and decreasing sequence of real numbers, such that $a_n < b_n$ for all n , and such that the difference $b_n - a_n$ can be made arbitrarily small, then there exists one and only one real number x such that $a_n < x < b_n$ for all n . Such an element x is of course the limit of both a_n and b_n .

This principle was taken as an axiom, that is as an unproven assumption, until the later XIX century. Also Cauchy and Bernard Bolzano (Prague, 1781 - 1848) used it without proof. As already said it is the nowadays known as Bolzano-Weierstrass theorem, and such a name acknowledges the first two mathematicians that recognized the need to state it explicitly. A direct consequence of the nested interval principle is the nowadays known as Bolzano Theorem or Intermediate Value Theorem: *a continuous function f on an interval $[a, b]$ attains all the values between $f(a)$ and $f(b)$.*

Until a model of real numbers (i.e. of complete field) is not given, the completeness of the real numbers should be part of the definition itself of the real numbers. In that case the only right thing to do is to define the real numbers as the (unique) field which contains the rational numbers and has the nested interval property. But this is how we argue nowadays, who know that we must choose an axiom and we can decide to choose “the real numbers has the nested interval property” as our axiom/starting point. This was not completely clear before the year 1872. Indeed, before of that year, the problem of defining the real numbers was obviously addressed, but in an ambiguous and non rigorous manner. For instance, Cauchy and Bolzano, and their contemporaries, went into a sort of “circular definition” of a real number: they substantially first defined the limit of a sequence as a real number, and then they defined a real number as the limit of a sequence of rational numbers. Indeed, Cauchy said that if a sequence is “internally convergent” (that is what we now call a Cauchy sequence), then it is also “externally convergent”, that is convergent to a real number ℓ . Hence, at this level, you must already know what a real number is, and so, if you use the convergence of a sequence for defining a real number, then you go into circularity.

⁹⁵For instance the rational sequence a_{2n} = the area of the inscribed polygon with n vertices, and a_{2n+1} = the area of the circumscribed polygon with n vertices, which converges to the area of the circle which may not be rational.

This lacking of rigor in defining the real numbers was first addressed by Charles Meray (France, 1835 - 1911) who, in the year 1872, published his studies about infinitesimal analysis and modified the definition of Cauchy and Bolzano. He indeed defined the convergence of a sequence just using the “Cauchy property” of the sequence, without referring to any number as its limit (as instead was done by Cauchy and Bolzano). Hence, he supposed that a convergent (in \mathbb{Q}) sequence of rational numbers defines a rational number and the non convergent (in \mathbb{Q}) Cauchy sequence defines a “fictitious number”, what we now call an irrational number. He also proved that such fictitious numbers can be ordered and well-behave with respect to the sum and product.

Weierstrass also tried to overcome the logic error made by Cauchy. Similarly to Meray, he realized the need of a definition of real number which was independent from the concept of limit. In some sense, going further Meray, he identified the real numbers with the Cauchy sequences of rational numbers, and so bypassing the problem of defining what the limit is. In our modern language, this is indeed another way than what we have done before with the half-sections, to complete the archimedean field of rational numbers. We can introduce the set F of all the Cauchy sequences of rational numbers. On F we introduce the following equivalence relation

$$\{a_n\} \sim \{b_n\} \iff \forall \varepsilon > 0 \exists n_\varepsilon \in \mathbb{N} \text{ such that } |a_n - b_n| \leq \varepsilon \forall n \geq n_\varepsilon,$$

and we consider the quotient set $\tilde{F} = F / \sim$. On \tilde{F} it is possible to introduce a sum, a product and an order in such a way that \tilde{F} is a complete ordered field. Moreover, \tilde{F} contains an isomorphic copy of \mathbb{Q} , the equivalence classes of sequences constantly equal to a rational number.

Weierstrass did not publish these studies, but they were published by some of his students/collaborators. Eduard Heine (1821-1881), in the year 1872, published an article where he put together the ideas of Weierstrass and some simplifications of recent studies by Georg Cantor (1845-1918). In particular, both Heine and Cantor were on the same streamline as Meray (and Weierstrass) in searching of a definition of real numbers using convergent sequences of rational numbers, but avoiding to refer to the concept of limit.

A completely different approach was instead published, in the same year 1872, by another German mathematician, Richard Dedekind (1831 - 1916), even if he was conscious of his ideas since 1858. The Dedekind’s method for “constructing” the real numbers is probably the most known nowadays⁹⁶, even if it is not very useful by an operative point of view. Dedekind realized that the property of “continuity” of the points of a line does not rely on a property of “density”⁹⁷, that is it is not equivalent to the sentence: “between two points there always exist a third point”, since the rational numbers are “dense” but do not form a “continuum”. Instead, Dedekind completely changed the point of view: the “continuity” of the line is not a property which can be formalized as the fact that it consists of a unique dense piece⁹⁸, but, on the contrary, as the fact that when you

⁹⁶It is substantially the method of half-sections we presented before.

⁹⁷For instance, Galileo (Pisa 1564-Arcetri 1642) and Leibniz were instead convinced of that.

⁹⁸In some sense “indivisible”.

cut the line you get exactly one point on the cut⁹⁹. For every section of the points of a line into two non empty classes, such that every point of the first class stays on the left of every point of the second one, there exists one and only one point which determines such a division¹⁰⁰. Dedekind however left this geometrical vision and transposed it in a completely analytical property of sections of rational numbers. A Dedekind section of the rational numbers is a couple of subset (A, B) as defined in Proposition 3.16. Inspired by the property of the line, Dedekind said that, for every section (A, B) of the rational numbers, there is one and only one real number which produces such a section (cut). Roughly speaking, we can say that there exists one and only one real number x such that

$$A \setminus \{x\} = \{q \in \mathbb{Q} \mid q < x\}, \quad B \setminus \{x\} = \{q \in \mathbb{Q} \mid q > x\}.$$

If A has a maximum or B has a minimum, then such a separating number is a rational number, otherwise it is an irrational number.

At the beginning of the XX century, Bertrand Russell (1872 - 1970)¹⁰¹ made a simple observation: if (A, B) is a Dedekind section of the real numbers, then B is uniquely determined by A and vice-versa. Hence it is sufficient to consider just the “half-section” A . And this is what we have done in the previous section¹⁰².

In the Example 3.26, we have described a non-archimedean field and we have shown that the element $1/x$ is smaller than any element of the form $1/n$. In the same way we can see that the element x is larger than any $n \in \mathbb{N}$.

If F is a non-archimedean field, it is natural to call *infinitesimal* all the elements a such that $0 < |a| < 1/n$ for all $n \in \mathbb{N} \setminus \{0\}$ and *infinite elements* all the elements a such that $|a| > n$ for all $n \in \mathbb{N}$. Of course, if a is infinitesimal then $1/a$ is an infinite and vice-versa. The existence of infinitesimal elements (as well as of infinite elements) is a characteristic property of the non-archimedean fields. Gottfried Wilhelm Leibniz (1646 - 1716) and Isaac Newton (1643-1727) were very involved in formalizing calculus rules for the use of infinitesimal quantities¹⁰³. By their rules for treating infinitesimal quantities, Newton and Leibniz discovered many results about what concerns limits, derivatives and

⁹⁹In some sense, as a property on how the line can be divided.

¹⁰⁰Note that the novelty of such a sentence is not rather on the uniqueness of the divisor point, but on its existence!

¹⁰¹Look to the coincidence: Bertrand Russell was born in the year 1872!

¹⁰²Note that the isomorphism in (3.12) just says that the section $(L, \mathbb{Q}_F \setminus L)$ and the section $(L', F \setminus L')$ define the same separating element (think to the case $\mathbb{Q}_F = \mathbb{Q}$, $F = \mathbb{R}$).

¹⁰³Actually, Leibniz and Newton may be considered the fathers of the infinitesimal analysis. Newton called his infinitesimal quantities “fluxions” and Leibniz “differentials”. They both introduced many of the notations and symbols we are now still using, for instance Newton : “ \dot{x} ” and “ x' ” for derivatives, and Leibniz “ dx ” for the infinitesimal increment of the abscissa x , “ $\int f dx$ ” for the integral. Leibniz also first used the term “function” for denoting exactly what we now call a function. We have to say that Newton and Leibniz, in the last years of their lives, were involved in a bitter dispute about the paternity of the “invention” of infinitesimal calculus: in particular Newton and his school accused Leibniz of plagiarism and an (English!) commission gave completely reason to Newton. Nowadays we can say that Newton certainly first arrived to his results and methods, but Leibniz was independent in his results and studies.

integration. Using the infinitesimals, all of such results were obtained just by some “algebraic calculations” and not by a limit procedure. A very simple example is the calculation of the derivative of the function $f(x) = x^2$. The derivative is the ratio between: the increment of the function when the abscissa is incremented by an infinitesimal dx and such infinitesimal increment dx itself. We then have

$$f'(x) = \frac{f(x + dx) - f(x)}{dx} = \frac{x^2 + 2xdx + (dx)^2 - x^2}{dx} = 2x + dx = 2x, \quad (3.14)$$

where the last equality is due to the fact that we neglect the infinitesimal quantity dx . This kind of procedure was strongly criticized by other mathematicians in the following years. Indeed it seems to be without logic foundations: these infinitesimal quantities are sometimes treated as non-zero elements (as in the second and in the third term of (3.14), where the denominator must be different from zero; and also in the third equality, where we divide by dx) and some other times as null quantities (as in the fourth equality in (3.14) where we substantially put $dx = 0$). The use of the infinitesimal quantities is certainly a powerful tool for “direct calculation” and for inferring the right result, but it is lacking of any rigorous foundation. This fact was not acceptable and the $\epsilon - \delta$ formulation of Weierstrass and Cauchy of the late XIX century will definitely rigorously define the limit, the derivative and the integration procedures without the use of any infinitesimal quantity: just using finite real numbers which can be taken as small as we want.

In the year 1966, 250 years after Leibniz’s death, Abraham Robinson (Germany/Poland 1918-USA 1974) published his famous book “Non-standard analysis”. Using deep mathematical logic tools, he was successful in constructing a suitable non-archimedean field containing the real numbers, denoted by \mathbb{R}^* and called the hyperreals, with the following properties: everything is true in the standard universe \mathbb{R} is also true in the non-standard universe \mathbb{R}^* (when, of course, concerning \mathbb{R} itself). Moreover the algebraic rules for the infinitesimals of \mathbb{R}^* are rigorous and well-behave as Leibniz has imagined. In particular, if $x \in \mathbb{R}$, then we can define the following relation

$$x' \in \mathbb{R}^*, x' \sim x \iff x - x' \text{ is infinitesimal in } \mathbb{R}^*,$$

and, for every finite¹⁰⁴ $y \in \mathbb{R}^*$ there exists a unique $x \in \mathbb{R}$ such that $y \sim x$ ¹⁰⁵ and it is denoted by $\text{st}(y)$: the standard part of y . Moreover, for every function $f : \mathbb{R} \rightarrow \mathbb{R}$ there is a unique extension $f^* : \mathbb{R}^* \rightarrow \mathbb{R}^*$. Hence, many of the usual definitions in the standard universe, such as continuity, uniform continuity and derivatives, become very easier and more intuitive in the non standard universe. For instance, if $f : \mathbb{R} \rightarrow \mathbb{R}$ is a function and $x_0 \in \mathbb{R}$ is a fixed real number:

¹⁰⁴Finite means that there exists $n \in \mathbb{N}$ such that $|y| \leq n$.

¹⁰⁵For $x \in \mathbb{R}$, the subset $\{y \in \mathbb{R}^* | y \sim x\} \subset \mathbb{R}^*$ is also called the monad of x , so using a term which was important in Leibniz’s philosophy.

$$\begin{aligned}
f \text{ is continuous at } x_0 \in \mathbb{R} &\iff \left(\mathbb{R}^* \ni x \sim x_0 \implies f^*(x) \sim f^*(x_0) \right), \\
f \text{ is uniformly continuous} &\iff f^*(x) \sim f^*(y) \quad \forall x, y \in \mathbb{R}^*, \quad x \sim y, \\
\text{if } f'(x_0) \text{ exists, then } f'(x_0) &= \text{st} \left(\frac{f^*(x_0 + y) - f^*(x_0)}{y} \right)
\end{aligned}$$

where y is any non-null infinitesimal.

Note in particular that, the non-standard formulation of the continuity in x_0 is very simple and it represents what every student would like to be true: if x is close to x_0 , then $f(x)$ is close to $f(x_0)$. Instead, the $\varepsilon - \delta$ formulation is rather more complicated: first you take an arbitrary criterium of vicinity for the images, ε , then you find a criterium of vicinity for the points, δ , and then you test again the criterium of vicinity for the images. Moreover, also the equivalent definition of derivative makes rigorous the calculation in (3.14)¹⁰⁶.

In this way, many statements and proofs of the standard analysis have an easier formulation in the non-standard analysis. However recall that whatever, about the standard analysis, you can prove using standard notations ($\varepsilon - \delta$ procedure) you can also prove using non-standard analysis and vice-versa. So, the non-standard analysis does not introduce more true facts (already proved or not) with respect to the ones in the standard analysis, but only provides a different way to look to them and, sometimes, a more intuitive way to prove them. Using non-standard analysis Robinson, together with Allen Bernstein, first proved as true a particular case of an open question about operators in Hilbert spaces. Such a proof was however immediately “translated” in the standard tools, even in a shorter way. However, the first non-standard proof was certainly illuminating.

¹⁰⁶Actually, it is also true that: if $\text{st}((f^*(x_0 + y) - f^*(x_0))/y)$ exists for all infinitesimal y and it is independent from y , then, f is derivable in x_0 and $f'(x_0) = \text{st}((f^*(x_0 + y) - f^*(x_0))/y)$.

4 Cardinality

In this chapter we are concerned with the power (or cardinality) of the sets. That is with the possibility of counting the elements of a set and of comparing two different sets by their quantity of elements. Such an argument is somehow linked to the theory of the topological spaces, which will be treated in the next Chapter.

4.1 The power of a set

Definition 4.1 *Let A and B two sets. We say that A and B are equivalent if there exists a bijective function $\varphi : A \rightarrow B$ and we write $A \sim B$. We say that A is finite if there exists $n \in \mathbb{N} \setminus \{0\}$ such that A is equivalent to $J_n = \{1, 2, \dots, n\}$ ¹⁰⁷. We say that the empty set \emptyset is finite. We say that A is infinite if it is not empty and not equivalent to J_n for all $n \in \mathbb{N} \setminus \{0\}$. We say that A is countable if it is infinite and equivalent to $\mathbb{N} \setminus \{0\}$. We say that A is uncountable if it is infinite but not countable. Finally, we say that A is at most countable if it is finite or countable.*

Remark 4.2 *It is easy to see that the relation $A \sim B$ between sets is an “equivalence relation”, i.e. reflexive, symmetric and transitive¹⁰⁸¹⁰⁹.*

Moreover, \mathbb{N} and $\mathbb{N} \setminus \{0\}$ are infinite and in particular countable¹¹⁰, and so a set is countable if and only if it is equivalent to \mathbb{N} (without a priori requiring to be infinite, which is just a consequence.).

If two sets are equivalent, then they in some sense have the same number of elements¹¹¹, however we say that they have the same power or the same cardinality.

It is evident the origin of the term “countable”: it takes account of the possibility of counting the elements of A . Also note that every infinite set A contains a countable subset. Indeed, since A is not empty, we can take an element $a_1 \in A$ and define an

¹⁰⁷This of course means that A has exactly n elements.

¹⁰⁸The quotation marks here stand for the fact that, for defining an equivalence relation, we need a set and we say that the equivalence relation is a suitable subset of the cartesian product. But, which is the set where we have defined the equivalence relation “ \sim ”? The natural answer would be “the set of all the sets”, but the existence of such a set cannot be accepted, otherwise we encounter various problems as the Russell Paradox. The only thing we can say is that, whenever we take a family of sets (i.e. a collection (a set) of sets which we put in such a collection), then, on such a collection, \sim is an equivalence relation.

¹⁰⁹On the Russell paradox. If we assume the existence of the set of all the sets, \mathcal{I} , then, since a property on the elements of a set must define a subset corresponding to the elements satisfying that property, we can consider the set $\hat{A} = \{A \in \mathcal{I} | A \notin A\}$. Hence, any answer to the question $\hat{A} \in \hat{A}$ leads to a contradiction (see also the proof of Theorem 4.14). Note that the contradiction comes from the fact that we have assumed the existence of the set \mathcal{I} . If instead we consider the set $\hat{B} = \{B \in \mathcal{P}(A) | B \notin B\}$ (that is the universe is now $\mathcal{P}(A)$), where A is a given set and $\mathcal{P}(A)$ is the set of all the subsets of A , then the question “ $\hat{B} \in \hat{B}$ yes or no?” does not leads to a contradiction. Indeed, as it is easy to see, $\hat{B} = \mathcal{P}(A)$ but $\mathcal{P}(A) \notin \mathcal{P}(A)$ and hence the answer $\hat{B} \notin \hat{B}$ does not leads to the contradiction $\hat{B} \in \hat{B}$.

¹¹⁰Just use the successor function of Peano.

¹¹¹This is rigorous only for finite set, for infinite set the concept of “having the same number of elements” needs a clarification.

injective function $\varphi_1 : J_1 \rightarrow A$, $1 \mapsto a_1$. Of course, such a function is not surjective, otherwise A would be equivalent to J_1 . Hence there exists an element $a_2 \in A$ such that $a_1 \neq a_2$. Hence, we have an injective function $\varphi_2 : J_2 \rightarrow A$, $i \mapsto a_i$ for $i = 1, 2$. Also φ_2 is not surjective, and so there exists $a_3 \in A$ such that $a_3 \neq a_1, a_2$. Proceeding in this way, by induction, for every $n > 1$ we get an injective function $\varphi_n : J_n \rightarrow A$ which is not surjective and such that, when restricted to J_{n-1} it coincides with φ_{n-1} . If we define $C = \bigcup_{n \in \mathbb{N} \setminus \{0\}} \varphi_n(J_n) \subseteq A$, then we have the bijection

$$\varphi : \mathbb{N} \setminus \{0\} \rightarrow C \quad n \mapsto \varphi_n(n) = a_n.$$

In some sense we can say that the countable sets are the “smallest” infinite sets. From the fact that (see Remark 4.2) $\mathbb{N} \setminus \{0\}$ is a proper subset of \mathbb{N} but nevertheless equivalent to \mathbb{N} , then we can also easily deduce that a set A is infinite if and only if it is equivalent to a proper own part¹¹².

Proposition 4.3 *i) The set of integers \mathbb{Z} is countable; ii) every infinite subset B of a countable set A is countable; iii) if A has an infinite subset, then A is infinite too, iv) if A is a countable index set (i.e. a countable family of indices) and, for every $a \in A$, B_a is a countable set, then $B = \bigcup_{a \in A} B_a$ is countable, v) if $\{A_i\}_{i=1, \dots, n}$ is a finite family of countable sets, then $A_1 \times A_2 \times \dots \times A_n$ is countable; vi) \mathbb{Q} is countable.*

Proof. i)

$$\varphi : \mathbb{N} \rightarrow \mathbb{Z} \quad n \mapsto \begin{cases} \frac{n}{2} & \text{if } n \text{ is even,} \\ -\frac{n+1}{2} & \text{if } n \text{ is odd.} \end{cases}$$

ii) Let $\varphi : \mathbb{N} \rightarrow A$ be a bijection. By induction, we define the following other bijection

$$\tilde{\varphi} : \mathbb{N} \rightarrow B, \quad \begin{cases} 0 \mapsto \varphi(n_0) & n_0 = \min\{n \in \mathbb{N} \mid \varphi(n) \in B\} \\ m \mapsto \varphi(n_m) & n_m = \min\{n > n_{m-1} \mid \varphi(n) \in B\} \quad m \geq 1 \end{cases}$$

iii) Obvious.

iv) Since B_a is countable, then its elements form a sequence which can be enumerated: $\{b_{a1}, b_{a2}, b_{a3}, \dots\} = B_a$. Moreover let $\varphi : \mathbb{N} \setminus \{0\} \rightarrow A$ be a bijection. We form an infinite table with a countable quantity of rows (one per very element $a \in A$), and such that every n -th row consists of the sequence of the elements of $B_{\varphi(n)}$. Then we can count the elements of such a table with the following (diagonal) bijection¹¹³:

¹¹²If A is infinite, then it contains a countable set B , which is then equivalent to both \mathbb{N} and $\mathbb{N} \setminus \{0\}$. From this we can deduce that, for every $x_0 \in B$, $B \sim B \setminus \{x_0\}$ and also that $A \sim A \setminus \{x_0\}$. On the contrary, if A is finite, it is easy to see that it cannot be equivalent to any proper subset because $J_n \not\sim J_m$ if $n \neq m$.

Of course, it is not necessary to remove only a finite number of elements, but it is also possible to remove a suitable infinite quantity of elements: think to the set of even natural numbers, which differ from \mathbb{N} by an infinitely many quantity of removed elements, and nevertheless it is equivalent to \mathbb{N} itself.

¹¹³We are counting along bottom-left to top-right diagonals, starting from top-left of the table. The reader is invited to draw a picture of the procedure.

$$\tilde{\varphi}(1) = b_{\varphi(1)1}, \tilde{\varphi}(2) = b_{\varphi(2)1}, \tilde{\varphi}(3) = b_{\varphi(1)2}, \tilde{\varphi}(4) = b_{\varphi(3)1}, \tilde{\varphi}(5) = b_{\varphi(2)2}, \tilde{\varphi}(6) = b_{\varphi(1)3},$$

and, by induction, whenever we have counted the term $\tilde{\varphi}(k) = b_{\varphi(1)m}$ for some k, m , then the following $m + 1$ terms are of the form $\tilde{\varphi}(k + j) = b_{\varphi(m+2-j)j}$ for all $j = 1, \dots, m + 1$, and moreover, it will be $\tilde{\varphi}(k + m + 2) = b_{\varphi(1)(m+2)}$. Hence the position-elements of the table form a countable set. Since some of such elements may be repeated more than one time (if it belongs to more than one B_a) then $\bigcup_{a \in A} B_a$ is an infinite set which is equivalent to a subset of position-element of the table, and hence it is countable by point ii).

v) We only prove the sentence for $n = 2$. For every $a^1 \in A_1$, let us define the set $A_2^{a^1} = \{(a_1, a) | a \in A_2\} \subseteq A_1 \times A_2$, and note that $A_1 \times A_2 = \bigcup_{a^1 \in A_1} A_2^{a^1}$. Hence, the sentence immediate follows from point iv).

vi) \mathbb{Q} is an infinite subset of $\mathbb{N} \times \mathbb{N}$ and so we conclude by points v) and ii), \square

Remark 4.4 *Of course, in Proposition 4.3, in points iv) and v) we can consider collections B_a and A_i respectively of at most countable sets with at least one countable, and the conclusions remain the same. Also, if one of B_a is countable, then the index set A may be finite.*

Proposition 4.5 *If $\{a_n\}_{n \geq 1}$ is a sequence of natural numbers belonging to $\{0, 1, \dots, 9\}$, then the series*

$$\frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3} + \dots + \frac{a_k}{10^k} + \dots$$

is convergent¹¹⁴ to a real number $x \in]0, 1[$. We write

$$x = 0.a_1a_2a_3a_4 \dots$$

and we call it the decimal expansion of x .

Every real number $x \in]0, 1[$ has a decimal expansion and such expansion is unique unless x is a rational number of the form $m/10^n$ for some $m, n \in \mathbb{N}$ in which case it has precisely two decimal expansions.

Proof. Take a_1 as the largest natural number in $\{0, 1, \dots, 9\}$ such that

$$0 \leq x - \frac{a_1}{10},$$

and note that we have¹¹⁵ $x - a_1/10 < 1/10$. By induction, take a_n as the largest natural number in $\{0, 1, \dots, 9\}$ such that

¹¹⁴By comparison with a geometric sequence.

¹¹⁵If $a_1 = 9$, then $x - 9/10 \geq 1/10$ would imply $x \geq 1$ which is a contradiction. If $a_1 \leq 8$, then $x - a_1/10 \geq 1/10$ would imply $x \geq (a_1 + 1)/10$ which is a contradiction to the maximality of a_1 , because $a_1 + 1 \in \{1, \dots, 9\}$.

$$0 \leq x - \sum_{k=1}^n \frac{a_k}{10^k},$$

and note again that $x - \sum_{k=1}^n (a_k/10^k) < 1/10^n$. It is easy to see that, in this way, we have constructed a decimal expansion of x .

If $x \neq m/10^k$, if $\{a_n\}, \{b_n\}$ are two decimal expansions of x and if we suppose that $a_1 < b_1$ then we have¹¹⁶

$$\frac{a_1}{10} < \frac{b_1}{10} < x,$$

and hence

$$\frac{1}{10} < x - \frac{a_1}{10} = \sum_{n=1}^{\infty} \frac{a_n}{10^n} - \frac{a_1}{10} = \sum_{n=2}^{\infty} \frac{a_n}{10^n} \leq 9 \left(\sum_{n=2}^{\infty} \frac{1}{10^n} \right) \leq \frac{1}{10},$$

which is a contradiction. Hence $a_1 = b_1$, and proceeding in this way, $a_n = b_n$ for all $n \geq 1$ ¹¹⁷. If instead $x = m/10^k$ for some m, k , then, until $n = k - 1$, with the same argument as before we must have $a_n = b_n$. Now, let us suppose $a_k < b_k$; with the same argument, it must be $a_k = b_k - 1$ ¹¹⁸. Hence we have

$$\sum_{n=1}^k \frac{b_n}{10^n} - \sum_{n=1}^k \frac{a_n}{10^n} = \frac{1}{10^k},$$

and then the only possibility is $a_n = 9, b_n = 0$ for all $n > k$, which exactly forms two decimal expansions.. \square

Remark 4.6 *Of course, a similar conclusion of Proposition 4.5 holds for very $x \in \mathbb{R}$, and, for $x > 0$ the expansion is of the form $a_0.a_1a_2a_3 \dots$ where a_0 is the largest natural number smaller than x and $0.a_1a_2a_3 \dots$ is the decimal expansion of $x - a_0$. Note that we get two expansions for $x = 1$: $0.999 \dots$ and $1.000 \dots$*

Theorem 4.7 *The set of real numbers is uncountable.*

Proof. We prove that the set S of points $x \in]0, 1[$ which have a unique decimal expansion is uncountable. From this we then get that $]0, 1[$ is uncountable (otherwise it

¹¹⁶Of course, in this case, we certainly have $x \neq a_1/10, b_1/10$. We also have $b_1 \geq 1$ because $a_1 \geq 0$. Also note that, being the series made by positive terms, every partial summation is not larger than the final sum. Moreover, any partial summation is of the form $m/10^k$.

¹¹⁷Roughly speaking, if for some k we have $a_k \neq b_k$, the rest of the series cannot fill the gap (having both series the same sum x).

¹¹⁸Otherwise, if $a_k \leq b_k - 2$, the rest of the series with a_i cannot fill the gap and reach the same sum x of the series with b_i .

can have only at most countable subsets, by point ii) of Proposition 4.3), and similarly \mathbb{R} is uncountable.

First of all, note that S is infinite. Indeed, for example, it contains the countable set $\{\pi/(10)^n | n \in \mathbb{N}\}$. Now, we prove that any function $f : \mathbb{N} \rightarrow S$ cannot be surjective, from which the conclusion follows¹¹⁹. For every n , let us denote $x_n = f(n) \in S \subseteq]0, 1[$ and write the decimal expansion of x_n (which is unique) as

$$x_n = 0.a_{n1}a_{n2}a_{n3} \cdots$$

For every $n \in \mathbb{N} \setminus \{0\}$, we define

$$b_n = \begin{cases} 8 & \text{if } a_{nn} \neq 8, \\ 1 & \text{if } a_{nn} = 8, \end{cases}$$

and consider the decimal expansion

$$y = 0.b_1b_2b_3 \cdots \in S.$$

Obviously $y \in S$, since its decimal expansion does not contain either 0 or 9, and so it is unique. We say that $f(n) \neq y$ for all $n \in \mathbb{N}$, which will conclude the proof. This is true since, for every n , the decimal expansion of y differs from the one of x_n in its n -th place, and x_n has a unique expansion. \square

In Remark 4.2 we have said that two equivalent sets have the same power. But what is the power of a set? It would be a common property of all the sets belonging to the same “equivalent class”, but as explained, the use of the quotation marks means that we have to be very careful when speaking about classes of equivalence of sets. Anyway we can certainly say when a set has a power strictly greater than the power of another set.

Definition 4.8 *Given two sets A and B we say that the power of A is strictly less than the power of B , and we write $m(A) < m(B)$, if A is equivalent to a subset of B and no subsets of A are equivalent to B ¹²⁰.*

One can ask whether, giving two arbitrary sets A and B , it is always true that A and B are equivalent, and hence have the same power ($m(A) = m(B)$), or that one of the two set has strictly smaller power than the other one. In other words the question is whether the powers of sets can always be compared. We can positively answer to such a question since the following two results are true. In particular the first result can be proved¹²¹, the second one must be taken as an axiom, since it comes from the Axiom of Choice¹²².

¹¹⁹ S is infinite and hence at least countable, but, being no such a surjective function, it is certainly more than countable.

¹²⁰It is easy to see that this is equivalent to the existence of a injection from A to B and no existence of a surjection.

¹²¹But we do not do that.

¹²²Axiom of Choice: given any set M , there exists a “choice function” f from the parts of M , $\mathcal{P}(M)$, to M such that for every non-empty set $A \subseteq M$, $f(A) \in A$.

Actually, without saying it, we have already used the axiom of choice several times. For instance in the proof of the fact that every infinite set A contains a countable set: in the possibility of choosing a point $a \in A$ such that $a \notin \{a_1, a_2, \dots, a_{n-1}\}$ for every step n . Also the Bolzano-Weierstrass theorem requires the use of the axiom of choice.

Theorem 4.9 *Let A and B two well-ordered sets. Then it is certainly true that $m(A) = m(B)$ or $m(A) < m(B)$ or $m(A) > m(B)$. In other words, the powers of two well-ordered sets are always comparable.*

Theorem 4.10 *(Well-ordering principle) Every set can be well-ordered. That is, for every set, there exists an order relation on it which makes it a well-ordered set.*

We now prove the following result, which is important for comparing sets, and also shows that the case stated in Definition 4.8 is exhaustive.

Theorem 4.11 *(Cantor-Bernstein theorem) Given any two sets A and B , suppose that A contains a subset A_1 equivalent to B and that B contains a subset B_1 equivalent to A . Then A and B are equivalent.*

To prove this theorem we first need the following Proposition

Proposition 4.12 *Let $\{A_n\}_n$ and $\{B_n\}_n$ be countable collections of pairwise disjoint sets respectively ($A_n \cap A_m = \emptyset$, $B_n \cap B_m = \emptyset$ for all $n \neq m$). Moreover, let us suppose that $A_n \sim B_n$ for all n . Hence*

$$\bigcup_n A_n \sim \bigcup_n B_n.$$

Proof. For every n we have the bijection $\varphi_n : A_n \rightarrow B_n$. By the pairwise disjointness, the following function is well-defined and a bijection

$$\varphi : \bigcup_n A_n \rightarrow \bigcup_n B_n, \quad a \mapsto \varphi_n(a), \quad \text{where } n \text{ is such that } a \in A_n.$$

□

Proof of Theorem 4.11. Let $\varphi_A : A \rightarrow B_1$ and $\varphi_B : B \rightarrow A_1$ be bijections. We define

$$A_2 = \varphi_B(\varphi_A(A)) = \varphi_B(B_1) \subseteq A_1 \subseteq A.$$

Since compositions of bijections are bijections, we have that A_2 is a subset of A_1 equivalent to A . We then define $A_3 = \varphi_B(\varphi_A(A_1)) \subseteq \varphi_B(\varphi_A(A)) = A_2$ which is a subset of A_2 equivalent to A_1 . By induction (defining $A_0 = A$), we have a sequence of nested subsets of A

$$A_{k+2} = \varphi_B(\varphi_A(A_k)) \subseteq A_{k+1} \quad \forall k \in \mathbb{N} \setminus \{0\},$$

such that A_{k+2} is equivalent to A_k and that $A_k \setminus A_{k+1}$ is equivalent to $A_{k+2} \setminus A_{k+3}$ for every k ¹²³. Now, we define

¹²³Because the latter is exactly equal to $\varphi_B(\varphi_A(A_k \setminus A_{k+1})) = \varphi_B(\varphi_A(A_k)) \setminus \varphi_B(\varphi_A(A_{k+1}))$.

$$D = \bigcap_k A_k$$

and write

$$A = \left[(A \setminus A_1) \cup (A_2 \setminus A_3) \cup (A_4 \setminus A_5) \cup \dots \right] \cup \\ A_1 = \left[\begin{array}{l} (A_1 \setminus A_2) \cup (A_3 \setminus A_4) \cup (A_5 \setminus A_6) \cup \dots \\ (A_2 \setminus A_3) \cup (A_4 \setminus A_5) \cup (A_6 \setminus A_7) \cup \dots \\ (A_1 \setminus A_2) \cup (A_3 \setminus A_4) \cup (A_5 \setminus A_6) \cup \dots \end{array} \right] \cup D$$

By Proposition 4.12 we then conclude that $A \sim A_1$ which of course implies $A \sim B$. \square

Remark 4.13 *By the Cantor-Bernstein theorem, we deduce that, whenever there exists an injection between two sets $\varphi : A \rightarrow B$, then $m(A) \leq m(B)$. Indeed, being φ injective, A is equivalent to a subset of B , $\varphi(A) \subseteq B$. Hence, if A contains a subset which is equivalent to B then $m(A) = m(B)$ by the Cantor-Bernstein theorem, otherwise $m(A) < m(B)$ by the definition.*

The fact that $\varphi : A \rightarrow B$ injection implies $m(A) \leq m(B)$ is obvious in the finite case. By Proposition 4.3, this is also true in the countable case. However, in the infinite uncountable case, it is not a-priori obvious. We need the Cantor-Bernstein theorem.

Now, the question is: is there a set with power bigger than all the others? The answer is “no, there is not”, as deduced from the following result.

Theorem 4.14 *Let A be a non-empty set. Then, the set of the parts of A*

$$\mathcal{P}(A) = \left\{ B \mid B \subseteq A \right\},$$

has power strictly greater than the power of A : $m(A) < m(\mathcal{P}(A))$.

Proof. Obviously the power of $\mathcal{P}(A)$ cannot be strictly smaller than the power of A , since there is the injection

$$\psi : A \rightarrow \mathcal{P}(A), \quad a \mapsto \{a\}.$$

By contradiction, let us suppose that there exists a bijection $\varphi : A \rightarrow \mathcal{P}(A)$, and let us define the following subset of A

$$X = \left\{ a \in A \mid a \notin \varphi(a) \right\} \in \mathcal{P}(A).$$

Possibly empty, but X , as subset of A , exists. Hence, there should exist $x \in A$ such that $x = \varphi^{-1}(X)$. Now, ask whether $x \in X$ or not. Both answers (yes or no) give a contradiction:

$$\begin{aligned}x \in X &\implies x \notin \varphi(x) = X, \\x \notin X &\implies x \in \varphi(x) = X.\end{aligned}$$

Hence, the bijection φ cannot exist and so $m(A) < m(\mathcal{P}(A))$. \square

Now, we conclude this section reporting some useful results and some comparison between important sets.

i) If M is uncountable and A is countable (or finite), then $M \cup A \sim M \sim M \setminus A$. For the second equivalence, observe that if $\tilde{A} = \{a_0, a_1, a_2, a_3, \dots\} \subseteq M$ is countable and $\tilde{A}_0 = \{a_i \in \tilde{A} | i \text{ is even}\}$, $\tilde{A}_1 = \{a_i \in \tilde{A} | i \text{ is odd}\}$ we then have $\tilde{A} \sim \tilde{A}_0 \sim \tilde{A}_1$. Hence, writing $M = \tilde{A} \cup (M \setminus \tilde{A})$, $M \setminus \tilde{A}_1 = \tilde{A}_0 \cup (M \setminus \tilde{A})$ we get $M \sim M \setminus \tilde{A}_1$. Conclude by showing that any countable $A \subseteq M$ can be seen as the set of the “odd” elements of a suitable countable set $\tilde{A} \subseteq M$.¹²⁴

ii) For any $-\infty \leq a < b \leq +\infty$ (extended) real numbers, $]a, b[$ is equivalent to \mathbb{R} . Just take a strictly increasing continuous function f on $]a, b[$ such that $\lim_{x \rightarrow a^+}(x) = -\infty$ and $\lim_{x \rightarrow b^-}(x) = +\infty$.

iii) For any $-\infty \leq a < b \leq +\infty$ (extended) real numbers, the closed or semi-open interval with a and b as extremes is equivalent to \mathbb{R} . For example, apply point i) to $[a, b] =]a, b[\cup \{a, b\}$.

iv) The countable (or finite) union of sets equivalent to \mathbb{R} is still equivalent to \mathbb{R} . Let $A_n \sim \mathbb{R}$ and suppose that they are pairwise disjoint (which is not restrictive). By point ii), $A_n \sim]n, n+1[$. Let $\varphi_n : A_n \rightarrow]n, n+1[$ be a bijection. Hence $\varphi : \bigcup_n A_n \rightarrow [0, +\infty[\setminus \mathbb{N}$, $a \mapsto \varphi_n(a)$, where n is such that $a \in A_n$, is a bijection.

v) The set of *algebraic*¹²⁵ real number is countable. The set of polynomials with rational coefficients is countable...

vi) The set of *transcendental*¹²⁶ real numbers is equivalent to \mathbb{R} . Use points i) and v).

vii) The set of real numbers in $]0, 1[$ with exactly one decimal expansion has the power of \mathbb{R} . Note that the set of numbers with two decimal expansions is countable (they are rationals). Conclude by points i) and ii).

ix) $m(\mathcal{P}(\mathbb{N})) = m(\mathbb{R})$. In the same way as in Proposition 4.5, we can prove that for every natural number $b \in \mathbb{N} \setminus \{0, 1\}$, every points $x \in]0, 1[$ admits an expansion

$$x = 0.\beta_1\beta_2\beta_3 \dots$$

where $\beta_i \in \{0, 1, 2, \dots, b-1\}$ and $x = \sum_{n=1}^{\infty} \beta_n / (b^n)$. Moreover, also in this case, the expansion is unique unless x is of the form m/b^k for some natural numbers m, k in which

¹²⁴Note that being M uncountable and A countable, then $M \setminus A$ is still infinite (otherwise M would be countable) and hence it contains another countable set.

¹²⁵A real number is “algebraic” if it is the root of a polynomial with rational coefficients.

¹²⁶A real number is “transcendental” if it is not algebraic.

case there are exactly two expansions. Hence, also in this case the set of points with exactly one expansion has the power of \mathbb{R} . Consider the binary expansions, that is take $b = 2$, and prove that \mathbb{R} is equivalent to the set of all sequences with value in $\{0, 1\}$. Conclude by noting that the set of the sequences with value in $\{0, 1\}$ is equivalent to the set of the parts of \mathbb{N} (consider the characteristic functions $\chi_A(n) = 1$ if $n \in A$, $\chi_A(n) = 0$ if $n \notin A$, for all $A \subseteq \mathbb{N}$).

x) The set of all sequences of rational numbers is equivalent to \mathbb{R} .

If $s : \mathbb{N} \rightarrow \mathbb{Q}$ is a sequence, define $\varphi(s) = \{(n, s(n)) | n \in \mathbb{N}\} \subseteq \mathcal{P}(\mathbb{N} \times \mathbb{Q})$, which is an injection. From this, by point ix) and the countability of $\mathbb{N} \times \mathbb{Q}$, we have

$$m(\mathbb{R}) = m(\{\text{sequences in } \{0, 1\}\}) \leq m(\{\text{sequences in } \mathbb{Q}\}) \leq m(\mathcal{P}(\mathbb{N} \times \mathbb{Q})) = m(\mathbb{R}).$$

xi) The set of all sequences of real numbers is equivalent to \mathbb{R} .

By considering the binary expansions, we have the equivalence of the set of sequences of real numbers with the set X of the sequences of sequences taking values in $\{0, 1\}$. If $\{s_k\}_k$ is an element of such a set (that is for every $k \in \mathbb{N}$, $s_k : \mathbb{N} \rightarrow \mathbb{R}$ is a sequence) we define $\varphi(\{s_k\}_k) = \{(k, n, s_k(n)) | k, n \in \mathbb{N}\} \subseteq \mathcal{P}(\mathbb{N} \times \mathbb{N} \times \mathbb{N})$, which is an injection. We then have

$$m(\mathbb{R}) = m(\{\text{sequences in } \{0, 1\}\}) \leq m(X) \leq m(\mathcal{P}(\mathbb{N} \times \mathbb{N} \times \mathbb{N})) = m(\mathbb{R}).$$

xii) For every $n \in \mathbb{N} \setminus \{0\}$, \mathbb{R}^n is equivalent to \mathbb{R} .

\mathbb{R}^n is equivalent to $]0, 1[^n$ which is equivalent to the set of n -uple of sequences of 0 and 1 not always null (i.e. every such a sequence has a non-null term, otherwise it would represent $x = 0$ which is not a point of $]0, 1[$), which is equivalent to the set¹²⁷ $(\mathcal{P}(\mathbb{N}) \setminus \emptyset)^n$ which is equivalent to $\mathcal{P}((\mathbb{N} \setminus \emptyset)^n)$ which is equivalent to \mathbb{R} .

xiii) The set of all continuous functions on $[0, 1]$ is equivalent to \mathbb{R} .

By continuity and the density of \mathbb{Q} in $[0, 1]$, a continuous function on $[0, 1]$ is completely determined by its real values on $\mathbb{Q} \cap [0, 1]$, which is a countable set. Hence there is an injection between the set of continuous functions and the set of sequences of real numbers. Hence, by point xi), the power of the continuous functions is not more than the power of \mathbb{R} . On the other hand, \mathbb{R} is injected into the set of continuous functions (as constant functions) and so the the powers coincide.

xiv) The set of all functions on $[0, 1]$ (continuous or discontinuous) has power strictly greater than the power of \mathbb{R} .

It contains the characteristic functions... Apply Theorem 4.14.

4.2 Historical notes and comments

Georg Cantor was the first to prove that the real numbers has a power larger than the power of the rational numbers. He also firstly proved that there are “infinitely many powers”, that is that there is not an upper bound on the possible power of a set. All such powers were called by Cantor “transfinite numbers” and he also constructed a real arithmetics and an order on such numbers.

¹²⁷The null sequence represents, via characteristic function, the empty set.

All these studies by Cantor (also Dedekind was a pioneer on the subject), even if opposed by many of his contemporaries, certainly led, in the following decades, to the beginning of the set theory that is the branch of mathematics which is interested on sets just as collections of objects, and on their possible relations. Such a branch of mathematics, or such a way to look to sets and their elements, has become in the XX century the main tool for research in mathematics, having led to the creation of the theory of many abstract structures such as groups, rings, fields, vector space, metric spaces, topological spaces and so on.

Concerning the power of sets, there is a natural question that may arise when studying such a subject. The question is “Is there a set whose power is strictly greater than the power of the natural numbers¹²⁸, and strictly less than the power of the real numbers?”. Also Cantor addressed such a question and he was not able to answer. Anyway he conjectured that the answer is “No, such a set does not exist”. Since the power of the real numbers is also called the power of the continuum, such a conjecture is known as the continuum hypothesis: “There is not a set A whose power satisfies $m(\mathbb{N}) < m(A) < m(\mathbb{R})$ ”.

In the year 1900, at the congress of mathematicians in Paris, the important German mathematician David Hilbert (Königsberg 1862 - Göttingen 1943) gave a list of twenty-three open problems on which, in his authoritative opinion, mathematicians of XX century should had to be interested. The first one of such problems was the validity or falsity of the continuum hypothesis. Despite to this fact, no mathematician was able to prove the validity (or the falsity) of the continuum hypothesis. In the year 1940, Kurt Gödel (1906-1978) showed that the affirmative answer to the continuum hypothesis (i.e. there is not the set) is consistent with the axiomatization of the set theory¹²⁹, that is no contradiction may arise if we accept the continuum hypothesis as true. However, in the year 1963, Paul Cohen (1934-2007) proved that also the negative answer to the continuum hypothesis (i.e. the set exists) is consistent with the set theory, that is no contradiction may arise if we reject the continuum hypothesis. So the continuum hypothesis is independent from the axioms of the set theory¹³⁰

The question whether a set A such that $m(\mathbb{N}) < m(A) < m(\mathbb{R})$ exists or not has then become a question whether the continuum hypothesis must be added as an axiom or not¹³¹. Both behaviors would lead to several different, but interesting, results.

¹²⁸Which is, as we know, the minimum infinite power.

¹²⁹The so called ZFC set theory.

¹³⁰As the Euclid’s fifth postulate on the parallel lines is independent from the first four ones.

¹³¹Another possible question may be find some other “natural axioms” which imply the validity (or the falsity) of the continuum hypothesis.

5 Topologies

In this chapter, we analyze the other component of the broken core in Figure 2, the one concerning the concept of limit and its consequences as well as generalizations, in one word the concept of topology. We start revisiting the concept of limit in \mathbb{R} .

5.1 On the topology in \mathbb{R}

Look to the definition of limit of a sequence of real numbers Definition 2.1, and consider the case $\ell \in \mathbb{R}$. We can ingenuously say that ℓ is the limit of the sequence a_n if “taking n larger and larger, then a_n becomes closer and closer to ℓ ”. Where is the ingenuity? It is of course in the meaning of the expressions “larger and larger” and “closer and closer”. For instance, when n becomes larger and larger, the sequence $a_n = 1/n$ becomes closer and closer to -1 . But -1 is not its limit! The problem here is that a_n does become closer and closer to -1 , but not so close as we want: it always remains from -1 at a distance more than 1. We can then modify our ingenuous sentence as “taking n larger and larger, we can make a_n be so close to ℓ as we want”. But also this one is not correct. Indeed, take the sequence $a_n = 0$ if n is even, $a_n = 1 - 1/n$ if n is odd. For this sequence it is true that taking n larger and larger (the larger and larger odd natural numbers), then a_n becomes close to 1 as we want. But 1 is not its limit! The problem here is that a_n becomes so close to 1 as we want not for all sufficiently large natural numbers, but only for all sufficiently large odd natural numbers. Hence, we finally modify our sentence as

“taking n sufficiently large we can definitely make a_n be as close to ℓ as we want”.
(5.1)

Are we satisfied with such a sentence? Certainly yes, if we know the meaning of “large” and “close”. Let us suppose that we already know what “large” means for a natural number, and concentrate on the meaning of “closeness” in \mathbb{R} . By our exterior experience, for giving a meaning to the word “closeness” we must have a concept of distance. And, by our identification of \mathbb{R} with the line, we have a natural concept of distance from two elements x, y of \mathbb{R} : it is the length of the segment which links the two points, in other words the distance is the non-negative real number $|x - y|$. Hence we say: *the distance between two real numbers is the absolute value of their difference*. So, we can view the distance in \mathbb{R} as a function from $\mathbb{R} \times \mathbb{R}$ to \mathbb{R} which maps the couple (x, y) to the real number $|x - y|$. Hence, making a_n definitely as close to ℓ as we want means making $|a_n - \ell|$ as small as we want for all $n \geq \bar{n}$, with a suitable \bar{n} , in other words it means that:

$$\forall \varepsilon > 0 \exists \bar{n} \in \mathbb{N} \text{ such that } |a_n - \ell| \leq \varepsilon \forall n \geq \bar{n}. \quad (5.2)$$

Note that in (5.2) we take $\varepsilon > 0$, that is we never require that $a_n = \ell$, which means to permit $\varepsilon = 0$, since this would be too strong. We only require that a_n stays “around” to ℓ , even if want to make such “aroundness” as small as we want (but never collapsing to

the single ℓ). A natural concept of points which stay around to ℓ is of course the concept of interval centered in ℓ . Hence we can rewrite (5.1) (or equivalently (5.2)) as

$$\forall \text{ interval } I \text{ centered on } \ell \exists \bar{n} \in \mathbb{N} \text{ such that } a_n \in I \forall n \geq \bar{n}.$$

But then some natural questions are: can we take another function $\varphi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ as a “distance” in \mathbb{R} ? Which properties has φ to satisfy in order to be considered as a distance in \mathbb{R} ? Can we define a distance in any other different set than \mathbb{R} ? To answer to such questions we have to go back to our exterior experience: the distance is the length of the segment which links the two points. With this flashback in mind we can certainly say what is the distance in \mathbb{R}^2 (the plane) and in \mathbb{R}^3 (the 3D physical space): indeed in such sets we know what is a segment linking two points and what is its length. Of course we can extend such a concept to \mathbb{R}^n for every n . But these facts are not enough for answering to the second question, which is the most important. Looking more precise to the concept of “segment” in the plane and in the 3D-space, and also thinking to our exterior experience, we can say that it is the shorter path linking the two points. Hence the distance between two elements, if coherent with our experience, should represent the minimum length among the lengths of all paths linking the two points¹³². Hence, thinking to distance as “minimum length”, it is natural to require that i) a distance is non-negative valued, ii) every path has strictly positive length, iii) the distance between x and y is the same of the distance between y and x ¹³³, iv) the distance between x and y is not larger than the sum of the distances between x and z and z and y respectively¹³⁴.

The following subsection will analyze the concept of distance, whereas a subsequent one will be concerned with the concept of “aroundness”.

5.2 Metric spaces

Let X be a non empty set. Looking to the points i)–iv) at the end of the previous subsection, we can consider, as a distance on X , a function $d : X \times X \rightarrow [0, +\infty[$ such that

$$\begin{aligned} a) \text{ (positive definition)} \quad & \forall x, y \in X, d(x, y) = 0 \iff x = y, \\ b) \text{ (symmetry)} \quad & d(x, y) = d(y, x) \quad \forall x, y \in X, \\ c) \text{ (triangular inequality)} \quad & d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z \in X. \end{aligned} \tag{5.3}$$

Point a) together with the choice of co-domain of d (i.e. $[0, +\infty[$) takes account of points i) and ii), point b) takes account of point iii) and point c) takes account of point iv).

¹³²This of course would imply that we know what a “path” inside a set and its length are, but we do not go further in such a rigorous argumentation.

¹³³All the paths linking x to y also link y to x and vice-versa.

¹³⁴This comes from the minimality: if it was the contrary, then instead of directly going from x to y it would be more convenient, in between, to pass through z .

Definition 5.1 A non empty set X is said a metric space if there exists a function $d : X \times X \rightarrow [0, +\infty[$ satisfying (5.3). The function d is said a distance or a metrics on X ¹³⁵.

Example 5.2 Here are some examples of metrics spaces.

- 1) \mathbb{R} with $d(x, y) = |x - y|$.
- 2) $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ with the distance of \mathbb{R} .
- 3) \mathbb{R} with the distance $d(x, y) = |\arctan x - \arctan y|$.
- 4) \mathbb{R}^n with $d(x, y) = \|x - y\|_{\mathbb{R}^n} = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$.
- 5) \mathbb{C} with the distance of \mathbb{R}^2 .
- 6) \mathbb{R}^2 with $d(x, y) = \max\{|x_1 - y_1|, |x_2 - y_2|\}$.
- 7) \mathbb{R}^2 with $d(x, y) = |x_1 - y_1| + |x_2 - y_2|$.
- 8) $C^0([a, b]; \mathbb{R}) = \left\{ f : [a, b] \rightarrow \mathbb{R} \mid f \text{ is continuous} \right\}$ with the metrics given by $d(f, g) = \|f - g\|_{\infty} = \max_{x \in [a, b]} |f(x) - g(x)|$.
- 9) $C^n([a, b]; \mathbb{R}) = \left\{ f : [a, b] \rightarrow \mathbb{R} \mid f \text{ is derivable } n \text{ times with continuous derivatives} \right\}$ with $d(f, g) = \|f - g\|_{\infty}$.
- 10) $C^n([a, b]; \mathbb{R}) = \left\{ f : [a, b] \rightarrow \mathbb{R} \mid f \text{ is derivable } n \text{ times with continuous derivatives} \right\}$ with $d(f, g) = \sum_{i=1}^n \|f^{(i)} - g^{(i)}\|_{\infty}$, where $f^{(i)}$ states for the i -th derivative¹³⁶.

The reader is invited to prove that all those are really metric spaces. The metrics in 1) and in 4) will be called “Euclidean” or also “standard” metrics. The metrics in 8) is also said the “uniform” metrics (or the uniform topology, or the metrics/topology of the uniform convergence)

Definition 5.3 Let X and Y be metric spaces with metrics d_X and d_Y respectively, $f : X \rightarrow Y$ be a function and $\bar{x} \in X$ be fixed.

i) We say that a sequence $\{x_n\}_{n \in \mathbb{N}} \subseteq X$ converges to \bar{x} in X , and we write $x_n \rightarrow \bar{x}$, if the sequence of real numbers $a_n = d_X(x_n, \bar{x})$ converges to $0 \in \mathbb{R}$ ¹³⁷.

ii) We say that f is continuous at \bar{x} if

$$f(x_n) \rightarrow f(\bar{x}) \text{ in } Y \text{ (i.e. } d_Y(f(x_n), f(\bar{x})) \rightarrow 0) \forall \text{ sequence } x_n \rightarrow \bar{x} \text{ in } X.$$

iii) We say that f is continuous if it is continuous at all points of X .

iv) Given $x_0 \in X$ and $r > 0$, the open and closed ball of radius r centered in x_0 are, respectively

$$B(x_0, r) = \left\{ x \in X \mid d_X(x, x_0) < r \right\}, \quad \bar{B}(x_0, r) = \left\{ x \in X \mid d(x, x_0) \leq r \right\}. \quad (5.4)$$

¹³⁵Actually, to be rigorous, we have to speak of a metric space as the couple (X, d) where X is a set and d is a function satisfying (5.3). We prefer the easier way of considering only X as a metric space, with the warning that when we say “ X is a metric space” we will always mean that the metrics d is already fixed.

¹³⁶With $f^{(0)} = f$.

¹³⁷And hence, by definition, $\forall \varepsilon > 0 \exists \bar{n}$ such that $|a_n| = d(x_n, \bar{x}) \leq \varepsilon$ for all $n \geq \bar{n}$.

v) If $\emptyset \neq A \subseteq X$ and $x_0 \in X$, then the diameter of A and the distance of x_0 from A are, respectively

$$\text{diam}(A) = \sup_{x,y \in A} d_X(x,y), \quad d(x_0, A) = \inf_{x \in A} d_X(x, x_0).$$

vi) A non-empty subset $A \subseteq X$ is said to be bounded if its diameter is bounded $\text{diam}(A) < +\infty$.

vii) A subset $C \subseteq X$ is said to be closed if it contains the limit of every converging sequence of its points:

$$\{x_n\} \subseteq C, \quad x_n \rightarrow \bar{x} \in X \implies \bar{x} \in C.$$

viii) If $A \subseteq X$ is a subset, then its closure in X , denoted by \bar{A} or by $\text{cl}(A)$, is the least closed set containing A , and it coincides with all points of X which are limit of a sequence of points of A .

viii) A subset $A \subseteq X$ is said dense in X if $\bar{A} = X$.

The reader is invited to prove that a) if a sequence is convergent then the limit is unique, b) Definition 5.3 ii) exactly recovers the usual definition $\varepsilon - \delta$ of continuity for function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, when they are endowed with the metrics in 1) and/or 4), c) definition v) of boundedness is equivalent to the fact that there exists a ball $B(x_0, r)$, for some $x_0 \in X$ and $r > 0$, which contains A , d) the definition of closedness vii) in the case of \mathbb{R}^n coincides with the usual definition, e) X and \emptyset are closed in X , and the closed balls in (5.4) are closed sets, f) in (5.4) the closed balls are the closure of the respective open balls, g) \mathbb{Q} is dense in \mathbb{R} , \mathbb{Q}^n is dense in \mathbb{R}^n , with the standard Euclidean metrics.

Remark 5.4 By the definition of continuity, we immediately get that, for every $\bar{x} \in X$, the function “distance from \bar{x} ”

$$d(\cdot, \bar{x}) : X \rightarrow [0, +\infty[, \quad x \mapsto d(x, \bar{x})$$

is continuous.

Moreover, if X is a metric space and $A \subseteq X$ is a subset, then also A is a metric space endowed with the metrics $d|_{A \times A} : A \times A \rightarrow [0, +\infty[, (a, b) \mapsto d(a, b)$ for all $a, b \in A$, which is also called the induced metrics by X on A .

5.3 Completeness

We have seen that a consequence of the completeness of \mathbb{R} (as ordered field, i.e. existence of supremum) is the fact that a sequence converges if and only if it is a Cauchy sequence (see Proposition 2.11). Actually, also the converse is true, as it is easy to prove: the fact

that every Cauchy sequence in \mathbb{R} converges implies that every set bounded from above has the supremum¹³⁸.

A metric space is not necessarily ordered and so speaking of infimum and supremum is meaningless. However, the notion of Cauchy sequence is certainly natural for a metric space. Hence we have the following definition.

Definition 5.5 *Let X be a metric space. A sequence $\{x_n\}_n$ of points of X is said to be a Cauchy sequence if*

$$\forall \varepsilon > 0 \exists \bar{n} \in \mathbb{N} \text{ such that } d(x_n, x_m) \leq \varepsilon \forall n, m \geq \bar{n}. \quad (5.5)$$

The metric space X is said to be complete if every Cauchy sequence is convergent¹³⁹.

In Example 5.2, all those metric spaces are complete except \mathbb{Q} in 2)¹⁴⁰, the metric space in 3) and the one in 9).

The metric space in 3) is not complete because the sequence $x_n = n$ is a Cauchy sequence¹⁴¹ but obviously divergent.

Let us prove that the space in 8) is complete. Note that the convergence $f_n \rightarrow f$ is exactly the uniform convergence on $[a, b]$. Let us consider a Cauchy sequence $\{f_n\}$. By definition of $d(f, g) = \|f - g\|_\infty$, for every $x \in [a, b]$ fixed, the sequence $f_n(x)$ is also a Cauchy sequence in \mathbb{R} . Hence, for the completeness of \mathbb{R} , it converges. Let us denote by $f(x)$ the limit of such a sequence. We are going to prove that $f \in C([a, b]; \mathbb{R})$ and that

$$\lim_{n \rightarrow +\infty} \|f_n - f\|_\infty = 0,$$

which will conclude the proof. Let us fix $\varepsilon > 0$ and take the corresponding \bar{n} as in (5.5). Hence, for $n, m \geq \bar{n}$, we have

$$\begin{aligned} \|f_n - f_m\|_\infty \leq \varepsilon &\implies \|f_n(x) - f_m(x)\| \leq \varepsilon \forall x \in [a, b] \\ \implies \lim_{m \rightarrow +\infty} \|f_n(x) - f_m(x)\| \leq \varepsilon \forall x \in [a, b] &\implies \|f_n(x) - f(x)\| \leq \varepsilon \forall x \in [a, b] \\ \implies \|f_n - f\|_\infty \leq \varepsilon. \end{aligned}$$

Since the convergence in metrics is the uniform convergence, then the limit is also continuous: $f \in C([a, b]; \mathbb{R})$, and the proof is finished.

¹³⁸Let m_0 be a majorant of $\emptyset \neq A \subseteq \mathbb{R}$ such that $m_0 - 1$ is not a majorant (it exists). Take $m_0 - 1 < a_1 \leq m_0$ and set $m_1 = m_0$ if $m_0 - (1/2)$ is not a majorant, $m_1 = m_0 - (1/2)$ otherwise. In any case m_1 is majorant and $m_1 - (1/2)$ is not. Hence take $m_1 - (1/2) < a_2 \leq m_1$ and define $m_2 = m_1$ if $m_1 - (1/4)$ is not a majorant, $m_2 = m_1 - (1/4)$ otherwise. In any case m_2 is a majorant and $m_2 - (1/4)$ is not. Proceeding in this way, the sequence a_n is a Cauchy sequence and so convergent to a point $a \in \mathbb{R}$ which cannot be nothing else but the supremum of A . However, note that, in this argumentation, we have somewhere used the archimedean property of \mathbb{R} .

¹³⁹The opposite is always true: any convergent sequence is a Cauchy sequence.

¹⁴⁰ \mathbb{N} and \mathbb{Z} are complete because they are discrete and hence, taking $\varepsilon < 1$ the Cauchy sequence is constant from \bar{n} on.

¹⁴¹ $d(x_n, x_m) = |\arctan n - \arctan m| \rightarrow 0$ as $n, m \rightarrow +\infty$.

Also the space in 10) is complete, and it can be seen arguing as before and applying Proposition 2.44.

The space in 9) is not complete since we can always uniformly approximate a non-derivable function by a sequence of derivable functions: such a sequence is then a Cauchy sequence but it converges to an element which does not belong to the space.

Theorem 5.6 (*Nested spheres theorem*) *Let x_n be a sequence of points of a metric space X , and $r_n > 0$ be a sequence of radii. We say that the closed balls $\overline{B}(x_n, r_n)$ are nested if*

$$\overline{B}(x_{n-1}, r_{n-1}) \supseteq \overline{B}(x_n, r_n) \supseteq \overline{B}(x_{n+1}, r_{n+1}) \quad \forall n \geq 1.$$

The metric space is complete if and only if every sequence of nested closed balls, with $r_n \rightarrow 0$, has nonempty intersection¹⁴².

Proof. If X is complete and we have a nested sequence of closed balls $\overline{B}(x_n, r_n)$, as in the statement, then the sequence of the centers x_n is a Cauchy sequence¹⁴³ and so convergent. Let \bar{x} be its limit. Let us prove that $\bar{x} \in \overline{B}(x_n, r_n)$ for all n , which will imply the conclusion, that is

$$\bar{x} \in \bigcap_n \overline{B}(x_n, r_n) \neq \emptyset.$$

Indeed, for $n \in \mathbb{N}$, by the fact that the balls are closed and nested and by the continuity of the distance function (see Remark 5.4) we have

$$d(x_n, x_m) \leq r_n \quad \forall m \geq n \implies r_n \geq \lim_{m \rightarrow +\infty, m \geq n} d(x_n, x_m) = d(x_n, \bar{x}) \implies \bar{x} \in \overline{B}(x_n, r_n).$$

Vice-versa, let X be a metric space with the nested closed balls property, and let x_n be a Cauchy sequence. Let us take $n_1 \in \mathbb{N}$ such that

$$d(x_n, x_{n_1}) \leq \frac{1}{2} \quad \forall n \geq n_1,$$

which is possible since $\{x_n\}$ is a Cauchy sequence. Let B_1 be the closed ball of radius 1 with center in x_{n_1} . Then, let us take $n_2 > n_1$ such that

$$d(x_n, x_{n_2}) \leq \frac{1}{2^2} \quad \forall n \geq n_2,$$

and define B_2 as the closed ball of radius 1/2 with centered in x_{n_2} . By induction, let n_k and x_{n_k} be such that

$$d(x_n, x_{n_k}) \leq \frac{1}{2^k} \quad \forall n \geq n_k, \quad B_k = \overline{B}\left(x_{n_k}, \frac{1}{2^{k-1}}\right).$$

¹⁴²Obviously, if the intersection is not empty, then it contains just one point only. Indeed, if x', x'' both belong to the intersection of the balls, then, by the definition of balls, $d(x', x'') \leq 2r_n \rightarrow 0 \implies x' = x''$.

¹⁴³Fixed \bar{n} , for all $n, m \geq \bar{n}$ it is $d(x_n, x_m) \leq 2r_{\bar{n}}$ because $x_n, x_m \in \overline{B}(x_{\bar{n}}, r_{\bar{n}})$, and $r_n \rightarrow 0$.

The sequence of closed balls B_k is then nested with radii converging to 0. Hence there exists \bar{x} belonging to all such balls. This easily implies that the subsequence $\{x_{n_k}\}$ converges to \bar{x} . But if a Cauchy sequence, as $\{x_n\}$ is, admits a convergent subsequence, then all the sequence converges to the same limit (see the proof of Proposition 2.11). \square

Remark 5.7 *Note that if the balls are not closed, then they may have empty intersection even if the space is complete. Take for instance $x_n = 1/n$ for $n > 0$ and the open balls, in \mathbb{R} , $]0, 2/n[$ which are nested open balls of radius $r_n = 1/n$ and center $x_n = 1/n$. Their intersection is obviously empty $\bigcap_{n>0}]0, 2/n[= \emptyset$.*

Moreover, an equivalent version of Theorem 5.6 can be given using closed nested subsets instead of closed balls. Indeed X is complete if and only if every nested sequence of closed subsets $C_{n-1} \supseteq C_n \supseteq C_{n+1}$, such that $\text{diam}(C_n) \rightarrow 0$, has non empty intersection. Also in this case note that the closedness is necessary, but also the condition $\text{diam}(C_n) \rightarrow 0$ is essential. Indeed, in \mathbb{R} , take the sequence of closed nested sets $C_n = [n, +\infty[$ and note that $\bigcap_n [n, +\infty[= \emptyset$. In this case, however, the diameters are infinite.

If instead, the nested closed subsets (balls) have finite diameters but do not satisfy $\text{diam}(C_n) \rightarrow 0$, then we anyway conclude that the intersection is not empty (but in this case it may contain more than one point).

Proposition 5.8 *i) Let X be a metric space. If $C \subseteq X$ is complete as metric space (with the distance induced by X), then C is closed in X .*

ii) Let X be a complete metric space and $C \subseteq X$ be a subset. Then C is complete as metric space (with the distance induced by X) if and only if it is closed in X .

Proof. i) Let x_n be a sequence of points of C which converges to \bar{x} in X . Then it is a Cauchy sequence in X and hence in C too. Hence, being C complete, it must be $\bar{x} \in C$.

ii) The necessity is the point i). Let us suppose C closed and let x_n be a Cauchy sequence of points of C . Then it is also a Cauchy sequence in X which is complete and so x_n converges to \bar{x} in X . But C is closed and hence $\bar{x} \in C$. \square

From Proposition 5.8 we immediately get that, given $B \subseteq \mathbb{R}$, the space

$$C^m([a, b], B) = \left\{ f : [a, b] \rightarrow B \mid n - \text{times continuously derivable} \right\}$$

endowed with the metrics as in Example 5.2 point 10), is complete if and only if B is closed¹⁴⁴.

Definition 5.9 *Let X and Y be two metric spaces with metrics d_X and d_Y respectively. A bijective function $f : X \rightarrow Y$ is said to be an isometry if it maintains the distances:*

$$d_X(x_1, x_2) = d_Y(f(x_1), f(x_2)) \quad \forall x_1, x_2 \in X.$$

In the case of existence of an isometry, the two spaces will be said isometric.

¹⁴⁴Indeed, it is closed if and only if B is closed. Anyway B must be not so bad, as set, for instance, finite disjoint union of closed intervals.

Let us note that if $f : X \rightarrow Y$ is a isometry, then also $f^{-1} : Y \rightarrow X$ is a isometry.

Definition 5.10 *Let X be a metric space. A metric space X^* is a completion of X if it is complete and it contains a dense isometric copy of X .*

Theorem 5.11 *Every metric space X has a completion X^* and such a completion is unique in the sense that if X^{**} is another completion, then there exists an isometry $j : X^* \rightarrow X^{**}$ which maps the dense copy of X in X^* onto the dense copy of X in X^{**} .*

Proof. We do not report the proof. We only say that, as already said for a possible completion of \mathbb{Q} in \mathbb{R} , X^* is constructed taking the equivalence classes of Cauchy sequences in X . \square

Remark 5.12 *It is easy to prove that, if X and Y are isometric, then X is complete if and only if Y is complete and, in any case, the completion X^* is isometric to the completion Y^* .*

It is rather evident that if F is an archimedean field, then it is also a metric space endowed with the distance

$$d(x, y) = \varphi(|x - y|_F),$$

where $|\cdot|_F$ is the absolute value in F (see (3.5)) and φ is the injective isomorphism from F to \mathbb{R} (see remark 3.24).

One may then ask whether completing F as ordered field and completing F as metric space give the same completion, that is \mathbb{R} . The answer is of course “yes, the completions are the same”.

Moreover, we know that the completion of \mathbb{Q} is \mathbb{R} and the completion of \mathbb{Q}^n is \mathbb{R}^n , when all of them are endowed by the Euclidean distance. However, we also know that \mathbb{Q} is equivalent to \mathbb{Q}^n , that is there exists a bijection $\varphi : \mathbb{Q} \rightarrow \mathbb{Q}^n$. Hence we can equip \mathbb{Q} with the following distance:

$$d(p, q) = \|\varphi(p) - \varphi(q)\|_n,$$

where $\|\cdot\|_n$ is the Euclidean norm in \mathbb{R}^n . It is obvious that $\varphi : (\mathbb{Q}, d) \rightarrow (\mathbb{Q}^n, \|\cdot\|_n)$ is an isometry. Hence, the completion of \mathbb{Q} with the distance d is isometric to (roughly speaking: it is equal to) \mathbb{R}^n . Hence, we deduce that the completion strictly depends on the metric structure.

Finally, the completion of the incomplete metric space \mathbb{R} with metrics $d(x, y) = |\arctan x - \arctan y|$ is the metric space given by the set $[-\infty, +\infty]$ and distance d^ acting as*

$$\begin{aligned} d^*(x, y) &= d(x, y) && \text{if } x, y \in \mathbb{R}, && d^*(-\infty, x) = \arctan x + \frac{\pi}{2} && \text{if } x \in \mathbb{R}, \\ d^*(+\infty, x) &= \frac{\pi}{2} - \arctan x && \text{if } x \in \mathbb{R}, && d^*(-\infty, +\infty) &= \pi. \end{aligned}$$

The completeness of a metric space is important because of the existence of the limits of the Cauchy sequences. Think for instance to problems of approximation or to problems of iterative procedures where at every new step you construct a new quantity x_n (a number, a vector, a function, an algorithm...). If such quantities are embedded in a complete metric space and they form a Cauchy sequence, then you are sure that you are converging to somewhere.

Another very important framework (which is however somehow linked to the previous situation) where the completeness of a metric space plays a crucial role is in the fixed point theory, that is in the theory concerning the solvability of the equation

$$F(x) = x$$

where $F : X \rightarrow X$ is a given function on a given metric space. Here we report, without the easy proof, the famous result¹⁴⁵ which goes under several names such as “Contraction Principle”, “Banach-Caccioppoli Theorem”, “Lipschitz-Picard Theorem” ...

Theorem 5.13 *Let X be a metric space and $F : X \rightarrow X$ be a function. We say that F is a contraction if there exists a constant $0 \leq L < 1$ such that*

$$d(F(x), F(y)) \leq Ld(x, y) \quad \forall x, y \in X.$$

We say that $x \in X$ is a fixed point of F if

$$F(x) = x$$

If X is complete and F is a contraction, then there exists one and only one fixed point for F .

Note that if we only have the inequality $d(F(x), F(y)) \leq d(x, y)$, i. e. it holds with $L = 1$ and it is not possible to take $L < 1$ uniformly for every pair (x, y) , then we may have existence of more than one fixed points (even infinitely many) or no existence of fixed points at all. Just take, as examples, the metric space \mathbb{R} with the usual metrics and the functions $F(x) = x$ (identity) and $F(x) = x + 1$ (rigid translation).

5.4 Compactness in metric spaces

Definition 5.14 *Let X be a metric space and $C \subseteq X$ be a subset. We say that C is compact if for every sequence $\{x_n\}$ of points of C there exists a subsequence $\{x_{n_k}\}$ and a point $\bar{x} \in C$ such that*

$$x_{n_k} \rightarrow \bar{x} \text{ as } k \rightarrow +\infty.$$

We say that X is a compact metric space if X is compact (as subset of X itself).

¹⁴⁵Certainly known to the reader.

Proposition 5.15 *Let X be a metric space. i) If $C \subseteq X$ is compact then C is closed and, as metric space with the induced metrics, it is complete. ii) If X is compact and $C \subseteq X$ is closed, then C is compact. iii) If $C \subseteq X$ is compact, then C is bounded.*

Proof. i) If x_n is a sequence of points of C converging to $\bar{x} \in X$, then by compactness, there is a subsequence converging to a point of C and such a point, by uniqueness of the limit, must be \bar{x} . If x_n is Cauchy sequence of point of C , then by compactness it has a convergent subsequence to $\bar{x} \in C$; but if a Cauchy sequence has a convergent subsequence to $\bar{x} \in C$ then the whole sequence converges to \bar{x} . ii) Let x_n be a sequence of points of C ; since X is compact, then there exists a subsequence x_{n_k} converging to a point $\bar{x} \in X$; but C is closed and x_{n_k} is contained in C and so $\bar{x} \in C$. iii) By absurd, let us suppose that C is not bounded; hence there exist two sequences of points of C , $\{x_n\}, \{y_n\}$, such that $d(x_n, y_n) \rightarrow +\infty$ as $n \rightarrow +\infty$; hence they respectively have a converging subsequence $x_{n_k} \rightarrow \bar{x} \in C$, $y_{n_k} \rightarrow \bar{y} \in C$ ¹⁴⁶, from which the contradiction

$$d(x_{n_k}, y_{n_k}) \leq d(x_{n_k}, \bar{x}) + d(\bar{x}, \bar{y}) + d(\bar{y}, y_{n_k}) \leq 1 + d(\bar{x}, \bar{y}) < +\infty$$

for all k sufficiently large. □

Remark 5.16 *The reader certainly knows that a subset of \mathbb{R}^n (with the Euclidean distance) is compact if and only if it is a closed and bounded subset. The necessity is stated in Proposition 5.15, whereas the sufficiency comes from the Bolzano-Weierstrass Theorem 2.9¹⁴⁷.*

However, for a general metric space, the property of being closed and bounded is not sufficient for being compact. Take for instance the metric space $C^0([-1, 1]; \mathbb{R})$ with the distance $d(f, g) = \|f - g\|_\infty$, and consider the sequence of continuous functions as in Example 2.41 point ii): all such functions belong on the closed unit ball $B = \{f : [-1, 1] \rightarrow \mathbb{R} \text{ continuous} | d(f, 0) = \|f\|_\infty \leq 1\}$ (which is closed and bounded) but no subsequence may uniformly converge because the pointwise limit is discontinuous.

As explained in Remark 5.16, the boundedness (together with closedness) is not sufficient for compactness. The right concept is the following one.

Definition 5.17 *Let X be a metric space. We say that $C \subseteq X$ is totally bounded if for every $\varepsilon > 0$ there exists a finite collection of balls of radius ε , $B(x_i, \varepsilon)$ $i = 1, \dots, n$, whose union covers C*

$$C \subseteq \bigcup_{i=1}^n B(x_i, \varepsilon)$$

¹⁴⁶We can suppose that they have the same subindex k because we can first extract a converging subsequence x_{n_j} from x_n and hence a converging subsequence $y_{n_{j_k}}$ from y_{n_j} , and finally put $n_k = n_{j_k}$.

¹⁴⁷Note that if C is closed and bounded in \mathbb{R}^n then it is contained in a closed n -cube $I_1 \times I_2 \times \dots \times I_n$ where I_i is closed and bounded interval of \mathbb{R} , and hence, for getting the “ n -dimensional version of the Bolzano-Weierstrass Theorem” you just need to repeat n -times the “1-dimensional version” component by component.

It is easy to prove that the total boundedness implies the boundedness. The following theorem is a first characterization of the compact metric spaces (and hence of the compact subsets of a metric space.)

Theorem 5.18 *A metric space X is compact if and only if it is complete and totally bounded.*

Proof. (Necessity) The necessity of the completeness is stated in Proposition 5.15. Let us prove that X must be totally bounded. If not, then there exists $\bar{\varepsilon} > 0$ such that every finite collection of balls with radius $\bar{\varepsilon} > 0$ does not cover X . Hence, take a point $x_1 \in X$, and note that $X \setminus B(x_1, \bar{\varepsilon}) \neq \emptyset$. Hence, take $x_2 \in X \setminus B(x_1, \bar{\varepsilon})$, and note that there must exist $x_3 \in X \setminus (B(x_1, \bar{\varepsilon}) \cup B(x_2, \bar{\varepsilon}))$. Proceeding in this way we can construct a sequence in X such that $d(x_n, x_{n+1}) \geq \bar{\varepsilon} > 0$. Hence no subsequence of $\{x_n\}$ may converge: contradiction.

(Sufficiency) Let $\{x_n\}$ be a sequence in X . We are going to prove that the totally boundedness implies the existence of a subsequence which is a Cauchy sequence, from which, by completeness we conclude. For every k take a finite collection \mathcal{B}_k of balls of radius $1/2^k$ covering X . For $k = 1$ take a ball $B_1 \in \mathcal{B}_1$ which contains infinitely many points of the sequence¹⁴⁸ x_n ; for $k = 2$ take $B_2 \in \mathcal{B}_2$ such that $B_1 \cap B_2$ contains infinitely many points of the sequence x_n ; in general, take $B_k \in \mathcal{B}_k$ such that $B_1 \cap B_2 \cap \cdots \cap B_k$ contains infinitely many points of $\{x_n\}$. Let n_1 be the first n such that $x_n \in B_1$ and in general n_k be the first $n > n_{k-1}$ such that $x_n \in B_1 \cap \cdots \cap B_k$. The subsequence x_{n_k} is then a Cauchy sequence because, for all k', k'' ,

$$x_{n_{k'}}, x_{n_{k''}} \in B_{\min\{k', k''\}} \implies d(x_{n_{k'}}, x_{n_{k''}}) \leq \frac{2}{2^{\min\{k', k''\}}} \rightarrow 0 \text{ as } k', k'' \rightarrow +\infty.$$

□

Remark 5.19 *By Remark 5.16 and Theorem 5.18 we deduce that the bounded subsets of \mathbb{R}^n are totally bounded. A direct proof of this fact is nothing else but the use of the archimedean property of \mathbb{R} : every bounded sets is contained in a hypercube $I_1 \times \cdots \times I_n$ which, for every $\varepsilon > 0$, is covered by a finite family of hypercubes $J_1 \times \cdots \times J_n$ with J_i interval of length ε . And, if you want to cover by balls, just note that any hypercube contains a ball and any ball contains a hypercube.*

The possibility of covering a compact metric space with a finite family of balls with fixed radius is also an important ingredient of the following other characterization of the compact metric space. Given an index set \mathcal{I} and a family of open balls of X (not necessarily all of the same radius), $\mathcal{F} = \{B_i | B_i \subseteq X \text{ open ball, } i \in \mathcal{I}\}$, a finite subfamily of \mathcal{F} is a family of open balls $\mathcal{B} = \{B_i | i \in I\}$ where $I \subseteq \mathcal{I}$ is a finite index subset.

Theorem 5.20 *A metric space X is compact if and only if for every family \mathcal{F} of open balls covering X there exists a finite subfamily \mathcal{B} still covering X .*

¹⁴⁸More precisely: it contains points of the sequence for infinitely many indices n .

Proof. (Necessity) Let us assume X compact. Let \mathcal{F} be a family of open balls covering X . For every n let \mathcal{B}_n be a finite family of open balls of radius $1/2^n$ covering X , which exists by Theorem 5.18. If we prove that there exists \bar{n} such that for every $B \in \mathcal{B}_{\bar{n}}$ there exists $S \in \mathcal{F}$ such that $B \subseteq S$, then we are done. Indeed, at that point, it will be sufficient to take the finite subfamily $\{S \in \mathcal{F} | \exists B \in \mathcal{B}_{\bar{n}}, B \subseteq S\}$. By absurd, let us suppose that for every n there exists a ball $B \in \mathcal{B}_n$ such that $B \not\subseteq S$ for all $S \in \mathcal{F}$ ¹⁴⁹. If $\{x_n\}$ is the sequence of the centers of such balls $B \in \mathcal{B}_n$, then, by compactness, there exists $\bar{x} \in X$ and a subsequence x_{n_k} converging to \bar{x} . Now, by covering, there exists $B(\bar{y}, r) = S \in \mathcal{F}$ such that $\bar{x} \in S$. But then, taking $\bar{r} > 0$ such that $B(\bar{x}, \bar{r}) \subseteq B(\bar{y}, r) = S$ ¹⁵⁰, for every k such that $\min\{1/2^{n_k}, d(x_{n_k}, \bar{x})\} < \bar{r}/2$ we have $B(x_{n_k}, 1/2^{n_k}) \subseteq S$ which is a contradiction.

(Sufficiency) X is totally bounded because, for every $\varepsilon > 0$ we have the covering family of open balls $\{B(x, \varepsilon) | x \in X\}$ from which we can extract a finite covering subfamily. Let us prove that X is complete and we conclude by Theorem 5.18. Let $\{x_n\}$ be a Cauchy sequence and suppose that it does not converge. Since a non-converging Cauchy sequence cannot have any converging subsequence (otherwise it is itself convergent), for every $x \in X$ there exist $\varepsilon_x > 0$ and $n_x \in \mathbb{N}$ such that

$$x_n \notin B(x, \varepsilon_x) \quad \forall n \geq n_x.$$

From the covering family $\{B(x, \varepsilon_x) | x \in X\}$ we can extract a covering finite subfamily $B(x_i, \varepsilon_{x_i})$ for $i = 1, \dots, N$ for some $N \in \mathbb{N}$. Taking $\bar{n} = \max\{n_{x_i} | i = 1, \dots, N\}$ we get the contradiction

$$x_n \notin \bigcup_{i=1}^N B(x_i, \varepsilon_{x_i}) = X \quad \forall n \geq \bar{n}.$$

□

The property of nested closed sets also holds in the case of compactness.

Proposition 5.21 *Let X be a metric space, and $\{C_n\}$ a collection of nested compact subset of X such that $\text{diam}(C_n) \rightarrow 0$. Then, the intersection $\bigcap_n C_n$ is not empty (and contains just one point).*

Proof. Since C_0 is compact, then it is also complete. Hence we only need to apply Theorem 5.6 (see also Remark 5.7) to the complete metric space C_0 with nested sequence of closed subsets C_n . □

5.5 The importance of being compact

Similarly to completeness, the compactness (which is a stronger property, as we know) is important for problems of approximation and of solving equations. For instance, if $F : X \rightarrow Y$ is a function, $\bar{y} \in Y$ is fixed, and we want to find a solution of the equation

¹⁴⁹Note that B is certainly contained in the union of the balls $S \in \mathcal{F}$ but not necessarily contained in a single specific ball $S \in \mathcal{F}$.

¹⁵⁰Which is possible because the balls are open.

$$F(x) = \bar{y},$$

then, one possible way to attack the problem (also by a numerical point of view) is to try to study a suitable approximating equation

$$F_n(x) = \bar{y}$$

where, for every $n \in \mathbb{N}$, $F_n : X \rightarrow Y$ is an approximating function. Let us suppose that the approximating F_n have the property:

$$x_n \rightarrow x \text{ in } X \implies F_n(x_n) \rightarrow F(x) \text{ in } \mathbb{R}. \quad (5.6)$$

Hence, if for every n , we can sufficiently easily calculate (or proving the existence of) a solution $x_n \in X$ for $F_n(x) = \bar{y}$, then whenever $\{x_n\}$ is contained in a compact subset of X , we can extract a subsequence converging to a point $\bar{x} \in X$ and so obtain

$$x_{n_k} \rightarrow \bar{x}, F_{n_k}(x_{n_k}) = \bar{y} \implies F(\bar{x}) = \bar{y} \implies \bar{x} \text{ is a solution.}$$

A typical case where the property (5.6) is satisfied is when F_n uniformly converges to F , that is

$$\forall \varepsilon > 0 \exists \bar{n} \text{ such that } n \geq \bar{n} \implies d_Y(F_n(x), F(x)) \leq \varepsilon \forall x \in X.$$

One of the major field of applications of such a method is in the theory of differential equations, as next example shows.

Example 5.22 As it is well-known, a solution of the Cauchy problem

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = x_0, \end{cases}$$

is a solution of the equation $F(y) = 0$, where $F : C^0([t_0 - \delta, t_0 + \delta], \mathbb{R}^n) \rightarrow C^0([t_0 - \delta, t_0 + \delta], \mathbb{R}^n)$ acts as

$$v \mapsto F(v) : t \mapsto v(t) - x_0 - \int_{t_0}^t f(s, v(s)) ds,$$

with a suitable $\delta > 0$. We can approximate the equation by considering

$$F_n(v) : t \mapsto \begin{cases} v(t) - x_0 - \int_{t_0}^t f\left(s, v\left(s - \frac{1}{n}\right)\right) ds & \text{if } t \geq t_0 - \delta + \frac{1}{n}, \\ 0 & \text{otherwise} \end{cases}$$

The existence of a solution y_n of the delayed approximating equation $F_n(y) = 0$ is easily proven by an iterative argument. If one is able to prove that all such delayed solutions belong to the same compact set of $C^0([t_0 - \delta, t_0 + \delta], \mathbb{R}^n)$, then one can prove the existence of a solution of the Cauchy problem. This is indeed one possible way for proving the existence theorem of Peano, under continuity hypothesis for f .

Another very common setting where compactness plays an essential role is in the problem of minimization and maximization of functions. We have the following result.

Proposition 5.23 *A function $f : X \rightarrow \mathbb{R}$ is said lower semicontinuous (respectively: upper semicontinuous) if for every $\bar{x} \in X$ ¹⁵¹:*

$$f(\bar{x}) \leq \liminf_{n \rightarrow +\infty} f(x_n), \quad \left(\text{respectively: } \limsup_{n \rightarrow +\infty} f(x_n) \leq f(\bar{x}) \right) \quad \forall \text{ sequence } x_n \rightarrow \bar{x} \text{ in } X.$$

If $C \subseteq X$ is compact and $f : X \rightarrow \mathbb{R}$ is lower semicontinuous (respectively, upper semicontinuous, continuous) then f reaches its minimum (respectively its maximum, its minimum and maximum) on C .

Proof. The proof is similar to the one of Theorem 2.21. □

One way to use the previous result is in the so-called *direct method of the calculus of variations*. We say that a function $F : X \rightarrow \mathbb{R}$ is coercive, if the following holds¹⁵²

$$A \subseteq X, \bar{A} \text{ not compact} \implies \exists \{x_n\}_n \subseteq A \text{ such that } F(x_n) \rightarrow +\infty.$$

Proposition 5.24 *Let X be a metric space and $F : X \rightarrow \mathbb{R}$ be lower semicontinuous and coercive. Then, F has a minimum on X , that is there exists $\bar{x} \in X$ such that*

$$F(\bar{x}) \leq F(x) \quad \forall x \in X.$$

Proof. Let us define $m = \inf_{x \in X} F(x) \in [-\infty, +\infty[$. By definition of infimum, there exists a sequence of points $x_n \in X$ such that

$$F(x_n) \rightarrow m \text{ as } n \rightarrow +\infty. \tag{5.7}$$

By the coercivity, the closure $\overline{\{x_n\}_n}$ of the sequence $\{x_n\}_n$ must be compact, otherwise there is a subsequence $\{x_{n_k}\}_k$ with $F(x_{n_k}) \rightarrow +\infty$, which is a contradiction to (5.7). Hence, by lower semicontinuity, F has a minimum on $\overline{\{x_n\}_n}$, which means that there exists $\bar{x} \in \overline{\{x_n\}_n}$ such that

$$F(\bar{x}) \leq F(x_n) \quad \forall n,$$

and so $F(\bar{x}) = m$ by (5.7) and definition of m . □

The typical setting of application of such a result is in the theory of minimization of functionals defined on some suitable functions space and representing some energy to be minimized. Such a theory is also extremely linked to the theory of partial differential equations as it is shown in the following example.

¹⁵¹Since the function f has \mathbb{R} as codomain, the definitions of \liminf and \limsup are as the standard ones given in Definition 2.13.

¹⁵²In the simpler case of $X = \mathbb{R}^n$, a typical situation is when $F(x_k) \rightarrow +\infty$ whenever $\|x_k\| \rightarrow +\infty$.

Example 5.25 Let us consider the Dirichlet problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (5.8)$$

where $\Omega \subset \mathbb{R}^n$ is an open regular set, $u : \bar{\Omega} \rightarrow \mathbb{R}$ is the unknown function, $\Delta = \sum_{i=1}^n (\partial^2 u) / (\partial x_i^2)$ is the *Laplacian*, and $f : \Omega \rightarrow \mathbb{R}$ is a given function.

If u is a (classical) solution¹⁵³, then, for every functions $\varphi \in C^1(\bar{\Omega})$ with $\varphi = 0$ on $\partial\Omega$ (a so-called test function), by the Green formulas we have

$$-\Delta u = f \implies - \int_{\Omega} \varphi \Delta u = \int_{\Omega} \varphi f \implies \int_{\Omega} \nabla u \cdot \nabla \varphi = \int_{\Omega} \varphi f, \quad (5.9)$$

where ∇u is the gradient $((\partial u) / (\partial x_1), \dots, (\partial u) / (\partial x_n))$. Also the opposite implication holds: if a C^2 function which vanishes on the boundary and satisfies the last inequality of (5.9) for all test functions φ , then it also satisfies the Dirichlet problem (5.8).

On the other hand, let us consider the functional

$$J : C^1(\Omega) \rightarrow \mathbb{R}, \quad J(u) = \int_{\Omega} (\|\nabla u\|^2 - 2fu) \, dx.$$

If u is a point of minimum of J , then u satisfies the first equation in (5.9) (the Laplace equation) in the so-called “weak” sense, that is

$$\int_{\Omega} \nabla u \cdot \nabla \varphi = \int_{\Omega} \varphi f \quad \forall \text{ test functions } \varphi.$$

Indeed, let $\varepsilon > 0$ and φ (test function) be arbitrary, if u is a minimum then $J(u) \leq J(u + \varepsilon\varphi)$ and so the function

$$\varepsilon \mapsto J(u + \varepsilon\varphi)$$

has a minimum in $\varepsilon = 0$. Hence¹⁵⁴

$$\frac{d}{d\varepsilon} J(u + \varepsilon\varphi)|_{\varepsilon=0} = 0 \implies \int_{\Omega} (\nabla u \cdot \nabla \varphi - f\varphi) \, dx = 0,$$

and we conclude by the arbitrariness of the test function φ .

If then J is lower semicontinuous and coercive, we get the existence of a minimum for J and consequently the existence of a weak solution of (5.8).

Remark 5.26 *The reader has certainly noted that for being a solution of the Dirichlet problem (5.8) it required to be of class C^2 , whereas for being a minimum of J it is required to be of class C^1 . Indeed, the validity of the last inequality in (5.9) for all test functions, can be taken as a weak definition of solution of (5.8), which only require to be C^1 . Hence*

¹⁵³i.e. u has the regularity C^2 and satisfies the equation point by point.

¹⁵⁴Deriving under the sign of integral.

we may look for a weak solution as a C^1 function which minimizes J , knowing that, if the minimizer is also C^2 then it classically solves (5.8).

The problem of proving the existence of a weak solution of (5.8) can be then viewed as the problem of proving that J is lower semicontinuous and coercive. To this end, we can also change the natural distance in the domain of J , and put a different distance which makes J lower semicontinuous and coercive. For applying this procedure it is necessary (among other things) to very well know which are the compact subsets of the domain with the new distance, in order to be sure that J is coercive.

However, let us note that (as indeed it happens in the case of the Dirichlet problem), it may be also necessary to enlarge the possible domain and to introduce on it a new kind of convergence which may be not generated by a distance, in other words to change the topology in a more general way than changing the distance. The concept of topology is addressed in the next section.

5.6 Topological spaces

In the previous paragraph we have studied the properties of the concept of “distance”. Such a concept is used for defining a criterium of “closeness” which is required for the concept of “limit”. In this paragraph we study the other possible way of treat “closeness”: via the concept of “aroundness” which, as we are going to see, generalizes the use of the distance.

In a metric space, the intuitive concept of “aroundness” is certainly given by the concept of “ball”: the points of a ball “stay around” the center of the ball itself. However it is more natural to extend the concept of “aroundness” to every set which contains a ball; in other words we say that a set U is a neighborhood of a point x if it contains an open ball centered in x ¹⁵⁵. Here are the essential properties of neighborhoods in a metric space, whose proofs is almost immediate. Denoting by¹⁵⁶ $\mathcal{F}(x) \subseteq \mathcal{P}(X)$ the family of neighborhoods of $x \in X$, we have¹⁵⁷

$$\begin{aligned}
 i) \quad & \mathcal{F}(x) \neq \emptyset \quad \forall x \in X; \quad U \in \mathcal{F}(x) \implies x \in U; \\
 ii) \quad & U, V \in \mathcal{F}(x) \implies U \cap V \in \mathcal{F}(x); \\
 iii) \quad & U \in \mathcal{F}(x), \quad U \subseteq V \implies V \in \mathcal{F}(x); \\
 iv) \quad & U \in \mathcal{F}(x) \implies \exists V \in \mathcal{F}(x) \text{ such that } U \in \mathcal{F}(y) \quad \forall y \in V; \\
 v) \quad & x \neq y \implies \exists U \in \mathcal{F}(x), \exists V \in \mathcal{F}(y) \text{ such that } U \cap V = \emptyset.
 \end{aligned}
 \tag{5.10}$$

If we are concerning with a concept of “aroundness” the properties i), ii) and iii) are almost obvious to be required; the property v) is obvious if we think to neighborhoods as

¹⁵⁵Equivalently: if it contains a closed ball centered in x ; note however that, by definition, a ball (closed or open) always has a strictly positive radius.

¹⁵⁶Recall that $\mathcal{P}(A)$ denotes the set of all the parts (subsets) of the set A .

¹⁵⁷By words, such properties are: i) every point has a neighborhood; every neighborhood of x contains x ; ii) the intersection of two neighborhoods of x is still a neighborhood of x ; iii) if a set contains a neighborhood of x then it is itself a neighborhood of x ; iv) every neighborhood U of x contains a neighborhood V of x such that U is a neighborhood of every points of V , v) two different points have two neighborhoods with non shared points.

something containing a ball (by the fact that for every two different points there exist two non intersecting balls centered on them); let us concentrate on iv). The meaning of it is, roughly speaking, that if U stays around to x then it also stays around to all surrounding points which are “sufficiently close” to x , the points of V indeed. Its importance will be clarified next, when speaking of open sets. Inspired by what happens in a metric space, we then give the following definition.

Definition 5.27 *Let X be a nonempty set. A topological structure on X is a function¹⁵⁸*

$$X \rightarrow \mathcal{P}(\mathcal{P}(X)), \quad x \mapsto \mathcal{F}(x)$$

such that, for every $x \in X$, $\mathcal{F}(x)$ satisfies the properties i), ii), iii), iv) of (5.10). In such a case we say that X is a topological space and, for every $x \in X$, $\mathcal{F}(x) \subseteq \mathcal{P}(x)$ is called the neighborhoods filter of x , and the elements of $\mathcal{F}(x)$ the neighborhoods of x .

If moreover, $\mathcal{F}(x)$ also satisfies v) in (5.10), we say that X is a separated or Hausdorff space.

The following proposition is obvious.

Proposition 5.28 *Every metric space is a topological space with topological structure (neighborhoods filter) given by*

$$\mathcal{F}(x) = \left\{ U \subseteq X \mid U \text{ contains an open ball centered in } x \right\}, \quad \forall x \in X.$$

Moreover, every metric space is a Hausdorff space.

Whence we have a topological structure, it is interesting to consider the sets which are neighborhood of all their points.

Definition 5.29 *Let X be a topological space. A non-empty set $A \subseteq X$ is said to be open if it is a neighborhood of all its points, that is*

$$A \in \mathcal{F}(x) \quad \forall x \in A.$$

Of course the whole space X is open¹⁵⁹, and, by definition, we can also assume \emptyset open. The following proposition shows that the open sets are indeed very “distributed”.

Proposition 5.30 *Let X be a topological space, $x \in X$ and $U \in \mathcal{F}(x)$. Then, U contains a non-empty open neighborhood of x , that is*

$$\exists A \in \mathcal{F}(x), \quad A \subseteq U, \quad A \text{ open.}$$

¹⁵⁸Here, $\mathcal{P}(\mathcal{P}(X))$ is the set of the parts of the parts of X ; the set of the families of subsets of X .

¹⁵⁹Because it is obviously a neighborhood of every point.

Proof. We define the non-empty set (non-empty because it contains x)

$$A = \left\{ y \in X \mid U \in \mathcal{F}(y) \right\},$$

and we conclude if we prove that A is open. Let $y \in A$, then $U \in \mathcal{F}(y)$ and hence, by property iv) of (5.10) there exists $V \in \mathcal{F}(y)$ such that $U \in \mathcal{F}(z)$ for all $z \in V$. We then deduce, by definition, that $V \subseteq A$ and so $A \in \mathcal{F}(y)$. \square

Definition 5.31 Let X be a topological space and $C \subseteq X$ be a subset. C is said to be closed if its complementary, $\complement C = X \setminus C$, is open. A point $x \in X$ is said to be an adherent point for C if

$$U \cap C \neq \emptyset \quad \forall U \in \mathcal{F}(x).$$

A point $x \in X$ is said to be an accumulation point for C if

$$\forall U \in \mathcal{F}(x) \exists y \in U \cap C \text{ such that } y \neq x.$$

A point $x \in C$ is said to be isolated in C if there exists $U \in \mathcal{F}(x)$ such that $U \cap C = \{x\}$. The closure or the adherence of C , denoted by \bar{C} or by $cl(C)$, is the set of all adherent points for C . The interior of C , denoted by $\overset{\circ}{C}$ or by $int(C)$ is the set of points $x \in C$ such that $C \in \mathcal{F}(x)$.

Proposition 5.32 a) The closure of a set is closed; the interior of a set is open. b) A set $C \subseteq X$ is closed (respectively, open) if and only if it coincides with its closure (respectively, interior). c) In a metric space X a subset C is closed in the sense of Definition 5.31 if and only if it is (sequentially) closed in the sense of Definition 5.3. Also the closures of the set by the two definitions coincide. d) In a metric space X the open and closed balls defined in (5.4) are open and closed in the sense of Definitions 5.29 and 5.31 respectively. e) A point $x \in X$ is isolated in X if and only if $U = \{x\}$ is a neighborhood of x , that is $\{x\} \in \mathcal{F}(x)$. f) X and \emptyset are closed.

Proof. We only prove a) for the interior and c) for closed sets. a) if $\overset{\circ}{A}$ is empty, then it is open by definition. Let $x \in \overset{\circ}{A}$. Hence $A \in \mathcal{F}(x)$ and so there exists $U \in \mathcal{F}(x)$ such that $A \in \mathcal{F}(y)$ for all $y \in U$. This implies $U \subseteq \overset{\circ}{A}$ and so $\overset{\circ}{A} \in \mathcal{F}(x)$. c) Let C be sequentially closed (i.e. as in Definition 5.3) and let x be an adherent point of C . Since any open ball $B(x, 1/n)$ belongs to $\mathcal{F}(x)$, then it must contain a point of C , let us say x_n . But then the sequence $\{x_n\}$ is convergent to x by construction and so $x \in C$. On the other hand, let us suppose C closed as in Definition 5.31, and suppose that $\{x_n\} \subseteq C$ converges to $x \in X$. Take any $U \in \mathcal{F}(x)$. By definition of neighborhoods in a metric space, there is a ball $B(x, r) \subseteq U$. By convergence, there exists n such that $x_n \in B(x, r) \subseteq U$, and so x is an adherent point for C , so $x \in C$, and hence C turns out to be sequentially closed. \square

Proposition 5.33 *The family \mathcal{T} of all open subsets of a topological space X has the following properties¹⁶⁰:*

- i) $X, \emptyset \in \mathcal{T}$;
- ii) I index set (even non countable), and $A_i \in \mathcal{T} \forall i \in I \implies \bigcup_{i \in I} A_i \in \mathcal{T}$; (5.11)
- iii) $N \in \mathbb{N}, A_i \in \mathcal{T} \forall i = 1, \dots, N \implies \bigcap_{i=1}^N A_i \in \mathcal{T}$.

Proof. The proof is easy. Only note that the property ii) in (5.10) easily extend to finite intersections of neighborhoods. \square

Note that, in general, the countable (or more than countable) intersection of open sets is not open. For instance, in \mathbb{R} with the usual metrics, the open intervals $]a, b[$ are open, but the countable intersection $\bigcap_{n \geq 1}](1/n), 1/n[$ is equal to the singleton $\{0\}$ which is not open in \mathbb{R} .

Definition 5.34 *Let X be a set. A family of subsets of X , $\mathcal{T} \subseteq \mathcal{P}(X)$, is called a topology on X if it satisfies the three properties of (5.11).*

Proposition 5.35 *Let X be a nonempty set, and \mathcal{T} be a topology on X . Hence, there exists a unique topological structure $x \mapsto \mathcal{F}(x)$ such that \mathcal{T} is exactly the family of the open subsets for it.*

Proof. Given \mathcal{T} topology on X we define, for every $x \in X$

$$\mathcal{F}(x) = \left\{ U \subseteq X \mid \exists A \in \mathcal{T} \text{ such that } x \in A \subseteq U \right\}.$$

It is not hard to see that $x \mapsto \mathcal{F}(x)$ is a topological structure on X . We define the family of open subsets for such a topological structure

$$\mathcal{T}' = \left\{ A \subseteq X \mid A \in \mathcal{F}(x) \forall x \in A \right\},$$

and we prove that $\mathcal{T} = \mathcal{T}'$. Indeed, if $A \in \mathcal{T}$ and $x \in A$, then we have $x \in A \subseteq A$ and so $A \in \mathcal{F}(x)$ which implies, by the arbitrariness of $x \in A$, $A \in \mathcal{T}'$. On the other hand, if $A \in \mathcal{T}'$, then, by definition, $A \in \mathcal{F}(x)$ for all $x \in A$, and so, for every $x \in A$ there exists $A_x \in \mathcal{T}$ such that $x \in A_x \subseteq A$. We then get

$$A = \bigcup_{x \in A} A_x \in \mathcal{T}.$$

Now we prove uniqueness. Let us suppose that $x \mapsto \mathcal{F}(x)$ and $x \mapsto \mathcal{F}'(x)$ are two topological structures with the same family of open sets:

$$\mathcal{T} = \left\{ A \subseteq X \mid A \in \mathcal{F}(x) \forall x \in A \right\} = \left\{ A \subseteq X \mid A \in \mathcal{F}'(x) \forall x \in A \right\} = \mathcal{T}'.$$

¹⁶⁰By words, ii) means that every union of open sets is open, and iii) means that every finite intersection of open sets is open.

Then, $\mathcal{F}(x) = \mathcal{F}'(x)$ for all $x \in X$. Indeed, if $U \in \mathcal{F}(x)$, by Proposition 5.30 there exists $A \in \mathcal{T}$ such that $x \in A \subseteq U$. But A also belongs to \mathcal{T}' , and so we get $U \in \mathcal{F}'(x)$ by definition. Hence $\mathcal{F}(x) \subseteq \mathcal{F}'(x)$ and, being the opposite inclusion proven in the same way, we conclude. \square

Remark 5.36 *From Proposition 5.35 we deduce that to give a topology on a set (i.e. to say which are the open sets) and to give a topological structure (i.e. to say which are the neighborhoods for every point) are equivalent: given a topology there exists a unique topological structure which generates that topology as family of open subsets; vice-versa, if two topological structures generates the same family of open subsets, then they are the same topological structure.*

By the previous argumentation, a topological space is often denoted by the couple (X, \mathcal{T}) , where X is a nonempty set and \mathcal{T} is a topology on it.

Of course, instead of giving a topology (i.e. giving the open subsets) we can equivalently give the family of closed subsets $\mathcal{C} = \{C \subseteq X \mid C \text{ is closed}\}$, because, via complementation, the family of closed subsets uniquely identifies the family of open subsets. A general family \mathcal{C} of subsets is the family of the closed subsets for a topology if and only if

- i) $X, \emptyset \in \mathcal{C}$;
- ii) I index set (even non countable), and $C_i \in \mathcal{C} \forall i \in I \implies \bigcap_{i \in I} C_i \in \mathcal{C}$; (5.12)
- iii) $N \in \mathbb{N}$, $C_i \in \mathcal{C} \forall i = 1, \dots, N \implies \bigcup_{i=1}^N C_i \in \mathcal{C}$.

To see it, just recall that the complement of the union (respectively, of the intersection) is the intersection (respectively, the union) of the complements.

We end this paragraph with the following consideration: in a metric space X the open balls are open subsets and neighborhoods of their centers. Moreover a subset $U \subseteq X$ is a neighborhood of $x \in X$ if and only if it contains an open ball centered in x . We then deduce that for study the topology \mathcal{T} , as well as the topological structure $x \mapsto \mathcal{F}(x)$, in a metric space it is sufficient to consider the family \mathcal{B} of all open balls of X , as well as the structure $x \mapsto \mathcal{B}(x)$ where $\mathcal{B}(x)$ is the family of open balls centered in x . We say that the family $\mathcal{B} \subseteq \mathcal{T}$ of the open balls is a basis for the topology and that the family $\mathcal{B}(x) \subseteq \mathcal{F}(x)$ is a basis for the filter of neighborhoods of x . More generally we have the following definition.

Definition 5.37 *Let X be a topological space. A subset $\mathcal{B} \subseteq \mathcal{T}$ is a basis for the topology \mathcal{T} if every element of \mathcal{T} is the union of elements of \mathcal{B} . Fixed $\bar{x} \in X$, a subset $\mathcal{B} \subseteq \mathcal{F}(\bar{x})$ is a basis for the filter of neighborhoods $\mathcal{F}(\bar{x})$ if every element of $\mathcal{F}(\bar{x})$ contains an element of \mathcal{B} as subset.*

Let X be a set. A family of subsets $\mathcal{B} \subseteq \mathcal{P}(X)$ is a basis for a topology in X if there exists a topology \mathcal{T} such that \mathcal{B} is a basis for \mathcal{T} . A structure¹⁶¹ $x \mapsto \mathcal{B}(x) \subseteq \mathcal{P}(X)$ is a basis for neighborhoods filters in X if there exists a topological structure $x \mapsto \mathcal{F}(x)$ such that $\mathcal{B}(x)$ is a basis for $\mathcal{F}(x)$ for all $x \in X$.

¹⁶¹Here and in the sequel, by “structure” we mean any function from X to $\mathcal{P}(\mathcal{P}(X))$.

The following Proposition may be proven in similar way as the previous propositions.

Proposition 5.38 *Let X be a nonempty set. A family of subsets $\mathcal{B} \subseteq \mathcal{P}(X)$ is a basis for a topology in X if and only if \mathcal{B} contains the empty set, \mathcal{B} covers X ¹⁶² and every finite intersection of elements of \mathcal{B} is a union of elements of \mathcal{B} .*

If \mathcal{B} is a basis for a topology in a set X , then there exists a unique topology \mathcal{T} for which \mathcal{B} is a basis and it is

$$\mathcal{T} = \left\{ A \subseteq X \mid A \text{ is a union of elements of } \mathcal{B} \right\}$$

A structure $x \mapsto \mathcal{B}(x)$ is a basis for neighborhoods filters in X if and only if, for every $x \in X$, $\mathcal{B}(x)$ is not empty, every element of $\mathcal{B}(x)$ contains x and the intersection of two elements of $\mathcal{B}(x)$ contains an element of $\mathcal{B}(x)$ as subset.

If $x \mapsto \mathcal{B}(x)$ is a basis for neighborhoods filters in X , then there exists a unique topological structure $x \mapsto \mathcal{F}(x)$ such that $\mathcal{B}(x)$ is a basis for $\mathcal{F}(x)$ for every $x \in X$, and it is

$$\mathcal{F}(x) = \left\{ U \subseteq X \mid \exists B \in \mathcal{B}(x) \text{ such that } B \subseteq U \right\}$$

The family $\mathcal{B} \subseteq \mathcal{P}(X)$ is a basis for a topology \mathcal{T} in X if and only if the structure

$$x \mapsto \mathcal{B}(x) = \left\{ B \in \mathcal{B} \mid x \in B \right\}$$

is a basis for the corresponding neighborhoods filters structure $x \mapsto \mathcal{F}(x)$.

Many of the usual definitions and results in the topological space theory can be given just using a basis for the topology (respectively, for the neighborhoods filters) instead of using the whole topology (respectively, the neighborhoods filters) as we are going to see in the next paragraph.

As already said, in a metric space, a basis for the topology is the family of open balls¹⁶³, and a basis for the neighborhoods filter at x is the family of the open balls centered in x . In the case of \mathbb{R}^n , with the Euclidean distance, this are the usual Euclidean open balls of \mathbb{R}^n . However, note that also the family of open cubes $Q =]a_1, b_1[\times \cdots \times]a_n, b_n[$ is a basis for the topology¹⁶⁴, as well as the family of open cylinders and so on.

For the particular case of \mathbb{R} , a basis for the topology is the family of open bounded intervals, and for the filter of neighborhoods in x , the family of open bounded intervals centered in x .

¹⁶²i.e. X is the union of elements of \mathcal{B} .

¹⁶³If A is open, then, for every $x \in A$, it contains an open ball centered at x , and so A is the union of such balls.

¹⁶⁴This is true because every open ball contains an open cube and every open cube contains an open ball.

5.7 Convergence and continuity in topological spaces

Looking back again to the metric space case, we have that a sequence $\{x_n\} \subseteq X$ converges to a point $x \in X$ if and only if one of the three following equivalent conditions is satisfied

- a) $d(x_n, x) \rightarrow 0$ as $n \rightarrow +\infty$;
- b) $\forall \varepsilon \exists \bar{n}$ such that $n \geq \bar{n} \implies d(x_n, \bar{x}) \leq \varepsilon$;
- c) \forall open ball B centered in $x \exists \bar{n}$ such that $n \geq \bar{n} \implies x_n \in B$.

Similarly, if X and Y are two metric spaces with metrics d_X and d_Y respectively, $f : X \rightarrow Y$ is a function and $\bar{x} \in X$ is a point, we have that f is continuous at \bar{x} if and only if one of the following three equivalent conditions is satisfied

- d) $d_Y(f(x_n), f(\bar{x})) \rightarrow 0 \forall$ sequence $\{x_n\}$ converging to \bar{x} in X ;
- e) $\forall \varepsilon > 0 \exists \delta > 0$ such that $d_X(x, \bar{x}) \leq \delta \implies d_Y(f(x), f(\bar{x})) \leq \varepsilon$;
- f) \forall open ball B in Y centered in $f(\bar{x}) \exists$ an open ball B' in X centered in \bar{x} such that $x \in B' \implies f(x) \in B$.

Hence, it is natural to give the following definition.

Definition 5.39 *Let X and Y be two topological spaces with topological structures $x \mapsto \mathcal{F}_X(x)$ and $y \mapsto \mathcal{F}_Y(y)$ respectively, let $\{x_n\}$ be a sequence in X , $f : X \rightarrow Y$ a function and $\bar{x} \in X$ a point.*

i) *The sequence $\{x_n\}$ converges to \bar{x} if*

$$\forall U \in \mathcal{F}_X(\bar{x}) \exists \bar{n} \text{ such that } n \geq \bar{n} \implies x_n \in U.$$

ii) *The function f is continuous at \bar{x} if*

$$\forall U \in \mathcal{F}_Y(f(\bar{x})) \exists V \in \mathcal{F}_X(\bar{x}) \text{ such that } x \in V \implies f(x) \in U.$$

iii) *The function f is said continuous if it is continuous at every point $x \in X$.*

It is not hard to see that the previous points i) and ii) can be equivalently stated taking U and V inside any basis for the filters of neighborhoods. For instance, for the continuity, if $\mathcal{B}_X(\bar{x})$ and $\mathcal{B}_Y(f(\bar{x}))$ are basis for the neighborhoods, we have that f is continuous at \bar{x} if and only if

$$iv) \quad \forall U \in \mathcal{B}_Y(f(\bar{x})) \exists V \in \mathcal{B}_X(\bar{x}) \text{ such that } x \in V \implies f(x) \in U.$$

Indeed we prove that ii) is equivalent to iv).

$$\begin{aligned} ii) \implies iv) \quad & U \in \mathcal{B}_Y(f(\bar{x})) \subseteq \mathcal{F}_Y(f(\bar{x})) \implies \exists V' \in \mathcal{F}_X(\bar{x}), f(V') \subseteq U \\ & \implies \exists V \in \mathcal{B}_X(\bar{x}), V \subseteq V', f(V) \subseteq U; \\ iv) \implies ii) \quad & U \in \mathcal{F}_Y(f(\bar{x})) \implies \exists U' \in \mathcal{B}_Y(f(\bar{x})), U' \subseteq U \\ & \implies \exists V \in \mathcal{B}_X(\bar{x}) \subseteq \mathcal{F}_X(\bar{x}), f(V) \subseteq U' \subseteq U. \end{aligned}$$

Proposition 5.40 *If X is a Hausdorff space and $\{x_n\}$ is a convergent sequence in X , then the limit is unique. This means that there exists a unique point \bar{x} such that $x_n \rightarrow \bar{x}$.*

Proof. By contradiction, let us suppose that there exist two different points \bar{x}, \tilde{x} which are limit of the sequence. By definition of Hausdorff space, there exist two neighborhoods $U \in \mathcal{F}(\bar{x}), V \in \mathcal{F}(\tilde{x})$ such that $U \cap V = \emptyset$. By definition of convergence, there exists \bar{n} such that $x_n \in U$ and $x_n \in V$ for all $n \geq \bar{n}$. This is a contradiction. \square

We now give an example of a non-separated space where a sequence has two different limits. Let us consider the interval $[-1, 1]$ with the following basis for a topological structure: if $x \in]-1, 1[$, $\mathcal{B}(x)$ is the usual family of open intervals centered in x ; otherwise $\mathcal{B}(-1) = \mathcal{B}(1)$ is the family of sets of the form $[-1, -1 + \varepsilon[\cup]1 - \varepsilon, 1]$ with $\varepsilon > 0$. The space is not separated because -1 and 1 do not have disjoint neighborhoods. The sequence $a_n = (-1)^n$ has two different limits: $\bar{x} = -1$ and $\tilde{x} = 1$. Note that such a topology on $[-1, 1]$ corresponds to a sort of “circular topology”. Indeed, if we take a circumference where, not surprising, neighborhoods are arches of circumference centered in the point, and if we “cut” the circumference in a point obtaining a segment, then such a segment has no topologically separated extremes.

We have already seen that in a metric space a subset is topologically closed if and only if it is sequentially closed. Unfortunately, this is not more true in a generic topological space. The topological closedness still implies the sequentially closedness (as it is easy to prove), but the contrary is false as the following example shows.

Example 5.41 Let us consider the set $X = \left\{ f : [0, 1] \rightarrow \mathbb{R} \mid f \text{ is a function} \right\}$ and on it we consider the following basis for a topological structure¹⁶⁵: for every $f \in X$

$$B \in \mathcal{B}(f) \iff \exists \varepsilon > 0, \emptyset \neq \Gamma \subset [0, 1] \text{ finite, such that } B = \left\{ g \in X \mid \max_{x \in \Gamma} |g(x) - f(x)| < \varepsilon \right\}$$

For convenience let us denote the generic element of the basis as $B_{\Gamma, \varepsilon}(f)$.

By this topological structure, we have that a sequence f_n converges to f in X if and only if it pointwise converges to f on $[0, 1]$. Indeed, if it converges in X , then taking $\varepsilon > 0$ and $x \in [0, 1]$ we have that for n sufficiently large, $f_n \in B_{\{x\}, \varepsilon}(f)$, which, by the arbitrariness of $\varepsilon > 0$ gives the convergence in x . On the other hand, if the sequence pointwise converges, then for every finite set $\Gamma \subseteq [0, 1]$ and for every $\varepsilon > 0$ we easily get that $f_n \in B_{\Gamma, \varepsilon}(f)$ for any sufficiently large n .

Now, let us consider the following subset of X :

$$C = \left\{ f \in X \mid f(x) \neq 0 \text{ for at most countable points } x \in [0, 1] \right\}.$$

C is not topologically closed. Indeed, the function $g \equiv 1$ does not belong to C , but for every $\Gamma \subseteq [0, 1]$ finite and for every $\varepsilon > 0$ the characteristic function of the set Γ ¹⁶⁶ belongs

¹⁶⁵It can be proved that it is so.

¹⁶⁶ $\chi_\Gamma(x) = 1$ if $x \in \Gamma$ and $\chi_\Gamma(x) = 0$ otherwise.

to $C \cap B_{\Gamma, \varepsilon}(g)$, and so g is adherent to C but it does not belong to C . Nevertheless, C is sequentially closed. Indeed if a sequence of elements of C , f_n , converges to $f \in X$ (and so pointwise converges on $[0, 1]$), then the limit function f cannot be different from zero outside the set

$$N = \bigcup_n \left\{ x \in [0, 1] \mid f_n(x) \neq 0 \right\}.$$

But N , being a countable union of at most countable sets, is itself at most countable, which means $f \in C$.

Remark 5.42 *In a metric space the sequential closedness is equivalent to the topological closedness essentially because in a metric space every point has a countable basis of neighborhoods¹⁶⁷, and this fact is of course compatible with testing the closedness along sequences only. A generic topological space has not a countable basis for the neighborhoods filters, as indeed it happens for the space in Example 5.41. Hence, in a topological space the informations brought by convergence sequences are too few.*

In the same way, the continuity of a function in a generic topological spaces cannot be testing along convergence sequences only, and the same thing happens for compactness, as we are going to see in the next paragraph.

Theorem 5.43 *Let (X, \mathcal{T}_X) and (Y, \mathcal{T}_Y) be two topological spaces. A function $f : X \rightarrow Y$ is continuous if and only if*

$$f^{-1}(A) \in \mathcal{T}_X \quad \forall A \in \mathcal{T}_Y,$$

where $f^{-1}(A) = \{x \in X \mid f(x) \in A\}$ is the anti-image of $A \subseteq Y$ via f . Roughly speaking, f is continuous if and only if the anti-images of the open subsets are open.

Similarly, f is continuous if and only if the anti-images of the closed subsets are closed.

Proof. By the definition of continuity it is easy to see that f is continuous at x if and only if $f^{-1}(U) \in \mathcal{F}_X(x)$ for all $U \in \mathcal{F}_Y(f(x))$. From this the thesis easily comes.

The last sentence comes from the equality

$$\mathcal{C}f^{-1}(A) = f^{-1}(\mathcal{C}A).$$

□

Remark 5.44 *The most trivial topologies in a set are $\mathcal{T}_1 = \{\emptyset, X\}$ and $\mathcal{T}_2 = \mathcal{P}(X)$. In the first case, for every point $x \in X$ the unique neighborhood is the whole space X , whereas, in the second case, any subset containing X is a neighborhood of x , in particular the singleton $\{x\}$. If X is endowed by the topology \mathcal{T}_1 and Y is a separated topological space, then a function $f : X \rightarrow Y$ is continuous if and only if it is constant. On the other hand, if X is endowed by the topology \mathcal{T}_2 and Y is any topological space, then every function $f : X \rightarrow Y$ is continuous. Of course the interesting cases are the intermediate ones, that is when the topology \mathcal{T} is strictly contained between \mathcal{T}_1 and \mathcal{T}_2 .*

¹⁶⁷The family of balls $B(x, 1/n)$.

To the continuity of some suitable functions it is also linked the so-called product topology. If (X_1, \mathcal{T}_1) and (X_2, \mathcal{T}_2) are two topological space, then we can naturally endow the cartesian product $X_1 \times X_2$ by the topology a basis of which is $\mathcal{B} = \{A_1 \times A_2 | A_1 \in \mathcal{T}_1, A_2 \in \mathcal{T}_2\}$. Moreover, if \mathcal{B}_1 and \mathcal{B}_2 are two basis for \mathcal{T}_1 and \mathcal{T}_2 respectively, then $\{B_1 \times B_2 | B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2\}$ is also a basis for the product topology. Note that, in general, \mathcal{B} is not a topology in $X_1 \times X_2$. For instance, in $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, when \mathbb{R} has the standard topology, the product topology is the usual Euclidean one and a basis of it is given by the cartesian product of open sets of the real line. However, a subset of \mathbb{R}^2 can be open without being the cartesian product of two subsets of \mathbb{R} .

If $X_1 \times X_2$ is endowed by the product topology, then the projections $\pi : X_1 \times X_2 \rightarrow X_1$ $(x_1, x_2) \mapsto x_1$ and $\pi_2 : X_1 \times X_2 \rightarrow X_2$ $(x_1, x_2) \mapsto x_2$ are continuous. In the general case of product set of the form $X = \prod_{i \in I} X_i$ with I any index set (possibly infinite), if X_i is a topological space for every $i \in I$, the product topology on it is defined as the smallest topology such that all the projections $\pi_i : X \rightarrow X_i$ are continuous¹⁶⁸.

If $C \subseteq X$ is a subset of a topological space (X, \mathcal{T}) , then C can be also regarded as a topological space, endowed with the induced topology $\mathcal{T}_C = \{A \cap C | A \in \mathcal{T}\}$, and topological structure $x \mapsto \mathcal{F}_C(x) = \{U \cap C | U \in \mathcal{F}(x)\}$. The induced topology is the smallest topology on C such that the injection $i : C \rightarrow X$ $x \mapsto x$ is continuous¹⁶⁹.

Definition 5.45 *A bijective function $f : X \rightarrow Y$ between two topological spaces is said to be a homeomorphism if f and the inverse f^{-1} are continuous. In such a case the two spaces are said to be homeomorphic.*

It is obvious that if X and Y are homeomorphic, then every possible property concerning topological aspects holds for X if and only if it holds for Y . Moreover, by Theorem 5.43, a continuous bijective function $f : X \rightarrow Y$ is a homeomorphism if and only if $f(A)$ is open in Y for every A open in X (or equivalently, $f(C)$ is closed for every C closed). Moreover, a bijective continuous function is not necessary a homeomorphism, that is its inverse may be non continuous. For instance, take $(X, \mathcal{T}) = (\mathbb{R}, \mathcal{P}(\mathbb{R}))$ and \mathbb{R} with the usual topology. Hence the identity map $i : X \rightarrow \mathbb{R}$ $x \mapsto x$ is bijective and continuous, but its inverse, which is still i , is not continuous because, for instance, $\{0\}$ ($= i^{-1}(\{0\})$) is open for X but not for \mathbb{R} .

5.8 Compactness in topological spaces and sequential compactness

As reported in Remark 5.42, testing topological properties along convergent sequences is not exhaustive in a generic topological space. This is certainly true also for the property

¹⁶⁸Smallest with respect to the inclusion. Such a topology must contain all the sets of the form $\pi_{\bar{i}}^{-1}(A) = A \times \prod_{I \in I, i \neq \bar{i}} X_i$ for all $\bar{i} \in I$ and $A \in \mathcal{T}_{\bar{i}}$. It can be proved that a smallest topology containing all these sets exists.

¹⁶⁹It must contain all the sets $i^{-1}(A) = A \cap C$ such that $A \in \mathcal{T}$, and in such a case, this family of subsets of C exactly forms a topology.

of compactness. In a topological space the sequential compactness¹⁷⁰ is too poor as property for being significative. Of course we also cannot speak of total boundedness and of completeness, because if the space is not a metric space such concepts are meaningless. Hence, inspired by Theorem 5.20, we give the following definition.

Definition 5.46 *Let X be a topological space. A subset $C \subseteq X$ is said to be compact if every family of open subsets which covers C has a finite subfamily still covering C .*

The space X is said to be a compact space if X is compact as subset of itself.

Proposition 5.47 *Let X be a topological space and $C \subseteq X$ a subset. i) If X is compact and C closed, then C is compact. ii) If X is separated and C compact, then C is closed.*

Proof. i) Let \mathcal{R} be an open covering of C . Since C is closed, $\mathcal{R} \cup \{X \setminus C\}$ is an open covering of X . Since X is compact, there exists a finite subcovering of X , and hence a finite subcovering of C .

ii) Let x be adherent to C and by contradiction suppose that $x \notin C$. Hence, because X is separated, for every $y \in C$, being $y \neq x$, it exist a neighborhood U_y of y and a neighborhood V_y of x such that $U_y \cap V_y = \emptyset$. For every y let us take $U'_y \subseteq U_y$ open neighborhood of y . Hence the family of the open subsets U'_y covers C and, being C compact, there exists a finite number of points $y_1, \dots, y_N \in C$ such that the union of U'_{y_i} $i = 1, \dots, N$ covers C . Hence we put

$$V = \bigcap_{i=1}^N V_{y_i} \in \mathcal{F}(x),$$

we have

$$C \cap V \subseteq \bigcup_{I=1}^N U_{y_i} \cap V = \emptyset,$$

which is a contradiction, because $C \cap V$ must be non-empty, being x adherent to C . \square

Proposition 5.48 *Let X and Y be two topological spaces and $f : X \rightarrow Y$ be a continuous function. Then, if $K \subseteq X$ is compact, $f(K) = \{y \in Y | \exists x \in K, f(x) = y\}$ is also compact in Y .*

Proof. Let $\bigcup_{i \in I} B_i \supseteq f(K)$ be an open covering. Hence, by continuity and definition of anti-image, $\bigcup_{i \in I} f^{-1}(B_i)$ is an open covering of K . The conclusion then easily follows. \square

Remark 5.49 *Proposition 5.48 says that the continuous functions map compact sets into compact sets. Putting together Propositions 5.47 and 5.48, we get that if X is compact*

¹⁷⁰That is the property that from every sequence we can extract a convergent subsequence.

and Y is Hausdorff, then every continuous function is “closed”, that is it maps closed subsets into closed subsets.

In the special case that f is also bijective, we immediately get that f is a homeomorphism, since the inverse function turns out to be continuous. Hence, bijective continuous functions on a compact space are homeomorphisms.

Moreover, if $f : X \rightarrow Y$ is continuous and bijective, Y is Hausdorff and X is possibly not compact but it is such that every point $x \in X$ has a compact neighborhood U such that $f(U)$ is a (necessarily compact) neighborhood of $f(x) \in Y$, then f is a homeomorphism¹⁷¹.

Note that not every topological space is such that every point has a compact neighborhood. A favorable case where this happens is when X is locally compact, that is every point has a neighborhood basis given by compact sets. A first important example of locally compact space is \mathbb{R}^n (a basis of compact neighborhoods is given by the closed balls centered at the point). However also note that the fact that X is locally compact is not sufficient for f being a homeomorphism whenever it is bijective and continuous. The request that a compact neighborhood is sent onto a neighborhood is essential. For instance the already given example where X is \mathbb{R} endowed with the trivial topology $\mathcal{T} = \mathcal{P}(\mathbb{R})$, Y is \mathbb{R} with the usual topology and $f = i$ the identity map, is a counterexample¹⁷².

Proposition 5.50 *Let $f : X \rightarrow \mathbb{R}$ be continuous. Then, for every compact subset $K \subseteq X$, f reaches its maximum and its minimum on K .*

Proof. It is an immediate consequence of Proposition 5.48 which assures that $f(K)$ is compact in \mathbb{R} and so closed and bounded. \square

It is evident that in Theorem 5.20 we can replace the family of open balls with a family of generic open subset, and this because the open balls form a basis for the topology. Hence, in a metric space, the sequential compactness of Definition 5.14 is equivalent to the topological compactness in the sense of Definition 5.46.

One may ask for the possible relation between sequential compactness and topological compactness. Unfortunately the answer is not satisfactory: for a generic topological space there is absolutely no relation between the two concepts: a set can be topologically compact without being sequentially compact and vice-versa sequentially compact without being topologically compact. The next two examples show these facts.

Example 5.51 (Topological compactness without sequential compactness.) Let us consider the set

¹⁷¹Note that f restricted to U is a homeomorphism between U and $f(U)$ with the topologies induced by X and Y respectively; it is sufficient to prove that, for every $x \in X$, f maps neighborhoods of x onto neighborhoods of $f(x)$; to this end it is sufficient to restrict ourselves to neighborhoods contained in U (which exist because U is a neighborhood of x): if $V \in \mathcal{F}_X(x)$ and $V \subseteq U$ then $V \in \mathcal{F}_U(x)$ and so $f(V) \in \mathcal{F}_{f(U)}(f(x))$ which implies the existence of $W \in \mathcal{F}_Y(f(x))$ such that $W \cap f(U) = f(V)$ and we conclude since $f(U) \in \mathcal{F}_Y(f(x))$ and so $f(V) \in \mathcal{F}_Y(f(x))$.

¹⁷² X is locally compact: for every point x , $\{x\}$ is a compact neighborhood. But $i(\{x\}) = \{x\}$ which is not a neighborhood of x in Y .

$$X = \left\{ f : [0, 1] \rightarrow [0, 1] \mid f \text{ function} \right\},$$

endowed with the same topology as the space in the Example 5.41. The space X is topologically compact. We do not give the proof of this fact. It comes from the following reasoning: X is homeomorphic to the product space (endowed with product topology):

$$[0, 1]^{[0, 1]} = \prod_{x \in [0, 1]} [0, 1],$$

that is the space of the continuum strings of points of $[0, 1]$, in other words, the space of the graphs of all functions $f : [0, 1] \rightarrow [0, 1]$, endowed with the product topology, i.e. the minimal one that makes the projections (evaluations) $\pi_x : X \rightarrow [0, 1]$, $\pi_x(f) = f(x)$, continuous for every $x \in X$. This is a product space of compact spaces ($[0, 1]$), and a famous result by Tychonoff assures that the product spaces endowed by the product topology is compact if and only if the spaces are compact.

Let us prove that it is instead not sequentially compact. Let us consider the following sequences in X ($a_n(x)$ is the n -th digit of the decimal expansion of $x \in [0, 1]$):

$$f_n(x) = \begin{cases} \frac{a_n(x)}{9} & \text{if } x \text{ has a unique decimal expansion,} \\ 0 & \text{otherwise.} \end{cases}$$

The sequence f_n cannot have any convergent subsequence in X . Indeed, we know that the convergence of a sequence in X is equivalent to the pointwise convergence. Let $\{n_k\}$ be any subsequence of indices, and take a point $\bar{x} \in [0, 1]$ which has a unique decimal expansion satisfying

$$a_{n_k}(\bar{x}) = 0 \text{ if } k \text{ is even, } a_{n_k}(\bar{x}) = 1 \text{ otherwise.}$$

Hence, the subsequence f_{n_k} does not converge in \bar{x} and so it does not converge in X .

Example 5.52 (Sequential compactness without topological compactness.) Let us consider the same space as in Example 5.51, and also consider the subset

$$C = \left\{ f \in X \mid f(x) \neq 0 \text{ for at most countable } x \in [0, 1] \right\}.$$

Since the space X is an Hausdorff space¹⁷³, then C is not topological compact because, as we already know from Example 5.41 it is not closed. Nevertheless it is sequentially compact. Indeed, if f_n is a sequence in C , we define the set

$$N = \bigcup_{n \in \mathbb{N}} \{x \in [0, 1] \mid f_n(x) \neq 0\},$$

¹⁷³If $f \neq g \in X$ then there exists $x_0 \in [0, 1]$ such that $f(x_0) \neq g(x_0)$. Hence, taking $\Gamma = \{x_0\}$ and $\varepsilon < |f(x_0) - g(x_0)|/2$, we have $B_{\Gamma, \varepsilon}(f) \cap B_{\Gamma, \varepsilon}(g) = \emptyset$.

which is at most countable. Outside of N , all the functions f_n are null and hence they certainly pointwise converge to the null function in $[0, 1] \setminus N$. Now, by the compactness of $[0, 1]$, from every sequence of real numbers $\{f_n(x)\}_n$ with $x \in N$, we can extract a convergent subsequence. Recalling that N is at most countable, via a standard “diagonal procedure” we can extract a subsequence f_{n_k} which pointwise converges on the countable set N ¹⁷⁴. Hence, the subsequence pointwise converges on $[0, 1]$, that is the subsequence converges in X , and it is obvious that the limit function belongs to C because it is null outside N .

Definition 5.53 *A non-empty set A is said a direct (or filtering) set if there is an order relation “ \leq ” on it (not necessarily total) such that given any two elements $\alpha, \beta \in A$ there exists an element $\gamma \in A$ such that both $\alpha \leq \gamma$ and $\beta \leq \gamma$.*

If A is a direct set and X is a topological space, a function $r : A \rightarrow X$ is said to be a net on X . If $\bar{x} \in X$, we say that the net r converges to \bar{x} if for all $U \in \mathcal{F}(x)$ there exists $\bar{\alpha} \in A$ such that $r(\alpha) \in U$ for every $\alpha \geq \bar{\alpha}$.

If B is another direct set, and if $\varphi : B \rightarrow A$ is a function such that it preserves the order¹⁷⁵, then the composed function $r \circ \varphi : B \rightarrow X$ is said to be a subnet of the net $r : A \rightarrow X$.

A first natural example of direct set is the set of natural numbers \mathbb{N} . However, \mathbb{N} is a very special case of direct set because it is totally ordered. An interesting example of non-totally ordered direct set is the filter of neighborhoods of a point x in a topological space X , with the order given by the inverse inclusion. It is clear that the concept of net is a generalization of the concept of sequence, as well as the concept of subnet is a generalization of the concept of subsequence¹⁷⁶. However, the convergence of a net is a very weaker properties than the convergence of a sequence. It is also clear that nets are much more than sequences and hence, testing topological properties (such as continuity, closedness, compactness and so on) may be sufficient and exhaustive. This is indeed what happens. For instance, it can be proven that a subset C of a topological space X is compact if and only if every net inside C has a convergent subnet to a point of C . In particular note that, by this result, every sequence in a compact set has a convergent subnet, which however may be not more a sequence (i.e. defined on \mathbb{N}).

Facing a general topological space X , it may be then interesting to know whether, in a compact set, from any sequence it is possible to extract a convergent subsequence, even if the space is not a metric space (or a metrizable space¹⁷⁷). This is indeed a very important issue in the theory of the differential equation where one wants to work with

¹⁷⁴Let us enumerate N : $\{x_1, x_2, x_3, \dots, x_n, \dots\}$. From the sequence of real numbers $\{f_n(x_1)\}_n \subseteq [0, 1]$ we extract a converging subsequence $\{f_{n_j^1}(x_1)\}_j$. From the sequence $\{f_{n_j^1}(x_2)\}_j \subseteq [0, 1]$ we extract a convergent subsequence $\{f_{n_j^{1,2}}(x_2)\}_j$. We proceed in this way step by step. The subsequence of functions $f_{n_j^{1,2,\dots,n}}$ (diagonal procedure) is a pointwise converging subsequence in N .

¹⁷⁵ $b_1 \leq b_2$ in B implies $\varphi(b_1) \leq \varphi(b_2)$ in A .

¹⁷⁶A sequence is a particular net defined on the direct set \mathbb{N} ; a subsequence of a sequence is a particular subnet of the sequence which is still defined on the direct set \mathbb{N} .

¹⁷⁷See one of the next paragraphs for the definition.

sequences of approximating solutions. In the theory of reflexive Banach spaces (which is a very important framework in the modern theory of partial differential equations), when the space is endowed with the so-called weak topology, this is fortunately true: from every bounded sequence it is possible to extract a weak-convergent subsequence.

5.9 Fast excursion on density, separability, Ascoli-Arzelà, Baire and Stone-Weierstrass

As for the case of metric spaces, if X is a topological space and $A \subseteq X$ is a subset, we say that A is dense in X if the closure of A , \overline{A} , coincides with X , $\overline{A} = X$. This means that every point of X is adherent to A , that is, for every $x \in X$ and for every $U \in \mathcal{F}(x)$ there exists $a \in U \cap A$. A very interesting case is when a topological space X , whose power as set may be more than countable, contains a dense countable subset. In this case we say that X is separable¹⁷⁸. A first known example of separable space is \mathbb{R}^n with \mathbb{Q}^n as dense countable subset. Another interesting example of separable space is the space $C^0([a, b])$ of the continuous real functions defined on the compact interval $[a, b]$, with the uniform convergence¹⁷⁹. The space

$$X = \left\{ f : [a, b] \rightarrow \mathbb{R} \mid \exists M > 0 \text{ such that } |f(x)| \leq M \forall x \in [a, b] \right\}$$

of bounded real functions on the compact interval $[a, b]$, endowed with the uniform topology, is instead not separable. Indeed, for every $x \in [a, b]$ and $r > 0$, consider the characteristic function $\chi_{x,r}$ of the set $[a, b] \cap]x - r, x + r[$ and define the non-empty open set

$$\mathcal{O}_{x,r} = \left\{ f \in X \mid \|f - \chi_{x,r}\|_\infty < \frac{1}{2} \right\},$$

which is a particular neighborhood of $\chi_{x,r}$, and note that such neighborhoods are pairwise disjoint: $(x, r) \neq (y, s) \implies \mathcal{O}_{x,r} \cap \mathcal{O}_{y,s} = \emptyset$. Now, by absurd, let us suppose that X is separable and let f_n be a dense countable family of elements of X . Hence, for every $\chi_{x,r}$ there exists $f_{n(x,r)} \in \mathcal{O}_{x,r}$. Hence, we have an injective¹⁸⁰ function from the set of couples (x, r) to the set of natural numbers \mathbb{N} . This is a contradiction because the set of those couples is uncountable.

The separability of a topological space is somehow linked to the property of having a countable basis for the topology. In particular, the following result holds.

Theorem 5.54 *If a topological space X has a countable basis for the topology, that is a basis formed by a countable quantity of subsets, then it is separable.*

¹⁷⁸Note the difference between the definitions of “separated” and “separable”: they are different and moreover there is no a-priori link between them.

¹⁷⁹This fact will be (partially) explained when the Stone-Weierstrass theorem will be stated.

¹⁸⁰By the pairwise-disjointness of the sets $\mathcal{O}_{x,r}$.

Proof. Let $\mathcal{B} = \{B_1, B_2, B_3, \dots\}$ be the countable basis, and consider the set

$$M = \{x_1, x_2, x_3, \dots\} \subseteq X,$$

such that $x_n \in B_n$ for all $n \in \mathbb{N} \setminus \{0\}$. By absurd, let us suppose that M is not dense in X . Then $X \setminus \overline{M}$ is open and not empty, hence it must be the union of elements of \mathcal{B} and hence it must contain some point of M . Contradiction. \square

In general, the opposite of the statement in Theorem 5.54 is not true, that is there exist separable topological spaces which do not have a countable basis. As usual, for the case of metric spaces, the situation is more favorable, as the following Theorem asserts.

Theorem 5.55 *A metric space X is separable if and only if it has a countable basis.*

Proof. The sufficiency is stated in Theorem 5.54. Let $M = \{x_1, x_2, x_3, \dots\} \subseteq X$ be the countable dense set. Let us consider the family of balls

$$\mathcal{B} = \left\{ B \left(x_n, \frac{1}{m} \right) \mid n, m \in \mathbb{N} \setminus \{0\} \right\} \subseteq \mathcal{P}(X).$$

The family \mathcal{B} is then countable and it is a basis for the topology. Indeed, for every open subsets $A \subseteq X$ and for every $x \in A$ we can find $n, m \in \mathbb{N} \setminus \{0\}$ such that

$$x_m \in B \left(x, \frac{1}{3n} \right) \subseteq B \left(x, \frac{1}{n} \right) \subseteq A,$$

from which the conclusion, because:

$$x \in B \left(x_m, \frac{1}{2n} \right) \subseteq A,$$

and so A is the union of elements of \mathcal{B} .

\square .

Definition 5.56 *A topological space X is said to satisfy the first axiom of countability if every point $x \in X$ has a countable neighborhoods basis. It is said to satisfy the second axiom of countability if it has a countable basis for the topology.*

It is evident that every metric space satisfies the first axiom of countability¹⁸¹. But in general it does not satisfy the second axiom: this happens if and only if the metric space is separable¹⁸². About the power of a topological space, it can be proved that every separable Hausdorff space satisfying the first axiom of countability (in particular, a separable metric space) has at most the power of the continuum, $m(\mathbb{R})$. More generally, a separable Hausdorff space has at most the power of $\mathcal{P}(\mathbb{R})$.

It is also evident that if a space X satisfies the first (respectively, second) axiom of countability, then every subset $C \subseteq X$, when regarded as a topological space with the induced topology, satisfies the first (respectively, second) axiom of countability too.

¹⁸¹For every point x the countable family of open balls $B(x, 1/n)$ is a neighborhoods basis.

¹⁸²And we have just seen that there exists non separable metric spaces, as the space of bounded real functions on $[a, b]$.

Proposition 5.57 *If X is a topological space satisfying the second axiom of countability, then from every open covering it is possible to extract an at most countable subcovering.*

Proof. Let \mathcal{O} be an open covering of X , and \mathcal{B} a countable basis for the topology. For every $x \in X$ we choose $O_x \in \mathcal{O}$ such that $x \in O_x$. Since \mathcal{B} is a basis for the topology, we then may choose $B_x \in \mathcal{B}$ such that $x \in B_x \subseteq O_x$. Being the topology \mathcal{B} countable, the subfamily $\{\mathcal{B}_x\}_x$ is also countable¹⁸³. Let us enumerate such a subfamily $\{B_1, B_2, B_3, \dots, B_n, \dots\}$, and, for every n take $O_n \in \mathcal{O}$ such that $B_n \subseteq O_n$. Hence, the family $\{O_n\}_n$ is an at most countable subcovering. \square

For the special case of the real line \mathbb{R} , we have the following more precise result.

Theorem 5.58 *Every open set $A \subseteq \mathbb{R}$ is the union of an at most countable family of pairwise disjoint open intervals.*

Proof. Since \mathbb{R} is a separable metric space, then it satisfies the second axiom of countability. If we prove that every open set is the union of a family of pairwise disjoint open intervals, then we are done by Proposition 5.57.

For every $x \in A$, the family \mathcal{I}_x of open intervals containing x and contained in A is not empty, because A is open. We then define

$$I_x = \bigcup_{I \in \mathcal{I}_x} I.$$

We recall the following easily proven fact: every union of intervals all containing the same point x is still an interval. From this, it is not difficult to see that, for every $x \in A$, I_x is an open interval and that, for every $x \neq y$ points of A , either $I_x = I_y$ or $I_x \cap I_y = \emptyset$. Hence, we can easily conclude. \square

We now consider the relation between continuous function on dense subsets.

Proposition 5.59 *Let X and Y be two topological spaces, with Y separated. If $f, g : X \rightarrow Y$ are two continuous function which coincide on a dense subset $A \subseteq X$, then they coincide on the whole space X .*

Proof. By absurd, let us suppose that there exists $\bar{x} \in X \setminus A$ such that $f(\bar{x}) \neq g(\bar{x})$. By separation of Y , take two neighborhoods $U \in \mathcal{F}_Y(f(\bar{x}))$, $V \in \mathcal{F}_Y(g(\bar{x}))$ such that $U \cap V = \emptyset$. By continuity, $f^{-1}(U), g^{-1}(V) \in \mathcal{F}_X(\bar{x})$, and so $f^{-1}(U) \cap g^{-1}(V) \in \mathcal{F}(\bar{x})$. By density, there exists $x \in A$, such that $x \in f^{-1}(U) \cap g^{-1}(V)$ and so $f(x) = g(x) \in U \cap V$. Contradiction. \square

For the special case of metric spaces, Proposition 5.59 has a form of inversion, that is, under suitable hypotheses, a continuous function defined on a dense set can be uniquely extended to the whole space.

¹⁸³Note that the points x may be not countable, and so there are different points $x \neq y$ such that $B_x = B_y$.

Definition 5.60 If $A \subseteq X$ is a subset of a metric space X , a function $f : A \rightarrow Y$, where Y is another metric space, is said to be uniformly continuous on A if

$$\forall \varepsilon > 0 \exists \delta > 0 \text{ such that } x, y \in A, d_X(x, y) \leq \delta \implies d_Y(f(x), f(y)) \leq \varepsilon.$$

Proposition 5.61 If $f : A \rightarrow Y$ is uniformly continuous, then it is also continuous. If $\{x_n\}$ is a Cauchy sequence in A , then $\{f(x_n)\}$ is a Cauchy sequence in Y .

If Y is Hausdorff and complete and f uniformly continuous, then there exists a unique uniformly continuous extension of f to \bar{A} . That is there exists a unique uniformly continuous function $\bar{f} : \bar{A} \rightarrow Y$ such that $\bar{f}(x) = f(x)$ for all $x \in A$.

If $K \subseteq X$ is compact and $f : K \rightarrow Y$ is continuous, then f is uniformly continuous.

Proof. We only give hints for the extension and for the uniform continuity on the compact set.

Extension. If $x \in \bar{A}$, then there exists a sequence of points $x_n \in A$ which converges to x . Hence it is a Cauchy sequence in X , and so is $\{f(x_n)\}$ in Y . By completeness, the latter sequence converges to y in Y . Define $\bar{f}(x) = y$ and show that it is a good definition (independent from the sequence x_n converging to x) and that it has all the requested properties.

Uniform continuity on a compact. By absurd, if f is not uniformly continuous, there exist $\varepsilon > 0$ and two sequences $\{x_n\}, \{y_n\}$ in K such that $d_X(x_n, y_n) \leq 1/n$ and $d_Y(x_n, y_n) \geq \varepsilon$. Hence, you may extract convergent subsequences and obtain a contradiction. \square

The importance of the previous result is mainly in the case when $Y = \mathbb{R}$ (or \mathbb{R}^n). Linked to the concept of uniform continuity is also the concept of equicontinuity, which is the essential ingredient of the following theorem whose proof, which is based on a diagonal procedure, we do not report.

Theorem 5.62 (Ascoli-Arzelà) Let $[a, b]$ be a compact interval, and $H \subseteq C^0([a, b])$ be a bounded and equicontinuous subset. That is

- i) (boundedness) $\exists M > 0$ such that $\|f\|_\infty \leq M$ for all $f \in H$;
- ii) (equicontinuity) $\forall \varepsilon > 0$ there exists $\delta > 0$ such that

$$x, y \in [a, b], |x - y| \leq \delta \implies |f(x) - f(y)| \leq \varepsilon \forall f \in H.$$

Then, the closure of H is compact (of course, sequentially compact, because $C^0([a, b])$ is a metric space).

Still in the special case of metric spaces, we have the following important result, known as the Baire (category) Lemma.

Theorem 5.63 Let X be a complete metric space. Then X cannot be the countable union of “nowhere dense” subsets¹⁸⁴.

¹⁸⁴A subset is “nowhere dense” if it is not dense in any open ball.

Proof. By absurd, let us suppose $X = \bigcup_{n \in \mathbb{N}} A_n$ with A_n nowhere dense. Take a closed ball B of radius 1. Since A_0 is not dense in B , then there exists a closed ball $B_0 \subseteq B$ of radius less than $1/2$ such that $A_0 \cap B_0 = \emptyset$. Since A_1 is not dense in B_0 , then there exists a closed ball $B_1 \subseteq B_0$ of radius less than $1/3$ such that $A_1 \cap B_1 = \emptyset$. In this way we construct a sequence of nested closed balls, with radii converging to zero, which, by completeness, must all contain a point $x \in X (= \bigcup_n A_n)$. By construction, such an element x cannot belong to any set A_n . Contradiction. \square

Remark 5.64 *There are several equivalent versions of Baire's Lemma. Some of them are: i) in a complete metric space the union of any sequence of closed subsets with empty interior is still a subset with empty interior; ii) if a complete metric space is the union of a sequence of closed subsets, then at least one of those subsets must have nonempty interior; iii) in a complete metric space any sequence of dense open subsets has the intersection which is still a dense subset.*

An immediate corollary of the Baire's Lemma is the following: a complete metric space without isolated points is uncountable (all the sets $\{x\}$ are nowhere dense).

Here we give an application of the Baire's Lemma to the proof of existence of continuous functions which are nowhere derivable. In the next paragraph we will give an explicitly example of such a function.

Proposition 5.65 *Let $[a, b] \subseteq \mathbb{R}$ be a compact interval. There exists a function $f \in C^0([a, b])$ which, for every $x \in [a, b]$, is not derivable in x .*

Proof. The idea of the proof is to prove that the subset C of $C^0([a, b])$ of continuous functions which are derivable at least in one point is contained in the union of a countable quantities of closed subsets with empty interior. Hence, by the completeness of $C^0([a, b])$, and by Baire's Lemma (see remark 5.64), we conclude that C cannot be the whole $C^0([a, b])$.

For every $n \in \mathbb{N} \setminus \{0\}$, we define the set

$$E_n = \left\{ f \in C^0([a, b]) \mid \exists x_0 \in [a, b] \text{ with } \left| \frac{f(x) - f(x_0)}{x - x_0} \right| \leq n \ \forall x \in [a, b], 0 < |x - x_0| < \frac{1}{n} \right\}.$$

We prove that E_n is closed. Let $\{f_k\} \subseteq E_n$ be a sequence of continuous functions uniformly converging on $[a, b]$ to a continuous function f . Moreover, for every k let $x_k \in [a, b]$ be a point for f_k as in the definition of E_n . By possibly extracting a subsequence (which we continue to denote by x_k) we have the convergence of x_k to a point $\bar{x} \in [a, b]$. By the uniform convergence, we have

$$f_k(x_k) \rightarrow f(\bar{x}) \text{ as } k \rightarrow +\infty.$$

Hence, taken $x \in [a, b]$ with $0 < |x - \bar{x}| < 1/n$, for every sufficiently large k (in such a way that $0 < |x - x_k| < 1/n$) we have

$$n \geq \left| \frac{f_k(x) - f_k(x_k)}{x - x_k} \right| \rightarrow \left| \frac{f(x) - f(\bar{x})}{x - \bar{x}} \right|,$$

from which $f \in E_n$, and E_n is closed.

Let us prove that E_n has empty interior¹⁸⁵. We take $\varepsilon > 0$ and $f \in E_n$ and we show that there exists a continuous function $g \notin E_n$ such that $\|f - g\|_\infty \leq 4\varepsilon$, which implies that E_n contains no balls and so its interior is empty. Let us construct such a continuous function g . Since f is uniformly continuous in $[a, b]$ (being continuous on a compact), there exists $\delta' > 0$ such that

$$x, y \in [a, b], |x - y| \leq \delta' \implies |f(x) - f(y)| \leq \varepsilon.$$

Let us take $\delta < \min\{\delta', \varepsilon/n\}$, and consider the partition of $[a, b]$ in subintervals of length δ :

$$x_0 = a, x_1 = a + \delta, x_2 = a + 2\delta, \dots, x_{m-1} = a + (m-1)\delta, x_m = b,$$

where m is the maximum integer not larger than $(b-a)/\delta$. On $[x_0, x_1]$ define g (better: its graph) as the segment in \mathbb{R}^2 between the points $(x_0, f(x_0) - \varepsilon)$ and $(x_1, f(x_1) + \varepsilon)$. Similarly, on $[x_1, x_2]$ as the segment between $(x_1, f(x_1) + \varepsilon)$ and $(x_2, f(x_2) - \varepsilon)$, and proceed in this way¹⁸⁶. To simplify notations let us suppose that $a = 0$ and $f(0) = \varepsilon$, and we prove that $|g(x) - f(x)| \leq 4\varepsilon$ for all $0 \leq x \leq \delta$; the extension to the general case and to the other subintervals is then easy. By our assumption for $x \in [0, \delta]$,

$$g(x) = \frac{f(\delta) + \varepsilon}{\delta}x, \text{ and } 0 \leq f(\delta) \leq 2\varepsilon$$

and so, for $x \in [0, \delta]$

$$\begin{aligned} |g(x) - f(x)| &= \left| \frac{f(\delta) + \varepsilon}{\delta}x - f(x) \right| \leq \frac{1}{\delta} |xf(\delta) - \delta f(x) + \varepsilon x| \\ &\leq \frac{1}{\delta} (|xf(\delta) - \delta f(\delta)| + |\delta f(\delta) - \delta f(x)| + \varepsilon x) = \frac{1}{\delta} (|f(\delta)||x - \delta| + \delta|f(x) - f(\delta)| + \varepsilon x) \\ &\leq \frac{1}{\delta} (2\varepsilon\delta + \delta\varepsilon + \varepsilon\delta) = 4\varepsilon. \end{aligned}$$

Now, with the same assumption, we prove that $|g'(x)| > n$ for all $x \in [0, \delta]$, which will show that $g \notin E_n$:

$$|g'(x)| = \frac{f(\delta) + \varepsilon}{\delta} \geq \frac{\varepsilon}{\delta} > n.$$

Hence, by the Baire's Lemma, the subset

$$E = \bigcup_{n=1}^{+\infty} E_n \subseteq C^0([a, b]),$$

¹⁸⁵Often, a set with empty interior is also called “meager” or “first category set”.

¹⁸⁶ g is the piecewise affine function with nodes $(x_i, f(x_i) - \varepsilon)$ if i is even, and $(x_i, f(x_i) + \varepsilon)$ otherwise.

has empty interior and so it is strictly contained in $C^0([a, b])$. Now if f is a continuous function derivable¹⁸⁷ in $x_0 \in [a, b]$, then $f \in E_n$ for some n . Indeed, if not we have, for every n , the existence of a point $x_n \in [a, b]$ with $0 < |x_n - x_0| \leq 1/n$ such that

$$\left| \frac{f(x_n) - f(x_0)}{x_n - x_0} \right| > n,$$

but then, because f is derivable in x_0 , we have

$$|f'(x_0)| = \lim_{n \rightarrow +\infty} \left| \frac{f(x_n) - f(x_0)}{x_n - x_0} \right| = +\infty$$

which is a contradiction. Hence, the set of continuous function which are derivable at least in one point is contained in E and so it is not the whole space $C^0([a, b])$. The continuous functions which do not belong to E are nowhere derivable. \square

Remark 5.66 *The proof of Proposition 5.65 not only shows the existence of continuous functions which are nowhere derivable, but also shows that such functions are “widely distributed”, dense, in $C^0([a, b])$. That happens because their complementary set is meager, it has empty interior.*

However, as the Stone-Weierstrass Theorem shows, the derivable functions are also dense in $C^0([a, b])$, but it can be proved that they are “much less” than the nowhere derivable functions, in the same way as the rational numbers are dense in the real line but they are “much less” than the irrational ones.

Definition 5.67 *Let $A \subseteq C^0([a, b])$ with $[a, b]$ compact. We say that A is an algebra if it is a vectorial subspace¹⁸⁸ of $C^0([a, b])$ and moreover if it is closed with respect to the multiplication of functions*

$$f, g \in A \implies fg \in A.$$

We say that A separates points if for every $x, y \in [a, b]$ there exists $f \in A$ such that $f(x) \neq f(y)$.

Theorem 5.68 *(Stone-Weierstrass Theorem). Let $[a, b]$ be a compact interval and $A \subseteq C^0([a, b])$ an algebra which separates points and contains all the constant functions. Then A is dense in $C^0([a, b])$.*

Corollary 5.69 *Let $[a, b]$ be a compact interval. Then the set of polynomials is dense in $C^0([a, b])$: every continuous function can be uniformly approximated by a polynomial.*

Proof. It is evident that the set of polynomials on the variable $x \in [a, b]$ is an algebra (it is certainly a vectorial subspace and the product of polynomials is still a polynomial), it obviously contains the constant functions (polynomials of degree zero) and separates points (just take the polynomial $y(x) = x$). \square

¹⁸⁷If $x_0 = a$ or $x_0 = b$ we mean the existence of the right and left derivative respectively.

¹⁸⁸With respect to the sum of functions $f + g$ and to the multiplication of a function with a scalar cf .

Remark 5.70 1) Two immediate consequences of Corollary 5.69: i) the derivable functions are dense in $C^0([a, b])$, ii) $C^0([a, b])$ is separable, because the polynomials with rational coefficients are obviously dense in the set of polynomials with real coefficients, and the polynomials with rational coefficients are countable.

2) We already know a result of approximation of functions with polynomials: the Taylor Theorem 2.53, which says that a C^∞ function whose derivatives satisfy a suitable property is uniformly approximable by a sequence of polynomials: the Taylor series. Which is the difference between Corollary 5.69 and Theorem 2.53? It is in the fact that the partial summation of the Taylor series is a very particular sequence of polynomials: every element of the sequence differs from the preceding one only for the term of maximum degree:

$$p_n(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + a_nx^n, \quad p_{n-1}(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1}. \quad (5.13)$$

This is the reason why only a subclass of continuous function can be represented by a power series, the class that we have called the analytical functions.

The uniform approximation stated by corollary 5.69 is instead given by sequences of polynomials which do not necessarily satisfy (5.13). Indeed, any term of the sequence may be very different from the preceding, in the coefficients and in the degree:

$$p_n(x) = a_0 + a_1x + \cdots + a_{r_{n-1}}x^{r_{n-1}} + a_{r_n}x^{r_n}, \quad p_{n-1}(x) = b_0 + b_1x + \cdots + b_{s_{n-1}}x^{s_{n-1}}.$$

This is why in that case we can approximate every continuous functions.

Obviously, the approximation via power series is much more “powerful”.

The Stone-Weierstrass Theorem also holds for continuous complex valued functions $f : [a, b] \rightarrow \mathbb{C}$, provided that the algebra A is also closed by complex conjugation. A typical example of such an algebra is the one generated by the functions of the form $t \mapsto e^{ikt}$ with $k \in \mathbb{Z}$ over the interval $[0, 2\pi]$ whose elements have the form

$$t \mapsto \sum_{k=-m}^m z_k e^{ikt}, \quad (5.14)$$

with $z_k \in \mathbb{C}$. Hence, every continuous function $f : [0, 2\pi] \rightarrow \mathbb{C}$ such that $f(0) = f(2\pi)$ can be uniformly approximated by complex trigonometric polynomials of the form (5.14). Passing to the real parts, we get the uniform approximation of real continuous functions periodic of period 2π by real trigonometric polynomials of the form

$$\frac{a_0}{2} + \sum_{k=1}^m (a_k \cos(kt) + b_k \sin(kt)),$$

with real coefficients a_i, b_i . Also in this case, we have to say that we do not get the Fourier series, which is a particular series of trigonometric polynomials: a similar argumentation as for Taylor series in Remark 5.70 holds (also compare with Theorem 2.57).

5.10 Topological equivalence and metrizability

As we have seen in the previous paragraphs, being a metric space is the best thing that may happen to a general topological space. Hence, a natural question arises: when the topology of a given topological space is induced by a metrics? in other words, given a topological space X with topology \mathcal{T} , when there exists a metric d on X such that the corresponding topology \mathcal{T}_d (i.e. the topology generated by the open balls as basis) coincides with the topology \mathcal{T} ?

Before answering to such a question, let us concentrate to another (somehow related) question. We start with a definition.

Definition 5.71 *Let X be a non empty set and d_1, d_2 two metrics on it. These metrics are said to be topologically equivalent if they generate the same topology $\mathcal{T}_{d_1} = \mathcal{T}_{d_2}$.*

Hence we have this question: given a non empty set X and two metrics on it when such two metrics are topologically equivalent? A first obvious answer is: when the identity map $i : (X, d_1) \rightarrow (X, d_2)$, $x \mapsto x$ is a homomorphism. Indeed, in such a case every open set for (X, d_1) is also open for (X, d_2) and vice-versa. A slightly different way to approach the question is to look to the convergence: every convergent sequence in (X, d_1) must also converge in (X, d_2) and vice-versa. This point of view leads to consider the neighborhoods of the points and to conclude that the metrics are equivalent if and only if every balls centered at x for the metrics d_1 contains a ball centered at the same x for the metrics d_2 and vice-versa. A sufficient conditions for that is the following one:

$$\forall x \in X \exists r > 0, \exists \alpha = \alpha(x, r) > 0, \exists \beta = \beta(x, r) > 0 \text{ such that} \quad (5.15)$$

$$\alpha d_1(y, z) \leq d_2(y, z) \leq \beta d_1(y, z) \quad \forall y, z \in B_{d_1}(x, r) \cap B_{d_2}(x, r).$$

Example 5.72 Let us consider \mathbb{R} endowed with the usual metrics $d_1(x, y) = |x - y|$ and with the metrics $d_2(x, y) = |\arctan(x) - \arctan(y)|$. Then the two metrics are topological equivalent. Indeed, we have

$$|\arctan x - \arctan y| \leq |x - y| \quad \forall x, y \in \mathbb{R},$$

and so we can always take $\beta = 1$. On the other hand, for any bounded set $C \subseteq \mathbb{R}$ if we define¹⁸⁹

$$\alpha = \inf_{x \in C} \frac{1}{1 + x^2},$$

then we have

$$\alpha|x - y| \leq |\arctan(x) - \arctan(y)| \quad \forall x, y \in C.$$

Note that in this case α cannot be taken uniformly on \mathbb{R} as instead it happens for β . This is the reason for introducing the local formulation (5.15).

¹⁸⁹Recall the derivative of the arctangent...

Remark 5.73 We have already seen that \mathbb{R} with the usual metrics is complete whereas with the arctangent metric is not complete. Hence the topological equivalence do not implies a “metric equivalence” that is (X, d_1) and (X, d_2) may be not isometric even if they are topologically equivalent. If we look to Cauchy sequences, the reason for that is the following: if they are topologically equivalent they have the same convergent sequences, but they may have different Cauchy sequences, because being a Cauchy sequence depends only on metric and not on topological properties.

Now, let us come back to the initial question about a topology given by a metrics. If the topology is induced by a metrics, then the topological space must have the characteristic properties of a metric space. First of all it must be a Hausdorff space, and it must satisfies the first axiom of countability. There is another important property which is satisfied by a metric space: the fact that it is normal.

Definition 5.74 A topological space X is said to be normal if every singleton $\{x\}$ is closed¹⁹⁰, and every pair of disjoint closed subsets are separable, that is

$$C_1, C_2 \subseteq X \text{ closed, } C_1 \cap C_2 = \emptyset \implies \exists A_1, A_2 \in \mathcal{T} \text{ such that } C_1 \subseteq A_1, C_2 \subseteq A_2, A_1 \cap A_2 = \emptyset.$$

It is evident that every normal space is Hausdorff, being points closed. The converse is not true: there are Hausdorff spaces which are not normal. For instance consider the interval $[0, 1]$ where every $x > 0$ has the usual neighborhoods, whereas 0 has a basis of neighborhoods given by sets of the form

$$[0, \varepsilon[\setminus \left\{ \frac{1}{n} \mid n \in \mathbb{N} \setminus \{0\} \right\}], \text{ with } \varepsilon > 0.$$

Such a space is a Hausdorff space but it is not normal. Indeed, the two sets

$$\{0\}, \left\{ \frac{1}{n} \mid n \in \mathbb{N} \setminus \{0\} \right\}$$

are both closed, disjoint, but not separable¹⁹¹.

Proposition 5.75 Every metric space is normal.

Proof. Every metric space is Hausdorff and so the points are closed. Let C_1, C_2 be two closed disjoint subsets. Then, for every $x \in C_1$ there exists $\rho_x > 0$ and for every $y \in C_2$ there exists $\rho_y > 0$ such that

$$d(x, y) \geq \rho_x \quad \forall y \in C_2, \quad d(x, y) \geq \rho_y \quad \forall x \in C_1$$

¹⁹⁰Roughly speaking: the points are closed

¹⁹¹The fact that the second one is closed comes from the fact that, apart from the elements $1/n$, any other possible adherent point should be 0, but 0 is not adherent because it has neighborhoods which do not intersect the set.

Indeed, if not, looking to the first inequality, for every n there exists $y_n \in C_2$ such that $d(x, y_n) \leq 1/n$, which implies $y_n \rightarrow x$ and so $x \in C_2$ which is a contradiction.

The following open sets are the requested ones

$$A_1 = \bigcup_{x \in C_1} B\left(x, \frac{\rho_x}{2}\right), \quad A_2 = \bigcup_{y \in C_2} B\left(x, \frac{\rho_y}{2}\right).$$

□

Definition 5.76 *A topological space X is said to be metrizable if there exists a metrics on it which generates the same topology.*

Taking also account of Proposition 5.75, we have the following statement.

Proposition 5.77 *If a topological space is metrizable, then it satisfies the first axiom of countability and it is normal.*

Unfortunately, the opposite of Proposition 5.77 is false: there exist topological spaces which satisfy the first axiom of countability and which are normal but nevertheless are not metrizable. A sort of opposite of Proposition 5.77 holds in the case of second axiom of countability.

Theorem 5.78 (*Urysohn's metrization theorem*). *If a topological space satisfies the second axiom of countability then it is metrizable if and only if it is normal.*

5.11 Historical notes

In the next chapter, we are going to give an explicitly example of a continuous function which is nowhere derivable. The first example of that kind was given in the year 1872 by Weierstrass, and was the function.

$$f(x) = \sum_{n=0}^{+\infty} b^n \cos(a^n \pi x),$$

with a and odd integer and $0 < b < 1$.

Before the XIX century, mathematicians strongly believed that any functions have derivatives, except for few isolated points. After the year 1861 Riemann was convinced that the continuous function

$$f(x) = \sum_{n=1}^{+\infty} \frac{\sin(n^2 x)}{n^2}$$

is not derivable for infinitely many values of x . Indeed, this is true, as proved by G.H. Hardy in 1916. However, in 1970 it was demonstrated that there are also infinitely many values of x at which the derivative really exists. Hence, the example of Riemann only

shows that continuous functions may be non-derivable in a set more than finite, but does not show that they may be nowhere derivable.

The example of Weierstrass was certainly a surprise for many mathematicians, but nevertheless it was closing an important question. However, another question was immediately arising: are there some “nice” properties which may guarantee the existence of derivative “at most all points”? It was rather clear that such a property is monotonicity. Many mathematicians searched for a proof that any continuous monotone function is derivable everywhere. But Weierstrass exhibited an example of a monotone function which is not derivable at any rational number. Weierstrass was moreover convinced that such a continuous monotone function is not derivable at any point. But that was wrong. In the year 1903 the French mathematician Henri Lebesgue (1875-1941) proved that a continuous and monotone function must be derivable at “almost every” point, that is the set where it is not derivable has measure zero.

As we already pointed out, the XX century was characterized by a search of abstraction in mathematics. After the works of Cantor, the interested was put not only on real numbers and related questions, but on general sets with general elements. The object of the study was the relation between the elements of a same set (ordering, operations, ...) and between different sets (functions), without being interested on the nature of the elements themselves: many of the historical questions were now seen as particular case of a more general setting and point of view. Such a level of abstraction rapidly showed its power and become one of the fundamental bricks of the modern mathematics. In particular, one of these new abstract fields is the now called “topology theory” and it has become probably the most important and widely present theory in all modern mathematics.

There are several theories about the born of topology. The father may be Cantor, or Poincaré or Brouwer or Fréchet and Hilbert. But probably everyone agrees to consider the german mathematicians Felix Hausdorff (1868-1942) the man who mainly contributed to the development of topology. With his work published in the year 1914, Hausdorff gave a systematic exposition of the aspects of the set theory where the nature of elements was not important and, in the second part of the same work, he gave precise formulation of what we now call “Hausdorff spaces” with an axiomatic definition of the filter of the neighborhoods of a point.

It is interesting to note that all was started with the so-called “arithmetization of analysis”: initially in the XIX centuries mathematicians were interested in the foundations of the mathematical analysis: the concept of number, the construction of real numbers, ordering of numbers, the concept of function of real numbers, convergence, sequences, series and limits. All was around the notion of “number”. At the end of the process, the concept of number was almost passed in a secondary level and covered by a very more general point of view. Moreover, even if topology speaks about “points”, such points have very few in common with the points of classical geometry.

6 Construction of some special functions

In this section we give a direct construction of two considerable examples of functions in \mathbb{R}^n .

6.1 A continuous nowhere derivable function

Let us take the function $\varphi : [-1, 1] \rightarrow \mathbb{R}$ defined by $\varphi(x) = |x|$, and extend it to the whole real line by periodicity¹⁹² with period 2. Let us call φ_0 such extended function. By construction, we have that φ_0 is bounded and Lipschitz continuous with Lipschitz constant 1:

$$0 \leq \varphi(x) \leq 1, \quad |\varphi_0(x) - \varphi_0(y)| \leq |x - y| \quad \forall x, y \in \mathbb{R}.$$

We now define the following function

$$f : \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) = \sum_{n=0}^{+\infty} \left(\frac{3}{4}\right)^n \varphi_0(4^n x).$$

The function f is our candidate to be continuous but nowhere derivable. Note that f is constructed by a series of a sort of “compressed” copies of φ_0 and so of functions which have a density of cusps bigger and bigger¹⁹³. For all $n \in \mathbb{N}$, let us define the n -th term of the series as φ_n .

We first have to prove that f is well defined (that is the series converges for all $x \in \mathbb{R}$) and that it is continuous. By construction, we have that, for every n , φ_n is bounded and Lipschitz continuous with Lipschitz constant 3^n :

$$|\varphi_n(x)| \leq \frac{3^n}{4^n}, \quad |\varphi_n(x) - \varphi_n(y)| \leq \frac{3^n}{4^n} |4^n x - 4^n y| = 3^n |x - y| \quad \forall x, y \in \mathbb{R}. \quad (6.1)$$

By the first estimate of (6.1) and by the Weierstrass criterium Proposition 2.47, we immediately get the uniform convergence of the series on the whole real line and so the existence and continuity of f .

Let us prove that f is nowhere derivable. Take $x \in \mathbb{R}$. We are going to construct a suitable sequence of points convergent to x , such that the absolute value of the incremental ratio computed along the sequence diverges to $+\infty$. This will show that the derivative of f at x does not exist. The arbitrariness of x will conclude the argument. For every $m \in \mathbb{N}$ we take

¹⁹²That is you replicate infinitely many times the graph of φ , getting then a saw-tooth function, with period 2.

¹⁹³In a interval of length 1, φ_0 has at most two cusps, whereas the n -th term of the series has at least 4^n cusps.

$$\begin{aligned}
x_m &= x + \frac{1}{2 \cdot 4^m} && \text{if }]4^m x, 4^m x_m[\text{ does not contain integer numbers,} \\
x_m &= x - \frac{1}{2 \cdot 4^m} && \text{otherwise.}
\end{aligned} \tag{6.2}$$

Note that, since $4^m/(2 \cdot 4^m) = 1/2$, then the definition of x_m (6.2) implies that there is never an integer number between x and x_m . Let us define the increment as

$$\varepsilon_m = \pm \frac{1}{2 \cdot 4^m},$$

with the sign taken accordingly to (6.2). For $n > m$ we have, by the periodicity with period 2 of φ_0 ,

$$\varphi_n(x_m) = \frac{3^n}{4^n} \varphi_0(4^n x_m) = \frac{3^n}{4^n} \varphi_0(4^n(x + \varepsilon_m)) = \frac{3^n}{4^n} \varphi_0(4^n x + 4^n \varepsilon_m) = \frac{3^n}{4^n} \varphi_0(4^n x) = \varphi_n(x).$$

Hence we have, also using the uniform convergence which guarantees the possibility of adding term by term,

$$\begin{aligned}
\left| \frac{f(x + \varepsilon_m) - f(x)}{\varepsilon_m} \right| &= \left| \sum_{n=0}^{+\infty} \frac{3^n \varphi_0(4^n(x + \varepsilon_m)) - \varphi_0(4^n x)}{4^n \varepsilon_m} \right| \\
&= \left| \sum_{n=0}^{m-1} \frac{3^n \varphi_0(4^n(x + \varepsilon_m)) - \varphi_0(4^n x)}{4^n \varepsilon_m} + \frac{3^m \varphi_0(4^m(x + \varepsilon_m)) - \varphi_0(x)}{4^m \varepsilon_m} \right|
\end{aligned} \tag{6.3}$$

By our construction of the sequence x_m , $4^m(x + \varepsilon_m) = 4^m x \pm 1/2$ and between $4^m x$ and $4^m x_m$ there are no integers. This implies¹⁹⁴

$$|\varphi_0(4^m x_m) - \varphi_0(4^m x)| = |4^m x_m - 4^m x| = \frac{1}{2}.$$

Hence, the absolute value of the second addendum of the last term of (6.3) is exactly equal to 3^m . Moreover, by (6.1), the absolute value of every term of the summary in (6.3) is less than or equal to 3^n . Hence, continuing with the inequalities, we have

$$\left| \frac{f(x + \varepsilon_m) - f(x)}{\varepsilon_m} \right| \geq 3^m - \sum_{n=0}^{m-1} 3^n = 3^m - \frac{1 - 3^m}{1 - 3} = \frac{3^m + 1}{2} \rightarrow +\infty \text{ as } m \rightarrow +\infty,$$

and the argument is concluded.

Figure 4 shows the graphs over the interval $[-1, 1]$ of the first four partial summations of f , corresponding to $n = 0, 1, 2, 3$.

¹⁹⁴Because, being there no integers, between $4^m x_m$ and $4^m x$, φ_0 acts just as the absolute value.

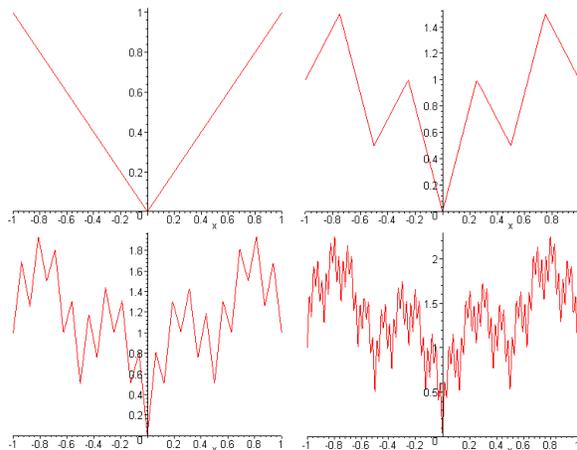


Figure 4: Approximation of a continuous nowhere derivable function

6.2 Abundance of smooth functions in \mathbb{R}^n

We have seen, in the last chapter, that a metric space X is a normal space, that is, for every pair of disjoint closed subsets C_1, C_2 , there exist two open disjoint sets A_1, A_2 such that $C_1 \subseteq A_1, C_2 \subseteq A_2$. It is easy to see that such a property is implied by the following one: for every pair of disjoint closed subsets, C_1, C_2 , there exists a continuous function $f : X \rightarrow [0, 1]$ such that $f(x) = 0$ if and only if $x \in C_1$ and $f(x) = 1$ if and only if $x \in C_2$ ¹⁹⁵. Actually, it can be proved that such a property is also equivalent to the one of being normal (whenever the points are closed).

In the particular case of $X = \mathbb{R}^n$, we get even more, that is the continuous function $f : \mathbb{R}^n \rightarrow [0, 1]$ separating the closed sets, may be also taken of class C^∞ , that is derivable at any order with continuous derivatives.

Proposition 6.1 (*Abundance of smooth functions*). *Given a closed set $C \subseteq \mathbb{R}^n$, there exists a C^∞ function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ which vanishes exactly on C ¹⁹⁶, that is*

$$f(x) = 0 \text{ if and only if } x \in C.$$

Proof. Let us consider the following function:

$$u_0 : \mathbb{R} \rightarrow \mathbb{R} \quad u(t) = \begin{cases} e^{-1/t} & \text{if } t > 0 \\ 0 & \text{if } t \leq 0. \end{cases}$$

¹⁹⁵If such a property holds, then $A_1 = f^{-1}([0, 1/2[)$ and $A_2 = f^{-1}(]1/2, 1])$ are two requested open sets.

¹⁹⁶Note that such a result holds for every closed set, that is also for closed sets very irregular, very bad, with many cusps on their boundaries; nevertheless, the requested function may be found of class C^∞ .

It is not hard to see that u_0 is of class C^∞ on \mathbb{R} .

Let $x \in \mathbb{R}^n$ and $r > 0$. Hence, there exists a C^∞ function $\varphi_{x,r} : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $\varphi_{x,r}(y) > 0$ if and only if y belongs to the open ball $B(x, r)$. Indeed, it is sufficient to take

$$\varphi_{x,r}(y) = u_0 \left(1 - \left| \frac{y-x}{r} \right|^2 \right).$$

Now, let $C \subseteq \mathbb{R}^n$ be a closed subset. Then $\mathbb{R}^n \setminus C$ is open, and so, being \mathbb{R}^n separable, it is the union of an at most countable family of open balls (see Proposition 5.57). Let $\{B(x_k, r_k)\}_k$ be such a family and consider the family of C^∞ functions φ_{x_k, r_k} as constructed before. Note that, every mixed partial derivative of any order of φ_{x_k, r_k} is still a C^∞ function which vanishes outside the corresponding open ball $B(x_k, r_k)$, in particular it is bounded on \mathbb{R}^n . Hence, for very k , we define¹⁹⁷

$$M_k = \max \left\{ \max_{x \in \mathbb{R}^n} |\partial^\alpha \varphi_{x_k, r_k}(x)| \mid \alpha \text{ multi-index, } |\alpha| \leq k \right\} > 0.$$

We then define

$$f : \mathbb{R}^n \rightarrow \mathbb{R} \quad f(x) = \sum_{n=0}^{+\infty} \frac{\varphi_{x_k, r_k}(x)}{2^k M_k}.$$

By definition of M_k , the series totally converge in \mathbb{R}^n (the norm of every term is less than or equal to 2^{-k}), and so, by the Weierstrass criterium, Proposition 2.47, applied to the convergence in \mathbb{R}^n , the series uniformly converges to a the well-defined continuous function f . Moreover, the series of the derivatives also uniformly converge on \mathbb{R}^n because, still by the definition of M_k , fixed a multi-index α , very term with index $k > |\alpha|$ is in norm less than or equal to 2^{-k} . Hence, by a result of derivation by series (see Remark 2.49), we conclude that f is well-defined and of class C^∞ .

Finally, it is evident that $f(x) = 0$ whenever $x \in C$ (because, by construction, all the function φ_{x_k, r_k} vanish on x), and that $f(x) > 0$ whenever $x \notin C$ (because at least one of the functions φ_{x_k, r_k} does not vanish on x and also by the fact that all such functions are nonnegative). \square

Corollary 6.2 *Given two disjoint closed sets C_1, C_2 in \mathbb{R}^n , then there exists a C^∞ function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $f(x) = 0$ if and only if $x \in C_1$, $f(x) = 1$ if and only if $x \in C_2$, and that $f(x) \in [0, 1]$ for all $x \in \mathbb{R}^n$.*

¹⁹⁷Here, by a multi-index α we mean an ordered string of n natural numbers $(\alpha_1, \alpha_2, \dots, \alpha_n)$, the length of α is $|\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_n$, and $\partial^\alpha \varphi$ stays for the mixed partial derivative of φ corresponding to α_1 partial derivatives with respect to x_1 , α_2 partial derivatives with respect to x_2, \dots , α_n partial derivatives with respect to x_n .

Proof. Let us take the functions f_1, f_2 respectively corresponding to the closed sets C_1 and C_2 as in Proposition 6.1. Recall that such functions are, by construction, nonnegative. By the hypothesis of disjointness, the function

$$f : \mathbb{R}^n \rightarrow \mathbb{R} \quad f(x) = \frac{f_1(x)}{f_1(x) + f_2(x)}$$

satisfies all the requests.

□

References

- [1] T. M. Apostol: *Calcolo - Volume Terzo, Analisi 2*, Bollati Boringhieri, Torino, 1992.
- [2] K. G. Binmore: *Foundations of Analysis: a Straightforward Introduction - Book 1 and Book 2*, Cambridge University Press, London, 1980.
- [3] C. B. Boyer: *Storia della Matematica*, Arnoldo Mondadori Editore, Milano, 1987.
- [4] D. Bressoud: *A Radical Approach to Real Analysis*, The Mathematical Association of America, Washington, 2007.
- [5] H. Brezis: *Analisi Funzionale*, Liguori, Napoli, 1982.
- [6] G. De Marco: *Analisi Due*, Zanichelli-Decibel, Bologna, 1993.
- [7] G. Gilardi: *Analisi uno*, McGraw-Hill, Milano, 1991.
- [8] A. N. Kolmogorov, S. V. Fomin: *Introductory Real Analysis*, Dover Publications, New York, 1970.
- [9] C. D. Pagani, S. Salsa: *Analisi Matematica - Volume 2*, Masson, Milano, 1991.
- [10] G. Prodi: *Analisi Matematica*, Boringhieri, Torino, 1984.
- [11] G. Prodi: *Lezioni di Analisi Matematica II*, ETS Editrice, Pisa, 1987.
- [12] W. Rudin: *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1976.
- [13] T. Valent: *Appunti di Analisi Superiore*, manuscripted notes of a course at University of Padova, late '80s early '90s.