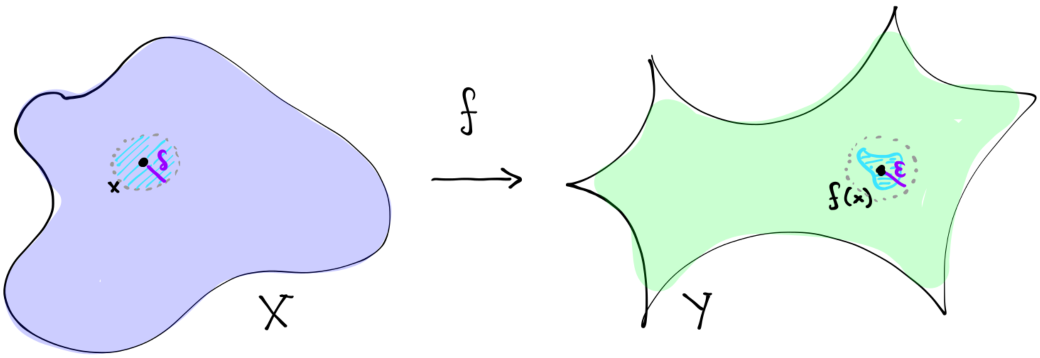


Metric spaces

Kalle Kytölä

February 12, 2024

Contents

Foreword

These lecture notes are primarily intended for the core BSc-level mathematics course **MS-C1541** *Metric Spaces* at Aalto University.

The structure of these notes is largely based on an earlier version of the course taught by Pekka Alestalo. Directly or indirectly, a lot of inspiration is also drawn from the textbook [**Väi99**], which has been dear to me since my own first year at the University of Helsinki more than two decades ago.

The notes are still in a preliminary and incomplete form, and I plan to frequently update them during the current course. I have received many valuable comments from the participants of the courses of 2021, 2022, and 2023. For numerous corrections in 2021, Sampo Paukkonen in particular deserves a special mention and many thanks. I am extremely grateful to the head teaching assistant for the course in 2022, Milo Orlich, who read the notes very carefully and made tireless and timely corrections of my mistakes. No doubt I have introduced some new mistakes after old ones have been corrected. I would therefore appreciate if you still help me — or perhaps more importantly the students who will use this material — by reporting mistakes, misprints, needs for clarification, etc., to me via the course's Zulip chat forum or by email (`kalle.kytola@aalto.fi`).

The elegant and succinct textbook [**Väi99**] matches the contents and level of this course very well, but it unfortunately only exists in Finnish and Swedish. As English textbooks for the course we recommend for example [**Car00**] or [**Rud76**], both of which cover significantly more material than the present course.

In 2024, at Aalto University, there will also be a course on general topology taught in period IV. For mathematically oriented students, that course would be a very natural continuation to the present one: it addresses topology at the level of abstraction and generality that has proven extremely fruitful in the past 100 years of mathematics research, and that constitutes one of the cornerstones of university level mathematics almost universally.

Symbols indicating material beyond core content:

Section symbols

♡ *stands for sections that are* **optional** *for the purposes of the present course; they are intended for those who prepare to study mathematics further or seek in depth understanding.*

Exercise symbols

✓ *stands for exercises which should be quick, easy, and* **routine**, *in view of the preceding definitions and/or basic results.*

$\ddots$ *stands for exercises whose solution can be* **lengthy** *in comparison to regular ones, but not necessarily more difficult.*

♯ *stands for exercises that are more* **challenging** *than regular ones; they may require some creativity and a good command of various topics in basic mathematics, in addition to the newly introduced material.*

Lecture I

Foundations: set theory and logic

Study tips

This chapter is long... Most of the material should, however, be familiar to you already. This more extensive background chapter is intended to serve two purposes:

- If there are things you are not yet familiar with, you can still catch up now and follow the rest of the course.
- We fix some notations and terminology that will be used in the subsequent chapters. If you are already very familiar with set theory and logic, perhaps just quickly check the terms in **bold** and the related notation.

Among the topics of this chapter, the images and preimages of sets under functions (the last part of Section I.2) is perhaps the only one likely to not have been covered in prior basic courses. It may correspondingly deserve some more attention.

The students planning further mathematical studies are advised to take this opportunity to look into countable and uncountable infinite sets (Appendix A.1), although those topics are not strictly necessary for the present course.

Some perspectives on set theory before we start

Mathematics, as it has been practiced especially in the 20th and 21st centuries, is (almost entirely) founded on the notions of *sets* and their *elements*. To begin to appreciate the expressive power of set theory, one may note that, e.g., the notion of a function can be reduced to that of sets,[1] and constructions of real numbers[2] and even of natural numbers[3] can be given based on sets. Such flexibility of set theory justifies taking it as the foundations of mathematics.

Historical reasons for looking for solid foundations for mathematics include the discovery of numerous apparent paradoxes, questions about the validity of certain arguments about numbers and functions, and even disagreements about what can be considered legitimate topics of mathematics in the first place. The late 19th century and early 20th century featured particularly notable controversies of this kind, and lead to the adoption of set theory as the (more or less) universally agreed-on foundations. In this course, we will encounter a few of those examples that had been

[1] A function *is* its "graph", a subset of the Cartesian product of the domain and codomain of the function. For more details, we refer to more serious treatments of set theory, e.g., [**Hal74**].

[2] One noteworthy construction of the real numbers is as Dedekind cuts of the set of rational numbers, see for example Wikipedia `https://en.wikipedia.org/wiki/Dedekind_cut`.

[3] Indeed, starting strictly from the set-theoretic foundations, one should *construct* natural numbers 1, 2, 3, ..., instead of assuming their existence! Again we refer to serious treatments of set theory about the details. Or, if you prefer a puzzle game, you can get a feeling for this from the *Natural Number Game* `https://adam.math.hhu.de`.

considered controversial before the foundations were clarified: for example the Cantor set (Appendix B.3) and the Weierstrass function (Figure IX.3), and countable and uncountable infinities (Appendix A.1). Being precise about the foundations probably became necessary when the level of abstraction increased sufficiently so that one could no longer rely on too informal intuitive reasoning. This historical rationale has its parallel in mathematics studies: as one proceeds to more advanced and abstract mathematics, precise formalism becomes more important, and it is eventually good to make peace with foundational questions, too.

Given that mathematics is founded on set theory, a student of mathematics must become familiar and comfortable with set-theoretic notation and reasoning. Pragmatically, set theory at the very least forms the common *language* in which virtually all of university-level mathematics is phrased: measure and integration, probability theory, group theory and algebra, differential geometry, real analysis, complex analysis, functional analysis, etc. etc.

The present course, specifically, falls almost entirely within what is often labeled as *point set topology* (as opposed to *algebraic topology*). This terminology is best understood by noting that the expression *an element of a set* is used exactly parallel to the expression *a point in space*.[4] To refer to a "space" as a "point set" is to emphasize that as a mathematical object a space is first and foremost the set consisting of its points, and only as such potentially equipped with some further structure.

I.1. Basics of set theory

Sets and elements

Sets consist of elements — a **set** is the collection of all **elements** that belong to it. It is customary to denote sets by upper case letters and their elements by lower case letters, as we mostly do also below. However, this is merely a typographical practice often adopted because it can serve as a helpful notational cue; it will be discarded when appropriate.[5]

We denote $a \in A$, if a is an element of a set A. If a is not an element of A, we denote $a \notin A$. The *binary relation* denoted by the symbol $\in$ (whose negation is denoted by the symbol $\notin$) is at the heart of set theory. In natural language, "$\in$" can be read as "to belong to"[6], for example

$$\begin{aligned} a \in A \quad &: \quad \text{(the element) } a \text{ belongs to (the set) } A, \qquad \text{(I.1)} \\ a \notin A \quad &: \quad \text{(the element) } a \text{ does not belong to (the set) } A. \end{aligned}$$

[4] We will in fact not precisely define the notions of *a point* and *a space* as such, but rather we use them as synonymous to *an element* and *a set* whenever we wish to draw attention to the fact that the set is also furnished with some additional structure. Various specific structures will be defined precisely later on: we will consider vector spaces, inner product spaces, normed spaces, metric spaces, topological spaces, ... — but the general term *space* itself will not be used in a definite mathematical sense.

[5] Indeed, for example for sets whose elements themselves are sets, it becomes impossible to strictly adhere to such a typographical "rule".

[6] Suomeksi "$\in$" luetaan "kuuluu (joukkoon)".

As an informal (not exactly well-defined) example of a set, one could consider the set of all fruits, which has as its elements apples, kiwis, mangos, etc. More mathematical examples of sets include:

- the set $\mathbb{Z}$ of all integers, having $\ldots, -2, -1, 0, 1, 2, \ldots$ as its elements;
- the open interval $(-\pi, \pi)$ is the set whose elements are the real numbers x satisfying $-\pi < x < \pi$;
- the set of all continuous real-valued functions on the real line;
- the set of all subsets of the plane $\mathbb{R}^2$;
- $\ldots$.

In particular, sets may have numbers as their elements (a probably very familiar case already). But sets may also have much more general types of elements: functions, other sets, etc. etc.[7]

Remark I.1 (What does it take to specify a set?)**.**
A set is known when its elements are known — i.e., when for every possible object we are able to decide whether the object is an element of the set or not.

In particular two sets A and B are equal, denoted $A = B$, if they have exactly the same elements. (This is known as the *axiom of extensionality* in set theory.)

The following (small) sets show up often:

- The **empty set** $\emptyset$ is the set which does not have any elements.
- A **singleton** is a set which has exactly one element. The singleton consisting of an element a is denoted $\{a\}$.

Remark I.2 (Do not confuse a singleton with its element!)**.**
Note that

$$a \in \{a\} \qquad \text{but} \qquad a \neq \{a\}\,.$$

Exercise (✓) I.1 (Empty set vs. the set consisting of the empty set, etc.)**.**
How many elements are in the following sets:

(a): $\emptyset$, (b): $\{\emptyset\}$, (c): $\{\{\emptyset\}\}$?

When possible[8], the simplest method of specifying a set is to list its elements:

- The notation $\{a_1, a_2, \ldots, a_n\}$ stands for the set consisting of $a_1, a_2, \ldots, a_n$.

Note that we have, e.g., $\{9, 1, 1\} = \{1, 9\}$ (this set has two elements: 1 and 9). A set is only the collection of elements belonging to it; the order of elements has no significance, and repetition is redundant.

[7] To give an example, problems in nonparametric statistics of real-valued data are basically optimization problems over the set of all those functions from the set of all measurable subsets of the set of real numbers to the unit interval, which satisfy the three defining properties of a probability measure. In traditional (parametric) statistics the optimization problem is resticted to a subset of the above set. Don't be fooled by apparent simplicity; the expressive power of set theory is vast, and actual mathematical applications make good use of that expressive power.

[8] For any finite set (i.e., a set with only finitely many elements), listing the elements is possible, at least in principle. If the elements of a countably infinite set follow a list-like pattern, a similar notation with ellipsis may be used, as we do in, e.g., (I.4), (I.5), and Example I.3. But it would be dubious to try to specify an uncountably infinite set by listing its elements!

Another common method of specifying a set is to use a logical condition as a criterion for whether an element belongs to the set:

- If A is a set and $P(x)$ is a logical proposition depending on a variable x, then the notation
$$\big\{x \in A \;\big|\; P(x)\big\} \tag{I.2}$$
stands for the set consisting of those elements x of the set A for which $P(x)$ is true.

Example I.3 (Examples of sets defined by a condition)**.**
In the following examples, various conditions are used to "extract" from the set $\mathbb{R}$ of all real numbers a subset of those that satisfy the condition in question:
$$\begin{aligned} \Big\{x \in \mathbb{R} \;\Big|\; \sin(x) = 0\Big\} &= \{\ldots, -2\pi, -\pi, 0, \pi, 2\pi, \ldots\} \\ \left\{x \in \mathbb{R} \;\middle|\; x^4 = \frac{1}{3}\right\} &= \left\{\frac{-1}{\sqrt{\sqrt{3}}}, \frac{1}{\sqrt{\sqrt{3}}}\right\} \\ \Big\{x \in \mathbb{R} \;\Big|\; x^2 < 0\Big\} &= \emptyset. \end{aligned}$$

We will also take some liberties to modify the above notational conventions in hopefully self-explanatory ways, such as:
$$\begin{aligned} \Big\{a_j \;\Big|\; j \in \{1, 2, \ldots, n\}\Big\} &= \{a_1, a_2, \ldots, a_n\}, \\ \big\{\sqrt{x} \;\big|\; x \in \mathbb{N}\big\} &= \Big\{1,\ \sqrt{2},\ \sqrt{3},\ 2,\ \sqrt{5},\ \ldots\Big\}, \qquad \text{etc.} \end{aligned}$$

Subsets

A set A is said to be a **subset** of another set B if every element of A is also an element of B; we then denote $A \subset B$. Being a subset is a binary relation among sets, denoted by the symbol $\subset$. Its negation is denoted by $\not\subset$, so $A \not\subset B$ means that there exists at least one element in the set A which is not an element of B.

Note the difference between the relations $\in$ and $\subset$.[9]

In natural language, the subset symbol $\subset$ could be read as[10]
$$A \subset B \qquad : \qquad \text{(the set) } A \text{ is contained in (the set) } B, \tag{I.3}$$
or simply as: A is a subset of B. For precise and unambiguous mathematical meaning, it is best to avoid mixing the natural language expressions "belongs to" and "is contained in", which stand for the relations $\in$ and $\subset$, respectively.

Evidently, the subset relation is **transitive**: if $A \subset B$ and $B \subset C$, then $A \subset C$.

[9] If you have some experience in programming with *Python* or *Mathematica* (or other programming languages which encourage extensive use of nested lists), you are probably familiar with bugs which occur if you mix up a test of whether one list (perhaps containing just one element) is an element rather than a sublist of another list. A mathematical equivalent of such a programming error is mixing up the relations "$\in$" and "$\subset$". Beware of that bug!

[10] Suomeksi "$\subset$" voidaan lukea "sisältyy" — erotuksena relaatiosta "$\in$", joka luettiin "kuuluu". Toisinaan selkeyden vuoksi on kuitenkin turvallisinta käyttää ilmaisua "on osajoukko".

Exercise (✓) I.2 (Subsets of the empty set, the set consisting of the empty set, etc.).
How many subsets do the following sets have:

$$\text{(a):}\quad \emptyset, \qquad \text{(b):}\quad \{\emptyset\}, \qquad \text{(c):}\quad \{\{\emptyset\}\}, \qquad \text{(d):}\quad \{\emptyset,\{\emptyset\}\} \quad ?$$

Inverted relation symbols

When it is more appropriate to mention, e.g., a set before its element or a set before its subset, the inverted relation symbols $\ni$ and $\supset$ are used so that

$$\begin{aligned} A \ni a \qquad &\text{means} \qquad a \in A, \\ B \supset A \qquad &\text{means} \qquad A \subset B. \end{aligned}$$

The relation $\supset$ can be read as "to contain", so that

$$B \supset A \qquad : \qquad \text{(the set) } B \text{ contains (the set) } A,$$

Remark I.4 (Equality of two sets).
Two sets A and B are equal if they have exactly the same elements. This occurs if and only if both are subsets of the other: $A \subset B$ and $B \subset A$.

While the above may initially seem like a useless remark, it lends itself to a strategy of proof that is very common. To *prove* that $A = B$, it is often practical to separately show $A \subset B$ and $A \supset B$, by first arguing that any element of A must necessarily be also an element of B, and then vice versa.

Familiar sets of numbers

The following examples of sets of numbers should be familiar:[11]

$$\begin{aligned} &\text{the set of natural numbers} && \mathbb{N} = \{1,2,3,4,\ldots\} && \text{(I.4)} \\ &\text{the set of integers} && \mathbb{Z} = \{\ldots,-2,-1,0,1,2,\ldots\} && \text{(I.5)} \\ &\text{the set of rational numbers} && \mathbb{Q} = \left\{ \frac{n}{m} \;\middle|\; n \in \mathbb{Z},\ m \in \mathbb{N} \right\} && \text{(I.6)} \\ &\text{the set of real numbers} && \mathbb{R}\,. && \text{(I.7)} \end{aligned}$$

To give just a few examples, we have

$$\begin{aligned} 2023 &\in \mathbb{N}, & -42 &\in \mathbb{Z}, & \sqrt{2} &\in \mathbb{R}, \\ 0 &\notin \mathbb{N}, & -\frac{5}{7} &\in \mathbb{Q}, & \sqrt{2} &\notin \mathbb{Q}. \end{aligned}$$

[11] Note that zero is not considered a natural number... in this course. Unfortunately, mathematical literature is very divided regarding conventions about this issue; different areas of mathematics and different authors adopt different conventions. As a rule of thumb, in analysis it is typical to follow our present convention $0 \notin \mathbb{N}$, whereas in algebra it is typical to follow the convention that zero is a natural number. In fact, if you ever attend my algebra courses, then a different convention for natural numbers will be used there. ´Sorry! In the present course, if we want to include zero, we use the notation $\mathbb{N}_0 = \{0,1,2,3,\ldots\}$ instead of $\mathbb{N}$.

The last example above is classical — it was known to the ancient Greeks.[12] We provide it here as one of our first examples of a careful mathematical proof. It is also a great illustration of the method of proof by contradiction.

Theorem I.5 (Square root of two is irrational)**.**
We have $\sqrt{2} \notin \mathbb{Q}$.

Proof. Suppose, by contrary, that $\sqrt{2} \in \mathbb{Q}$. In that case we can write $\sqrt{2} = \frac{n}{m}$ with $n, m \in \mathbb{N}$. If the integers n and m have common factors, they may be cancelled, so we may assume that m and n here are chosen coprime (i.e., no integer greater than one divides both).

Multiplying by m, we obtain that $\sqrt{2}m = n$. Then squaring both sides we see that $2m^2 = n^2$. This shows that n^2 is even, which is only possible if n is even, i.e., $n = 2k$ for some $k \in \mathbb{N}$. But in this case we find $m^2 = \frac{1}{2}n^2 = \frac{1}{2}(2k)^2 = 2k^2$, which similarly implies that m^2 is even, and thus also m is even. Therefore 2 divides both n and m, which contradicts the choice of these coprime integers.

We conclude that $\sqrt{2}$ could not have been rational in the first place, so $\sqrt{2} \notin \mathbb{Q}$. □

The irrationality of π, while well known, is more difficult to prove.

Exercise (♯ ∴) I.3 (Irrationality of π)**.**
Prove that $\pi \notin \mathbb{Q}$.

The above sets of numbers provide a few good examples of subsets, for example $\mathbb{N} \subset \mathbb{Z}$, $\mathbb{Z} \subset \mathbb{Q}$, and $\mathbb{Q} \subset \mathbb{R}$; or more concisely

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

By transitivity, further subset relations such as $\mathbb{N} \subset \mathbb{Q}$, $\mathbb{Z} \subset \mathbb{R}$, and $\mathbb{N} \subset \mathbb{R}$ are apparent from this.

Intervals

Intervals of various types are extremely frequently used subsets of the real line $\mathbb{R}$. For intervals between points $a, b \in \mathbb{R}$, $a \leq b$, we use the following notation and terminology:

closed interval $$[a,b] = \left\{ x \in \mathbb{R} \;\middle|\; a \leq x \leq b \right\} \tag{I.8}$$

(bounded) **open interval** $$(a,b) = \left\{ x \in \mathbb{R} \;\middle|\; a < x < b \right\} \tag{I.9}$$

$$(a,b] = \left\{ x \in \mathbb{R} \;\middle|\; a < x \leq b \right\} \tag{I.10}$$

$$[a,b) = \left\{ x \in \mathbb{R} \;\middle|\; a \leq x < b \right\}. \tag{I.11}$$

We also use the following notation for unbounded intervals,

$$(-\infty, b] = \left\{ x \in \mathbb{R} \;\middle|\; x \leq b \right\}, \qquad [a, +\infty) = \left\{ x \in \mathbb{R} \;\middle|\; a \leq x \right\}, \tag{I.12}$$

$$(-\infty, b) = \left\{ x \in \mathbb{R} \;\middle|\; x < b \right\}, \qquad (a, +\infty) = \left\{ x \in \mathbb{R} \;\middle|\; a < x \right\}, \tag{I.13}$$

[12]The Pythagoreans held it that only rational numbers were of divine origin, and a (probably dubious) legend has it that they killed Hippasus for revealing the irrationality of the length ($\sqrt{2}$) of the diagonal of the unit square to the outside world. As for written records, a proof of the irrationality of $\sqrt{2}$ is contained in Euclid's *Elements* from ca. 300 BCE.

and sometimes $(-\infty, +\infty) = \mathbb{R}$ is used the whole real line. In addition to the bounded open intervals (I.9), also the unbounded open intervals (I.13) are considered **open intervals**. By contrast, according to the above convention, a closed interval always contains its finite endpoints, so there are no unbounded closed intervals.

Operations with sets

New sets can be formed from old ones by set-theoretic operations. For example, for two sets A and B, we define

$$\text{the } \textbf{union} \text{ of } A \text{ and } B\text{:} \qquad A \cup B = \big\{ x \;\big|\; x \in A \text{ or } x \in B \big\} \tag{I.14}$$

$$\text{the } \textbf{intersection} \text{ of } A \text{ and } B\text{:} \qquad A \cap B = \big\{ x \;\big|\; x \in A \text{ and } x \in B \big\} \tag{I.15}$$

$$\text{the } \textbf{set difference} \text{ of } A \text{ and } B\text{:} \qquad A \setminus B = \big\{ x \;\big|\; x \in A \text{ and } x \notin B \big\}\,. \tag{I.16}$$

If $B \subset A$, then $A \setminus B$ is also called the **complement** of B in A.

Example I.6 (A few examples of set operations)**.**
If we let

$$A = \{1, 2, 3, 4, 5, 6, \ldots\} \qquad \text{(natural numbers)}$$
$$B = \{\ldots, -6, -4, -2, 0, 2, 4, 6, \ldots\} \qquad \text{(even integers),}$$

then the intersection, union, and set differences are

$$A \cap B = \{2, 4, 6, \ldots\}\,, \qquad A \setminus B = \{1, 3, 5, \ldots\}\,,$$
$$A \cup B = \{\ldots, -6, -4, -2, 0, 1, 2, 3, 4, \ldots\}\,, \qquad B \setminus A = \{\ldots, -6, -4, -2, 0\}\,.$$

The set operations follow more or less obvious rules of calculation: for any sets A, B, C we have

$$\begin{cases} A \cup B = B \cup A \\ A \cap B = B \cap A \end{cases} \qquad \text{(commutativity)} \tag{I.17}$$

$$\begin{cases} A \cup (B \cup C) = (A \cup B) \cup C \\ A \cap (B \cap C) = (A \cap B) \cap C \end{cases} \qquad \text{(associativity)} \tag{I.18}$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \qquad \text{(distributivity)} \tag{I.19}$$

$$\begin{cases} A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C) \\ A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C). \end{cases} \qquad \text{(De Morgan's laws)} \tag{I.20}$$

The properties (I.17), at least, should be obvious from the definitions (I.14) and (I.15). We leave it as an exercise to verify most of the remaining properties, but as an example we prove the first De Morgan's law below.

Proof of the first De Morgan's law. The most straightforward proof strategy is the following: look at what condition characterizes an element of the set on the left-hand side according to the definitions, and then successively rewrite equivalent versions of this condition until we reach the defining condition for elements on the set on the right-hand side. The same proof strategy is convenient for all of the rules above. In the case of the first De Morgan's law, the appropriate chain of equivalent conditions is

$$\begin{aligned}
x \in A \setminus (B \cup C) \quad & \iff \quad x \in A \quad \text{and} \quad x \notin B \cup C \\
& \iff \quad x \in A \quad \text{and} \quad \text{not } (x \in B \text{ or } x \in C) \\
& \iff \quad x \in A \quad \text{and} \quad \big(\text{not } x \in B \text{ and not } x \in C\big) \\
& \iff \quad x \in A \quad \text{and} \quad \big(x \notin B \ \text{ and } \ x \notin C\big) \\
& \iff \quad \big(x \in A \ \text{ and } \ x \notin B\big) \quad \text{and} \quad \big(x \in A \ \text{ and } \ x \notin C\big) \\
& \iff \quad x \in (A \setminus B) \cap (A \setminus C).
\end{aligned}$$

Thus an element x belongs to the set $A \setminus (B \cup C)$ if and only if it belongs to the set $(A \setminus B) \cap (A \setminus C)$. This is the asserted equality (I.20) of these two sets. □

Exercise I.4 (Proofs of set-theoretic identities)**.**
Prove the remaining formulas among (I.17) – (I.20).

We can also form unions and intersections of more than two sets — in fact of arbitrary collections of sets. By an indexed collection $(A_j)_{j \in J}$ of sets, we mean that a set A_j is given for each index $j \in J$ (the set J is called the index set of the collection). The **union** of the collection is defined as

$$\bigcup_{j \in J} A_j = \Big\{ x \ \Big| \ x \in A_j \text{ for some } j \in J \Big\}. \tag{I.21}$$

The **intersection** is defined if the collection is non-empty, i.e., if $J \neq \emptyset$, and then it is

$$\bigcap_{j \in J} A_j = \Big\{ x \ \Big| \ x \in A_j \text{ for all } j \in J \Big\}. \tag{I.22}$$

Unions and intersections of two sets, (I.14) – (I.15), are recovered from (I.21) – (I.22) in the special case that the index set has two elements, for example $J = \{1, 2\}$.

When the index set is finite, in particular in the case $J = \{1, 2, \ldots, n\}$, as alternative notations we often use

$$\begin{aligned}
\bigcup_{j \in \{1,\ldots,n\}} A_j \; &= \; \bigcup_{j=1}^{n} A_j \; = \; A_1 \cup \cdots \cup A_n \\
\bigcap_{j \in \{1,\ldots,n\}} A_j \; &= \; \bigcap_{j=1}^{n} A_j \; = \; A_1 \cap \cdots \cap A_n.
\end{aligned}$$

When the index set is $J = \mathbb{N}$, it is common to use the alternative notations

$$\bigcup_{j \in \mathbb{N}} A_j = \bigcup_{j=1}^{\infty} A_j \qquad \text{and} \qquad \bigcap_{j \in \mathbb{N}} A_j = \bigcap_{j=1}^{\infty} A_j.$$

Example I.7 (Union and intersection of nested intervals)**.**
Consider the example $A_n = \left[\frac{-1}{n}, \frac{1}{n}\right]$ for $n \in \mathbb{N}$. Then the union and intersection of the collection $(A_n)_{n \in \mathbb{N}}$ are

$$\begin{aligned}
\bigcup_{n=1}^{\infty} \left[\frac{-1}{n}, \frac{1}{n}\right] &= [-1, 1] && \text{(closed interval)} \\
\bigcap_{n=1}^{\infty} \left[\frac{-1}{n}, \frac{1}{n}\right] &= \{0\} && \text{(singleton).}
\end{aligned}$$

The first of the above claims is easy, but let us justify the second claim in detail. Suppose that $x \in \bigcap_{n=1}^{\infty} \left[\frac{-1}{n}, \frac{1}{n}\right]$. According to the definition of an intersection, this means that $x \in \left[\frac{-1}{n}, \frac{1}{n}\right]$ for all $n \in \mathbb{N}$. At the very least x must then be a real number. But note that if $x \neq 0$, then by choosing a natural number $n_0 > \frac{1}{|x|}$, we have $|x| > \frac{1}{n_0}$, and so $x \notin \left[\frac{-1}{n_0}, \frac{1}{n_0}\right]$. This shows that a non-zero real number $x \neq 0$ cannot belong to all of the sets $\left[\frac{-1}{n}, \frac{1}{n}\right]$, $n \in \mathbb{N}$, and thus does not belong to the intersection $x \notin \bigcap_{n=1}^{\infty} \left[\frac{-1}{n}, \frac{1}{n}\right]$. On the other hand, $x = 0$ belongs to each of the sets: $0 \in \left[\frac{-1}{n}, \frac{1}{n}\right]$ for all $n \in \mathbb{N}$. By definition of the intersection, then, $0 \in \bigcap_{n=1}^{\infty} \left[\frac{-1}{n}, \frac{1}{n}\right]$. We have thus showed that the intersection consists of the single element 0, i.e., $\bigcap_{n=1}^{\infty} \left[\frac{-1}{n}, \frac{1}{n}\right] = \{0\}$ as claimed.

Example I.8 (Another union and intersection of nested intervals)**.**
Consider now the example $B_n = \left(0, \frac{1}{n}\right)$ for $n \in \mathbb{N}$. Then the union and intersection of the collection $(B_n)_{n\in\mathbb{N}}$ are

$$\bigcup_{n=1}^{\infty} \left(0, \frac{1}{n}\right) = (0,1) \qquad \text{(open interval)}$$

$$\bigcap_{n=1}^{\infty} \left(0, \frac{1}{n}\right) = \emptyset \qquad \text{(empty set).}$$

The precise justifications of these claims are left as an exercise.

Exercise I.5 (Details of Example I.8)**.**
Provide careful reasoning to justify the claims made in Example I.8.

Example I.9 (Yet another union and intersection of nested intervals)**.**
Consider now the example $C_n = \left[0, 1 - \frac{1}{2n}\right]$ for $n \in \mathbb{N}$. Then the union and intersection of the collection $(C_n)_{n\in\mathbb{N}}$ are

$$\bigcup_{n=1}^{\infty} \left[0, 1 - \frac{1}{2n}\right] = [0,1) \qquad \text{(half-open interval)}$$

$$\bigcap_{n=1}^{\infty} \left[0, 1 - \frac{1}{2n}\right] = \left[0, \frac{1}{2}\right] \qquad \text{(closed interval).}$$

Note in particular that 1 does not belong to the union. The precise justifications are again left as an exercise.

Exercise I.6 (Details of Example I.9)**.**
Provide careful reasoning to justify the claims made in Example I.9.

Exercise I.7 (De Morgan laws for general unions and intersections)**.**
Let X be a set, $J \neq \emptyset$ a nonempty index set, and $A_j \subset X$ subsets for each $j \in J$. Prove De Morgan's laws for arbitrary unions and intersections:

$$X \setminus \bigcup_{j\in J} A_j = \bigcap_{j\in J} (X \setminus A_j) \qquad \text{and} \qquad X \setminus \bigcap_{j\in J} A_j = \bigcup_{j\in J} (X \setminus A_j).$$

Cartesian products

Another operation of set theory is forming Cartesian products of sets.

If A and B are sets, then their **Cartesian product** $A \times B$ is the set whose elements are **ordered pairs** (a, b) whose first member belongs to the former set, $a \in A$, and second member to the latter set, $b \in B$. In symbols, the Cartesian product is

$$A \times B = \left\{ (a, b) \;\middle|\; a \in A,\ b \in B \right\}. \tag{I.23}$$

A familiar example is the plane. A pair of real coordinates specifies a point in the plane, so the plane is the Cartesian product of the real line $\mathbb{R}$ with itself:

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \Big\{ (x,y) \;\Big|\; x \in \mathbb{R},\ y \in \mathbb{R} \Big\}.$$

Another familiar example is the rectangle, which (as a set) is the Cartesian product of two closed intervals

$$[\alpha,\beta] \times [\gamma,\delta] = \Big\{ (x,y) \;\Big|\; \alpha \le x \le \beta,\ \gamma \le y \le \delta \Big\}.$$

Also other interesting geometric shapes arise as Cartesian products of sets: the Cartesian product of a disc and an interval is a solid cylinder, and the Cartesian product of two circles is (a parametrization of) the torus (the surface of a donut).

Cartesian products of three sets are defined similarly as ordered triples $A \times B \times C = \big\{ (a,b,c) \;\big|\; a \in A,\ b \in B,\ c \in C \big\}$, etc. The familiar n-dimensional space is an n-fold Cartesian product of the real line with itself

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}} = \Big\{ (x_1, \ldots, x_n) \;\Big|\; x_1, \ldots, x_n \in \mathbb{R} \Big\}.$$

The elements of $\mathbb{R}^n$ are ordered n-tuples[13] of real numbers, consisting of the n coordinates of a point in this space.

One can consider Cartesian products of arbitrary collections of sets, but in this course finite Cartesian products, such as the ones above, should be sufficient.

I.2. Functions

What is a function, precisely? According to the familiar operational description, a function f associates a (single) value $f(x)$ to each argument x. For a precise meaning, we must specify what is the set of acceptable arguments, and what is the set of possible values. The mathematical terms for these two sets are the **domain** of the function and the **codomain**[14] of the function, respectively.

The notations

$$f : X \to Y \qquad \text{or} \qquad X \xrightarrow{f} Y$$

indicate that f is a **function**, whose **domain** is a set X, and whose **codomain** is a set Y, i.e., to each $x \in X$ a value $f(x) \in Y$ is assigned by f. The notation

$$x \mapsto f(x) \qquad \text{(read: “}x\text{ maps to }f(x)\text{”)}$$

[13] One uses the terms **pair**, **triple**, **quadruple**, **quintuple**, **sextuple**, ... for ordered collections of two, three, four, five, six, ... elements. For general n it has become conventional to refer to an ordered collection of n elements as an **n-tuple**.

[14] Sometimes the term **range** is used instead of *codomain*. Beware, however, as it is also common to use the term *range* to mean the set of actually attained values of the function, which may be smaller than the set of possible values that we are a priori willing to consider. The Finnish terms ”määrittelyjoukko” and ”arvojoukko” are often used for the domain and codomain, and the latter unfortunately suffers from exactly the same ambiguity of usage as the term *range*.

The terms *domain* and *codomain* are in my opinion not very illuminating, but they are well-established mathematical terminology. Were they not, I would advocate for using *source* and *target* instead. In Finnish, following [**Väi99**], I would recommend the unambiguous and descriptive terms ”lähtö”/”lähtöjoukko” and ”maali”/”maalijoukko”.

may be used to emphasize this assignment (but the domain X and the codomain Y must also be specified or should be otherwise clear from the context). We may view this as a mapping from the points x of the domain X to points of the codomain Y; the term **mapping** is considered synonymous to the term **function**. Functions are often defined by giving a formula (but also any other way that ensures that a unique value is assigned to each argument is acceptable).

Example I.10 (An example of a function).
We may define a function

$$r \colon \mathbb{R}^3 \to [0, +\infty)$$

by the formula

$$r((x, y, z)) = \sqrt{x^2 + y^2 + z^2} \qquad \text{for } (x, y, z) \in \mathbb{R}^3.$$

Here, the domain of the function r was chosen as $\mathbb{R}^3$, and the codomain was chosen as $[0, +\infty)$. We must only make sure that for each $(x, y, z) \in \mathbb{R}^3$, the formula above is meaningful (as it is), and gives a value in the set $[0, +\infty)$ (as it does).

Often giving a name or a symbol for a function is purposeful: the letters $f, g, \ldots$ are most commonly used, r was chosen in Example I.10, and you have without a doubt seen also

$$\log \colon (0, \infty) \to \mathbb{R}, \qquad \sin \colon [0, 2\pi) \to [-1, 1], \qquad \text{etc.}$$

Occasionally, it is more meaningful to talk of functions $X \to Y$ (i.e., functions with domain X and codomain Y) without explicitly naming them. The "maps to" notation "$\mapsto$" may then be particularly convenient.

Example I.11 (A few more examples of functions).

- The exponential function $\mathbb{R} \to (0, +\infty)$ is given by the formula $x \mapsto e^x$.
- The formula $\theta \mapsto (\cos(\theta), \sin(\theta))$ defines a function $(-\pi, +\pi] \to \mathbb{R}^2$.

Sometimes we simply consider established symbols as the names of functions, and indicate by a "dot" at which slot the argument is to be inserted to get the value.

Example I.12 (Yet a few more examples of functions).

- The square root function $\sqrt{\cdot} \colon [0, +\infty) \to [0, +\infty)$ is given by $x \mapsto \sqrt{x}$.
- The absolute value function $|\cdot| \colon \mathbb{R} \to [0, +\infty)$ is given by $x \mapsto |x|$.

Example I.13 (The identity function).
Let X be a set. Then the function

$$\mathrm{id}_X \colon X \to X \qquad \mathrm{id}_X(x) = x \quad \text{for } x \in X \tag{I.24}$$

is known as the **identity function** on X.

Note that functions defined by the same formula are not the same functions unless also their domains and codomains are the same!

Example I.14 (Different functions given by the same formula).
The following functions are given by the rule $x \mapsto x^2$:

- $f_1 \colon \mathbb{R} \to \mathbb{R}$ given by $f_1(x) = x^2$,
- $f_2 \colon \mathbb{R} \to [0, +\infty)$ given by $f_2(x) = x^2$,

- $f_3\colon [0,+\infty) \to \mathbb{R}$ given by $f_3(x) = x^2$,
- $f_4\colon (-\infty,0] \to \mathbb{R}$ given by $f_4(x) = x^2$,
- $f_5\colon (-\infty,0] \to [0,+\infty)$ given by $f_5(x) = x^2$.

Nevertheless, each of the above is a different function: $f_1 \neq f_2$, $f_3 \neq f_5$, etc.

The examples we have used so far should appear unintimidating. The concept of a function is nevertheless very general: the domain and range of a function may be arbitrary sets. We will in particular encounter functions whose arguments are functions (i.e., the domain is some suitable set of functions), functions whose values are sets (i.e., the codomain is some set whose elements are sets), etc. Be forewarned.

Surjective, injective, and bijective functions

The following properties of functions are often important.

Definition I.15 (Surjectivity, injectivity, and bijectivity)**.**
Let $f\colon X \to Y$ be a function.

We say that f is **surjective** if for every $y \in Y$ there exists an $x \in X$ such that $y = f(x)$.

We say that f is **injective** if for any $x_1, x_2 \in X$ which are different, $x_1 \neq x_2$, also the corresponding function values are different, $f(x_1) \neq f(x_2)$.

We say that f is **bijective** if it is both surjective and injective.

Consider a function $f\colon X \to Y$ and a given $y \in Y$. If f is injective, there can be at most one $x \in X$ such that $f(x) = y$ (since for any $x' \neq x$ we have $f(x') \neq f(x) = y$). For a surjective function there always exists at least one such x. For bijective functions, therefore, there exists exactly one such x, i.e., the condition $y = f(x)$ gives rise to a well-defined mapping $y \mapsto x$, the **inverse function**

$$f^{-1}\colon Y \to X.$$

Exercise I.8 (Examples of injectivity and surjectivity)**.**
Consider the functions

$$f_1\colon \mathbb{R} \to \mathbb{R}, \qquad f_2\colon \mathbb{R} \to [0,+\infty), \qquad f_3\colon [0,+\infty) \to \mathbb{R},$$
$$f_4\colon (-\infty,0] \to \mathbb{R}, \qquad f_5\colon (-\infty,0] \to [0,+\infty),$$

from Example I.14, each given by the formula $x \mapsto x^2$. Check that:

(a) The functions f_2 and f_5 are surjective, whereas f_1, f_3, and f_4 are not.
(b) The functions f_3, f_4, and f_5 are injective, whereas f_1 and f_2 are not.
(c) The function f_5 is bijective, whereas f_1, f_2, f_3, and f_4 are not.
(d) The inverse function $f_5^{-1}\colon [0,+\infty) \to (-\infty,0]$ of f_5 is given by $f_5^{-1}(y) = -\sqrt{y}$, whereas f_1, f_2, f_3, and f_4 do not have inverse functions.

Example I.16 (The identity function is bijective)**.**
For any set X, the identity function $\mathrm{id}_X\colon X \to X$ of Example I.13 is bijective. It is its own inverse function[15]: $\mathrm{id}_X^{-1} = \mathrm{id}_X$.

[15] The property of the identity function that it is its own inverse may seem special, but there are also other functions satisfying this. Such functions are called *involutions*. We leave it to the

Compositions of functions

Suppose that

$$f\colon X \to Y \qquad \text{and} \qquad g\colon Y \to Z$$

are two functions. The setup is such that the codomain of f is the same set as the domain of g (both are Y). Then we may form a new function,

$$g \circ f : X \to Z$$

called the **composition** of f and g by the rule that at a point $x \in X$ of its domain the value is

$$\big(g \circ f\big)(x) \;=\; g\big(f(x)\big) \in Z.$$

The composition is illustrated in the following diagram:[16]

$$\begin{tikzcd} X \arrow[r, "f"] \arrow[rr, bend right, "g\circ f"'] & Y \arrow[r, "g"] & Z \end{tikzcd}$$

$$x \overset{f}{\longmapsto} f(x) \overset{g}{\longmapsto} g\big(f(x)\big)\;.$$

Example I.17 (A composition of functions)**.**
Consider the functions

$$f\colon\; (0,2\pi) \to [-1,1) \qquad\qquad f(\theta) = \cos(\theta) \quad \text{for } \theta \in (0,2\pi)$$

$$g\colon\; [-1,1) \to \mathbb{R} \qquad\qquad g(u) = \frac{1}{(1-u)^2} \quad \text{for } u \in [-1,1).$$

Their composition is

$$g \circ f\colon\; (0,2\pi) \to \mathbb{R} \qquad\qquad \big(g\circ f\big)(\theta) = \frac{1}{\big(1-\cos(\theta)\big)^2} \quad \text{for } \theta \in (0,2\pi).$$

Example I.18 (Compositions with an inverse function)**.**
Suppose that

$$f\colon\; X \to Y$$

is bijective, and consider its inverse function

$$f^{-1}\colon\; Y \to X.$$

Then the composition of f and f^{-1} is (as you should verify from the definitions)

$$f^{-1} \circ f\colon\; X \to X \qquad\qquad \big(f^{-1}\circ f\big)(x) = x \quad \text{for } x \in X.$$

The composition of f^{-1} and f, on the other hand, is (as you should also verify)

$$f \circ f^{-1}\colon\; Y \to Y \qquad\qquad \big(f\circ f^{-1}\big)(y) = y \quad \text{for } y \in Y.$$

So both compositions are identity functions (see Example I.13) but on different sets:

$$f^{-1}\circ f \;=\; \mathrm{id}_X \qquad\qquad f \circ f^{-1} \;=\; \mathrm{id}_Y.$$

---

reader to construct examples: the formulas $x \mapsto -x$, $y \mapsto \frac{1}{y}$, $z \mapsto \frac{-1}{z}$, $\beta \mapsto -\frac{1}{2}\log\big(\tanh(\beta)\big)$ may serve as an inspiration, but remember to specify the domains and codomains!

[16]The line below the diagram (with the $\mapsto$ -symbols) should make it clear why the composition of f and g is denoted by $g\circ f$ in this order. We take an element $x \in X$, first apply f to it to get $f(x) \in Y$, and then apply g to the result to get $g\big(f(x)\big) \in Z$. Some people prefer to read the composition symbol $\circ$ as "after", so $g\circ f$ is read as "g after f". This succinctly indicates the order in which the two functions are applied.

Exercise (✓) I.9 (Composition of injective functions is injective)**.**
Show that if both $f\colon X \to Y$ and $g\colon Y \to Z$ are injective, then $g \circ f\colon X \to Z$ is also injective.

Exercise (✓) I.10 (Composition of surjective functions is surjective)**.**
Show that if both $f\colon X \to Y$ and $g\colon Y \to Z$ are surjective, then $g \circ f\colon X \to Z$ is also surjective.

Exercise I.11 (Injectivity of composition)**.**
Let $f\colon X \to Y$ and $g\colon Y \to Z$ be functions such that $g \circ f\colon X \to Z$ is injective.

(a) Show that f is also injective.
(b) Construct an example which shows that g does not have to be injective.

Exercise I.12 (Surjectivity of composition)**.**
Let $f\colon X \to Y$ and $g\colon Y \to Z$ be functions such that $g \circ f\colon X \to Z$ is surjective.

(a) Show that g is also surjective.
(b) Construct an example which shows that f does not have to be surjective.

The following exercise shows that the conditions in Example I.18 characterize the inverse function.

Exercise I.13 (Characterization of the inverse function)**.**
Suppose that $f\colon X \to Y$ and $g\colon Y \to X$ are two functions such that

$$g \circ f \;=\; \mathrm{id}_X \qquad\qquad f \circ g \;=\; \mathrm{id}_Y .$$

Prove that the both f and g are bijective, and we have $g = f^{-1}$ and $f = g^{-1}$.

Restricting the domain

Given a function $f\colon X \to Y$ and a subset $A \subset X$ of its domain, it is possible to consider a function defined only in the subset by otherwise the same formula. This function $A \to Y$ is denoted by $f|_A$ and is called the **restriction** of f to the subset A; it is given by

$$\begin{aligned} f|_A\colon\; & A \to Y \\ f|_A(a) = f(a) & \qquad \text{for } a \in A. \end{aligned} \tag{I.25}$$

Example I.19 (Examples of restrictions of functions)**.**
Consider again the functions

$$f_1\colon \mathbb{R} \to \mathbb{R}, \qquad f_2\colon \mathbb{R} \to [0,+\infty), \qquad f_3\colon [0,+\infty) \to \mathbb{R},$$
$$f_4\colon (-\infty,0] \to \mathbb{R}, \qquad f_5\colon (-\infty,0] \to [0,+\infty),$$

from Example I.14, each given by the formula $x \mapsto x^2$. Then we have

$$f_3 = f_1|_{[0,+\infty)}, \qquad f_4 = f_1|_{(-\infty,0]}, \qquad f_5 = f_2|_{(-\infty,0]}.$$

Images and preimages under functions

Let $f\colon X \to Y$ be a function.

If $A \subset X$ is a subset of the domain of f, then the **image** of A under f is the subset

$$f[A] \;=\; \big\{ f(a) \;\big|\; a \in A \big\} \tag{I.26}$$

of the codomain of f consisting of those points $y \in Y$ which are obtained as the value at some point $a \in A \subset X$, i.e., $y = f(a)$.

Example I.20 (Some images).
Consider the function

$$f\colon \mathbb{R} \to \mathbb{R} \qquad \text{given by} \qquad f(x) = x^2 \quad \text{for } x \in \mathbb{R}.$$

The image of the subset $A = \{-7, -5, 5, 7\} \subset \mathbb{R}$ is

$$f\big[\{-7, -5, 5, 7\}\big] \;=\; \{25, 49\}\,.$$

In particular the image of a subset may contain fewer elements than the subset itself.

Example I.21 (The image of a singleton).
The image of a singleton $\{x\} \subset X$ is $f\big[\{x\}\big] = \{f(x)\}$.

Exercise (✓) I.14 (Singleton image).
If the image of $A \subset X$ is a singleton $f[A] = \{y\}$, then what can be said about the behavior of the function f on the subset A?

Example I.22 (Surjectivity by images).
A function $f\colon X \to Y$ is surjective if and only if the image of the whole domain is the whole codomain, $f[X] = Y$.

If $B \subset Y$ is a subset of the codomain of f, then the **preimage** of B under f is the subset

$$f^{-1}[B] \;=\; \big\{x \in X \;\big|\; f(x) \in B\big\} \tag{I.27}$$

of the domain of f consisting of those points $x \in X$ at which the value belongs to the subset $B \subset Y$. The notation f^{-1} is used for both the inverse function and the preimage, but we try to consistently use square brackets in the latter case.[17] The inverse function only exists if f is a bijection, whereas preimages make sense for arbitrary functions f.

Example I.23 (Some preimages).
Consider the function

$$f\colon \mathbb{R} \to \mathbb{R} \qquad \text{given by} \qquad f(x) = x^2 \quad \text{for } x \in \mathbb{R}.$$

The preimage of the (singleton) subset $A = \{5\} \subset \mathbb{R}$ is

$$f^{-1}\big[\{5\}\big] \;=\; \big\{-\sqrt{5},\ \sqrt{5}\,\big\}.$$

In particular the preimage of a singleton can contain more than one element.

The preimage of the (singleton) subset $A = \{-5\} \subset \mathbb{R}$ is the empty set,

$$f^{-1}\big[\{-5\}\big] \;=\; \emptyset.$$

In particular the preimage of a nonempty set can be empty.

[17] In most mathematical literature ordinary parentheses are used also for preimages, or parentheses are altogether omitted. You are in any case expected to figure out which one is which, since the inverse function is applied to *elements* of the codomain, whereas in the preimage is applied to *subsets* of the codomain.

Similarly, we try to consistently use square brackets for the images of subsets under a function. Elsewhere in literature ordinary parentheses or no parentheses at all are used. Again you are in any case expected to figure out which one is which, since the function is applied to *elements* of the domain, whereas in the image is applied to *subsets* of the domain.

Example I.24 (Injectivity by preimages).
A function $f\colon X \to Y$ is injective if and only if the preimage of every singleton $\{y\} \subset Y$ is either a singleton or the empty set.

Example I.25 (Bijectivity by preimages).
A function $f\colon X \to Y$ is bijective if and only if the preimage of every singleton $\{y\} \subset Y$ is a singleton.

Exercise I.15 (Images and preimages of unions and intersections).
Let X and Y be sets and $f\colon X \to Y$ a function.

(a) Show that for any $C, D \subset Y$, the preimages satisfy

$$f^{-1}[C \cup D] = f^{-1}[C] \cup f^{-1}[D].$$

(b) Show that for any $C, D \subset Y$, the preimages satisfy

$$f^{-1}[C \cap D] = f^{-1}[C] \cap f^{-1}[D].$$

(c) Show that for any $A, B \subset X$, the images satisfy

$$f[A \cup B] = f[A] \cup f[B].$$

(d) Give an example in which for the images of subsets $A, B \subset X$ we have

$$f[A \cap B] \neq f[A] \cap f[B].$$

Hint: *In parts (a)–(c) it is possible to argue by a chain of equivalent conditions*

$$x \in \textit{left-hand side set} \quad \Longleftrightarrow \quad \cdots \quad \Longleftrightarrow \quad x \in \textit{right-hand side set.}$$

A comparison of (b) and (d) above indicates that preimages behave better with set operations than images. Another example is the following.

Exercise I.16 (Images and preimages of complements).
Let X and Y be sets and $f\colon X \to Y$ a function.

(a) Suppose that $C \subset D \subset Y$. Show that the preimages satisfy

$$f^{-1}[D \setminus C] \;=\; f^{-1}[D] \setminus f^{-1}[C].$$

(b) Give an example in which for the images of subsets $A \subset B \subset X$ we have

$$f[B \setminus A] \;\neq\; f[B] \setminus f[A].$$

The formulas of Exercise I.15(a)–(c) generalize to arbitrary unions and intersections (more than two sets), and it is easy to modify the proofs to account for this general case. Specifically, if $f\colon X \to Y$ is a function and $B_j \subset Y$ for $j \in J$, then the preimages satisfy

$$f^{-1}\Big[\bigcup_{j\in J} B_j\Big] \;=\; \bigcup_{j\in J} f^{-1}[B_j], \qquad\qquad f^{-1}\Big[\bigcap_{j\in J} B_j\Big] \;=\; \bigcap_{j\in J} f^{-1}[B_j].$$

Likewise if $A_j \subset X$ for $j \in J$, then the images satisfy

$$f\Big[\bigcup_{j\in J} A_j\Big] \;=\; \bigcup_{j\in J} f[A_j].$$

I.3. Logic and related notation

Mathematics in general concerns statements that can be logically proven to be true, starting from some specified assumptions. In other words, we want to deduce conclusions of interest from known premises. The rules for valid deduction are the subject of logic. Logic itself could be formalized, but for our purposes it suffices to introduce a little bit of notation and typical examples. The main message is that in this course and in all mathematics, we must make unambiguous definitions and statements, and we must provide valid arguments to justify the statements made.

Quantifiers

Statements involving free variables

Commonly, statements (predicates, logical formulas) involve one or more variables, and whether the statements are true or false depends on those variables. A statement may be thought of as a function of those variables, whose possible values are true and false.

Example I.26 (An example statement depending on a variable)**.**
The statement

$$\text{“}k \text{ is even”}$$

is true if the variable k has the value $k = -8$, but it is false if $k = 33$.

Those variables, of which the truth value of a statement depends, are called **free variables**. Without specifying values for the free variables, the statement itself is neither true nor false. A statement without free variables is called a **sentence**, and a sentence is either true or false as it stands. Self-contained meaningful assertions must be sentences!

In mathematical text we often indicate that we (temporarily) fix the values of (some) variables by expressions such as: “fix $n \in \mathbb{N}$” (after which the varible n is no longer thought of as free, but having some a priori arbitrary but fixed natural number value) or “let $\varepsilon > 0$” (after which the varible ε is no longer thought of as free, but fixed to some a priori arbitrary positive real value). Fixing a value for a variable in effect makes it a named constant.[18]

Statements involving quantifiers

The following two example statements,

$$\text{“for all real numbers } x \text{ we have } x^2 \geq 0\text{”}$$

and

$$\text{“there exists a real number } x \text{ such that } x^2 = -1\text{”},$$

[18]This is the way a computer treats variables when *executing* (i.e., *running*) a program: at any time they are used, they must have assigned values or otherwise errors arise. On the other hand, when *writing* a program, the computer scientist or a programmer may think of variables much like a mathematician or a logician would.

both involve a variable x which has not been fixed, but the statements are in fact sentences. In particular they have definite truth values: the former sentence is true and the latter is false. The variable x in the above formulas is not free, because a *quantifier* applies to it: each statement above makes a claim about the *quantity* of values of x for which another statement, which involves x as a free variable, is true.

There are two quantifiers in ordinary logic. The **universal quantifier** denoted by the symbol $\forall$ means that the statement that follows it is true *for all values of the variable*; in natural language its symbol $\forall$ is read as "for all". The **existential quantifier** denoted by the symbol $\exists$ means that the statement that follows it is true *for at least one value of the variable*; in natural language its symbol $\exists$ is read as "there exists". Both quantifiers must be followed first by the variable which they quantify, and they should (in careful usage) also be followed by a specification of the set in which the variable is allowed to take values, and then another statement (possibly) involving the variable. For clarity we separate the other statement (in which the variable appears as free) by a colon (:). For example the two example statements that we started with can be written as

$$\forall\, x \in \mathbb{R} : \quad x^2 \geq 0$$

and

$$\exists\, x \in \mathbb{R} : \quad x^2 = -1.$$

Example I.27 (Revisiting subsets)**.**
The meaning of the subset relation $A \subset B$ is, in concise logical notation with quantifiers,

$$\forall\, x \in A : \quad x \in B.$$

Example I.28 (Revisiting unions and intersections)**.**
Let $(A_j)_{j \in J}$ be a collection of sets. Then the meaning of the statement that an element x belongs to the union of this collection, $x \in \bigcup_{j \in J} A_j$, can be expressed in concise logical notation with quantifiers as

$$\exists\, j \in J : \quad x \in A_j.$$

Similarly, if the collection is nonempty ($J \neq \emptyset$), then the meaning of the statement that an element x belongs to the intersection of this collection, $x \in \bigcap_{j \in J} A_j$, can be expressed as

$$\forall\, j \in J : \quad x \in A_j.$$

Note that the quantified variable appears as a free variable in the statement that follows, but that statement may itself have quantifiers for other variables. This is in fact very common in, e.g., definitions related to metric space topology (continuity, limits of sequences, etc.) — we will see plenty of examples later. For now, as an example statement with multiple quantifiers, consider

$$\forall\, \alpha \in \mathbb{R} : \quad \exists\, C > 0 : \quad \forall\, x \in \mathbb{R} : \quad \left| e^{3+5\alpha x - x^2} \right| \leq C$$

(this sentence is true, by the way). Literally this statement reads "for all real numbers α there exists a positive number C such that for all real numbers x we have $|e^{3+5\alpha x - x^2}| \leq C$", and allowing for a bit more liberties of expression and interpretation could also be read "whatever the value of the real paramenter α, the expression $e^{3+5\alpha x - x^2}$ involving a real variable x is bounded in absolute value by some positive constant C". The use of logical notation makes the statements more concise and unambiguous. By contrast, in natural language one has to be quite careful to achieve an unambiguous meaning.[19]

Note that the order of quantifiers matters! For instance the statement

$$\forall\, x \in \mathbb{R} : \quad \exists\, C > 0 : \quad \forall\, \alpha \in \mathbb{R} : \quad \left| e^{3+5\alpha x - x^2} \right| \leq C$$

is false, although it "only" differs from the true statement of the previous example by the order of quantifiers.

Exercise I.17 (Check the truth values of the above examples).
Prove that the statement $\forall\, \alpha \in \mathbb{R} : \exists\, C > 0 : \forall\, x \in \mathbb{R} : |e^{3+5\alpha x - x^2}| \leq C$ is true and that the statement $\forall\, x \in \mathbb{R} : \exists\, C > 0 : \forall\, \alpha \in \mathbb{R} : |e^{3+5\alpha x - x^2}| \leq C$ is false.

Implications and equivalences

Implications

Suppose that P and Q are two statements (predicates) possibly involving free variables (the same variables in both). Then one can form a new statement out of them, called an **implication** and denoted by

$$P \implies Q,$$

which in natural language can be read in any of the following ways

$$\begin{aligned} P \text{ implies } Q \;; \\ \text{if } P\text{, then } Q \;; \\ P \text{ only if } Q \;; \\ P \text{ is a sufficient condition for } Q \;. \end{aligned}$$

The natural-language descriptions (the most transparent of which is probably "if P, then Q") should already explain exactly how the truth value of an implication is determined: $P \Rightarrow Q$ is true if either P and Q are both true, or if P is false.[20]

Example I.29 (Revisiting subsets again).
The subset relation $A \subset B$ simply means that for any x we have the implication

$$x \in A \implies x \in B$$

A typical use of an implication is in theorem statements: P may represent the assumptions of a theorem (the hypotheses / the premise) and Q the conclusion, so the theorem statement itself is the implication $P \Rightarrow Q$ ("if the assumptions hold, then also the conclusion holds").

[19] This does not mean that natural language is strictly worse. Among advantages of natural language are that it allows to emphasize certain aspects of the statement and to draw attention to interpretations (for example interpreting α as a parameter, C as a constant, and x as a variable in the above expression). None of this affects the validity of logical statements, but to claim that such things are irrelevant would be rather extreme...

[20] In particular, if P does not hold, the implication is claiming nothing whatsoever about the truth or falsity of Q. The only situation in which the implication is false is if P is true but Q is nevertheless false.

Note, however, that the implication $P \Rightarrow Q$ is merely a statement about the relationship of the truth values of the two statements P and Q — it does not signify any *causal* relationship.

Instead of $P \Rightarrow Q$ we also occasionally switch the places of P and Q and invert the arrow direction, i.e., we alternatively write

$$Q \quad \Longleftarrow \quad P,$$

which in natural language can be read in any of the following ways

$$\begin{aligned} Q \text{ is implied by } P \; ; \\ Q \text{ if } P \; ; \\ Q \text{ is a necessary condition for } P \; . \end{aligned}$$

We have by now given altogether seven different phrases in natural language for the implication $P \Rightarrow Q$ (and we have still restricted ourselves to just unambiguous expressions that are common in mathematical text). The redundancy at least allows for different ways of explaining something, but ultimately the logical meaning of each of these phrases is exactly the same.

Equivalences

Two statements P and Q, possibly depending on some free variables (same in both), are said to be **logically equivalent** if they have the same truth value for all values of the free variables. This precisely amounts to requiring that both implications $P \Rightarrow Q$ and $P \Leftarrow Q$ hold (Q is true if P is true and vice versa). We denote **logical equivalence** by

$$P \quad \Longleftrightarrow \quad Q,$$

and in natural language read it as

$$\begin{aligned} P \text{ and } Q \text{ are (logically) equivalent } ; \\ P \text{ if and only if } Q \; ; \\ P \text{ is a necessary and sufficient condition for } Q \; . \end{aligned}$$

Example I.30 (Revisiting the equality of sets)**.**
The meaning of the equality of two sets $A = B$ is that for any x we have

$$x \in A \quad \Longleftrightarrow \quad x \in B.$$

Remark I.31 (Warning: a convention used in definitions)**.**
When a new term is defined, we usually say that the term is used if some conditions are satisfied. In this context we actually mean that the term is used *if and only if* the conditions are satisfied! But it is conventional in definitions to only state the "if" part, leaving it implicit that the newly introduced term is *not* used unless the conditions are satisfied.

For example our definition of a bijective function was: "we say that f is bijective if it is both injective and surjective". We of course *also* meant that we *do not* call f bijective *unless* it is both injective and surjective.

So while the natural language word "if" in usual mathematical text signifies just one implication, in definitions it is conventionally used to signify equivalence.

Negations

If P is a logical statement (predicate), its **negation**

$$\neg P$$

is the logical statement which always has the opposite truth value: if P is true, then $\neg P$ is false and if P is false, then $\neg P$ is true. In natural language, $\neg$ is read as "not".

Clearly the statement $\neg(\neg P)$ is logically equivalent to P (a double negation).

We have already introduced specific symbols for the negations of a few common statements, e.g.,

$$\begin{array}{lcl} \neg\big(a = b\big) & \text{is denoted} & a \neq b \\ \neg\big(a \in A\big) & \text{is denoted} & a \notin A \\ \neg\big(A \subset B\big) & \text{is denoted} & A \not\subset B. \end{array}$$

Contrapositives

The implications

$$\begin{aligned} P &\implies Q \\ \neg Q &\implies \neg P \end{aligned} \tag{I.28}$$

are logically equivalent — you should make sure that you understand why![21] The second implication is known as the **contrapositive** of the first.

Let us take an example from Section I.2.

Example I.32 (Revisiting injectivity).
Recall definition of an injective function $f : X \to Y$ from Definition I.15. In logical symbols, injectivity means

$$\forall\, x_1, x_2 \in X : \qquad x_1 \neq x_2 \implies f(x_1) \neq f(x_2). \tag{I.29}$$

An equivalent form of this condition can be obtained by taking a contrapositive of the implication above. This is particularly helpful, since the contrapositive will involve negations of the propositions $x_1 \neq x_2$ and $f(x_1) \neq f(x_2)$, which are simply $x_1 = x_2$ and $f(x_1) = f(x_2)$, respectively. Using the contrapositive form, we find that injectivity equivalently means

$$\forall\, x_1, x_2 \in X : \qquad f(x_1) = f(x_2) \implies x_1 = x_2. \tag{I.30}$$

The original formulation (I.29) of injectivity can be read: "At any two different points, the function has different values". The formulation (I.30) obtained by contrapositive can be read: "Whenever the values at two points are the same, the points themselves must be the same". With a bit of thinking, one easily convinces oneself about the equivalence of these, but routine use of logical symbols clearly facilitates correct rewriting.

The equivalence of the two implications in (I.28) in particular underlies the idea of a *proof by contrapositive* or an *indirect proof* — of which the *proof by contradiction* is essentially a special case.

[21] The most straightforward way to convince oneself about this is to write a truth table, i.e., to consider the truth value of both implications for all the four possibilities of the truth values of the propositions P and Q (P false and Q false, P false and Q true, P true and Q false, P true and Q true).

To elaborate, suppose our goal is to prove that from assumption P, conclusion Q follows. Another way of doing that is to prove that if we assume that the desired conclusion Q does not hold, then it implies that the assumption P cannot hold, either. This is the proof by contrapositive.

In particular, if there are no (explicit) assumptions, we can consider the assumptions represented by an identically true proposition P. Then this proof strategy is proof by contradiction. Indeed, by showing the contrapositive, we find that if the desired conclusion Q would not hold, then the negation of the identically true assumption P would hold, which is absurd; a contradition. The proof of Theorem I.5 is a classic example.

Negations and quantifiers

Since typical definitions in metric space topology involve many logical quantifiers, it is essential to be comfortable working with them. In particular one has to be able to routinely form the negation of a statement involving a quantifier.

Consider a statement involving the universal quantifier, say

$$\forall\, x \in X : \quad P(x).$$

Its negation, $\neg\,\big(\forall\, x \in X :\ P(x)\big)$, is logically equivalent with

$$\exists\, x \in X : \quad \neg\, P(x);$$

you should make sure you understand why!

A common use of this observation is that to *disprove* the validity of a claim (for example a belief that we might have) that $P(x)$ holds for all $x \in X$, it is sufficient to find just one example of an $x \in X$ for which $P(x)$ is false. Such an example is called a **counterexample**, because it is sufficient to invalidate the whole original claim (which started with the universal quantifier).

And even if we are seeking to prove rather than to disprove a claim of this form, in an indirect proof (contrapositive) we would first form the negation.

Symmetrically, consider a statement involving the existential quantifier, say

$$\exists\, x \in X : \quad Q(x).$$

Its negation, $\neg\,\big(\exists\, x \in X :\ Q(x)\big)$, is logically equivalent with (think about it)

$$\forall\, x \in X : \quad \neg\, Q(x).$$

On the surface this does not appear as practical as the negation of a statement with a universal quantifier. But in indirect proofs, or when disproving a claim (say a false belief) of this form, we would have to form the negation. Another typical use of this appears when a statement involves many quantifiers; we may want to consider the negation to get rid of some universal quantifiers, but the existential quantifiers are affected as well.

Example I.33 (There are arbitrarily large natural numbers)**.**
The logical statement

$$\forall\, q \in \mathbb{Q} : \quad \exists\, n \in \mathbb{N} : \quad n > q \tag{I.31}$$

has the interpretation that there are arbitrarily large natural numbers ("for any rational number q there exists a larger natural number n").

The statement (I.31) is true: it is the so called **Archimedean property** of the rational numbers. Let us examine a proof of this statement using a proof by contradiction.

For a proof by contradiction, one would suppose that (I.31) is false, and show that this leads to a contradiction. If (I.31) is false, then its negation

$$\neg \Big(\forall\, q \in \mathbb{Q} : \quad \exists\, n \in \mathbb{N} : \quad n > q \Big).$$

is true. The original statement (I.31) begins with a universal quantifier ($\forall$), and its negation takes the form

$$\exists\, q \in \mathbb{Q} : \quad \neg \big(\exists\, n \in \mathbb{N} : \quad n > q \big).$$

In this form, after the first existential quantifier ($\exists$) we have the negation of a statement which begins with an existential quantifier ($\exists$), which we may further unravel. The negation of (I.31) is thus written as

$$\exists\, q \in \mathbb{Q} : \quad \forall\, n \in \mathbb{N} : \quad \neg \big(n > q \big).$$

The negation $\neg\,(n > q)$ is obviously $n \leq q$, so the negation of (I.31) has been rewritten as

$$\exists\, q \in \mathbb{Q} : \quad \forall\, n \in \mathbb{N} : \quad n \leq q,$$

which in natural language claims that there exists some rational number q such that all natural numbers n are bounded from above by it. This clearly already sounds absurd, but we leave it to the reader to now derive the contradiction.[22] The point here was to illustrate how we typically end up considering negations of statements with quantifiers, and how they can be systematically unraveled.

I.4. ♡ In-depth topics related to set theory

The reader interested in deepening their understanding may wish to look into Appendix A.1 for some of the following topics related to the subject of this lecture:

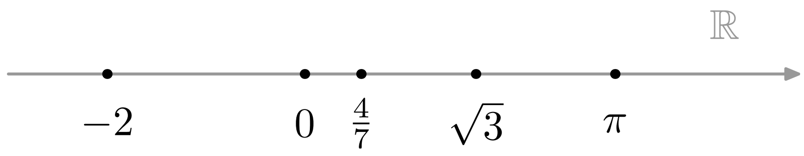
- the notion of cardinality as a way of comparing the sizes of sets;
- countable infinity, i.e., the smallest infinite cardinality;
- the uncountability of the set $\mathbb{R}$ of real numbers.

The notions of countable and uncountable cardinalities will be crucial in various advanced topics of mathematics, notably in measure theory and probability theory. Familiarity with these notions will be assumed in the corresponding courses.

[22] Hint: *If such a rational number $q \in \mathbb{Q}$ exists, write it as $q = \frac{m}{k}$ with $m \in \mathbb{Z}$ and $k \in \mathbb{N}$. Then use for example the natural number $n = |m| + 1$ as a counterexample to the statement $\forall\, n \in \mathbb{N} :\ n \leq q$.*

Lecture II

Real numbers



There may not be much of a feeling of novelty to the real numbers, but we spend a lecture on this topic for two reasons: to appreciate some subtleties of real numbers themselves, and to get concrete examples of some metric (and topological) properties and notions that will arise in more general contexts in the subsequent lectures.

We start by discussing the absolute value of real numbers and the related notion of distance[1] on the real line. This simple notion underlies for example the convergence of sequences of real numbers. We look at basic properties of convergence of real-number sequences.[2] We also show that both rational and irrational numbers are dense on the real line.

We then briefly turn to foundational questions again: what exactly are real numbers? A standard rigorous mathematical approach is to start from their axiomatic properties, most of which are entirely unsurprising. We only mention the other axioms in passing, but discuss the *completeness axiom* in more detail, since it is the most subtle and the most consequential for topology as well. It involves the notions of *supremum* (least upper bound) and *infimum* (greatest lower bound).

II.1. Absolute value, distances between numbers, and triangle inequality

At the core of topological considerations of numbers is the concept of distances on the real line. They are based on the absolute value — which itself is the distance of a given number to the origin.

The **absolute value** of a real number $x \in \mathbb{R}$ is defined by

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases} \tag{II.1}$$

Another equivalent expression for the absolute value is

$$|x| = \sqrt{x^2}, \tag{II.2}$$

[1] This will be, in a thin disguise, our first example of a metric, although the general definition will have to wait until Lecture V.

[2] The convergence of real-number sequences will turn out to be a special case of the convergence of sequences in metric spaces treated in Lecture VIII. But the case of real sequences is an instructive concrete example, which is important in its own right.

where $\sqrt{\cdot}$ denotes the non-negative square root of a non-negative number. Clearly the absolute value satisfies

$$|x| = \big|-x\big|, \qquad |x| \geq 0, \qquad x \leq |x|, \qquad -x \leq |x|$$

for any $x \in \mathbb{R}$, and

$$\big|xy\big| \;=\; |x|\,|y|$$

for any $x, y \in \mathbb{R}$. Also it is practical to note the following equivalent formulations of bounds on absolute values:

$$|x| < c \iff -c < x < c \qquad \text{and} \qquad |x| \leq c \iff -c \leq x \leq c.$$

The **distance**[3] between two real numbers $x, y \in \mathbb{R}$ is the absolute value of their difference,

$$|x - y|. \tag{II.3}$$

The following **triangle inequalities** are simple and intuitive results. But as we will see later from Lecture V on, a rather straightforward generalization of them serves as the basis of a very fruitful general theory of metric spaces.

Lemma II.1 (Triangle inequalities on $\mathbb{R}$)**.**
For all $x, y \in \mathbb{R}$ we have

$$\Big||x| - |y|\Big| \;\leq\; \big|x + y\big| \;\leq\; |x| + |y|. \tag{II.4}$$

Proof. There are two inequalities in (II.4). It turns out that the first of them can be obtained as a consequence of the second, so it is a good idea to prove the second inequality first.

To prove the second inequality, fix $x, y \in \mathbb{R}$. Calculate the square of the absolute value of the sum, and estimate each term from above by its absolute value to get

$$\begin{aligned} \big|x + y\big|^2 = \big(x + y\big)^2 &= x^2 + 2xy + y^2 \\ &\leq |x|^2 + 2|x||y| + |y|^2 \\ &= \Big(|x| + |y|\Big)^2. \end{aligned}$$

Taking square roots of the above ($\sqrt{\cdot}$ is an increasing function and thus respects the inequality), we obtain exactly the second asserted inequality

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

To prove the first inequality, we use the second one that we just proved above. Namely, we first apply the second inequality to the numbers $x + y$ and $-y$, to get

$$|x| \;=\; \big|(x + y) + (-y)\big| \;\leq\; |x + y| + |-y| \;=\; |x + y| + |y|.$$

Then by subtracting $|y|$ from both sides, we get

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

A similar application of the second inequality to the numbers $x + y$ and $-x$ (or just interchanging the roles of x and y above) yields

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

Combining these gives the first asserted inequality

$$\Big||x| - |y|\Big| \;\leq\; |x + y|,$$

[3] Later, in Lecture V, we will see that this notion of a distance (II.3) is a special case of a *metric*, and it makes the real line a *metric space*. For now, however, we will work exclusively with the concrete case of real numbers, and we will not refer to the general notion.

and the proof is complete. □

Another common form of the triangle inequality, which features the distance $|x-y|$ in exactly this form is

$$|x-y| \leq |x| + |y|. \tag{II.5}$$

It is obtained from the second inequality in (II.4) by plugging in x and $-y$, and noting $|-y| = |y|$.

There is a straightforward generalization of the triangle inequality to sums with finitely many terms.

Lemma II.2 (Triangle inequality for finite sums)**.**
Let $x_1, \ldots, x_n \in \mathbb{R}$. Then we have

$$\big|x_1 + \cdots + x_n\big| \;\leq\; |x_1| + \cdots + |x_n|. \tag{II.6}$$

Exercise II.1 (Proof of Lemma II.2)**.**
Prove Lemma II.2 by induction over $n \in \mathbb{N}$, using Lemma II.1.

Remark II.3 (Triangle inequality for infinite series)**.**
Let $x_1, x_2, x_3, \ldots \in \mathbb{R}$. The natural generalization of (II.6) to infinite series would be

$$\Big|\sum_{j=1}^{\infty} x_j\Big| \;\overset{?}{\leq}\; \sum_{j=1}^{\infty} |x_j|.$$

With suitable interpretations, this generalization is indeed valid. However, it immediately requires consideration of the convergence of the infinite series on both sides, and/or suitable interpretations in the case that one or both of the series diverge. For now, we will not try to use these types of generalizations, but we invite the reader to think about what can and cannot be said about them.

II.2. Sequences on the real line

Sequences of real numbers

A **sequence** of real numbers is an "infinite list"

$$(a_1, a_2, a_3, \ldots)$$

of real numbers $a_1, a_2, a_3, \ldots \in \mathbb{R}$. A precise definition of a sequence is a function

$$a \colon \mathbb{N} \to \mathbb{R},$$

and the "list" consists of the values of this function: $a_n = a(n)$ for $n \in \mathbb{N}$. We usually use notations such as

$$(a_1, a_2, a_3, \ldots) \;=\; (a_n)_{n=1}^{\infty} = (a_n)_{n\in\mathbb{N}}$$

for a sequence. Also just the lazy notation (a_n) is quite common, for instance when the indexing by natural numbers n is clear from the context.

Moreover, we occasionally let the indexing start from something other than $n = 1$. The notation is then modified in obvious ways: e.g., $(a_0, a_1, a_2, \ldots) = (a_n)_{n=0}^{\infty}$ or $(a_{25}, a_{26}, a_{27}, \ldots) = (a_n)_{n=25}^{\infty}$; again precisely interpreted as suitable functions, e.g.,

$\{0, 1, 2, \ldots\} \to \mathbb{R}$ or $\{25, 26, 27, \ldots\} \to \mathbb{R}$. The definitions in the following subsections must then be adapted in obvious ways. Some properties of interest (particularly limits) do not even depend on any finitely many initial members of the sequence, so for these the choice of the starting index is largely irrelevant.

Monotonicity properties of number sequences

The real line has an order. Sequences which respect that order have some particularly nice properties.

Definition II.4 (Monotonicity properties of sequences of real numbers)**.**
A real-number sequence $(a_n)_{n\in\mathbb{N}}$ is

- **increasing**, if we have[4] $a_{n+1} \geq a_n$ for all $n \in \mathbb{N}$;
- **strictly increasing**, if we have[5] $a_{n+1} > a_n$ for all $n \in \mathbb{N}$;
- **decreasing**, if we have[6] $a_{n+1} \leq a_n$ for all $n \in \mathbb{N}$;
- **strictly decreasing**, if we have[7] $a_{n+1} < a_n$ for all $n \in \mathbb{N}$;
- **monotone**, if it is either increasing or decreasing.

Example II.5 (An increasing sequence)**.**
The sequence $(1, 2, 4, 8, 16, 32, \ldots)$ is (strictly) increasing: its n:th member is given by the formula $a_n = 2^{n-1}$, and we have $a_{n+1} = 2^n > 2^{n-1} = a_n$ for all $n \in \mathbb{N}$.

Example II.6 (Another increasing sequence)**.**
Let $(a_n)_{n\in\mathbb{N}}$ be defined by the formula $a_n = \frac{n}{n+1}$ for $n \in \mathbb{N}$. We claim that this sequence is increasing.

One way to verify this is to calculate the difference of consequtive members,

$$a_{n+1} - a_n \;=\; \frac{n+1}{n+2} - \frac{n}{n+1} \;=\; \frac{(n+1)^2 - n(n+2)}{(n+2)(n+1)} \;=\; \frac{1}{(n+2)(n+1)} \;>\; 0,$$

which gives $a_{n+1} > a_n$ (by adding a_n to both sides in the inequality).

Another (perhaps easier?) way is to calculate the ratio of consequtive terms

$$\frac{a_n}{a_{n+1}} \;=\; \frac{n/(n+1)}{(n+1)/(n+2)} \;=\; \frac{n(n+2)}{(n+1)^2} \;=\; \frac{n^2+2n}{n^2+2n+1} \;<\; 1,$$

which gives $a_n < a_{n+1}$ (by multiplying both sides by the positive number a_{n+1}).

Example II.7 (A decreasing sequence)**.**
The sequence $(1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}, \frac{1}{25}, \ldots)$ is (strictly) decreasing: its n:th member is given by the formula $a_n = \frac{1}{n^2}$, and we have $a_{n+1} = \frac{1}{(n+1)^2} < \frac{1}{n^2} = a_n$ for all $n \in \mathbb{N}$.

Boundedness properties of number sequences

Definition II.8 (Boundedness properties of sequences of real numbers)**.**
A real-number sequence $(a_n)_{n\in\mathbb{N}}$ is

[4] In other words, a sequence is increasing if $a_1 \leq a_2 \leq a_3 \leq \cdots$.
[5] In other words, a sequence is strictly increasing if $a_1 < a_2 < a_3 < \cdots$.
[6] In other words, a sequence is decreasing if $a_1 \geq a_2 \geq a_3 \geq \cdots$.
[7] In other words, a sequence is strictly decreasing if $a_1 > a_2 > a_3 > \cdots$.

- **bounded from above**, if there exists some $u \in \mathbb{R}$ (called an **upper bound** for the sequence) such that $a_n \leq u$ for all $n \in \mathbb{N}$;
- **bounded from below**, if there exists some $\ell \in \mathbb{R}$ (called a **lower bound** for the sequence) such that $a_n \geq \ell$ for all $n \in \mathbb{N}$;
- **bounded**, if there exists some $r \in \mathbb{R}$ such that $|a_n| \leq r$ for all $n \in \mathbb{N}$.

Exercise II.2 (Boundedness from both above and below).
Prove that a real-number sequence $(a_n)_{n\in\mathbb{N}}$ is bounded if and only if it is both bounded from above and bounded from below.

Example II.9 (A sequence bounded from below).
The sequence $(1, 2, 4, 8, 16, 32, \ldots)$ is bounded from below by, for example, the constant $\ell = 1$.[8] To see this, note that the n:th member of the sequence is given by the formula $a_n = 2^{n-1}$, and we have $a_n = 2^{n-1} \geq 1 = \ell$ for all $n \in \mathbb{N}$.

This sequence is not bounded from above. Indeed for any $u \in \mathbb{R}$, we can find a natural number $n > \log_2(u) + 1$, and then $a_n = 2^{n-1} > 2^{\log_2(u)} = u$, so u does not work as an upper bound. Since the $u \in \mathbb{R}$ we tried was arbitrary, no upper bound exists for this sequence.

Example II.10 (A bounded sequence).
Let $(a_n)_{n\in\mathbb{N}}$ be defined by the formula $a_n = \frac{n}{n+1}$ for $n \in \mathbb{N}$. We claim that this sequence is bounded by, for example, the constant $r = 1$.[9] To see this, note that for any $n \in \mathbb{N}$ we have

$$|a_n| = \left|\frac{n}{n+1}\right| = \frac{n}{n+1} \leq 1 = r.$$

II.3. Limits of sequences on the real line

The definition of the limit of a sequence

Definition II.11 (Limit of a sequence of numbers).
A sequence $(a_n)_{n\in\mathbb{N}}$ of real numbers **converges** to a **limit** $\alpha \in \mathbb{R}$ if for every $\varepsilon > 0$ there exists an index $n_\varepsilon \in \mathbb{N}$ such that $|a_n - \alpha| < \varepsilon$ whenever $n \geq n_\varepsilon$.

If a sequence $(a_n)_{n\in\mathbb{N}}$ converges to a limit α in the sense of the above definition, we denote

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

Alternatively, it is common to say that $a_n \to \alpha$ as $n \to \infty$, or that the sequence $(a_n)_{n\in\mathbb{N}}$ tends to α (as $n \to \infty$).

Remark II.12 (Meaning of convergence as a logical statement).
Definition II.11 is written in plain English (a natural language), as usual. It nevertheless has a precise logical meaning, and it is instructive to unravel the definition using logical symbols (a formal language). The meaning of the statement $a_n \to \alpha$ as $n \to \infty$ is:

$$\forall \varepsilon > 0 : \quad \exists n_\varepsilon \in \mathbb{N} : \quad \forall n \geq n_\varepsilon : \quad |a_n - \alpha| < \varepsilon. \tag{II.7}$$

A few advantages of the formal statement (II.7) are that it is succinct, unambiguous, and understandable to mathematicians irrespective of whether their mother tongue is English, Finnish, Swedish, or some other natural language (being familiar with the logical symbols, you can easily "read it" in your native language).

---

[8] As a lower bound here we could equally well use $\ell = 0$ or $\ell = -7$ or indeed any number $\ell \leq 1$.
[9] As a bound here we could equally well use $r = 42$ or $r = 10^{23}$ or indeed any number $r \geq 1$.

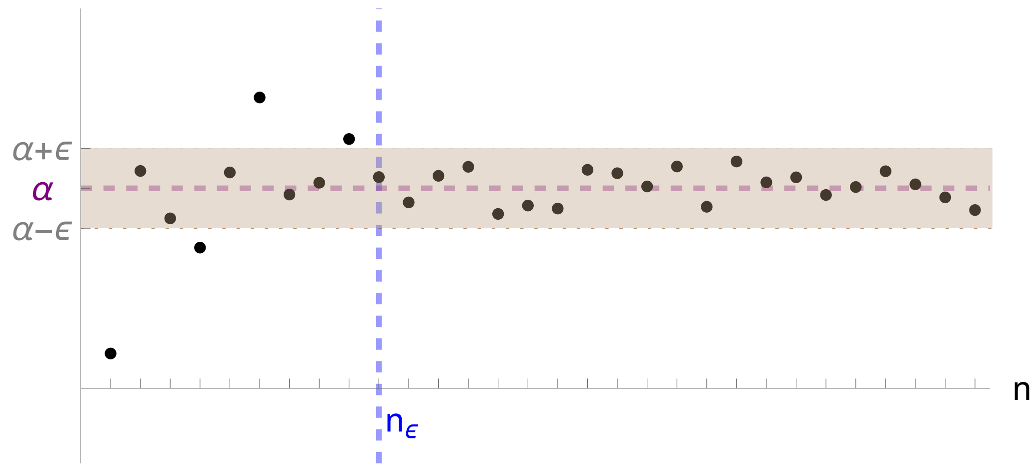


FIGURE II.1. An illustration of the limit of a number sequence. The limit of the sequence $(a_n)_{n\in\mathbb{N}}$ is $\alpha \in \mathbb{R}$ if for however small an error $\varepsilon > 0$ we are willing to tolerate, the entire tail of the sequence starting from some index $n_\varepsilon \in \mathbb{N}$ lies within that error range from the value α; i.e., for all $n \geq n_\varepsilon$ we have $\alpha - \varepsilon < a_n < \alpha + \varepsilon$.

The logical definition is crucial, because it gives the precise meaning to the word *limit* (and *converge*). But of course it is still also important to have an intuitive idea about the notion as well! So how should you think about Definition II.11? In Figure II.1 and its caption we attempt to illustrate and describe this.

Exercise II.3 (Non-zero limit implies members are eventually non-zero)**.**
Suppose that $(a_n)_{n\in\mathbb{N}}$ is a sequence of real numbers which tends to a non-zero limit

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

Show that there exists some $N \in \mathbb{N}$ such that $a_n \neq 0$ for all $n \geq N$.

Hint: *We have* $|a_n| = |\alpha - (\alpha - a_n)| \geq |\alpha| - |\alpha - a_n|$ *by triangle inequality, Lemma II.1.*

Basic properties of limits of sequences

Having given the precise definition of limit, we are ready to state and prove some basic properties of limits.

The first one says that convergence implies boundedness. For perspective, let us already note that we will return to the analogous implication in a more general setup in Lecture VIII, Exercise VIII.2.

Theorem II.13 (A convergent sequence of number is bounded)**.**
If a sequence $(a_n)_{n\in\mathbb{N}}$ *of real numbers converges, then it is bounded.*

Exercise II.4 (Proof of Theorem II.13)**.**
Prove Theorem II.13.

Exercise II.5 (Boundedness does not imply convergence)**.**
Show that there exists a bounded sequence which is not convergent. Conclude that the implication in the converse direction compared to Theorem II.13 does not hold.

Lemma II.14 (Preservation of inequalities).

Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two convergent sequences of real numbers, with respective limits $\alpha = \lim_{n\to\infty} a_n$ and $\beta = \lim_{n\to\infty} b_n$. If we have

$$a_n \leq b_n \qquad \text{for all } n \in \mathbb{N},$$

then we also have the corresponding inequality

$$\alpha \;=\; \lim_{n\to\infty} a_n \;\leq\; \lim_{n\to\infty} b_n \;=\; \beta$$

for the limits.

Proof. Let $\varepsilon > 0$. Since $\frac{\varepsilon}{2} > 0$, it follows from the assumptions $\lim_{n\to\infty} a_n = \alpha$ and $\lim_{n\to\infty} b_n = \beta$ and the definition of limits that there exists numbers $n', n'' \in \mathbb{N}$ such that

$$\big|a_n - \alpha\big| < \frac{\varepsilon}{2} \quad \text{for } n \geq n', \qquad\qquad \big|b_n - \beta\big| < \frac{\varepsilon}{2} \quad \text{for } n \geq n''.$$

Consider now $m = \max\{n', n''\} \in \mathbb{N}$. Then $m \geq n'$ and $m \geq n''$, so both of the estimates above apply. Using furthermore the assumption $a_m \leq b_m$, we find

$$\begin{aligned} 0 \;\leq\; b_m - a_m \;&=\; b_m - \beta + \beta - \alpha + \alpha - a_m \\ &\leq\; \big|b_m - \beta\big| + \beta - \alpha + \big|\alpha - a_m\big| \\ &<\; \frac{\varepsilon}{2} + \beta - \alpha + \frac{\varepsilon}{2} \\ &=\; \beta - \alpha + \varepsilon. \end{aligned}$$

Rearranging this inequality, we get $\alpha < \beta + \varepsilon$. Since this was shown to hold for an arbitrary $\varepsilon > 0$, it must be that $\alpha \leq \beta$.[10] □

Note that Lemma II.14 is about the preservation of non-strict inequalities ($\leq$). Strict inequalities ($<$) may not be preserved in the limit, as the following exercise shows.

Exercise II.6 (Non-preservation of strict inequalities).

Find an example of two convergent sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ such that $a_n < b_n$ for all $n \in \mathbb{N}$ but the sequences have equal limits, $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$.

Exercise II.7 (An improvement to Lemma II.14).

Show that in Lemma II.14 it suffices to assume that the inequality holds for all sufficiently large n; more precicely that there exists $N \in \mathbb{N}$ such that $a_n \leq b_n$ for all $n \geq N$.

For sequences bounded from above or below, if a limit exists, it must also lie within the same bounds.

Corollary II.15 (Preservation of bounds).

Let $(a_n)_{n\in\mathbb{N}}$ be a convergent sequence of real numbers, with limit $\alpha = \lim_{n\to\infty} a_n$. If for some $u \in \mathbb{R}$ we have

$$a_n \leq u \qquad \text{for all } n \in \mathbb{N},$$

[10] This is a very common step in proofs, so let us look at it in detail now that we first use it. The claim can be reformulated as the implication

$$\forall\, \varepsilon > 0: \;\; \alpha < \beta + \varepsilon \qquad \Longrightarrow \qquad \alpha \leq \beta.$$

This is easy to prove by *contrapositive*: the claim is logically equivalent to (recall contrapositives and negations of statements with universal quantifier)

$$\alpha > \beta \qquad \Longrightarrow \qquad \exists\, \varepsilon > 0: \;\; \alpha \geq \beta + \varepsilon$$

In this equivalent formulation the claim is evident: if $\alpha > \beta$, then $\alpha - \beta > 0$ and we can use $\varepsilon = \alpha - \beta$ which indeed satisfies $\alpha = \beta + \varepsilon \geq \beta + \varepsilon$.

then we also have the corresponding inequality for the limit,

$$\alpha = \lim_{n\to\infty} a_n \leq u.$$

Similarly, if for some $\ell \in \mathbb{R}$ *we have* $a_n \geq \ell$ *for all* $n \in \mathbb{N}$*, then we also have* $\lim_{n\to\infty} a_n \geq \ell$*.*

Exercise II.8 (Proof of Corollary II.15).
Prove the first part of Corollary II.15, by using a constant sequence $(b_n)_{n\in\mathbb{N}}$ in Lemma II.14. Prove the second part by a similar idea.

Exercise II.9 (An improvement to Corollary II.15).
Formulate the following improvement precisely and prove it:
In Corollary II.15 it suffices to assume that the inequality holds for all sufficiently large n.
Hint: *Compare with Exercise II.7.*

Note that from Definition II.11 it is *not immediately obvious* that there could not exist several different numbers that satisfy what is required of a limit. In other words, if the condition (II.7) holds for $\alpha \in \mathbb{R}$, then could it also hold for some other real number $\alpha' \in \mathbb{R}$ in place of α?[11] Fortunately the following result lifts any concerns about the possibility of such an ambiguity.[12]

Corollary II.16 (Uniqueness of limits of real-number sequences).
Let $(a_n)_{n\in\mathbb{N}}$ *be a sequence of real numbers. If both* $\alpha \in \mathbb{R}$ *and* $\alpha' \in \mathbb{R}$ *are limits of this sequence, then* $\alpha = \alpha'$*.*

Proof. Trivially, the inequality $a_n \leq a_n$ holds for all $n \in \mathbb{N}$. If we now suppose that both α and α' are limits of the sequence $(a_n)_{n\in\mathbb{N}}$, and we use the former limit for the left-hand side and the latter for the right-hand side, from Lemma II.14 we get that $\alpha \leq \alpha'$. Similarly using the former for the right-hand side and the latter for the left-hand side, we get $\alpha' \leq \alpha$. The combination of these gives $\alpha = \alpha'$. □

We next state a useful result, which is in some sense an improved version of the preservation of inequalities — in it we do not even need to assume the existence of the limit, but the existence is a part of the conclusion! It states that if a sequence is "squeezed" in between two convergent sequences with equal limits, then the sequence itself also has to converge to the same limit. This is often called *the squeeze theorem* (although we label it as a lemma only). It also has various other affectionate nicknames. Some call it *the sandwich principle* — since it talks about a sequence "sandwiched" in between two others. Also *the lemma of two policemen* is descriptive: the idea being that one sequence "guards" the sequence of interest from above, preventing its escape to the upwards direction, and another sequence "guards" the sequence of interest from below, preventing its escape to the downwards direction. You can already tell that the result must be important, given how imaginative nomenclature it has inspired!

[11] The notation $\lim_{n\to\infty} a_n$ *suggests* that the limit is uniquely determined, but we should actually *prove* that this is so, in order to be sure that the notation is unambiguously defined!

[12] It would not have been difficult to prove the uniqueness of limits more directly, immediately after the definition. In fact, in Lecture VIII we treat sequences and limits in a more general setup, and we give a more direct (as well as more general) proof of the uniqueness. Here, however, we find it instructive to give a proof based on the preservation of inequalities.

Lemma II.17 (Squeeze theorem).

Let $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, $(c_n)_{n\in\mathbb{N}}$ be three sequences of real numbers. Suppose that for all $n \in \mathbb{N}$ we have

$$a_n \;\leq\; b_n \;\leq\; c_n.$$

Suppose also that the sequences $(a_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ are convergent and have the same limit

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n = \beta \in \mathbb{R}.$$

Then the sequence $(b_n)_{n\in\mathbb{N}}$ is also convergent and its limit is the same,

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

Exercise II.10 (Proof of Lemma II.17).

Prove the squeeze theorem (Lemma II.17).

Example II.18 (An example of the squeeze theorem).

Consider the sequence $(b_n)_{n\in\mathbb{N}}$ given by

$$b_n = 3 + 4^{-n}\,\sin(5n) \qquad \text{for } n \in \mathbb{N}.$$

Since $-1 \leq \sin(\theta) \leq +1$ for all $\theta \in \mathbb{R}$, we get

$$3 - 4^{-n} \;\leq\; b_n \;\leq\; 3 + 4^{-n} \qquad \text{for all } n \in \mathbb{N}.$$

We have $\lim_{n\to\infty}(3-4^{-n}) = 3$ and $\lim_{n\to\infty}(3+4^{-n}) = 3$, so we can use the sequences $(a_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ defined by $a_n = 3-4^{-n}$ and $c_n = 3+4^{-n}$ in the squeeze theorem (Lemma II.17) to conclude that

$$\lim_{n\to\infty} b_n = 3.$$

Rules of calculation with limits

Given two real-number sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we often consider new sequences formed from these for example by[13]

$$\big(a_n + b_n\big)_{n\in\mathbb{N}}\,, \qquad \big(a_n\,b_n\big)_{n\in\mathbb{N}}\,, \qquad \big(a_n/b_n\big)_{n\in\mathbb{N}}\,.$$

Theorem II.19 (Rules of calculation with limits).

Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two convergent sequences with respective limits $\alpha = \lim_{n\to\infty} a_n$ and $\beta = \lim_{n\to\infty} b_n$. Then we have

$$\lim_{n\to\infty}(a_n + b_n) \;=\; \alpha + \beta, \tag{II.8}$$

$$\lim_{n\to\infty}(a_n\,b_n) \;=\; \alpha\beta, \tag{II.9}$$

and if moreover $\beta \neq 0$, then

$$\lim_{n\to\infty}\frac{a_n}{b_n} \;=\; \frac{\alpha}{\beta}. \tag{II.10}$$

[13] The last of these is only well defined if $b_n \neq 0$ for all $n \in \mathbb{N}$ — otherwise we encounter a division by zero in a_n/b_n. But in fact, since we will be concerned with calculating limits as $n \to \infty$, we do not necessarily need to care if such a problem occurs just finitely many times along the sequence. If there are only finitely many indices $n \in \mathbb{N}$ for which $b_n = 0$, then we can still consider the sequence $(a_n/b_n)_{n=n_0}^{\infty}$ starting from a sufficiently large index n_0 such that $b_n \neq 0$ for all $n \geq n_0$. It is meaningful to consider the limit of this sequence, which we still write as $\lim_{n\to\infty} a_n/b_n$.

Before addressing the proof, we note two simple but practical consequences. From (II.8) and (II.9) above, we get the following special cases by choosing $(b_n)_{n\in\mathbb{N}}$ to be a constant sequence $b_n = c$ for all $n \in \mathbb{N}$.

Corollary II.20 (Additive and multiplicative constants in limits)**.**
Let $(a_n)_{n\in\mathbb{N}}$ be a convergent sequence with limit $\alpha = \lim_{n\to\infty} a_n$, and let $c \in \mathbb{R}$ be a constant. Then we have

$$\lim_{n\to\infty} (a_n + c) = \alpha + c, \qquad\qquad \lim_{n\to\infty} (c\, a_n) = c\alpha.$$

Proof of Theorem II.19. The proofs of all cases are quite similar, so we only do (II.9) and leave the other two as exercises.

So suppose that $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are two convergent sequences with limits $\alpha = \lim_{n\to\infty} a_n$ and $\beta = \lim_{n\to\infty} b_n$. We must prove that the sequence $(a_n\, b_n)_{n\in\mathbb{N}}$ converges to $\alpha\beta$, and we will do this directly using the Definition II.11 of limits. For this purpose, we first estimate

$$\begin{aligned}
& \big|a_n\, b_n - \alpha\beta\big| \\
= \; & \Big|a_n\, b_n - \alpha\, b_n + \alpha b_n - \alpha\beta\Big| && \text{(added and subtracted the same term)} \\
= \; & \Big|(a_n - \alpha)\, b_n + \alpha\,(b_n - \beta)\Big| && \text{(rearranged terms)} \\
\leq \; & \big|(a_n - \alpha)\, b_n\big| \; + \; \big|\alpha\,(b_n - \beta)\big| && \text{(triangle inequality)} \\
= \; & |b_n|\,\big|a_n - \alpha\big| \; + \; |\alpha|\,\big|b_n - \beta\big|.
\end{aligned}$$

By Theorem II.13, the convergent sequence $(b_n)_{n\in\mathbb{N}}$ is bounded, so there exists a constant $M > 0$ such that $|b_n| \leq M$ for all $n \in \mathbb{N}$. Plugging this in the above estimate, we obtain[14]

$$\big|a_n\, b_n - \alpha\beta\big| \;\leq\; M\,\big|a_n - \alpha\big| \;+\; |\alpha|\,\big|b_n - \beta\big|. \tag{II.11}$$

Now let $\varepsilon > 0$. Because $\lim_{n\to\infty} a_n = \alpha$, and because $\frac{\varepsilon}{3M}$ is a positive number, there exists an $n' \in \mathbb{N}$ such that we have

$$|a_n - \alpha| \;<\; \frac{\varepsilon}{3M} \qquad \text{for } n \geq n'.$$

Similarly, because $\lim_{n\to\infty} b_n = \beta$, and because $\frac{\varepsilon}{3\,(|\alpha|+1)}$ is a positive number, there exists an $n'' \in \mathbb{N}$ such that we have

$$|b_n - \beta| \;<\; \frac{\varepsilon}{3\,(|\alpha|+1)} \qquad \text{for } n \geq n''.$$

Then if $n \geq \max\{n', n''\}$, we have both $n \geq n'$ and $n \geq n''$, so both of the above estimates apply, and from (II.11) we get

$$\begin{aligned}
\big|a_n\, b_n - \alpha\beta\big| \;&\leq\; M\, \underbrace{\big|a_n - \alpha\big|}_{\leq \varepsilon/3M} + |\alpha|\; \underbrace{\big|b_n - \beta\big|}_{\leq \varepsilon/3(|\alpha|+1)} \\
&\leq\; \underbrace{\frac{M}{3M}}_{\leq 1/3}\,\varepsilon \;+\; \underbrace{\frac{|\alpha|}{3\,(|\alpha|+1)}}_{\leq 1/3}\,\varepsilon \\
&\leq\; \frac{2}{3}\,\varepsilon \;<\; \varepsilon.
\end{aligned}$$

If we set $n_\varepsilon = \max\{n', n''\}$, then from the above we got $\big|a_n\, b_n - \alpha\beta\big| < \varepsilon$ for all $n \geq n_\varepsilon$. Since $\varepsilon > 0$ was arbitrary, this by definition shows that $\lim_{n\to\infty}(a_n\, b_n) = \alpha\beta$, proving (II.9).

The proofs of (II.8) and (II.10) are left as exercises. □

[14]After this, the idea is that assumed convergence of $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ implies that both $|a_n - \alpha|$ and $|b_n - \beta|$ become small for n large. This should allow us to show that the expression (II.11) becomes small for large n. The rest of the proof is about making this idea precise.

Exercise II.11 (Completing the proof of Theorem II.19).
Prove (II.8) and (II.10).

Here are some exercises, in which you may use the properties of limits above together with some well-known limits from earlier courses.

Exercise II.12 (Calculating limits of sequences).
Calculate the limits of the real-number sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, where

$$a_n = \frac{3n+4}{5n+6} \qquad \text{and} \qquad b_n = \frac{(1+2n)(1+3n)\,e^{-n}}{n^3\,e^{-2n}+n^2\,e^{-n}} \qquad \text{for } n \in \mathbb{N}.$$

Exercise II.13 (Examining properties of some sequences).
Let

$$a_n = \frac{n+1}{2n+1} \qquad \text{and} \qquad b_n = \sqrt[n]{3^n+2^n} \qquad \text{for } n \in \mathbb{N}.$$

Find out whether the sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are increasing, decreasing, bounded from above, bounded from below. Determine the limits of these sequence if they exist.

II.4. Density of rational and irrational numbers

We will next look at an interesting property of the real line: both the rational numbers and the irrational numbers are dense in $\mathbb{R}$. In particular, between any two rational numbers there is an irrational number, and between any two irrational numbers there is a rational number. The fact that the countably infinite (and thus set-theoretically rather "small") set of rational numbers is already dense in the (uncountably infinite) set of real numbers is a topological property of $\mathbb{R}$ known as *separability.*

Density of rational numbers

Let us show that between any two (different) real numbers, there exists a rational number.

Theorem II.21 (Density of rational numbers).
Let $x, y \in \mathbb{R}$ with $x < y$. Then there exists a $q \in \mathbb{Q}$ such that $x < q < y$.

Proof. Since $y-x > 0$, we can choose an $n \in \mathbb{N}$ such that $n > \frac{1}{y-x}$. Now in the set $\big\{ \frac{m}{n} \;\big|\; m \in \mathbb{Z} \big\} \subset \mathbb{Q}$ of all integer multiples of $\frac{1}{n}$, the distance between consecutive elements is $\frac{1}{n}$. In particular since the interval $(x, y) \subset \mathbb{R}$ has length $y - x > \frac{1}{n}$, at least one such integer multiple $\frac{m}{n}$ must lie in this interval. In other words, the rational number $q = \frac{m}{n} \in \mathbb{Q}$ lies in the desired interval, $q \in (x, y)$. □

Although we only apparently showed that an open interval $(x, y) \subset \mathbb{R}$ of the real axis contains one rational number, it almost immediately follows that the interval in fact contains infinitely many rational numbers.

Corollary II.22 (An open interval contains infinitely many rational numbers)**.**
On any nonempty open interval $(x, y) \subset \mathbb{R}$, there are infinitely many rational numbers, i.e., $\#\big((x, y) \cap \mathbb{Q}\big) = \infty$.

Proof. By Theorem II.21, we may first choose one rational number $q_1 \in \mathbb{Q}$ such that $x < q_1 < y$. Then applying the same result to the interval (x, q_1), we can choose $q_2 \in \mathbb{Q}$ such that $x < q_2 < q_1 < y$. Inductively, once $q_1, \ldots, q_n \in \mathbb{Q}$ have been chosen so that

$$x < q_n < q_{n-1} < \cdots < q_2 < q_1 < y,$$

we choose a rational number $q_{n+1} \in \mathbb{Q}$ from the interval (x, q_n). This yields an infinite sequence $(q_n)_{n\in\mathbb{N}}$ of rational numbers on the interval (x, y). Finally note that by construction the sequence is strictly decreasing, so no two members of it are equal, and we have indeed found infinitely many different rational numbers on the interval. □

Density of irrational numbers

Similarly, between any two (different) real numbers, there exists an irrational number.

Theorem II.23 (Density of irrational numbers)**.**
Let $x, y \in \mathbb{R}$ with $x < y$. Then there exists a $z \in \mathbb{R} \setminus \mathbb{Q}$ such that $x < z < y$.

Proof. Let us first use Corollary II.22 to choose two different rational numbers $q, r \in \mathbb{Q}$ so that $x < q < r < y$. It suffices to show that the subinterval $(q, r) \subset (x, y)$ contains an irrational number.

Let

$$z \;=\; q + \frac{r-q}{\sqrt{2}}.$$

Clearly we have $z > q$, since $r - q > 0$ and $\frac{1}{\sqrt{2}} > 0$. On the other hand $\frac{1}{\sqrt{2}} < 1$, so we also have

$$z \;=\; q + \frac{r-q}{\sqrt{2}} \;<\; q + 1 \cdot (r-q) \;=\; r.$$

From these we get that the number z lies in the subinterval of interest, $x < q < z < r < y$. It remains to show that z is irrational. From the definition of z, we can solve

$$\sqrt{2} \;=\; \frac{r-q}{z-q}.$$

Now argue by contradiction: if z were rational, then the right-hand side above would also be rational (recall that q, r were chosen rational), so $\sqrt{2}$ would be rational, contradicting Theorem I.5. We thus conclude the desired irrationality $z \in \mathbb{R} \setminus \mathbb{Q}$. □

Again from knowing that any non-empty open interval $(x, y) \subset \mathbb{R}$ of the real axis contains one irrational number, we easily get that such an interval in fact contains infinitely many irrational numbers.

Corollary II.24 (An open interval contains infinitely many irrational numbers)**.**
On any nonempty open interval $(x, y) \subset \mathbb{R}$, there are infinitely many irrational numbers, i.e., $\#\big((x, y) \setminus \mathbb{Q}\big) = \infty$.

Exercise II.14 (Proof of Corollary II.24)**.**
Prove Corollary II.24, using Theorem II.23 and an argument similar to Corollary II.22.

II.5. Axioms of the real numbers

So far we have talked about real numbers, but we have not in fact really said what they are! Upon a close inspection, these seemingly innocent old friends turn out to be not entirely trivial to understand precisely...

As with many other mathematical structures, a standard rigorous approach is to take certain properties of real numbers as axioms, and require that any other statement made about real numbers is logically deduced from these axioms.[15] [16] [17] We want the axioms to be as modest as possible, so that the starting point is uncontroversial. And since everything else is deduced from the starting point by logical reasoning, also everything else will be uncontroversial — which of course is the point of mathematics!

The set of real numbers is denoted by $\mathbb{R}$, and it is equipped with two binary operations:

$$
\begin{aligned}
\text{addition "}+\text{":} \quad & \mathbb{R} \times \mathbb{R} \to \mathbb{R} \\
& (x, y) \mapsto x + y, \\
\text{multiplication "}\cdot\text{":} \quad & \mathbb{R} \times \mathbb{R} \to \mathbb{R} \\
& (x, y) \mapsto x \cdot y \\
& (\text{omitted in usual notation, so } x \cdot y = xy),
\end{aligned}
$$

and a binary relation

$$\text{less than "}<\text{".}$$

The axioms of real numbers fall into three types: the *field axioms* concern only the operations of addition and multiplication, the *order axioms* concern the order relation $<$ and how it behaves under addition and multiplication, and finally in many ways the most subtle is the *completeness axiom*.

On the field axioms and order axioms

The field axioms and order axioms concern properties of the real numbers that are without a doubt familiar to the reader already. We list these axioms and briefly discuss them in Appendix B.1. For a broader mathematical perspective, it may be worthwhile to take a look and to appreciate also these familiar properties of addition, multiplication, and order.

---

[15] In practice, it would be very tedious to give full proofs of all properties of real numbers that we use, as this would include much of what you have learned about mathematics since kindergarten... We will therefore not ask you to provide proofs of totally commonplace statements about the real numbers — but instead to just realize that they must in principle be logical consequences of the axioms. In Appendix B.1 all of the axioms of $\mathbb{R}$ are at least stated, and a few detailed derivations of simple properties from them are exemplified.

[16] In addition to merely giving the axioms, it would be desirable to provide a (set-theoretic) construction of real numbers, for which the axioms can be shown to be true. Such constructions indeed exist, but we will not enter the details in the present course.

[17] Importantly, a sort of uniqueness property of real numbers holds: any two constructions in which the axioms hold yield results that are in all essential ways the same (isomorphic). So we do not need to care about the chosen construction — the axioms of real numbers themselves are a good starting point.

Here we content ourselves to just mentioning that there are various other *fields* regularly used in mathematics — for example:

- the field of rational numbers $\mathbb{Q}$;
- the field of complex numbers $\mathbb{C}$;
- finite fields, the simplest example of which is the field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ of p elements, where p is a prime number;
- algebraic number fields, a simple example of which is the field $\mathbb{Q}(\sqrt{33})$ of rational numbers adjoined with a square root of 33;
- the field $\mathbb{Q}_p$ of p-adic numbers, where p is a prime number;
- the field $\mathbb{K}(q)$ of rational functions in a single variable q over another field $\mathbb{K}$;
- $\cdots$

All of these satisfy the same *field axioms* as $\mathbb{R}$, so the rules governing addition and multiplication with all their consequences are similar.

The field $\mathbb{Q}$ of rational numbers (as well as some but not all algebraic number fields) moreover has an order relation compatible with the addition and multiplication, in that it satisfies also the same *order axioms* as $\mathbb{R}$. The "only" difference between the real numbers and the rational numbers must therefore stem from the *completeness axiom*, which we discuss in more detail below.

Completeness axiom

We will next address the one really subtle axiom of the real numbers: the *completeness axiom.* But before stating it, we first need to introduce a few notions that will make an appearance there.

Supremum and infimum

Let us start from notions of boundedness for subsets of the real line. If you compare the following with Definition II.8, a common theme becomes evident.

Definition II.25 (Upper and lower bounds)**.**
Let $A \subset \mathbb{R}$ be a subset.

A real number $u \in \mathbb{R}$ is an **upper bound** for A if for all $a \in A$ we have $a \leq u$.

A real number $\ell \in \mathbb{R}$ is a **lower bound** for A if for all $a \in A$ we have $a \geq \ell$.

Example II.26 (Examples of upper and lower bounds)**.**

- Consider the open interval $(-\pi, +\pi) \subset \mathbb{R}$. The number π is an upper bound for the set $(-\pi, +\pi)$. But also any larger number $u > \pi$ is an upper bound for $(-\pi, +\pi)$. Likewise, any number $\ell \leq -\pi$ is a lower bound for the set $(-\pi, +\pi)$.
- Consider the set $\mathbb{N} \subset \mathbb{R}$ of natural numbers. The number 1 is a lower bound for $\mathbb{N}$, and so is any number $\ell < 1$. The set $\mathbb{N}$ does not have any upper bounds: for any $u \in \mathbb{R}$ there exists a natural number $n \in \mathbb{N}$ such that $n > u$.
- The sets $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ do not have any upper or lower bounds.
- Consider the empty set $\emptyset \subset \mathbb{R}$. Any real number $u \in \mathbb{R}$ is an upper bound for $\emptyset$.[18] Likewise, any real number $\ell \in \mathbb{R}$ is a lower bound for $\emptyset$.

[18]Indeed, the definition (Definition II.25) of an upper bound for a set only requires something for each element of the set. The empty set $\emptyset$ has no elements, so its upper bounds are subject to no requirements — anything goes!

From these examples we learn that upper and lower bounds for arbitrary sets do not need to exist. Moreover, when they exist, they are never unique.

Definition II.27 (Sets bounded from above and below)**.**
A subset $A \subset \mathbb{R}$ is **bounded from above** if it has an upper bound, and **bounded from below** if it has a lower bound.

The following definition introduces the key notion in the completeness axiom: the least upper bound or supremum. The supremum and its counterpart infimum feature in a number of contexts in real analysis — some applications are exemplified in Appendix B.4.

Definition II.28 (Supremum and infimum)**.**
Let $A \subset \mathbb{R}$ be a subset of the real line. A number $u^* \in \mathbb{R}$ is called the **least upper bound** or **supremum** of A, if u^* is an upper bound for A (i.e. $\forall a \in A : \ a \leq u^*$) and for all upper bounds u of A we have $u^* \leq u$. We then denote $u^* = \sup A$.

Let $B \subset \mathbb{R}$ be a subset of the real line. A number $\ell^* \in \mathbb{R}$ is called the **greatest lower bound** or **infimum** of B, if ℓ^* is a lower bound for B (i.e. $\forall b \in B : \ b \geq \ell^*$) and for all lower bounds ℓ of B we have $\ell^* \geq \ell$. We then denote $\ell^* = \inf B$.

Note that if a set $A \subset \mathbb{R}$ is not bounded from above, then it does not have any upper bounds, and it certainly does not have a least upper bound, i.e., a supremum. Also note that for the empty set $\emptyset \subset \mathbb{R}$, any real number is an upper bound, so there is no least upper bound (given any upper bound it is always possible to find an even smaller upper bound). These observations indicate that the supremum as defined above can only exist for subsets $A \subset \mathbb{R}$ which are nonempty ($A \neq \emptyset$) and bounded from above. For similar reasons, the infimum as defined above can only exist for subsets $B \subset \mathbb{R}$ which are nonempty ($B \neq \emptyset$) and bounded from below. The definition of supremum and infimum makes sense without including these assumptions, but for the existence of supremum and infimum, the additional assumptions are necessary.[19]

Implicit in the notations $\sup A$ and $\inf B$ above (as well as in our use of the definite article in Definition II.28) is that the least upper bound and the greatest lower bound, if they exist, are uniquely determined by the above properties. This is indeed not difficult to check, and we leave it as an exercise to the reader.

The notions of supremum and infimum are crucial in this course as well as mathematics more generally. For later uses, let us give slightly different practical characterizations for them.

Lemma II.29 (Characterization of supremum)**.**
A number $u^* \in \mathbb{R}$ is the supremum of a subset $A \subset \mathbb{R}$ if and only if the following two conditions hold:

[19]The fact that these additional assumptions are also *sufficient* for the existence of supremum (resp. infimum) is exactly the content of the completeness axiom, in its formulation (C1) (resp. (C1'))

- *For all $a \in A$ we have $a \leq u^*$;*[20]
- *For all $\varepsilon > 0$ there exists some $a \in A$ such that $a > u^* - \varepsilon$.*[21]

Exercise II.15 (Proof of Lemma II.29)**.**
Prove Lemma II.29.
Hint: *Compare the defining properties of supremum in Definition II.28 with the interpretations of the two conditions in the lemma, given in the footnotes.*

Exercise II.16 (Characterization of infimum)**.**
Formulate and prove the counterpart of Lemma II.29 for infimum, i.e., two conditions characterizing the infimum of a subset $B \subset \mathbb{R}$.

Supremum and infimum are generalizations of the notions of maximum (i.e., the largest element) and minimum (i.e., the smallest element). For comparison, let us recall also these notions.

Definition II.30 (Maximum and minimum)**.**
Let $A \subset \mathbb{R}$ be a subset.

An element $a' \in A$ is the **maximum** (i.e. the **largest element**) of A, if for all $a \in A$ we have $a \leq a'$. We then denote $a' = \max A$.

An element $a'' \in A$ is the **minimum** (i.e. the **smallest element**) of A, if for all $a \in A$ we have $a \geq a''$. We then denote $a'' = \min A$.

The crucial difference between the maximum and the supremum (or an upper bound more generally) is that the maximum of A is required to be an element of the set A itself. A similar remark applies to the difference between the minimum and the infimum.

Note, however, that not every subset $A \subset \mathbb{R}$ has a maximum (resp. minimum) — even if we assume A to be non-empty and bounded from above (resp. from below).

Example II.31 (Open intervals have no maximum and minimum)**.**
The open interval $(-\pi, \pi)$ has no largest element and no smallest element ($\nexists \max(-\pi, \pi)$, $\nexists \min(-\pi, \pi)$). In fact, open intervals never have a maximum or minimum.

Example II.32 (A bounded nonempty set)**.**
The set $A = \left\{ \frac{1}{n} \;\middle|\; n \in \mathbb{N} \right\}$ is bounded and non-empty. The largest element of this set is $\max A = 1$. The set A has no smallest element ($\nexists \min A$). The least upper bound of this set is $\sup A = 1$ and the greatest lower bound of it is $\inf A = 0$. Note that in this case the infimum is not an element of the set, $0 \notin A$.

Exercise II.17 (Examine properties of some sets)**.**
Consider the sets

$$A_1 = [0, \sqrt{2}] \cap \mathbb{Q}, \qquad A_2 = \left\{ e^{-x} \;\middle|\; x \in \mathbb{R} \right\}, \qquad A_3 = \left\{ \frac{7n(-1)^n + 4}{6n} \;\middle|\; n \in \mathbb{N} \right\}.$$

Answer the following questions about each of the sets A_j, $j \in \{1, 2, 3\}$.

- Is the set A_j bounded from above?
- Is it bounded from below?

[20]This condition simply says that u^* is an upper bound for A.
[21]This condition says that any smaller number $u^* - \varepsilon$ cannot be an upper bound for A.

- Does it have a maximum, and if it does, what is $\max A_j$?
- Does it have a supremum, and if it does, what is $\sup A_j$?
- Does it have a minimum, and if it does, what is $\min A_j$?
- Does it have an infimum, and if it does, what is $\inf A_j$?

Exercise II.18 (The relationship between supremum and maximum).
Let $A \subset \mathbb{R}$.

(a) Prove that if $\max A$ exists, then we have $\sup A = \max A$.
(b) Prove that if $\sup A \in A$ (which in particular assumes that the supremum exists), then the maximum exists and we have $\max A = \sup A$.

Exercise II.19 (The relationship between infimum and minimum).
Formulate and prove an analogous result for the infimum and minimum.

Exercise II.20 (Some considerations of geometric sums).
Consider the real-number sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ defined by

$$a_n = \sum_{j=0}^{n} \Big(\frac{4}{5}\Big)^j = 1 + \frac{4}{5} + \frac{16}{25} + \cdots + \frac{4^n}{5^n} \quad \text{and} \quad b_n = \sum_{j=0}^{n} \Big(\frac{-1}{2}\Big)^j.$$

(a) Find simplified formulas for a_n and b_n. Find the limits $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$, and prove directly using the definition of limits of real-number sequences that your answers indeed are those limits.
Hint: *To simplify, recall finite geometric sums.*
(b) Does the set

$$A = \big\{a_n \bigm| n \in \mathbb{N}\big\} \subset \mathbb{R}$$

have a supremum, and if it does, what is $\sup(A)$? Does the set A have a maximum, and if it does, what is $\max(A)$?
(c) Does the set

$$B = \big\{b_n \bigm| n \in \mathbb{N}\big\} \subset \mathbb{R}$$

have an infimum, and if it does, what is $\inf(A)$? Does the set B have a minimum, and if it does, what is $\min(A)$?

Formulations of the completeness axiom

There are various equivalent ways of formulating the completeness axiom of $\mathbb{R}$. We begin by giving the statements of three formulations; the proof of their logical equivalence can be found in Appendix B.2.

The **completeness axiom** of $\mathbb{R}$ refers to any of the following (equivalent) statements:

(C1): Every non-empty subset $A \subset \mathbb{R}$ which is bounded from above has a least upper bound $\sup A \in \mathbb{R}$.
(C2): Every increasing real-number sequence $(a_n)_{n\in\mathbb{N}}$ which is bounded from above has a limit $\lim_{n\to\infty} a_n \in \mathbb{R}$.
(C3): Every collection $(I_n)_{n\in\mathbb{N}}$ of closed intervals $I_n \subset \mathbb{R}$, which is nested in the sense that $I_{n+1} \subset I_n$ for every $n \in \mathbb{N}$, has a nonempty intersection

$$\bigcap_{n\in\mathbb{N}} I_n \neq \emptyset.$$

Recall that in Example I.8 we saw that the intersection of nested *open intervals* can be empty. The formulation (C3) must be phrased with nested *closed intervals*.

Furthermore, it is easy to see that the first two formulations above are equivalent with the following two, respectively:

- **(C1'):** Every non-empty subset $A \subset \mathbb{R}$ which is bounded from below has a greatest lower bound $\inf A \in \mathbb{R}$.
- **(C2'):** Every decreasing real-number sequence $(a_n)_{n\in\mathbb{N}}$ which is bounded from below has a limit $\lim_{n\to\infty} a_n \in \mathbb{R}$.

Exercise II.21 (Equivalence of (C1) and (C1')).
Prove that (C1) and (C1') are logically equivalent.

Exercise II.22 (Equivalence of (C2) and (C2')).
Prove that (C2) and (C2') are logically equivalent.

We have thus given altogether five different (but logically equivalent) formulations of the completeness axiom. Once their equivalence is proven[22], we may use any of them as we like. Which formulation is the most convenient depends on what we are trying to do. As an example application of the formulation (C2) of the completeness axiom, let us discuss the existence of real numbers given by decimal expansions.

Example II.33 (Decimal expansions).
Let $(d_1, d_2, d_3, \ldots) = (d_k)_{k\in\mathbb{N}}$ be a sequence of digits $d_k \in \{0, 1, 2, \ldots, 8, 9\}$. The precise definition of the **decimal number**

$$0.d_1 d_2 d_3 \ldots$$

is the sum of the series

$$\sum_{k=1}^{\infty} d_k \, 10^{-k}. \tag{II.12}$$

In particular, the existence of this decimal number requires the series to be convergent (which we will soon see is a consequence of the completeness axiom).

The partial sums of the series

$$a_n = \sum_{k=1}^{n} d_k \, 10^{-k}$$

clearly form an increasing sequence $(a_n)_{n\in\mathbb{N}}$: the added terms are nonnegative, so

$$a_{n+1} - a_n \; = \; \sum_{k=1}^{n+1} d_k \, 10^{-k} - \sum_{k=1}^{n} d_k \, 10^{-k} \; = \; d_{n+1} \, 10^{-n-1} \; \geq \; 0.$$

[22]This should probably be read as: "after reading Appendix B.2 and doing Exercises II.21 and II.22". Although if you'd rather move on in the main material at the expense of a complete mathematical understanding of this point, you can admit the equivalence on faith (i.e., without fully convincing yourself of the proof of the equivalences).

The sequence of the partial sums is also bounded from above, because

$$
\begin{aligned}
a_n \;&=\; \sum_{k=1}^{n} d_k \, 10^{-k} \\
&\leq\; \sum_{k=1}^{n} 9 \cdot 10^{-k} && \text{(since } d_k \leq 9\text{)} \\
&=\; \frac{9}{10} \sum_{\ell=0}^{n-1} 10^{-\ell} && \text{(change summation index to } \ell = k-1\text{)} \\
&=\; \frac{9}{10} \, \frac{1-10^{-n}}{1-\frac{1}{10}} && \text{(finite geometric sum)} \\
&\leq\; \frac{9}{10} \, \frac{1}{9/10} \;=\; 1.
\end{aligned}
$$

According to the formulation (C2) of the completeness axiom, the sequence $(a_n)_{n \in \mathbb{N}}$ of partial sums converges, because it is increasing and bounded from above. This means that the series (II.12) converges.

We conclude that every decimal expansion indeed represents some real number.

Conventions to extend the notion of supremum and infimum

In Definition II.28 we considered the supremum only for subsets A which are nonempty and bounded from above. If A fails one of these properties, we will use the following conventions:

- If a set $A \subset \mathbb{R}$ is not bounded from above (and thus has no upper bounds), we interpret the least upper bound as the symbol $+\infty$ (not a number),
$$\sup A \;=\; +\infty.$$
- For the empty set $\emptyset \subset \mathbb{R}$ (for which any real number is an upper bound), we interpret the least upper bound as the symbol $-\infty$ (not a number),
$$\sup \emptyset \;=\; -\infty.$$

Symmetrically for the infimum we use the conventions:

- If a set $A \subset \mathbb{R}$ is not bounded from below (and thus has no lower bounds), we interpret the greatest lower bound as the symbol $-\infty$ (not a number),
$$\inf A \;=\; -\infty.$$
- For the empty set $\emptyset \subset \mathbb{R}$ (for which any real number is a lower bound), we interpret the greatest lower bound as the symbol $+\infty$ (not a number),
$$\inf \emptyset \;=\; +\infty.$$

II.6. ♡ In-depth topics about the real numbers

The reader interested in deepening their understanding may wish to look into some of the following topics related to the subject of this lecture:

- The field axioms and order axioms of the real numbers, Appendix B.1.
- The proof of equivalence of the different formulations of the completeness axiom, Appendix B.2.

- The Cantor set — a fractal subset of the unit interval with interesting set-theoretic, topological, and measure theoretic properties, Appendix B.3.
- Applications of supremum and infimum, Appendix B.4.

Lecture III

Sequences and functions on the real line

The main theme of this section is continuous functions of a real variable. We will, for example, verify that polynomial and rational functions are continuous. We examine continuous functions on closed intervals in some detail, and show in particular that they always have a maximum and a minimum — an important existence result for optimization tasks.[1] We also prove the intermediate value theorem (Bolzano's theorem), according to which continuous functions on intervals "can not skip values".

One of the main objectives here is to start from a precise definition of continuity, and to rigorously prove the above (probably familiar) facts. Another goal is to start drawing attention to the underlying topological reasons behind such important results. In particular, the completeness axiom of the real line crucially underlies many of these results. On the other hand, later in the course we obtain a more general perspective, and will be able to appreciate the role of, e.g., compactness (of closed intervals) and connectedness (of intervals) in the results.

As a tool we will still use real-number sequences, and in particular their judiciously chosen subsequences. In a key role will be the observation that from a bounded sequence it is always possible to pick some convergent subsequence.[2]

III.1. Real-number sequences

Subsequences

Consider a sequence of real numbers,

$$(x_1, x_2, x_3, x_4, x_5, \ldots).$$

The idea of a subsequence is that we leave out some members from the original sequence. For example if we leave out the first member of the above sequence, we obtain the (sub)sequence

$$(x_2, x_3, x_4, x_5, x_6, \ldots),$$

and if we leave out all members of even indices, we obtain the (sub)sequence

$$(x_1, x_3, x_5, x_7, x_9, \ldots).$$

[1] Mathematically, optimization is formulated as the task of finding the maximum (or minimum) of some objective function: e.g., "profit", "accuracy", "efficiency", ... (or "cost", "error", "risk", ...). If the maximum (or minimum) does not exist, the problem is not even well posed and it is meaningless to look for the optimum!

[2] This fact makes use of the completeness axiom of the real numbers. In later chapters, this property will be recognized as a compactness statement.

Yet another example is leaving out all except prime number indices, in which case we obtain the (sub)sequence

$$(x_2, x_3, x_5, x_7, x_{11}, x_{13}, x_{17}, x_{19}, x_{23}, x_{29}, x_{31}, x_{37}, \ldots).$$

We anyway require that infinitely many members remain after leaving out some. Also the convention is that the remaining members are relabeled by the original index set $\mathbb{N}$ again[3] — this ensures that the subsequence itself will be a sequence (in the sense of Section II.2).

The precise definition of a subsequence is the following.

Definition III.1 (Subsequence)**.**
Let $(x_n)_{n\in\mathbb{N}}$ be a sequence, and $\varphi\colon \mathbb{N} \to \mathbb{N}$ a strictly increasing function. Then the sequence

$$\big(x_{\varphi(n)}\big)_{n\in\mathbb{N}} = (x_{\varphi(1)}, x_{\varphi(2)}, x_{\varphi(3)}, \ldots)$$

is called a **subsequence** of the sequence $(x_n)_{n\in\mathbb{N}}$.

Example III.2 (Examples of picking subsequences)**.**
The first example informally described above, of leaving out the first member, is obtained by $\varphi_{\mathrm{shift}}\colon \mathbb{N} \to \mathbb{N}$ defined by the formula $\varphi_{\mathrm{shift}}(n) = n + 1$ for $n \in \mathbb{N}$. The corresponding subsequence of $(x_n)_{n\in\mathbb{N}} = (x_1, x_2, x_3, \ldots)$ is

$$\big(x_{\varphi_{\mathrm{shift}}(n)}\big)_{n\in\mathbb{N}} = (x_{\varphi_{\mathrm{shift}}(1)}, x_{\varphi_{\mathrm{shift}}(2)}, x_{\varphi_{\mathrm{shift}}(3)}, \ldots) = (x_2, x_3, x_4, \ldots).$$

The second example, of leaving out the even indices and keeping the odd, is obtained by $\varphi_{\mathrm{odd}}\colon \mathbb{N} \to \mathbb{N}$ defined by the formula $\varphi_{\mathrm{odd}}(n) = 2n - 1$ for $n \in \mathbb{N}$. The corresponding subsequence of $(x_n)_{n\in\mathbb{N}} = (x_1, x_2, x_3, \ldots)$ is

$$\big(x_{\varphi_{\mathrm{odd}}(n)}\big)_{n\in\mathbb{N}} = (x_{\varphi_{\mathrm{odd}}(1)}, x_{\varphi_{\mathrm{odd}}(2)}, x_{\varphi_{\mathrm{odd}}(3)}, \ldots) = (x_1, x_3, x_5, \ldots).$$

Another common notation for a subsequence is

$$\big(x_{n_k}\big)_{k\in\mathbb{N}} \;=\; (x_{n_1}, x_{n_2}, x_{n_3}, \ldots)\ ,$$

where k is used for the new indices, and from the original sequence we have picked members with indices $n_1 < n_2 < n_3 < \cdots$, corresponding to $n_k = \varphi(k)$ in the above definition.

It is easy to show (by induction) that for a strictly increasing function $\varphi\colon \mathbb{N} \to \mathbb{N}$ we have

$$\varphi(n) \;\geq\; n \qquad \text{for all } n \in \mathbb{N}, \tag{III.1}$$

which implies that a member in a subsequence corresponds to a member in the original sequence which is at least as far in the tail of the original sequence (its index in the original sequence is at least as large as its index in the subsequence). This observation makes the proof of the following lemma very easy.

Lemma III.3 (Subsequences of a convergent sequence inherit the same limit)**.**
Suppose that $(x_n)_{n\in\mathbb{N}}$ *is a real-number sequence which converges to a limit* $\lim_{n\to\infty} x_n = x$. *Then any subsequence* $(x_{\varphi(n)})_{n\in\mathbb{N}}$ *also converges to the same limit,* $\lim_{n\to\infty} x_{\varphi(n)} = x$.

[3] It is possible to consider index sets other than $\mathbb{N}$, but even in such cases we require that the indices are consecutive integers. For simplicity, let us focus on the case when the index set is $\mathbb{N}$.

Exercise III.1 (Proof of Lemma III.3)**.**
Prove Lemma III.3.

Example III.4 (The alternating sign sequence)**.**
Consider the sequence $(a_n)_{n\in\mathbb{N}}$ where $a_n = (-1)^n$, i.e.,

$$(a_n)_{n\in\mathbb{N}} = (-1, +1, -1, +1, -1, \ldots).$$

The subsequence of even members, obtained with $\varphi_{\text{even}}(n) = 2n$, is

$$(a_{2n})_{n\in\mathbb{N}} = (+1, +1, +1, +1, \ldots),$$

and it obviously converges to $+1$.

The subsequence of odd members, obtained with $\varphi_{\text{odd}}(n) = 2n - 1$, is

$$(a_{2n-1})_{n\in\mathbb{N}} = (-1, -1, -1, -1, \ldots),$$

and it obviously converges to -1.

From the fact that these two subsequences have different limits, it follows by contrapositive of Lemma III.3 that the original sequence $(a_n)_{n\in\mathbb{N}}$ does not converge. (Of course this non-convergence could also have been proven directly.)

Example III.5 (A sequence without convergent subsequences)**.**
Consider the sequence $(a_n)_{n\in\mathbb{N}}$ where $a_n = -n$, i.e.,

$$(a_n)_{n\in\mathbb{N}} = (-1, -2, -3, -4, \ldots).$$

It is easy to see that this sequence has no convergent subsequences (in fact it has no bounded subsequences, so this follows from Theorem II.13, by contrapositive).

Existence of monotone subsequences

It turns out that from any real-number sequence, it is possible to pick a subsequence which is monotone (either increasing or decreasing). This observation is somewhat interesting on its own right, but its most crucial role is as a lemma towards the next very important result, Theorem III.7.

Lemma III.6 (Any real-number sequence has a monotone subsequence)**.**
Let $(x_n)_{n\in\mathbb{N}}$ be a real-number sequence. Then there exists a subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ which is monotone.

Proof. Define the set

$$J \;:=\; \Big\{ n \in \mathbb{N} \;\Big|\; x_m \leq x_n \text{ for all } m > n \Big\}$$

of those indices, whose corresponding value is never exceeded later in the sequence. We consider two cases: the set J is either finite or infinite.

Suppose first that the set J is infinite. In this case, by enumerating the elements of J in an increasing order

$$J = \{n_1, n_2, n_3, \ldots\} \qquad \text{with} \qquad n_1 < n_2 < n_3 < \cdots,$$

we obtain indices such that the corresponding subsequence

$$(x_{n_1}, x_{n_2}, x_{n_3}, \ldots)$$

is decreasing: indeed we have $x_{n_{k+1}} \leq x_{n_k}$ for all $k \in \mathbb{N}$ because $n_k \in J$ and $n_{k+1} > n_k$. In this case we have thus found a monotone (decreasing) subsequence.

Suppose then that the set J is finite. In this case there exists[4] some $n_1 \in \mathbb{N}$ such that all of the elements of J are smaller than n_1. By the definition of the set J, this implies that starting from the index n_1, every member of the sequence is later exceeded; or more formally

$$\forall\, n \geq n_1 : \quad \exists\, m > n : \quad x_m > x_n.$$

We can use this property to recursively find members which form a good subsequence. When the first k indices $n_1 < \cdots < n_k$ have been chosen, the property above implies that there exists some $n_{k+1} > n_k$ such that $x_{n_{k+1}} > x_{n_k}$, which we will use as the $k+1$:st index. The resulting subsequence $(x_{n_1}, x_{n_2}, x_{n_3}, \ldots)$ will be strictly increasing by construction: $x_{n_1} < x_{n_2} < x_{n_3} < \cdots$. In this case, too, we have found a monotone (increasing) subsequence. The proof is thus complete. □

Bounded sequences have convergent subsequences

The following result is crucially important in real analysis. We will see some of its familiar but powerful consequences already in this lecture.

Theorem III.7 (Bounded real-number sequences have convergent subsequences)**.** *Any bounded sequence $(x_n)_{n\in\mathbb{N}}$ of real numbers has a subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ which is convergent.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a bounded sequence of real numbers. Recall from Definition II.8 that boundedness of $(x_n)_{n\in\mathbb{N}}$ means that for some $r \in \mathbb{R}$ we have $|x_n| \leq r$ for all $n \in \mathbb{N}$. Obviously any subsequence of a bounded sequence is also bounded: the same bound r will a fortiori work for the subsequence. By Lemma III.6, the real-number sequence $(x_n)_{n\in\mathbb{N}}$ has a monotone subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$. As a bounded monotone sequence of real numbers, the sequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ converges by the completeness axiom, formulation (C2) or (C2'). We have thus shown that $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence. □

III.2. Functions of real variable

We will now consider real-valued functions of a real variable. Being a function of a real variable means that the domain of the function is taken to be a subset $A \subset \mathbb{R}$ of the real axis. Real-valuedness means that we (can) take the codomain of the function to be $\mathbb{R}$. Therefore we are interested in functions

$$f : A \to \mathbb{R} \qquad \text{where } A \subset \mathbb{R}.$$

Important special cases of the choice of the domain include, e.g., closed intervals $A = [a, b] \subset \mathbb{R}$ and the whole real axis $A = \mathbb{R}$.

Continuity of a function of real variable

The intuitive idea of continuous functions should be familiar from calculus courses. As a precise definition, we take the following.[5]

[4] If $J \neq \emptyset$, then $n_1 = \max J + 1$ works, and if $J = \emptyset$ then $n_1 = 1$ works.

[5] Later, in Lecture VI, we define continuity in a more general setup. In Lecture VIII that general definition will be shown to be logically equivalent to the definition used here, when specialized to the case of real-valued functions of a real variable. So continuity will still have an unambiguous meaning, despite the two apparently different definitions. Ultimately, you can use either one of the two equivalent definitions.

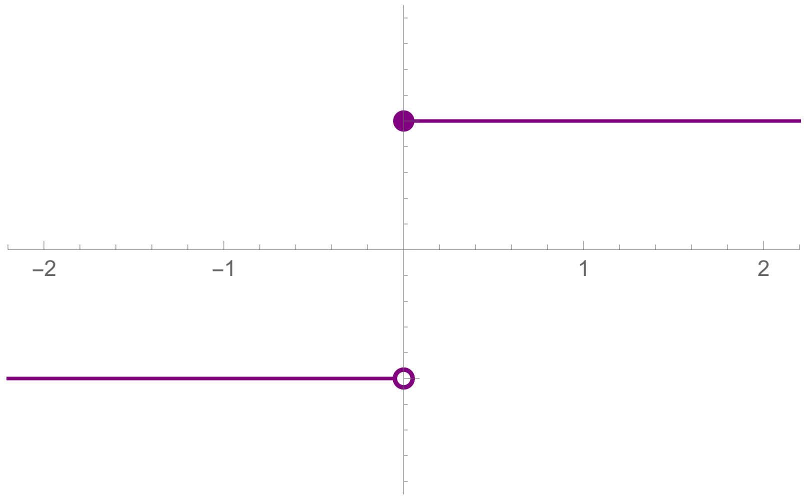


FIGURE III.1. The step function of Exercise III.2 is a basic example of a function that is not continuous.

Definition III.8 (Continuity of a real-valued function of a real variable)**.**
Let $A \subset \mathbb{R}$, and let $f \colon A \to \mathbb{R}$ be a function. We say that f is **continuous** if the following holds: whenever $(a_n)_{n\in\mathbb{N}}$ is a sequence in A which is convergent and has its limit in the domain, i.e., $\lim_{n\to\infty} a_n \in A$, then we have

$$\lim_{n\to\infty} f(a_n) = f\Big(\lim_{n\to\infty} a_n\Big).$$

To elaborate — given a sequence $(a_n)_{n\in\mathbb{N}}$ in the domain A of the function $f \colon A \to \mathbb{R}$, we map the members $a_n \in A$ of the sequence to real numbers $f(a_n) \in \mathbb{R}$, and form a real-number sequence $\big(f(a_n)\big)_{n\in\mathbb{N}}$ of these images under f, i.e., of the corresponding function values. Continuity of f (Definition III.8) is the requirement that if the original sequence is convergent and its limit is in the domain A, then also the image sequence is convergent and its limit is the image of the limit of the original sequence. In somewhat imprecise terms, "we are allowed to interchange the order of taking the limit and applying the function".[6]

Certainly a function which abruptly jumps from one value to another should be the simplest example of a function that is not continuous; see Figure III.1. In the next exercise you check the discontinuity of such a step function directly from the definition.

Exercise III.2 (Step function is discontinuous)**.**
Consider the function

$$f \colon \mathbb{R} \to \mathbb{R}, \qquad f(x) := \begin{cases} -1 & \text{for } x < 0 \\ +1 & \text{for } x \geq 0. \end{cases}$$

Show that it is not a continuous function $\mathbb{R} \to \mathbb{R}$.
Hint: *By Definition III.8 it suffices to find* some *convergent sequence* $(x_n)_{n\in\mathbb{N}}$ *on* $\mathbb{R}$ *such that* $\lim_n f(x_n) \neq f\big(\lim_n x_n\big)$.

[6]Note, however, that we still have to assume convergence of the original sequence to a limit in the domain.

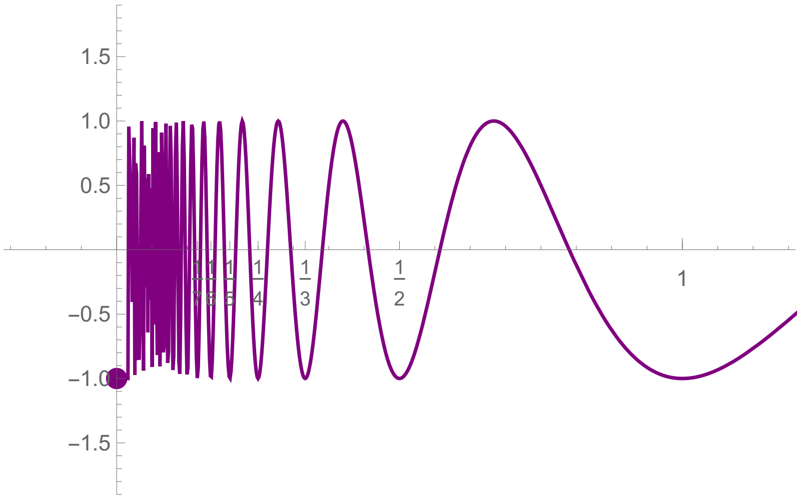


FIGURE III.2. The function of Example III.9.

Example III.9 (Another discontinuous function)**.**
Consider the function

$$f\colon\ [0,+\infty) \to \mathbb{R}, \qquad\qquad f(x) := \begin{cases} -1 & \text{for } x = 0 \\ \cos(\pi + 2\pi/x) & \text{for } x > 0. \end{cases}$$

Is this function continuous or not? (For a plot, see Figure III.2.)

Intuitively it seems that any potential problem should occur at $x = 0$, if at all. So let us try out some sequences $(x_n)_{n\in\mathbb{N}}$ which tend to 0.

First try the sequence with members $x_n = \frac{1}{n}$. This sequence has a limit

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \frac{1}{n} = 0.$$

The values of f at members of the sequence are

$$f(x_n) = \cos\Big(\pi + \frac{2\pi}{1/n}\Big) = \cos(\pi + 2\pi n) = -1 \qquad \text{for } n \in \mathbb{N}.$$

Therefore we find

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} (-1) = -1 = f(0) = f\big(\lim_{n\to\infty} x_n\big).$$

This is as required in the definition of continuity. (Note that $x_n = \frac{1}{n} \in [0,+\infty)$ for all $n \in \mathbb{N}$, so this is a sequence in the domain of definition of f, and also the limit $\lim_n x_n = 0$ remains in the domain, $0 \in [0,+\infty)$.)

If you were to try for example the sequence with members $x_n = 2^{-n}$, which also tends to zero, you would again find $\lim f(x_n) = f\big(\lim x_n\big)$.

But Definition III.8 requires a similar conclusion for *all* sequences (which are convergent in the domain of f). So a single counterexample will be sufficient to show discontinuity! For such a counterexample, take the sequence with members $x_n = \frac{4}{4n-1}$. One again notes that $x_n \in [0,+\infty)$ for all $n \in \mathbb{N}$, and $\lim x_n = 0$. Now the function values along this sequence are

$$f(x_n) = \cos\Big(\pi + \frac{2\pi}{4/(4n-1)}\Big) = \cos\Big(\pi + \frac{\pi}{2}(4n-1)\Big) = \cos\Big(\frac{\pi}{2} + 2\pi n\Big) = 0$$

for $n \in \mathbb{N}$. Therefore

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} 0 = 0 \neq -1 = f(0) = f\big(\lim_{n\to\infty} x_n\big).$$

This shows that f is not continuous.

Exercise III.3 (Indicator function of rational numbers).
Consider the function

$$f\colon\ \mathbb{R}\to\mathbb{R}, \qquad\qquad f(x) := \begin{cases} 1 & \text{for } x\in\mathbb{Q} \\ 0 & \text{for } x\in\mathbb{R}\setminus\mathbb{Q}. \end{cases}$$

Is it continuous?

Note that continuity depends on the domain of definition of the function: a discontinuous function restricted to a smaller set can become continuous, as you will see in the next example.

Exercise III.4 (The restriction of a discontinuous function can be continuous).
Consider the step funtion $f\colon \mathbb{R}\to\mathbb{R}$ of Exercise III.2, and let $\widetilde{f} = f|_{\mathbb{R}\setminus\{0\}}$ be its restriction to the set $\mathbb{R}\setminus\{0\}$ non-zero real numbers. Show that $\widetilde{f}\colon \mathbb{R}\setminus\{0\}\to\mathbb{R}$ is continuous.

In the converse direction things are stable: a function never loses its continuity when restricting the domain.

Exercise III.5 (Restrictions of continuous functions are continuous).
Suppose $f\colon A\to\mathbb{R}$ is a continuous function and $\widetilde{A}\subset A$ is a subset. Show that the restriction $\widetilde{f} = f|_{\widetilde{A}}$ is a continuous function $\widetilde{f}\colon \widetilde{A}\to\mathbb{R}$.

Operations on continuous functions

Because the definition of continuity relies on limits of real-number sequences, our earlier results about limits easily yield some fundamental results about continuous functions. Particularly important is the fact that the following pointwise operations preserve continuity.

Definition III.10 (Pointwise operations on real-valued functions).
Let $A\subset\mathbb{R}$, and let $f,g\colon A\to\mathbb{R}$ be two real-valued functions on A.

The **pointwise sum** of f and g is the function

$$f+g\colon\ A\to\mathbb{R} \qquad\qquad \big(f+g\big)(x) := f(x)+g(x) \quad \text{for } x\in A.$$

The **pointwise product** of f and g is the function

$$fg\colon\ A\to\mathbb{R} \qquad\qquad \big(fg\big)(x) := f(x)\,g(x) \quad \text{for } x\in A.$$

The **pointwise quotient** of f and g is the function

$$f/g\colon\ A'\to\mathbb{R} \qquad\qquad \big(f/g\big)(x) := \frac{f(x)}{g(x)} \quad \text{for } x\in A',$$

where $A' := \big\{x\in A \;\big|\; g(x)\neq 0\big\}$.

Note that the pointwise quotient f/g is only defined on the subset $A'\subset A$ of the domain, where the function g is non-vanishing — in order to avoid ill-defined division by zero.

Scalar multiplication of a function $f\colon A \to \mathbb{R}$ is a special case of the pointwise product: it is natural to interpret a real number $c \in \mathbb{R}$ also as the constant function $x \mapsto c$ on A; the function $cf\colon A \to \mathbb{R}$ is defined by $(cf)(x) := c\, f(x)$ for $x \in A$.

Theorem III.11 (Continuity-preserving operations)**.**
Let $A \subset \mathbb{R}$, and let $f, g\colon A \to \mathbb{R}$ be two continuous real-valued functions on A. Then also

(i) *the pointwise sum function $f + g\colon A \to \mathbb{R}$ is continuous;*
(ii) *the pointwise product function $fg\colon A \to \mathbb{R}$ is continuous;*
(iii) *the pointwise quotient function $f/g\colon A' \to \mathbb{R}$ is continuous on the subset $A' := \big\{x \in A \;\big|\; g(x) \neq 0\big\}$.*

Proof. The proofs of all cases are essentially similar, so let us provide the details only for (ii), and leave (i) and (iii) as exercises.

proof of (ii): Assume $f, g\colon A \to \mathbb{R}$ are continuous. We must prove that $fg\colon A \to \mathbb{R}$ is continuous, and we will do this directly using Definition III.8 of continuity. So suppose $(x_n)_{n\in\mathbb{N}}$ is a sequence such that $x_n \in A$ for all $n \in \mathbb{N}$, and $(x_n)_{n\in\mathbb{N}}$ converges to a limit in A. Let us denote the limit by $x = \lim_{n\to\infty} x_n \in A$. By definition of continuity, we must consider the sequence $\big((fg)(x_n)\big)_{n\in\mathbb{N}}$ of values of the pointwise product function fg and show that it converges to the limit $(fg)(x)$. By definition of pointwise products, the values are $(fg)(x_n) = f(x_n)\, g(x_n)$. By the assumptions of continuity of f and g we know that $\lim_{n\to\infty} f(x_n) = f(x)$ and $\lim_{n\to\infty} g(x_n) = g(x)$. With these and (II.9) of Theorem II.19, we get

$$\lim_{n\to\infty} \Big(f(x_n)\, g(x_n)\Big) = \Big(\lim_{n\to\infty} f(x_n)\Big)\Big(\lim_{n\to\infty} g(x_n)\Big) = f(x)\, g(x).$$

By definition of the pointwise products the right-hand side above is $(fg)(x)$, so we have shown

$$\lim_{n\to\infty} \Big((fg)(x_n)\Big) = (fg)(x),$$

which is what was needed to prove continuity. □

Exercise III.6 (Proof of Theorem III.11 (i) and (iii))**.**
Prove parts (i) and (iii) of the above theorem.

Exercise III.7 (Continuity of the tangent trigonometric function)**.**
Is the function

$$\tan\colon \mathbb{R} \setminus \Big\{\frac{\pi}{2} + n\pi \;\Big|\; n \in \mathbb{Z}\Big\} \to \mathbb{R}$$

continuous?
Hint: *The continuity of the trigonometric functions* $\sin\colon \mathbb{R} \to \mathbb{R}$ *and* $\cos\colon \mathbb{R} \to \mathbb{R}$ *can be considered known here (a precise justification will be given in Lecture IX). Recall that* $\tan(x) = \frac{\sin(x)}{\cos(x)}$.

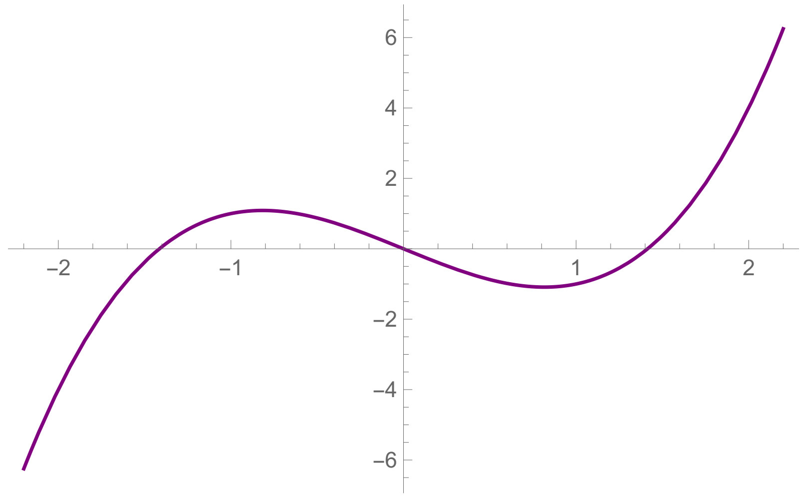
Corollary III.12 (Polynomials and rational functions are continuous)**.**
Let $a_0, a_1, a_2, \ldots, a_n \in \mathbb{R}$ and consider the polynomial function

$$P(x) := a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \qquad \textit{for } x \in \mathbb{R}.$$

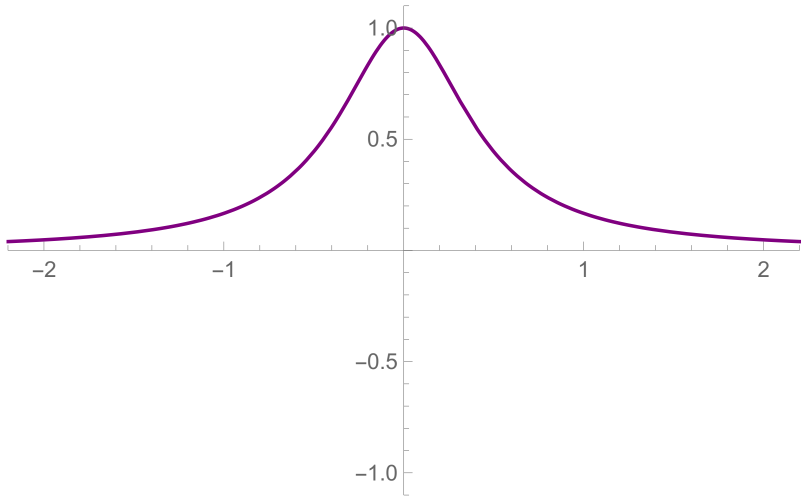
Then $P\colon \mathbb{R} \to \mathbb{R}$ is a continuous function.

Let also $b_0, b_1, b_2, \ldots, b_m \in \mathbb{R}$ and

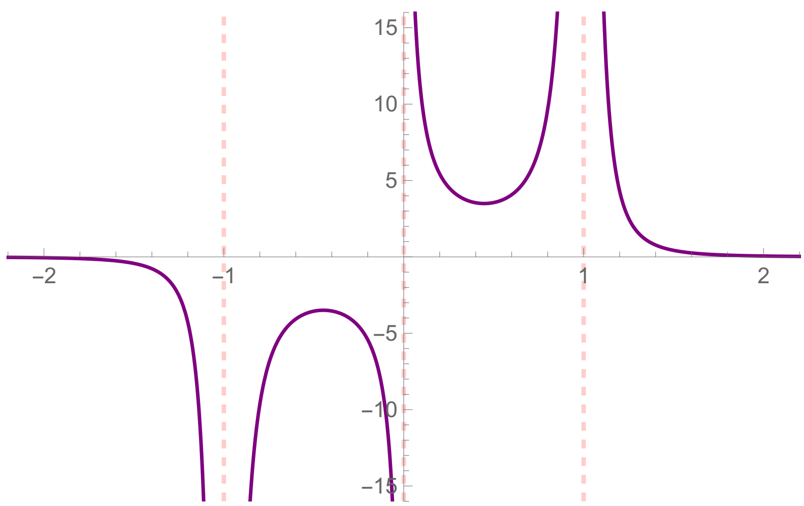
$$Q(x) := b_0 + b_1 x + b_2 x^2 + \cdots + b_m x^m \qquad \textit{for } x \in \mathbb{R}.$$



(a) The formula $x \mapsto x^3 - 2x$ defines a polynomial function $\mathbb{R} \to \mathbb{R}$.



(b) The formula $x \mapsto \frac{1}{1+5x^2}$ defines a rational function $\mathbb{R} \to \mathbb{R}$.



(c) The formula $x \mapsto \frac{1}{x(x^2-1)^2}$ defines a rational function $\mathbb{R} \setminus \{-1, 0, +1\} \to \mathbb{R}$. This rational function has poles at $x = 0$ and $x = \pm 1$, but the domain of definition is the complement of the poles, and the function is continuous.

FIGURE III.3. Polynomials and rational functions are continuous.

Consider the rational function

$$R(x) = \frac{P(x)}{Q(x)} \qquad \text{for } x \in A,$$

where $A := \{x \in \mathbb{R} \mid Q(x) \neq 0\}$. *Then* $R\colon A \to \mathbb{R}$ *is a continuous function.*

Proof. Let us first observe that the continuity of the identity function $x \mapsto x$ (as a function $\mathbb{R} \to \mathbb{R}$) is trivial from Definition III.8. The function $x \mapsto x^2$ is a pointwise product of identity functions, and therefore also continuous by Theorem III.11(ii). Continuing to take pointwise products with the identity function, we find that the monomial functions $x \mapsto x^k$ are continuous for all $k \in \mathbb{N}$ (easy induction).

Then note that for any $c \in \mathbb{R}$ the continuity of the constant function $x \mapsto c$ (as a function $\mathbb{R} \to \mathbb{R}$) is also obvious from the definition. Taking $c = a_k \in \mathbb{R}$ and further taking pointwise product with a monomial function, we find that $x \mapsto a_k x^k$ is continuous, for any $k \in \mathbb{Z}_{\geq 0}$ (for the case $k = 0$ one does not even need a product with monomial function). The polynomial $P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$ is a finite sum of terms of this type, so its continuity follows from applying Theorem III.11(i) repeatedly (more precisely, by induction on the degree n).

Since polynomial functions given by $P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$ and $Q(x) = b_0 + b_1 x + b_2 x^2 + \cdots + b_m x^m$ are continuous, we get from Theorem III.11(iii) that the rational function given by the pointwise quotient $R(x) = \frac{P(x)}{Q(x)}$ is continuous on the subset $A = \{x \in \mathbb{R} \mid Q(x) \neq 0\} \subset \mathbb{R}$ where the denominator is non-vanishing. □

Later in the course we will develop more powerful tools for verifying the continuity of functions. But knowing the continuity of polynomial and rational functions at least assures us that some nontrivial continuous functions exist; see Figure III.3 for a few concrete examples.

III.3. Continuous functions on a closed interval

Let us now examine in more depth the case when the domain of the function is a closed interval on the real axis, $A = [a, b] \subset \mathbb{R}$. This case appears very frequently in applications. We address some key properties of continuous functions

$$f\colon [a, b] \to \mathbb{R}$$

on a closed interval. Later in the course we will understand them from more general perspectives.

Boundedness and attaining extremum values

Let us recall a few key concepts: a function $f\colon A \to \mathbb{R}$ is **bounded** if there exists an $M \geq 0$ such that $|f(x)| \leq M$ for all $x \in A$. A function $f\colon A \to \mathbb{R}$ has a **maximum** if there exists a point $x_{\max} \in A$ such that $f(x) \leq f(x_{\max})$ for all $x \in A$, and it has a **minimum** if there exists a point $x_{\min} \in A$ such that $f(x) \geq f(x_{\min})$ for all $x \in A$.

Remark III.13 (Reinterpretation in terms of the range of the function)**.**
Consider the image $f[A] = \{f(x) \mid x \in A\} \subset \mathbb{R}$ of the whole domain A (this is often called the range of f). It is straightforward to verify that the function f is bounded if and only if

the set $f[A] \subset \mathbb{R}$ is bounded, and the function f has a maximum (resp. minimum) if and only if the set $f[A]$ has a maximum (resp. minimum).

The following important properties are related to compactness (see Lecture XI) of the closed interval.

Theorem III.14 (Extrema of continuous functions on a closed interval)**.**
Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function. Then,

(i) *f is bounded,*
(ii) *f has a maximum and a minimum.*

Before giving the proof, let us note that for both conclusions of Theorem III.14 it is important that the domain was assumed to be a *closed* interval. The examples in Figure III.4 illustrate what can go wrong on non-closed intervals.

Exercise III.8 (Continuous functions on non-closed intervals)**.**

(a) Prove that the function of Figure III.4(a) is continuous but not bounded.
(b) Prove that the function of Figure III.4(b) is continuous but has no minimum.
(c) Find an example of a continuous function on an interval, which has no maximum.

Proof of Theorem III.14 Let us prove the two assertions separately.

proof of (i): Suppose, by way of contradiction, that f is not bounded. Then for every $n \in \mathbb{N}$, it is possible to choose some $x_n \in [a,b]$ such that $|f(x_n)| \geq n$. Let us form a sequence $(x_n)_{n\in\mathbb{N}}$ using such choices. This sequence is bounded, since $a \leq x_n \leq b$ for all $n \in \mathbb{N}$. Therefore, by Theorem III.7, it has a convergent subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$. If we denote the limit of such a convergent subsequence by $x = \lim_{n\to\infty} x_{\varphi(n)}$, then by the preservation of bounds we have $a \leq x \leq b$ (see Corollary II.15). By continuity, then, $\big(f(x_{\varphi(n)})\big)_{n\in\mathbb{N}}$ converges to $f(x)$. But by the choice of x_n we have $|f(x_{\varphi(n)})| \geq \varphi(n) \geq n$ for all $n \in \mathbb{N}$, which means that the sequence $\big(f(x_{\varphi(n)})\big)_{n\in\mathbb{N}}$ is not bounded and so cannot be convergent (by the contrapositive of Theorem II.13). This is a contradiction. We conclude that f had to be bounded.

proof of (ii): Let us only prove that f has a maximum — the existence of a minimum can be concluded similarly (or by considering the maximum of the continuous function $-f$).

By (i), f is bounded, so the supremum of its values is finite,

$$C \;:=\; \sup\big\{f(x) \;\big|\; x \in [a,b]\big\} \;\in\; \mathbb{R}.$$

For all $n \in \mathbb{N}$ there then exists some $x_n \in [a,b]$ such that $f(x_n) \geq C - \frac{1}{n}$. Let us form a sequence $(x_n)_{n\in\mathbb{N}}$ using such choices. As in part (i), this sequence is bounded and therefore has a convergent subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ whose limit $x = \lim_{n\to\infty} x_{\varphi(n)}$ is also on the interval $[a,b]$. By the choice of C (supremum of the values) and of x_n we have

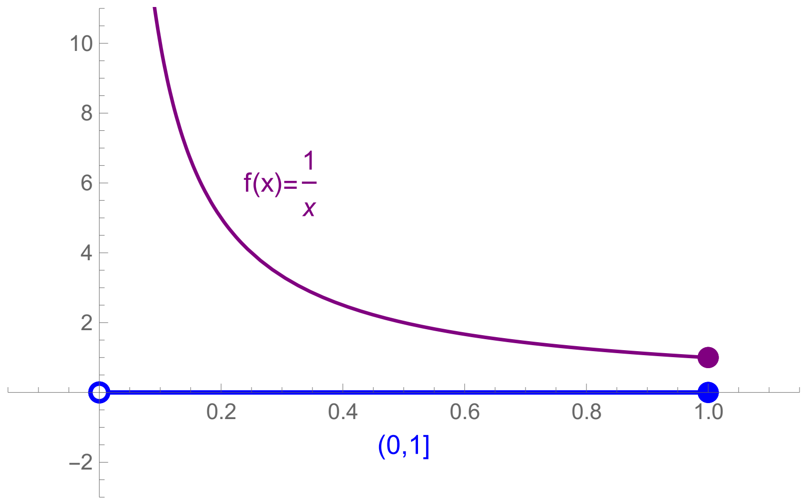
$$C \;\geq\; f(x_{\varphi(n)}) \;\geq\; C - \frac{1}{\varphi(n)} \;\geq\; C - \frac{1}{n}.$$

The squeeze theorem (Lemma II.17) thus gives $\lim_{n\to\infty} f(x_{\varphi(n)}) = C$. On the other hand, by continuity we have

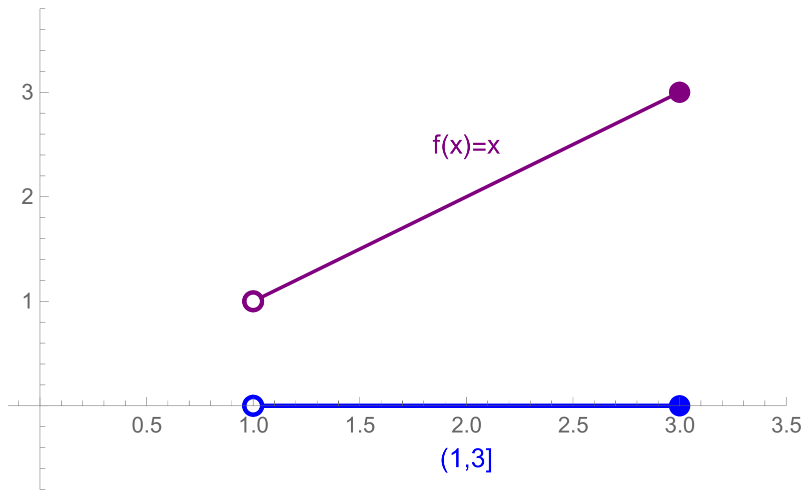
$$f(x) = f\big(\lim_{n\to\infty} x_{\varphi(n)}\big) = \lim_{n\to\infty} f(x_{\varphi(n)}) = C.$$

Having thus shown that

$$f(x) = \sup\big\{f(x) \;\big|\; x \in [a,b]\big\}\,,$$



(a) The function $f\colon (0,1] \to \mathbb{R}$ given by $f(x) = 1/x$ is continuous (since it is a rational function) but f is not bounded.



(b) The function $f\colon (1,3] \to \mathbb{R}$ given by $f(x) = x$ is continuous (since it is a polynomial function) but f has no minimum.

FIGURE III.4. Continuous functions on general intervals do not need to be bounded and do not need to have minima and maxima.

we conclude (as in Exercise II.18(b)) that $f\colon [a,b] \to \mathbb{R}$ indeed attains its maximum at the point $x \in [a,b]$. $\square$

Intermediate value theorem

The following property is related to connectedness (see Lecture XII) of the closed interval.

Theorem III.15 (Intermediate value theorem / Bolzano's theorem)**.**
Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function. Assume that $f(a)f(b) < 0$.[7] *Then there exists a point $z \in (a,b)$ such that $f(z) = 0$.*

Proof. We may assume that $f(a) < 0$ and $f(b) > 0$; the other case is similar (or is obtained from this one by considering the continuous function $-f$ with the same zeroes as f).

Consider the subset

$$N = \Big\{ x \in [a,b] \;\Big|\; f(x) < 0 \Big\}$$

of the interval $[a,b]$ where the function f is negative. Since $f(a) < 0$, we have at least $a \in N$, so this subset is non-empty, $N \neq \emptyset$. The subset is also bounded from above, because the right endpoint b of the interval is an upper bound for N. So let

$$z = \sup N.$$

Since $a \in N$ and z is the least upper bound for N, we have $a \leq z$. On the other hand, since b is an upper bound for N and z is the least upper bound, we have $z \leq b$. Combining these, we see that $z \in [a,b]$.

By a characterization of the supremum (Lemma II.29), there exists a sequence $(x_n^-)_{n\in\mathbb{N}}$ in the set N such that $\lim_{n\to\infty} x_n^- = z$. By continuity of f and by preservation of inequalities, we have

$$f(z) \;=\; f\big(\lim_{n\to\infty} x_n^-\big) \;=\; \lim_{n\to\infty} \underbrace{f(x_n^-)}_{<0} \;\leq\; 0.$$

Then consider (for example) the sequence (x_n^+) with members $x_n^+ = z + \frac{1}{n}$ (note that for large enough n we have $x_n^+ = z + \frac{1}{n} \leq b$, so with large enough indices the members of the sequence are in the domain $[a,b]$ of f). This sequence also has the property $\lim_{n\to\infty} x_n^+ = z$, and along it we have $f(x_n^+) \geq 0$ since $x_n^+ \notin N$. By continuity of f and by preservation of bounds, we have

$$f(z) \;=\; f\big(\lim_{n\to\infty} x_n^+\big) \;=\; \lim_{n\to\infty} \underbrace{f(x_n^+)}_{\geq 0} \;\geq\; 0.$$

Combining the two inequalities obtained above, $f(z) \leq 0$ and $f(z) \geq 0$, we conclude $f(z) = 0$. Since $f(a) < 0$ and $f(b) > 0$, this excludes the possibility that z is one of the two endpoints of the interval; we must therefore have $z \in (a,b)$. This finishes the proof. □

Corollary III.16 (Continuous images of intervals)**.**
If $I \subset \mathbb{R}$ is an interval and $f\colon I \to \mathbb{R}$ is continuous, then the image $f[I]$ is also an interval.

Proof. Let $f\colon I \to \mathbb{R}$ be a continuous function defined on an interval $I \subset \mathbb{R}$. To show that the image $f[I]$ is an interval, it suffices to show that for any two values $\alpha, \beta \in f[I]$ in the image with $\alpha < \beta$, any value $\gamma \in (\alpha,\beta)$ in between them also belongs to the image, $\gamma \in f[I]$.

So suppose $\alpha, \beta \in f[I]$ with $\alpha < \beta$, and let $\gamma \in (\alpha,\beta)$. By definition of images, there exists some $a, b \in I$ such that $f(a) = \alpha$ and $f(b) = \beta$. We have either $a < b$ or $a > b$. Both cases are handled similarly, so let us only present the details in the case $a < b$. Then by Exercise III.5 and Theorem III.11, the rule $x \mapsto f(x) - \gamma$ defines a continuous function on the closed subinterval $[a,b] \subset I$. Moreover, its values at the endpoints of $[a,b]$ have opposite signs: $f(a) - \gamma = \alpha - \gamma < 0$ and $f(b) - \gamma = \beta - \gamma > 0$. By Theorem III.15, then, there exists a point $z \in (a,b) \subset I$ such that $f(z) - \gamma = 0$, i.e., $f(z) = \gamma$. This shows $\gamma \in f[I]$. □

[7] This is just a concise way of saying that $f(a)$ and $f(b)$ are non-zero and have opposite signs: either $f(a) < 0$ and $f(b) > 0$, or $f(a) > 0$ and $f(b) < 0$.

Example III.17 (The image of an interval under the sine function)**.**
Consider the sine function

$$f_1\colon\ (0,\pi) \to \mathbb{R} \qquad\qquad f_1(x) = \sin(x) \quad \text{for } x \in (0,\pi).$$

This function is continuous.[8] The domain of definition is the open interval $(0,\pi)$, and its image under f_1 is the half-open interval $f_1\big[(0,\pi)\big] = (0,1]$.

If instead we considered the sine function on the longer interval

$$f_2\colon\ (0,2\pi) \to \mathbb{R} \qquad\qquad f_2(x) = \sin(x) \quad \text{for } x \in (0,2\pi),$$

then the image would be the closed interval $f_2\big[(0,2\pi)\big] = [-1,1]$.

The examples above show in particular that under a continuous function, the image of an open interval can be a closed or a half-open interval. Of course the image of an open interval can also be an open interval.[9]

However, by the next exercise the image of a closed interval under a continuous function is necessarily always a closed interval![10]

Exercise III.9 (The continuous image of a closed interval is a closed interval)**.**
Using the combination of Corollary III.16 and Theorem III.14, prove that if $f\colon [a,b] \to \mathbb{R}$ is continuous, then $f\big[[a,b]\big] \subset \mathbb{R}$ is a closed interval.

The following property is often used in non-degeneracy statements of formulas given by integrals; especially for certain function-space norms in the next chapter.

Exercise III.10 (Integrals of non-zero nonnegative functions are positive)**.**
Let $[a,b]$ be a closed interval which is nondegenerate, $a < b$. Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function.

(a) Show that if there exists an $x_0 \in [a,b]$ such that $f(x_0) > 0$, then there exists some $\delta > 0$ such that $f(x) > 0$ for all $x \in (x_0 - \delta, x_0 + \delta) \cap [a,b]$.
Hint: *Argue by contradiction: if this property is not true, you can construct a sequence which violates the assumed continuity of f.*

(b) Improve the conclusion of part (a) as follows. Show that if there exists an $x_0 \in [a,b]$ such that $f(x_0) > 0$, then there exists some $c > 0$ and some $\delta' > 0$ such that $f(x) > c$ for all $x \in (x_0 - \delta', x_0 + \delta') \cap [a,b]$.
Hint: *You can let $c = \frac{1}{2} f(x_0)$, and apply part (a) to the function defined by $\tilde{f}(x) = f(x) - c$.*

(c) Show that if f is non-negative, i.e., $f(x) \geq 0$ for all $x \in [a,b]$, and is not the constant function 0, then its integral is positive

$$\int_a^b f(x)\,\mathrm{d}x \;>\; 0. \tag{III.2}$$

Hint: *Use part (b) and a lower bound for the integral based on what you know of the values of f.*

[8] The appropriate tools for the precise justification of the continuity of trigonometric functions will be developed in Lecture IX; for now let us just admit the continuity as a fact.

[9] Consider for example the identity function $\mathrm{id}_{(a,b)}\colon (a,b) \to (a,b)$.

[10] The attentive reader may worry that constant functions are continuous and their images are singletons. This is indeed correct, but it does not contradict the assertion — a singleton is a special type of a closed interval, corresponding to the degenerate case when the interval endpoints coincide. Namely, when $a = b$, we have $[a,b] := \big\{x \in \mathbb{R} \,\big|\, a \leq x \leq b\big\} = \{a\}$.

Lecture IV

Normed spaces and inner product spaces

In this section we start to study more general spaces than just the real line $\mathbb{R}$. We will consider vector spaces with some additional structure, which suffices to determine a topology. Specifically, we consider vector spaces equipped with an inner product (and the norm induced by it) or with just a norm. Such setups are general enough to include a number of important spaces arising in applications.

The familiar d-dimensional Euclidean space $\mathbb{R}^d$ in particular has an inner product and a norm induced by it. But even this finite-dimensional vector space $\mathbb{R}^d$ could be equipped with some slightly different norms — and for some applications it is meaningful to do so. There are also straightforward infinite-dimensional counterparts of such finite-dimensional normed spaces.

Among the most important examples for applications of normed spaces and inner product spaces are various spaces of functions. We will already introduce a few function space examples that we keep returning to throughout the rest of this course, and that you are very likely to encounter in whatever field of mathematics you end up using in the future. But the number of relevant examples of such function spaces is so vast that we are in fact only scratching the surface of a deep topic: in courses on Hilbert spaces, Banach spaces, Measure theory, Probability theory, Partial differential equations, etc. you will encounter many more. The basic structures introduced here will be present in almost all of those![1]

IV.1. Vector spaces

The topic of vector spaces by itself belongs to algebra or linear algebra. But because vector spaces with some additional structure feature so crucially in analysis as well, we review the notion here. For concreteness, we only define *real vector spaces*, i.e., vector spaces over the field $\mathbb{R}$ of real numbers.[2] Algebraists would typically study vector spaces over arbitrary fields. The term **vector** will be used for elements of

[1] If one deliberately tries to find vector spaces in analysis that are *not* either inner product spaces or at least normed spaces, then among the easiest relevant examples are certain spaces of distributions, i.e., of generalized functions. Even they have a topology, and while the topology does not come from a norm, it is not too far from that — in typical cases it is obtained from a collection of seminorms.

[2] Let us admit, however, that also *complex vector spaces*, i.e., vector spaces over the field $\mathbb{C}$ of complex numbers, appear frequently in analysis, probability, etc. The most important additional structures they can be equipped with are, again, (complex) inner products and norms. The complex case would require only minor modifications to the treatment in this chapter, but we choose to focus on only the real case for the sake of clarity. We trust that when you will need to work with complex inner product spaces or normed spaces, your experience with the real counterparts makes it easy.

the vector space, and the term **scalar** for elements of the field over which the vector space is defined — in our case for real numbers.

Axioms of real vector spaces

For brevity, the term vector space will from here on always mean a real vector space — defined precisely by the following axiomatic properties.

Definition IV.1 (Vector space)**.**
A **(real) vector space** is a set V equipped with two operations,

$$\begin{array}{cc} \text{vector addition} & \text{scalar multiplication} \\ \mathsf{V} \times \mathsf{V} \to \mathsf{V} & \mathbb{R} \times \mathsf{V} \to \mathsf{V} \\ (\vec{u}, \vec{v}) \mapsto \vec{u} + \vec{v} & (c, \vec{v}) \mapsto c\,\vec{v}\,, \end{array}$$

such that the following properties hold:

Commutativity of vector addition:
$$\forall\, \vec{u}, \vec{v} \in \mathsf{V}: \quad \vec{u} + \vec{v} = \vec{v} + \vec{u} \tag{IV.1}$$
Associativity of vector addition:
$$\forall\, \vec{u}, \vec{v}, \vec{w} \in \mathsf{V}: \quad \vec{u} + (\vec{v} + \vec{w}) = (\vec{u} + \vec{v}) + \vec{w} \tag{IV.2}$$
Neutral element for vector addition:
$$\exists\, \vec{0} \in \mathsf{V}: \quad \forall\, \vec{v} \in \mathsf{V}: \quad \vec{v} + \vec{0} = \vec{v} \tag{IV.3}$$
Opposite vectors:
$$\forall\, \vec{v} \in \mathsf{V}: \quad \exists\, (-\vec{v}) \in \mathsf{V}: \quad \vec{v} + (-\vec{v}) = \vec{0} \tag{IV.4}$$
Compatibility of scalar multiplication:
$$\forall\, c, d \in \mathbb{R}: \quad \forall\, \vec{v} \in \mathsf{V}: \quad c\,(d\,\vec{v}) = (cd)\,\vec{v} \tag{IV.5}$$
Unit scalar multiplication:
$$\forall\, \vec{v} \in \mathsf{V}: \quad 1\,\vec{v} = \vec{v} \tag{IV.6}$$
Distributivity of vector addition:
$$\forall\, c \in \mathbb{R}: \quad \forall\, \vec{u}, \vec{v} \in \mathsf{V}: \quad c\,(\vec{u} + \vec{v}) = c\,\vec{u} + c\,\vec{v} \tag{IV.7}$$
Distributivity of scalar addition:
$$\forall\, c, d \in \mathbb{R}: \quad \forall\, \vec{v} \in \mathsf{V}: \quad (c + d)\,\vec{v} = c\,\vec{v} + d\,\vec{v}. \tag{IV.8}$$

These axioms are probably every bit as unsurprising as the field axioms of the real numbers given in Appendix B.1. But again, from such a simple general starting point, a very fruitful theory emerges.

Above, we used arrows on top of vectors, e.g., $\vec{u}, \vec{v}, \vec{w} \in \mathsf{V}$. In the long run, maintaining that practice would be cumbersome, so we will soon adopt the convention to label elements of the vector space without such decorations, e.g., by $u, v, w \in \mathsf{V}$. The case of the zero vector $\vec{0}$ perhaps deserves a special warning: it is commonly denoted by simply 0, and the reader is expected to figure out from the context whether 0 stands for the scalar zero $0 \in \mathbb{R}$ or the vector zero $0 \in \mathsf{V}$ (two very different objects).

First examples of real vector spaces

The most obvious example of a real vector space is the following.

Example IV.2 (The standard d-dimensional space $\mathbb{R}^d$)**.**
Let $d \in \mathbb{N}$. The set

$$\mathbb{R}^d = \Big\{ (x_1, \ldots, x_d) \;\Big|\; x_1, \ldots, x_d \in \mathbb{R} \Big\} \tag{IV.9}$$

is a real vector space, when equipped with the coordinatewise vector addition

$$(x_1, \ldots, x_d) + (y_1, \ldots, y_d) \;=\; (x_1 + y_1, \ldots, x_d + y_d) \qquad \text{for } (x_1, \ldots, x_d),\ (y_1, \ldots, y_d) \in \mathbb{R}^d$$

and coordinatewise scalar multiplication

$$c\,(x_1, \ldots, x_d) \;=\; (c\,x_1, \ldots, c\,x_d) \qquad \text{for } c \in \mathbb{R} \text{ and } (x_1, \ldots, x_d) \in \mathbb{R}^d$$

The zero vector is $\vec{0} = (0, 0, \ldots, 0) \in \mathbb{R}^d$, and the opposite vector to $(x_1, \ldots, x_d) \in \mathbb{R}^d$ is $(-x_1, \ldots, -x_d)$. Verifying the properties (IV.1) – (IV.8) for this case is straightforward.

The previous example has a straightforward infinite-dimensional counterpart as well.

Example IV.3 (The space of real-number sequences)**.**
The set[3]

$$\mathbb{R}^{\mathbb{N}} \;=\; \Big\{ (x_1, x_2, x_3, \ldots) \;\Big|\; x_1, x_2, x_3, \ldots \in \mathbb{R} \Big\} \tag{IV.10}$$

of all real-number sequences is a real vector space, when equipped with the coordinatewise vector addition

$$(x_1, x_2, x_3, \ldots) + (y_1, y_2, y_3, \ldots) \;=\; (x_1 + y_1,\ x_2 + y_2,\ x_3 + y_3, \ldots) \qquad \text{for } (x_1, x_2, x_3, \ldots),\ (y_1, y_2, y_3, \ldots) \in \mathbb{R}^{\mathbb{N}}$$

and coordinatewise scalar multiplication

$$c\,(x_1, x_2, x_3, \ldots) \;=\; (c\,x_1,\ c\,x_2,\ c\,x_3, \ldots) \qquad \text{for } c \in \mathbb{R} \text{ and } (x_1, x_2, x_3, \ldots) \in \mathbb{R}^{\mathbb{N}}.$$

The zero vector is the sequence of zeroes $\vec{0} = (0, 0, 0, \ldots)$, and the opposite vector to a sequence $(x_1, x_2, x_3, \ldots)$ is the sequence $(-x_1, -x_2, -x_3, \ldots)$. Verifying that the properties (IV.1) – (IV.8) is again straightforward.

Example IV.4 (Space of polynomials)**.**
Consider the space

$$\mathbb{R}[x] \;=\; \left\{ \sum_{k=0}^{d} c_k x^k \;\middle|\; d \in \mathbb{N}_0,\ c_0, c_1, \ldots, c_d \in \mathbb{R} \right\} \tag{IV.11}$$

of all polynomials in variable x, with real coefficients. The addition and scalar multiplication are defined coefficientwise, as usual for polynomials. It is easy to verify properties (IV.1) – (IV.8), so the space $\mathbb{R}[x]$ of real polynomials in one variable is a real vector space.[4]

---

[3]If X and Y are sets, the notation Y^X is often used for the set of all functions $X \to Y$. The notation $\mathbb{R}^{\mathbb{N}}$ correspondingly stands for all functions $\mathbb{N} \to \mathbb{R}$ — which in view of the precise definition of sequences is exactly the set of all real-number sequences.

[4]In fact, besides just addition and scalar multiplication of polynomials, we can also define multiplication of polynomials as usual. So the space $\mathbb{R}[x]$ has even further algebraic structure than just that of a vector space over $\mathbb{R}$ — it is what is known as an algebra over $\mathbb{R}$. We refer to courses on algebra for details.

Example IV.5 (Spaces of real-valued functions).
Let X be any set. Consider the space

$$\mathbb{R}^X \;=\; \Big\{ \text{ functions } f\colon X \to \mathbb{R} \Big\} \tag{IV.12}$$

of all functions whose domain is X and codomain is $\mathbb{R}$. We equip it with the *pointwise addition* and *pointwise scalar multiplication* as follows. If $f, g \in \mathbb{R}^X$, i.e., if $f, g\colon X \to \mathbb{R}$ are two functions, then we define the new function $f + g\colon X \to \mathbb{R}$ by the formula

$$\big(f+g\big)(x) \;=\; f(x) + g(x) \qquad \text{for } x \in X. \tag{IV.13}$$

Likewise, if $c \in \mathbb{R}$ is a scalar and $f \in \mathbb{R}^X$, i.e., $f\colon X \to \mathbb{R}$ is a function, then we define the new function $cf\colon X \to \mathbb{R}$ by the formula

$$\big(cf\big)(x) \;=\; c\,f(x) \qquad \text{for } x \in X. \tag{IV.14}$$

It is straightforward to check that the properties (IV.1) – (IV.8) hold, so the space $\mathbb{R}^X$ of real-valued functions on an arbitrary set X is a real vector space.[5]

In fact, Example IV.3 is readily a special case of the above, corresponding to the choice of domain $X = \mathbb{N}$: functions $\mathbb{N} \to \mathbb{R}$ are exactly real-number sequences.

Also Example IV.2 can be seen as a special case of this: $\mathbb{R}^d$ can be interpreted as the set of real-valued functions on the finite set $\{1, 2, \ldots, d\}$. A vector $(x_1, \ldots, x_d) \in \mathbb{R}$ corresponds to the function $f\colon \{1, 2, \ldots, d\} \to \mathbb{R}$ such that $f(j) = x_j$ for $j \in \{1, 2, \ldots, d\}$. This way we can identify the spaces $\mathbb{R}^d$ and $\mathbb{R}^{\{1,2,\ldots,d\}}$.

Vector subspaces

We saw a number of examples of vector spaces above, but other important examples are often constructed as vector subspaces of other vector spaces. The precise definition is the following.

Definition IV.6 (Vector subspace).
Let V be a vector space. A subset $\mathsf{W} \subset \mathsf{V}$ is called a **vector subspace** of V, if the following conditions hold:

- We have $\vec{0} \in \mathsf{W}$.
- If $\vec{u}, \vec{v} \in \mathsf{W}$, then we also have $\vec{u} + \vec{v} \in \mathsf{W}$.
- If $c \in \mathbb{R}$ and $\vec{v} \in \mathsf{W}$, then we also have $c\,\vec{v} \in \mathsf{W}$.

A vector subspace is therefore stable under both vector addition and scalar multiplication, and in particular naturally inherits these operations. The properties in Definition IV.1 remain valid on a subspace $\mathsf{W} \subset \mathsf{V}$, so in particular W is itself a vector space. This is an often practical way of constructing vector spaces: instead of checking the eight properties in Definition IV.1, it suffices to verify the three (usually easy) properties of Definition IV.6, if the space has been realized as a subset of some known vector space.

Example IV.7 (The subspace spanned by finitely many vectors).
If V is a vector space and $\vec{v}_1, \ldots, \vec{v}_n \in \mathsf{V}$ are vectors in it, then the set

$$\mathsf{W} \;=\; \left\{ c_1\,\vec{v}_1 + \cdots + c_n\,\vec{v}_n \;\middle|\; c_1, \ldots, c_n \in \mathbb{R} \right\} \tag{IV.15}$$

[5] In fact, besides just pointwise addition and scalar multiplication of real-valued functions, it is possible to also define pointwise multiplication of such functions. So the space $\mathbb{R}^X$ also has even more algebraic structure than just that of a vector space over $\mathbb{R}$ — it is also an algebra over $\mathbb{R}$.

of all linear combinations of them satisfies the properties in Definition IV.6, i.e., it is a vector subspace $\mathsf{W} \subset \mathsf{V}$. It is called the linear span of the vectors $\vec{v}_1, \ldots, \vec{v}_n$.[6]

Interesting function spaces are often defined as subspaces of the space of all functions with a given domain and a codomain: instead of considering the space of arbitrary functions, we usually care about sufficiently well-behaved functions. The following typical example is very important.

Example IV.8 (The space of continuous functions on an interval)**.**
Let $[a, b] \subset \mathbb{R}$ be a closed interval. Consider the set

$$\mathcal{C}\big([a, b]\big) \;=\; \Big\{\; f \colon [a, b] \to \mathbb{R} \text{ continuous function } \Big\}. \tag{IV.16}$$

This set of continuous real-valued functions on $[a, b]$ is obviously a subset of the space $\mathbb{R}^{[a,b]}$ of *all* real-valued functions on $[a, b]$ from Example IV.5. Since the zero-function (constant) is continuous, and since pointwise sums and scalar multiples of continuous functions are continuous (by Theorem III.11), we see that $\mathcal{C}([a, b])$ is a vector subspace in $\mathbb{R}^{[a,b]}$.

Let us finally give a few examples of subspaces of the sequence space of Example IV.3.

Example IV.9 (The space of absolutely summable sequences)**.**
Consider the subset

$$\ell^1 \;=\; \left\{ (x_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \sum_{n=1}^{\infty} |x_n| < \infty \right\} \;\subset\; \mathbb{R}^{\mathbb{N}}$$

of absolutely summable real sequences in the vector space $\mathbb{R}^{\mathbb{N}}$ of all real-number sequences (Example IV.3). The zero sequence $(0, 0, 0, \ldots)$ is of course absolutely summable, since $\sum_{n=1}^{\infty} |0| = \sum_{n=1}^{\infty} 0 = 0 < \infty$. Also if $(x_n)_{n\in\mathbb{N}} \in \ell^1$ is an absolutely summable sequence, then multiplying it by a scalar $c \in \mathbb{R}$ (coordinatewise), we get the sequence $c\,(x_n)_{n\in\mathbb{N}} = (cx_n)_{n\in\mathbb{N}}$ which is absolutely summable since

$$\sum_{n=1}^{\infty} |cx_n| \;=\; \sum_{n=1}^{\infty} |c|\,|x_n| \;=\; |c| \underbrace{\sum_{n=1}^{\infty} |x_n|}_{<\infty} \;<\; \infty.$$

Finally, if $(x_n)_{n\in\mathbb{N}}, (y_n)_{n\in\mathbb{N}} \in \ell^1$ are two absolutely summable sequences, then their sum (coordinatewise) is the sequence $(x_n)_{n\in\mathbb{N}} + (y_n)_{n\in\mathbb{N}} = \big(x_n + y_n\big)_{n\in\mathbb{N}}$ which is also absolutely summable since

$$\sum_{n=1}^{\infty} \underbrace{|x_n + y_n|}_{\leq |x_n|+|y_n|} \;\leq\; \sum_{n=1}^{\infty} \big(|x_n| + |y_n|\big) \;=\; \underbrace{\sum_{n=1}^{\infty} |x_n|}_{<\infty} + \underbrace{\sum_{n=1}^{\infty} |y_n|}_{<\infty} \;<\; \infty.$$

These three observations show that $\ell^1 \subset \mathbb{R}^{\mathbb{N}}$ is a vector subspace.

Exercise IV.1 (The space of bounded sequences)**.**
Consider the set

$$\ell^\infty \;=\; \Big\{ (x_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \;\Big|\; \exists r \geq 0 : \; \forall n \in \mathbb{N} : \; |x_n| \leq r \Big\} \;\subset\; \mathbb{R}^{\mathbb{N}}$$

of all bounded real-number sequences.

(a) Show that $\ell^\infty \subset \mathbb{R}^{\mathbb{N}}$ is a vector subspace.

---

[6]More generally, if $S \subset \mathsf{V}$ is any subset, then the subset consisting of all *finite* linear combinations of elements of S is still a vector subspace. This is denoted $\mathrm{span}(S) \subset \mathsf{V}$, and called the linear span of the subset $S \subset \mathsf{V}$ — it is the smallest vector subspace of V, which contains all vectors of S. Checking this is not difficult, but it is better suited for a course on linear algebra.

(b) Let ℓ^1 be as in Example IV.9. Show the inclusion $\ell^1 \subset \ell^\infty$, and conclude that the space of all absolutely summable real-number sequences is a vector subspace of the space of all bounded real-number sequences.

Exercise IV.2 (The space of convergent sequences).

Show that the set of those real-number sequences which are convergent is a vector subspace of the space ℓ^∞ of all bounded real-number sequences.

Example IV.10 (The space of sequences tending to zero).

Show that the set of those real-number sequences which converge to 0 is a vector subspace of the space of all convergent real-number sequences.

Exercise (✓) IV.3 (The set of sequences tending to a given value).

Show that the set of those real-number sequences which converge to $-\sqrt{\pi}$ is not a vector subspace of the space of all real-number sequences.

Hint: *Instead of $-\sqrt{\pi}$, you could use any non-zero value as the limit, and the conclusion would remain the same. This hints that even the easiest vector subspace property fails.*

IV.2. Normed spaces

A vector space, by itself, is a purely algebraic notion. Applying the vector addition inductively, one can define finite sums of vectors, but it would be meaningless to form infinite series, for example. Infinite series, as well as for example continuity of functions, require some additional structure — in fact something that gives rise to a topology on the vector space. A sufficient and often both intuitive and practical additional structure is a *norm*, which gives a notion of lengths of vectors. By considering length of differences, one then gets a notion of distances, and from distances one builds topological notions, as we will do in the subsequent lectures.

Axioms of normed spaces

Definition IV.11 (Normed space).

Let V be a real vector space. A function

$$\begin{aligned} \mathsf{V} &\to [0,\infty) \\ \vec{u} &\mapsto \|\vec{u}\| \end{aligned} \tag{IV.17}$$

is called a **norm** on V, if it satisfies the following conditions

(N1) $\|\vec{u}+\vec{v}\| \leq \|\vec{u}\| + \|\vec{v}\|$ for all $\vec{u},\vec{v} \in \mathsf{V}$; (**triangle inequality**)
(N2) $\|c\,\vec{u}\| = |c|\,\|\vec{u}\|$ for all $c \in \mathbb{R}$ and $\vec{u} \in \mathsf{V}$; (**homogeneity**)
(N3) $\|\vec{u}\| = 0$ only if $\vec{u} = \vec{0} \in \mathsf{V}$. (**non-degeneracy**)[7]

A vector space V equipped with a norm on it is called a **normed space**.

[7] We in fact have $\|\vec{u}\| = 0$ if and only if $\vec{u} = \vec{0}$. In (N3) we only require the "only if" implication to minimize work to check these properties in examples. The "if" direction follows from the other required properties. Namely, we have $\vec{0} = 0\,\vec{0}$ (you can derive this from the vector space axioms of Definition IV.1 as an algebraic exercise), so by (N2) we get

$$\|\vec{0}\| \;=\; \|0\,\vec{0}\| \;=\; |0|\,\|\vec{0}\| \;=\; 0\,\|\vec{0}\| \;=\; 0.$$

To concisely and unambiguously mention both the vector space V and the norm $\|\cdot\|$ on it, we may refer to the *pair* $(\mathsf{V}, \|\cdot\|)$ as a normed space.

Let us immediately mention the example that is probably familiar already:

Example IV.12 (The Euclidean norm)**.**
The d-dimensional Euclidean space is the vector space $\mathbb{R}^d$ of Example IV.2 equipped with the norm given by

$$\|\vec{x}\| \;=\; \sqrt{x_1^2 + \cdots + x_d^2} \qquad \text{for } \vec{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{IV.18}$$

One could verify the properties (N1), (N2), (N3) for (IV.18) directly, but we will also later see how they follow from a more general result.

Examples of normed spaces

Note that the same vector space could be equipped with many different norms. In fact, even the standard d-dimensional vector space $\mathbb{R}^d$ admits other norms besides the Euclidean norm of Example IV.12, and these are sometimes relevant in applications. We will start with some such examples, and then give perhaps more interesting examples of norms on (infinite-dimensional) function spaces and sequence spaces.

Examples of norms on finite-dimensional vector spaces

Example IV.13 (The ℓ^∞-norm on $\mathbb{R}^d$)**.**
Define, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$\|x\|_\infty \;=\; \max\big\{|x_1|, \ldots, |x_d|\big\} = \max_{j=1,\ldots,d} |x_j|. \tag{IV.19}$$

We call this the ℓ^∞**-norm** on $\mathbb{R}^d$; so let us verify that this indeed is a norm.

(N1): Let $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$ be two vectors in $\mathbb{R}^d$. By definition, their vector sum is $x+y = (x_1+y_1, \ldots, x_d+y_d)$. For any $j \in \{1, \ldots, d\}$ we have $|x_j+y_j| \le |x_j|+|y_j|$ by the triangle inequality of real numbers (Lemma II.1), and we may further estimate

$$|x_j + y_j| \;\le\; |x_j| + |y_j| \;\le\; \max_{j=1,\ldots,d} |x_j| + \max_{j=1,\ldots,d} |y_j|.$$

Since this estimate holds for each j, we get

$$\max_{j=1,\ldots,d} |x_j + y_j| \;\le\; \max_{j=1,\ldots,d} |x_j| + \max_{j=1,\ldots,d} |y_j|,$$

which is exactly property (N1) for the ℓ^∞ norm, $\|x + y\|_\infty \le \|x\|_\infty + \|y\|_\infty$.

(N2): Let $c \in \mathbb{R}$ and $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. By definition of scalar multiplication, we have $c\,x = (c\,x_1, \ldots, c\,x_d)$. For any $j \in \{1, \ldots, d\}$ we have

$$|c\,x_j| \;=\; |c|\,|x_j| \;\le\; |c| \max_{j=1,\ldots,d} |x_j|.$$

Since this estimate holds for each j, we get

$$\max_{j=1,\ldots,d} |c\,x_j| \;\le\; |c| \max_{j=1,\ldots,d} |x_j|,$$

and by considering the index j for which the maximal $|x_j|$ is attained, we see that equality in fact holds here. The equality is exactly property (N2) for the ℓ^∞ norm, $\|c\,x\|_\infty = |c|\,\|x\|_\infty$.

(N3): If for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ we have $\|x\|_\infty = 0$, i.e., $\max\{|x_1|, \ldots, |x_d|\} = 0$, then for each j we must have $x_j = 0$, which implies that $x = (0, \ldots, 0)$. This establishes (N3).

Exercise IV.4 (The ℓ^1-norm on $\mathbb{R}^d$)**.**
Define, for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$,

$$\|x\|_1 \;=\; |x_1| + \cdots + |x_d| \;=\; \sum_{j=1}^{d} |x_j|. \tag{IV.20}$$

We call this the ℓ^1**-norm** on $\mathbb{R}^d$. Prove that it satisfies (N1), (N2), (N3).

Exercise IV.5 (The "unit circles" in $\mathbb{R}^2$ with different norms)**.**
Consider the two-dimensional vector space $\mathbb{R}^2$ equipped with the norms of Example IV.13 and Exercise IV.4.

(a) Draw the "unit circle" with respect to the ℓ^∞-norm,

$$\Big\{(x,y) \in \mathbb{R}^2 \;\Big|\; \|(x,y)\|_\infty = 1\Big\}.$$

(b) Draw the "unit circle" with respect to the ℓ^1-norm,

$$\Big\{(x,y) \in \mathbb{R}^2 \;\Big|\; \|(x,y)\|_1 = 1\Big\}.$$

Exercise (✓) IV.6 (A non-norm)**.**
Consider the function

$$\mathbb{R}^2 \to [0,+\infty) \qquad \text{given by} \qquad \mathbb{R}^2 \ni (x_1, x_2) \mapsto |x_1 + x_2|.$$

Which of the properties (N1), (N2), (N3) hold for it? Is it a norm on the vector space $\mathbb{R}^2$?

More generally, one may consider the following ℓ^p-norms.

Example IV.14 (The ℓ^p-norm on $\mathbb{R}^d$)**.**
Let $p \in [1, \infty)$ Define, for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$,

$$\|x\|_p \;=\; \Big(\sum_{j=1}^{d} |x_j|^p\Big)^{1/p}. \tag{IV.21}$$

We call this the ℓ^p**-norm** on $\mathbb{R}^d$. It satisfies (N1), (N2), (N3), but verifying (N1) is more difficult than in the cases above (Exercise IV.7 is for those who like a challenge). This is a generalization which includes, at specific values of $p \geq 1$, some cases we have seen already:

- In the case $p = 1$, from (IV.21) we clearly recover the ℓ^1-norm (IV.20) of Exercise IV.4.
- In the case $p = 2$, from (IV.21) we recover the Euclidean norm of Example IV.12,

$$\|x\|_2 \;=\; \Big(\sum_{j=1}^{d} x_j^2\Big)^{1/2} \;=\; \sqrt{\langle x, x\rangle}\,.$$

- In Exercise IV.8 you will show that in the limit $p \to \infty$, we recover the ℓ^∞-norm of Example IV.13.

Exercise (♯) IV.7 (The ℓ^p-norm is a norm)**.**
Prove that (IV.21) defines a norm on $\mathbb{R}^d$.

Exercise IV.8 (The ℓ^p-norm as $p \to \infty$)**.**
On $\mathbb{R}^d$, consider the ℓ^∞-norm $\|\cdot\|_\infty$ of Example IV.13, and the ℓ^p-norms $\|\cdot\|_p$ of Example IV.14, for $p \geq 1$.

(a) Prove that for any $\vec{x} \in \mathbb{R}^d$ and any $p \geq 1$, we have $\|\vec{x}\|_\infty \leq \|\vec{x}\|_p$.
(b) Prove that for any $\vec{x} \in \mathbb{R}^d$ and any $p \geq 1$, we have $\|\vec{x}\|_p \leq d^{1/p}\,\|\vec{x}\|_\infty$.
(c) Use (a), (b), and the calculation of $\lim_{p\to\infty} d^{1/p}$ to prove that for any $\vec{x} \in \mathbb{R}^d$ we have

$$\lim_{p\to\infty} \|\vec{x}\|_p \;=\; \|\vec{x}\|_\infty.$$

By now, we have given many examples of different norms on the vector space $\mathbb{R}^d$. These norms are, however, in some sense comparable to each other. As an example, you may prove:

Exercise IV.9 (A comparison of the ℓ^1-norm and the ℓ^∞-norm)**.**
Let $\|\cdot\|_\infty$ and $\|\cdot\|_1$ denote the norms of Example IV.13 and Exercise IV.4 on $\mathbb{R}^d$. Prove that for any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we have

$$\|x\|_\infty \;\leq\; \|x\|_1 \;\leq\; d\,\|x\|_\infty. \tag{IV.22}$$

We will later see other similar comparisons between the different norms.[8]

Examples of norms on function spaces

Also many important function spaces are equipped with norms. These are, arguably, the most important normed spaces for applications in analysis.

Let us start with an extremely fundamental example that will feature in many applications in this course, too.

Example IV.15 (The supremum norm on the space of continuous functions)**.**
Consider the vector space $\mathcal{C}([a,b])$ of continuous real-valued functions on a closed interval $[a,b] \subset \mathbb{R}$ as in Example IV.15. For $f \in \mathcal{C}([a,b])$, let us define[9]

$$\begin{aligned} \|f\|_\infty \;&=\; \sup\Big\{|f(t)| \;\Big|\; t \in [a,b]\Big\} \\ &=\; \sup_{t\in[a,b]} |f(t)|. \end{aligned} \tag{IV.23}$$

Note that since the continuous function f on a closed interval $[a,b]$ is bounded (Theorem III.14(i)), the supremum here is finite, $\|f\|_\infty \in [0,\infty)$. In fact, since the continuous function $t \mapsto |f(t)|$ on the closed interval $[a,b]$ attains its maximum (Theorem III.14(ii)), we could equivalently have written $\|f\|_\infty = \max\big\{|f(t)| \;\big|\; t \in [a,b]\big\}$. In particular, we at least have $\|f\|_\infty \in [0,\infty)$ as required for a norm.

We call $\|\cdot\|_\infty$ the the **supremum norm**, the **sup-norm**, or the **uniform norm** — but let us verify that it indeed is a norm.

(N1): Let $f, g \in \mathcal{C}([a,b])$. Recall that addition $f+g$ in $\mathcal{C}([a,b])$ is defined pointwise by $(f+g)(t) = f(t) + g(t)$. For any $t \in [a,b]$, using the triangle inquality of real numbers (Lemma II.1), we get

$$\begin{aligned} \big|(f+g)(t)\big| \;&=\; \big|f(t) + g(t)\big| \\ &\leq\; |f(t)| + |g(t)| \;\leq\; \sup_{t\in[a,b]} |f(t)| + \sup_{t\in[a,b]} |g(t)| \;=\; \|f\|_\infty + \|g\|_\infty. \end{aligned}$$

Since this holds for any $t \in [a,b]$, we also get

$$\|f+g\|_\infty \;=\; \sup_{t\in[a,b]} \big|(f+g)(t)\big| \;\leq\; \|f\|_\infty + \|g\|_\infty,$$

which is exactly the triangle inequality (N1) for the norm $\|\cdot\|_\infty$.

(N2): Let $f \in \mathcal{C}([a,b])$ and $c \in \mathbb{R}$. Recall that scalar multiplication cf in $\mathcal{C}([a,b])$ is defined pointwise by $(cf)(t) = c\,f(t)$. For $t \in [a,b]$, we have

$$\big|(cf)(t)\big| \;=\; \big|c\,f(t)\big| \;=\; |c|\,|f(t)| \;\leq\; |c| \sup_{t\in[a,b]} |f(t)| \;=\; |c|\,\|f\|_\infty,$$

[8] After Lecture XI we will also have the tools to prove generally that *all* possible norms on a finite-dimensional space are necessarily comparable.

[9] The second line of the formula here means exactly the same as the first line, but it is a more concise notation and as such often very convenient.

and since this holds for all $t \in [a,b]$, we get

$$\|cf\|_\infty \;=\; \sup_{t\in[a,b]} \big|(cf)(t)\big| \;\leq\; |c|\,\|f\|_\infty.$$

This inequality is only a half of the desired equality (N2), but we can use it as follows. If $c = 0$, then $cf = 0$ is the constant function zero, so clearly $\|cf\|_\infty = \|0\|_\infty = 0 = |0|\,\|f\|_\infty$ as desired. Assume therefore that $c \neq 0$. Then by using the already established inequality with scalar $\frac{1}{c}$ and function cf, we get have

$$\|f\|_\infty \;=\; \left\|\frac{1}{c}\,(cf)\right\|_\infty \;\leq\; \frac{1}{|c|}\,\|cf\|_\infty,$$

which upon rearranging yields the inequality $\|cf\|_\infty \geq |c|\,\|f\|_\infty$ in the other direction. The equality (N2) follows.

(N3): If $f \in \mathcal{C}([a,b])$ satisfies $\|f\|_\infty = 0$, then $\sup_{t\in[a,b]} |f(t)| = 0$ so we cannot have $|f(t)| > 0$ for any t, i.e., we have $|f(t)| = 0$ for every t, and f is the constant function zero.

Exercise (✓) IV.10 (Another non-norm)**.**
Consider the function

$$\mathcal{C}([-1,+1]) \to [0,+\infty) \qquad \text{given by} \qquad f \mapsto |f(0)|.$$

Which of the properties (N1), (N2), (N3) hold for it? Is it a norm on the vector space $\mathcal{C}([-1,+1])$ of continuous functions on the interval $[-1,+1]$?

Exercise IV.11 (Yet another norm on the space of continuous functions)**.**
Consider again the vector space $\mathcal{C}\left([a,b]\right)$ of continuous real-valued functions on a nondegenerate closed interval $[a,b] \subset \mathbb{R}$. For $f \in \mathcal{C}\left([a,b]\right)$, let us define

$$\|f\|_1 \;=\; \int_a^b |f(t)|\,\mathrm{d}t. \tag{IV.24}$$

Verify that $\|\cdot\|_1$ is a norm on $\mathcal{C}\left([a,b]\right)$. We call it the L^1**-norm**.
Hint: *For one of the properties, using Exercise III.10 is the key.*

Examples of normed sequence spaces

Exercise IV.12 (The sequence space ℓ^1)**.**
Consider the vector space

$$\ell^1 \;=\; \left\{ x = (x_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \sum_{n=1}^{\infty} |x_n| < \infty \right\}$$

of absolutely summable real sequences (Example IV.9). For $x = (x_n)_{n\in\mathbb{N}} \in \ell^1$, define

$$\big\|x\big\|_1 \;=\; \sum_{n=1}^{\infty} |x_n|. \tag{IV.25}$$

Show that $\|\cdot\|_1$ is a norm on ℓ^1.

Exercise IV.13 (The sequence space ℓ^∞)**.**
Consider the vector space

$$\ell^\infty \;=\; \left\{ x = (x_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \exists r \geq 0 :\; \forall n \in \mathbb{N} :\; |x_n| \leq r \right\}$$

of bounded real sequences (Exercise IV.1). For $x = (x_n)_{n\in\mathbb{N}} \in \ell^\infty$, define

$$\big\|x\big\|_\infty \;=\; \sup\left\{ |x_n| \;\middle|\; n \in \mathbb{N} \right\}. \tag{IV.26}$$

Show that $\|\cdot\|_\infty$ is a norm on ℓ^∞.

IV.3. Inner product spaces

Section IV.2 was concerned with how to equip a vector space with a norm, which gives the lenghts of vectors, and therefore eventually a topology. We next consider a stronger structure, which gives a bit more geometric structure to vector spaces. Namely, we introduce inner products on (real) vector spaces, and we will see that inner products always induce norms, and that they moreover give rise to a notion of angles between vectors.

Examples of inner product spaces will include both familiar finite-dimensional cases and some (infinite-dimensional) function spaces.

Axioms of inner product spaces

Definition IV.16 (Inner product space)**.**
Let V be a real vector space. A binary operation

$$\begin{aligned} \mathsf{V} \times \mathsf{V} &\to \mathbb{R} \\ (\vec{u}, \vec{v}) &\mapsto \langle \vec{u}, \vec{v} \rangle \end{aligned} \tag{IV.27}$$

is called an **inner product** on V if it satisfies the following conditions:

(IP1) $\langle \vec{u}, \vec{v} \rangle = \langle \vec{v}, \vec{u} \rangle$ for all $\vec{u}, \vec{v} \in \mathsf{V}$; (**symmetricity**)
(IP2) $\langle c\,\vec{u}, \vec{v} \rangle = c\,\langle \vec{u}, \vec{v} \rangle$ for all $c \in \mathbb{R}$ and $\vec{u}, \vec{v} \in \mathsf{V}$; (**homogeneity**)
(IP3) $\langle \vec{u} + \vec{v}, \vec{w} \rangle = \langle \vec{u}, \vec{w} \rangle + \langle \vec{v}, \vec{w} \rangle$ for all $\vec{u}, \vec{v}, \vec{w} \in \mathsf{V}$; (**additivity**)
(IP4) $\langle \vec{u}, \vec{u} \rangle \geq 0$ for all $\vec{u} \in \mathsf{V}$; (**positive semi-definiteness**)
(IP5) $\langle \vec{u}, \vec{u} \rangle = 0$ only if $\vec{u} = \vec{0} \in \mathsf{V}$. (**non-degeneracy**)

A vector space V equipped with an inner product on it is called an **inner product space**.

To concisely and unambiguously mention both the vector space V and the inner product $\langle \cdot, \cdot \rangle$ on it, we may refer to the *pair* $(\mathsf{V}, \langle \cdot, \cdot \rangle)$ as an inner product space.

The conditions (IP2) and (IP3) say, respectively, that the inner product respects scalar multiplication and vector addition in its first argument. Together these amount to the inner product being **linear** in its first argument: we have

$$\langle c_1 \vec{u}_1 + c_2 \vec{u}_2,\, \vec{v} \rangle = c_1 \langle \vec{u}_1, \vec{v} \rangle + c_2 \langle \vec{u}_2, \vec{v} \rangle \qquad \text{for all } \vec{u}_1, \vec{u}_2, \vec{v} \in \mathsf{V} \text{ and } c_1, c_2 \in \mathbb{R}.$$

If we furthermore combine this with symmetricity (IP1), we also get

$$\langle \vec{u},\, d_1 \vec{v}_1 + d_2 \vec{v}_2 \rangle = d_1 \langle \vec{u}, \vec{v}_1 \rangle + d_2 \langle \vec{u}, \vec{v}_2 \rangle \qquad \text{for all } \vec{u}, \vec{v}_1, \vec{v}_2 \in \mathsf{V} \text{ and } d_1, d_2 \in \mathbb{R}.$$

This yields **bilinearity** of the inner product: an inner product is linear in both of its arguments.

The condition (IP5) requires that we have $\langle \vec{u}, \vec{u} \rangle = 0$ only if $\vec{u} = \vec{0}$. We could equally well have written "if and only if" here; the converse direction direction follows from the other properties already.[10] Namely, in any vector space we have $\vec{0} = 0\,\vec{0}$, so using (IP2) we get that the inner product of the zero vector $\vec{0}$ with any vector $\vec{v}$

[10] So why did we formulate the condition (IP5) in an apparently weaker way? As the two choices lead to logically equivalent definitions, this is ultimately a matter of taste. But including fewer requirements in definitions is considered preferable, because when one wants to construct examples (in this case of inner product spaces), there are fewer conditions to check (explicitly).

vanishes: $\langle \vec{0}, \vec{v} \rangle = \langle 0\,\vec{0}, \vec{v} \rangle = 0\,\langle \vec{0}, \vec{v} \rangle = 0$. In particular, the zero vector has zero inner product with itself, giving the "if" implication. So in an inner product space we have

$$\langle \vec{u}, \vec{u} \rangle = 0 \quad \Longleftrightarrow \quad \vec{u} = \vec{0}.$$

Examples of inner product spaces

Let us again start from the most familiar example.

Example IV.17 (Euclidean spaces)**.**
In the vector space $\mathbb{R}^d$ of Example IV.2, the formula

$$\langle \vec{x}, \vec{y} \rangle \; = \; \sum_{j=1}^{d} x_j y_j \qquad \text{for } \; \vec{x} = (x_1, \ldots, x_d), \; \vec{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d \tag{IV.28}$$

defines the usual inner product — below we will verify that the five properties indeed hold for it. The corresponding inner product space is called the d-dimensional **Euclidean space**.

In this familiar case, the properties (IP1) – (IP5) can be checked as follows.

(IP1): For any $\vec{x} = (x_1, \ldots, x_d), \vec{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d$ we have

$$\langle \vec{x}, \vec{y} \rangle \; = \; \sum_{j=1}^{d} x_j y_j \; = \; \sum_{j=1}^{d} y_j x_j \; = \; \langle \vec{y}, \vec{x} \rangle.$$

(IP2): For any $\vec{x} = (x_1, \ldots, x_d), \vec{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d$ and $c \in \mathbb{R}$, we have first of all $c\,\vec{x} = (cx_1, \ldots, cx_d)$ and then

$$\langle c\,\vec{x}, \vec{y} \rangle \; = \; \sum_{j=1}^{d} (c\,x_j)\, y_j \; = \; c \sum_{j=1}^{d} x_j y_j \; = \; c\,\langle \vec{x}, \vec{y} \rangle.$$

(IP3): For any $\vec{x} = (x_1, \ldots, x_d), \vec{y} = (y_1, \ldots, y_d), \vec{z} = (z_1, \ldots, z_d) \in \mathbb{R}^d$, we have first of all $\vec{x} + \vec{y} = (x_1 + y_1, \ldots, x_d + y_d)$ and then

$$\begin{aligned} \langle \vec{x} + \vec{y}, \; \vec{z} \rangle \; &= \; \sum_{j=1}^{d} (x_j + y_j)\, z_j \; = \; \sum_{j=1}^{d} \big(x_j z_j + y_j z_j \big) \\ &= \; \sum_{j=1}^{d} x_j z_j + \sum_{j=1}^{d} y_j z_j \; = \; \langle \vec{x}, \vec{z} \rangle + \langle \vec{y}, \vec{z} \rangle. \end{aligned}$$

(IP4): For any $\vec{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ we have, by nonnegativity of squares of real numbers,

$$\langle \vec{x}, \vec{x} \rangle \; = \; \sum_{j=1}^{d} x_j x_j \; = \; \sum_{j=1}^{d} x_j^2 \; \geq \; 0,$$

(IP5): From the formula above, we see that $\langle \vec{x}, \vec{x} \rangle$ can only be zero for $\vec{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, if for each $j \in \{1, \ldots, d\}$ we have $x_j^2 = 0$. This only occurs if each component x_j is zero and the vector is $\vec{x} = (0, \ldots, 0) = \vec{0}$.

Let us then give an infinite dimensional example.

Example IV.18 (An inner product in a function space)**.**
Consider the vector space $\mathcal{C}([a,b])$ of continuous real-valued functions on a closed interval $[a,b] \subset \mathbb{R}$, which is nondegenerate, $a < b$. Define $\langle \cdot, \cdot \rangle \colon \mathcal{C}([a,b]) \times \mathcal{C}([a,b]) \to \mathbb{R}$ by

$$\langle f, g \rangle \;=\; \int_a^b f(t)g(t)\,\mathrm{d}t \qquad \text{for } \; f, g \in \mathcal{C}\big([a,b]\big). \tag{IV.29}$$

This is called the L^2-inner product of functions. Let us verify that this is indeed an inner product on $\mathcal{C}([a,b])$. Note that the verifications are very similar to those in Example IV.17; only in (IP5) an essentially new observation is used.

(IP1): For any $f, g \in \mathcal{C}([a,b])$ we have

$$\langle f, g \rangle \;=\; \int_a^b f(t)g(t)\,\mathrm{d}t \;=\; \int_a^b g(t)f(t)\,\mathrm{d}t \;=\; \langle g, f \rangle.$$

(IP2): For any $f, g \in \mathcal{C}([a,b])$ and $c \in \mathbb{R}$, recalling the definition $(cf)(t) = c\,f(t)$ of pointwise scalar multiplication of functions, we get

$$\langle c\,f, g \rangle \;=\; \int_a^b \big(cf\big)(t)\; g(t)\;\mathrm{d}t \;=\; \int_a^b c\,f(t)\,g(t)\;\mathrm{d}t \;=\; c\int_a^b f(t)g(t)\,\mathrm{d}t \;=\; c\,\langle f, g \rangle.$$

(IP3): For any $f, g, h \in \mathcal{C}([a,b])$, recalling the definition $(f+g)(t) = f(t) + g(t)$ of pointwise addition of functions, we get

$$\begin{aligned}
\langle f + g, h \rangle \;&=\; \int_a^b \big(f+g\big)(t)\; h(t)\;\mathrm{d}t \\
&=\; \int_a^b \Big(f(t) + g(t)\Big)\; h(t)\;\mathrm{d}t \\
&=\; \int_a^b \Big(f(t)h(t) + g(t)h(t)\Big)\;\mathrm{d}t \\
&=\; \int_a^b f(t)h(t)\,\mathrm{d}t + \int_a^b g(t)h(t)\,\mathrm{d}t \;=\; \langle f, h \rangle + \langle g, h \rangle.
\end{aligned}$$

(IP4): For any $f \in \mathcal{C}([a,b])$, we have

$$\langle f, f \rangle \;=\; \int_a^b \underbrace{f(t)f(t)}_{=\,f(t)^2\,\geq\,0}\;\mathrm{d}t \;\geq\; 0.$$

(IP5): From the formula above, we see that $\langle f, f \rangle$ can only be zero for $f \in \mathcal{C}([a,b])$ if $\int_a^b f(t)^2\,\mathrm{d}t = 0$. Since $t \mapsto f(t)^2$ is a continuous nonnegative function on $[a,b]$, by the contrapositive of Exercise III.10, the vanishing of this integral implies that $f(t)^2 = 0$ for all $t \in [a,b]$. But this implies that $f(t) = 0$ for all $t \in [a,b]$, i.e., that f is the zero function.

The combination of the following two exercises amount to a complete characterization of inner products in finite-dimensional vector spaces.

Exercise IV.14 (Constructing inner products on $\mathbb{R}^d$)**.**
Let $A \in \mathbb{R}^{d\times d}$ be a symmetric[11] positive definite[12] $d \times d$ matrix. Show that the formula $\langle x, y \rangle = x^\top A\, y$ for $x, y \in \mathbb{R}^d$ defines an inner product on $\mathbb{R}^d$.

Exercise IV.15 (Characterizing inner products on $\mathbb{R}^d$)**.**
Let $\langle \cdot, \cdot \rangle$ be an inner product on $\mathbb{R}^d$ (not necessarily the standard Euclidean inner product, but any inner product that satisfies the requirements in Definition IV.16). Let $\vec{e}_1, \ldots, \vec{e}_d$

[11]A square matrix $A \in \mathbb{R}^{d\times d}$ is symmetric if it is its own transpose, $A^\top = A$.

[12]A square matrix $A \in \mathbb{R}^{d\times d}$ is said to be **positive semi-definite** if $x^\top A\, x \geq 0$ for all vectors $x \in \mathbb{R}^d$, and **positive definite** if $x^\top A\, x > 0$ for all nonzero vectors $x \in \mathbb{R}^d \setminus \{\vec{0}\}$.

be the standard basis vectors of $\mathbb{R}^d$. For all $i, j \in \{1, \ldots, d\}$, define $a_{i,j} = \langle \vec{e}_i, \vec{e}_j \rangle$, and let $A \in \mathbb{R}^{d \times d}$ denote the matrix with these entries $a_{i,j}$.

(a) Show that with this definition, the inner product of any two vectors $x, y \in \mathbb{R}^d$ can be written as $\langle x, y \rangle = x^\top A\, y$.
(b) Prove that the matrix A is symmetric and positive definite.

The norm induced by an inner product

An inner product in particular yields a notion of lenght of vectors, i.e., a norm. In an inner product space V, we define the **norm** $\|v\|$ of $v \in \mathsf{V}$ as

$$\|v\| \;=\; \sqrt{\langle v, v \rangle}. \tag{IV.30}$$

To emphasize that (IV.30) is not just any norm but specifically related to the inner product, we call it *the norm induced by the inner product* $\langle \cdot, \cdot \rangle$. After some preliminaries, we will show that (IV.30) indeed gives a norm, i.e., that it satisfies the properties (N1), (N2), (N3).

Example IV.19 (The Euclidean norm is iduced by the Euclidean inner product)**.**
The norm of Example IV.12 on the d-dimensional vector space $\mathbb{R}^d$ is exactly the norm induced by the inner product of Example IV.17: indeed the Euclidean norm of $\vec{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ reads

$$\|\vec{x}\| \;=\; \sqrt{\langle \vec{x}, \vec{x} \rangle} \;=\; \sqrt{x_1^2 + \cdots + x_d^2}\,.$$

Example IV.20 (The L^2-norm on the space of continuous functions on a closed interval)**.**
Recall from Example IV.18 that the space $\mathcal{C}([a, b])$ of continuous real-valued functions on a nondegenerate closed interval $[a, b] \subset \mathbb{R}$ is an inner product space with the inner product (IV.29). We denote the norm induced by it by $\|\cdot\|_2$. Explicitly for a continuous function $f\colon [a, b] \to \mathbb{R}$, this norm reads

$$\|f\|_2 \;=\; \sqrt{\langle f, f \rangle} \;=\; \left(\int_a^b f(t)^2 \,\mathrm{d}t \right)^{1/2}.$$

The following exercise gives a property that is specific to norms induced by inner products. It is not valid in general normed spaces, as the examples afterwards demonstrate.

Exercise IV.16 (Parallelogram rule)**.**
Let V be an inner product space. Show that for any two vectors $u, v \in \mathsf{V}$ we have the following parallelogram rule

$$\|u + v\|^2 + \|u - v\|^2 \;=\; 2\,\|u\|^2 + 2\,\|v\|^2. \tag{IV.31}$$

We also note that the inner products can be recovered from the norms they induce.

Exercise IV.17 (Polarization formula)**.**
Let V be an inner product space, and let $u, v \in \mathsf{V}$ be two vectors in it. Find a formula for the inner product $\langle u, v \rangle$ using only the norms of the vectors $u + v$ and $u - v$.

The parallelogram rule can be used to show that some norms are not induced by any inner product.

Exercise IV.18 (The ℓ^∞-norm is not induced by an inner product).
Consider the norm $\|\cdot\|_\infty$ of Example IV.13 on $\mathbb{R}^d$ with $d \geq 2$. Find two vectors $u, v \in \mathbb{R}^d$ such that $\|u+v\|_\infty^2 + \|u-v\|_\infty^2 \neq 2\,\|u\|_\infty^2 + 2\,\|v\|_\infty^2$. Use Exercise IV.16 to conclude that the norm $\|\cdot\|_\infty$ is not induced by any inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^d$.

Exercise IV.19 (The ℓ^1-norm is not induced by an inner product).
Consider the norm $\|\cdot\|_1$ of Exercise IV.4 on $\mathbb{R}^d$ with $d \geq 2$. Show that the norm $\|\cdot\|_1$ is not induced by any inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^d$.

Exercise IV.20 (The sup-norm is not induced by an inner product).
Prove that if the interval $[a,b]$ is non-degenerate, $a<b$, then the sup-norm $\|\cdot\|_\infty$ on $\mathcal{C}([a,b])$ is not induced by any inner product.

Exercise IV.21 (The L^1-norm is not induced by an inner product).
Prove that if the interval $[a,b]$ is non-degenerate, $a<b$, then the L^1-norm $\|\cdot\|_1$ on $\mathcal{C}([a,b])$ is not induced by any inner product.

A crucial property of inner products and the norms induced by them is the following inequality.

Theorem IV.21 (Cauchy–Schwarz inequality in inner product spaces).
Let V be an inner product space. Then for any $u, v \in \mathsf{V}$ we have

$$\big|\langle u, v\rangle\big| \;\leq\; \|u\|\,\|v\|. \tag{IV.32}$$

Proof. If either $u=0$ or $v=0$, then both sides of (IV.32) are zero and the inequality holds trivially. We may therefore assume that $u, v \in \mathsf{V}\setminus\{0\}$.

Observe that for any $t \in \mathbb{R}$, by (IP4) for the vector $u+t\,v \in \mathsf{V}$ we have $0 \leq \langle u+t\,v,\; u+t\,v\rangle$. Using bilinearity and symmetricity of the inner product, we can write this as

$$\begin{aligned} 0 &\leq \langle u+t\,v,\; u+t\,v\rangle \\ &= \langle u,u\rangle + \langle u, t\,v\rangle + \langle t\,v, u\rangle + \langle t\,v, t\,v\rangle \\ &= \|u\|^2 + 2t\,\langle u,v\rangle + t^2\,\|v\|^2. \end{aligned}$$

This last expression is a polynomial at^2+bt+c, with coefficients $a=\|v\|^2$, $b=2\,\langle u,v\rangle$, and $c=\|u\|^2$. The fact that polynomial is non-negative for all $t\in\mathbb{R}$ means that it has either no real roots or just one real root, so its discriminant is non-positive, $b^2-4ac\leq 0$. By plugging in the coefficients a,b,c, this reads

$$4\,\langle u,v\rangle^2 - 4\,\|u\|^2\,\|v\|^2 \;\leq\; 0.$$

Dividing by 4 and rearranging we get $\langle u,v\rangle^2 \leq \|u\|^2\,\|v\|^2$. Then by taking square roots we get (IV.32). □

Exercise IV.22 (When does equality occur in the Cauchy–Schwarz inequality?).

(a) Show that if $u, v \in \mathsf{V}$ are such that $u = c\,v$ for some $c \in \mathbb{R}$, then equality occurs in the Cauchy–Schwarz inequality (IV.32) for these vectors.
(b) By carefully examining the steps of the proof of Theorem IV.21, show that for two vectors $u, v \in \mathsf{V}$, equality can occur in the Cauchy–Schwarz inequality only if one of them is a real scalar multiple of the other.

The Cauchy–Schwarz inequality is a tool used in numerous circumstances, and already the following analogue of Lemma II.1 indicates its importance: it gives property (N1) for the norm induced by an inner product.

Corollary IV.22 (Triangle inequality in inner product spaces)**.**
Let V *be an inner product space. Then for any* $u, v \in \mathsf{V}$ *we have*

$$\|u+v\| \;\leq\; \|u\| + \|v\|. \tag{IV.33}$$

Proof. The proof is very similar to that of Lemma II.1. We first expand the square of $\|u+v\|$ using bilinearity, and then estimate it using (IV.32) as follows

$$\begin{aligned}
\|u+v\|^2 &= \langle u+v,\ u+v\rangle && \text{(definition)}\\
&= \langle u,u\rangle + \langle u,v\rangle + \langle v,u\rangle + \langle v,v\rangle && \text{(bilinearity)}\\
&= \|u\|^2 + 2\,\langle u,v\rangle + \|v\|^2 && \text{(symmetricity)}\\
&\leq \|u\|^2 + 2\,\big|\langle u,v\rangle\big| + \|v\|^2 && \text{(taking absolute value)}\\
&\leq \|u\|^2 + 2\,\|u\|\,\|v\| + \|v\|^2 && \text{(Cauchy–Schwarz)}\\
&= \big(\|u\| + \|v\|\big)^2. && \text{(binomial formula)}
\end{aligned}$$

Taking the square roots, we get (IV.33). □

Corollary IV.23 (The norm induced by an inner product is a norm)**.**
In an inner product space $(\mathsf{V}, \langle\cdot,\cdot\rangle)$*, the formula* (IV.30) *defines a norm (in the sense of Definition IV.11).*

Proof. Indeed, for the norm $\|v\| = \sqrt{\langle v,v\rangle}$ induced by the inner product, we already saw in Corollary IV.22 that (N1) holds. The other two properties are easy. From (IP2) and (IP1) we get $\langle c\,v, c\,v\rangle = c^2\,\langle v,v\rangle$, so

$$\|c\,v\| = \sqrt{c^2\,\langle v,v\rangle} = |c|\,\sqrt{\langle v,v\rangle} = |c|\,\|v\|,$$

establishing (N2). The property (N3) follows directly from (IP5). □

Remark IV.24 (All inner product spaces are normed spaces)**.**
It follows from Corollary IV.23 above that any inner product space becomes a normed space, when it is equipped with the norm induced by the inner product as in (IV.30).

Let us also indicate some direct applications of the Cauchy–Schwarz inequality.

Exercise IV.23 (A sum bound based on Cauchy–Schwarz inequality)**.**
Let $x_1, \ldots, x_d \in \mathbb{R}$. Prove the following inequality

$$|x_1| + \cdots + |x_d| \;\leq\; \sqrt{d}\,\sqrt{x_1^2 + \cdots + x_d^2}.$$

Hint: *Consider the Euclidean space* $\mathbb{R}^d$*, and apply the Cauchy–Schwarz inequality with* $u = (1,1,\ldots,1)$ *and a suitably chosen* v*.*

Example IV.25 (Cauchy–Schwarz inequality in a function space)**.**
Recall from Examples IV.18 and IV.20 that the space $\mathcal{C}([a,b])$ of continuous real-valued functions on a nondegenerate closed interval $[a,b] \subset \mathbb{R}$ is an inner product space. The Cauchy–Schwarz inequality in the inner product space $\mathcal{C}([a,b])$ amounts to the following integral inequality

$$\left|\int_a^b f(t)\,g(t)\,\mathrm{d}t\right| \;\leq\; \left(\int_a^b f(t)^2\,\mathrm{d}t\right)^{1/2} \left(\int_a^b g(t)^2\,\mathrm{d}t\right)^{1/2}$$

for all continuous functions $f,g\colon [a,b] \to \mathbb{R}$.

As a special case, if $h \in \mathcal{C}([a,b])$ and we take $f = |h|$ and $g = 1$, we get

$$\int_a^b |h(t)|\,\mathrm{d}t \;\leq\; \sqrt{b-a}\,\left(\int_a^b h(t)^2\,\mathrm{d}t\right)^{1/2}.$$

Exercise IV.24 (A comparison of the ℓ^1-norm and the Euclidean norm)**.**
Let $\|\cdot\|$ denote the ordinary Euclidean norm on $\mathbb{R}^d$, induced by the standard inner product, and let $\|\cdot\|_1$ denote the ℓ^1-norm of Exercise IV.4. Prove that for any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we have

$$\|x\| \;\leq\; \|x\|_1 \;\leq\; \sqrt{d}\,\|x\|.$$

Hint: *Recall Exercise IV.23.*

Exercise IV.25 (A comparison of the ℓ^∞-norm and the Euclidean norm)**.**
Show that there exists a constant $C_d > 0$ such that for any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we have

$$\frac{1}{C_d}\,\|x\| \;\leq\; \|x\|_\infty \;\leq\; C_d\,\|x\|.$$

Exercise IV.26 (The sequence space ℓ^2)**.**
Consider the set

$$\ell^2 \;=\; \left\{ x = (x_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \sum_{n=1}^{\infty} x_n^2 < \infty \right\}$$

of square-summable real sequences. Use the Cauchy–Schwarz inequality in $\mathbb{R}^d$ and considerations of suitable limits to show that first of all $\ell^2 \subset \mathbb{R}^{\mathbb{N}}$ is a vector subspace and moreover that ℓ^2 can be equipped with an inner product.

Angles and orthogonality

Let V be an inner product space.

Suppose that $u, v \in \mathsf{V} \setminus \{0\}$ are two non-zero vectors. Then by the Cauchy–Schwarz inequality (Theorem IV.21) we have

$$-1 \;\leq\; \frac{\langle u, v\rangle}{\|u\|\,\|v\|} \;\leq\; 1.$$

The function $\cos\colon [0,\pi] \to [-1,1]$ is bijective, so there exists a unique $\theta \in [0,\pi]$ such that $\frac{\langle u,v\rangle}{\|u\|\,\|v\|} = \cos(\theta)$, i.e., such that

$$\langle u, v\rangle \;=\; \cos(\theta)\,\|u\|\,\|v\|. \tag{IV.34}$$

This $\theta \in [0,\pi]$ is called the **angle** between the vectors u and v. In the Euclidean spaces $\mathbb{R}^2$ and $\mathbb{R}^3$ this of course coincides with the familiar notion of (undirected[13]) angle.

The case of the right angle $\theta = \frac{\pi}{2}$ is special. Since $\cos\left(\frac{\pi}{2}\right) = 0$, it occurs when the inner product between two vectors vanishes. Correspondingly, we make the following definitions.

Definition IV.26 (Orthogonality and orthonormality)**.**
Two vectors $u, v \in \mathsf{V}$ are said to be **orthogonal** if their inner product vanishes, $\langle u, v\rangle = 0$.

A collection $S \subset \mathsf{V}$ is said to be **orthogonal** if any two different vectors $u, v \in S$, $u \neq v$, are orthogonal.

[13] In the plane $\mathbb{R}^2$ there is a convention about what is the positive orientation, and it is natural to define also the directed angle from $u \in \mathbb{R}^2 \setminus \{0\}$ to $v \in \mathbb{R}^2 \setminus \{0\}$, which is a number $\phi \in (-\pi, \pi]$. But the sign of this directed angle changes if we interchange u and v. In a general inner product space, two linearly independent vectors u, v span a plane, but this plane may be given either one of two possible orientations. The undirected angle does not depend on the choice of the orientation.

A collection $S \subset \mathsf{V}$ is said to be **orthonormal** if it is an orthogonal collection and all $v \in S$ have unit norm, $\|v\| = 1$.

The cases of angles $\theta = 0$ and $\theta = \pi$ are also special: they correspond to vectors that lie on the same line through the origin.

Lemma IV.27 (Collinear vectors)**.**
Let $u, v \in \mathsf{V} \setminus \{0\}$ be two non-zero vectors in an inner product space V, and let θ be the angle between them. Then we have

- *$\theta = 0$ if and only if $u = c\,v$ for some $c > 0$;*
- *$\theta = \pi$ if and only if $u = c\,v$ for some $c < 0$.*

Exercise IV.27 (Proof of Lemma IV.27)**.**
Prove the two statements made in Lemma IV.27.
Hint: *The "if" directions of both statements are easy. For the "only if" statements, use Exercise IV.22.*

Exercise IV.28 (The dihedral angle of a regular tetrahedron)**.**
Calculate the dihedral angle (the angle between two adjacent faces) of a regular tetrahedron using the following idea:

Place the vertices of the tetrahedron in the 4-dimensional space $\mathbb{R}^4$ at points $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$, and $\vec{e}_4$ (standard basis vectors).

(i) Verify that the lengths of the edges $\|\vec{e}_i - \vec{e}_j\|$, for $i \neq j$, are all equal.
(ii) By symmetry, the angle in question is the same as (for example) the angle between the vectors
$$\vec{u} = \vec{e}_4 - \frac{1}{2}(\vec{e}_1 + \vec{e}_2) \qquad \text{and} \qquad \vec{v} = \vec{e}_3 - \frac{1}{2}(\vec{e}_1 + \vec{e}_2),$$
so it is easily calculated.

Hint: *Draw a figure! Answer:* $\arccos(1/3)$.

There are various function spaces that have inner products, and in many cases appropriate orthogonal polynomials in such spaces play an important role. To get some flavor of this, the following exercise deals with the first few[14] Legendre polynomials.

Exercise IV.29 (Some Legendre polynomials)**.**
Consider the three first *Legendre polynomials*
$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ P_2(x) &= (3x^2 - 1)/2. \end{aligned}$$

(a) Show that these polynomials are pairwise *orthogonal* with respect to the inner product
$$\langle f, g \rangle = \int_{-1}^{1} f(x) g(x) \, \mathrm{d}x,$$
i.e., that we have $\langle P_i, P_j \rangle = 0$ for $i \neq j$.
(b) Determine coefficients $c_0, c_1, c_2 \in \mathbb{R}$ such that $\|c_k P_k\| = 1$ for all $k \in \{0, 1, 2\}$, where as norm we use the one arising from the inner product in part (a).

[14] If you are more ambitious, you can try to continue and find polynomials of higher degree which satisfy similar orthogonality.

Also the idea of Fourier series is based on orthogonality of trigonometric functions in a suitable function space.

Exercise IV.30 (Ideas behind Fourier series)**.**

(a) Consider a countable orthonormal collection $(e_n)_{n\in\mathbb{N}}$ in an inner product space V. Assume that a vector $v \in \mathsf{V}$ can be expressed, for some $m \in \mathbb{N}$, as a (finite) linear combination

$$v = \alpha_1 \, e_1 + \cdots + \alpha_m \, e_m \qquad \text{with some coefficients } \alpha_n \in \mathbb{R}.$$

Calculate the inner products between the vector v and the vectors e_n from the orthonormal collection, and deduce a formula for the coefficients α_n.

(b) Define functions $c_0, s_1, c_1, s_2, c_2, s_3, c_3, \ldots$ of a real variable x by the formulas

$$\begin{aligned} c_0(x) &= \frac{1}{\sqrt{2}}, \\ s_n(x) &= \sin(nx) \qquad (\text{for } n \in \mathbb{N}), \\ c_n(x) &= \cos(nx) \qquad (\text{for } n \in \mathbb{N}). \end{aligned}$$

Prove that the (countably infinite) collection of these functions is orthonormal in the space $\mathcal{C}([-\pi,\pi])$ of continuous functions on $[-\pi,\pi]$, with respect to the rescaled L^2 inner product

$$\langle f, g \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) g(x) \, \mathrm{d}x.$$

(c) The idea of Fourier series is founded on the following result: Every continuously differentiable 2π-periodic function $f \colon \mathbb{R} \to \mathbb{R}$ can be represented as

$$f(x) = \alpha_0 + \sum_{n=1}^{\infty} \big(\alpha_n \cos(nx) + \beta_n \sin(nx) \big).$$

Since we have not yet addressed infinite series, consider for simplicity the case of a function f, for which the above series contains only finitely many terms, i.e., for some $m \in \mathbb{N}$ we have

$$f(x) = \alpha_0 + \sum_{n=1}^{m} \big(\alpha_n \cos(nx) + \beta_n \sin(nx) \big).$$

Using parts (a) and (b), derive the following formula for the coefficients

$$\alpha_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) \, \mathrm{d}x \qquad (n \in \mathbb{N}),$$

and find similar formulas for the coefficients β_n ($n \in \mathbb{N}$), and α_0.

Lecture V

Metric spaces

In this lecture we reach the level of generality at which we will work almost exclusively from here on: we introduce the notion of a *metric space*. The spaces we have considered so far — the real line, the Euclidean space, more general inner product spaces and normed spaces — all turn out to be at least metric spaces. But there is a vast amount of other relevant examples of metric spaces, too! What is important is that fruitful general theory for metric spaces can be developed, which can then be applied to all of the specific cases.

In the subsequent lectures we will start developing that theory. The goal of this lecture is to just define metric spaces, give sufficiently many interesting examples of them, and introduce a few first concepts that make sense in general metric spaces.

V.1. Axioms of metric spaces

Intuitively, a metric space is a space where there is a reasonable notion of *distance* between points. The distances are encoded in a function called a metric. The only requirements, informally described, are: *(symmetricity)* that the distance between two points is the same both ways, *(triangle inequality)* that the distance from one point to another cannot be reduced by taking a "shortcut" through a third point, and *(separation)* that the distance between two different points is always strictly positive, while the distance from a point to itself is zero.

The precise definition is the following.

Definition V.1 (Metric space)**.**
Let X be a set. A function

$$\mathsf{d}\colon X \times X \to [0,\infty) \tag{V.1}$$

is a **metric** on X if the following conditions hold:

(M-s): For any $x, y \in X$ we have ("symmetricity")

$$\mathsf{d}(x,y) \;=\; \mathsf{d}(y,x). \tag{M-s}$$

(M-Δ): For any $x, y, z \in X$ we have ("triangle inequality")

$$\mathsf{d}(x,z) \;\leq\; \mathsf{d}(x,y) + \mathsf{d}(y,z). \tag{M-Δ}$$

(M-0): We have ("separation")

$$\mathsf{d}(x,y) = 0 \;\text{ if and only if }\; x = y. \tag{M-0}$$

A set X equipped with a metric d is called a **metric space**.

To concisely and unambiguously mention both the set X and the metric d on it, we may refer to the *pair* (X, d) as a metric space.

V.2. Examples of metric spaces

Metric spaces are very general, and most spaces that we have seen so far are in fact also metric spaces. Let us start from something familiar and safe.

Example V.2 (The real line is a metric space)**.**
On the real line $\mathbb{R}$, we use the familiar notion of distances between points, encoded in the metric $\mathsf{d}_{\mathbb{R}} \colon \mathbb{R} \times \mathbb{R} \to [0, \infty)$ on the set $\mathbb{R}$ given by the formula

$$\mathsf{d}_{\mathbb{R}}(x, y) \;=\; |y - x| \qquad \text{for } x, y \in \mathbb{R}. \tag{V.2}$$

Checking the properties (M-s), (M-Δ), (M-0) directly is not difficult (it is a good exercise!), but the properties also follow from the more general Remark V.5 below.

We call (V.2) the **standard metric on** $\mathbb{R}$. Unless otherwise mentioned, we consider the real axis $\mathbb{R}$ equipped with this metric.

Although the above standard metric on $\mathbb{R}$ is the reasonable one that we typically use without explicit mention, let us take this opportunity to point out that it is *possible* to equip the set $\mathbb{R}$ also with some other more exotic metrics.

Exercise V.1 (Another metric on the real line)**.**
Prove that the formula

$$\mathsf{d}(x, y) \;=\; \log\Big(1 + |x - y|\Big)$$

also defines a metric $\mathsf{d} \colon \mathbb{R} \times \mathbb{R} \to [0, \infty)$ on the set $\mathbb{R}$.

How creative can you get with metrics on the line?

Exercise V.2 (Yet other metrics on the real line)**.**
Find further examples of metrics on $\mathbb{R}$, besides those in Example V.2 and Exercise V.1.
<u>Hint</u>: *First try to come up with some creative new formula! If that is too hard, a simple solution is to "change the units of distances" by multiplying a given metric*[1] *by some positive constant. Another idea that always works is to "allow for teleportation at a fixed cost".*[2][3] *Yet another idea is to define a simple $\{0, 1\}$-valued metric (see Example V.7).*

All of this is to illustrate that the same set can be equipped with many possible metrics. But on $\mathbb{R}$ the standard metric (V.2) is of course the most important choice!

Let us then look at the next most familiar case.

Example V.3 (The Euclidean spaces are metric spaces)**.**
In the d-dimensional Euclidean space $\mathbb{R}^d$, the usual notion of distances comes from the

---

[1]Note that you can apply the "change of units" to for example the standard metric or the more exotic metric on Exercise V.1.

[2]Example VII.39 will give a precise meaning to this.

[3]Also the "teleportation" modification can be done to either the standard metric or some more exotic metric to start with.

Euclidean norm $\|\cdot\|_2$ of (IV.18); the corresponding metric $\mathsf{d}_{\mathbb{R}^d}\colon \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ is given by the familiar formula

$$\mathsf{d}_{\mathbb{R}^d}(\vec{x},\vec{y}) \;=\; \|\vec{y}-\vec{x}\|_2 \;=\; \sqrt{(y_1-x_1)^2+\cdots+(y_d-x_d)^2} \tag{V.3}$$

for $\vec{x}=(x_1,\ldots,x_d)\in\mathbb{R}^d$ and $\vec{y}=(y_1,\ldots,y_d)\in\mathbb{R}^d$. One can check the properties (M-s), (M-Δ), (M-0) directly, but they also follow from the more general Remark V.5 below.

When we refer to $\mathbb{R}^d$ as a Euclidean space, we consider it equipped with this metric (V.3).

Just as with the real line, the Euclidean metric above is the usual reasonable choice, but there are other possible metrics as well. Let us give just one such example for now.

Example V.4 (The Manhattan metric in the plane)**.**
In the plane $\mathbb{R}^2$, the formula

$$\mathsf{d}\big((x,y),\,(x',y')\big) \;=\; |x'-x|+|y'-y| \qquad \text{for } (x,y),\,(x',y')\in\mathbb{R}^2 \tag{V.4}$$

defines a metric $\mathsf{d}\colon \mathbb{R}^2\times\mathbb{R}^2\to[0,\infty)$ — the required properties (M-s), (M-Δ), (M-0) can again be obtained from Remark V.5 applied to the norm $\|\cdot\|_1$ of Exercise IV.4.

The metric (V.4) is sometimes referred to as the *Manhattan metric*, because the distance between two points is obtained as if one had to separately move along horizontal "streets" (the term $|x'-x|$) and vertical "avenues" (the term $|y'-y|$).

In the examples above we did not do detailed checks of the properties (M-s), (M-Δ), (M-0). The reason is that they can all be handled at once with the following general observation, which in fact immediately gives also many further examples of metric spaces.

Remark V.5 (All normed spaces are metric spaces)**.**
Suppose that $(\mathsf{V},\|\cdot\|)$ is a normed space. Then the **metric induced by the norm** $\|\cdot\|$ is the metric $\mathsf{d}_{\mathsf{V}}\colon \mathsf{V}\times\mathsf{V}\to[0,\infty)$ defined by the formula

$$\mathsf{d}_{\mathsf{V}}(\vec{u},\vec{v}) \;=\; \|\vec{v}-\vec{u}\| \qquad \text{for } \vec{u},\vec{v}\in\mathsf{V}. \tag{V.5}$$

Let us check that d_{V} is indeed a metric on V in the sense of Definition V.1.

(M-s): Let $\vec{u},\vec{v}\in\mathsf{V}$. Noting that $\vec{u}-\vec{v}=(-1)\,(\vec{v}-\vec{u})$ and using the homogeneity property (N2) of norms, we straightforwardly get the desired symmetricity of d_{V}

$$\mathsf{d}_{\mathsf{V}}(\vec{v},\vec{u}) \;=\; \|\vec{u}-\vec{v}\| \;=\; \big\|(-1)\,\big(\vec{v}-\vec{u}\big)\big\| \;=\; |-1|\,\|\vec{v}-\vec{u}\| \;=\; \|\vec{v}-\vec{u}\| \;=\; \mathsf{d}_{\mathsf{V}}(\vec{u},\vec{v}).$$

(M-Δ): Let $\vec{u},\vec{v},\vec{w}\in\mathsf{V}$. Then using the triangle inequality (N1) for the norm $\|\cdot\|$, we get the desired triangle inequality of d_{V} as follows

$$\begin{aligned}
\mathsf{d}_{\mathsf{V}}(\vec{u},\vec{w}) \;&=\; \|\vec{w}-\vec{u}\| && \text{(definition of } \mathsf{d}_{\mathsf{V}}) \\
&=\; \big\|(\vec{w}-\vec{v})+(\vec{v}-\vec{u})\big\| && \text{(added and subtracted } \vec{v}) \\
&\le\; \|\vec{w}-\vec{v}\|+\|\vec{v}-\vec{u}\| && \text{(N1)} \\
&=\; \mathsf{d}_{\mathsf{V}}(\vec{v},\vec{w})+\mathsf{d}_{\mathsf{V}}(\vec{u},\vec{v})\;. && \text{(definition of } \mathsf{d}_{\mathsf{V}})
\end{aligned}$$

(M-0): Suppose that $\vec{u},\vec{v}\in\mathsf{V}$ are such that $\mathsf{d}_{\mathsf{V}}(\vec{u},\vec{v})=0$. By definition of d_{V}, this means that $\|\vec{v}-\vec{u}\|=0$. By property (N3) of the norm $\|\cdot\|$, this implies that $\vec{v}-\vec{u}=0$. Adding $\vec{u}$ to both sides we find $\vec{v}=\vec{u}$, as desired. The converse implication is easy: we have $\mathsf{d}_{\mathsf{V}}(\vec{v},\vec{v})=\|\vec{v}-\vec{v}\|=\|\vec{0}\|=0$ for any $\vec{v}\in\mathsf{V}$.

The verification of the properties of metric in Examples V.2, V.3, and V.4 now follow directly by applying Remark V.5 to the normed spaces $(\mathbb{R}, |\cdot|)$, $(\mathbb{R}^d, \|\cdot\|_2)$, and $(\mathbb{R}^2, \|\cdot\|_1)$, respectively.

We also get lots of other examples — just check again every example of a normed space from Lecture IV (including all examples of inner product spaces, in view of Remark IV.24). Let us just take one such example explicitly.

Example V.6 (A metric on the space of continuous function on a closed interval)**.**
The space $\mathcal{C}([a,b])$ on continuous real-valued functions on a closed interval $[a,b]$, equipped with the supremum norm $\|\cdot\|_\infty$ of Example IV.15, is a normed space — and thus in particular a metric space by Remark V.5. The metric (V.5) induced by the supremum norm (IV.23) takes the form

$$\mathsf{d}(f,g) \;=\; \|g-f\|_\infty \;=\; \sup_{t\in[a,b]} \big|g(t)-f(t)\big| \qquad \text{for } f,g\in\mathcal{C}([a,b]). \tag{V.6}$$

Let us finally mention an extremely simple metric you could use on any set.

Example V.7 (The 0/1 metric)**.**
Let X be any set. The formula

$$\mathsf{d}_{0/1}(x,y) \;=\; \begin{cases} 0 & \text{if } x=y \\ 1 & \text{if } x\neq y \end{cases} \qquad \text{for } x,y\in X \tag{V.7}$$

is easily seen to define a metric on X (Exercise V.3). It is called the 0/1**-metric** on X. It equips the set X with a structure that in some cases (discrete spaces) is reasonable, but in other cases is bizarre — there are no points “near each other”, the distance between all pairs of distinct points is the same!

Nevertheless, $(X, \mathsf{d}_{0/1})$ is a valid example of a metric space, it has some relevant applications, and it often provides simple and illuminating counterexamples.

Exercise (✓) V.3 (The 0/1 metric is a metric)**.**
Check the properties (M-s), (M-Δ), (M-0) for the metric $\mathsf{d}_{0/1}$ defined in Example V.7.

Metric inherited to subsets

Let (X, d) be a metric space, and $X' \subset X$ a subset. Then the subset naturally inherits a metric from the whole space — since d gives distances between all pairs of points in X, it certainly also gives distances between all pairs of points in the subset $X' \subset X$. Formally, the **inherited metric** $\mathsf{d}' \colon X' \times X' \to [0,\infty)$ on X' is given by

$$\mathsf{d}'(x',y') \;=\; \mathsf{d}(x',y') \qquad \text{for } x',y'\in X'\subset X, \tag{V.8}$$

i.e., the restriction of $\mathsf{d} \colon X\times X \to [0,\infty)$ to the subset $X'\times X' \subset X\times X$. The properties (M-s), (M-Δ), and (M-0) for d' directly follow as special cases of the corresponding properties of d.

We now get lots of new examples of metric spaces!

Example V.8 (Subsets of the real line are metric spaces)**.**
Since the real line $\mathbb{R}$ is a metric space (Example V.2), all its subsets also become metric spaces with the metrics they inherit by (V.8).

In particular, any of the following are metric spaces:

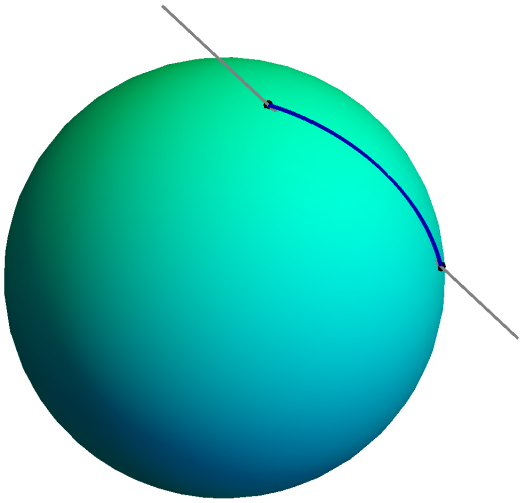


FIGURE V.1. The unit sphere $\mathbf{S}^2 \subset \mathbb{R}^3$ inherits a metric from $\mathbb{R}^3$, but it also has another natural metric corresponding to the minimal length of paths along the surface.

- intervals such as $[0,1] \subset \mathbb{R}$, $(-\infty, 0) \subset \mathbb{R}$, $(-\pi, \pi] \subset \mathbb{R}$;
- the set of rational numbers $\mathbb{Q} \subset \mathbb{R}$;
- the set of irrational numbers $\mathbb{R} \setminus \mathbb{Q} \subset \mathbb{R}$;
- the Cantor set $C \subset \mathbb{R}$ (Appendix B.3);
- etc. etc.

Example V.9 (Subsets of the plane are metric spaces)**.**

Since the Euclidean plane $\mathbb{R}^2$ is a metric space (Example V.3), all its subsets also become metric spaces with the metrics they inherit by (V.8).

In particular, all lines, circles, curves, rectangles, triangles, disks, half-spaces, cones, the Koch snowflake, the Sierpinski carpet, ... in the plane $\mathbb{R}^2$ are metric spaces.

You have to admit that the number of relevant examples of metric spaces is vast.

Since we happen to live on this (approximately) spherical planet called Earth, and we care about distances between places on its surface (say Otaniemi and Paris), let us still briefly comment on the topic of metrics on the sphere.

Example V.10 (The unit sphere in $\mathbb{R}^3$)**.**

Consider the unit sphere

$$\mathbf{S}^2 = \Big\{ (x,y,z) \in \mathbb{R}^3 \;\Big|\; x^2 + y^2 + z^2 = 1 \Big\} \subset \mathbb{R}^3$$

in the 3-dimensional space $\mathbb{R}^3$. Since the Euclidean space $\mathbb{R}^3$ is a metric space (Example V.3), the sphere $\mathbf{S}^2$ also inherits a metric as its subset. Note, however, that in this metric inherited from $\mathbb{R}^3$, the distances between points on the sphere are their shortest distances in the 3-dimensional space (i.e., along straight lines) — see the gray line in Figure V.1.

This notion of distances on $\mathbf{S}^2$ is relevant for example for neutrinos passing through the Earth (admitting that we only use neutrino detectors relatively close to the surface of the Earth).

Example V.11 (Another metric on the unit sphere in $\mathbb{R}^3$)**.**
For the purposes of, say, aviation or nautical seafaring, the natural metric on the sphere

$$\mathbf{S}^2 \;=\; \Big\{(x,y,z)\in\mathbb{R}^3 \;\Big|\; x^2+y^2+z^2=1\Big\} \;\subset\; \mathbb{R}^3$$

would be given by shortest distances along the surface of the sphere — see the blue path in Figure V.1. The easiest precise definition of this metric is in terms of the infimum of the arclengths of paths in $\mathbf{S}^2$ connecting the two points, see Appendix B.4, but we leave the details to the interested reader.

Already Examples V.8, V.9, V.10, and V.11 as well as Example V.7 give a lot of metric spaces which are *not* normed spaces. Among these simple examples, Examples V.11 and V.7 are not even subsets of a normed space equipped with the metric induced by the norm as in (V.5). The abundance of such relevant examples shows an advantage of the more general concept of metric spaces compared to that of normed spaces.

V.3. First notions in metric spaces

We would like to introduce some notions that are immediately meaningful in metric spaces. Before that, let us however record the following lemma as a preparation.

Lemma V.12 (Triangle inequality lower bound)**.**
Let (X,d) be a metric space, and $x,y,z\in X$ any three points. Then we have

$$\mathsf{d}(x,z) \geq \big|\mathsf{d}(x,y)-\mathsf{d}(y,z)\big|. \tag{V.9}$$

Proof. Note that $\big|\mathsf{d}(x,y)-\mathsf{d}(y,z)\big|$ equals either $\mathsf{d}(x,y)-\mathsf{d}(y,z)$ or $\mathsf{d}(y,z)-\mathsf{d}(x,y)$, depending on the sign of the quantity inside the absolute value. We will separately prove both

$$\mathsf{d}(x,y)-\mathsf{d}(y,z) \;\leq\; \mathsf{d}(x,z) \qquad \text{and} \tag{V.10}$$
$$\mathsf{d}(y,z)-\mathsf{d}(x,y) \;\leq\; \mathsf{d}(x,z). \tag{V.11}$$

Together these will show the asserted inequality $\big|\mathsf{d}(x,y)-\mathsf{d}(y,z)\big| \leq \mathsf{d}(x,z)$.

First apply the triangle inequality (M-Δ) to the three points $x,y,z\in X$ in the form

$$\mathsf{d}(x,y) \;\leq\; \mathsf{d}(x,z)+\mathsf{d}(z,y) \qquad \text{(“going via } z \text{ is not a shortcut from } x \text{ to } y\text{”)}.$$

Subtracting $\mathsf{d}(z,y)$ from both sides yields $\mathsf{d}(x,y)-\mathsf{d}(y,z)\leq\mathsf{d}(x,z)$. Since $\mathsf{d}(z,y)=\mathsf{d}(y,z)$ by (M-s), this establishes (V.10) — the first of the two inequalities we set out to prove.

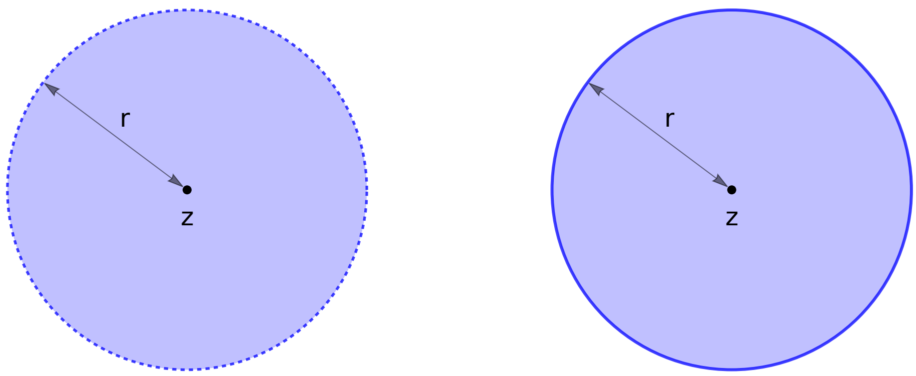
Then apply the triangle inequality (M-Δ) to the three points $x,y,z\in X$ in the form

$$\mathsf{d}(y,z) \;\leq\; \mathsf{d}(y,x)+\mathsf{d}(x,z) \qquad \text{(“going via } x \text{ is not a shortcut from } y \text{ to } z\text{”)}.$$

Subtracting $\mathsf{d}(y,x)$ from both sides yields $\mathsf{d}(y,z)-\mathsf{d}(y,x)\leq\mathsf{d}(x,z)$. Since $\mathsf{d}(y,x)=\mathsf{d}(x,y)$ by (M-s), this establishes (V.11) — the second of the two inequalities we set out to prove.

The proof is thus complete. □

Then we are ready to introduce some notions that are meaningful in the general setup of metric spaces.



(a) Open ball $\mathcal{B}_r(z)$ (b) Closed ball $\overline{\mathcal{B}}_r(z)$

FIGURE V.2. A schematic illustration of open and closed balls of radius $r > 0$ centered at $z \in X$.

Balls

For $z \in X$ and $r > 0$, the set

$$\mathcal{B}_r(z) = \left\{ x \in X \;\middle|\; \mathsf{d}(z, x) < r \right\} \tag{V.12}$$

is called the **open ball**[4] of radius r centered at z. Note that while in Euclidean spaces these are balls in an ordinary sense, in general such "balls" can be quite different. Note that by (M-0), the center always belongs to the ball, $z \in \mathcal{B}_r(z)$ for any $r > 0$.

Similarly, the **closed ball** of radius $r \geq 0$ centered at $z \in X$ is

$$\overline{\mathcal{B}}_r(z) = \left\{ x \in X \;\middle|\; \mathsf{d}(z, x) \leq r \right\}; \tag{V.13}$$

the only difference to (V.12) being that also points at exactly distance r from z are included. In particular the open ball is a subset in the closed ball of the same center and radius, $\mathcal{B}_r(z) \subset \overline{\mathcal{B}}_r(z)$. By (M-0) we have $z \in \overline{\mathcal{B}}_r(z)$ for any $r \geq 0$.

Furthermore the set

$$\left\{ x \in X \;\middle|\; \mathsf{d}(z, x) = r \right\} = \overline{\mathcal{B}}_r(z) \setminus \mathcal{B}_r(z) \tag{V.14}$$

of points exactly at distance $r > 0$ from z could be called a **sphere** (compare with Example V.10), but this notion is less important for the theory than that of balls.

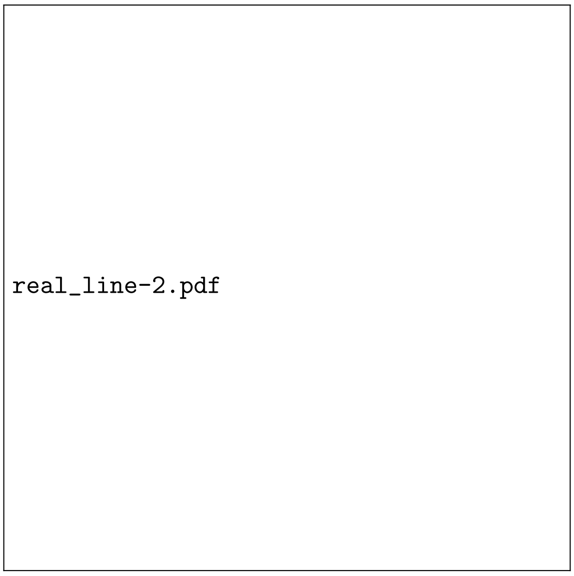
Example V.13 (Balls on the real line are intervals)**.**

Consider the real line $\mathbb{R}$ with its standard metric $\mathsf{d}_{\mathbb{R}}$ as in (V.2). Let $z \in \mathbb{R}$ and $r > 0$.

The open ball of radius r centered at z now consists of all those real numbers x whose distance from z is strictly less than r — it is an open interval

$$\mathcal{B}_r(z) = \left\{ x \in \mathbb{R} \;\middle|\; \underbrace{\mathsf{d}_{\mathbb{R}}(z, x) < r}_{|x-z|<r} \right\} = (z - r, z + r) \subset \mathbb{R}.$$

[4]A note to the Finnish speakers: the word *ball* translates in this context as "kuula". By contrast, "pallo" is the translation of the word *sphere*. This terminology is chosen because the former commonly refers to a solid object (filled throughout) as in (V.12) and (V.13), whereas the latter can refer to something that is hollow (or filled with air) as in (V.14).



The closed ball of radius r centered at z consists of all those real numbers x whose distance from z is less than or equal to r; it is a closed interval

$$\overline{\mathcal{B}}_r(z) \;=\; \Big\{ x \in \mathbb{R} \;\Big|\; \underbrace{\mathsf{d}_{\mathbb{R}}(z,x) \leq r}_{|x-z| \leq r} \Big\} \;=\; [z-r, z+r] \;\subset\; \mathbb{R}.$$

Example V.14 (Balls in a function space with the uniform norm)**.**
Consider the space $\mathcal{C}([a,b])$ of continuous real-valued functions on a closed interval $[a,b]$, equipped with the metric induced by the supremum norm as in Example V.6.

Let $f \in \mathcal{C}([a,b])$ and $r > 0$. In view of the expression (V.6) for the metric on $\mathcal{C}([a,b])$ and the definition of balls (V.12) in metric spaces, the open ball $\mathcal{B}_r(f)$ of radius r around the function $f\colon [a,b] \to \mathbb{R}$ can be visualized as consisting of all those continuous functions $g\colon [a,b] \to \mathbb{R}$ whose graphs stay in a "tube" of radius r from the graph of f.

Example V.15 (Balls with respect to the 0/1-metric)**.**
Let X be a set and $\mathsf{d}_{0/1}$ the 0/1-metric (V.7) on X. Let $z \in X$.

Since all distinct points are at distance 1 from each other, there are no other points besides the center z itself in any open ball of radius $r \leq 1$, i.e., the open ball is a singleton $\mathcal{B}_r(z) = \{z\}$ whenever $0 < r \leq 1$. On the other hand, all of the points in the space are within any radius $r > 1$ from z, so the open ball is the whole space $\mathcal{B}_r(z) = X$ whenever $r > 1$.

For closed balls one gets $\overline{\mathcal{B}}_r(z) = \{z\}$ when $0 \leq r < 1$, and $\overline{\mathcal{B}}_r(z) = X$ when $r \geq 1$.

A simple observation which is used frequently is that balls of smaller radius are contained in balls of larger radius with the same center. More precisely, for any $z \in X$ and for $r_1 \leq r_2$, directly from (V.12) and (V.13) we see that $\mathcal{B}_{r_1}(z) \subset \mathcal{B}_{r_2}(z)$ and $\overline{\mathcal{B}}_{r_1}(z) \subset \overline{\mathcal{B}}_{r_2}(z)$.

Another simple but useful observation is a description of balls in a subspace in terms of balls in the full space.

Exercise (✓) V.4 (Balls in a subspace are intersections of balls with the subspace)**.**
Let (X, d) be a metric space, and let $X' \subset X$ be a subset, which we equip with the inherited metric d' given by (V.8). Let $z' \in X' \subset X$ and let $r > 0$. Denote the balls of radius r centered at z' in the subspace and in the full space, respectively, by

$$\mathcal{B}'_r(z') = \big\{ x' \in X' \;\big|\; \mathsf{d}'(z', x') < r \big\} \qquad \text{and} \qquad \mathcal{B}_r(z') = \big\{ x \in X \;\big|\; \mathsf{d}(z', x) < r \big\}.$$

Show that

$$\mathcal{B}'_r(z') \;=\; \mathcal{B}_r(z') \cap X'. \tag{V.15}$$

Diameter and distance

For $A \subset X$ a subset, we define the **diameter** as

$$\operatorname{diam}(A) \;=\; \sup\Big\{ \mathsf{d}(a_1, a_2) \;\Big|\; a_1, a_2 \in A \Big\}. \tag{V.16}$$

The idea is that we are trying to choose two points $a_1, a_2 \in A$ from the set with maximal possible distance between them, but the maximum in general does not need to exist, so we take the supremum instead.

We say that a subset $A \subset X$ in a metric space (X, d) is **bounded** if there exists some $u \in \mathbb{R}$ such that $\mathsf{d}(a_1, a_2) \leq u$ for all $a_1, a_2 \in A$.[5] For a non-empty bounded set $A \subset X$, the supremum defining the diameter (V.16) is a nonnegative real number. If a set $A \subset X$ is not bounded, then we use the convention $\operatorname{diam}(A) = +\infty$.

Example V.16 (Balls are bounded)**.**
Let $z \in X$ and $r > 0$. What is the diameter of $A = \mathcal{B}_r(z)$?

If $x_1, x_2 \in \mathcal{B}_r(z)$, then by definition $\mathsf{d}(z, x_1) < r$ and $\mathsf{d}(z, x_2) < r$. By triangle inequality (M-Δ) we get

$$\mathsf{d}(x_1, x_2) \;\leq\; \mathsf{d}(x_1, z) + \mathsf{d}(z, x_2) < r + r = 2r.$$

In view of the definition (V.16) this shows that

$$\operatorname{diam}\big(\mathcal{B}_r(z)\big) \leq 2r.$$

In particular this shows that any open ball $\mathcal{B}_r(z)$ is bounded.

One should not, however, expect the diameter of a ball of radius r to be exactly $2r$ in general. For example in a space with 0/1 metric, if $0 < r \leq 1$, we have $\mathcal{B}_r(z) = \{z\}$ (see Example V.15) and correspondingly $\operatorname{diam}\big(\mathcal{B}_r(z)\big) = \operatorname{diam}(\{z\}) = 0$.

Lemma V.17 (Sets contained in balls are bounded)**.**
If $A \subset \mathcal{B}_r(z)$ for some $z \in X$ and $r > 0$, then the set A is bounded.

Proof. If $A \subset A' \subset X$, then from definition (V.16) it is clear that $\operatorname{diam}(A) \leq \operatorname{diam}(A')$. In particular assuming $A \subset \mathcal{B}_r(z)$, we get $\operatorname{diam}(A) \leq \operatorname{diam}(\mathcal{B}_r(z)) \leq 2r$ by the example above. This shows that $\operatorname{diam}(A) < \infty$ and establishes boundedness of A. □

Also the following converse implication holds, and it is frequently useful.

Exercise V.5 (Bounded sets are contained in some balls)**.**
Suppose that $z \in X$. Show that if $A \subset X$ is bounded, then there exists some $r > 0$ (sufficiently large) such that $A \subset \mathcal{B}_r(z)$.
Hint: *The case $A = \emptyset$ is trivial (why?). If $A \neq \emptyset$, pick some $a \in A$, and choose (for example) $r = \operatorname{diam}(A) + \mathsf{d}(z, a) + 1$. Use the triangle inequality to conclude.*

---

[5] The astute reader will notice that for subsets of $\mathbb{R}$ we have now given two different definitions of boundedness of subsets: Definition II.27 based on the order relation $\leq$ on $\mathbb{R}$, and the definition here based on viewing $\mathbb{R}$ as a metric space. In more careful terminology, the two notions might be called *order boundedness* and *metric boundedness*. We are lucky, however, because the two conditions are in fact logically equivalent for subsets of $\mathbb{R}$ (the astute reader can also quickly prove this). So our abuse of terminology does not ultimately risk any confusion.

The **distance** between non-empty subsets $A, B \subset X$ is defined as

$$\mathrm{dist}(A,B) \;=\; \inf \Big\{ \mathsf{d}(a,b) \;\Big|\; a \in A,\; b \in B \Big\} . \tag{V.17}$$

The idea is that we are trying to choose points $a \in A$ and $b \in B$ from the two sets with minimal possible distance among them, but the minimum in general does not need to exist, so we take the infimum instead.

Obviously for any $a, b \in X$, the distance between the corresponding singletons is just $\mathrm{dist}(\{a\}, \{b\}) = \mathsf{d}(a,b)$, so in this way the distance is a more general notion than the metric itself.[6]

We always have $\mathrm{dist}(A,B) \geq 0$ for non-empty subsets $A, B \subset X$, since $\mathsf{d}(a,b) \geq 0$ for all $a \in A$, $b \in B$. If $A \cap B \neq \emptyset$, then the existence of a point $x \in A \cap B$ and the property $\mathsf{d}(x,x) = 0$ implies that $\mathrm{dist}(A,B) = 0$. But it is possible to have vanishing distance even between two disjoint sets!

Example V.18 (Disjoint sets at zero distance)**.**
Consider the real line $\mathbb{R}$ with its usual metric $\mathsf{d}_{\mathbb{R}}$ (Example V.2). Let $A = (-\infty, 0] \subset \mathbb{R}$ and $B = (0,1]$. The sets A and B are disjoint, $A \cap B = \emptyset$. We nevertheless have $\mathrm{dist}(A,B) = 0$ (think about the precise justification!).

The general notion of course captures also familiar problems such as the following.

Exercise V.6 (Distance from a circle to a line)**.**
Consider the Euclidean plane $\mathbb{R}^2$ (Example V.3). Let

$$\begin{aligned} C \;&=\; \Big\{ (x,y) \in \mathbb{R}^2 \;\Big|\; x^2 + y^2 = 1 \Big\} \;\subset\; \mathbb{R}^2, \\ L \;&=\; \Big\{ (x,y) \in \mathbb{R}^2 \;\Big|\; y = 2 - x \Big\} \;\subset\; \mathbb{R}^2. \end{aligned}$$

Calculate the distance between the circle C and the line L, and argue precisely that your answer $\mathrm{dist}(C,L) = \sqrt{2} - 1$ indeed is the infimum in the definition (V.17).

Interior, exterior, boundary, and closure

The following notions are not strictly necessary for the main theory developed in the present course, but they are very frequently used in calculus, and various areas of analysis and its applications.

Interior, exterior, and boundary

Given a subset A of a metric space X, an often useful idea is to classify the points x of the space X according to how they lie in relation to the subset A: whether they are well inside, well outside, or neither really. The precise definitions and terminology are as follows.

Definition V.19 (Interior and exterior)**.**
A point $x \in X$ is said to be an **interior point** of $A \subset X$ if for some $r > 0$ we have $\mathcal{B}_r(x) \subset A$. A point x is said to be an **exterior point** of A if for

[6] Despite some parallels of properties and interpretation, the distance is *not* a metric on the set $\mathscr{P}(X)$ of all subsets of X — it is nonnegative and symmetric, but generally fails the triangle inequality and nondegeneracy, as shown by the examples right afterwards.

some $r > 0$ we have $\mathcal{B}_r(x) \subset X \setminus A$. The set of all interior points of A is denoted by A°, and the set of all exterior points by $\mathrm{ext}(A)$.

From the definition it is clear that $A^\circ \subset A$ and $\mathrm{ext}(A) \subset X \setminus A$. In particular the interior and the exterior are mutually disjoint, $A^\circ \cap \mathrm{ext}(A) = \emptyset$.

Definition V.20 (Boundary).
The **boundary** of subset $A \subset X$ is defined as

$$\partial A = X \setminus \big(A^\circ \cup \mathrm{ext}(A)\big)$$

The points $x \in \partial A$ are called **boundary points** of $A \subset X$.

In other words, a point x is a boundary point of A if it is neither an interior point nor an exterior point. The following characterization is therefore obvious (consider the contrapositive).

Lemma V.21 (A direct characterization of boundary points).
A point $x \in X$ is a boundary point of $A \subset X$ if and only if for all $r > 0$ the ball $\mathcal{B}_r(x)$ contains both a point of A and a point of its complement, i.e., $\mathcal{B}_r(x) \cap A \neq \emptyset$ and $\mathcal{B}_r(x) \cap (X \setminus A) \neq \emptyset$.

In summary we have defined the interior, the exterior, and the boundary of A so that they form a partition

$$X = A^\circ \cup \mathrm{ext}(A) \cup \partial A.$$

of the space into three mutually disjoint subsets.

Example V.22 (The interior of a closed interval).
Consider a closed interval $[a, b] \subset \mathbb{R}$ with endpoints $a < b$. Then for any point $x \in (a, b)$ we can take $r = \min\{x - a, b - x\}$ to be the minimum of the distances of x to the two endpoints, and we have $(x - r, x + r) \subset [a, b]$ (it is a recommended exercise to write down the details). Therefore any such x is an interior point, showing that $(a, b) \subset [a, b]^\circ$. On the other hand, it is obvious that points in the complement $x \in \mathbb{R} \setminus [a, b]$ are not interior points (from the general fact that $A^\circ \subset A$). Also the end-points a and b are not interior points: for any $r > 0$ the open ball $(a - r, a + r)$ centered at a contains the point $a - \frac{r}{2} \notin [a, b]$, so $a \notin [a, b]^\circ$, and similarly $b \notin [a, b]^\circ$. Combining these considerations, we have obtained

$$[a, b]^\circ \;=\; (a, b).$$

In words, the interior of a closed interval is the maximal open interval contained in it.

Example V.23 (The boundary of a closed interval).
Consider again a closed interval $[a, b] \subset \mathbb{R}$ with endpoints $a < b$. Then if $x \in \mathbb{R} \setminus [a, b]$ is a point in its complement, there are two possibilities: $x < a$ or $x > b$. In the former case setting $r = a - x$ we have $(x - r, x + r) \subset \mathbb{R} \setminus [a, b]$ and in the latter case setting $r = x - b$ we have $(x - r, x + r) \subset \mathbb{R} \setminus [a, b]$. We conclude that have $\mathbb{R} \setminus [a, b] \subset \mathrm{ext}([a, b])$. But the opposite inclusion holds on general grounds (the general fact that $\mathrm{ext}(A) \subset X \setminus A$), so we in fact have $\mathrm{ext}([a, b]) = \mathbb{R} \setminus [a, b]$.

Knowing the interior $[a, b]^\circ = (a, b)$ and exterior $\mathrm{ext}([a, b]) = \mathbb{R} \setminus [a, b]$, we directly get that the boundary $\partial\, [a, b] = \mathbb{R} \setminus \big([a, b]^\circ \cup \mathrm{ext}([a, b])\big)$ (the set of points that are neither interior nor exterior) is

$$\partial\, [a, b] \;=\; \{a, b\}\,.$$

In other words, the boundary of a closed interval consists of the two endpoints of it.

Exercise V.7 (Interiors and boundaries of other types of intervals).
Let $a < b$. For $A_1 = (a, b) \subset \mathbb{R}$, $A_2 = [a, b) \subset \mathbb{R}$, and $A_3 = (a, b] \subset \mathbb{R}$, show that we still have the same interior and boundary as for the closed interval $[a, b] \subset \mathbb{R}$, i.e., that $A_j^\circ = (a, b)$ and $\partial A_j = \{a, b\}$ for $j \in \{1, 2, 3\}$.

Exercise V.8 (Interiors and boundary of a rectangle).
Let $a_1 < b_1$ and $a_2 < b_2$. Consider the rectangle $R = [a_1, b_1] \times [a_2, b_2]$ in the Euclidean plane $\mathbb{R}^2$. What is the interior R° of the rectangle $R \subset \mathbb{R}^2$? What is its boundary ∂R?

If, in the following exercise, you do detailed justifications starting from the definitions of interior, exterior, and boundary, then the arguments become more tedious than one would hope. In the long run, you will not want to work this way! Instead, with the tools of Lecture VII, you will be able to give arguments that are much shorter and easier to check.

Exercise V.9 (The interior and boundary of a region in the plane).
Consider the subset

$$A = \Big\{ (x, y) \in \mathbb{R}^2 \;\Big|\; x \geq y^2 \Big\}$$

of the Euclidean plane $\mathbb{R}^2$. Draw a picture of the subset $A \subset \mathbb{R}^2$. What is the interior A°? What is the boundary ∂A?

Closure

We define the closure of a subset in terms of the previous notions.

Definition V.24 (Closure).
The **closure** of a subset $A \subset X$ is $\overline{A} = A \cup \partial A$.

Observe that since the interior A°, exterior $\mathrm{ext}(A)$, and boundary ∂A of A form a partition of the space X, an equivalent definition of the closure is

$$\overline{A} = X \setminus \mathrm{ext}(A).$$

V.4. Product spaces

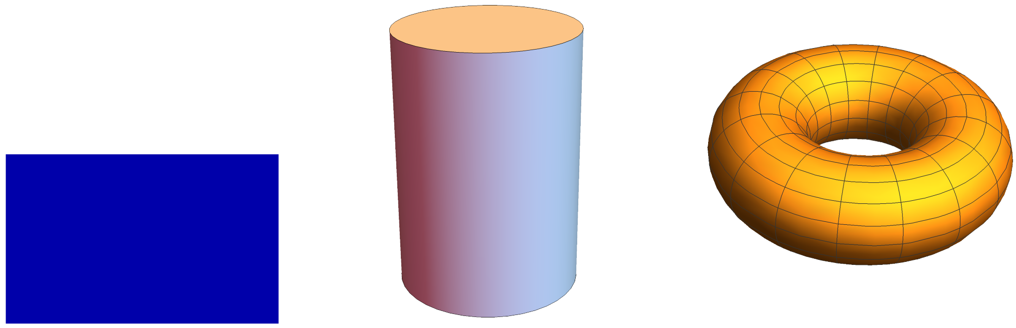
Recall that the Cartesian product of two sets X and Y is the set

$$X \times Y \;=\; \Big\{ (x, y) \;\Big|\; x \in X,\; y \in Y \Big\}$$

of ordered pairs (x, y), whose first component x belongs to the first set X and second component y belongs to the second set Y.

It is very common that we want to form Cartesian products of metric spaces.

Some examples are provided by geometric objects, see Figure V.3. The *plane* $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is the Cartesian product of the real line $\mathbb{R}$ with itself (more generally, the d-dimensional space $\mathbb{R}^d$ is a d-fold Cartesian product of copies of $\mathbb{R}$), a *rectangle* $[a_1, b_1] \times [a_2, b_2]$ is the Cartesian product of two intervals, and a *(solid) cylinder* is



(a) Rectangle (b) Cylinder (c) Torus

FIGURE V.3. Some familiar geometric objects are Cartesian products, for example rectangles $[a_1, b_1] \times [a_2, b_2]$ and (solid) cylinders $\mathbf{D}^2 \times [a, b]$. Also, in very natural coordinates, the torus is $\mathbf{S}^1 \times \mathbf{S}^1$.

the Cartesian product $\mathbf{D}^2 \times [a, b]$ of a disk $\mathbf{D}^2 = \big\{(x, y) \in \mathbb{R}^2 \;\big|\; x^2 + y^2 \leq 1\big\}$ with an interval $[a, b]$. Moreover, the *torus* is naturally viewed as the Cartesian product $\mathbf{S}^1 \times \mathbf{S}^1$ of a circle $\mathbf{S}^1$ with itself.[7]

Product spaces are also important, because for a function of several variables, the domain of definition is naturally a Cartesian product (or a subset of a Cartesian product), in which the sets of allowed values for each variable are the factors of the Cartesian product. Examples we have seen already include an *inner product* $\langle \cdot, \cdot \rangle$, which is a real-valued function of a pair of vectors in an inner product space V, i.e., a function $\langle \cdot, \cdot \rangle \colon \mathsf{V} \times \mathsf{V} \to \mathbb{R}$ defined on the Cartesian product of V with itself, and a *metric* d, which is a function of a pair of points in a metric space X, i.e., a function $\mathsf{d} \colon X \times X \to [0, \infty)$ defined on the Cartesian product of X with itself. In applications to physics, for example, one might consider a field ϕ (e.g., pressure, temperature, magnetic field, $\ldots$) which depends on both spatial and temporal variables so that $\phi(t, x)$ represents the field value at time $t \in [0, \infty)$ and position $x \in D \subset \mathbb{R}^3$ so that ϕ is a function defined on $[0, \infty) \times D$.[8]

We therefore want to be able to equip the Cartesian product $X \times Y$ of two metric spaces (X, d_X) and (Y, d_Y) with enough structure so that for example continuity of functions of several variables makes sense — we would like to equip $X \times Y$ with a metric. It turns out that there are a few different reasonable choices of metrics on the product space $X \times Y$, each choice having its own advantages and each natural in certain situations, but none of the choices generally better than the others. At first this might seem like a problem — if the choice of metric is not canonical, does it mean for example that the continuity of functions of several variables depends on what choice one makes? As we will see in Lecture VII, the answer is that while the choice of *metric* on a product space is not canonical, the choice of *topology* is! All of the different reasonable metrics are topologically equivalent, so for example

[7] As a subset of $\mathbb{R}^3$, the torus is actually not a Cartesian product — the circle $\mathbf{S}^1 \subset \mathbb{R}^2$ lives in the plane, so the Cartesian product of circles lives in $\mathbb{R}^4$. But it is still very natural to parametrize the torus surface in $\mathbb{R}^3$ by coordinates, which make it correspond to $\mathbf{S}^1 \times \mathbf{S}^1$.

[8] The precise temporal and spatial domains would of course depend on the application. This is just one quite typical example, chosen for concreteness.

To convince yourself about the prominence of such examples in different areas of physics, just look at the wave equation, the heat equation, Maxwell's equations, the Schrödinger equation, $\ldots$.

continuity of functions (and other topological properties) does not depend on the choice!

For simplicity, we will only discuss Cartesian products of two metric spaces (X, d_X) and (Y, d_Y). Treating Cartesian products of finitely many metric spaces works similarly, only the necessary notation would become a bit clumsier.[9]

Before delving into the choices of metrics, we introduce projections and component functions; two key notions for concretely working with product spaces.

Projections to the components of the product

On the Cartesian product $X \times Y$ we have two functions, which pick the first and second "coordinates" of a point $(x, y) \in X \times Y$,

$$\begin{aligned} \mathrm{pr}_1 : \quad X \times Y &\to X \\ (x, y) &\mapsto x \end{aligned} \qquad\qquad \begin{aligned} \mathrm{pr}_2 : \quad X \times Y &\to Y \\ (x, y) &\mapsto y. \end{aligned} \tag{V.18}$$

These are called the **projections** to X and Y, respectively.

Component functions

Let X, Y, Z be sets. For a function

$$f \colon Z \to X \times Y$$

with values in the product space $X \times Y$, we can consider the **component functions** (or coordinate functions) obtained by composing the function f with one of the projections (V.18):

$$\begin{aligned} f_1 &= \mathrm{pr}_1 \circ f : \quad Z \to X \\ f_2 &= \mathrm{pr}_2 \circ f : \quad Z \to Y. \end{aligned}$$

The function f can be written in terms of these component functions as

$$f(z) = \big(f_1(z),\, f_2(z)\big) \qquad \text{for } z \in Z. \tag{V.19}$$

In particular, $f \colon Z \to X \times Y$ is uniquely determined by its component functions $f_1 \colon Z \to X$ and $f_2 \colon Z \to Y$ via (V.19).

Also conversely, given any two functions $f_1 \colon Z \to X$ and $f_2 \colon Z \to Y$, one can define a function $f \colon Z \to X \times Y$ by the formula (V.19).

[9] One could also consider Cartesian products of infinitely many metric spaces — not just as an academic curiosity, but also as something that is often relevant (for example in probability theory). The mathematical details become just slightly more complicated for countably infinite Cartesian products, but the conclusion remains similar: there are various natural metrics, each giving rise to the same topology. In uncountably infinite Cartesian products, however, the correct topology does not necessarily come from a metric anymore. Uncountably infinite Cartesian products are one important source of *non-metrizable topologies*, and they illustrate the advantages of considering general topological spaces instead of metric spaces, as in Appendix D. For simplicity, in this course we restrict our attention to just finite Cartesian products.

Metrics on product space

Let (X, d_X) and (Y, d_Y) be two metric spaces. As three natural metrics on the Cartesian product $X \times Y$, we consider

$$\mathsf{d}_1, \mathsf{d}_2, \mathsf{d}_\infty : \ (X \times Y) \times (X \times Y) \ \to \ [0, \infty) \tag{V.20}$$

given by

$$\mathsf{d}_1\big((x,y),\ (x',y')\big) \ = \ \mathsf{d}_X(x,x') + \mathsf{d}_Y(y,y'), \tag{V.21}$$

$$\mathsf{d}_2\big((x,y),\ (x',y')\big) \ = \ \sqrt{\mathsf{d}_X(x,x')^2 + \mathsf{d}_Y(y,y')^2}, \tag{V.22}$$

$$\mathsf{d}_\infty\big((x,y),\ (x',y')\big) \ = \ \max\big\{\mathsf{d}_X(x,x'),\ \mathsf{d}_Y(y,y')\big\}, \tag{V.23}$$

for $(x,y) \in X \times Y$ and $(x',y') \in X \times Y$. Of course we need to prove that each of these three satisfies the properties (M-s), (M-Δ), (M-0), and therefore indeed is a metric on the set $X \times Y$.

Lemma V.25 ($\mathsf{d}_1, \mathsf{d}_2, \mathsf{d}_\infty$ are metrics on the product space)**.**
Each of the formulas (V.21) – (V.23) *defines a metric on* $X \times Y$*.*

Proof. We give the proof for (V.22) and leave the two other cases as exercises.

Clearly (V.22) defines a function $\mathsf{d}_2 \colon (X \times Y) \times (X \times Y) \to [0, +\infty)$ since the sum of squared distances is nonnegative and the square root of a nonnegative number is nonnegative. According to Definition V.1, there are then three properties to verify.

(M-s): Let $(x,y),(x',y') \in X \times Y$. Observe that by the symmetricity of the metrics on X and Y, we have $\mathsf{d}_X(x,x') = \mathsf{d}_X(x',x)$ and $\mathsf{d}_Y(y,y') = \mathsf{d}_Y(y',y)$. Therefore we find

$$\begin{aligned}\mathsf{d}_2\big((x,y),(x',y')\big) &= \sqrt{\mathsf{d}_X(x,x')^2 + \mathsf{d}_Y(y,y')^2} \\ &= \sqrt{\mathsf{d}_X(x',x)^2 + \mathsf{d}_Y(y',y)^2} \ = \ \mathsf{d}_2\big((x',y'),(x,y)\big).\end{aligned}$$

(M-Δ): Here it is practical to start from an inequality of real numbers which may not appear to have anything to do with our metric spaces X and Y or the Cartesian product $X \times Y$. So recall Corollary IV.22 for the inner product space $\mathbb{R}^2$ (the Euclidean plane: case $d = 2$ of Examples IV.17 and IV.12). For any real numbers $a, a', b, b' \in \mathbb{R}$, applying that corollary to the vectors $(a,b) \in \mathbb{R}^2$ and $(a',b') \in \mathbb{R}^2$ yields $\|(a,b) + (a',b')\| \leq \|(a,b)\| + \|(a',b')\|$, i.e.,

$$\sqrt{(a+a')^2 + (b+b')^2} \leq \sqrt{a^2+b^2} + \sqrt{a'^2 + b'^2}.$$

Now let $(x_1,y_1),(x_2,y_2),(x_3,y_3) \in X \times Y$. By the triangle inequalities for metrics on X and Y we get $\mathsf{d}_X(x_1,x_3) \leq \mathsf{d}_X(x_1,x_2) + \mathsf{d}_X(x_2,x_3)$ and $\mathsf{d}_Y(y_1,y_3) \leq \mathsf{d}_Y(y_1,y_2) + \mathsf{d}_Y(y_2,y_3)$. Plugging these estimates in the defining formula of d_2, using the fact that $\sqrt{\cdot}$ is increasing, and finally using the inequality above with $a = \mathsf{d}_X(x_1,x_2)$, $a' = \mathsf{d}_X(x_2,x_3)$, $b = \mathsf{d}_Y(y_1,y_2)$, $b' = \mathsf{d}_Y(y_2,y_3)$, we get

$$\begin{aligned}\mathsf{d}_2\big((x_1,y_1),(x_3,y_3)\big) &= \sqrt{\mathsf{d}_X(x_1,x_3)^2 + \mathsf{d}_Y(y_1,y_3)^2} \\ &\leq \sqrt{\big(\mathsf{d}_X(x_1,x_2) + \mathsf{d}_X(x_2,x_3)\big)^2 + \big(\mathsf{d}_Y(y_1,y_2) + \mathsf{d}_Y(y_2,y_3)\big)^2} \\ &\leq \sqrt{\mathsf{d}_X(x_1,x_2)^2 + \mathsf{d}_Y(y_1,y_2)^2} + \sqrt{\mathsf{d}_X(x_2,x_3)^2 + \mathsf{d}_Y(y_2,y_3)^2} \\ &= \mathsf{d}_2\big((x_1,y_1),(x_2,y_2)\big) + \mathsf{d}_2\big((x_2,y_2),(x_3,y_3)\big).\end{aligned}$$

This proves the triangle inequality for d_2 on $X \times Y$.

(M-0): Let $(x,y),(x',y') \in X \times Y$. We consider the two implications of the "if and only if" separately.

If $(x,y) = (x',y')$, then we have both $x = x'$ and $y = y'$. Then by the separation for metrics on X and Y we get $\mathsf{d}_X(x,x') = \mathsf{d}_X(x,x) = 0$ and $\mathsf{d}_Y(y,y') = \mathsf{d}_Y(y,y) = 0$. We get

$$\mathsf{d}_2\big((x,y),(x',y')\big) \ = \ \sqrt{\mathsf{d}_X(x,x')^2 + \mathsf{d}_Y(y,y')^2} = \ \sqrt{0^2 + 0^2} \ = \ \sqrt{0} \ = \ 0.$$

Conversely, if $\mathsf{d}_2\big((x,y),(x',y')\big) = 0$, then we must have $\mathsf{d}_X(x,x')^2 + \mathsf{d}_Y(y,y')^2 = 0$ (since the square root of a positive number would be positive), and therefore both $\mathsf{d}_X(x,x') = 0$ and $\mathsf{d}_Y(y,y') = 0$ (since if either of these were nonzero, the sum of their squares would be positive). From these, by the separation for metrics on X and Y, we then get $x = x'$ and $y = y'$. These together yield $(x,y) = (x',y')$.

The properties required for d_2 to be a metric on $X \times Y$ have thus been verified. □

Exercise V.10 (d_1 is a metric on $X \times Y$)**.**
Check that (V.21) defines a metric on $X \times Y$.

Exercise V.11 (d_∞ is a metric on $X \times Y$)**.**
Check that (V.23) defines a metric on $X \times Y$.

Lecture VI

Continuity of functions

Throughout this section, let (X, d_X) and (Y, d_Y) be two metric spaces. We will consider functions

$$f\colon X \to Y.$$

Our main task is to extend the notion of continuity, already familiar in simpler setups (for example real valued functions on the real line), to this general setup. The general definition, affectionately known as the "ε-δ condition", intuitively requires that if one does not change the argument x of a continuous function f by too much, then the value $f(x)$ of the function does not change by too much either.

Here it is worth pointing out this "ε-δ condition" that is taken as the general definition of continuity does not immediately reduce to the definition of continuity that was used in the special case of real functions in Lecture III. However, in this and subsequent lectures we will also obtain a number of equivalent characterizations of the general notion of continuity, and one such characterization in Lecture VIII will in particular show that the general definition indeed is logically equivalent, in the special case of real functions, to the definition used before.

Continuous functions are arguably the most important topological notion: they are in a sense exactly those functions which "respect the topology".[1] In later chapters we will study what happens to various topological properties of spaces when continuous functions are applied: we will show, e.g., that the image of a connected space under a continuous function is also connected (Lecture XII), that the image of a compact space under a continuous function is also compact (Lecture XI), etc. But let us not get ahead of ourselves. We must first define the notion of continuity! The definition will be followed by the first simple examples of continuous functions. We will also introduce Lipschitz-continuity — a quantitative, sufficient condition for continuity. With it, we easily get more substantial examples of continuous functions in many of the spaces that have been introduced in earlier chapters.

[1] To provide some context by way of analogies and points of comparison, let us make the following meta-mathematical observation. Whenever some new mathematical structure is defined, it becomes natural to ask what are the mappings (=functions) which either "respect" the structure or "(fully) preserve" the structure. An example from linear algebra would be vector spaces: the functions that respect the vector space structure are linear maps, and functions which preserve the vector space structure are bijective linear maps. An example from algebra would be groups: the functions that respect the group structure are group homomorphisms, and functions which preserve the group structure are group isomorphisms. An example from differential geometry is that of a smooth manifold: the mappings that respect the manifold structure are smooth functions, and mappings which preserve the manifold structure are diffeomorphisms. The reader should become attentive to such examples in whatever topic of mathematics they study. For the present course, the core underlying structure is topology (although to a large extent we focus on the concrete case of metric spaces which have a more specific structure given by the metric). The mappings which respect the topological structure are continuous functions. In the next lecture we will take a look at the topological structure and we will furthermore introduce also mappings which (fully) preserve the topological structure: these are called homeomorphisms.

VI.1. Definition of continuity

The following general definition of continuity makes sense whenever both the domain X and the codomain Y of a function are equipped with metrics, denoted d_X and d_Y, respectively.

Definition VI.1 (Continuity)**.**
Let $f : X \to Y$ be a function, and $x \in X$ a point. The function f is said to be **continuous at** x if for any $\varepsilon > 0$ there exists a $\delta > 0$ such that we have $\mathsf{d}_Y\big(f(x'), f(x)\big) < \varepsilon$ whenever $x' \in X$ and $\mathsf{d}_X(x', x) < \delta$.

We say that $f : X \to Y$ is a **continuous function** if f is continuous at every point $x \in X$.

Here you should really care about very small numbers $\varepsilon > 0$ and $\delta > 0$. Indeed, in more intuitive terms, the definition of continuity at x should be thought of as follows. However small positive amount $\varepsilon > 0$ we are willing to allow the values of our function f to move, movement by less than this much can be guaranteed by requiring that the arguments move by less than some small amount $\delta > 0$. A crucial thing to note is that δ is allowed to depend on ε (as well as on the point x where continuity is considered).[2] The less we are willing to let the values move, the less we can allow the arguments to move — i.e., the smaller the given $\varepsilon > 0$, the smaller we should choose δ.

Remark VI.2 (Rewriting the ε-δ condition in logically equivalent forms)**.**
We may rewrite the above definitions in logical notation as follows:

$$f \text{ is continuous at } x \text{ means}$$
$$\forall\, \varepsilon > 0 : \; \exists\, \delta > 0 : \; \forall\, x' \in X : \qquad \mathsf{d}_X(x', x) < \delta \;\Rightarrow\; \mathsf{d}_Y\big(f(x'), f(x)\big) < \varepsilon, \tag{VI.1}$$

and

$$f \text{ is a continuous function means}$$
$$\forall\, x \in X : \; \forall\, \varepsilon > 0 : \; \exists\, \delta > 0 : \; \forall\, x' \in X : \quad \mathsf{d}_X(x', x) < \delta \;\Rightarrow\; \mathsf{d}_Y\big(f(x'), f(x)\big) < \varepsilon. \tag{VI.2}$$

This may be further rewritten in various useful ways using the notion of balls — the images of balls in X or the preimages of balls in Y. Specifically, we observe[3] that (VI.1) is equivalent to any of the following:

$$\forall\, \varepsilon > 0 : \; \exists\, \delta > 0 : \; \forall\, x' \in X : \qquad x' \in \mathcal{B}_\delta(x) \;\Rightarrow\; f(x') \in \mathcal{B}_\varepsilon(f(x)), \tag{VI.1a}$$
$$\forall\, \varepsilon > 0 : \; \exists\, \delta > 0 : \qquad f\big[\mathcal{B}_\delta(x)\big] \;\subset\; \mathcal{B}_\varepsilon(f(x)), \tag{VI.1b}$$
$$\forall\, \varepsilon > 0 : \; \exists\, \delta > 0 : \qquad \mathcal{B}_\delta(x) \;\subset\; f^{-1}\big[\mathcal{B}_\varepsilon(f(x))\big]. \tag{VI.1c}$$

Similarly, by simply preceding each of these with the quantifier $\forall x \in X$, we obtain equivalent ways to phrase the condition (VI.2).

[2] Just note the order of quantifiers: for all $\varepsilon > 0$ there exists a $\delta > 0$ (which may depend on the earlier chosen ε) such that...

[3] The equivalence of (VI.1) and (VI.1a) is seen by noting that by the definition of open balls, $\mathsf{d}_X(x', x) < \delta$ is equivalent to $x' \in \mathcal{B}_\delta(x)$, while $\mathsf{d}_Y\big(f(x'), f(x)\big) < \varepsilon$ is equivalent to $f(x') \in \mathcal{B}_\varepsilon(f(x))$. The equivalence of (VI.1a) with (VI.1b) and (VI.1c) is then seen by noting that the implication $x' \in \mathcal{B}_\delta(x) \;\Rightarrow\; f(x') \in \mathcal{B}_\varepsilon(f(x))$ is equivalent to $f\big[\mathcal{B}_\delta(x)\big] \;\subset\; \mathcal{B}_\varepsilon(f(x))$ by definition of images, and this in turn is equivalent to $\mathcal{B}_\delta(x) \;\subset\; f^{-1}\big[\mathcal{B}_\varepsilon(f(x))\big]$ by definition of preimages.

At this point, let us give just a few extremely simple examples, without even full details; we will anyway revisit the two first simple examples more carefully with a later characterization of continuity, and we invite the reader to treat the details of the third one as a warm-up exercise.

Example VI.3 (Constant functions are continuous)**.**
Let X, Y be any metric spaces, and $c \in Y$. Consider the constant function $f \colon X \to Y$ given by

$$f(x) = c \qquad \text{for all } x \in X.$$

We claim that f is continuous. Indeed, we leave it for the reader to verify more formally that at any point x, for any given $\varepsilon > 0$, we can in fact use any $\delta > 0$ (for example $\delta = 1$ works) — the values of a constant function in fact never move at all, however much the argument is moved.

Example VI.4 (The identity function is continuous)**.**
Let X be any metric space. Consider the identity function $\mathrm{id} \colon X \to X$ given by

$$\mathrm{id}(x) = x \qquad \text{for } x \in X.$$

We claim that id is continuous. Indeed, we leave it for the reader to verify more formally that at any point x, for any given $\varepsilon > 0$, we can in fact use exactly $\delta = \varepsilon > 0$ — the values of the identity function move by only (and exactly) as much as the argument is moved.[4]

The following is equally easy and also frequently useful. It is a good opportunity to also recall the metric inherited to subspaces.

Exercise VI.1 (Inclusion is continuous)**.**
Let X be a metric space and $X' \subset X$ a subset. Recall that X' becomes a metric space with the metric (V.8) it inherits from X. Consider the inclusion function $\iota \colon X' \to X$ given by

$$\iota(x') = x' \qquad \text{for } x' \in X'.$$

Show that $\iota \colon X' \to X$ is continuous.

VI.2. Lipschitz continuity

The general ε-δ definition of continuity (Definition VI.1) is qualitative. It does not quantitatively say exactly how much of movement of the argument corresponds to how much of movement of the values (it merely asserts that the latter can be made arbitrarily small by choosing the former sufficiently small). There are various quantitative notions of continuity, which are of practical importance especially in analysis, and which also serve as convenient sufficient conditions by which we can often verify continuity more concretely. The simplest and most important of these is Lipschitz-continuity.

Definition VI.5 (Lipschitz property)**.**
Let $M \geq 0$. A function $f \colon X \to Y$ is said to be M-**Lipschitz** if for all $x_1, x_2 \in X$ we have

$$\mathsf{d}_Y\big(f(x_1), f(x_2)\big) \;\leq\; M \, \mathsf{d}_X(x_1, x_2). \tag{VI.3}$$

[4]Of course also any $\delta < \varepsilon$ could also have been chosen here instead.

The parameter M in the above definition is often called the **Lipschitz constant**. Equation (VI.3) has the interpretation that the mapping f cannot stretch distances by more than a factor M.

We say that a function f is **Lipschitz** or **Lipschitz continuous** if it is M-Lipschitz for some $M \geq 0$. The justification of the latter term is in the lemma below, according to which the Lipschitz property is a sufficient condition for continuity.

Theorem VI.6 (Lipschitz functions are continuous)**.**
If $f\colon X \to Y$ is M-Lipschitz, then f is continuous.

Proof. Assume that f is M-Lipschitz, and without loss of generality assume that $M > 0$.[5] [6] To show continuity, we must verify the condition in Definition VI.1. Let therefore $x \in X$, and let $\varepsilon > 0$ be given. We claim that the choice $\delta = \frac{\varepsilon}{M} > 0$ works. Indeed, if $\mathsf{d}_X(x', x) < \delta = \frac{\varepsilon}{M}$, then by the Lipschitz property (VI.3) we get

$$\mathsf{d}_Y\big(f(x'), f(x)\big) \;\leq\; M\, \mathsf{d}_X(x', x) \;<\; M\,\delta = M\,\frac{\varepsilon}{M} = \varepsilon.$$

Since this worked for an arbitrary $\varepsilon > 0$, the continuity of f at x is established. Since the point $x \in X$ was arbitrary, the continuity of the function $f\colon X \to Y$ is established. □

Real functions with bounded derivative

For real functions, the Lipschitz property is closely related to the boundedness of the derivative. Real functions with bounded derivative constitute a common and important class of Lipschitz functions.

Here we frequently refer to the continuity of a real-valued function f of a real variable, and the continuity of its derivative f'. For the purposes of this subsection in particular, let us admit that continuity can be understood in the sense of Definition III.8, which will indeed later be shown to be equivalent to Definition VI.1 (see Corollary VIII.11).

Example VI.7 (Functions with bounded derivative are Lipschitz)**.**
Let $I \subset \mathbb{R}$ be an interval and $f\colon I \to \mathbb{R}$ a continuously differentiable function such that $|f'(x)| \leq M$ for all $x \in I$. We then claim that f is M-Lipschitz,

If $x_1, x_2 \in I$ with $x_1 < x_2$, then the Mean Value Theorem states that for some $x \in (x_1, x_2) \subset I$ we have

$$f'(x) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

We may rearrange this to the form $f(x_2) - f(x_1) = f'(x)\ (x_2 - x_1)$, and then take absolute values and use the bound $|f'(x)| \leq M$ for the derivative to get

$$\big|f(x_2) - f(x_1)\big| \;=\; |f'(x)|\,|x_2 - x_1| \;\leq\; M\,|x_2 - x_1|.$$

The metric on $\mathbb{R}$ is taken to be the usual one, so $\big|f(x_2) - f(x_1)\big| = \mathsf{d}_{\mathbb{R}}\big(f(x_1), f(x_2)\big)$, and the interval $I \subset \mathbb{R}$ inherits a metric from the real axis so that $|x_2 - x_1| = \mathsf{d}_I(x_1, x_2)$. Thus the conclusion is exactly of the form (VI.3)

$$\mathsf{d}_{\mathbb{R}}\big(f(x_1), f(x_2)\big) \;\leq\; M\,\mathsf{d}_I(x_1, x_2).$$

It is easy to see that the above inequality also holds with $x_1 > x_2$ (just rename the variables) and with $x_1 = x_2$ (both sides are zero). Therefore f is M-Lipschitz.

[5] If a function is M-Lipschitz for some M, then it is also M'-Lipschitz for any $M' > M$, so we may assume the Lipschitz constant chosen positive.

[6] Also one may note that the case $M = 0$ is trivial anyway: it requires that the function is constant.

An important particular case of the above is the following.

Example VI.8 (Continuously differentiable functions on closed intervals are Lipschitz).
Let $[a,b] \subset \mathbb{R}$ be a closed interval and $f : [a,b] \to \mathbb{R}$ a continuously differentiable function. Then the continuous function $f' : [a,b] \to \mathbb{R}$ on the closed interval $[a,b]$ is bounded (according to Theorem III.14), so there exists some $M \geq 0$ such that $|f'(x)| \leq M$ for all $x \in [a,b]$, and the reasoning in Example VI.7 applies. Therefore any continuously differentiable function on a closed interval is automatically Lipschitz (with some Lipschitz constant M).

The assumption that the interval is closed was important in the previous example.

Exercise VI.2 (A continuously differentiable function on an interval).
Consider the function $f : (-1,1) \to \mathbb{R}$ defined by

$$f(x) = \frac{1}{1-x^2} \qquad \text{for } x \in (-1,1).$$

(a) Is the function $f : (-1,1) \to \mathbb{R}$ continuous?
(b) Calculate the derivative $f'(x)$. Is $f' : (-1,1) \to \mathbb{R}$ continuous?
(c) Is the function $f : (-1,1) \to \mathbb{R}$ Lipschitz continuous? Does the reasoning in Example VI.8 apply to this case?

Exercise VI.3 (The square root function is not Lipschitz).
Prove that the function

$$q\colon\ [0,1] \to [0,1] \qquad\qquad q(x) = \sqrt{x} \quad \text{for } x \in [0,1]$$

is not M-Lipschitz for any $M \geq 0$ (although it is continuous).

Why is this not a contradiction with what was stated in Example VI.8?

More examples of Lipschitz functions

Let us give some examples of Lipschitz functions, where the space X on which the functions are defined is not just the real axis or an interval.

In the following example, the domain X of our function is a function space. So we are considering functions of functions (albeit still quite simple ones, more complicated examples will follow).

Example VI.9 (The evaluation map on a function space).
Let $[a,b] \subset \mathbb{R}$ be a closed interval and, as in Example IV.15, consider the space $\mathcal{C}([a,b])$ of continuous real-valued functions on this interval equipped with the supremum-norm $\|\cdot\|_\infty$ and the metric induced by this norm.

Fix a point $z \in [a,b]$. Then we may define a real-valued function

$$\mathrm{ev}_z \colon \mathcal{C}([a,b]) \to \mathbb{R}$$

on this function space by the formula

$$\mathrm{ev}_z(f) = f(z) \qquad \text{for } f \in \mathcal{C}([a,b]),$$

i.e., the value of the function ev_z at the point f (note that points of a function space are themselves functions) is the value of f at the chosen point $z \in [a,b]$. This function ev_z is called **evaluation at the point** z.

We claim that $\mathrm{ev}_z \colon \mathcal{C}([a,b]) \to \mathbb{R}$ is continuous, in fact Lipschitz continuous with constant $M = 1$. To see this, let $f_1, f_2 \in \mathcal{C}([a,b])$ be two points in the function space — thus themselves functions $f_1, f_2 \colon [a,b] \to \mathbb{R}$. On the codomain $\mathbb{R}$ of the function ev_z we use the usual metric, so the distance of the values is

$$\mathsf{d}_{\mathbb{R}}\big(\mathrm{ev}_z(f_1), \mathrm{ev}_z(f_2)\big) = \big|\mathrm{ev}_z(f_2) - \mathrm{ev}_z(f_1)\big| = \big|f_2(z) - f_1(z)\big|.$$

By the vector space structure on $\mathcal{C}([a,b])$ we have $f_2(z) - f_1(z) = \big(f_2 - f_1\big)(z)$, the value of the function $f_2 - f_1 \in \mathcal{C}([a,b])$ at z. The distance of the values can then be estimated as

$$\begin{aligned}\mathsf{d}_{\mathbb{R}}\big(\mathrm{ev}_z(f_1), \mathrm{ev}_z(f_2)\big) &= \big|\big(f_2 - f_1\big)(z)\big| \\ &\leq \sup_{x\in[a,b]} \big|\big(f_2 - f_1\big)(x)\big| = \|f_2 - f_1\|_\infty = \mathsf{d}_{\mathcal{C}([a,b])}(f_1, f_2).\end{aligned}$$

This is the Lipschitz property of the function $\mathrm{ev}_z \colon \mathcal{C}([a,b]) \to \mathbb{R}$ with Lipschitz constant $M = 1$. In particular it follows from Theorem VI.6 that evaluation ev_z is a continuous function.

The next exercise features another function on the same function space.

Exercise VI.4 (The maximum map on a function space)**.**
Let $[a,b] \subset \mathbb{R}$ be a closed interval and consider again the space $\mathcal{C}([a,b])$ of continuous real-valued functions on this interval equipped with the supremum-norm $\|\cdot\|_\infty$. For any $f \in \mathcal{C}([a,b])$, the maximum $\max_{x\in[a,b]} f(x)$ exists, and we can thus define a function

$$\mathrm{Max} \colon \mathcal{C}([a,b]) \to \mathbb{R}$$

on this function space by the formula

$$\mathrm{Max}(f) := \max_{x\in[a,b]} f(x) \qquad \text{for } f \in \mathcal{C}([a,b]).$$

Show that $\mathrm{Max} \colon \mathcal{C}([a,b]) \to \mathbb{R}$ is 1-Lipschitz, and conclude that it is in particular continuous.

In the following, the domain X is an inner product space (with its natural metric).

Example VI.10 (Inner product is continuous in either of its arguments)**.**
Let V be an inner product space, with inner product denoted by $\langle\cdot,\cdot\rangle$. Fix a vector $w \in \mathsf{V}$. Define a function

$$f \colon \mathsf{V} \to \mathbb{R}$$

by the formula

$$f(v) = \langle w, v\rangle \qquad \text{for } v \in \mathsf{V}.$$

We claim that this function $f \colon \mathsf{V} \to \mathbb{R}$ is Lipschitz continuous with $M = \|w\|$.

Recall first that the domain V of the above function f is naturally a metric space with the metric $\mathsf{d}_{\mathsf{V}}(u,v) = \|u - v\| = \sqrt{\langle u-v, u-v\rangle}$ (combine Remarks IV.24 and V.5). Also the codomain $\mathbb{R}$ of the function f is naturally a metric space, with the usual metric $\mathsf{d}_{\mathbb{R}}(x,y) = |x - y|$ of the real axis. Now for $v_1, v_2 \in \mathsf{V}$, calculate the quantity in the definition of Lischitz property and estimate it as follows

$$\begin{aligned}\mathsf{d}_{\mathbb{R}}\big(f(v_1), f(v_2)\big) &= \big|f(v_2) - f(v_1)\big| && \text{(definition of } \mathsf{d}_{\mathbb{R}}) \\ &= \big|\langle w, v_2\rangle - \langle w, v_1\rangle\big| && \text{(defining formula for } f) \\ &= \big|\langle w, v_2 - v_1\rangle\big| && \text{(linearity of inner product)} \\ &\leq \|w\|\,\|v_2 - v_1\| && \text{(Cauchy-Schwarz inequality)} \\ &= \|w\|\,\mathsf{d}_{\mathsf{V}}(v_1, v_2). && \text{(definition of } \mathsf{d}_{\mathsf{V}})\end{aligned}$$

This is precisely the Lipschitz property $\mathsf{d}_{\mathbb{R}}\big(f(v_1), f(v_2)\big) \leq M\,\mathsf{d}_{\mathsf{V}}(v_1, v_2)$ with Lipschitz constant $M = \|w\|$.

Since $v \mapsto \langle w, v\rangle$ has the Lipschitz property, it is continuous by Theorem VI.6. This shows that the inner product is a continuous function of its second argument. But by symmetricity of inner product, $\langle w, v\rangle = \langle v, w\rangle$, also $v \mapsto \langle v, w\rangle$ is continuous. Thus the inner product is continuous in either of its arguments.[7]

One special case (albeit one that may seem a bit abstract at first) of Example VI.10 in a function space is the following. Again the domain is a certain function space, so we are considering functions of functions.

Example VI.11 (♡ Fourier coefficients of a function depend continuously on the function)**.**
The Fourier series coefficients of a function $f \in \mathcal{C}([-\pi, \pi])$ are up to a normalization constant the L^2-inner products of the function f with the corresponding trigonometric basis functions of Exercise IV.30.[8] Therefore, applying Example VI.10, it follows that each Fourier coefficient is a continuous function of the function f, when the space of functions $\mathcal{C}([-\pi, \pi])$ is equipped with the metric induced by the norm induced by the L^2-inner product (see Example IV.18).

Since there are (countably) infinitely many basis functions, and corresponding Fourier series coefficients for each basis function, this example featured (countably) infinitely many continuous functions of continuous functions.

Also the next exercise deals with functions of functions, but with a norm that does not come from an inner product.

Exercise VI.5 (The integral map on a function space)**.**
Let $[a, b] \subset \mathbb{R}$ be a closed interval and consider again the space $\mathcal{C}([a, b])$ of continuous real-valued functions on this interval equipped with the supremum-norm $\|\cdot\|_\infty$. For any $f \in \mathcal{C}([a, b])$, the integral $\int_a^b f(x)\,\mathrm{d}x$ is well-defined. We can therefore define a function

$$\mathrm{Int}\colon \mathcal{C}([a, b]) \to \mathbb{R}$$

on this function space by the formula

$$\mathrm{Int}(f) := \int_a^b f(x)\,\mathrm{d}x \qquad \text{for } f \in \mathcal{C}([a, b]).$$

Show that $\mathrm{Int}\colon \mathcal{C}([a, b]) \to \mathbb{R}$ is M-Lipschitz with $M = b - a$, and conclude that it is in particular continuous.
Hint: *In this exercise, you may need a triangle inequality for integrals and a monotonicity of integrals. The triangle inequality for integrals states that*

$$\Big| \int_\alpha^\beta g(t)\,\mathrm{d}t \Big| \;\le\; \int_\alpha^\beta \big|g(t)\big|\,\mathrm{d}t$$

whenever $g\colon [\alpha, \beta] \to \mathbb{R}$ is a Riemann-integrable function (in particular if g is continuous). The monotonicity of integrals states that if $g, h\colon [\alpha, \beta] \to \mathbb{R}$ are Riemann-integrable functions such that $g(t) \le h(t)$ for all $t \in [\alpha, \beta]$, then

$$\int_\alpha^\beta g(t)\,\mathrm{d}t \;\le\; \int_\alpha^\beta h(t)\,\mathrm{d}t.$$

[7] The stronger statement that the inner product is "jointly continuous" in its two arguments is also true, i.e., the function $\mathsf{V} \times \mathsf{V} \to \mathbb{R}$ defined by $(v, w) \mapsto \langle v, w\rangle$ is continuous. Here, the domain $\mathsf{V} \times \mathsf{V}$ of this function of course needs to be equipped with a metric. In Section V.4 we introduced a few reasonable choices metrics for such Cartesian products of metric spaces. The conclusion that $(v, w) \mapsto \langle v, w\rangle$ is continuous is valid for any of these choices (we will address the reasons for this in Section VII.5) — for now we just invite the interested reader to check the continuity directly with any of the choices.

[8] The same exercise also contains the key idea for *why* the Fourier coefficients are defined as such inner products.

The following two exercises are again meant as reminders and concrete illustrations of the fact that continuity and Lipschitz-continuity depend on the metrics chosen for the domain and the codomain.

Exercise ($\sharp$) VI.6 (Revisiting functions on function spaces)**.**
Consider again the space $\mathcal{C}([a,b])$, but now equipped with the norm $\|\cdot\|_1$ and the metric induced by it. The functions $\mathrm{ev}_z, \mathrm{Max}, \mathrm{Int}\colon \mathcal{C}([a,b]) \to \mathbb{R}$ of the Example VI.9, Exercise VI.4, and Exercise VI.5 are still well-defined. One of these three is still Lipschitz continuous with the new choice of metric — which one?
(Extra: Why are the other two not even continuous?)

Exercise ($\sharp$) VI.7 (Re-revisiting functions on function spaces)**.**
Consider again the space $\mathcal{C}([a,b])$, but now equipped with the norm $\|\cdot\|_2$ and the metric induced by it. The functions $\mathrm{ev}_z, \mathrm{Max}, \mathrm{Int}\colon \mathcal{C}([a,b]) \to \mathbb{R}$ of the Example VI.9, Exercise VI.4, and Exercise VI.5 are still well-defined. One of these three is still Lipschitz continuous with the new choice of metric — which one?
(Extra: Why are the other two not even continuous?)
Hint: *Recall Example VI.10.*

Distance to a point

On any metric space (X, d), distances (defined by the metric d) give rise to natural functions. We next address the continuity of these. The key here is the triangle inequality lower bound of Lemma V.12.

Example VI.12 (Distances to a given point are 1-Lipschitz)**.**
Let (X, d) be a metric space and let $a \in X$ be a point in the space. Consider the function

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

defined by the distance to the given point a,

$$f(x) = \mathsf{d}(x, a) \qquad \text{for } x \in X.$$

We claim that this function is 1-Lipschitz, and therefore in particular continuous.

On the codomain $[0,\infty) \subset \mathbb{R}$ of this function we use the metric inherited from the real axis, so for any two points $x_1, x_2 \in X$ of the domain, the distance between the corresponding values is

$$\mathsf{d}_{\mathbb{R}}\big(f(x_1), f(x_2)\big) = \big|f(x_2) - f(x_1)\big| = \big|\mathsf{d}(x_2, a) - \mathsf{d}(x_1, a)\big| \leq \mathsf{d}(x_1, x_2)$$

where the last inequality follows from (V.9) with $x = x_2$, $y = a$, $z = x_1$. This shows that $x \mapsto \mathsf{d}(x, a)$ is 1-Lipschitz $X \to [0, \infty)$.

Exercise ($\checkmark$) VI.8 (The norm is a continuous function)**.**
Let $(\mathsf{V}, \|\cdot\|)$ be a normed space. Consider the norm, seen as a function

$$v \mapsto \|v\|.$$

Why is this function $\mathsf{V} \to [0, \infty)$ continuous?

$\heartsuit$ Other quantitative notions of continuity

Besides Lipschitz-continuity, there are also other frequently used quantitative notions of continuity of functions between metric spaces. In Appendix C.1 we briefly introduce a few that you are likely to encounter in subsequent mathematics studies.

That appendix can also be treated as a source of additional exercises about continuity; most of it should in fact be easier than some of the examples and exercises above.

VI.3. Continuity considerations in product spaces

Let (X, d_X) and (Y, d_Y) be two metric spaces. Recall from Section V.4 that the Cartesian product $X \times Y$ can be equipped with for example one of the metrics $\mathsf{d}_1, \mathsf{d}_2, \mathsf{d}_\infty$ defined in (V.21), (V.22), and (V.23).

Continuity of projections

Recall also from (V.18) that on the product space, we define the two coordinate projections $\mathrm{pr}_1 \colon X \times Y \to X$ and $\mathrm{pr}_2 \colon X \times Y \to Y$ by $(x, y) \mapsto x$ and $(x, y) \mapsto y$, respectively.

Exercise VI.9 (Continuity of the projections)**.**
Show that the projections $\mathrm{pr}_1 \colon X \times Y \to X$ and $\mathrm{pr}_2 \colon X \times Y \to Y$ are continuous...

(a) ...when the product space $X \times Y$ is equipped with the metric d_1 of (V.21);
(b) ...when the product space $X \times Y$ is equipped with the metric d_2 of (V.22);
(c) ...when the product space $X \times Y$ is equipped with the metric d_∞ of (V.23).

Hint: *The simplest approach is to prove that the projections are* 1*-Lipschitz in each case.*

In Section VII.5 we will understand that it is actually enough to check just one of the three cases in the above exercise; the continuity in the other two cases then follows from the topological equivalence of these three metrics.

Continuity by component functions

Let (X, d_X), (Y, d_Y), (Z, d_Z) be three metric spaces. Recall from (V.19) the idea of specifying a function $Z \to X \times Y$ in terms of its two component functions.

Exercise VI.10 (Continuity via component functions)**.**
Let $f_1 \colon Z \to X$ and $f_2 \colon Z \to Y$ be two functions. Consider the function

$$f \colon Z \to X \times Y \qquad \text{given by} \qquad f(z) = \big(f_1(z), f_2(z)\big) \ \text{ for } z \in Z.$$

Prove that when the Cartesian product space $X \times Y$ is equipped with the metric d_1 of (V.21) and when the functions f_1 and f_2 are continuous, then the function f is also continuous.

Do the same when the product space $X \times Y$ is equipped with the metric d_2 of (V.22).

Do the same when the product space $X \times Y$ is equipped with the metric d_∞ of (V.23).

Again, in Section VII.5 we will understand that checking just one of the three cases above is in fact sufficient.

Continuity of the metric

According to Definition V.1, a metric d on X is a function $X \times X \to [0, \infty)$. The domain $X \times X$ of this function is a Cartesian product of two metric spaces (both equal to X, both equipped with the metric d), and this Cartesian product be made into a metric space by one of the metrics (V.20). Of course the codomain $[0, \infty)$ is also naturally a metric space, with the metric it inherits as a subset of the real line. Therefore it makes sense to consider the continuity of the function

$$\mathsf{d} \colon X \times X \to [0, \infty).$$

Exercise VI.11 (Continuity of the metric).
Let (X, d) be a metric space. Consider the metric $\mathsf{d} \colon X \times X \to [0, \infty)$ as a function between metric spaces, where the codomain $[0, \infty) \subset \mathbb{R}$ inherits a metric from $\mathbb{R}$, and the domain $X \times X$ is equipped with the metric d_1 of (V.21). Prove that $X \times X \to [0, \infty)$ is 1-Lipschitz, and conclude in particular that d is continuous.

What changes if we equip the domain $X \times X$ with the metric d_2 of (V.22) or the metric d_∞ of (V.23)?

VI.4. More exercises

Exercise VI.12 (Continuity of the sum of continuous functions).
Let (X, d) be a metric space and $(\mathsf{V}, \|\cdot\|)$ a normed space. Prove that if $f, g \colon X \to \mathsf{V}$ are two continuous functions, then the function $f + g \colon X \to \mathsf{V}$ is also continuous.

Remark: The function $f + g \colon X \to \mathsf{V}$ is the pointwise sum of the vector valued functions f and g, defined by the formula $(f + g)(x) := f(x) + g(x)$, where the right-hand side uses vector addition in vector space V.

Exercise VI.13 (A continuous function defined by an integral).
Define a function $f \colon \mathcal{C}([0, 10]) \to \mathcal{C}([0, 10])$ by setting[9]

$$[f(x)](t) = \int_0^t s\, x(s)\, \mathrm{d}s, \qquad \text{for } x \in \mathcal{C}([0, 10]) \text{ and } t \in [0, 10].$$

Prove that f is M-Lipschitz for a suitable $M \geq 0$, when the space $\mathcal{C}([0, 10])$ (both as the domain and codomain of f) is equipped with the metric induced by the norm $\|\cdot\|_\infty$ — i.e., show that for all $x, y \in \mathcal{C}([0, 10])$ we have

$$\|f(x) - f(y)\|_\infty \leq M \, \|x - y\|_\infty.$$

<u>Hint</u>: *The "triangle inequality for integrals"* $\left| \int_a^b h(s)\, \mathrm{d}s \right| \leq \int_a^b |h(s)|\, \mathrm{d}s$ *can be used here.*

Exercise VI.14 (Taking the derivative is not Lipschitz with respect to the uniform norm).
Let

$$\mathcal{C}^1([-1, 1]) = \big\{ x \colon [-1, 1] \to \mathbb{R} \;\big|\; x' \text{ is continuous} \big\}$$

be the set of all continuously differentiable functions on the interval $[-1, 1]$. Interpret it as a subset $\mathcal{C}^1([-1, 1]) \subset \mathcal{C}([-1, 1])$, where the space $\mathcal{C}([-1, 1])$ of continuous functions is equipped again with the metric induced by the norm $\|\cdot\|_\infty$. Consider the function $g \colon \mathcal{C}^1([-1, 1]) \to \mathcal{C}([-1, 1])$ given by

$$[g(x)](t) = x'(t), \qquad \text{for } x \in \mathcal{C}^1([-1, 1]) \text{ and } t \in [-1, 1].$$

[9] Don't be confused! We define $f(x)$ to be the *function* $[0, 10] \to \mathbb{R}$, whose value at $t \in [0, 10]$ is obtained by the above integral. Make sure you really understand what the domain and codomain of f are.

Show that g is not M-Lipschitz for any $M \geq 0$.
Hint: *In the Lipschitz condition, choose* $y = \vec{0}$ *(zero function). Find a sequence* $(x_n)_{n\in\mathbb{N}}$ *of functions for which* $\|x_n\|_\infty \leq 1$ *for all* $n \in \mathbb{N}$*, but* $\|x_n'\|_\infty \to \infty$ *as* $n \to \infty$*. Think about why this suffices to prove non-Lipschitzness.*

Lecture VII

Metric space topology

In this chapter, we start working with metric spaces with a bit higher level of abstraction. We define in particular the notion of open sets in metric spaces. This turns out to lead to an extremely rich theory — in fact all general topology, ultimately, is about open sets (the reader interested in this is referred to Appendix D to get an idea of the starting point of general topology). This is probably hard to believe at first sight, but in the chapters that follow we hope you get to appreciate that the simple-looking concept of open sets is in fact very profound.

Already in this chapter, we will see one substantial result indicating the usefulness of open sets: we characterize the continuity of functions in terms of preimages of open sets. We will also take a brief look at homeomorphisms (functions that preserve the topology) — with examples that hopefully start to make it clearer what topology is really about. And after considering to what extent the choice of metric is relevant to the notion of open sets, we revisit the issue of choices of metrics in product spaces, and show that the different reasonable metrics we have considered in product spaces can in fact be used interchangeably for many purposes (more precisely, for all topological properties).

VII.1. Open and closed sets

Let (X, d) be a metric space, throughout.

Open sets

Definition VII.1 (Open set)**.**
A subset $U \subset X$ in a metric space X is **open** if for all $x \in U$ there exists an $r_x > 0$ such that $\mathcal{B}_{r_x}(x) \subset U$.

In other words, all points of an open set must have a little bit of room around them within the set itself: a small ball around the point is contained in the set. Given a point $x \in X$, we call any open set $U \subset X$ such that $x \in U$ a **neighborhood** of the point x. Topology can be seen (almost by its definition) as the study of neighborhoods of points.

Remark VII.2 (Being an open set is a relative property!)**.**
Let us already note that a set U being open is *not* an intrinsic property of U, but a property

that $U \subset X$ has in relation to the metric space X where it is considered a subset.[1] Indeed, Definition VII.1 explicitly refers to balls $\mathcal{B}_r(x)$ in X.

Nevertheless, when the metric space X is clear from the context, it is convenient to simply speak of open sets without explicitly mentioning in which ambient space X they are viewed as subsets.

The following two examples are trivial — but important!

Example VII.3 (The empty set is open)**.**
The empty set $\emptyset$ is a subset in any metric space X. It is obvious (by careful thinking[2]) that $\emptyset \subset X$ is open in the sense of Definition VII.1.

Example VII.4 (The whole space is open)**.**
In any metric space X, the whole space X is a subset. It is obvious (by careful thinking[3]) that $X \subset X$ is open in the sense of Definition VII.1.

The following could be considered our first nontrivial example.[4]

Example VII.5 (Open balls are open sets)**.**
Suppose $z \in X$ and $r > 0$. We claim that the open ball $\mathcal{B}_r(z) \subset X$ is an open subset.

We check this directly from Definition VII.1. So let $x \in \mathcal{B}_r(z)$. By definition (V.12) this means that $\mathsf{d}(z,x) < r$. Now let $r_x = r - \mathsf{d}(z,x)$ and note that $r_x > 0$ because $\mathsf{d}(z,x) < r$. Now we only need to check that $\mathcal{B}_{r_x}(x) \subset \mathcal{B}_r(z)$. So let $y \in \mathcal{B}_{r_x}(x)$, which by definition means $\mathsf{d}(x,y) < r_x$. Using triangle inequality (M-Δ), we now get

$$\mathsf{d}(z,y) \;\leq\; \mathsf{d}(z,x) + \mathsf{d}(x,y) \;<\; \mathsf{d}(z,x) + r_x \;=\; \mathsf{d}(z,x) + \big(r - \mathsf{d}(z,x)\big) \;=\; r,$$

which shows that $y \in \mathcal{B}_r(z)$. Since $y \in \mathcal{B}_{r_x}(x)$ was arbitrary, we conclude $\mathcal{B}_{r_x}(x) \subset \mathcal{B}_r(z)$, which is what we needed to check.

Exercise (✓) VII.1 (Punctured space is open)**.**
Let $z \in X$. The complement $X \setminus \{z\}$ of the singleton $\{z\}$ is called a **punctured space**. Show that the punctured space $X \setminus \{z\} \subset X$ is open.

Exercise VII.2 (The complement of a closed ball is open)**.**
Let $z \in X$ and $r \geq 0$. Show that the complement $X \setminus \overline{\mathcal{B}}_r(z)$ of the closed ball $\overline{\mathcal{B}}_r(z)$ is open.

If you want an easy exercise on the familiar real line, then here is one.

Exercise (✓) VII.3 (Open intervals are open sets)**.**
Show that open intervals $I \subset \mathbb{R}$ are open sets in $\mathbb{R}$ (with its standard metric).

---

[1] Later we will have concrete examples $U \subset X' \subset X$, where if we view X' as a metric space with the metric inherited from X, then $U \subset X'$ may be an open subset, even when $U \subset X$ is not an open subset.

[2] There are no points in the empty set $\emptyset$, so Definition VII.1 contains no requirements (there is only one requirement for every point x of the set), and in particular there is no required property that would not be satisfied! So $\emptyset$ is open.

[3] Let $x \in X$. Take any $r_x > 0$, for example $r_x = 1$. The open ball $\mathcal{B}_{r_x}(x)$ is *by definition* (V.12) contained in the space X, i.e., $\mathcal{B}_{r_x}(x) \subset X$. In view of Definition VII.1, this shows that X is open.

[4] If the phrase "open balls are open sets" sounds tautological, remember that we have given a precise meaning to both the term *open ball* (V.12) and the term *open set* (Definition VII.1), and it is the precise definitions that carry logical content, not the words we have chosen to use. . . The terminology has actually been chosen *because of* the properties that we are just about to establish!

The following result shows how we can construct new open sets out of given ones. Again despite its simplicity, it captures something very consequential.

Theorem VII.6 (Unions and intersections of open sets)**.**

(a) *The union of any collection of open sets in X is also open.*
(b) *The intersection of any finite collection of open sets in X is also open.*

Note the difference: in (a) we allow arbitrary collections of open sets, whereas in (b) we only allow finite collections of open sets! The following example shows why.

Example VII.7 (Infinite intersections of open sets are not always open)**.**
Consider the metric space $(\mathbb{R}, \mathsf{d}_{\mathbb{R}})$ — the real line with its standard metric.

For $n \in \mathbb{N}$, the open interval $U_n = \big(-\frac{1}{n}, \frac{1}{n}\big)$ is just the open ball of radius $\frac{1}{n}$ centered at $0 \in \mathbb{R}$ (recall Example V.13), and therefore $U_n \subset \mathbb{R}$ is an open set (by Example VII.5). The intersection of the countably infinite collection $(U_n)_{n\in\mathbb{N}}$ of open sets is $\bigcap_{n\in\mathbb{N}} U_n = \{0\}$, a singleton (the reasoning is similar to Example I.7). But the singleton $\{0\} \subset \mathbb{R}$ is not an open set![5]

This shows that the requirement of finiteness of the collection of open sets is needed in Theorem VII.6(b) — the conclusion would not be generally true for infinite collections.

Proof of Theorem VII.6. We prove statements (a) and (b) separately.

proof of (a): Let $(U_j)_{j\in J}$ be a collection of open sets $U_j \subset X$. Denote the union of this collection by

$$U = \bigcup_{j\in J} U_j.$$

We prove that $U \subset X$ is open, directly from Definition VII.1. So let $x \in U = \bigcup_{j\in J} U_j$. By definition of the union, there exists some $j_0 \in J$ such that $x \in U_{j_0}$. Then since $U_{j_0} \subset X$ was assumed open, there exists some $r > 0$ such that $\mathcal{B}_r(x) \subset U_{j_0}$. But now

$$\mathcal{B}_r(x) \subset U_{j_0} \subset \bigcup_{j\in J} U_j = U.$$

The existence of such an $r > 0$ for any $x \in U$ by definition shows that U is open.

proof of (b): Let $U_1, U_2, \ldots, U_m \subset X$ be open sets (finitely many!). Consider their intersection

$$V = U_1 \cap U_2 \cap \cdots \cap U_m.$$

We prove that $V \subset X$ is open, directly from Definition VII.1. So let $x \in V = \bigcap_{j=1}^m U_j$. By definition of the intersection, for all $j \in \{1, \ldots, m\}$ we have $x \in U_j$. Then since each $U_j \subset X$ was assumed open, there exists $r_j > 0$ such that $\mathcal{B}_{r_j}(x) \subset U_j$. Now let $r = \min\{r_1, \ldots, r_m\}$. We have $r > 0$, as the minimum of finitely many positive numbers. For each $j \in \{1, \ldots, m\}$, since $r \leq r_j$, we have $\mathcal{B}_r(x) \subset \mathcal{B}_{r_j}(x) \subset U_j$. By definition of intersection, this implies $\mathcal{B}_r(x) \subset \bigcap_{j=1}^m U_j = V$. The existence of such an $r > 0$ for any $x \in V$ by definition shows that V is open. □

At this moment, the reader may want to peek at the definition of general topological spaces in Appendix D. Namely, Theorem VII.6 and Examples VII.3 and VII.4 show that the collection $\mathscr{T}(X) = \{U \subset X \text{ open}\}$ of all open subsets of a metric space X forms a topology on X. The open subsets in a metric space are determined by the metric, so this topology is called the **topology induced by the metric**.

[5] Indeed, there is no $r > 0$ such that the open ball $\mathcal{B}_r(0)$ of radius r centered at $0 \in \{0\}$ would not contain also other real numbers besides 0: we always have $\mathcal{B}_r(0) \not\subset \{0\}$.

Let us continue with one more example on open sets.

Example VII.8 (Open sets in the 0/1-metric)**.**
Let X be a set. Consider the metric space $(X, \mathsf{d}_{0/1})$ obtained by equipping X with the 0/1-metric of Example V.7.

For any $z \in X$, we observe based on Example V.15 that the singleton $\{z\}$ is in fact a certain open ball: for $r = 1$ (or in fact any $0 < r \leq 1$) we have $\{z\} = \mathcal{B}_r(z)$. In particular it then follows from Example VII.5 that the singleton $\{z\} \subset X$ is an open set.

Now consider any subset $A \subset X$. Of course the subset is the union of the singletons of its points, $A = \bigcup_{a \in A} \{a\}$, and since each singleton $\{a\} \subset X$ is open, this union is open by Theorem VII.6(i). We conclude that *every* subset $A \subset X$ is open in a space equipped with the 0/1-metric!

Isolated points and discrete topology

Example VII.8 shows some peculiar properties of the 0/1-metric.

In general, in a metric space (X, d), if $z \in X$ is a point such that the singleton $\{z\}$ is open, then we say that z is an **isolated point**. The first observation in Example VII.8 therefore says that in a space with the 0/1-metric, every point is an isolated point.

If in a metric space (X, d) every subset $A \subset X$ is an open set, then the space is said to have (the) **discrete topology**. The second observation in Example VII.8 therefore says that a space equipped with the 0/1-metric has discrete topology. Just like in the example, it is easy to see that a space has discrete topology if and only if its every point is an isolated point.

Despite the fact that we thus first encountered discrete topology in a somewhat exotic space, there are actually many quite familiar examples of it.

Exercise VII.4 (Finite metric spaces have discrete topology)**.**
Suppose that X is a finite set, and d is a metric on X. Show that every point $z \in X$ is an isolated point, and that (X, d) has discrete topology.

Exercise VII.5 (The set of integers has discrete topology)**.**
Consider the real line $\mathbb{R}$ with its standard metric $\mathsf{d}_{\mathbb{R}}$. Since the set $\mathbb{Z} \subset \mathbb{R}$ of integers is a subset in the metric space $(\mathbb{R}, \mathsf{d}_{\mathbb{R}})$, it becomes a metric space with the metric inherited from $\mathbb{R}$ by (V.8). Show that every point $n \in \mathbb{Z}$ is an isolated point in $\mathbb{Z}$, and that $\mathbb{Z}$ has discrete topology.

Do we have similar conclusions for $\mathbb{N} \subset \mathbb{R}$?

Open sets in subspaces of metric spaces

The following observation is the foundation of subspace topology.

Theorem VII.9 (Open sets in a subspace of a metric space)**.**
Let (X, d) be a metric space, and let $X' \subset X$ be a subset, which we equip with the inherited metric d' given by (V.8)*.*

Then a subset $U' \subset X'$ is open if and only if there exists an open set $U \subset X$ such that $U' = U \cap X'$.

Proof. We prove the "if" and "only if" implications separately.

Throughout the proof, we use the notation as in Exercise V.4 to make a distinction between balls in the subspace X' and balls in the full space X: for any $z' \in X'$ and $r > 0$ we denote the open balls of radius r centered at z' in the subspace and the full space, respectively, by

$$\mathcal{B}'_r(z') = \big\{x' \in X' \;\big|\; \mathsf{d}'(z', x') < r\big\} \qquad \text{and} \qquad \mathcal{B}_r(z') = \big\{x \in X \;\big|\; \mathsf{d}(z', x) < r\big\}.$$

proof of "if": Suppose that $U \subset X$ is open. Let $U' = U \cap X'$. We will prove that $U' \subset X'$ is an open set in the subpace directly from the definition of open sets. So let $x' \in U'$. Since $U' \subset U$, we also have $x' \in U$. Then since U is an open set in X, there exists some $r > 0$ such that $\mathcal{B}_r(x') \subset U$. By Exercise V.4, the ball in the subspace X' with the same radius r and center x' can be expressed as an intersection with the subspace. Therefore we get

$$\mathcal{B}'_r(x') \;=\; \mathcal{B}_r(x') \cap X' \;\subset\; U \cap X' \;=\; U'.$$

This proves that $U' \subset X'$ is open.

proof of "only if": Suppose that $U' \subset X'$ is open. By definition, this means that for every $x' \in U'$ there exists an $r_{x'} > 0$ such that $\mathcal{B}'_{r_{x'}}(x') \subset U'$. Now define U as the following union of balls in the full space X, with centers x' in the set U', and radii $r_{x'}$ as above:

$$U \;=\; \bigcup_{x' \in U'} \mathcal{B}_{r_{x'}}(x').$$

Each of the open balls $\mathcal{B}_{r_{x'}}(x') \subset X$ is open by Example VII.5, so as their union, $U \subset X$ is open by Theorem VII.6. It only remains to check that $U \cap X' = U'$. By distributivity of unions over intersections and by the observation of Exercise V.4 again, we immediately get

$$U \cap X' \;=\; \Big(\bigcup_{x' \in U'} \mathcal{B}_{r_{x'}}(x')\Big) \cap X' \;=\; \bigcup_{x' \in U'} \big(\mathcal{B}_{r_{x'}}(x') \cap X'\big) \;=\; \bigcup_{x' \in U'} \mathcal{B}'_{r_{x'}}(x').$$

The proof will therefore be complete once we verify the equality $\bigcup_{x' \in U'} \mathcal{B}'_{r_{x'}}(x') = U'$. To do this almost obvious last step in detail, let us verify inclusions in both directions. The direction $\bigcup_{x' \in U'} \mathcal{B}'_{r_{x'}}(x') \subset U'$ is clear, since each of the balls in the union were chosen so that $\mathcal{B}'_{r_{x'}}(x') \subset U'$. The direction $\bigcup_{x' \in U'} \mathcal{B}'_{r_{x'}}(x') \supset U'$ is clear, since each point $x' \in U'$ belongs at least to $\mathcal{B}'_{r_{x'}}(x')$ and therefore belongs to the union. The proof is complete. $\square$

Closed sets

The definition of closed sets below will directly refer to that of open sets (Definition VII.1). All questions about closed sets therefore in principle translate to questions about open sets. Nevertheless, the notion of a closed set turns out to be convenient and useful on its own right as well.

Definition VII.10 (Closed set)**.**
A subset $A \subset X$ in a metric space X is **closed** if its complement $X \setminus A \subset X$ is open.

We issue the same warning as for open sets.

Remark VII.11 (Being a closed set is a relative property!)**.**
Note that a set A being open is *not* an intrinsic property of A, but a property that $A \subset X$ has in relation to the metric space X where it is considered a subset.

Let us again start by two trivial but important examples.

Example VII.12 (The empty set is closed).
The empty set $\emptyset$ is a subset in any metric space X. It is closed, since its complement $X \setminus \emptyset = X$ is the whole space, which is open by Example VII.4.

Example VII.13 (The whole space is closed).
In a metric space X, the whole space is a subset $X \subset X$. It is closed, since its complement $X \setminus X = \emptyset$ is the empty set, which is open by Example VII.3.

The following could be considered our first nontrivial example of a closed set.

Example VII.14 (Closed balls are closed sets).
Let $z \in X$ and $r \geq 0$. The closed ball $\overline{\mathcal{B}}_r(z) \subset X$ is closed, since by Exercise VII.2 its complement $X \setminus \overline{\mathcal{B}}_r(z) \subset X$ is open.

Example VII.15 (Singletons are closed sets).
Let $z \in X$. The singleton $\{z\} \subset X$ is closed, since by Exercise VII.1 its complement $X \setminus \{z\} \subset X$ is open.

As an exercise, you can prove the following directly from definition.

Exercise VII.6 (Finite sets are closed).
Let $A = \{z_1, \ldots, z_m\} \subset X$ be a finite subset in a metric space X. Prove that $A \subset X$ is closed.

The following counterpart of Theorem VII.6 shows how we can construct new closed sets out of given ones. Note that here the finiteness requirement is on unions, whereas arbitrary intersections are allowed.

Theorem VII.16 (Unions and intersections of closed sets).

(a) *The intersection of any collection of closed sets in X is also closed.*
(b) *The union of any finite collection of closed sets in X is also closed.*

Exercise VII.7 (Proof of Theorem VII.16).
Prove Theorem VII.16.
Hint: *Use Exercise I.7 and the corresponding result for open sets.*

You can now consider Exercise VII.6 again, in view of Theorem VII.16 and Example VII.15. The right tools make things easier.

Warning: clopen sets!

You should *not* think of open and closed as opposites![6] They are mathematical terms whose meaning was given in Definitions VII.1 and VII.10.

There are, in particular, sets which are both open and closed.

[6] Sometimes this warning is summarized as "sets are not doors". A door, arguably, is either open or closed, and never both at the same time. Remark VII.17 and Example VII.18 show that sets in metric spaces don't obey such a dichotomy.

Remark VII.17 (Clopen sets).
According to Examples VII.3 and VII.12, the empty set $\emptyset \subset X$ is both open and closed.

According to Examples VII.4 and VII.13, the whole space $X \subset X$ is both open and closed.

Depending on the metric space, there can exist also other subsets which are both open and closed (see disconnected spaces in Lecture XII). Such sets are sometimes called **clopen** (although this notion is not needed often).

In any metric space (X, d) with discrete topology[7], any subset $A \subset X$ is open (this is the definition of discrete topology!), but also its complement $X \setminus A \subset X$ is open (it is also a subset!), so A is by definition also closed.

There are also sets which are neither open nor closed.

Example VII.18 (The set of rational numbers is neither open nor closed on the real line).
Consider the real line $\mathbb{R}$ with its usual metric, and consider the subset $\mathbb{Q} \subset \mathbb{R}$ of rational numbers in it. We claim that $\mathbb{Q}$ is neither open nor closed.

To see that $\mathbb{Q}$ is not open, take a rational number $q \in \mathbb{Q}$, for example $q = 0$. Now for an arbitrary $r > 0$, the open ball $\mathcal{B}_r(q) = (q - r, q + r)$ is a non-empty open interval, and thus by Theorem II.23 contains some irrational numbers and therefore we have $\mathcal{B}_r(q) \not\subset \mathbb{Q}$. This shows that $\mathbb{Q}$ is not open.

To see that $\mathbb{Q}$ is not closed, we must by definition show that its complement $\mathbb{R} \setminus \mathbb{Q}$ is not open. So take an irrational number $z \in \mathbb{R} \setminus \mathbb{Q}$, for example $z = \sqrt{2}$. Now for an arbitrary $r > 0$, the open ball $\mathcal{B}_r(z) = (z - r, z + r)$ is a non-empty open interval, and thus by Theorem II.21 contains some rational numbers and therefore we have $\mathcal{B}_r(z) \not\subset \mathbb{R} \setminus \mathbb{Q}$. This shows that $\mathbb{R} \setminus \mathbb{Q}$ is not open, i.e., that $\mathbb{Q}$ is not closed.

Exercise VII.8 (Half-open intervals are neither open nor closed on the real line).
Let $a < b$. Show that $(a, b] \subset \mathbb{R}$ is neither open nor closed.

Check that the same holds also for $[a, b) \subset \mathbb{R}$.

To reiterate, *open* and *closed* mean exactly what they were defined to mean. They are *not* the opposites of each other!

Interior, exterior, boundary, and closure again

Recall from Section V.3 that given a subset $A \subset X$ in a metric space (X, d), the space X partitions into a disjoint union of the interior A°, the exterior $\mathrm{ext}(A)$, and the boundary ∂A of A; see Definitions V.19 and V.20. The first two of these are themselves open sets in X.

Theorem VII.19 (The interior and exterior are open).
Let $A \subset X$. Then both A° and $\mathrm{ext}(A)$ are open sets in X.

Proof. Let us prove that the interior $A^\circ \subset X$ is open. Let $x \in A^\circ$ be an interior point. By definition, there exists some $r > 0$ such that $\mathcal{B}_r(x) \subset A$. We will show that in fact $\mathcal{B}_r(x) \subset A^\circ$, which by definition will give the openness of the interior $A^\circ \subset X$. So let $y \in \mathcal{B}_r(x)$. But we have seen that the open ball $\mathcal{B}_r(x)$ is open (Example VII.5), so there exists some $r' > 0$ such that

[7] Recall that examples of spaces with discrete topology include the set $\mathbb{Z}$ of integers, the set $\mathbb{N}$ of natural numbers, any finite metric space X, and any set X equipped with the $0/1$ metric.

$\mathcal{B}_{r'}(y) \subset \mathcal{B}_r(x)$. Therefore we have

$$\mathcal{B}_{r'}(y) \;\subset\; \mathcal{B}_r(x) \;\subset\; A,$$

showing that $y \in A^\circ$. We conclude that $\mathcal{B}_r(x) \subset A^\circ$. This proves that $A^\circ \subset X$ is open.

The openness of the exterior $\mathrm{ext}(A) \subset X$ can be proven similarly, but it also follows simply by observing that the exterior of A is the interior of its complement, $\mathrm{ext}(A) = (X \setminus A)^\circ$. □

Corollary VII.20 (The boundary is closed)**.**
Let $A \subset X$. Then the boundary ∂A is a closed set in X.

Proof. By Theorem VII.19, both A° and $\mathrm{ext}(A)$ are open, so their union $A^\circ \cup \mathrm{ext}(A)$ is also open by Theorem VII.6(a). The boundary $\partial A = X \setminus \big(A^\circ \cup \mathrm{ext}(A)\big)$ is by definition the complement of this open set, so it is closed (by definition). □

Recall also Definition V.24: the closure of $A \subset X$ is $\overline{A} = A^\circ \cup \partial A$, and it can be equivalently expressed as $\overline{A} = X \setminus \mathrm{ext}(A)$. The exterior is open by Theorem VII.19, so the closure has thus been expressed as the complement of an open set. We immediately get the following.

Corollary VII.21 (The closure is closed)**.**
Let $A \subset X$ be a subset of a metric space X. Then its closure $\overline{A} \subset X$ is a closed subset.

In the following exercise you will establish a few basic properties of the closure. You will in particular show that the closure $\overline{A}$ of A is in fact the *smallest* closed set containing A.

Exercise VII.9 (Characterizing properties of the closure)**.**
Let X be a metric space and $A \subset X$ a subset. Prove the following:

(a) We have $A \subset \overline{A}$.
(b) If $F \subset X$ is a closed set such that $A \subset F$, then we have $\overline{A} \subset F$.
(c) The closure $\overline{A}$ is the smallest closed set in X which contains A.

The interior of a subset admits a characterization of a similar kind.

Exercise VII.10 (Characterizing properties of the interior)**.**
Give a precise formulation to the statement: the interior A° is the greatest open set in X which is contained in A. Prove the statement.
Hint: *Compare with Exercise VII.9.*

VII.2. Characterization of continuity

It turns out that the ε-δ definition of continuity (Definition VI.1) is equivalent to a condition formulated purely in terms of open sets. Let us just briefly mention why this is significant, although we will not elaborate further now. It means that the notion of continuity of a function only depends on the metrics used on its domain and codomain to the extent that these metrics affect which sets are open. Properties that can be phrased purely in terms of open sets are topological (by definition). Namely, in general topology the notion of open sets is taken as the axiomatic starting point. What we below see as a characterization of continuity between metric spaces is then taken as the definition of continuity between topological spaces (the ε-δ definition would be meaningless in the absence of metrics).

Besides thus enabling (further) generalization of the notion of continuity, the characterization below has concrete benefits as well. It occasionally serves as an easier way to verify that a given function is continuous (for an example of this, see Theorem VII.29). Even more commonly it is used with functions that are known to be continuous, to give quick arguments that certain sets defined in terms of those functions are open (or closed). We will provide examples of such reasoning soon, and almost every lecture from here on contains further examples.

Note that the following characterizes the continuity of a function, rather than just the continuity of a function at a given point.

Theorem VII.22 (Characterization of continuity with open (and closed) sets)**.** *Let (X, d_X) and (Y, d_Y) be metric spaces and $f: X \to Y$ a function. Then the following are equivalent:*

(i) *the function $f: X \to Y$ is continuous (in the sense of Definition VI.1);*
(ii) *for every open set $V \subset Y$, the preimage $f^{-1}[V] \subset X$ is open;*
(iii) *for every closed set $A \subset Y$, the preimage $f^{-1}[A] \subset X$ is closed.*

Proof. We will show the equivalence of all three conditions by separately proving the implications (i) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (i), and the equivalence (ii) $\Leftrightarrow$ (iii).

proof of (i) $\Rightarrow$ (ii): Suppose that $f: X \to Y$ is continuous. Let $V \subset Y$ be an open set. We must show that the preimage $f^{-1}[V] \subset X$ is open, i.e., we must find neighborhoods for its points contained in the set itself. So let $x \in f^{-1}[V]$ be a point. By definition of preimages, this means $f(x) \in V$. By openness of V there exists some $\varepsilon > 0$ such that $\mathcal{B}_\varepsilon(f(x)) \subset V$. According to the reformulation (VI.1c) of continuity of f at the point x, there then exists a $\delta > 0$ such that $\mathcal{B}_\delta(x) \subset f^{-1}[\mathcal{B}_\varepsilon(f(x))] \subset f^{-1}[V]$. Thus $\mathcal{B}_\delta(x)$ is a neighborhood of x contained in $f^{-1}[V]$. We conclude that $f^{-1}[V] \subset X$ is open.

proof of (ii) $\Rightarrow$ (i): Suppose that the preimage $f^{-1}[V]$ of each open set $V \subset Y$ is open. We must show that f is continuous. So let $x \in X$, and let $\varepsilon > 0$. The open ball $\mathcal{B}_\varepsilon(f(x)) \subset Y$ is an open set, so by assumption its preimage $f^{-1}[\mathcal{B}_\varepsilon(f(x))] \subset X$ is also open. Note that we have $x \in f^{-1}[\mathcal{B}_\varepsilon(f(x))]$, since $f(x) \in \mathcal{B}_\varepsilon(f(x))$. Therefore by openness of $f^{-1}[\mathcal{B}_\varepsilon(f(x))]$, there exists some $\delta > 0$ such that $\mathcal{B}_\delta(x) \subset f^{-1}[\mathcal{B}_\varepsilon(f(x))]$. According to the reformulation (VI.1c), we have thus shown the continuity of f at the point x. The point $x \in X$ was arbitrary, so the continuity of $f: X \to Y$ follows.

proof of (ii) $\Leftrightarrow$ (iii): The equivalence of (ii) and (iii) is easy to see by passing to complements.

For example, assume that (ii) holds, i.e., that the preimage $f^{-1}[V]$ of every open set $V \subset Y$ is open. We must then show that the preimage of any closed set in Y is closed. So let $A \subset Y$ be closed. To show that its preimage $f^{-1}[A] \subset X$ is closed, we must by definition show that its complement $X \setminus f^{-1}[A]$ is open. But by routine set theory, the complement of the preimage is the preimage of the complement, $X \setminus f^{-1}[A] = f^{-1}[Y \setminus A]$. However, the complement $Y \setminus A$ of the closed set A is by definition open. The preimage $f^{-1}[Y \setminus A]$ of this open set is by assumption (ii) open. We have thus shown that the preimage $f^{-1}[A]$ of an arbitrary closed set $A \subset Y$ is closed.

The other direction of this equivalence is entirely similar. $\square$

Observe carefully that this characterization of continuity talks about *preimages* — not *images* — of open and closed sets!

Remark VII.23 (The images of open sets under continuous functions are not always open)**.**
Consider the continuous[8] function $f\colon \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$. Consider the open interval $U = (-7, 7)$. Note that $U \subset \mathbb{R}$ is an open set (recall Exercise VII.3). Its image under f is $f[U] = [0, 49)$. This half-open interval is not an open set in $\mathbb{R}$ (recall Exercise VII.8).

In particular, the images of open sets under continuous functions are not always open!

To familiarize ourselves with this characterization, let us revisit some very simple examples of continuous functions in view of it (this time in fact providing full details).

Example VII.24 (Constant functions are continuous, again)**.**
Let X, Y be metric spaces, and $c \in Y$. Consider the constant function $f\colon X \to Y$ given by

$$f(x) = c \qquad \text{for all } x \in X,$$

as in Example VI.3.

Suppose that $V \subset Y$ is an open subset, and consider the preimage $f^{-1}[V] \subset X$. If $c \in V$, then the preimage is the whole space, $f^{-1}[V] = X$. If $c \notin V$, then the preimage is empty, $f^{-1}[V] = \emptyset$. Both the whole space and the empty set are open subsets of X, so the preimage $f^{-1}[V] \subset X$ is, in either case, open. The continuity of f now follows by the characterization of Theorem VII.22.

Example VII.25 (Identity function is continuous, again)**.**
Let X be a metric space, and consider the identity function $\mathrm{id}\colon X \to X$ given by

$$\mathrm{id}(x) = x \qquad \text{for } x \in X,$$

as in Example VI.4.

Note that the preimage of any subset $V \subset X$ under the identity function id is the subset itself, $\mathrm{id}^{-1}[V] = V$. Therefore in particular for an open subset $V \subset X$, the preimage $\mathrm{id}^{-1}[V] = V$ is also open.[9] The continuity of id now follows by the characterization of Theorem VII.22.

Verifying the openness and closedness of sets

The following examples illustrate a very common and practical way of verifying that sets defined by conditions involving continuous functions are either open or closed.

Let us start with a little bit to think about.

Exercise VII.11 (A closed ball is a closed set, again)**.**
With the tools that we now have (particularly Theorem VII.22 and Example VI.12), how would you prove that the closed ball $\overline{\mathcal{B}}_r(x) \subset X$ in a metric space (X, d) is a closed set? Compare with Example VII.14.

---

[8] Recall that polynomial functions are continuous by Corollary III.12. The only caveat is that we need to wait until Lecture VIII to prove that the notion of continuity used in that corollary agrees with the current general notion of continuity. For the moment, rest assured that it does.

[9] It is worth noting that we are not merely using the same set X as the domain and codomain of the function, but we are moreover equipping it with the same metric, and consequently the same collection of open sets. We later consider the possibility of equipping the domain X and codomain X (same set) with different metrics.

Example VII.26 (An ellipse)**.**
Let $a, b > 0$. Consider the ellipse in the plane

$$E = \left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \right\}.$$

We claim that $E \subset \mathbb{R}^2$ is a closed subset. To see this, note that its definition involves the function

$$f\colon \mathbb{R}^2 \to \mathbb{R} \qquad \text{given by} \qquad f(x,y) = \frac{x^2}{a^2} + \frac{y^2}{b^2}.$$

In fact, by its very definition, the ellipse is the preimage

$$E = \left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; f(x,y) = 1 \right\} = f^{-1}\big[\,\{1\}\,\big]$$

of the singleton $\{1\} \subset \mathbb{R}$.

The function f is a polynomial. For the purposes of this example, let us consider it known that polynomials are continuous functions — this easy result will be proven later in this course as well. The singleton $\{1\} \subset \mathbb{R}$ is a closed set (apply Exercise VII.6). Thus the ellipse E is closed, as the preimage of the closed set $\{1\} \subset \mathbb{R}$ under the continuous function f.

Reasoning similarly, we can conclude that the inside region of the ellipse E,

$$\left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; \frac{x^2}{a^2} + \frac{y^2}{b^2} < 1 \right\} = f^{-1}\big[(-\infty, 1)\big],$$

is open. Indeed, it is the preimage of the open set $(-\infty, 1) \subset \mathbb{R}$ under the same continuous function f. The case of the outside region of the ellipse, $f^{-1}\big[(1,\infty)\big] \subset \mathbb{R}^2$, is again similar.

Example VII.27 (Continuous functions on an interval vanishing at the two endpoints)**.**
Let $[a,b] \subset \mathbb{R}$ be a closed interval and consider the space $\mathcal{C}\left([a,b]\right)$ of continuous real-valued functions on this interval equipped with the supremum-norm $\|\cdot\|_\infty$ and the metric induced by this norm.

Consider also the subset

$$A = \left\{ f \in \mathcal{C}\left([a,b]\right) \;\middle|\; f(a) = 0,\ f(b) = 0 \right\} \;\subset\; \mathcal{C}\left([a,b]\right)$$

consisting of those functions which vanish at both endpoints, a and b.

We claim that this set $A \subset \mathcal{C}\left([a,b]\right)$ is closed. To see this, recall first that $f(a) = \mathrm{ev}_a(f)$ and $f(b) = \mathrm{ev}_b(f)$, where $\mathrm{ev}_z\colon \mathcal{C}\left([a,b]\right) \to \mathbb{R}$ are the evaluation functions studied in Example VI.9. In terms of these, the set A can be written as an intersection of preimages

$$\begin{aligned} A &= \left\{ f \in \mathcal{C}\left([a,b]\right) \;\middle|\; \mathrm{ev}_a(f) = 0,\ \mathrm{ev}_b(f) = 0 \right\} \\ &= \mathrm{ev}_a^{-1}\big[\,\{0\}\,\big] \cap \mathrm{ev}_b^{-1}\big[\,\{0\}\,\big]. \end{aligned}$$

The singleton $\{0\} \subset \mathbb{R}$ is closed, so since $\mathrm{ev}_a, \mathrm{ev}_b\colon \mathcal{C}\left([a,b]\right) \to \mathbb{R}$ are continuous, the preimages $\mathrm{ev}_a^{-1}[\{0\}], \mathrm{ev}_b^{-1}[\{0\}] \subset \mathcal{C}\left([a,b]\right)$ are also closed. Therefore $A \subset \mathcal{C}\left([a,b]\right)$ is closed, as the intersection of two closed sets.

To further emphasize the point that the characterization of continuity in Theorem VII.22 is useful, let us revisit the proof of the "if" implication in Theorem VII.9 about open sets in a subspace. The more abstract approach makes things easier — below we obtain a quick proof without the need to pick points x', choose radii $r_{x'}$, or juggle with balls in the subspace and balls in the full space.

Remark VII.28 (An alternative quick proof of the "if" part of Theorem VII.9)**.**
Let X be a metric space, and let $X' \subset X$ be a subset, equipped with the inherited metric (V.8). Let us provide a different proof of the "if" implication of Theorem VII.9.

So, let $U \subset X$ be an open set in the full space, and consider $U' = U \cap X'$. Recall from Exercise VI.1 that the subspace inclusion $\iota\colon X' \to X$ is continuous. An obvious observation about the preimages under the inclusion is that $\iota^{-1}[U] = U \cap X' = U'$. So $U' \subset X'$ is indeed open, as the preimage of the open set $U \subset X$ under the continuous function ι.

You can practice this important technique for example by the following.

Exercise VII.12 (Verifying openness and closedness).
Show that the set

$$U = \left\{ (x,y,z) \in \mathbb{R}^3 \;\middle|\; x^2 < y^2 + z^2 - xyz + 3 \right\}$$

is open and that the set

$$F = \left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; x^2 + y^2 \leq 2 \text{ and } x \leq 1 + \frac{1}{3}\sin(y) \right\}$$

is closed.
Hint: *Express the sets U and F appropriately using preimages; for the latter it is best to use two different functions and an intersection of preimages.*
The continuity of the functions involved can be considered known.

Exercise VII.13 (Verifying closedness of a set of probability densities).
Consider the space $\mathcal{C}([-1,1])$ of all continuous functions $f\colon [-1,1] \to \mathbb{R}$, equipped with the metric induced by the sup-norm $\|\cdot\|_\infty$. Consider the subset[10]

$$D \;=\; \left\{ p \in \mathcal{C}([-1,1]) \;\middle|\; p(x) \geq 0 \;\forall x \in [-1,1],\; \int_{-1}^{1} p(x)\,\mathrm{d}x = 1 \right\}.$$

Show that $D \subset \mathcal{C}([-1,1])$ is a closed set.
Hint: *You may use the facts that evaluation functions $f \mapsto f(x)$ (for an arbitrary $x \in [-1,1]$), and the integration function $f \mapsto \int_{-1}^{1} f(x)\,\mathrm{d}x$ are continuous functions $\mathcal{C}([-1,1]) \to \mathbb{R}$ with the chosen metric (recall Example VI.9 and Exercise VI.5). Otherwise the ideas are similar to Exercise VII.12.*

Exercise VII.14 (The cone of positive functions on a closed interval).
Let $[a,b] \subset \mathbb{R}$ be a closed interval and, as in Example IV.15, consider the space $\mathcal{C}([a,b])$ of continuous real-valued functions on this interval equipped with the supremum-norm $\|\cdot\|_\infty$ and the metric induced by this norm.

Consider the subset

$$P = \left\{ f \in \mathcal{C}([a,b]) \;\middle|\; f(x) > 0 \text{ for all } x \in [a,b] \right\} \;\subset\; \mathcal{C}([a,b])$$

consisting of all positive functions. Show that $P \subset \mathcal{C}([a,b])$ is an open set.
Hint: *You may want to apply a small modification of Exercise VI.4.*

Composition of continuous functions

We have so far still had only quite limited examples of continuous functions. To deal with more intricate ones, it is important to be able to construct new continuous functions from simpler building blocks. The operation of composition of functions is one of the most important constructions of this kind.

Theorem VII.29 (Compositions of continuous functions are continuous).
Let (X, d_X), (Y, d_Y), (Z, d_Z) be three metric spaces, and let

$$f\colon X \to Y \qquad \text{and} \qquad g\colon Y \to Z$$

[10] This subset P could be interpreted as the set of all continuous probability density functions supported on the interval $[-1,1]$.

be continuous functions. Then the composition $g \circ f : X \to Z$*, defined by*

$$\big(g \circ f\big)(x) := g(f(x)) \quad \text{for } x \in X,$$

is also a continuous function.

$$\begin{tikzcd} X \arrow[r, "f"] \arrow[rr, bend right, "g\circ f"'] & Y \arrow[r, "g"] & Z \end{tikzcd}$$

Proof. By the characterization of Theorem VII.22, it suffices to show that for all open subsets $W \subset Z$, the preimages $(g \circ f)^{-1}[W] \subset X$ are also open. Observe first the general set-theoretic fact that $(g \circ f)^{-1}[W] = f^{-1}\big[g^{-1}[W]\big]$ for any subset $W \subset Z$.

Let $W \subset Z$ be open. By the assumed continuity of g we have that its preimage $V = g^{-1}[W] \subset Y$ is also open. Then by the assumed continuity of f we have that the preimage $f^{-1}[V] \subset X$ of this open set is also open. On the other hand, we have

$$f^{-1}[V] = f^{-1}\big[g^{-1}[W]\big] = (g \circ f)^{-1}[W].$$

We conclude that $(g \circ f)^{-1}[W] \subset X$ is open. This shows that the composed function $g \circ f : X \to Z$ is continuous. □

If you write down a proof of Theorem VII.29 directly using the ε-δ definition of continuity, you will notice that it becomes lengthier and more convoluted. The simplicity of proofs like these is one advantage of the characterization of continuity in Theorem VII.22.

VII.3. Homeomorphism

Slightly informally, homeomorphisms are functions that preserve topology. Let's start by intuitively describing what this means.

If a metric space X can be mapped continuously onto another metric space Y, then one can interpret that the "shape" of the space X can be continuously deformed to that of Y without tearing anything apart (no discontinuities). If the mapping has an inverse function which is also continuous, and thus intuitively also continuously deforms Y to X without tearing anything apart, the two "shapes" can be mutually obtained from each other by continuous deformations. This relation among spaces is known as homeomorphism. The precise definition is the following.

Definition VII.30 (Homeomorphism)**.**
Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is called a **homeomorphism** if it is bijective and both $f : X \to Y$ and its inverse $f^{-1} : Y \to X$ are continuous.

If a homeomorphism $f : X \to Y$ exists, the spaces X and Y are said to be **homeomorphic**, and we denote $X \approx Y$.

Lemma VII.31 (Homeomorphism is an equivalence relation)**.**
Homeomorphism is an equivalence relation among metric spaces, i.e., the following properties hold:
(i) *For any metric space* X *we have* $X \approx X$.
(ii) *If* $X \approx Y$*, then we also have* $Y \approx X$.
(iii) *If* $X \approx Y$ *and* $Y \approx Z$*, then we also have* $X \approx Z$.

Proof. Each of the properties is very easy to prove:

(i): The identity function $\mathrm{id}_X \colon X \to X$ is bijective, continuous, and its inverse $\mathrm{id}_X^{-1} = \mathrm{id}_X$ is also continuous. Therefore id_X provides a homeomorphism $X \approx X$.

(ii): If $f \colon X \to Y$ is a homeomorphism, then also its inverse $f^{-1} \colon Y \to X$ is a homeomorphism: f^{-1} is bijective (since f is bijective), f^{-1} is continuous (as the inverse of the homeomorphism f), and its inverse $(f^{-1})^{-1} = f$ is continuous (since f is a homeomorphism).

(iii): If $f \colon X \to Y$ and $g \colon Y \to Z$ are homeomorphisms, then the composition $g \circ f \colon X \to Z$ is also a homeomorphism: $g \circ f$ is bijective (Exercises I.9 and I.10), $g \circ f$ is continuous (as the composition of the continuous functions f and g), and its inverse $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$ is continuous (as the composition of the continuous functions g^{-1} and f^{-1}). □

Let us start with some very simple examples of homeomorphisms among intervals on the real line. In all of the following, as usual, we equip the intervals with the standard metric they inherit from the real line.

Example VII.32 (All nondegenerate closed intervals are mutually homeomorphic)**.**
Let $[a_1, b_1] \subset \mathbb{R}$ and $[a_2, b_2] \subset \mathbb{R}$ be two closed intervals which are both nondegenerate, $a_1 < b_1$ and $a_2 < b_2$. Then the function

$$f \colon [a_1, b_1] \to [a_2, b_2] \qquad f(x) = a_2 + \frac{b_2 - a_2}{b_1 - a_1}(x - a_1) \quad \text{for } x \in [a_1, b_1]$$

is a homeomorphism: it is clearly bijective and M-Lipschitz with $M = \frac{b_2 - a_2}{b_1 - a_1} > 0$, and its inverse $y \mapsto a_1 + \frac{b_1 - a_1}{b_2 - a_2}(y - a_2)$ is $\frac{1}{M}$-Lipschitz, so both f and f^{-1} are continuos (by Theorem VI.6). This shows that any two nondegenerate closed intervals are homeomorphic to each other, $[a_1, b_1] \approx [a_2, b_2]$.

Exercise (✓) VII.15 (All bounded open intervals are mutually homeomorphic)**.**
Show that if $a_1 < b_1$ and $a_2 < b_2$, then the bounded open intervals $(a_1, b_1) \subset \mathbb{R}$ and $(a_2, b_2) \subset \mathbb{R}$ are homeomorphic, $(a_1, b_1) \approx (a_2, b_2)$.

Exercise (✓) VII.16 (Bounded half-open intervals are mutually homeomorphic)**.**
Show that if $a_1 < b_1$ and $a_2 < b_2$, then there are homeomorphisms $(a_1, b_1] \approx (a_2, b_2]$ and also $(a_1, b_1] \approx [a_2, b_2)$.

Exercise VII.17 (Bounded and unbounded half-open intervals are homeomorphic)**.**
Show that $[0, 1)$ and $[0, \infty)$ are homeomorphic.
Hint: *You can use the fact that rational functions are continuous. This was stated in Corollary III.12, and we admit for now that the definitions of continuity given in Lectures III and VI agree (this will be proven in Corollary VIII.11).*

Remark VII.33 (All bounded and unbounded half-open intervals are homeomorphic)**.**
By Exercise VII.17 we have a homeomorphism $[0, \infty) \approx [0, 1)$. By Exercise VII.16 we also have homeomorphism $[0, 1) \approx [a, b)$ and $[0, 1) \approx (a, b]$ for any $a < b$. By Lemma VII.31(iii) we get that the non-negative real axis $[0, \infty)$ is homeomorphic to any bounded open interval, $[0, \infty) \approx [a, b)$ and $[0, \infty) \approx (a, b]$. It is very easy to also see that $[0, \infty) \approx [a, \infty)$ and $[0, \infty) \approx (-\infty, b]$, so in fact all bounded and unbounded half-open intervals are mutually homeomorphic.

Similarly one can prove that bounded open intervals and unbounded open intervals are homeomorphic. Let us provide one such homeomorphism.

Example VII.34 (A homeomorphism from an open interval to the real line)**.**
Consider the open interval $\big(-\frac{\pi}{2},\frac{\pi}{2}\big)$. The tangent function

$$\tan\colon \Big(-\frac{\pi}{2},\frac{\pi}{2}\Big) \to \mathbb{R}$$

is bijective; its inverse is the arcustangent $\arctan\colon \mathbb{R} \to \big(-\frac{\pi}{2},\frac{\pi}{2}\big)$. We consider it known that both tan and arctan are continuous.[11] Therefore we get a homeomorphism $\big(-\frac{\pi}{2},\frac{\pi}{2}\big) \approx \mathbb{R}$.

One might suspect that closed intervals are not homeomorphic to open intervals, and that neither open nor closed intervals are homeomorphic to half-open intervals. This is indeed the case. You could try to prove it directly, but after Lectures XI and XII we have better general tools which make the proofs significantly easier.

All of the above examples are very easy homeomorphisms. Let us now give one example of something natural that turns out *not* to be a homeomorphism.

Example VII.35 (A continuous map from interval to circle is not a homeomorphism)**.**
Consider the unit circle

$$\mathbf{S}^1 \;=\; \Big\{(x,y)\in\mathbb{R}^2 \;\Big|\; x^2+y^2=1\Big\}.$$

and the half-open interval $[0,2\pi)$. It is very natural to define a mapping

$$f\colon\; [0,2\pi)\to\mathbf{S}^1 \qquad\qquad f(\theta)=\big(\cos(\theta),\,\sin(\theta)\big) \quad \text{for } \theta\in[0,2\pi).$$

This f is bijective. The trigonometric functions $\theta\mapsto\cos(\theta)$ and $\theta\mapsto\sin(\theta)$ have derivatives bounded by 1 and as such are 1-Lipschitz (see Example VI.7), and from this it is not difficult to show that also f is Lipschitz (the Lipschitz constant $M=1$ works but $M=2$ may be slightly easier to prove) and in particular f is continuous.

But f is *not* a homeomorphism between $[0,2\pi)$ and $\mathbf{S}^1$!

Indeed, the inverse $f^{-1}\colon\mathbf{S}^1\to[0,2\pi)$ is not continuous. In $[0,2\pi)$, consider an open ball[12] of small positive radius $\varepsilon<2\pi$ centered at $0\in[0,2\pi)$ — it is of the form $[0,\varepsilon)\subset[0,2\pi)$. The image $f\big[[0,\varepsilon)\big]\subset\mathbf{S}^1$ contains $(1,0)=f(0)\in\mathbf{S}^1$, but does not contain the point $\big(\cos(2\pi-\delta),\sin(2\pi-\delta)\big)=f(2\pi-\delta)\in\mathbf{S}^1$ for any small $\delta\in(0,2\pi-\varepsilon)$. Since these points are at distances $\mathsf{d}\big(f(0),f(2\pi-\delta)\big)<M\delta$ from $f(0)=(1,0)$, we see that no open ball in $\mathbf{S}^1$ centered at $f(0)$ is contained in $f\big[[0,\varepsilon)\big]$, so the image $f\big[[0,\varepsilon)\big]\subset\mathbf{S}^1$ is not open. Applying this to preimages under the inverse function $f^{-1}\colon\mathbf{S}^1\to[0,2\pi)$, we see that $(f^{-1})^{-1}\big[[0,\varepsilon)\big]=f\big[[0,\varepsilon)\big]$ is not open, and since $[0,\varepsilon)\subset[0,2\pi)$ was an open set, this shows (by Theorem VII.22) that the inverse $f^{-1}\colon\mathbf{S}^1\to[0,2\pi)$ is not a continuous function. In particular, f is not a homeomorphism.

The above reasoning only shows that this particular function f is not a homeomorphism between the half-open interval $[0,2\pi)$ and the circle $\mathbf{S}^1$ — it does not say that $[0,2\pi)$ and $\mathbf{S}^1$ are not homeomorphic (i.e., that no homeomorphism exists between them). In Lecture XI we will have good tools to prove that such a homeomorphism indeed is impossible, and we have $[0,2\pi)\not\approx\mathbf{S}^1$. The intuitive idea, however, should be clear: a bijection in the direction $\mathbf{S}^1\to[0,2\pi)$ cannot be continuous — it has to tear apart the circle at some point.

Since all of the above examples have pertained to merely “one-dimensional spaces” (“lines and curves”), let us take one slightly different example.

[11] A precise justification of their continuity would use results from Lecture IX to the power series representations of trigonometric functions.

[12] Note that we are taking open balls in the metric space $[0,2\pi)$ equipped with the metric it inherits from $\mathbb{R}$. Following the ideas of Theorem VII.9, these open balls are just intersections of open balls on $\mathbb{R}$ with $[0,2\pi)\subset\mathbb{R}$. In particular the ball of radius $\varepsilon<2\pi$ centered at 0 is $\mathcal{B}_\varepsilon(0)\;=\;(-\varepsilon,\varepsilon)\cap[0,2\pi)\;=\;[0,\varepsilon)$.

Exercise VII.18 (The open unit disk is homeomorphic to the plane).

Prove that the open unit disk $\mathcal{B}_1(\vec{0}) \subset \mathbb{R}^2$ in the Euclidean plane $\mathbb{R}^2$ and the whole plane $\mathbb{R}^2$ are homeomorphic, $\mathcal{B}_1(\vec{0}) \approx \mathbb{R}^2$.

Hint: *One idea is a radial mapping in polar coordinates, using Exercise VII.17. Detailed proofs of continuity may require some work until we develop better tools later...*

The following exercise gives a (mostly quantitative) sufficient condition for the homeomorphism property. It is evidently related to the observation of Theorem VI.6 that the Lipschitz property is a sufficient condition for continuity.

Exercise VII.19 (Surjective bilipschitz functions are homeomorphisms).

Let (X, d_X) and (Y, d_Y) be two metric spaces, and let $M \geq 1$. A function $f\colon X \to Y$ is said to be M**-bilipschitz** if for all $x_1, x_2 \in X$ we have

$$\frac{1}{M}\,\mathsf{d}_X(x_1, x_2) \;\leq\; \mathsf{d}_Y\big(f(x_1), f(x_2)\big) \;\leq\; M\,\mathsf{d}_X(x_1, x_2). \tag{VII.1}$$

(a) Show that any M-bilipschitz function is injective.
(b) Show that any surjective M-bilipschitz function is a homeomorphism.

Fun: homeomorphism classes of letters

Let us borrow a fun example from [**Väi99**]: consider the capital letters

$$\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}, \mathsf{E}, \mathsf{F}, \mathsf{G}, \mathsf{H}, \mathsf{I}, \mathsf{J}, \mathsf{K}, \mathsf{L}, \mathsf{M}$$
$$\mathsf{N}, \mathsf{O}, \mathsf{P}, \mathsf{Q}, \mathsf{R}, \mathsf{S}, \mathsf{T}, \mathsf{U}, \mathsf{V}, \mathsf{W}, \mathsf{X}, \mathsf{Y}, \mathsf{Z}$$

as subsets of the plane $\mathbb{R}^2$, so they in particular inherit a metric and become metric spaces. Which of these are homeomorphic?

Some "obvious" homeomorphisms are $\mathsf{D} \approx \mathsf{O}$ (both homeomorphic to the circle $\mathbf{S}^1$) and $\mathsf{J} \approx \mathsf{S}$ (both homeomorphic to the interval $[0, 1]$) and $\mathsf{Y} \approx \mathsf{T}$. Let us not spoil any more of the fun — you should try to classify the letters[13] into equivalence classes under homeomorphisms (feel free to resort to intuitive arguments, since we did not even give precise definitions of the subsets corresponding to the letters...[14]).

The quintessential topology joke

The donut is homeomorphic to the coffee cup.

`https://www.shapeways.com/product/6CJQ9GXWW/topology-joke`

Homeomorphisms as mappings that preserve topology

Observe the following.

Lemma VII.36 (A homeomorphism preserves the openness of sets).

Let $f\colon X \to Y$ be a homeomorphism. Then a subset $A \subset X$ is open if and only if its image $f[A] \subset Y$ is open.

[13] If you didn't get enough, include numbers $\mathsf{0}, \mathsf{1}, \mathsf{2}, \mathsf{3}, \mathsf{4}, \mathsf{5}, \mathsf{6}, \mathsf{7}, \mathsf{8}, \mathsf{9}$ as well! (Did $\mathsf{4}$ and Q offer a small surprise, and why?) Then the lowercase letters $\mathsf{a}, \mathsf{b}, \mathsf{c}, \ldots$ (how about å, ä, ö?), the Greek alphabet $\alpha, \beta, \gamma, \ldots$, simplified Chinese characters, traditional Chinese characters, ...

[14] This would require a fully precise definition of the font used.

Proof. Suppose that $A \subset X$ is such that $f[A] \subset Y$ is open. Then by Theorem VII.22 and the continuity of $f \colon X \to Y$, the preimage $f^{-1}\big[f[A]\big] \subset X$ is also open. By bijectivity of f we have $f^{-1}\big[f[A]\big] = A$, so $A \subset X$ is open.

Suppose then that $A \subset X$ is open. Then by Theorem VII.22 and the continuity of the inverse function $f^{-1} \colon Y \to X$, its preimage $(f^{-1})^{-1}[A] \subset Y$ is also open. But we have $(f^{-1})^{-1}[A] = f[A]$, so the image $f[A] \subset Y$ is open. □

Also the converse holds.

Exercise (✓) VII.20 (Bijections preserving openness are homeomorphisms)**.**
Suppose that $f \colon X \to Y$ is a bijection with the property that subsets $A \subset X$ are open if and only if their images $f[A] \subset Y$ are open. Prove that then f is a homeomorphism.

In view of the above, homeomorphisms are precisely bijective mappings between metric spaces which preserve the openness of subsets. The collection of all open subsets of a space is called the topology of the space, see Appendix D for a formal definition. In view of this, homeomorphisms are precisely bijections that preserve topology.

VII.4. Equivalence of metrics

A given set X can be equipped with various different metrics. The choice of metric d on X determines the open sets (i.e., the topology on X). In general, different metrics lead to different notions of open sets. But when different choices of metric actually yield the same open sets, they can for many purposes be used interchangeably.

Definition VII.37 (Topologically equivalent metrics)**.**
Let X be a set. Let d_1 and d_2 be two metrics on X, and let

$$\begin{aligned} \mathscr{T}_1 &= \big\{ U_1 \subset X \;\big|\; U_1 \text{ is open with respect to } \mathsf{d}_1 \big\} \\ \mathscr{T}_2 &= \big\{ U_2 \subset X \;\big|\; U_2 \text{ is open with respect to } \mathsf{d}_2 \big\} \end{aligned}$$

be the corresponding collections of open subsets with respect to each of the two. We call the metrics d_1 and d_2 **(topologically) equivalent** if the collections of open sets determined by both are the same, $\mathscr{T}_1 = \mathscr{T}_2$.

So, any property that can be phrased just in terms of open sets (i.e., any topological property), is going to be unchanged if we replace one metric with another (topologically) equivalent one. As a very important example, note that Theorem VII.22 characterizes continuity of functions between two metric spaces in terms of open sets in them, so continuity is not affected by changing a metric on the domain or on the codomain to a topologically equivalent one.

The most important practical situation in which we will use this notion is metrics on product spaces, see Section V.4. Let us nevertheless first give a few simpler examples.

The following simple observation shows that there are always infinitely many metrics equivalent with any given one (barring the trivial cases when the set X is either empty or a singleton, in which cases there are no nozero distances).

Example VII.38 (Constant multiple of the same metric)**.**
Let $\mathsf{d}_1 \colon X \times X \to [0, \infty)$ be a metric on X, and $c > 0$ a constant. Define $\mathsf{d}_2 \colon X \times X \to [0, \infty)$ by $\mathsf{d}_2(x, y) = c\, \mathsf{d}_1(x, y)$ for all $x, y \in X$. Properties (M-s), (M-Δ), and (M-0) for d_2 clearly follow from the correspodning properties for d_1, so d_2 is also a metric on X.

Let us denote the open balls of radius $r > 0$ centered at $z \in X$ with respect to d_1 and d_2 by

$$\mathcal{B}^1_r(z) = \big\{ x \in X \;\big|\; \mathsf{d}_1(z, x) < r \big\} \qquad\qquad \mathcal{B}^2_r(z) = \big\{ x \in X \;\big|\; \mathsf{d}_2(z, x) < r \big\} .$$

We clearly have $\mathcal{B}^1_r(z) = \mathcal{B}^2_{cr}(z)$ for any $r > 0$ and $z \in X$. In view of Definition VII.1 it is therefore clear that a subset $U \subset X$ is open with respect to d_1 if and only if it is open with respect to d_2, i.e., $U \in \mathscr{T}_1 \;\Leftrightarrow\; U \in \mathscr{T}_2$, where $\mathscr{T}_1$ and $\mathscr{T}_2$ are the collections of open sets determined by the metrics d_1 and d_2. In other words, $\mathscr{T}_1 = \mathscr{T}_2$, so the metrics d_1 and d_2 are topologically equivalent.

Such a multiplication of distances by a constant $c > 0$ can be thought of as a "change of units of distances": if one person measures distances in miles and another in kilometres, they disagree on the numerical values of distances, but they (hopefully) agree on the "shape" of the space they both live in.

From the following we see that any metric is topologically equivalent to a bounded metric.

Example VII.39 (Truncation of a metric)**.**
Let $\mathsf{d}_1 \colon X \times X \to [0, \infty)$ be a metric on X, and $c > 0$ a constant. Define $\mathsf{d}_2 \colon X \times X \to [0, \infty)$ by $\mathsf{d}_2(x, y) = \min \{\mathsf{d}_1(x, y),\, c\}$ for all $x, y \in \mathsf{d}$. Properties (M-s) and (M-0) for d_2 clearly follow from the correspodning properties for d_1. Also the triangle inequality (M-Δ) for d_2 holds by virtue of the easy observation

$$\min \{\mathsf{d}_1(x, y) + \mathsf{d}_1(y, z),\; c\} \le \min \{\mathsf{d}_1(x, y),\, c\} + \min \{\mathsf{d}_1(y, z),\, c\}$$

and the triangle inequality for d_1. Therefore d_2 is also a metric on X.

Denote the open balls of radius $r > 0$ centered at $z \in X$ with respect to d_1 and d_2 by $\mathcal{B}^1_r(z)$ and $\mathcal{B}^2_r(z)$ again. For $0 < r \le c$ we have $\mathcal{B}^1_r(z) = \mathcal{B}^2_r(z)$ for any $z \in X$. In Definition VII.1 of open sets we could have restricted the radius of balls to be below the constant $c > 0$ (if we found larger radius balls around points, then we certainly could use smaller ones as well). It is therefore clear that a subset $U \subset X$ is open with respect to d_1 if and only if it is open with respect to d_2. As before, this says that metrics d_1 and d_2 are topologically equivalent.

Such a truncation of distances to at most a given constant $c > 0$ can be thought of as "allowing teleportation at a fixed cost c": for any distances greater than c it is then advantageous to resort to the teleportation option. But the local "geometry" (at distances below c) remains unchanged, and this is what determines the topology.

The equivalence of metrics can be phrased in terms of homeomorphisms. Let us start by a lemma.

Lemma VII.40 (Continuity of the identity function with different metrics)**.**
Let X be a set. Let d_1 and d_2 be two metrics on X, and let $\mathscr{T}_1$ and $\mathscr{T}_2$ be the corresponding collections of open sets. Consider the identity mapping of the set X

$$\mathrm{id}_X \colon\; X \to X \qquad\qquad \mathrm{id}_X(x) = x \quad \text{for } x \in X,$$

as a mapping between the metric spaces (X, d_1) and (X, d_2); i.e., the domain X equipped with metric d_1 and the codomain X with the metric d_2. Then id_X is continuous if and only if $\mathscr{T}_2 \subset \mathscr{T}_1$.

Proof. We prove the "if" and "only if" separately.

So suppose first that id_X is continuous. If $U \in \mathscr{T}_2$, i.e., if U is an open subset of the codomain, then by continuity of id_X and Theorem VII.22 we get that the set $\mathrm{id}_X^{-1}[U] = U$ is open in the domain, i.e., $U \in \mathscr{T}_1$. This shows $\mathscr{T}_2 \subset \mathscr{T}_1$.

Conversely, suppose that $\mathscr{T}_2 \subset \mathscr{T}_1$. We prove continuity of id_X by the characterization of Theorem VII.22. So let U be an open subset of the codomain, i.e., $U \in \mathscr{T}_2$. The assumption $\mathscr{T}_2 \subset \mathscr{T}_1$ implies that then $U \in \mathscr{T}_1$ also. Since $\mathrm{id}_X^{-1}[U] = U$, this shows that the preimage of U under id_X is open in the domain. As this was done for an arbitrary open set U in the codomain, we have established the continuity of id_X. □

We now get the characterization of the equivalence of two metrics.

Theorem VII.41 (A characterization of equivalence of metrics)**.**
Let X be a set. Let d_1 and d_2 be two metrics on X. Consider the identity mapping of the set X

$$\mathrm{id}_X \colon\ X \to X \qquad\qquad \mathrm{id}_X(x) = x \quad \textit{for } x \in X,$$

as a mapping between the metric spaces (X, d_1) and (X, d_2) Then id_X is a homeomorphism if and only if the metrics d_1 and d_2 are topologically equivalent.

Proof. Note that the identity function $\mathrm{id}_X \colon X \to X$ is certainly a bijection, so it is a homeomorphism if and only if both id_X and id_X^{-1} are continuous. Therefore we must analyze how the choice of metrics affects the continuity of these.

The first part is already done: by Lemma VII.40, id_X is continuous if and only if $\mathscr{T}_2 \subset \mathscr{T}_1$.

Set-theoretically, the inverse function of the identity is itself, i.e., $\mathrm{id}_X^{-1} \colon X \to X$ is also the identity function. But in order to address continuity, the domain and codomain are not viewed merely as sets — they are equipped with their own metrics. The inverse function exchanges the roles of the domain and the codomain, so the domain of id_X^{-1} is equipped with the metric d_2 and the codomain of id_X^{-1} is equipped with the metric d_1. Applying Lemma VII.40 to this shows that id_X^{-1} is continuous if and only if $\mathscr{T}_1 \subset \mathscr{T}_2$.

Combining the two observations, we see that both id_X and id_X^{-1} are continuous if and only if $\mathscr{T}_1 = \mathscr{T}_2$, which proves the assertion. □

VII.5. The topology on product spaces

Recall from Section V.4 that the Cartesian product $X \times Y$ of two metric spaces X and Y can be equipped with various different metrics.

Equivalence of the product space metrics

An important observation is that the choice of metric does not matter for many purposes (for topological properties) — the metrics given by (V.21) – (V.23) are (topologically) equivalent.

Theorem VII.42 (The natural metrics on the product space are equivalent)**.**
The metrics $\mathsf{d}_1, \mathsf{d}_2, \mathsf{d}_\infty$ on $X \times Y$ given by (V.21) – (V.23) *are topologically equivalent.*

Proof. We will prove the equivalence of d_1 with d_∞, and sketch the proof of equivalence of d_1 with d_2. The equivalence of d_2 with d_∞ is obtained by combining these (by transitivity).

For both equivalences, according to Theorem VII.41, we need to show that the identity map of $X \times Y$ is a homeomorphism, when the domain and codomain are separately equipped with the corresponding metrics. By Exercise VII.19, the homeomorphism property will follow, if we show that the identity map is bilipschitz (the identity map is obviously surjective).

Let us first consider the identity map

$$\mathrm{id}\colon X \times Y \to X \times Y,$$

when the domain is equipped with the metric d_1 of (V.21) and the codomain with the metric d_∞ of (V.23). We claim that it is 2-bilipschitz. Let $(x_1, y_1), (x_2, y_2) \in X \times Y$. The definition of the metrics and an easy estimate yields

$$\begin{aligned} \mathsf{d}_\infty\big((x_1,y_1),\,(x_2,y_2)\big) \;&=\; \max\big\{\mathsf{d}_X(x_1,x_2),\; \mathsf{d}_Y(y_1,y_2)\big\} \\ &\le\; \mathsf{d}_X(x_1,x_2) + \mathsf{d}_Y(y_1,y_2) \\ &=\; \mathsf{d}_1\big((x_1,y_1),\,(x_2,y_2)\big). \end{aligned}$$

Similarly estimating in the other direction, we find

$$\begin{aligned} \mathsf{d}_1\big((x_1,y_1),(x_2,y_2)\big) \;&=\; \mathsf{d}_X(x_1,x_2) + \mathsf{d}_Y(y_1,y_2) \\ &\le\; 2\,\max\big\{\mathsf{d}_X(x_1,x_2),\; \mathsf{d}_Y(y_1,y_2)\big\} \\ &=\; 2\,\mathsf{d}_\infty\big((x_1,y_1),(x_2,y_2)\big). \end{aligned}$$

We have thus shown

$$\mathsf{d}_\infty\big((x_1,y_1),\,(x_2,y_2)\big) \;\le\; \mathsf{d}_1\Big(\mathrm{id}(x_1,y_1),\; \mathrm{id}(x_2,y_2)\Big) \;\le\; 2\,\mathsf{d}_\infty\big((x_1,y_1),\,(x_2,y_2)\big).$$

The 2-bilipschitzness of id follows, because multiplying the leftmost expression by $\frac{1}{2}$ can only make it even smaller.

The details of the other bilipschitzness verification is left as Exercise VII.21. $\square$

Exercise VII.21 (Completing the proof of Theorem VII.42)**.**
Consider the identity map

$$\mathrm{id}\colon X \times Y \to X \times Y,$$

when the domain is equipped with the metric d_1 of (V.21) and the codomain with the metric d_2 of (V.22). Prove that it is bilipschitz.

Continuity of projections

Recall from (V.18) that on the product space, we define the two coordinate projections $\mathrm{pr}_1\colon X \times Y \to X$ and $\mathrm{pr}_2\colon X \times Y \to Y$ by $(x, y) \mapsto x$ and $(x, y) \mapsto y$, respectively. In Exercise VI.9 you have verified that these are continuous when their domain $X \times Y$ is equipped with any of the metrics (V.21) – (V.23). In view of the equivalence of these metrics, Theorem VII.42, and the fact that continuity can be phrased in terms of open sets only, we now see that it is actually sufficient to check the continuity for just one of these cases and the others then follow automatically.

Continuity in terms of component functions

Let (X, d_X), (Y, d_Y), (Z, d_Z) be three metric spaces. Recall from (V.19) the idea of specifying a function $Z \to X \times Y$ in terms of its two component functions: a pair of functions $f_1\colon Z \to X$ and $f_2\colon Z \to Y$ determines a function to the product space,

$$f\colon Z \to X \times Y \quad \text{given by} \quad f(z) = \big(f_1(z), f_2(z)\big) \;\text{ for } z \in Z.$$

In Exercise VI.10 you have verified that if f_1 and f_2 are continuous and the product space $X \times Y$ is equipped with any of the metrics (V.21) – (V.23), then f is continuous. In view of the equivalence of these metrics, Theorem VII.42, and the fact that continuity can be phrased in terms of open sets only, we now see that it is actually sufficient to check the continuity for just one of these cases and the others then follow automatically.

Continuity of the metric

Let (X, d) be a metric space. In Exercise VI.11 you have shown that the metric $\mathsf{d}\colon X \times X \to [0, \infty)$ is a continuous function, when the domain $X \times X$ is equipped with the metric d_1 of (V.21). In view of the equivalence of product space metrics, Theorem VII.42, and the fact that continuity can be phrased in terms of open sets only, we now see that the continuity conclusion remains valid if the domain of d is equipped with the metric d_2 of (V.22) or the metric d_∞ of (V.23).

Exercises

Exercise VII.22 (The infinite cylinder is homeomorphic to the annulus)**.**
Consider the cylinder surface $C \subset \mathbb{R}^3$ and the annulus $A \subset \mathbb{R}^2$ given by

$$C = \left\{ (x, y, z) \in \mathbb{R}^3 \;\middle|\; x^2 + y^2 = 1 \right\}$$
$$A = \left\{ (\xi, \eta) \in \mathbb{R}^2 \;\middle|\; 1 < \xi^2 + \eta^2 < 2 \right\}$$

(we named the coordinates in $\mathbb{R}^3$ and $\mathbb{R}^2$ differently here to avoid confusion).
Prove that the cylinder C and the annulus A are homeomorphic, $C \approx A$.
Hint: *To construct a homomorphism between C and A, you can use some homeomorphism $(1, 2) \approx \mathbb{R}$ to make the radial direction of the annulus correspond to the z-coordinate of the cylider, while doing something simpler with the angular parts.*

Lecture VIII

Sequences in metric spaces

One of the straightforward but important notions that are easy to make sense of in metric spaces is sequences and their limits.

In this chapter we define limits of sequences in metric spaces, obtain some basic results about them, and use them to give further equivalent characterizations of continuity of functions and of closed sets and closures. These characterization are practical in certain situations, because working with sequnces is quite intuitive and easy. In particular, we will see how to use the characterization of continuity to prove that pointwise sums and products of continuous real-valued functions are continuous.

VIII.1. Sequences and limits in metric spaces

Throughout this section, let (X, d) be a metric space.

A **sequence** in the metric space (X, d) is an "infinite list"

$$(x_1, x_2, x_3, \ldots)$$

of points $x_1, x_2, x_3, \ldots \in X$. The precise meaning of this is again a function from the natural numbers to this space,

$$\xi \colon \mathbb{N} \to X,$$

and the "list" consists of the values $x_n = \xi(n)$ of that function, for $n \in \mathbb{N}$. As before, the sequence is usually denoted by

$$(x_1, x_2, x_3, \ldots) \qquad \text{or} \qquad (x_n)_{n=1}^{\infty} \qquad \text{or} \qquad (x_n)_{n \in \mathbb{N}}.$$

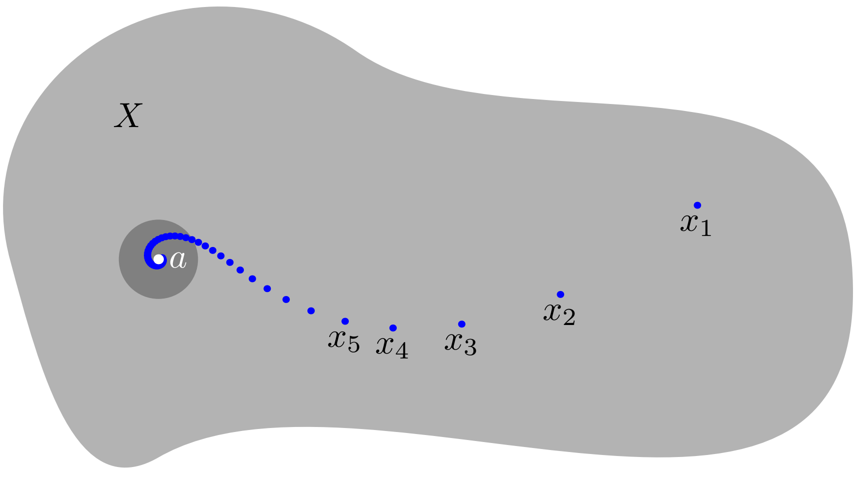


FIGURE VIII.1. A schematic illustration of a sequence $(x_n)_{n\in\mathbb{N}}$ and its limit a in a metric space X.

It is straightforward to generalize the notion of a limit of a sequence to metric spaces. Indeed, the intuitive meaning in the following definition remains exactly the same as in Definition II.11: however close we want to get to the limit, we achieve this by going far enough into the tail of the sequence. The only change is that for the distances we now use the metric d of the space.

Definition VIII.1 (Limit of a sequence in a metric space).
A sequence $(x_n)_{n\in\mathbb{N}}$ in a metric space X has a **limit** $a \in X$ if for all $\varepsilon > 0$ there exists an $n_\varepsilon \in \mathbb{N}$ such that $\mathsf{d}(x_n, a) < \varepsilon$ whenever $n \geq n_\varepsilon$. We then denote

$$\lim_{n\to\infty} x_n = a \qquad\qquad \text{or} \qquad\qquad x_n \to a \text{ as } n \to \infty.$$

Remark VIII.2 (Definition of a limit in logical symbols).
According to Definition VIII.1, the logical meaning of the statement

$$\lim_{n\to\infty} x_n = a$$

is

$$\forall \varepsilon > 0 : \quad \exists n_\varepsilon \in \mathbb{N} : \quad \forall n \geq n_\varepsilon : \quad \mathsf{d}(x_n, a) < \varepsilon.$$

Instead of saying that a sequence $(x_n)_{n\in\mathbb{N}}$ has a limit a, we occasionally say that $(x_n)_{n\in\mathbb{N}}$ **converges to** a, or that x_n **tends to** a (as n tends to ∞). If a sequence converges to some limit (especially in situations when we do not specify the limit explicitly), we say that the sequence **converges**, or that it is **convergent**.

By comparing Definition VIII.1 with Definition II.11 it is evident that the notion of a limit of a real number sequence as defined in Lecture II coincides with that of a limit of a sequence in the metric space $(\mathbb{R}, \mathsf{d}_\mathbb{R})$, the real line with its standard metric.

Note again that when a limit exists, it is unique. The proof here is different from the proof of uniqueness of real-number sequences (Corollary II.16). Indeed that proof in Lecture II used the order relation $\leq$ on the real line (and its preservation under limits), whereas a general metric space does not even have an order relation! Of course the general result that we prove here would also apply to the real line, so we did not really *need* the old proof — it was only given in Lecture II as an instructive application of the preservation of inequalities.

Lemma VIII.3 (Uniqueness of limits of sequences in metric spaces).
If a sequence $(x_n)_{n\in\mathbb{N}}$ in X has both $a \in X$ and $b \in X$ as its limits, then we necessarily have $a = b$.

Proof. Suppose that both $a \in X$ and $b \in X$ are limits of $(x_n)_{n\in\mathbb{N}}$. Let $\varepsilon > 0$. Then since $\frac{\varepsilon}{2} > 0$, by definition of limits there exists $n', n'' \in \mathbb{N}$ such that $\mathsf{d}(x_n, a) < \frac{\varepsilon}{2}$ for $n \geq n'$ and $\mathsf{d}(x_n, b) < \frac{\varepsilon}{2}$ for $n \geq n''$. In particular letting $n_0 = \max\{n', n''\}$ and applying the triangle inequality (M-Δ) to $a, x_{n_0}, b \in X$, we get

$$0 \;\leq\; \mathsf{d}(a,b) \;\leq\; \underbrace{\mathsf{d}(a, x_{n_0})}_{<\varepsilon/2} + \underbrace{\mathsf{d}(x_{n_0}, b)}_{<\varepsilon/2} \;<\; \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \;=\; \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this implies $\mathsf{d}(a,b) = 0$. By (M-0) we then get $a = b$. $\square$

Lemma VIII.3 in particular ensures that the notation $\lim_{n\to\infty} x_n$ for a convergent sequence $(x_n)_{n\in\mathbb{N}}$ is indeed unambiguous.

There are various conditions which are equivalent with the definition of a limit given above. Depending on the situation, it may be more convenient to verify or apply one of these instead of using Definition VIII.1 directly. The last of the equivalent conditions below refers only to the notion of an open set, and would actually serve as the definition of a limit in general topological spaces, where we do not in general have a metric at our disposal.

Theorem VIII.4 (Characterizations of limits in metric spaces).

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X and $x \in X$ a point. Then the following are equivalent:

(i) $\lim_{n\to\infty} x_n = x$;
(ii) $\lim_{n\to\infty} \mathsf{d}(x_n, x) = 0$;
(iii) *for any neighborhood U of x, only finitely many members of the sequence $(x_n)_{n\in\mathbb{N}}$ are not in U.*

Proof. We will prove the equivalence (i) $\Leftrightarrow$ (ii) and the implications (i) $\Rightarrow$ (iii) and (iii) $\Rightarrow$ (i) separately. The equivalence of all three conditions follows.

proof of (i) $\Leftrightarrow$ (ii): The logical equivalence of (i) and (ii) is clear from Definitions VIII.1 and II.11, once one notices that by non-negativity of metrics, the condition $\mathsf{d}(x_n, x) < \varepsilon$ is equivalent to $\big|\mathsf{d}(x_n, x) - 0\big| < \varepsilon$.

proof of (i) $\Rightarrow$ (iii): If $U \subset X$ is a neighborhood of x, then there exists an $\varepsilon > 0$ such that $\mathcal{B}_\varepsilon(x) \subset U$. If we assume $\lim_{n\to\infty} x_n = x$, then for some $n_\varepsilon \in \mathbb{N}$, and all $n \geq n_\varepsilon$ we have $x_n \in \mathcal{B}_\varepsilon(x) \subset U$. Therefore only the finitely many first members $x_1, x_2, \ldots, x_{n_\varepsilon - 1}$ are possibly not in U.

proof of (iii) $\Rightarrow$ (i): Assume that for any neighborhood U of x, only finitely many members of the sequence $(x_n)_{n\in\mathbb{N}}$ are not in U. Let $\varepsilon > 0$. The open ball $\mathcal{B}_\varepsilon(x)$, in particular, is a neighborhood of x, and therefore by assumption for only finitely many indices $n_1 < \cdots < n_m$ we have $x_{n_k} \notin \mathcal{B}_\varepsilon(x)$. For $n \geq n' = n_m + 1$ we then have $x_n \in \mathcal{B}_\varepsilon(x)$, i.e., $\mathsf{d}(x_n, x) < \varepsilon$. This shows that $\lim_{n\to\infty} x_n = x$. □

Boundedness of sequences in a metric space (X, d) is defined as follows.

Definition VIII.5 (Boundedness of a sequence).

A sequence $(x_n)_{n\in\mathbb{N}}$ in X is **bounded** if there exists some $z \in X$ and $r > 0$ such that $x_n \in \mathcal{B}_r(z)$ for all $n \in \mathbb{N}$.

Exercise VIII.1 (Boundedness of a sequence is boundedness of the set of its points).

Prove that a sequence $(x_n)_{n\in\mathbb{N}}$ in X is bounded (in the sense of Definition VIII.5) if and only if the subset $A = \{x_n \mid n \in \mathbb{N}\} \subset X$ is bounded (in the sense of Lecture V).

Hint: *Recall Example V.16 and Exercise V.5.*

Exercise VIII.2 (Convergent sequences are bounded).

Show that if $(x_n)_{n\in\mathbb{N}}$ is a sequence in X which is convergent, then $(x_n)_{n\in\mathbb{N}}$ is bounded.

Exercise (✓) VIII.3 (A sequence on the real line with 0/1-metric).

Consider the metric space $(\mathbb{R}, \mathsf{d}_{0/1})$, where the metric on the real line $\mathbb{R}$ is taken to be the discrete 0/1-metric $\mathsf{d}_{0/1}$. Show that the sequence $(x_n)_{n\in\mathbb{N}}$ in $(\mathbb{R}, \mathsf{d}_{0/1})$ defined by the formula $x_n = \frac{1}{n}$, for $n \in \mathbb{N}$, does not have $0 \in \mathbb{R}$ as its limit.

Example VIII.6 (A sequence converging to the zero function in L^1)**.**
Consider the space $\mathcal{C}([0,1])$ of all continuous functions $f\colon [0,1] \to \mathbb{R}$ with the L^1-norm $\|f\|_1 = \int_0^1 |f(t)|\,\mathrm{d}t$ and the metric induced by it.

For $n \in \mathbb{N}$, define the element $f_n \in \mathcal{C}([0,1])$ by

$$f_n(t) \;=\; \begin{cases} 1 - nt & \text{for } 0 \le t < \frac{1}{n} \\ 0 & \text{for } \frac{1}{n} \le t \le 1 \end{cases}$$

(the continuity of f_n is straightforward to check).

We claim that the limit of the sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{C}([0,1])$ is the zero function $\underline{0} \in \mathcal{C}([0,1])$. This is done by an easy calculation:

$$\mathsf{d}(\underline{0}, f_n) \;=\; \|f_n - \underline{0}\|_1 \;=\; \int_0^1 \big|f(t) - 0\big|\,\mathrm{d}t \;=\; \int_0^{1/n} (1 - nt)\,\mathrm{d}t \;=\; \frac{1}{2n}.$$

Since the right hand side tends to 0 (in $\mathbb{R}$) as $n \to \infty$, we get $\lim_{n\to\infty} \mathsf{d}(f_n, \underline{0}) = 0$. It then follows from Theorem VIII.4 that the sequence $(f_n)_{n\in\mathbb{N}}$ converges to the zero function $\underline{0} \in \mathcal{C}([0,1])$.

VIII.2. Characterization of closure and closed sets

Sequences provide intuitive and commonly used characterizations of closure and closed sets.

If $A \subset X$, we say that $(a_n)_{n\in\mathbb{N}}$ is a **sequence in** A if $a_n \in A$ for every $n \in \mathbb{N}$.

The following characterizations of closed sets and closure are very commonly used, both in calculus and in more advanced mathematics.

Theorem VIII.7 (Characterization of closed sets by sequences)**.**
A subset $A \subset X$ is closed if and only if for every sequence $(a_n)_{n\in\mathbb{N}}$ in A which is convergent in X, the limit is in $A \subset X$.

Proof. We prove "if" and "only if" separately.

proof of "if": We will prove this by contrapositive. Assume that $A \subset X$ is *not* closed, i.e., that its complement $X \setminus A$ is not open. By Definition VII.1 this means that for some point $x \in X \setminus A$ of the complement, there does not exist any $r > 0$ such that the ball $\mathcal{B}_r(x)$ would be contained in the complement $X \setminus A$. In particular, for $n \in \mathbb{N}$, using $r = \frac{1}{n}$ as the radius, we find that $\mathcal{B}_{1/n}(x) \not\subset X \setminus A$, that is, $\mathcal{B}_{1/n}(x) \cap A \neq \emptyset$. This means that for every $n \in \mathbb{N}$ we can choose a point $a_n \in \mathcal{B}_{1/n}(x) \cap A$, i.e, a point $a_n \in A$ such that $\mathsf{d}(x, a_n) < \frac{1}{n}$. Such points form a sequence $(a_n)_{n\in\mathbb{N}}$ in A. Also $0 \le \mathsf{d}(x, a_n) < \frac{1}{n}$ for all $n \in \mathbb{N}$, so by the squeeze theorem (Lemma II.17) we have $\lim_{n\to\infty} \mathsf{d}(x, a_n) = 0$, which by Theorem VIII.4 implies $\lim_{n\to\infty} a_n = x$. But $x \in X \setminus A$, so we have found a sequence $(a_n)_{n\in\mathbb{N}}$ in A whose limit x is not in A. By contrapositive we have proven the "if" implication.

proof of "only if": We will also prove this by contrapositive. Assume that there exists a sequence $(a_n)_{n\in\mathbb{N}}$ in A which is convergent in X, and its limit $x = \lim_{n\to\infty} a_n \in X$ is not in A, i.e., $x \in X \setminus A$. Now for any $r > 0$, since $\mathcal{B}_r(x)$ is a neighborhood of x, by the characterization of limits in Theorem VIII.4 we have that all but finitely many members of the sequence $(a_n)_{n\in\mathbb{N}}$ lie in $\mathcal{B}_r(x)$. In particular there exists some such member $a_n \in \mathcal{B}_r(x)$, and since $a_n \in A$, we get $\mathcal{B}_r(x) \cap A \neq \emptyset$, i.e., $\mathcal{B}_r(x) \not\subset X \setminus A$. We have thus shown that no open ball centered at x is contained in the complement $X \setminus A$ of A, so the complement $X \setminus A$ is not open, and therefore A is not closed. By contrapositive we have proven the "only if" implication. $\square$

Recall that the closure $\overline{A} \subset X$ of a subset $A \subset X$ is defined as the complement $\overline{A} = X \setminus \mathrm{ext}(A)$ of its exterior, or equivalently as the union $\overline{A} = A \cup \partial A$ of A with its boundary ∂A. The following is another commonly used characterization of the closure.

Theorem VIII.8 (Characterization of closure by sequences)**.**
For a subset $A \subset X$ and a point $x \in X$, the following are equivalent:
(a) $x \in \overline{A}$*;*
(b) *there exists a sequence* $(a_n)_{n\in\mathbb{N}}$ *in* A *such that* $\lim_{n\to\infty} a_n = x$*.*

The proof is quite similar to that of Theorem VIII.7. We leave the details as an exercise.

Exercise VIII.4 (Proof of Theorem VIII.8)**.**
Prove Theorem VIII.8.
Hint: *First observe that $\overline{A} = X \setminus \mathrm{ext}(A)$. From this it is easy to see that $x \in \overline{A}$ if and only if for every $r > 0$ we have $\mathcal{B}_r(x) \cap A \neq \emptyset$.*

The following is an easy consequence of combining Theorems VIII.8 and VIII.7.

Corollary VIII.9 (A characterization of closed sets by closures)**.**
For a subset $A \subset X$, the following are equivalent:
(i) $A \subset X$ *is closed;*
(ii) $\overline{A} \subset A$*;*
(iii) $\overline{A} = A$*.*

VIII.3. Characterization of continuity

We next give a characterization of continuity of functions using sequences and limits. This characterization is convenient for many purposes, and in fact, for real-valued functions of a real variable, we took this characterizing property as the definition of continuity in Lecture III. The theorem below will now show that continuity of real functions as defined in Chapter III is indeed equivalent to (a special case of) the notion of continuity of functions between metric spaces defined in Lecture VI.

Theorem VIII.10 (Characterization of continuity at a point)**.**
Let (X, d_X) and (Y, d_Y) be two metric spaces, $f \colon X \to Y$ a function, and $x \in X$ a point. Then the following are equivalent:
(i) f *is continuous at* x*;*
(ii) *For any sequence* $(x_n)_{n\in\mathbb{N}}$ *in* X *with limit* $\lim_{n\to\infty} x_n = x$*, the sequence* $\big(f(x_n)\big)_{n\in\mathbb{N}}$ *in* Y *has the limit* $\lim_{n\to\infty} f(x_n) = f(x)$*.*

Proof. We prove implications in both directions separately.

proof of (i) ⇒ (ii): Assume that $f \colon X \to Y$ is continuous at $x \in X$. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X such that $\lim_{n\to\infty} x_n = x$. We must show that the sequence $\big(f(x_n)\big)_{n\in\mathbb{N}}$ in Y converges to $f(x)$. So let $\varepsilon > 0$. By continuity of f at x, there exists a $\delta > 0$ such that when

$\mathsf{d}_X(x, x') < \delta$, we have $\mathsf{d}_Y\big(f(x), f(x')\big) < \varepsilon$. Since $\delta > 0$, by definition of limits, there exists an $n' \in \mathbb{N}$ such that for $n \geq n'$ we have $\mathsf{d}_X(x, x_n) < \delta$. Combining these, we get that for $n \geq n'$ we have $\mathsf{d}_Y\big(f(x), f(x_n)\big) < \varepsilon$. Since $\varepsilon > 0$ was arbitrary, this shows that $\lim_{n\to\infty} f(x_n) = f(x)$, and the implication (i) $\Rightarrow$ (ii) is thus proven.

proof of (ii) $\Rightarrow$ (i): Let us prove this by contrapositive. Assume that $f: X \to Y$ is *not* continuous at x. Then for some $\varepsilon > 0$ there does not exist any $\delta > 0$ such that $f[\mathcal{B}_\delta(x)]$ would be contained in $\mathcal{B}_\varepsilon(f(x))$. In particular for each $n \in \mathbb{N}$, using $\delta = \frac{1}{n}$, we have that $f[\mathcal{B}_{1/n}(x)] \not\subset \mathcal{B}_\varepsilon(f(x))$, i.e., there exists some $x_n \in \mathcal{B}_{1/n}(x)$ such that $f(x_n) \notin \mathcal{B}_\varepsilon(f(x))$. Now for a sequence $(x_n)_{n\in\mathbb{N}}$ of such choices, we have $0 \leq \mathsf{d}_X(x, x_n) < \frac{1}{n}$, so by the squeeze theorem we get $\lim_{n\to\infty} \mathsf{d}_X(x, x_n) = 0$ and therefore $\lim_{n\to\infty} x_n = x$ (by Theorem VIII.4). On the other hand since $f(x_n) \notin \mathcal{B}_\varepsilon(f(x))$, we have $\mathsf{d}_Y\big(f(x), f(x_n)\big) \geq \varepsilon$ for all $n \in \mathbb{N}$. In particular the sequence $\big(f(x_n)\big)_{n\in\mathbb{N}}$ does not converge to $f(x)$. By contrapositive, the existence of such a sequence $(x_n)_{n\in\mathbb{N}}$ proves the implication (ii) $\Rightarrow$ (i). $\square$

Corollary VIII.11 (Characterization of continuity using limits)**.**
Let (X, d_X) and (Y, d_Y) be two metric spaces and $f: X \to Y$ a function. Then the following are equivalent:

(i) *$f: X \to Y$ is a continuous function.*
(ii) *For every sequence $(x_n)_{n\in\mathbb{N}}$ in X which converges in X, we have that the sequence $\big(f(x_n)\big)_{n\in\mathbb{N}}$ converges in Y to the limit*

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

Proof. Since the continuity of a function $f: X \to Y$ is defined as its continuity at every point x of its domain X (Definition VI.1), the assertion follows directly from Theorem VIII.10. $\square$

From Theorem VIII.10 and Corollary VIII.11 it is now clear that continuity of real-valued functions on subsets of the real line, as defined in Lecture III indeed coincides with the general metric space definition of continuity specialized to the case of the metric space $(\mathbb{R}, \mathsf{d}_\mathbb{R})$ and its subspaces with their inherited metrics.

Example VIII.12 (Evaluation is not continuous with respect to the L^1-norm)**.**
Consider the space $\mathcal{C}([0,1])$ of all continuous functions $f: [0,1] \to \mathbb{R}$ with the L^1-norm $\|f\|_1 = \int_0^1 |f(t)|\,\mathrm{d}t$ and the metric induced by it. We will show that the evaluation at $z \in [0,1]$,

$$\mathrm{ev}_z : \mathcal{C}([0,1]) \to \mathbb{R} \qquad\qquad \mathrm{ev}_z(f) = f(z) \quad \text{for } f \in \mathcal{C}([0,1]),$$

is not continuous in this metric space (compare with Example VI.9, where the $\|\cdot\|_\infty$-norm was used instead). We will specifically treat the case $z = 0$ below. Similar arguments could be used to show that evaluations at other points $z \in [0,1]$ are not continuous, either. For notational clarity we nevertheless write z instead of 0 below, in particular because we also use the zero function $\underline{0} \in \mathcal{C}([0,1])$ and its value $\underline{0}(z) = 0$ at $z = 0$.

In Example VIII.6 we exhibited a sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{C}([0,1])$ such that at $z = 0$ we have $f_n(z) = 1$ for every $n \in \mathbb{N}$, and $\lim_{n\to\infty} f_n = \underline{0} \in \mathcal{C}([0,1])$ (the limit of the sequence is the zero function $\underline{0}$). Applying ev_z to this sequence and taking the limit, we find

$$\lim_{n\to\infty} \mathrm{ev}_z(f_n) \;=\; \lim_{n\to\infty} f_n(z) \;=\; \lim_{n\to\infty} 1 \;=\; 1,$$

while on the other hand ev_z applied to the limit of the sequence gives

$$\mathrm{ev}_z\Big(\lim_{n\to\infty} f_n\Big) \;=\; \mathrm{ev}_z\big(\underline{0}\big) \;=\; \underline{0}(z) \;=\; 0.$$

The two calculations together show that $\lim_{n\to\infty} \mathrm{ev}_z(f_n) \neq \mathrm{ev}_z\big(\lim_{n\to\infty} f_n\big)$, which by Corollary VIII.11 shows that $\mathrm{ev}_z : \mathcal{C}([0,1]) \to \mathbb{R}$ is not continuous.

Exercise VIII.5 (Taking the derivative is not continuous with respect to the uniform norm).
Let
$$\mathcal{C}^1([-1,1]) = \Big\{ x\colon [-1,1] \to \mathbb{R} \;\Big|\; x' \text{ is continuous} \Big\}$$
be the set of all continuously differentiable functions on the interval $[-1,1]$. Interpret it as a subset $\mathcal{C}^1([-1,1]) \subset \mathcal{C}([-1,1])$, where the space $\mathcal{C}([-1,1])$ of continuous functions is equipped again with the metric induced by the norm $\|\cdot\|_\infty$. As in Exercise VI.14, consider the function $g\colon \mathcal{C}^1([-1,1]) \to \mathcal{C}([-1,1])$ given by
$$[g(x)](t) = x'(t), \qquad \text{for } x \in \mathcal{C}^1([-1,1]) \text{ and } t \in [-1,1].$$
Show that $g\colon \mathcal{C}^1([-1,1]) \to \mathcal{C}([-1,1])$ is not continuous.

Exercise VIII.6 (Coordinatewise convergence is not sufficient for convergence in ℓ^1).
Consider the space
$$\ell^1 \;=\; \left\{ x = (x_j)_{j\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \sum_{j=1}^{\infty} |x_j| < \infty \right\}$$
of absolutely summable real sequences as in Example IV.9. Recall from Exercise IV.12 that the formula $\|x\|_1 = \sum_{j=1}^{\infty} |x_j|$ for $x = (x_j)_{j\in\mathbb{N}} \in \ell^1$ defines a norm on ℓ^1. We equip ℓ^1 with the metric induced by the norm $\|\cdot\|_1$.

(a) Show that if a sequence $(x^{(n)})_{n\in\mathbb{N}}$ of elements $x^{(n)} = (x_j^{(n)})_{j\in\mathbb{N}} \in \ell^1$ converges in ℓ^1 to $x = (x_j)_{j\in\mathbb{N}}$, then for every $k \in \mathbb{N}$, the sequence $(x_k^{(n)})_{n\in\mathbb{N}}$ of the k:th coordinates of $x^{(n)}$'s converges to $\lim_{n\to\infty} x_k^{(n)} = x_k$ (limit in $\mathbb{R}$).
Hint: *You can start by showing that the k:th coordinate projection function $(x_j)_{j\in\mathbb{N}} \mapsto x_k$ is a 1-Lipschitz function $\ell^1 \to \mathbb{R}$.*

(b) For $n \in \mathbb{N}$ let $x^{(n)} = (x_j^{(n)})_{j\in\mathbb{N}} \in \ell^1$ be the element given by
$$x_j^{(n)} = \begin{cases} 1 & \text{if } j = n \\ 0 & \text{if } j \neq n. \end{cases}$$
Show that for any $k \in \mathbb{N}$ we have $\lim_{n\to\infty} x_k^{(n)} = 0$ but in the space $(\ell^1, \|\cdot\|_1)$ the sequence $(x^{(n)})_{n\in\mathbb{N}}$ does not converge.
Hint: *If the sequence would converge in ℓ^1, then part (a) together with the first calculation of (b) identifies the only possibility for a limit $x \in \ell^1$. Now show directly from the definition of limits that we do not have convergence to that candidate limit.*

Operations on continuous real-valued functions

Let X be a set. Pointwise operations on real-valued functions on X are defined in the same way as in Definition III.10. The **pointwise sum** of $f\colon X \to \mathbb{R}$ and $g\colon X \to \mathbb{R}$ is the function
$$f+g\colon\; X \to \mathbb{R} \qquad\qquad \big(f+g\big)(x) := f(x) + g(x) \quad \text{for } x \in X.$$
The **pointwise product** of $f\colon X \to \mathbb{R}$ and $g\colon X \to \mathbb{R}$ is the function
$$fg\colon\; X \to \mathbb{R} \qquad\qquad \big(fg\big)(x) := f(x)\, g(x) \quad \text{for } x \in X.$$
The **pointwise quotient** of $f\colon X \to \mathbb{R}$ and $g\colon X \to \mathbb{R}$ is the function
$$f/g\colon\; X' \to \mathbb{R} \qquad\qquad \big(f/g\big)(x) := \frac{f(x)}{g(x)} \quad \text{for } x \in X',$$
where $X' := \big\{ x \in X \;\big|\; g(x) \neq 0 \big\}$. Again, the pointwise quotient f/g is only defined on the subset $X' \subset X$ of the domain, where the function g is non-vanishing.

If the set X is equipped with a metric d, so that (X, d) becomes a metric space, then we may consider in particular *continuous* real-valued functions $X \to \mathbb{R}$ (the real line $\mathbb{R}$ is equipped with its standard metric). By the following theorem, continuity of functions is preserved under the pointwise operations above.

Theorem VIII.13 (Continuity preserving operations).
Let $f, g \colon X \to \mathbb{R}$ be two continuous real-valued functions on a metric space X. Then also

(i) *the pointwise sum function $f + g \colon X \to \mathbb{R}$ is continuous,*
(ii) *the pointwise product function $fg \colon X \to \mathbb{R}$ is continuous,*
(iii) *the pointwise quotient function $f/g \colon X' \to \mathbb{R}$ is continuous on the subset $X' = \big\{ x \in X \;\big|\; g(x) \neq 0 \big\}$.*

Exercise (✓) VIII.7 (Proof of Theorem VIII.13).
Convince yourself that to prove Theorem VIII.13, using the characterization of continuity by sequences, Corollary VIII.11, one can follow the arguments in the proof of Theorem III.11 *mutatis mutandis.*

Example VIII.14 (Polynomials and rational functions in several variables).
Consider the Euclidean plane $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. Recall from Exercise VI.9 that the coordinate projections $(x, y) \mapsto x = \mathrm{pr}_1(x, y)$ and $(x, y) \mapsto y = \mathrm{pr}_2(x, y)$ are continuous functions $\mathbb{R}^2 \to \mathbb{R}$. Applying pointwise products to these successively, we find that for any $i, j \in \mathbb{N}_0$ the monomial function

$$(x, y) \mapsto x^i y^j$$

is continuous $\mathbb{R}^2 \to \mathbb{R}$. Applying further pointwise products with a constant function $(x, y) \mapsto c_{i,j} \in \mathbb{R}$, and a pointwise sum to finitely many different such scalar multiples of monomials, we find that an arbitrary two-variable polynomial

$$p(x, y) \;=\; \sum_{i,j=0}^{m} c_{i,j}\, x^i y^j$$

is continuous $p \colon \mathbb{R}^2 \to \mathbb{R}$. Taking two polynomials $p \colon \mathbb{R}^2 \to \mathbb{R}$ and $q \colon \mathbb{R}^2 \to \mathbb{R}$, the rational function defined by

$$r(x, y) \;=\; \frac{p(x, y)}{q(x, y)}$$

is continuous on the subset $\big\{ (x, y) \in \mathbb{R}^2 \;\big|\; q(x, y) \neq 0 \big\}$ where the denominator is non-vanishing, as the pointwise quotient of the continuous polynomial functions.

Similarly on the d-dimensional Euclidean space $\mathbb{R}^d$, any polynomial

$$p(x_1, \ldots, x_d) \;=\; \sum_{j_1, \ldots, j_d = 0}^{m} c_{j_1, \ldots, j_d}\, x_1^{j_1} \cdots x_d^{j_d}$$

defines a continuous function $p \colon \mathbb{R}^d \to \mathbb{R}$, and a rational function

$$r(x_1, \ldots, x_d) \;=\; \frac{p(x_1, \ldots, x_d)}{q(x_1, \ldots, x_d)}$$

is continuous on the subset of $\mathbb{R}^d$ where its denominator polynomial q is non-vanishing.

VIII.4. Sequences in product spaces

The convergence of sequences in product spaces can be characterized in terms of their “coordinate sequences”. If $(z_n)_{n \in \mathbb{N}}$ is a sequence in the product space $X \times Y$

of two metric spaces X and Y, then we may, for $n \in \mathbb{N}$, write the member of the sequence as $z_n = (x_n, y_n)$, where

$$x_n = \mathrm{pr}_1(z_n) \in X \qquad \text{and} \qquad y_n = \mathrm{pr}_2(z_n) \in Y$$

are the projections. We thus obtain the "coordinate sequences" $(x_n)_{n\in\mathbb{N}}$ in X and $(y_n)_{n\in\mathbb{N}}$ in Y of the sequence $(z_n)_{n\in\mathbb{N}} = \big((x_n, y_n)\big)_{n\in\mathbb{N}}$. We will verify below that the convergence and limit of the sequence $(z_n)_{n\in\mathbb{N}}$ in the product space $X \times Y$ can be characterized in terms of the convergence and limits of $(x_n)_{n\in\mathbb{N}}$ in X and $(y_n)_{n\in\mathbb{N}}$ in Y. This characterization probably appears familiar from calculus courses.

Notice first that convergence does not depend on which of the topologically equivalent metrics we equip the product space with.

Corollary VIII.15 (Convergence of sequences only depends on the topology)**.**
If d_1 *and* d_2 *are two topologically equivalent metrics on* X*, then a sequence* $(x_n)_{n\in\mathbb{N}}$ *converges to* $x \in X$ *in the metric space* (X, d_1) *if and only if it converges to* $x \in X$ *in the metric space* (X, d_2)*.*

Proof. The characterization (iii) of limits in Theorem VIII.4 only uses the notion of neighborhoods. In view of that, it is clear that if two metrics determine the same open sets, the notion of limits of sequences for both metrics is the same. □

Theorem VIII.16 (Convergence of sequences in a product space)**.**
Let (X, d_X) *and* (Y, d_Y) *be two metric spaces, and* $X \times Y$ *their product space equipped with any metric* d *which is topologically equivalent to those in* (V.20)*.*

Consider a sequence $\big((x_n, y_n)\big)_{n\in\mathbb{N}}$ *in the product space* $X \times Y$*. Then the following are equivalent:*

(i) *The sequence* $\big((x_n, y_n)\big)_{n\in\mathbb{N}}$ *converges to* $(x, y) \in X \times Y$*.*
(ii) *The coordinate sequences* $(x_n)_{n\in\mathbb{N}}$ *and* $(y_n)_{n\in\mathbb{N}}$ *converge in* X *and* Y*, respectively, to limits*

$$\lim_{n\to\infty} x_n = x \in X \qquad \textit{and} \qquad \lim_{n\to\infty} y_n = y \in Y.$$

Proof. We must prove the implications in both directions.

proof of (i) ⇒ (ii): Assume that $\big((x_n, y_n)\big)_{n\in\mathbb{N}}$ converges to (x, y) in the product space $X \times Y$. The projections $\mathrm{pr}_1 \colon X \times Y \to X$ and $\mathrm{pr}_2 \colon X \times Y \to Y$ are continuous by Exercise VI.9. Therefore by Theorem VIII.10 in X we have

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \mathrm{pr}_1\big((x_n, y_n)\big) = \mathrm{pr}_1\Big(\lim_{n\to\infty} (x_n, y_n)\Big) = \mathrm{pr}_1\big((x, y)\big) = x.$$

Similarly we get $\lim_{n\to\infty} y_n = y$ in Y.

proof of (ii) ⇒ (i): Assume that $\lim_{n\to\infty} x_n = x \in X$ and $\lim_{n\to\infty} y_n = y \in Y$. We must prove that $\big((x_n, y_n)\big)_{n\in\mathbb{N}}$ converges to $(x, y) \in X \times Y$. Note that by Corollary VIII.15 this convergence does not depend on the specific choice of the metric among topologically equivalent ones, so we may assume that the metric d_1 of (V.21) is used.

Let $\varepsilon > 0$. Since the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ converge to x and y, respectively, there exists some $n', n'' \in \mathbb{N}$ such that

$$\mathsf{d}_X(x_n, x) < \frac{\varepsilon}{2} \qquad \text{for } n \geq n',$$
$$\mathsf{d}_Y(y_n, y) < \frac{\varepsilon}{2} \qquad \text{for } n \geq n''.$$

Then for $n \geq n_\varepsilon = \max\{n', n''\}$, we have

$$\mathsf{d}_1\big((x_n, y_n), (x, y)\big) = \underbrace{\mathsf{d}_X(x_n, x)}_{<\varepsilon/2} + \underbrace{\mathsf{d}_Y(y_n, y)}_{<\varepsilon/2} < \varepsilon.$$

This shows the convergence $\lim_{n\to\infty}(x_n, y_n) = (x, y)$ in $X \times Y$. $\square$

Lecture IX

Function sequences

Let X and Y be two sets. If for every index $n \in \mathbb{N}$ we are given a function

$$f_n \colon X \to Y,$$

then the sequence

$$(f_1, f_2, f_3, \ldots) = (f_n)_{n \in \mathbb{N}}$$

is called a **function sequence**. Such sequences are important in various contexts, for example:

- *Function series:* In analysis one often deals with series representations of functions (for example Fourier series, power series, ...). To give a meaning to such an infinite series of functions, one must begin by considering, for each $n \in \mathbb{N}$, the partial sum of the first n terms. The sequence of the partial sums is a function sequence.
- *Numerical analysis:* In practice one commonly performs mathematical tasks not exactly but numerically (perhaps on a computer), to a good (but not perfect) numerical accuracy. Often one is interested in the numerics of a function and the parameters of the numerical approximation can be adjusted. Then one can form a sequence of functions out of ever improving choices of the numerical accuracy parameters, and one thus gets a sequence of functions.
- *Statistics and stochastics:* A random variable is (according to its mathematically idealized definition) literally a function defined on the "sample space". In statistics and probability one often considers sequences of random variables: time series, sequences indexed by increasing sample size, ... Such sequences of random variables are function sequences by their very definition.
- *etc.*

We will discuss two different notions of convergence of function sequences: pointwise convergence and uniform convergence. Both of these refer to the convergence of the values of the functions in an appropriate sense. Therefore the codomain Y needs to be equipped with a metric. By contrast, on the domain X it is not necessary to have a metric or any other structure.

IX.1. Pointwise convergence and uniform convergence

Let X be a set, and let (Y, d_Y) be a metric space. We consider sequences $(f_n)_{n \in \mathbb{N}}$ of functions $f_n \colon X \to Y$. Observe that for any given $x \in X$, the sequence $\big(f_n(x)\big)_{n \in \mathbb{N}}$ of the values of these functions at the point x is a sequence in the metric space Y, so it is meaningful to speak of its convergence, as in Lecture VIII. The notion of pointwise convergence of function sequences is therefore extremely straightforward.

Definition IX.1 (Pointwise convergence of a function sequence)**.**
A sequence $(f_n)_{n\in\mathbb{N}}$ of functions $X \to Y$ **converges pointwise** to a function $f\colon X \to Y$, if

$$\text{for all } x \in X \text{ we have} \qquad \lim_{n\to\infty} f_n(x) = f(x) \in Y.$$

Alternatively, in the above situation, we may say that f_n **tends to** f **pointwise** as $n \to \infty$, or simply write that $f_n \to f$ pointwise as $n \to \infty$.

Pointwise convergence thus literally means that at every point of the domain, the values of the functions converge.[1] If we unravel the definition of convergence of the values, we can rewrite pointwise convergence as follows.

Remark IX.2 (Meaning of pointwise convergence as a logical statement)**.**
In view of Definition VIII.1, $f_n \to f$ pointwise as $n \to \infty$ means that

$$\forall x \in X : \quad \forall \varepsilon > 0 : \quad \exists n' \in \mathbb{N} : \quad \forall n \geq n' : \quad \mathsf{d}_Y\big(f_n(x), f(x)\big) < \varepsilon.$$

Using the characterization of limits from Theorem VIII.4, $f_n \to f$ pointwise as $n \to \infty$ is also equivalent to

$$\forall x \in X : \quad \lim_{n\to\infty} \mathsf{d}_Y\big(f_n(x), f(x)\big) = 0.$$

The other notion of convergence of function sequences that we examine here is uniform convergence.

Definition IX.3 (Uniform convergence of a function sequence)**.**
A sequence $(f_n)_{n\in\mathbb{N}}$ of functions $X \to Y$ **converges uniformly** to a function $f\colon X \to Y$, if we have

$$\lim_{n\to\infty} \sup\Big\{\mathsf{d}_Y\big(f_n(x), f(x)\big) \;\Big|\; x \in X\Big\} \;=\; 0.$$

In the above situation, we alternatively say that f_n **tends to** f **uniformly** as $n \to \infty$, or simply write that $f_n \to f$ uniformly as $n \to \infty$.

A common situation is that of real-valued functions, i.e., taking the codomain to be $Y = \mathbb{R}$ with its standard metric. In that special case the condition for uniform convergence reads

$$\lim_{n\to\infty} \sup\Big\{\big|f_n(x) - f(x)\big| \;\Big|\; x \in X\Big\} \;=\; 0.$$

Exercise IX.1 (Meaning of uniform convergence as a logical statement)**.**
Prove that the statement $f_n \to f$ uniformly as $n \to \infty$ is logically equivalent to

$$\forall \varepsilon > 0 : \quad \exists n' \in \mathbb{N} : \quad \forall x \in X : \quad \forall n \geq n' : \quad \mathsf{d}_Y\big(f_n(x), f(x)\big) < \varepsilon.$$

[1] A reader with an abstract mindset may ask whether this notion of convergence is in fact obtained by equipping the space Y^X of all functions $X \to Y$ with a topology. It indeed is! The appropriate choice is the general product topology. However, if X is uncountably infinite, then the topology does not come from any metric on Y^X. In this course we therefore don't emphasize the abstract topological setup for pointwise convergence, but we take a more pedestrian point of view.

Compare Remark IX.2 and Exercise IX.1, and note the difference!

In pointwise convergence (Remark IX.2), the n' is allowed to depend on the point $x \in X$ as well, whereas in uniform convergence (Exercise IX.1) the *same* choice of n' must work simultaneously for all $x \in X$ — i.e., a *uniform* choice can be made for the whole space X at once. Once one internalizes this key difference between the two definitions, the following becomes rather evident.

Lemma IX.4 (Uniform convergence implies pointwise convergence)**.**
If $f_n \to f$ uniformly as $n \to \infty$, then we also have that $f_n \to f$ pointwise as $n \to \infty$.

Proof. Suppose that $f_n \to f$ uniformly as $n \to \infty$. To prove pointwise convergence, first fix a point $x \in X$. We must then show that $\lim_{n\to\infty} f_n(x) = f(x)$.

For the distance $\mathsf{d}_Y\big(f_n(x), f(x)\big)$, observe the obvious inequalities

$$0 \;\leq\; \mathsf{d}_Y\big(f_n(x), f(x)\big) \;\leq\; \sup\Big\{\mathsf{d}_Y\big(f_n(x'), f(x')\big) \;\Big|\; x' \in X\Big\} \underset{n\to\infty}{\longrightarrow} 0,$$

where the right hand side tends to 0 as $n \to \infty$ by virtue of the assumed uniform convergence $f_n \to f$. From these inequalities and the squeeze theorem (Lemma II.17), we get that $\lim_{n\to\infty} \mathsf{d}_Y\big(f_n(x), f(x)\big) = 0$. By the characterization of limits in Theorem VIII.4, this implies $\lim_{n\to\infty} f_n(x) = f(x)$, as desired. Pointwise convergence $f_n \to f$ as $n \to \infty$ follows. □

In other words, Lemma IX.4 says that uniform convergence of a function sequence is "stronger" (a more stringent requirement) than pointwise convergence. One practical use of this is the following. When we want to show uniform convergence of a function sequence $(f_n)_{n\in\mathbb{N}}$ to some limit function f, this limit function must also be the pointwise limit function. Therefore in order to identify the only possible candidate for the limit function, we start simply by calculating the pointwise limits $\lim_{n\to\infty} f_n(x)$ at all points x. This idea will be employed, e.g., in Examples IX.5, IX.6, and IX.7, below.

Strictly speaking Lemma IX.4 only says that uniform convergence is "at least as strong" as pointwise convergence — to conclude that it is a "strictly stronger" notion, we should show that the converse implication does not hold. And indeed, the following example shows it doesn't.

Example IX.5 (An example of pointwise but not uniform convergence)**.**
In this example, the domain of the functions is taken to be the interval $X = [0, 1]$, and the codomain is taken to be the real axis $Y = \mathbb{R}$ with its standard metric.

For $n \in \mathbb{N}$, define the function

$$f_n \colon [0,1] \to \mathbb{R} \qquad \text{by} \qquad f_n(x) = x^n \quad \text{for } x \in [0,1],$$

see Figure IX.1. Let us verify that the sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise. Below we handle the points $x \in [0, 1]$ in two separate cases: $x = 1$ and $x < 1$.

Consider the case $x = 1$. The function values at this point are simply $f_n(1) = 1^n = 1$ for any $n \in \mathbb{N}$. Obviously this constant sequence of values converges:

$$\lim_{n\to\infty} f_n(1) \;=\; \lim_{n\to\infty} 1^n \;=\; \lim_{n\to\infty} 1 \;=\; 1.$$

A pointwise limit function f must therefore assume the value $f(1) = 1$ at the point $x = 1$.

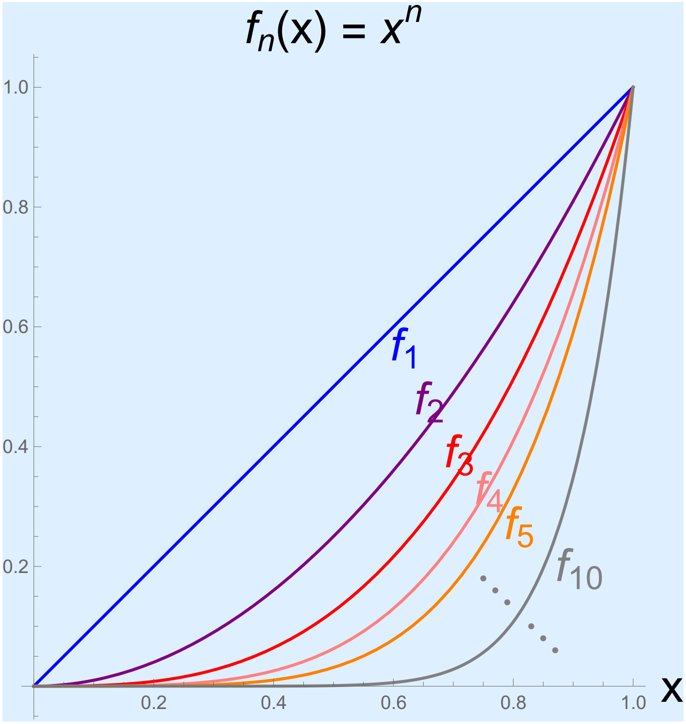


FIGURE IX.1. The functions $f_n\colon [0,1] \to \mathbb{R}$ given by $f_n(x) = x^n$.

Consider then the case $x \in [0, 1)$, i.e., $0 \le x < 1$. The function values at such a point are $f_n(x) = x^n$ for $n \in \mathbb{N}$, and the calculation of the limit is easy

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} x^n = 0.$$

A pointwise limit function f must therefore assume the value $f(x) = 0$ at any point $x \in [0, 1)$.

By the above, we have first of all narrowed down the only possible candidate for a pointwise limit function as the function

$$f\colon [0,1] \to \mathbb{R} \qquad \text{given by} \qquad f(x) = \begin{cases} 0 & \text{if } 0 \le x < 1 \\ 1 & \text{if } x = 1. \end{cases} \tag{IX.1}$$

By splitting to the two cases above, we have in fact verified that $\lim_{n\to\infty} f_n(x) = f(x)$ for any $x \in [0, 1]$, so indeed we have $f_n \to f$ pointwise as $n \to \infty$.

But does the function sequence $(f_n)_{n\in\mathbb{N}}$ converge uniformly? If it does converge uniformly to some function $\tilde{f}\colon [0,1] \to \mathbb{R}$, then by Lemma IX.4 it also converges pointwise to $\tilde{f}$. But since we showed that $(f_n)_{n\in\mathbb{N}}$ converges pointwise to the function f given in (IX.1), so in this case we must have[2] $\tilde{f} = f$. Therefore the remaining question is: do we have $f_n \to f$ uniformly as $n \to \infty$?

We will show that $(f_n)_{n\in\mathbb{N}}$ does *not* converge uniformly to f. According to Definition IX.3, we should show that the supremum

$$\sup\Big\{ \mathsf{d}_Y\big(f_n(x), f(x)\big) \;\Big|\; x \in [0,1] \Big\}$$

is not tending to zero as $n \to \infty$. To get a lower bound for the supremum for a given $n \in \mathbb{N}$, it is sufficient to find some point at which $\mathsf{d}_Y\big(f_n(x), f(x)\big)$ is not small. Consider for example the point

$$x_n = \sqrt[n]{1/2} = 2^{-1/n}.$$

[2]Here we are using the fact that a pointwise limit function is unique if it exists. This is an easy consequence of the uniqueness of limits of sequences in metric spaces (Lemma VIII.3), as the careful reader will quickly verify.

Clearly $0 < x_n < 1$, so the value of the pointwise limit function (IX.1) at this point is $f(x_n) = 0$. On the other hand, the value of the n:th member f_n of the function sequence at this point is $f_n(x_n) = (2^{-1/n})^n = 2^{-1} = 1/2$. We get[3]

$$\sup\Big\{\mathsf{d}_Y\big(f_n(x), f(x)\big)\;\Big|\; x \in [0,1]\Big\} \;\geq\; \mathsf{d}_Y\big(f_n(x_n), f(x_n)\big) \;=\; \big|\underbrace{f(x_n)}_{=0} - \underbrace{f_n(x_n)}_{=1/2}\big| \;=\; \frac{1}{2}.$$

Since this lower bound was obtained for an arbitrary $n \in \mathbb{N}$, the supremum here cannot be tending to 0 as $n \to \infty$. Therefore the sequence $(f_n)_{n\in\mathbb{N}}$ does not converge uniformly.

Example IX.6 (An example of uniform convergence)**.**
Consider then the following case, where the domain of the functions is taken to be the interval $X = [-\frac{6}{7}, \frac{6}{7}]$, and the codomain is taken to be the real axis $Y = \mathbb{R}$ with its standard metric.

For $n \in \mathbb{N}$, define the function

$$f_n \colon \Big[-\frac{6}{7}, \frac{6}{7}\Big] \to \mathbb{R} \qquad \text{by} \qquad f_n(x) = x^n \quad \text{for } x \in \Big[-\frac{6}{7}, \frac{6}{7}\Big].$$

The formula for f_n is the same as in Example IX.5, but the domain is different! In this case, we will verify that this sequence converges uniformly to the zero function $f \colon [-\frac{6}{7}, \frac{6}{7}] \to \mathbb{R}$, given by $f(x) = 0$ for all $x \in [-\frac{6}{7}, \frac{6}{7}]$.

First fix $n \in \mathbb{N}$. Then observe that for any $x \in [-\frac{6}{7}, \frac{6}{7}]$ we have the bound

$$\big|f_n(x) - f(x)\big| = \big|f_n(x) - 0\big| = |x^n| = |x|^n \leq \Big(\frac{6}{7}\Big)^n.$$

Since this holds for any x in the domain, we get

$$\sup\Big\{\big|f_n(x) - f(x)\big| \;\Big|\; x \in \Big[-\frac{6}{7}, \frac{6}{7}\Big]\Big\} \leq \Big(\frac{6}{7}\Big)^n$$

We have obtained

$$0 \;\leq\; \sup\Big\{\big|f_n(x) - f(x)\big| \;\Big|\; x \in \Big[-\frac{6}{7}, \frac{6}{7}\Big]\Big\} \;\leq\; \Big(\frac{6}{7}\Big)^n \underset{n\to\infty}{\longrightarrow} 0,$$

where the right hand side is tending to zero as $n \to \infty$, because $\frac{6}{7} < 1$. Applying the squeeze theorem (Lemma II.17), uniform convergence $f_n \to f$ as $n \to \infty$ follows. By Lemma IX.4 uniform convergence implies also pointwise convergence $f_n \to f$ as $n \to \infty$, to the same limit function f (the zero function).

Exercise IX.2 (Visualizing the uniform convergence in an example)**.**
Plot the fifteen or twenty first functions f_n in Example IX.6 to visualize the uniformly converging function sequence $(f_n)_{n\in\mathbb{N}}$. Compare with Example IX.5 and Figure IX.1.

Example IX.7 (An example of neither type of convergence)**.**
Consider then the following case, where the domain of the functions is taken to be the interval $X = [-1, 1]$, and the codomain is taken to be the real axis $Y = \mathbb{R}$ with its standard metric.

For $n \in \mathbb{N}$, define the function

$$f_n \colon [-1,1] \to \mathbb{R} \qquad \text{by} \qquad f_n(x) = x^n \quad \text{for } x \in [-1,1].$$

The formula for f_n is the same as in Examples IX.5 and IX.6, but the domain is again different! In this case, we will verify that this sequence converges neither pointwise nor uniformly.

Indeed, consider the point $x = -1$ in the domain $[-1, 1]$. The value of the n:th function at this point is $f_n(-1) = (-1)^n$. The sequence $\big(f_n(-1)\big)_{n\in\mathbb{N}} = \big((-1)^n\big)_{n\in\mathbb{N}}$ of values does not converge. Then we cannot have pointwise convergence $f_n \to f$ to any function f. And

[3]In fact a little bit more careful thinking would show that the supremum here is exactly 1, for any $n \in \mathbb{N}$. You could treat this improvement as an exercise.

by the contrapositive of Lemma IX.4, it is consequently also impossible to have uniform convergence.

Exercise IX.3 (A function sequence).
Define $f_n \colon [0,1] \to \mathbb{R}$ by setting

$$f_n(x) = n\,\sin(x/n), \qquad \text{for } x \in [0,1] \text{ and } n \in \mathbb{N}.$$

(a) Determine the limit function $f \colon [0,1] \to \mathbb{R}$ of the sequence $(f_n)_{n\in\mathbb{N}}$,

$$f(x) = \lim_{n\to\infty} f_n(x), \qquad x \in [0,1].$$

(b) Does the sequence $(f_n)_{n\in\mathbb{N}}$ converge uniformly on the interval $[0,1]$ to the function f?

Hint: *Either the* 1*:st order Taylor polynomial of the fucntion* sin *or L'Hospital's rule can be helpful.*

Exercise IX.4 (A different function sequence).
Define $f_n \colon \mathbb{R} \to \mathbb{R}$ by setting

$$f_n(x) = n\,\sin(x/n), \qquad \text{for } x \in \mathbb{R} \text{ and } n \in \mathbb{N}$$

(the formula is the same as in Exercise IX.3, but the domain is different!). Does this function sequence $(f_n)_{n\in\mathbb{N}}$ converge pointwise? Does it converge uniformly?

Exercise IX.5 (Yet another function sequence).
Define, for $n \in \mathbb{N}$, a function $f_n \colon [0,100] \to \mathbb{R}$ by setting

$$f_n(x) = n^2 - n^2\cos(x/n) \qquad \text{for } x \in [0,100].$$

(a) Determine the limit function $f \colon [0,100] \to \mathbb{R}$ of the sequence $(f_n)_{n\in\mathbb{N}}$,

$$f(x) = \lim_{n\to\infty} f_n(x) \qquad \text{for } x \in [0,100].$$

(b) Does the sequence (f_n) converge uniformly on the interval $[0,100]$ to the function f?

IX.2. Preservation of continuity in uniform limits

As a byproduct of Example IX.5 we see that even when each member f_n of a function sequence is a continuous function, the pointwise limit function f can be discontinuous. So, continuity is not in general preserved under pointwise convergence! The situation with uniform convergence is better, and this is one reason why uniform convergence is so useful.

Of course, to talk about continuity of the functions $X \to Y$, also the domain X needs to have a metric (or at least a topology). In the following we therefore assume that *both* (X, d_X) and (Y, d_Y) are metric spaces.

Theorem IX.8 (Uniform limit of continuous functions is continuous).
Suppose that $(f_n)_{n\in\mathbb{N}}$ is a sequence of continuous functions $f_n \colon X \to Y$ which converges uniformly to a limit function $f \colon X \to Y$. Then the limit function f is also continuous.

Proof. We assume the hypotheses: that each f_n is continuous, and that $f_n \to f$ uniformly.

We must prove the continuity of the limit function $f \colon X \to Y$. For this, let $z \in X$ and let $\varepsilon > 0$. We must find a $\delta > 0$ so that whenever $\mathsf{d}_X(z,x) < \delta$ we have $\mathsf{d}_Y\big(f(z), f(x)\big) < \varepsilon$.

By the assumed uniform convergence, since $\frac{\varepsilon}{4} > 0$, there exists an $n_0 \in \mathbb{N}$ such that

$$\sup\left\{\mathsf{d}_Y\big(f_n(x), f(x)\big) \;\middle|\; x \in X\right\} < \frac{\varepsilon}{4} \qquad \text{for all } n \geq n_0. \tag{IX.2}$$

Below we will in fact only need this for $n = n_0$.

Also since $f_{n_0}\colon X \to Y$ is (by assumption) continuous at z, and since $\frac{\varepsilon}{2} > 0$, there exists a $\delta > 0$ such that

$$\mathsf{d}_Y\big(f_{n_0}(z),\, f_{n_0}(x)\big) < \frac{\varepsilon}{2} \qquad \text{whenever } \mathsf{d}_X(z,x) < \delta. \tag{IX.3}$$

We claim that this δ works for us. To estimate the distance from $f(z)$ to $f(x)$ in Y, we use triangle inequality for the four points $f(z), f_{n_0}(z), f_{n_0}(x), f(x) \in Y$. For any $x \in X$ such that $\mathsf{d}_X(z,x) < \delta$ we thus get

$$\begin{aligned} &\mathsf{d}_Y\big(f(z), f(x)\big) \\ \leq\; &\underbrace{\mathsf{d}_Y\Big(f(z),\, f_{n_0}(z)\Big)}_{<\,\varepsilon/4 \text{ by (IX.2)}} + \underbrace{\mathsf{d}_Y\Big(f_{n_0}(z),\, f_{n_0}(x)\Big)}_{<\,\varepsilon/2 \text{ by (IX.3)}} + \underbrace{\mathsf{d}_Y\Big(f_{n_0}(x),\, f(x)\Big)}_{<\,\varepsilon/4 \text{ by (IX.2)}} \;<\; \frac{\varepsilon}{4} + \frac{\varepsilon}{2} + \frac{\varepsilon}{4} \;=\; \varepsilon. \end{aligned}$$

Since for an arbitrary $\varepsilon > 0$ we found a such $\delta > 0$, we obtain continuity of f at z. Since also $z \in X$ was arbitrary, this shows the continuity of the function $f\colon X \to Y$. □

IX.3. Function series

Let us now focus on the case of real-valued functions, i.e., taking the codomain to be $Y = \mathbb{R}$ (with its standard metric). The domain can be an arbitrary set X. The addition, multiplication, and division in $\mathbb{R}$ allow us to define the pointwise sums, products, and quotients of functions $X \to \mathbb{R}$ (for quotients we of course need the denominator to be non-zero).

Let $(g_k)_{k\in\mathbb{N}}$ be a sequence of functions $g_k\colon X \to \mathbb{R}$. Pointwise addition of the values allows us to define, for any $n \in \mathbb{N}$,

$$f_n\colon\; X \to \mathbb{R} \qquad\qquad f_n(x) \;=\; \sum_{k=1}^{n} g_k(x) \quad \text{for } x \in X \tag{IX.4}$$

This is the nth partial sum of the **function series** $\sum_{k=1}^{\infty} g_k$. We say that this function series **converges** to $F\colon X \to \mathbb{R}$ **pointwise** or **uniformly** if the sequence $(f_n)_{n\in\mathbb{N}}$ of its partial sums converges to F pointwise or uniformly, respectively.

The following result, known as **Weierstrass' M test**, is a practical criterion by which one can often verify the uniform convergence of a function series, and to prove the continuity of the function represented by the series.

Exercise IX.6 (Weierstrass' M test)**.**

Let $I \subset \mathbb{R}$ be an interval and $g_k\colon I \to \mathbb{R}$ continuous functions, for $k \in \mathbb{N}$. Assume that there exists constants $M_k \geq 0$, $k \in \mathbb{N}$, such that the series $\sum_{k=1}^{\infty} M_k$ is convergent and that for each $k \in \mathbb{N}$ we have $|g_k(x)| \leq M_k$ for all $x \in I$. Prove that the infinite series

$$F(x) = \sum_{k=1}^{\infty} g_k(x),$$

defines a continuous function $F\colon I \to \mathbb{R}$.

Hint: *Check the uniform convergence of the partial sums and apply a general result, or perform a direct argument of the same kind. The continuity of the partial sums (finitely many terms) follows inductively from Theorem III.11 or Theorem VIII.13. The remainder of a convergent series tends to zero as the starting index increases.*

Power series

Power series are a very common form of function series. They play a particularly prominent role in complex analysis, but they are also used heavily in other branches of mathematics. Important basic results about them can be proven using the Weierstrass M-test, see Appendix E for details. Here we content ourselves to mentioning that this way we get for example proofs of continuity of the exponential function and trigonometric functions. Let us state this as a lemma here.

Lemma IX.9 (The exponential function is continuous)**.**
The series

$$e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k, \tag{IX.5}$$

converges for all $x \in \mathbb{R}$, i.e., this function series converges pointwise on $\mathbb{R}$. Moreover, it converges uniformly on any closed interval $[a, b] \subset \mathbb{R}$.

The series (IX.5) *defines a continuous function $\mathbb{R} \to \mathbb{R}$ called the* **exponential function**.

Proof. See Appendix E. □

Lemma IX.10 (The trigonometric functions sin and cos are continuous)**.**
The two series

$$\sin(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1} \tag{IX.6}$$

$$\cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}. \tag{IX.7}$$

converge for all $x \in \mathbb{R}$, i.e., the function series converge pointwise on $\mathbb{R}$. Moreover, they converge uniformly on any closed interval $[a, b] \subset \mathbb{R}$.

Both series (IX.6) *and* (IX.7) *define continuous functions $\mathbb{R} \to \mathbb{R}$.*

Proof. See Appendix E. □

Fourier series

Here is a typical application of the Weierstrass' M-test.

Exercise IX.7 (A Fourier series application of the Weierstrass M-test)**.**
Using Exercise IX.6, prove that the Fourier series

$$\sum_{k=1}^{\infty} \frac{4\,(-1)^k \cos(kx)}{k^2}$$

defines a continuous function on $\mathbb{R}$.
Remark: Apart from an additive constant, this is the Fourier series of the 2π-periodic extension of $f(x) = x^2$, $-\pi \leq x \leq \pi$, see Figure IX.2. This can be used to calculate the sum of the series $\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2}$ (by first determining the additive constant and then substituting $x = 0$).

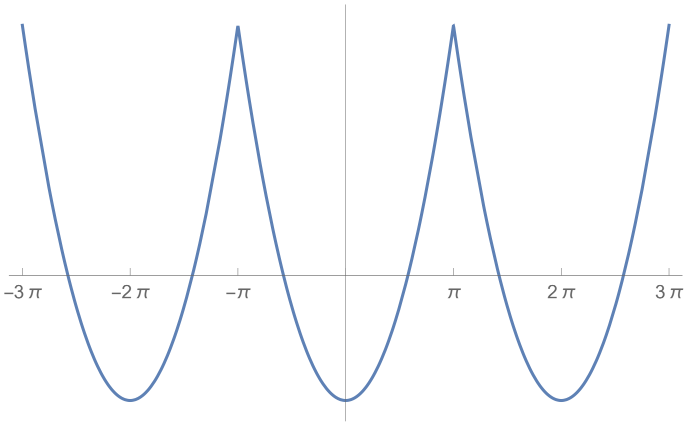


FIGURE IX.2. The continuous function represented by the Fourier series $\sum_{k=1}^{\infty} \frac{4\,(-1)^k \cos(kx)}{k^2}$ of Exercise IX.7 is a 2π-periodic extension of a parabola.

♡ *The Weierstrass function*

The function $f: \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \sum_{k=1}^{\infty} 2^{-k} \sin(4^k x) \qquad \text{for } x \in \mathbb{R}$$

is known as the Weierstrass function, see Figure IX.3 for a plot. Since $|\sin(\theta)| \leq 1$ for any $\theta \in \mathbb{R}$, it is easy to apply Weierstrass' M-test (Exercise IX.6) with the convergent geometric series $\sum_{k=1}^{\infty} 2^{-k} = 1 < \infty$ to show that the series is uniformly convergent and determines a continuous function $f: \mathbb{R} \to \mathbb{R}$. Furthermore, it is possible to show that the function is nowhere differentiable, i.e., it does not have a derivative $f'(x)$ at any point $x \in \mathbb{R}$.[4]

Exercise (✓) IX.8 (Continuity of the Weierstrass function)**.**
Apply Weierstrass' M-test (Exercise IX.6) to show that the Weierstrass function given by $f(x) = \sum_{k=1}^{\infty} 2^{-k} \sin(4^k x)$ is continuous $\mathbb{R} \to \mathbb{R}$.

Exercise IX.9 (Non-differentiability of the Weierstrass function at the origin)**.**
Consider the Weierstrass function given by $f(x) = \sum_{k=1}^{\infty} 2^{-k} \sin(4^k x)$ for $x \in \mathbb{R}$.

(a) Calculate $f(0)$.
(b) For $n \in \mathbb{N}$, let $x_n = \pi\, 4^{-n}$. Show that $f(x_n) \geq \frac{1}{\sqrt{2}}\, 2^{-n}$.
(c) By considering the difference quotients $\frac{f(x_n)-f(0)}{x_n-0}$, show that f has no derivative at 0.

Exercise (♯) IX.10 (Nowhere differentiability of the Weierstrass function)**.**
Prove that the Weierstrass function is not differentiable at any point $x \in \mathbb{R}$.

[4]Besides just looking at the plot in Figure IX.3, you can start to see why the function might be non-differentiable by observing the following. If you differentiate the series term by term, you formally obtain $f'(x) = \sum_{k=1}^{\infty} 2^k \cos(4^k x)$. This series does not converge at any point $x \in \mathbb{R}$ (the terms do not even tend to zero). While this is not yet a proof of non-differentiability, it gives a fair idea of the underlying problem. Exercise IX.9 indicates how you can get started with the precise proof of non-differentiability.

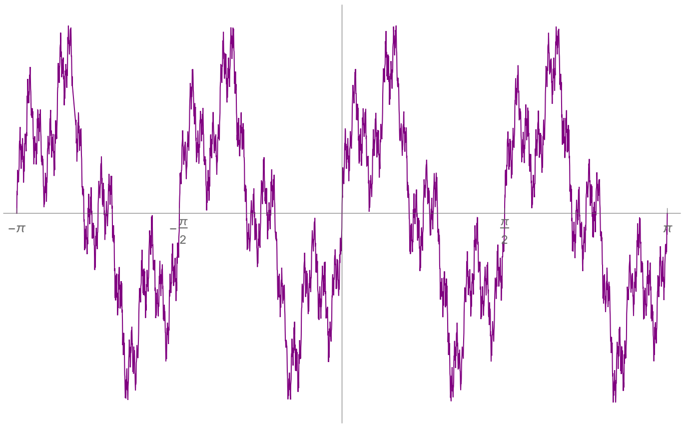


FIGURE IX.3. The continuous function represented by the Fourier series $\sum_{k=1}^{\infty} 2^{-k}\sin(4^k x)$. This function is continuous, but nowhere differentiable.

IX.4. ♡ Uniform convergence and integration

Another major advantage of uniform convergence is that it commutes with integration in the following sense.

Theorem IX.11 (Integrals of uniform limits)**.**
Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of Riemann-integrable functions $f_n\colon [a,b]\to\mathbb{R}$ which converges uniformly to a Riemann-integrable function $f\colon [a,b]\to\mathbb{R}$. Then we have

$$\lim_{n\to\infty}\int_a^b f_n(x)\,\mathrm{d}x \;=\; \int_a^b \big(\lim_{n\to\infty} f_n(x)\big)\,\mathrm{d}x.$$

Sketch of proof. Denote

$$s_n \;=\; \sup_{x\in[a,b]} \big|f_n(x)-f(x)\big|.$$

By uniform convergence $f_n\to f$, we have $s_n\to 0$ as $n\to\infty$. For any $x\in[a,b]$ and $n\in\mathbb{N}$, with the triangle inequality we get

$$f(x)-s_n \;\le\; f_n(x) \;\le\; f(x)+s_n.$$

Integrating over x and using monotonicity of integrals, we then get

$$\int_a^b f(x)\,\mathrm{d}x \;-\; (b-a)\,s_n \;\le\; \int_a^b f_n(x)\,\mathrm{d}x \;\le\; \int_a^b f(x)\,\mathrm{d}x \;+\; (b-a)\,s_n.$$

Here $(b-a)\,s_n\to 0$ as $n\to\infty$, so by the squeeze theorem we obtain

$$\lim_{n\to\infty}\int_a^b f_n(x)\,\mathrm{d}x \;=\; \int_a^b f(x)\,\mathrm{d}x,$$

which is the asserted result. □

The same does not hold for pointwise convergence.

Exercise IX.11 (A function sequence)**.**
Consider the sequence $(f_n)_{n\in\mathbb{N}}$ of functions $f_n\colon [0,\infty)\to\mathbb{R}$ given by

$$f_n(x)=\begin{cases} x-n+1, & \text{for } n-1\le x\le n,\\ n+1-x, & \text{for } n\le x\le n+1,\\ 0, & \text{otherwise.}\end{cases}$$

(a) Sketch the graphs of a few of the functions. Determine the pointwise limit function f of the function sequence $(f_n)_{n\in\mathbb{N}}$. (After guessing the limit with the help of the sketches, prove pointwise convergence from the definition).

(b) Does the function sequence $(f_n)_{n\in\mathbb{N}}$ converge uniformly on $[0,\infty)$ to the limit function f?

(c) Do we have

$$\lim_{n\to\infty}\int_0^\infty f_n(x)\,\mathrm{d}x \;=\; \int_0^\infty f(x)\,\mathrm{d}x\ ?$$

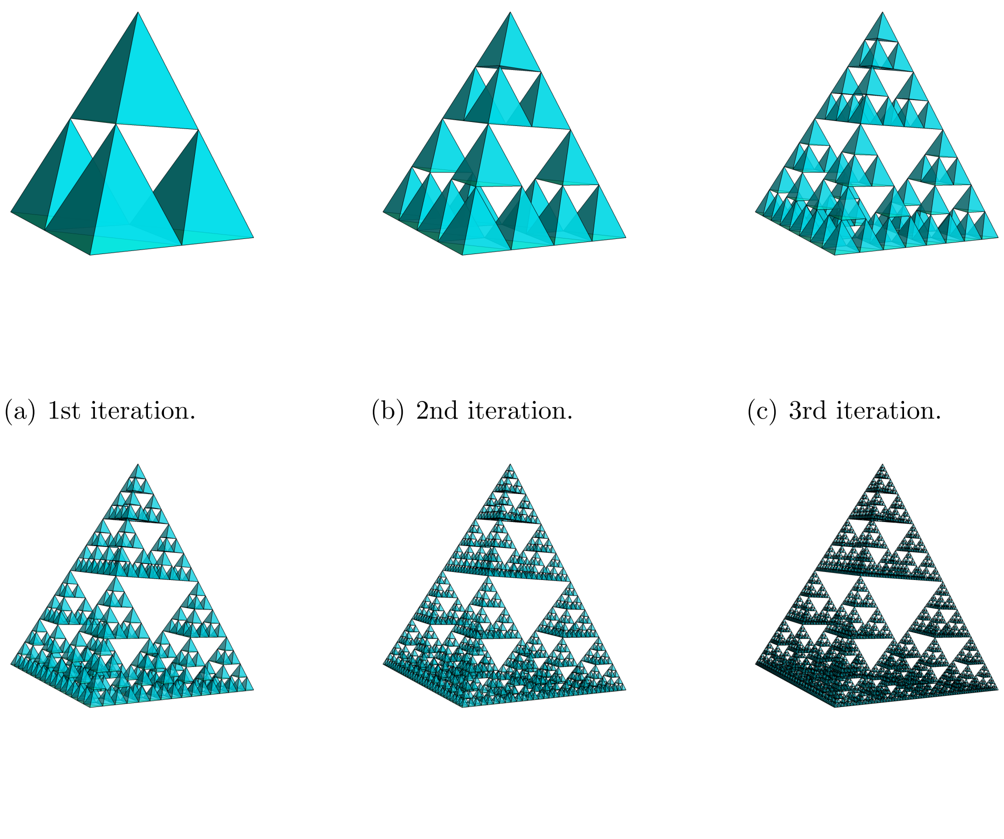
(d) By a minor modification one obtains a similar sequence $(\tilde{f}_n)_{n\in\mathbb{N}}$ of functions which has the same limit function f, but the integrals of $\tilde{f}_n$ tend to $+\infty$ as $n\to\infty$. Find such a modification (an illustration suffices).

Lecture X

Completeness

In this chapter we introduce a property of metric spaces called *completeness*. Not every metric space is complete, but many of the most important ones are. We will prove that the real line $\mathbb{R}$ and the Euclidean spaces $\mathbb{R}^d$ are complete, as is the space $\mathcal{C}([a,b])$ of continuous functions on a closed interval, when it is equipped with the metric induced by the sup-norm.

An intuitive way of thinking about completeness is that no points are "missing" from a complete metric space; that any sequence that "looks" convergent does indeed have a limit. Of course precise definitions are needed to make sense of this hand-wavy intuition. Nevertheless, the intuition already suggests that metric spaces which are complete are in important ways better behaved than those that are not, and the better behavior makes them more prominent in applications.



(a) 1st iteration. (b) 2nd iteration. (c) 3rd iteration.

(d) 4th iteration. (e) 5th iteration. (f) 6th iteration.

FIGURE X.1. Iterations that converge to a Sierpinski tetrahedron; a fixed point of a certain contractive self-map of a suitable complete metric space of subsets of $\mathbb{R}^3$.

Although we will only provide a handful of examples of complete spaces here, a plethora of further examples are used in (ordinary and) partial differential equations, probability, number theory, quantum physics, and many other fields. Especially widely used are inner product spaces and normed spaces which are complete in their natural metrics; such spaces are called Hilbert spaces and Banach spaces, respectively.

We will give one fundamental and powerful theorem about complete metric spaces: the contraction mapping principle, also known as Banach's fixed point theorem. The applications of this theorem alone are already vast; it gives the existence and uniqueness of solutions to many types of problems, and in addition provides a straightforward iterative scheme of obtaining rapidly improving approximations to the exact solution. We exemplify its uses by solving both numerical equations and differential equations. Figure X.1 illustrates a use of the contraction mapping principle to the construction of a fractal.[1]

X.1. Cauchy sequences

Let (X, d) be a metric space.

Recall from Definition VIII.1 that a sequence $(x_n)_{n\in\mathbb{N}}$ in X is said to converge to a limit $x \in X$ if for all $\varepsilon > 0$ there exists an $n_\varepsilon \in \mathbb{N}$ such that for $n \geq n_\varepsilon$ we have $\mathsf{d}(x_n, x) < \varepsilon$. Now note that according to this definition, *before* we can check this criterion of the convergence of the sequence $(x_n)_{n\in\mathbb{N}}$, we *already* have to know what the limit $x \in X$ should be! Without a doubt, it would often be more convenient if we could decide about the convergence of the sequence by looking only at the members of the sequence themselves. Such a criterion is provided by Cauchy sequences, and in favorable situations it turns out that the criterion already suffices to answer questions about convergence.

Definition X.1 (Cauchy sequence).
A sequence $(x_n)_{n\in\mathbb{N}}$ in a metric space (X, d) is said to be **Cauchy** if for any $\varepsilon > 0$ there exists an $n_\varepsilon \in \mathbb{N}$ such that

$$\mathsf{d}(x_k, x_\ell) < \varepsilon \qquad \text{for all } k, l \geq n_\varepsilon. \tag{X.1}$$

In other words, a sequence is Cauchy, if we can make the mutual distances among the members of the sequence arbitrarily small, by going far enough to the tail of the sequence.

Cauchy sequences and convergent sequences

Roughly, the idea of Definition X.1 is that a "Cauchy sequence looks (internally) like a convergent sequence". As a first observation in support of such an intuitive

[1]The Cantor set of Appendix B.3 can be viewed as a simpler similar example, and you will have no trouble finding a number of beautiful further example fractals illustrated online; e.g., `https://www.youtube.com/watch?v=dzymDsEUjKA`.

interpretation we show that any convergent sequence must be Cauchy (so "actually convergent sequences also look convergent internally").

Lemma X.2 (Every convergent sequence is Cauchy).
If a sequence $(x_n)_{n\in\mathbb{N}}$ in a metric space X is convergent, then it is Cauchy.

Proof. Suppose that $(x_n)_{n\in\mathbb{N}}$ converges in X, and denote its limit by $x = \lim_{n\to\infty} x_n \in X$. Let $\varepsilon > 0$. By the assumed convergence, there exists $n' \in \mathbb{N}$ such that

$$\mathsf{d}(x_n, x) < \frac{\varepsilon}{2} \qquad \text{for all } n \geq n'.$$

Then if $k, \ell \geq n'$, we have by triangle inequality

$$\mathsf{d}(x_k, x_\ell) \leq \underbrace{\mathsf{d}(x_k, x)}_{<\varepsilon/2} + \underbrace{\mathsf{d}(x, x_\ell)}_{<\varepsilon/2} < \varepsilon.$$

This shows that the sequence $(x_n)_{n\in\mathbb{N}}$ is Cauchy. □

The converse implication does not hold generally. Phrased intuitively again: there are metric spaces with sequences which "have the (internal) appearance of being convergent but which fail to actually converge within the space".

Example X.3 (Not every Cauchy sequence is convergent).
Consider the punctured real line $X = \mathbb{R} \setminus \{0\} \subset \mathbb{R}$ with the metric d that it inherits from the whole real line $\mathbb{R}$. Consider the sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R} \setminus \{0\}$ given by $x_n = \frac{1}{n}$ for $n \in \mathbb{N}$. For any $\varepsilon > 0$ we may set $n_\varepsilon = \lceil 1/\varepsilon \rceil$, and then for any $k, \ell \geq n_\varepsilon$ we have $\mathsf{d}(x_k, x_\ell) = \left|\frac{1}{\ell} - \frac{1}{k}\right| < \frac{1}{\min\{k,\ell\}} \leq \frac{1}{n_\varepsilon} \leq \varepsilon$. This shows that the sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R} \setminus \{0\}$ is Cauchy. However, it is easy to see that it does not converge in $\mathbb{R} \setminus \{0\}$ (why?).

In the above example it is (too) easy to see how the Cauchy sequence failed to converge because of a "missing" point: the space was defined by deliberately removing the point $0 \in \mathbb{R}$ from the space real line. In more complicated spaces which are not constructed as subspaces of some nice spaces, the phenomenon can be much harder to visualize, but this intuition of "missing points" is mostly good.

Cauchy sequences and bounded sequences

Above we saw that in general a Cauchy sequence may fail to be actually convergent, although it "appears (internally) convergent". Next we check that at least a Cauchy sequence must be bounded.

Lemma X.4 (Every Cauchy sequence is bounded).
If a sequence $(x_n)_{n\in\mathbb{N}}$ in a metric space X is Cauchy, then it is bounded.

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in X. To show that $(x_n)_{n\in\mathbb{N}}$ is bounded, we will find an $R > 0$ such that $x_n \in \mathcal{B}_R(x_1) \subset X$ for all $n \in \mathbb{N}$.

Let us fix $\varepsilon = 1$.[2] Using the Cauchy property (X.1) with this $\varepsilon = 1 > 0$, we get that there exists an $n_\varepsilon \in \mathbb{N}$ such that $\mathsf{d}(x_k, x_\ell) < \varepsilon$ for all $k, \ell \geq n_\varepsilon$. Among the finitely many first

[2] Any positive number ε would work here just as well — what is important is that one choice is fixed. For the convenience of the reader who prefers to choose $\varepsilon = 42$ or $\varepsilon = \frac{1}{361}$ or some other value instead, we keep referring to this *fixed* value simply by ε in the remaining proof (this should also help keep track of the original sources of the different terms).

indices $n \leq n_\varepsilon$, there is one which maximizes distance to x_1 — we now set

$$R' = \max\left\{ \mathsf{d}(x_1, x_n) \;\middle|\; n \leq n_\varepsilon \right\} \qquad \text{and} \qquad R = R' + \varepsilon.$$

Because of the maximum above, for any $n \leq n_\varepsilon$ we clearly have $\mathsf{d}(x_1, x_n) \leq R' < R$, so $x_n \in \mathcal{B}_R(x_1)$. It remains to consider $n > n_\varepsilon$. But for $n > n_\varepsilon$ we may use the triangle inequality and the choices of n_ε and R' to get

$$\mathsf{d}(x_1, x_n) \leq \underbrace{\mathsf{d}(x_1, x_{n_\varepsilon})}_{\leq R'} + \underbrace{\mathsf{d}(x_{n_\varepsilon}, x_n)}_{<\varepsilon} < R' + \varepsilon = R$$

We have thus showed that $x_n \in \mathcal{B}_R(x_1)$ for all $n \in \mathbb{N}$, so that the sequence $(x_n)_{n\in\mathbb{N}}$ is indeed bounded (recall Lemma V.17). □

In this case it is hardly surprising that the converse implication does not hold.

Exercise X.1 (Not every bounded sequence is Cauchy)**.**
Find an example of a sequence in some metric space, which is bounded but is not Cauchy.

Logical relationships among three different properties of sequences

We may summarize general logical implications (and lack thereof) between three properties of sequences in metric spaces by the following diagram:

$$\begin{tikzcd}
\text{convergent} \arrow[r, Rightarrow, bend left, "\text{Lemma X.2}"] & \text{Cauchy} \arrow[r, Rightarrow, bend left, "\text{Lemma X.4}"] \arrow[l, dotted, bend left, "/" description, "\text{Example X.3}"] & \text{bounded} \arrow[l, dotted, bend left, "/" description, "\text{Exercise X.1}"]
\end{tikzcd} .$$

In particular, Lemmas X.2 and X.4 combined give another proof of the boundedness of convergent sequences (in Exercise VIII.2 you most likely had a more direct proof).

A criterion for the convergence of a Cauchy sequence

A Cauchy sequence "appears convergent", and it turns out that if just any of its subsequences has a limit, then the whole sequence is in fact convergent.

Lemma X.5 (Cauchy sequences with convergent subsequences are convergent)**.**
Let $(x_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in X. If it has some subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ which converges, then the whole sequence $(x_n)_{n\in\mathbb{N}}$ also converges to the same limit,

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} x_{\varphi(n)}.$$

Proof. Suppose that $(x_n)_{n\in\mathbb{N}}$ is Cauchy and $(x_{\varphi(n)})_{n\in\mathbb{N}}$ is a convergent subsequence. Denote the limit of the subsequence by $\xi = \lim_{n\to\infty} x_{\varphi(n)} \in X$.

We must prove that $(x_n)_{n\in\mathbb{N}}$ also converges to ξ. So let $\varepsilon > 0$. Since $\varepsilon/2 > 0$, by the Cauchy property of $(x_n)_{n\in\mathbb{N}}$, we know that there exists some $n' \in \mathbb{N}$ such that

$$\mathsf{d}(x_k, x_\ell) < \frac{\varepsilon}{2} \qquad \text{for } k, \ell \geq n'. \tag{X.2}$$

On the other hand, by the convergence of the subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$, we know that there exists an $m' \in \mathbb{N}$ such that

$$\mathsf{d}(x_{\varphi(m)}, \xi) < \frac{\varepsilon}{2} \qquad \text{for } m \geq m'. \tag{X.3}$$

Let us select some $m_0 \geq m'$ such that also $\varphi(m_0) \geq n'$ (possible, since there are arbitrarily large subsequence indices), and denote $n_0 = \varphi(m_0)$. With this choice we have $\mathsf{d}(x_{n_0}, \xi) < \varepsilon/2$, by (X.3) and the observations $n_0 = \varphi(m_0)$ and $m_0 \geq m'$. Also since $n_0 = \varphi(m_0) \geq n'$, by (X.2) we have for any $n \geq n'$ that $\mathsf{d}(x_n, x_{n_0}) < \varepsilon/2$. By triangle inequality, for $n \geq n'$ we then get

$$\mathsf{d}(x_n, \xi) \;\leq\; \underbrace{\mathsf{d}(x_n, x_{n_0})}_{<\varepsilon/2} + \underbrace{\mathsf{d}(x_{n_0}, \xi)}_{<\varepsilon/2} \;<\; \varepsilon.$$

This shows the convergence of $(x_n)_{n\in\mathbb{N}}$ to ξ. □

Preservation of Cauchy sequences under Lipschitz functions

We next record the simple but useful observation that Lipschitz images of Cauchy sequences are Cauchy.

Lemma X.6 (Lipschitz maps preserve the Cauchy property)**.**
Let (X, d_X) and (Y, d_Y) be two metric spaces and $f\colon X \to Y$ a Lipschitz function. Then if $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in X, we have that $\big(f(x_n)\big)_{n\in\mathbb{N}}$ is a Cauchy sequence in Y.

Proof. Let $K > 0$ be a Lipschitz constant for the Lipschitz function $f\colon X \to Y$. Suppose that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in X, and consider the image sequence $\big(f(x_n)\big)_{n\in\mathbb{N}}$ in Y.

Let $\varepsilon > 0$. Since $(x_n)_{n\in\mathbb{N}}$ is Cauchy and $\frac{\varepsilon}{K} > 0$, there exists an $n' \in \mathbb{N}$ such that $\mathsf{d}_X(x_k, x_\ell) < \frac{\varepsilon}{K}$ whenever $k, \ell \geq n'$. By the K-Lipchitz property, we then get for $k, \ell \geq n'$ that

$$\mathsf{d}_Y\big(f(x_k), f(x_\ell)\big) \;\leq\; K\, \underbrace{\mathsf{d}_X(x_k, x_\ell)}_{<\varepsilon/K} \;<\; K\,\frac{\varepsilon}{K} \;=\; \varepsilon.$$

This shows that $\big(f(x_n)\big)_{n\in\mathbb{N}}$ is Cauchy. □

Let us remark that a similar result holds under the milder assumption that f is uniformly continuous (uniform continuity will be defined in Lecture XI, Lipschitz functions and Hölder functions are special cases of uniformly continuous functions). The easy proof of this generalization could be treated as an exercise (after a peek at Definition XI.24).

X.2. Completeness

Definition of completeness

The idea of a space with “no missing points” is captured by the following definition.

Definition X.7 (Complete metric space)**.**
A metric space X is **complete** if every Cauchy sequence in X converges in X.

We can readily give an example of a space that is *not* complete.

Example X.8 (The punctured real axis is not complete)**.**
In Example X.3 we saw that the punctured real axis $X = \mathbb{R} \setminus \{0\}$ (equipped with the metric inherited from the real axis) has a Cauchy sequence which is not convergent. Therefore $\mathbb{R} \setminus \{0\}$ is not a complete metric space.

A slightly more interesting counterexample is the following.

Example X.9 (The set of rational numbers is not complete)**.**
Consider the set $\mathbb{Q}$ of rational numbers, with the metric $\mathsf{d}_{\mathbb{Q}}$ it inherits from the real line as a subset $\mathbb{Q} \subset \mathbb{R}$, i.e., $\mathsf{d}_{\mathbb{Q}}(q_1, q_2) = |q_2 - q_1|$ for $q_1, q_2 \in \mathbb{Q}$. We claim that the metric space $(\mathbb{Q}, \mathsf{d}_{\mathbb{Q}})$ is not complete. To see this, if suffices to find a Cauchy sequence in $\mathbb{Q}$ which does not converge in $\mathbb{Q}$.

Let us form a sequence $(q_n)_{n\in\mathbb{N}}$ whose n:th member q_n is the truncated decimal expansion of $\sqrt{2}$ with n decimals included, i.e.,

$$q_1 = 1.4, \qquad q_2 = 1.41, \qquad q_3 = 1.414, \qquad q_4 = 1.4142, \qquad q_5 = 1.41421, \qquad \ldots,$$

or more formally $q_n = \lfloor 10^n \sqrt{2} \rfloor 10^{-n}$ for $n \in \mathbb{N}$. Each of these truncated decimal expansions is a rational number, $q_n \in \mathbb{Q}$, so $(q_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{Q}$. Let us check directly that this sequence is Cauchy.[3] For $k, \ell \geq n$ we have that q_k and q_ℓ may only differ in decimals after the n:th one, so as in Example II.33 we get $|q_k - q_\ell| \leq \sum_{j=n+1}^{\infty} 9 \times 10^{-j} = 10^{-n}$. Therefore for a given $\varepsilon > 0$ it suffices to set $n_\varepsilon = \lceil \log_{10}(2/\varepsilon) \rceil$ to ensure that for all $k, \ell \geq n_\varepsilon$ we have $|q_k - q_\ell| \leq \frac{\varepsilon}{2} < \varepsilon$.

Viewed as a sequence in $\mathbb{R}$, the sequence $(q_n)_{n\in\mathbb{N}}$ converges to $\sqrt{2}$. It is therefore impossible that $(q_n)_{n\in\mathbb{N}}$ converges in $\mathbb{Q} \subset \mathbb{R}$, because if it would, its limit $q \in \mathbb{Q}$ would by the uniqueness of limits (Lemma VIII.3) in $\mathbb{R}$ be $q = \sqrt{2} \notin \mathbb{Q}$; a contradiction. Therefore the Cauchy sequence $(q_n)_{n\in\mathbb{N}}$ in $\mathbb{Q}$ does not converge in $\mathbb{Q}$, and the metric space $(\mathbb{Q}, \mathsf{d}_{\mathbb{Q}})$ is not complete.

Completeness of the real line

The real axis is our simplest (and in some sense the most fundamental) example of a complete metric space. It is worth noting that the completeness of $\mathbb{R}$ as a metric space is in fact intimately related to the completeness axiom of $\mathbb{R}$ (think about which part of the proof below relies on the completeness axiom of $\mathbb{R}$).

Theorem X.10 (The real line is a complete metric space)**.**
The real line with its standard metric, i.e., $(\mathbb{R}, \mathsf{d}_{\mathbb{R}})$*, is a complete metric space.*

Proof. According to Definition X.7, we must show that every Cauchy sequence in $\mathbb{R}$ converges.

So let $(x_n)_{n\in\mathbb{N}}$ be a Cauchy sequence. By Lemma X.4, the sequence $(x_n)_{n\in\mathbb{N}}$ must then be bounded. By Theorem III.7, the bounded real-number sequence $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence. By Lemma X.5, the Cauchy sequence $(x_n)_{n\in\mathbb{N}}$ which has a convergent subsequence must itself be convergent. This finishes the proof. □

Let us issue one warning: completeness is *not* a purely topological property — it genuinely depends on the metric. The following exercise concretely shows that.

[3] As an alternative to this direct proof, it would also be possible to argue as follows. Since $\mathbb{Q} \subset \mathbb{R}$, the sequence $(q_n)_{n\in\mathbb{N}}$ may also be viewed as a sequence in $\mathbb{R}$. It is known to converge to $\sqrt{2}$ in $\mathbb{R}$ (recall Example II.33), so by Lemma X.2 it has to be a Cauchy sequence in $\mathbb{R}$. But since the metric $\mathsf{d}_{\mathbb{Q}}$ we are using in $\mathbb{Q}$ is the one inherited from $\mathbb{R}$, it is then also Cauchy in $\mathbb{Q}$.

Exercise X.2 (Completeness depends on the metric).
Consider the open interval $\left(\frac{-\pi}{2}, \frac{\pi}{2}\right) \subset \mathbb{R}$ with the metric it inherits from the real line. Recall that the function

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

is a homeomorphism, and therefore on $\mathbb{R}$ the metric defined by

$$\widetilde{\mathsf{d}}_{\mathbb{R}}(x,y) = \left|\tan^{-1}(y) - \tan^{-1}(x)\right| \qquad \text{for } x, y \in \mathbb{R}$$

is equivalent to the usual metric (i.e., it induces the same topology).

Show that the metric space $(\mathbb{R}, \widetilde{\mathsf{d}}_{\mathbb{R}})$ is not complete, although it is homeomorphic to the complete metric space $(\mathbb{R}, \mathsf{d}_{\mathbb{R}})$.

However, a homeomorphism which is bilipschitz does in fact preserve completeness.

Exercise X.3 (Completeness is preserved under bilipschitz-equivalence).
Suppose that (X, d_X) and (Y, d_Y) are two metric spaces and $f\colon X \to Y$ is a surjective bilipschitz map, see (VII.1). By Exercise VII.19, f is in particular a homeomorphism. Show that then (X, d_X) is complete if and only if (Y, d_Y) is complete.

Important complete spaces

Let us now look at a few other important complete metric spaces. Note that each of the proofs of completeness below ultimately relies on the completeness of $\mathbb{R}$.

Euclidean spaces are complete

Theorem X.11 (The Euclidean spaces $\mathbb{R}^d$ are complete).
For any $d \in \mathbb{N}$, the Euclidean space $\mathbb{R}^d$ equipped with its usual metric is a complete metric space.

Proof. Let us only prove the case $d = 2$ and leave the general case as an exercise.

Let $(v_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $\mathbb{R}^2$ (equipped with the Euclidean metric). For each member $v_n \in \mathbb{R}^2$ of the sequence, let us denote the projections to the two coordinates by $x_n = \mathrm{pr}_1(v_n) \in \mathbb{R}$ and $y_n = \mathrm{pr}_2(v_n) \in \mathbb{R}$, so that

$$v_n = (x_n, y_n) \in \mathbb{R}^2.$$

Then $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are sequences in $\mathbb{R}$. Since the projections $\mathrm{pr}_1, \mathrm{pr}_2\colon \mathbb{R}^2 \to \mathbb{R}$ are Lipschitz, it follows from Lemma X.6 that $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are Cauchy sequences.

Since $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are Cauchy sequences in the complete metric space $\mathbb{R}$, they converge to some limits

$$x = \lim_{n\to\infty} x_n, \qquad y = \lim_{n\to\infty} y_n.$$

By Theorem VIII.16, then, the original sequence $(v_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^2$ converges to the point $v = (x, y) \in \mathbb{R}$ whose coordinates are the above limits of the coordinate sequences. □

Exercise X.4 (The product of complete metric spaces is complete).
Modify the proof of Theorem X.11 to show the following: If (X, d_X) and (Y, d_Y) are two complete metric spaces, then the product space $X \times Y$ equipped with any of the metrics (V.20) is also a complete metric space.

Exercise X.5 (The remaining part of the proof of Theorem X.11).
Complete the proof of Theorem X.11, i.e., prove that $\mathbb{R}^d$ is complete for any $d > 2$, too.

A complete space of continuous functions

The following complete metric space is very important in many applications.

Theorem X.12 (A complete space of continuous functions)**.**
Consider the space

$$\mathcal{C}([a,b]) = \Big\{ f\colon [a,b] \to \mathbb{R} \;\Big|\; f \text{ is continuous} \Big\}$$

of continuous functions on a closed interval $[a,b] \subset \mathbb{R}$*, equipped with the metric* d_∞ *induced by the supremum-norm* $\|\cdot\|_\infty$ *(see Example IV.15). Then* $\big(\mathcal{C}([a,b]), \mathsf{d}_\infty\big)$ *is a complete metric space.*

Let us immediately point out that the conclusion of Theorem X.12 depends crucially on the fact that we equip $\mathcal{C}([a,b])$ with the supremum norm. We have seen also two other norms on $\mathcal{C}([a,b])$: the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ of Examples IV.11 and IV.20. The space $\mathcal{C}([a,b])$ equipped with the metric induced by either of these norms is *not* complete!

Exercise (♯) X.6 (Incomplete spaces of continuous functions)**.**

(a) Prove that the space $\mathcal{C}([a,b])$ equipped with the metric d_1 induced by the norm $\|\cdot\|_1$ of Example IV.11 is not complete.
(b) Prove that the space $\mathcal{C}([a,b])$ equipped with the metric d_2 induced by the norm $\|\cdot\|_2$ of Example IV.20 is not complete.

Proof of Theorem X.12. Let $(f_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $\mathcal{C}([a,b])$ equipped with the norm

$$\|f\|_\infty = \sup\big\{|f(x)| \;\big|\; x \in [a,b]\big\}$$

and the corresponding metric

$$\mathsf{d}_\infty(f,g) = \|g-f\|_\infty.$$

First recall from Example VI.9 that for any point $x \in [a,b]$, the evaluation mapping $f \mapsto f(x) = \mathrm{ev}_x(f)$ is 1-Lipschitz $\mathcal{C}([a,b]) \to \mathbb{R}$. Therefore by Lemma X.6, the real-number sequence $\big(\mathrm{ev}_x(f_n)\big)_{n\in\mathbb{N}} = \big(f_n(x)\big)_{n\in\mathbb{N}}$ is also Cauchy. By completeness of $\mathbb{R}$ (Theorem X.10), the Cauchy sequence $\big(f_n(x)\big)_{n\in\mathbb{N}}$ on $\mathbb{R}$ converges. Let us denote its limit by

$$f(x) := \lim_{n\to\infty} f_n(x).$$

The limits $f(x) \in \mathbb{R}$ thus obtained for different $x \in [a,b]$ define a function $f\colon [a,b] \to \mathbb{R}$. By construction, the sequence $(f_n)_{n\in\mathbb{N}}$ of functions converges (at least) pointwise to the function f. We will show that the convergence is in fact uniform.

Let $\varepsilon > 0$. By the Cauchy property of the sequence $(f_n)_{n\in\mathbb{N}}$, there exists an $n' \in \mathbb{N}$ such that

$$\|f_k - f_\ell\|_\infty < \frac{\varepsilon}{2} \qquad \text{whenever} \qquad k,\ell \geq n'.$$

This implies that for any $x \in [a,b]$ and any $k,\ell \geq n'$, we have

$$\big|f_k(x) - f_\ell(x)\big| \;\leq\; \|f_k - f_\ell\|_\infty \;<\; \frac{\varepsilon}{2}. \tag{X.4}$$

Now note that $z \mapsto |f_k(x) - z|$ is continuous $\mathbb{R} \to \mathbb{R}$ (in fact 1-Lipschitz) and we have $\lim_{\ell\to\infty} f_\ell(x) = f(x)$. Therefore we may fix $k \geq n'$ and take the limit as $\ell \to \infty$ in (X.4) and conclude (by preservation of bounds, Corollary II.15) for any $x \in [a,b]$ that

$$\big|f_k(x) - f(x)\big| \;=\; \Big|f_k(x) - \lim_{\ell\to\infty} f_\ell(x)\Big| \;=\; \lim_{\ell\to\infty} \big|f_k(x) - f_\ell(x)\big| \;\leq\; \frac{\varepsilon}{2}.$$

As this holds for all $x \in [a,b]$, we in fact get

$$\sup\Big\{ \big|f_k(x) - f(x)\big| \;\Big|\; x \in [a,b] \Big\} \;\leq\; \frac{\varepsilon}{2} \;<\; \varepsilon \tag{X.5}$$

when $k \geq n'$. This shows that the function sequence $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f. As a uniform limit of continuous functions, the function f is also continuous (Theorem IX.8), i.e., we have $f \in \mathcal{C}([a,b])$. Moreover, (X.5) can then be read as saying $\|f_k - f\|_\infty < \varepsilon$ when $k \geq n'$, so we conclude that $\|f_k - f\|_\infty \to 0$ as $k \to \infty$, and thus $\lim_{k\to\infty} f_k = f$ in the metric space $\big(\mathcal{C}([a,b]), \mathsf{d}_\infty\big)$. This proves that an arbitrary Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ in $\big(\mathcal{C}([a,b]), \mathsf{d}_\infty\big)$ converges, which is the asserted completeness property. □

Completeness for subspaces

In general, subsets of a complete metric space need not be complete. Indeed, the subsets $\mathbb{R} \setminus \{0\} \subset \mathbb{R}$ and $\mathbb{Q} \subset \mathbb{R}$ of the complete metric space $\mathbb{R}$ (see Theorem X.10) were seen to be not complete in Examples X.8 and X.9, respectively. The following result fully answers which subsets of complete metric spaces are complete as subspaces.

Theorem X.13 (Subspaces of a complete metric space).
Let (X, d) be a complete metric space. Then a subset $A \subset X$ is complete as a subspace[4] *if and only if $A \subset X$ is closed.*

Exercise X.7 (Proof of Theorem X.13).
Prove Theorem X.13.
Hint: *Recall Theorem VIII.7.*

Example X.14 (The set of rational numbers is not complete, again).
The set $\mathbb{Q} \subset \mathbb{R}$ of rational numbers is a subset of the metric space $\mathbb{R}$, which is complete (by Theorem X.10). Since $\mathbb{Q} \subset \mathbb{R}$ is not closed (recall Example VII.18), Theorem X.13 implies that $\mathbb{Q}$ is not complete (as we in fact already saw in Example X.9).

Example X.15 (The space of continuous functions vanishing at the endpoints of an interval).
Let $a < b$. Consider the space

$$X = \Big\{ f \in \mathcal{C}([a,b]) \;\Big|\; f(a) = 0,\; f(b) = 0 \Big\}$$

of those continuous functions $f\colon [a,b] \to \mathbb{R}$ which vanish at both endpoints a and b of the closed interval $[a,b]$. Equip it with the metric it inherits as a subset $X \subset \mathcal{C}([a,b])$, when on $\mathcal{C}([a,b])$ we use the metric induced by the sup-norm $\|\cdot\|_\infty$.

Using the evaluation functions $\mathrm{ev}_a\colon \mathcal{C}([a,b]) \to \mathbb{R}$ and $\mathrm{ev}_b\colon \mathcal{C}([a,b]) \to \mathbb{R}$ we can write X as the intersection of preimages

$$X \;=\; \mathrm{ev}_a^{-1}[\{0\}] \;\cap\; \mathrm{ev}_b^{-1}[\{0\}].$$

The singleton subset $\{0\} \subset \mathbb{R}$ is closed and the evaluation functions are continuous (Example VI.9), so both preimages above are closed subsets in $\mathcal{C}([a,b])$, and $X \subset \mathcal{C}([a,b])$ is closed as their intersection (Theorem VII.16).

Now since $\mathcal{C}([a,b])$ (with the metric induced by $\|\cdot\|_\infty$) is a complete metric space by Theorem X.12, and $X \subset \mathcal{C}([a,b])$ is closed, it follows from Theorem X.13 that X is a complete metric space (with the metric it inherits).

---

[4] This means that the subset is equipped with the metric that it inherits from the whole space X, as in (V.8).

X.3. Banach's fixed point theorem

Iterating a self-map

A **self-map** of a given space X is a function whose domain and codomain are both the same space X, i.e.,

$$f\colon X \to X. \tag{X.6}$$

Given a self-map f of X and an initial point $x_0 \in X$, we may define a sequence $(x_n)_{n\in\mathbb{N}_0}$ in X by

$$\begin{aligned} x_1 &= f(x_0) \\ x_2 &= f(x_1) = f\big(f(x_0)\big) \\ x_3 &= f(x_2) = f\big(f(x_1)\big) = f\Big(f\big(f(x_0)\big)\Big) \\ &\;\;\vdots \end{aligned}$$

i.e., recursively using

$$x_n = f(x_{n-1}) \in X \qquad \text{for } n \in \mathbb{N}. \tag{X.7}$$

The sequence $(x_n)_{n\in\mathbb{N}_0}$ thus obtained is called the **sequence of successive iterates** of x_0 under f. The nth member of this sequence is the image of x_0 under the function $X \to X$ obtained by composing f with itself n times; it can be denoted by

$$(\underbrace{f \circ \cdots \circ f}_{n \text{ times}})(x_0) \;=:\; f^{\circ n}(x_0). \tag{X.8}$$

The sequence of successive iterates can then be succinctly denoted $\big(f^{\circ n}(x_0)\big)_{n\in\mathbb{N}_0}$.

Definition X.16 (Fixed point)**.**
Let $f\colon X \to X$ be a function. A point $x^* \in X$ is called a **fixed point** of f if $f(x^*) = x^*$.

If a fixed point x^* of f is used as the initial point $x_0 = x^*$ of a sequence of successive iterates, then one simply obtains the constant sequence $(x^*, x^*, x^*, \ldots)$.

The contraction mapping principle

Definition X.17 (Contraction)**.**
Let (X, d) be a metric space. A function $f\colon X \to X$ is called a **contraction** on X if it is K-Lipschitz for some $K < 1$.

In other words, a contraction is a self-map a of metric space which does not stretch distances and instead contracts them by a nontrivial factor.

The following result is known as the **contraction mapping principle** or alternatively **Banach's fixed point theorem**. The scope of its applications is vast! We will look at a few example applications soon.

Theorem X.18 (Banach's fixed point theorem)**.**
Let (X, d) be a non-empty complete metric space and $f\colon X \to X$ a contraction.

Then there exists a unique fixed point $x^* \in X$ of f. Moreover, for any $x_0 \in X$, the sequence $\big(f^{\circ n}(x_0)\big)_{n\in\mathbb{N}}$ of successive iterates of x_0 converges to x^*.

Proof. Let us fix a $K < 1$ such that the contraction $f : X \to X$ is K-Lipschitz.

The uniqueness of the fixed point is deduced from the contraction property as follows. If both $x^* \in X$ and $\tilde{x}^* \in X$ are fixed points of f, then using first the fixed point equations $f(x^*) = x^*$ and $f(\tilde{x}^*) = \tilde{x}^*$, and then the K-Lipschitzness of f, we find

$$\mathsf{d}(x^*, \tilde{x}^*) \;=\; \mathsf{d}\big(f(x^*), f(\tilde{x}^*)\big) \;\leq\; K\,\mathsf{d}\big(x^*, \tilde{x}^*\big).$$

Moving the term on the right to the left-hand side this implies that $(1-K)\mathsf{d}(x^*, \tilde{x}^*) \leq 0$, which, since $1 - K \neq 0$, is only possible if $\mathsf{d}(x^*, \tilde{x}^*) = 0$. By axiom (M-0), this means $x^* = \tilde{x}^*$, so two different points could not have been fixed points.

Let us then use the contraction property to show that the sequence $(x_n)_{n\in\mathbb{N}_0} = \big(f^{\circ n}(x_0)\big)_{n\in\mathbb{N}_0}$ of successive iterates of any point $x_0 \in X$ is Cauchy. Denote $r = \mathsf{d}(x_0, x_1) = \mathsf{d}(x_0, f(x_0))$. Now note that by the K-Lipchitz property, for any $n \in \mathbb{N}$ we have

$$\mathsf{d}(x_n, x_{n+1}) \;=\; \mathsf{d}\big(f(x_{n-1}), f(x_n)\big) \;\leq\; K\,\mathsf{d}(x_{n-1}, x_n).$$

By an easy induction starting from $\mathsf{d}(x_0, x_1) = r$ this implies that $\mathsf{d}(x_n, x_{n+1}) \leq rK^n$ for any $n \in \mathbb{N}_0$. Let $k, \ell \in \mathbb{N}$ with $k < \ell$. To estimate the distance between members x_k and x_ℓ of the sequence, we apply the triangle inequality (iterated finitely many times) and get

$$\begin{aligned}
\mathsf{d}(x_k, x_\ell) &\leq \mathsf{d}(x_k, x_{k+1}) + \mathsf{d}(x_{k+1}, x_{k+2}) + \cdots + \mathsf{d}(x_{\ell-1}, x_\ell) \\
&\leq rK^k + rK^{k+1} + \cdots + rK^{\ell-1} \\
&= r\sum_{j=k}^{\ell-1} K^j \;=\; rK^k \frac{1-K^\ell}{1-K} \;\leq\; \frac{r}{1-K}\,K^k.
\end{aligned}$$

Now let $\varepsilon > 0$. Since $\frac{r}{1-K}\,K^n \to 0$ as $n \to \infty$, there exists some n_ε so that $\frac{r}{1-K}\,K^n < \varepsilon$ for any $n \geq n_\varepsilon$. The above estimate then shows that for any $k, \ell \geq n_\varepsilon$ we have $\mathsf{d}(x_k, x_\ell) < \varepsilon$. This shows that the sequence $(x_n)_{n\in\mathbb{N}_0}$ is Cauchy.

By completeness of X, the sequence $(x_n)_{n\in\mathbb{N}_0} = \big(f^{\circ n}(x_0)\big)_{n\in\mathbb{N}_0}$ of successive iterates of any point $x_0 \in X$ converges, since we saw above that it is Cauchy. Let $x' = \lim_{n\to\infty} x_n$ be the limit of such a sequence. We claim that x' is a fixed point of f. To see this, consider both the sequence $(x_n)_{n\in\mathbb{N}_0}$ which converges to x' and the sequence $\big(f(x_n)\big)_{n\in\mathbb{N}_0}$ of its images under f. The contraction f is continuous (in fact Lipschitz continuous), so by Corollary VIII.11 the limit of the sequence $\big(f(x_n)\big)_{n\in\mathbb{N}_0}$ is

$$\lim_{n\to\infty} f(x_n) \;=\; f\Big(\lim_{n\to\infty} x_n\Big) \;=\; f(x').$$

On the other hand, $f(x_n) = x_{n+1}$, so the sequence of images is simply a shifted version of the original sequence

$$\big(f(x_0), f(x_1), f(x_2), \ldots\big) \;=\; (x_1, x_2, x_3, \ldots),$$

and as a subsequence of $(x_0, x_1, x_2, \ldots)$, it in particular converges to the same limit x'. We thus have

$$\lim_{n\to\infty} f(x_n) \;=\; x'.$$

Now the uniqueness of limits (Lemma VIII.3) and a comparison of the two limit expressions obtained above shows that

$$x' = f(x'),$$

i.e., the limit x' is a fixed point of f.

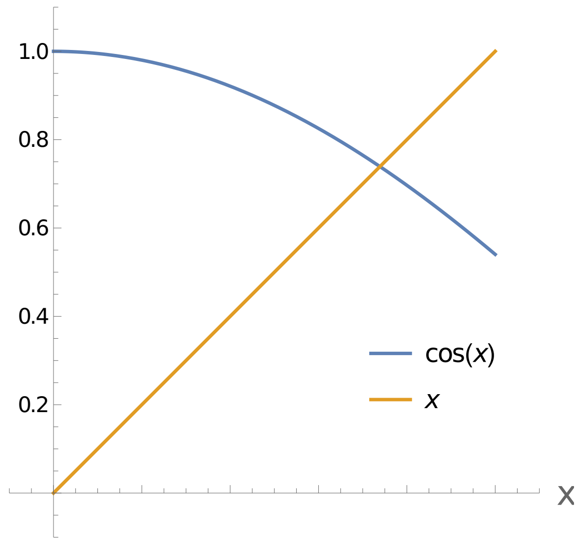
Since we have already shown that there can only exist one fixed point, we conclude from the above argument that the sequences of iterates with all initial values $x_0 \in X$ converge to the same fixed point $x^* \in X$ of f. We finish by noting that since we assumed that $X \neq \emptyset$, it is possible to choose at least some initial value $x_0 \in X$ and form its sequence of iterates, and thus obtain a limit which is a fixed point. This shows the existence of a fixed point $x^* \in X$. The proof is complete. $\square$

Applications

Banach's fixed point theorem, Theorem X.18, can be used to show the existence and uniqueness of solutions of equations of a suitable form, and also to form an iterative approximation scheme for the solutions. Let us look at a few examples.

Example X.19 (A simple application of Banach's fixed point theorem)**.**
Consider the equation

$$\cos(x) = x \qquad \text{on the interval } x \in [0,1]. \tag{X.9}$$



By plotting the left and right sides of the equation, it is easy to convince oneself that there is a unique solution to this equation (what is the easiest way to prove this rigorously?). However, familiar methods do not give a closed form solution to (X.9). Here we will not only show the existence and uniqueness of a solution, but also find a method of producing increasingly good approximations of it, with rigorous bounds on the error of the approximations.

Since we have $\cos(x) \in [0,1]$ for $x \in [0,1]$, the formula

$$f(x) = \cos(x) \qquad \text{defines a function} \qquad f\colon [0,1] \to [0,1].$$

Solutions to (X.9) are exactly the fixed points of f.

Now let us first note that by Theorems X.13 and X.10, the closed interval $[0,1]$ is a complete metric space, since it is a closed subset of the complete metric space $\mathbb{R}$.

To show that the function f above is a contraction, we find a bound for the absolute value of its derivative that is below one. Indeed, the derivative is $f'(x) = -\sin(x)$, and it satisfies the bound $|f'(x)| \leq \sin(1)$ for all $x \in [0,1]$ (the derivative is negative and decreasing, so attains its minimum at $x = 1$). Therefore denoting $K = \sin(1)$ we obtain as in Example VI.7 that f is K-Lipschitz, and since $K = \sin(1) < 1$, the function f is a contraction.

Since $[0,1]$ is complete and non-empty, and f is a contractive self-map of it, we may apply Banach's fixed point theorem. The first part of Banach's fixed point theorem guarantees that (X.9) indeed has a unique solution $x = x^* \in [0,1]$ (the unique fixed point x^* of f). Moreover, by the second part, given any $x_0 \in [0,1]$, we know that the sequence $(x_n)_{n\in\mathbb{N}_0} = \big(f^{\circ n}(x_0)\big)_{n\in\mathbb{N}_0}$ converges to the solution x^*. This allows us to form increasinly good approximate solutions. For example by letting $x_0 = 0$, the successive iterates are[5]

$$\begin{aligned}
x_0 &= 0 \\
x_1 &= \cos(x_0) = 1 \\
x_2 &= \cos(x_1) = \cos(1) \approx 0.540302 \\
x_3 &= \cos(x_2) = \cos\big(\cos(1)\big) \approx 0.857553 \\
x_4 &= \cos(x_3) = \cos\big(\cos(\cos(1))\big) \approx 0.65429 \\
x_5 &= \cos(x_4) = \cos\big(\cos(\cos(\cos(1)))\big) \approx 0.79348 \\
x_6 &= \cos(x_5) = \cos\big(\cos(\cos(\cos(\cos(1))))\big) \approx 0.701369 \\
&\ \vdots
\end{aligned}$$

A good numerical value of the true solution x^* is

$$x^* \approx 0.73908513321$$

and for example the 25th and 50th members of the iterative approximation scheme started from $x_0 = 0$ yield quite reasonable approximations[6]

$$\begin{aligned}
x_{25} &\approx 0.73910557193 \\
x_{50} &\approx 0.73908513217.
\end{aligned}$$

We can also control the error made when using the iterates x_n as approximations of the true solution x^*: from the proof of Theorem X.18 we see that $|x_n - x^*| \leq \frac{|x_1 - x_0|}{1-K} K^n$. Plugging in values $x_1 = 1$, $x_0 = 0$, and $K = \sin(1) \approx 0.841471$ (in the form $0.841 < K < 0.842$) we find that

$$|x_n - x^*| \leq \frac{|x_1 - x_0|}{1 - K} K^n < \frac{1}{1 - 0.842} (0.841)^n = \frac{(0.841)^n}{0.158}.$$

The error estimates obtained here are a bit conservative: they yield for example

$$|x_{25} - x^*| < 0.0835 \qquad \text{and} \qquad |x_{50} - x^*| < 0.0011;$$

correctly indeed, but the reality is even better... Let us just emphasize that such upper bounds for errors are fully rigorous and very easily obtained!

The following exercise is another application.

Exercise X.8 (Solving an equation by fixed point iteration)**.**
The equation $\tan x = 1/x$ has a unique solution on the interval $x \in (0, \pi/2)$ (sketch the graphs!). This equation can be rewritten as a fixed point equation $f(x) = x$ in (at least) two ways.

(a) Which way leads to a convergent fixed point iteration, when the initial value is $x_0 = 1$? (The careful choices of domain and codomain of the functions require a bit of work; maybe you can be a bit imprecise about those at first.)
(b) By fixed point iteration, determine an approximate solution to the equation, accurate to two decimal places.

Hint: *If a function f has an inverse function, then at a fixed point $f(a) = a$ we also have $f^{-1}(a) = a$. However, according to the formula*

$$(f^{-1})'(x) = \frac{1}{f'(f^{-1}(x))},$$

only one of the quantities $|f'(a)|$ and $|(f^{-1})'(a)| = 1/|f'(a)|$ can be < 1, and therefore lead to a contraction in the vicinity of the fixed point a.

[5]With an old-fashioned calculator, you can quickly obtain the iterates by entering the initial value 0, and then repeatedly pressing the cos-button.

[6]See for yourself, by repeatedly hitting the cos-button of an old-fashioned calculator, how the approximations converge to the solution.

The examples so far were about solving equations in a single real variable. Perhaps you already knew other methods that could have been used to numerically solve for example the equation (X.9).[7] But Banach's fixed point theorem applies in much more general spaces! Solving the following differential equation by other methods would be more difficult, wouldn't it?

Example X.20 (Solving a differential equation)**.**
Consider the differential equation

$$y'(s) = \sin\big(s\, y(s)\big) \qquad \text{for } s \in [0,1], \tag{X.10}$$

with the initial condition

$$y(0) = 1. \tag{X.11}$$

Finding an explicit formula for a solution to (X.10)–(X.11) seems hopeless (even with the help of computer algebra systems). We will nevertheless again show the existence and uniqueness of a solution, and find a method of producing increasingly good approximations of the true solution.

Observe first that the differential equation (X.10) is equivalent with its integrated form

$$\underbrace{\int_0^t y'(s)\,\mathrm{d}s}_{y(t)-y(0)} = \int_0^t \sin\big(s\, y(s)\big)\,\mathrm{d}s \qquad \text{for all } t \in [0,1],$$

and the initial value problem (X.10)–(X.11) is therefore equivalent to the integral equation

$$y(t) = 1 + \int_0^t \sin\big(s\, y(s)\big)\,\mathrm{d}s \qquad \text{for all } t \in [0,1]. \tag{X.12}$$

Let us now define a mapping

$$F\colon\ \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$$

by the formula[8]

$$\big[F(x)\big](t) = 1 + \int_0^t \sin\big(s\, x(s)\big)\,\mathrm{d}s \qquad \text{for } t \in [0,1].$$

The integral equation (X.12) is nothing but the fixed point equation $y = F(y)$ for F. Therefore this fixed point equation is in fact equivalent to the initial value problem (X.10)–(X.11).

We equip the space $\mathcal{C}([0,1])$ with the metric induced by the norm $\|\cdot\|_\infty$, and recall that $\mathcal{C}([0,1])$ is then a complete metric space by Theorem X.12. We will show that the function F above is a contraction of $\mathcal{C}([0,1])$. So, suppose that $x_1, x_2 \in \mathcal{C}([0,1])$. In order to estimate the distance between $F(x_1)$ and $F(x_2)$, first bound the absolute value of the function $F(x_2) - F(x_1) \in \mathcal{C}([0,1])$ at a point $t \in [0,1]$:

$$\begin{aligned}
\Big|\big[F(x_2) - F(x_1)\big](t)\Big| &= \Big|\big[F(x_2)\big](t) - \big[F(x_1)\big](t)\Big| \\
&= \left|1 + \int_0^t \sin\big(s\, x_2(s)\big)\,\mathrm{d}s - 1 - \int_0^t \sin\big(s\, x_1(s)\big)\,\mathrm{d}s\right| \\
&= \left|\int_0^t \Big(\sin\big(s\, x_2(s)\big) - \sin\big(s\, x_1(s)\big)\Big)\,\mathrm{d}s\right| \\
&\leq \int_0^t \Big|\sin\big(s\, x_2(s)\big) - \sin\big(s\, x_1(s)\big)\Big|\,\mathrm{d}s.
\end{aligned}$$

---

[7]For example, write (X.9) in the form $\cos(x) - x = 0$, and use Newton's iteration to find a zero of the left-hand side.

[8]Note that the argument of the function $F\colon \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$ is itself a function $x \in \mathcal{C}([0,1])$, and the value is also, $F(x) \in \mathcal{C}([0,1])$.

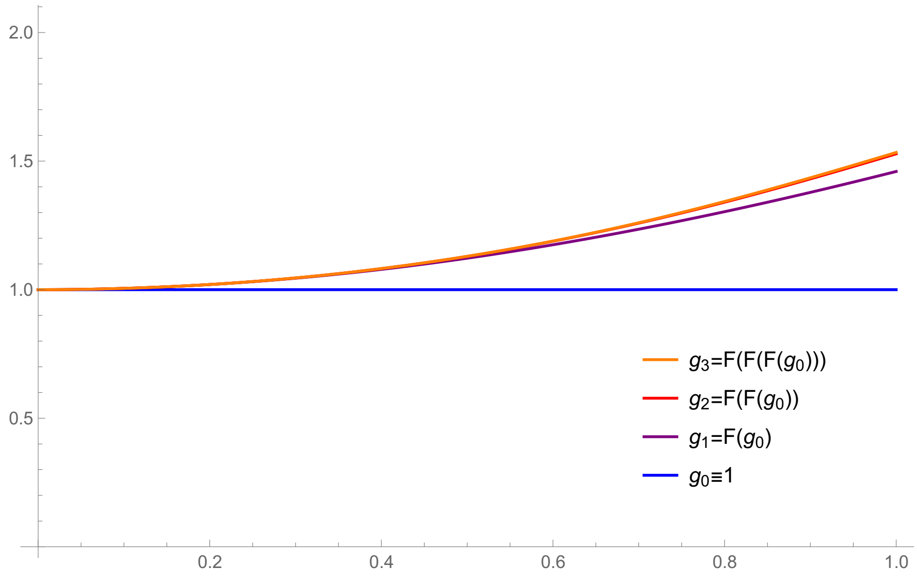


FIGURE X.2. A sequence of iterations of a self-map F of $\mathcal{C}([0,1])$ that converges in $\mathcal{C}([0,1])$ to the solution of the differential equation $y'(s) = \sin\big(s\,y(s)\big)$ with $y(0) = 1$. The iteration here is started from the constant function 1.

Using $|\sin(\theta_2) - \sin(\theta_1)| \leq |\theta_2 - \theta_1|$ (the function sin has derivative $|\sin'(\theta)| = |\cos(\theta)| \leq 1$ bounded by 1, and is therefore 1-Lipchitz by Example VI.7), we see that the integrand in the last expression is bounded by

$$\Big|\sin\big(s\,x_2(s)\big) - \sin\big(s\,x_1(s)\big)\Big| \leq \big|s\,x_2(s) - s\,x_1(s)\big| = |s|\,\big|x_2(s) - x_1(s)\big| \\ \leq s\,\|x_2 - x_1\|_\infty.$$

Plugging this estimate into the previous expression, we find

$$\Big|\big[F(x_2) - F(x_1)\big](t)\Big| \leq \|x_2 - x_1\|_\infty \int_0^t s\,\mathrm{d}s \;=\; \|x_2 - x_1\|_\infty\,\frac{t^2}{2} \;\leq\; \frac{1}{2}\,\|x_2 - x_1\|_\infty.$$

Since this holds for any $t \in [0,1]$, by taking the supremum over t we get

$$\big\|F(x_2) - F(x_1)\big\|_\infty \;\leq\; \frac{1}{2}\,\|x_2 - x_1\|_\infty,$$

which shows that $F\colon \mathcal{C}([0,1]) \to \mathcal{C}([0,1])$ is $\frac{1}{2}$-Lipschitz.

Since F thus is a contraction of the non-empty complete metric space $\mathcal{C}([0,1])$, Banach's fixed point theorem applies. It guarantees first of all that there exists a unique solution $y \in \mathcal{C}([0,1])$ to the initial value problem (X.10)–(X.11) (namely the unique fixed point y of F). This already is an important conclusion.

Moreover, Banach's fixed point theorem gives the following method of approximating the solution y: choose any $x_0 \in \mathcal{C}([0,1])$, and form the sequence $(x_n)_{n\in\mathbb{N}_0} = \big(F^{\circ n}(x_0)\big)_{n\in\mathbb{N}_0}$ of successive iterates. Then we have $\lim_{n\to\infty} F^{\circ n}(x_0) = y$ in $\mathcal{C}([0,1])$ with respect to the norm $\|\cdot\|_\infty$, i.e., uniformly. For example, we may take $x_0\colon [0,1] \to \mathbb{R}$ to be the constant function $x_0(s) = 1$ for all $s \in [0,1]$. The first few iterates are plotted in Figure X.2.

The idea in Example X.20 works very generally with ordinary differential equations: the Picard–Lindelöf theorem guarantees the existence and uniqueness of solutions, and gives an iterative approximation scheme for them. Its proof is a very straightforward modification of the above particular case.

X.4. ♡ Hilbert spaces and Banach spaces

Very powerful theory can be developed about normed spaces and inner product spaces which are complete as metric spaces with the induced metric. Due to the prominence of the theory and the scope of the applications, such spaces have been named as follows.

Definition X.21 (Hilbert space)**.**
An inner product space $(\mathsf{V}, \langle\cdot,\cdot\rangle)$, which is complete as a metric space with respect to the metric induced by the inner product, is called a **Hilbert space**.

Definition X.22 (Banach space)**.**
An normed space $(\mathsf{V}, \|\cdot\|)$, which is complete as a metric space with respect to the metric induced by the norm, is called a **Banach space**.

Remark X.23 (Hilbert spaces are Banach spaces)**.**
Every Hilbert space has a norm induced by its inner product. The metric induced by the inner product is (by definition) the same as the metric induced by the norm, so it is complete in the latter as well. Therefore every Hilbert space is in particular a Banach space.

The converse is not true: there are Banach spaces which are not Hilbert spaces; see Example X.25 below.

Our first example is familiar.

Example X.24 (Euclidean spaces are (finite-dimensional) Hilbert spaces)**.**
By Theorem X.11, the d-dimensional vector space $\mathbb{R}^d$ with its usual inner product is complete, and therefore a Hilbert space. By Remark X.23 it is also a Banach space.

The theory of Hilbert spaces and Banach spaces becomes more important in the infinite-dimensional setting, however. We have seen one important example of an infinite-dimensional Banach space already.

Example X.25 (A Banach space of continuous functions)**.**
By Theorem X.12, the space $\mathcal{C}([a,b])$ of continuous functions on a closed interval $[a,b]$, equipped with the supremum norm $\|\cdot\|_\infty$, is complete. Therefore it is a Banach space.

The following is an example of an infinite-dimensional Hilbert space.

Example X.26 (An infinite-dimensional Hilbert space)**.**
The space

$$\ell^2 = \left\{ (a_n)_{n\in\mathbb{N}} \;\middle|\; \sum_{n=1}^{\infty} a_n^2 < \infty \right\}$$

of square summable sequences has an inner product by Example IV.26.

This space is complete by Exercise X.9 below. Therefore ℓ^2 is a Hilbert space.

Exercise (♯) X.9 (Proof of completeness of ℓ^2)**.**
Prove that ℓ^2 is complete.

Remark X.27 (Complex Hilbert spaces and complex Banach spaces)**.**

In this course we have only considered real vector spaces. For complex vector spaces, there are notions of norm and inner product as well. Just like their real counterparts, they also induce metrics. By complex Hilbert spaces and complex Banach spaces, one means complex inner product spaces and complex normed spaces which are complete as metric spaces with the induced metric. In fact the theory of complex Hilbert spaces and Banach spaces is perhaps slightly simpler (not more complex, eh?) than that of real Hilbert spaces and Banach spaces. Complex Hilbert spaces and Banach spaces are also very common in applications.

Complex Hilbert spaces in particular have an indispensable role in the foundations of quantum mechanics and quantum field theory.

Both Hilbert spaces and Banach spaces are used very widely in mathematics and its applications. You will learn more about them in your subsequent studies.

Lecture XI

Compactness

Compactness is one of the most important topological notions; both for the development of mathematical theory and for practical applications. Compact spaces share many good properties analogous to those of finite sets, and it is quite appropriate to think of compactness as a topological analogy of finiteness.

One important aspect of the analogy with finite sets is that continuous real-valued functions on (nonempty) compact sets are bounded and have maxima and minima. In Lecture III we already saw that continuous functions on closed intervals have these properties, and closed intervals on the real line indeed serve as simple and familiar examples of compact sets. The existence of minima and maxima of continuous functions on compact sets is used in optimization, but also much more generally throughout mathematics — for example for the existence of solutions to partial differential equations (via calculus of variations), for the existence of holomorphic functions with various properties (e.g., conformal maps in the Riemann mapping theorem), and so on.

Besides the existence of extrema, another typical application of compactness is as a key step towards proving the existence of limits. In this role, compactness is for example used in the proof of the central limit theorem in probability, and in proofs of existence of thermodynamical limits in statistical physics and constructive quantum field theory.

Compactness is also applied more indirectly through its other consequences. Example applications include the Riemann integrability of continuous functions on closed intervals, (non-)homeomorphism results for various spaces, etc. We will look at a few arguments of this type already in this chapter.

Moreover, the nice properties that compact spaces have make them prominent as objects of study in their own right. Complex analysts love compact Riemann surfaces, algebraists (and particle physicists) love compact Lie groups, and so on.

As mentioned above, closed intervals will be among our first examples of compact sets. More generally, in the familiar setup of Euclidean spaces $\mathbb{R}^d$, we will see that a subset $A \subset \mathbb{R}^d$ is compact if and only if it is closed and bounded. This observation is known as the Bolzano-Weierstrass theorem or the Heine-Borel theorem, and it is indispensable in calculus.

Let us finish this introduction with a warning related to our conventions and terminology. In general topology, there are in fact two different notions of compactness: *sequential compactness* and *covering compactness*.[1] For metric spaces the two are

[1]In general topology, *compactness* usually refers to the latter, *covering compactness*. In the full generality of topological spaces, neither of these notions implies the other! Since our focus is on metric spaces, where the two are equivalent, we do not bother making a principled distinction in terminology. We content ourselves to only occasionally including the appropriate epithet when

nevertheless equivalent. Since our focus is on metric spaces, we will choose sequential compactness as our main definition of compactness, and we later show that (for metric spaces) it is equivalent to covering compactness. In the context of metric spaces, we can then use the term compactness for either one, without risk of confusion.

XI.1. Sequential compactness

Definition and examples of (sequential) compactness

Definition XI.1 (Sequential compactness).
A metric space X is said to be **(sequentially) compact** if every sequence in X has a convergent subsequence.

It is quite common that we care about compactness of subsets in a given metric space X. Naturally, such a subset $A \subset X$ is said to be compact if it is a compact metric space when equipped with the inherited metric.

Remark XI.2 (Compactness is an absolute notion).
This is a natural opportunity to emphasize important differences between two types of topological or metric properties of subsets. Alternatively, you may want to return to this remark only after having first checked out a few examples of compact and non-compact sets.

Let (X, d) be a metric space and let $A \subset X$ be a subset of it, equipped with the inherited metric d_A as in (V.8). We have introduced various properties of such subsets: e.g., openness, closedness, compactness, and completeness. Some properties of subsets of a metric space are "relative", in that they depend on both the subset A and the whole space X, whereas others are "absolute", in that they are meaningful properties of the subset A alone, as long as A is equipped with its metric d_A.

From the above it should be clear that compactness is an absolute notion. A subset $A \subset X$ is compact if and only if every sequence in A has a subsequence which converges in A. At no point does deciding about this require anything else besides the metric d_A on this subset.

Openness and closedness, by contrast, are relative notions. The meaning of "$A \subset X$ is an open set" or "$A \subset X$ is a closed set" depends on the ambient space X; see Remarks VII.2 and VII.11. For example the set $(0, \pi]$ is open in the space $(-\pi, \pi]$ but not open in $\mathbb{R}$.

Let us reiterate this important distinction. To say that A is an open or a closed set requires specifying the underlying metric space X in which it is viewed a subset — the meaning of the statement depends on this! By contrast, to say that A is compact does not require further specifying the context (a metric or a topology on A alone suffices).

Using results from Lecture III, we immediately get a good example.

Example XI.3 (Closed intervals are compact).
Let $a \leq b$ and consider the closed interval $[a, b] \subset \mathbb{R}$. If $(x_n)_{n \in \mathbb{N}}$ is a sequence on $[a, b]$, then $(x_n)_{n \in \mathbb{N}}$ is in particular a bounded real-number sequence, since $a \leq x_n \leq b$ for every $n \in \mathbb{N}$. By Theorem III.7, there then exists a subsequence $(x_{\varphi(n)})_{n \in \mathbb{N}}$ which converges in $\mathbb{R}$. But by preservation of bounds (Corollary II.15), the limit also satisfies $a \leq \lim_{n \to \infty} x_{\varphi(n)} \leq b$,

it adds to clarity. This remark and warning about terminology is mainly intended for those who will pursue more advanced studies that make use of general topology.

and so lies also in $[a,b]$. This shows that any sequence in $[a,b]$ has a subsequence which is convergent (in $[a,b]$), i.e., that the closed interval $[a,b]$ is compact.

Example XI.4 (The real line is not compact).
Consider the real line $\mathbb{R}$. For example the sequence $(x_n)_{n\in\mathbb{N}}$ given by $x_n = -n$ for $n \in \mathbb{N}$ has no convergent subsequences (it does not even have bounded subsequences). Therefore $\mathbb{R}$ is not compact.

Example XI.5 (Open intervals are not compact).
Let $a < b$ and consider the open interval $(a,b) \subset \mathbb{R}$. Consider the sequence $(x_n)_{n\in\mathbb{N}}$ in (a,b) given by $x_n = a + \frac{b-a}{2n}$ for $n \in \mathbb{N}$. This sequence has no convergent subsequences in (a,b) (by uniqueness of limits in $\mathbb{R}$, since in $\mathbb{R}$ the sequence converges to $a \notin (a,b)$). This shows that the open interval (a,b) is not compact.

Exercise (✓) XI.1 (Half-open intervals are not compact).
Let $a < b$. Show that the half-open intervals $[a,b) \subset \mathbb{R}$ and $(a,b] \subset \mathbb{R}$ are not compact.

Compactness versus completeness

The relationship of compactness to completeness is quite easy. We therefore use this connection to the topic of Lecture X as a small warm-up here.

Theorem XI.6 (Compactness implies completeness).
Every compact metric space is complete.

Proof. Suppose X is a compact metric space. Let $(x_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in X. By compactness, $(x_n)_{n\in\mathbb{N}}$ has some convergent subsequence, and by Lemma X.5 a Cauchy sequence with a convergent subsequence is necessarily convergent. This shows the completeness of X. □

Remark XI.7 (Completeness does not imply compactness).
The converse implication to Theorem XI.6 does not hold: for example the real line $\mathbb{R}$ is complete (Theorem X.10) but not compact (Example XI.4).

Compactness and closed subsets

Compactness is also quite closely related to the closedness of subsets. We present two results, roughly speaking establishing implications both ways — except that in the first implication below we must assume that the whole space is compact.

Theorem XI.8 (Closed subsets of compact spaces are compact).
Suppose that X is a compact metric space and $A \subset X$ is a closed subset. Then A is compact.

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in A. Since $A \subset X$, the sequence $(x_n)_{n\in\mathbb{N}}$ can be viewed as a sequence in X, too, and by the compactness of X, it has a subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ which converges to a limit $x = \lim_{n\to\infty} x_{\varphi(n)} \in X$. By the closedness of $A \subset X$ and the fact that $(x_{\varphi(n)})_{n\in\mathbb{N}}$ is a sequence in A, it follows by Theorem VIII.7 that $x = \lim_{n\to\infty} x_{\varphi(n)} \in A$. This shows that any sequence in A has a subsequence which converges in A, i.e., that A is compact. □

Exercise XI.2 (♡ The Cantor set is compact)**.**
Prove that the Cantor set C, defined in Appendix B.3, is compact.
Hint: *Start by observing that* $C \subset [0, 1]$.

Theorem XI.9 (Compact subsets are necessarily closed)**.**
Let X be a metric space and $A \subset X$ a compact subset. Then $A \subset X$ is closed.

Proof. We prove the closedness of $A \subset X$ by the characterization with sequences. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in A which converges in X to a limit $x = \lim_{n\to\infty} x_n \in X$. By the compactness of A, the sequence $(x_n)_{n\in\mathbb{N}}$ in A has some subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ which converges in A to a limit $a = \lim_{n\to\infty} x_{\varphi(n)} \in A$. But in X, as a subsequence of the convergent sequence $(x_n)_{n\in\mathbb{N}}$, it converges to the same limit $x \in X$ as the full sequence, $\lim_{n\to\infty} x_{\varphi(n)} = x$. By uniqueness of limits in X, we conclude $x = a$, and in particular the limit $x = a \in A$ of the sequence $(x_n)_{n\in\mathbb{N}}$ is in A. This shows closedness of A (by the characterization in Theorem VIII.7). □

Exercise XI.3 (Decreasing sequences of compact sets)**.**
(a) Let X be a metric space and $(A_n)_{n\in\mathbb{N}}$ a sequence of subsets of the space X such that
- for each $n \in \mathbb{N}$, the subset $A_n \subset X$ is non-empty and compact;
- the sequence is decreasing in the sense that for each $n \in \mathbb{N}$ we have $A_{n+1} \subset A_n$.

Prove that the intersection of these subsets
$$A = \bigcap_{n=1}^{\infty} A_n$$
is (i) non-empty; and (ii) compact.

(b) Give an example of a decreasing sequence of non-empty closed subsets, whose intersection is empty.

Preservation of compactness under continuous mappings

Continuous functions preserve the compactness of subsets in the following sense.

Theorem XI.10 (Compactness is preserved under continuous mappings)**.**
Let X and Y be metric spaces, $f\colon X \to Y$ a continuous function, and $A \subset X$ a compact subset. Then the image $f[A] \subset Y$ is compact.

Proof. Let $(y_n)_{n\in\mathbb{N}}$ be a sequence in the image $f[A]$. To prove the compactness of $f[A]$, we must show the existence of a convergent subsequence. For each $n \in \mathbb{N}$, by definition of the image, we have $y_n = f(x_n)$ for some $x_n \in A$. Then the sequence $(x_n)_{n\in\mathbb{N}}$ in A has, by the compactness of A, a subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ which converges in A to a limit $x = \lim_{n\to\infty} x_{\varphi(n)} \in A$. By the continuity of f and Corollary VIII.11 we then have
$$\lim_{n\to\infty} y_{\varphi(n)} \;=\; \lim_{n\to\infty} f(x_{\varphi(n)}) \;=\; f\big(\lim_{n\to\infty} x_{\varphi(n)}\big) \;=\; f(x).$$
This shows that the subsequence $(y_{\varphi(n)})_{n\in\mathbb{N}}$ converges to $f(x) \in f[A]$. Compactness of $f[A]$ is thus established. □

A nice corollary of the preservation of compactness is the following simplified criterion for the homeomorphism property, under the assumption that the domain is compact.

Corollary XI.11 (Continuous bijections defined on a compact space)**.**
Suppose that X and Y are metric spaces, and that X is compact. If a function $f\colon X \to Y$ is continuous and bijective, then it is a homeomorphism.

Proof. Suppose that $f\colon X \to Y$ is continuous and bijective. To show that f is a homeomorphism, we must show that the inverse function $f^{-1}\colon Y \to X$ is continuous. We will do this by the characterization of Theorem VII.22, by showing that preimages under the inverse function $f^{-1}\colon Y \to X$ of closed subsets are closed.

So let $A \subset X$ be closed. Since X is compact, by Theorem XI.8 the closed subset $A \subset X$ is compact as well. Note that the preimage of A under the inverse function f^{-1} is the same as the image of A under f, i.e., $(f^{-1})^{-1}[A] = f[A]$. Since A is compact and f is continuous, the image $f[A]$ is compact by Theorem XI.10. By Theorem XI.9, the compact subset $f[A] \subset Y$ is necessarily closed. We have thus shown that the preimage $(f^{-1})^{-1}[A] \subset Y$ of an arbitrary closed subset $A \subset X$ under $f^{-1}\colon Y \to X$ is closed. This proves the continuity of f^{-1}, and establishes that f is a homeomorphism. □

With the preservation of compactness under continuous functions, we also get some easy proofs of non-homeomorphism between spaces.

Example XI.12 (Closed intervals are not homeomorphic to other types of intervals)**.**
Let $a < b$ and $c < d$. Then there do not exist any continuous surjections $f\colon [a,b] \to (c,d)$ or $f\colon [a,b] \to [c,d)$ or $f\colon [a,b] \to (c,d]$, since $[a,b]$ is compact and its image $f\big[[a,b]\big]$ under a continuous function f is compact (Theorem XI.10), but the codomains (c,d), $[c,d)$, $(c,d]$ here are not compact (Example XI.5 and Exercise XI.1). In particular, there do not exist such homeomorphisms either. This shows that closed intervals are not homeomorphic to either open intervals or half-open intervals,

$$[a,b] \not\approx (c,d), \qquad [a,b] \not\approx [c,d).$$

The preservation of compactness under continuous functions gives also other easy non-homeomorphism proofs. We will give more examples after we have good characterizations of compact subsets in familiar spaces, especially in the Euclidean spaces.

Continuous real-valued functions on compact spaces

The next result is totally analogous to Theorem III.14. Also its proof goes through almost verbatim! The result is, however, so important that even at the risk of being repetitive, we provide the proof here. The similarity of the proofs also shows that compactness was really the only essential property of the closed interval responsible for the result (for example one-dimensionality or connectedness, or any such special property of the closed interval did not play a role!).

Theorem XI.13 (Continuous real-valued functions on compact spaces)**.**
Let X be a compact metric space and $f\colon X \to \mathbb{R}$ a continuous function. Then,

(i) *f is bounded, i.e., $f[X] \subset \mathbb{R}$ is a bounded subset;*
(ii) *if $X \neq \emptyset$, then f has a minimum and maximum.*

Proof of Theorem XI.13 Let us prove the two assertions separately.

proof of (i): Suppose, by way of contradiction, that f is not bounded. Then for every $n \in \mathbb{N}$, it is possible to choose some $x_n \in X$ such that $|f(x_n)| \geq n$. Let us form a sequence $(x_n)_{n\in\mathbb{N}}$ using such choices. By the compactness of X, it has a convergent subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$. If we denote the limit of such a convergent subsequence by $x = \lim_{n\to\infty} x_{\varphi(n)} \in X$, then by the continuity of f and Corollary VIII.11 we get that $\big(f(x_{\varphi(n)})\big)_{n\in\mathbb{N}}$ converges to $f(x)$. But by the choice of x_n we have $|f(x_{\varphi(n)})| \geq \varphi(n) \geq n$ for all $n \in \mathbb{N}$, which means that the sequence $\big(f(x_{\varphi(n)})\big)_{n\in\mathbb{N}}$ is not bounded and so cannot be convergent. This is a contradiction. We conclude that f had to be bounded.

proof of (ii): Let us only prove that f has a minimum — the existence of maximum can be concluded similarly (or by considering the minimum of the continuous function $-f$).

By (i), f is bounded, so the infimum of its values is finite,

$$C \;:=\; \inf\big\{f(x) \;\big|\; x\in X\big\} \;\in\; \mathbb{R}.$$

For all $n\in\mathbb{N}$ there then exists some $x_n \in X$ such that $f(x_n) \le C + \frac{1}{n}$. Let us form a sequence $(x_n)_{n\in\mathbb{N}}$ using such choices. By the compactness of X, this sequence has a convergent subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$. Let us denote the limit of such a subsequence by $x = \lim_{n\to\infty} x_{\varphi(n)} \in X$. By the choice of C (infimum of the values) and of x_n we have

$$C \;\le\; f(x_{\varphi(n)}) \;\le\; C + \frac{1}{\varphi(n)} \;\le\; C + \frac{1}{n}.$$

The squeeze theorem (Lemma II.17) thus gives $\lim_{n\to\infty} f(x_{\varphi(n)}) = C$. On the other hand, by continuity and Corollary VIII.11 we get that

$$f(x) = f\big(\lim_{n\to\infty} x_{\varphi(n)}\big) = \lim_{n\to\infty} f(x_{\varphi(n)}) = C.$$

We conclude that

$$f(x) = \inf\big\{f(x) \;\big|\; x\in[a,b]\big\},$$

which implies that the minimum of $f\colon X\to\mathbb{R}$ is attained at $x\in X$. $\square$

Corollary XI.14 (Compact metric spaces are bounded).
Any compact metric space is bounded.

Proof. Let (X,d) be a compact metric space. The case $X=\emptyset$ is trivial, so let us assume that $X\neq\emptyset$. Then pick a point $z\in X$, and define a function $f\colon X\to[0,\infty)$ by $f(x)=\mathsf{d}(z,x)$ for $x\in X$. By Example VI.12, this f is continuous. By the compactness of X and Theorem XI.13, f is bounded, which implies that there exists some $R\ge 0$ such that $|f(x)|\le R$ for all $x\in X$, i.e., that $\mathsf{d}(z,x)\le R$ for all $x\in X$. This shows boundedness of X (to elaborate, one gets $X\subset\overline{\mathcal{B}}_R(z)\subset\mathcal{B}_{R+1}(z)$ and then one can apply Lemma V.17). $\square$

Exercise XI.4 (Highest point in a compact set in $\mathbb{R}^3$).
Let $A\subset\mathbb{R}^3$ be a compact subset. Show that it contains a highest point $a\in A$, i.e., a point $a=(a_1,a_2,a_3)$ such that we have $a_3\ge x_3$ for all $x=(x_1,x_2,x_3)\in A$.

Exercise XI.5 (Fixed-point-free continuous self-maps of a compact space).
Let X be a compact metric space and $g\colon X\to X$ a continuous function that has no fixed points. Show that there exists a $c>0$ such that $\mathsf{d}\big(g(x),x\big)\ge c$ for all $x\in X$.
Hint: *In this exercise you may want to construct a suitable continuous function $X\to[0,\infty)$ by a composition of appropriately chosen functions $X\to X\times X$ and $X\times X\to[0,\infty)$.*

XI.2. Compact subsets in Euclidean spaces

From calculus courses you might already be familiar with the following statement, the Bolzano-Weierstrass theorem, which characterizes compactness of subsets of the Euclidean spaces $\mathbb{R}^d$.

Theorem XI.15 (Bolzano-Weierstrass theorem[2])**.**
A subset $A \subset \mathbb{R}^d$ is compact if and only if it is closed and bounded.

Proof. Let us prove separately the two implications, "if" and "only if".

proof of "only if": Assume that $A \subset \mathbb{R}^d$ is compact. Then it is closed by Theorem XI.9 and bounded by Corollary XI.14. This finishes the proof of the "only if" part.

proof of "if": We only consider the case $d = 2$ in detail; the modifications needed for general $d \in \mathbb{N}$ are straightforward.[3]

Assume that $A \subset \mathbb{R}^2$ is closed and bounded. We must show that it is compact. So let $(z_n)_{n\in\mathbb{N}}$ be a sequence in A. Write $z_n = (x_n, y_n)$, for $n \in \mathbb{N}$, where $x_n = \mathrm{pr}_1(z_n)$ and $y_n = \mathrm{pr}_2(z_n)$ are the coordinates. Since $A \subset \mathbb{R}^2$ is bounded, we have that $(z_n)_{n\in\mathbb{N}}$ is bounded sequence in $\mathbb{R}^2$, and the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are bounded in $\mathbb{R}$ (since the coordinate projections are Lipschitz).

The bounded real-number sequence $(x_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ by Theorem III.7. Let us denote its limit by $x = \lim_{n\to\infty} x_{\varphi(n)} \in \mathbb{R}$.

Now the sequence $(z_{\varphi(n)})_{n\in\mathbb{N}}$ in $\mathbb{R}^2$ is such that the sequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ of its first coordinates converge. The second coordinate sequence $(y_{\varphi(n)})_{n\in\mathbb{N}}$ is a bounded real-number sequence, so it has some convergent subsequence $(y_{\varphi(\psi(n))})_{n\in\mathbb{N}}$. Let us denote its limit by $y = \lim_{n\to\infty} y_{\varphi(\psi(n))} \in \mathbb{R}$. The corresponding first coordinate sequence $(x_{\varphi(\psi(n))})_{n\in\mathbb{N}}$ is a subsequence of the convergent sequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$, so it converges to the same limit $\lim_{n\to\infty} x_{\varphi(\psi(n))} = x$. We get that the subsequence $(z_{\varphi(\psi(n))})_{n\in\mathbb{N}}$ of the sequence $(z_n)_{n\in\mathbb{N}}$ in $A \subset \mathbb{R}^2$ has both of its coordinate sequences converging. By Theorem VIII.16 we then get $\lim_{n\to\infty} z_{\varphi(\psi(n))} = (x, y) \in \mathbb{R}^2$.

Now we recall that $(z_n)_{n\in\mathbb{N}}$ was a sequence in $A \subset \mathbb{R}^2$, and therefore also $(z_{\varphi(\psi(n))})_{n\in\mathbb{N}}$ is a sequence in $A \subset \mathbb{R}^2$. Since A is closed, by assumption, this sequence which converges in $\mathbb{R}^2$ must have its limit $\lim_{n\to\infty} z_{\varphi(\psi(n))} = (x, y)$ in A (according to Theorem VIII.7). We have thus shown that an arbitrary sequence $(z_n)_{n\in\mathbb{N}}$ in $A \subset \mathbb{R}^2$ has a subsequence which converges in A. This by definition shows that A is compact. $\square$

Example XI.16 (The unit circle is compact)**.**
Consider the unit circle in the plane,

$$\mathbf{S}^1 \;=\; \left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; x^2 + y^2 = 1 \right\}.$$

It is clearly bounded (it is contained in the closed ball of radius 1 centered at the origin). It is also closed as the preimage of the singleton $\{1\} \subset \mathbb{R}$ under the polynomial $(x, y) \mapsto x^2 + y^2$, which is a continuous function $\mathbb{R}^2 \to \mathbb{R}$. As a closed and bounded subset of the Euclidean plane $\mathbb{R}^2$, the unit circle $\mathbf{S}^1 \subset \mathbb{R}^2$ is compact (by the Bolzano-Weierstrass theorem).

The compactness of the unit circle gives a quick proof that the natural continuous mapping $[0, 2\pi) \to \mathbf{S}^1$ given by $\theta \mapsto \big(\cos(\theta), \sin(\theta)\big)$ is not a homeomorphism — and in fact that no homeomorphism $f\colon [0, 2\pi) \to \mathbf{S}^1$ exists! Namely, the inverse of such a homeomorphism would be a continuous surjection $g\colon \mathbf{S}^1 \to [0, 2\pi)$, and the image $g[\mathbf{S}^1]$ of the compact set $\mathbf{S}^1$ would be compact (Theorem XI.10), but the half-open interval $[0, 2\pi)$ is not compact (Exercise XI.1). This simple way to obtain a general non-existence result is a good illustration of the power of compactness.

[2] When compactness is interpreted as covering compactness instead of sequential compactness, the same statement is often called the Heine-Borel theorem. So once we prove the equivalence of the two notions, using either name for this result is reasonable.

[3] The case $d = 1$ is basically just Example XI.3. The case of $d > 2$ is only notationally more cumbersome, but involves no essentially new ideas. But thinking about it is a very good exercise, by which you can make sure you understood the essence of the proof here!

We emphasize that the Bolzano-Weierstrass theorem is a statement about compact subsets of the Euclidean space $\mathbb{R}^d$, but in other metric spaces the situation can be different. The following exercise provides a concrete example of a closed and bounded subset of a familiar metric space, which is not compact.

Exercise XI.6 (A closed and bounded non-compact set)**.**
Consider the space $\mathcal{C}([0,1])$ of continuous real-valued functions on $[0,1]$, with the metric d induced by the sup-norm $\|\cdot\|_\infty$

$$\mathsf{d}(f,g) = \|g-f\|_\infty = \sup\Big\{ |g(x)-f(x)| \;\Big|\; x \in [0,1] \Big\}.$$

Prove that the closed unit ball $\overline{B}(\vec{0},1)$ is not compact.

Hint: *Construct a sequence (f_n) for which $\|f_n\|_\infty \leq 1$ for all $n \in \mathbb{N}$ and*

$$\|f_k - f_n\|_\infty \geq 1$$

whenever $k \neq n$. Such a sequence cannot have a convergent subsequence (why?).

Exercise XI.7 (All norms in finite-dimensional vector spaces are bilipschitz equivalent)**.**
Consider the d-dimensional vector space $\mathbb{R}^d$ with different norms. Denote the Euclidean norm on $\mathbb{R}^d$ by $\|\cdot\|$, and denote another arbitrary norm on $\mathbb{R}^d$ by $|||\cdot|||$. Unless explicitly otherwise specified, use the metric induced by the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^d$.

(a) Prove that the function $f\colon \mathbb{R}^d \to [0,\infty)$ defined by $f(\vec{v}) = |||\vec{v}|||$ is Lipschitz.
Remark: Note that on the domain $\mathbb{R}^d$ of f we use the metric induced by $\|\cdot\|$, so the Lipschitzness of the norm $|||\cdot|||$ is *not* the same as the statement of Exercise VI.8.
(b) Prove that there exists constants $A, B > 0$ such that

$$A\,\|\vec{v}\| \;\leq\; |||\vec{v}||| \;\leq\; B\,\|\vec{v}\| \qquad \text{for all } \vec{v} \in \mathbb{R}^d.$$

Hint: Start by proving that f attains its minimum on the set $S = \big\{\vec{u} \in \mathbb{R}^d \;\big|\; \|\vec{u}\| = 1\big\}$.
(c) Using (b), prove that the metrics on $\mathbb{R}^d$ induced by the norms $\|\cdot\|$ and by $|||\cdot|||$ are bilipschitz equivalent in the sense of (VII.1).
(d) Use (c) to show that all metrics induced by a norm on $\mathbb{R}^d$ are topologically equivalent.

XI.3. Compactness of product spaces

It is often important to know that Cartesian products of compact spaces are compact; some typical example applications will be given later. The very general result of this type[4] is known as Tychonoff's theorem, but for our purposes the following concrete case is sufficient.

Theorem XI.17 (Products of compact spaces are compact)**.**
If (X,d_X) and (Y,d_Y) are compact metric spaces, then the space $X\times Y$ equipped with any of the metrics (V.20) *is also compact.*

Sketch of proof. The proof idea is similar to the $d=2$ case of Theorem XI.15. We leave the details as an exercise. □

Exercise XI.8 (Proof of Theorem XI.17)**.**
Write a detailed proof of Theorem XI.17.

[4]Tychonoff's theorem concerns arbitrary (in particular also uncountably infinite) products of compact topological spaces. The appropriate notion in this context is covering compactness (which is equivalent to sequential compactness in metric spaces).

Example XI.18 (Diameters of compact sets).
Let (X, d) be a metric space, and $A \subset X$ a non-empty compact subset. Recall that the diameter of A is defined by (V.16) as

$$\operatorname{diam}(A) \;=\; \sup\Big\{\mathsf{d}(a_1, a_2) \;\Big|\; a_1, a_2 \in A\Big\}.$$

The function $f = \mathsf{d}|_{A\times A}$,

$$f\colon\; A \times A \to [0, \infty), \qquad\qquad f(a_1, a_2) = \mathsf{d}(a_1, a_2) \quad \text{for } (a_1, a_2) \in A \times A,$$

is continuous (as in Exercise VI.11) and the set $A \times A$ is compact (by compactness of A and Theorem XI.17), so f attains its maximum on $A\times A$ (by Theorem XI.13). If $(a_1', a_2') \in A \times A$ is a point of the product space where the maximum is attained, then the pair of points $a_1', a_2' \in A$ is such that

$$\mathsf{d}(a_1', a_2') \;=\; \operatorname{diam}(A).$$

In particular, for a compact set the diameter is the maximal distance among pairs of points in the set.

Exercise XI.9 (Distances between compact sets).
Let (X, d) be a metric space, and $A, B \subset X$ two subsets of it. Recall that the distance between the subsets is defined (V.17) as $\operatorname{dist}(A, B) = \inf\big\{\mathsf{d}(a, b) \;\big|\; a \in A,\ b \in B\big\}$.

Show that if A and B are non-empty and compact, then there exists $a' \in A$ and $b' \in B$ such that

$$\operatorname{dist}(A, B) = \mathsf{d}(a', b').$$

Conclude in particular that if the non-empty compact sets are disjoint, $A \cap B = \emptyset$, then the distance between them is positive (compare and contrast with Example V.18).

Exercise XI.10 (Maximal perimeter triangle with vertices in a compact set).
Let $K \subset \mathbb{R}^2$ be a compact subset of the Euclidean plane $\mathbb{R}^2$. Show that it contains points $a, b, c \in K$, for which the perimeter

$$\|a - b\| + \|b - c\| + \|c - a\|$$

of the triangle is maximal among all triangles whose vertices lie in K.

XI.4. Covering compactness

In general topology, the standard definition of compactness is different from the sequential compactness we discussed so far. It is based on open covers.

We give the definitions of open covers and covering compactness for a subset of a metric space rather than a whole space, because this is often how it is used.

Definition XI.19 (Open cover).
Let X be a metric space and $A \subset X$ a subset. A collection $(U_j)_{j\in J}$ of subsets of X is called an **open cover** of A if for each $j \in J$ the subset $U_j \subset X$ is open, and we have

$$A \;\subset\; \bigcup_{j\in J} U_j.$$

The key auxiliary result relating sequential compactness to open covers is the following.

Lemma XI.20 (Lebesgue number lemma)**.**

Let X be a metric space, $A \subset X$ a compact subset, and $(U_j)_{j\in J}$ an open cover of A. Then there exists a positive number $\lambda > 0$ such that for every $a \in A$ there exists some index $j \in J$ such that $\mathcal{B}_\lambda(a) \subset U_j$.

A number $\lambda > 0$ as above it is often called a **Lebesgue number** of the open cover $(U_j)_{j\in J}$ of A.

Remark XI.21 (Lebesgue number lemma as a logical statement)**.**

Let X be a metric space, $A \subset X$ a subset, and $(U_j)_{j\in J}$ an open cover of A. The statement of Lebesgue's number lemma can be written in logical symbols as:

$$\exists\, \lambda > 0 : \quad \forall\, a \in A : \quad \exists\, j \in J : \quad \mathcal{B}_\lambda(a) \subset U_j.$$

The proof of Lemma XI.20 will proceed by contradiction, so it is instructive to write also the negation of the statement above. By routine manipulation of negations of statements with quantifiers, we find that the negation of the above statement is:

$$\forall\, \lambda > 0 : \quad \exists\, a \in A : \quad \forall\, j \in J : \quad \mathcal{B}_\lambda(a) \not\subset U_j.$$

Proof of Lemma XI.20. We will use a proof by contradiction. So suppose that no such $\lambda > 0$ exists. In other words, for every $\lambda > 0$ there exists some $a \in A$ such that for all $j \in J$ we have $\mathcal{B}_\lambda(a) \not\subset U_j$. For $n \in \mathbb{N}$, consider $\lambda = \frac{1}{n}$ and a correspondingly chosen $a_n \in A$ such that $\mathcal{B}_{1/n}(a_n) \not\subset U_j$ for all $j \in J$. Form a sequence $(a_n)_{n\in\mathbb{N}}$ using such choices, and note that by the compactness of A there exists a subsequence $(a_{\varphi(n)})_{n\in\mathbb{N}}$ which converges to some $a = \lim_{n\to\infty} a_{\varphi(n)} \in A$. Now since $a \in A \subset \bigcup_{j\in J} U_j$, there exists some $j \in J$ such that $a \in U_j$. Since $U_j \subset X$ is open, there exists an $r > 0$ such that $\mathcal{B}_r(a) \subset U_j$. Now since $\lim_{n\to\infty} a_{\varphi(n)} = a$ and $\frac{2r}{3} > 0$, there exists some $n' \in \mathbb{N}$ such that $\mathsf{d}(a_{\varphi(n)}, a) < \frac{2r}{3}$ whenever $n \geq n'$. Let us choose some integer $n \geq \max\{n', \frac{3}{r}\}$. Then $\mathsf{d}(a_{\varphi(n)}, a) < \frac{2r}{3}$, and by triangle inequality we get that $\mathcal{B}_{r/3}(a_{\varphi(n)}) \subset \mathcal{B}_r(a)$. But since $\mathcal{B}_r(a) \subset U_j$, this implies $\mathcal{B}_{r/3}(a_{\varphi(n)}) \subset U_j$. On the other hand, since $\varphi(n) \geq n \geq \frac{3}{r}$, we have $\frac{r}{3} \geq \frac{1}{\varphi(n)}$ and therefore $\mathcal{B}_{1/\varphi(n)}(a_{\varphi(n)}) \subset \mathcal{B}_{r/3}(a_{\varphi(n)}) \subset U_j$. This is a contradiction, because the point $a_{\varphi(n)}$ was chosen so that $\mathcal{B}_{1/\varphi(n)}(a_{\varphi(n)}) \not\subset U_j$ for all $j \in J$. This contradiction finishes the proof. □

We then give the definition of covering compactness for a subset of a metric space.[5] [6]

Definition XI.22 (Covering compactness)**.**

A subset $A \subset X$ in a metric space X is **covering compact** if for every open cover $(U_j)_{j\in J}$ of A there exists a finite subset $J' \subset J$ such that $(U_j)_{j\in J'}$ is also an open cover of A (such a finite collection is called a **finite subcover**.)

In metric spaces sequential compactness and covering compactness are equivalent, and we refer to both simply as compactness (but recall the warning: in general topology, neither implies the other!).

Theorem XI.23 (Equivalence of sequential and covering compactness)**.**

A metric space X is sequentially compact if and only if it is covering compact.

[5] Phrased this way, it may seem that covering compactness would be a relative rather than an absolute notion, see Remark XI.2. But with Theorem VII.9 one can quite easily see that also covering compactness is an absolute notion. Of course (for metric spaces) also the proof of its equivalence with the absolute notion of sequential compactness is another way to reach the same conclusion.

[6] The covering compactness of the whole space X then simply means the covering compactness of $X \subset X$.

Proof. We prove both implications separately.

proof of "if": Assume X is covering compact, i.e., every open cover of X has a finite subcover. To show that X is sequentially compact, let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X. We must find a convergent subsequence for $(x_n)_{n\in\mathbb{N}}$.

Note that if some point $z \in X$ is such that every neighborhood $U \subset X$ of z contains infinitely many members of the sequence $(x_n)_{n\in\mathbb{N}}$, i.e., if $\#\{n \in \mathbb{N} \,|\, x_n \in U\} = \infty$ for all open sets $U \subset X$ with $z \in U$, then it is possible to extract a subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ which converges to z.[7] It suffices to show that some such point exists.

We will use a proof by contradiction to show the existence of such z. So assume the opposite, that for every $z \in X$ there exists some neighborhood $U_z \subset X$ of z which contains only finitely many members of the sequence $(x_n)_{n\in\mathbb{N}}$, i.e., $\#\{n \in \mathbb{N} \,|\, x_n \in U_z\} < \infty$. The collection $(U_z)_{z\in X}$ is obviously[8] an open cover of X. By the assumed covering compactness, there then exists some finite number $m \in \mathbb{N}$ of points $z_1, \ldots, z_m \in X$ such that the corresponding neighborhoods form a finite subcover, $X \subset U_{z_1} \cup \cdots \cup U_{z_m}$. But this is impossible: we would then have $\mathbb{N} = \{n \in \mathbb{N} \,|\, x_n \in X\} = \bigcup_{j=1}^{m} \{n \in \mathbb{N} \,|\, x_n \in U_{z_j}\}$, which would express the infinite index set $\mathbb{N}$ as a finite union of the finite sets $\{n \in \mathbb{N} \,|\, x_n \in U_{z_j}\}$. This is a contradiction, which proves the result.

proof of "only if": Assume X is sequentially compact, i.e., every sequence in X has a convergent subsequence. To show that X is covering compact, let $(U_j)_{j\in J}$ be an open cover of X. We must show that there exists a finite subcover of $(U_j)_{j\in J}$. Note first, however, that Lemma XI.20 guarantees the existence of a positive number $\lambda > 0$ such that for every $x \in X$ there exists some $j \in J$ such that $\mathcal{B}_\lambda(x) \subset U_j$. We fix such a $\lambda > 0$ and will use it below.

We will use a proof by contradiction to show the existence of the finite subcover. So assume the opposite: that no finite subset $J_0 \subset J$ is such that $X \subset \bigcup_{j\in J_0} U_j$. Construct a sequence recursively as follows. First pick[9] some point $x_1 \in X$ and choose an index $j_1 \in J$ so that $\mathcal{B}_\lambda(x_1) \subset U_{j_1}$. Since U_{j_1} does not suffice to cover X, we can choose a point $x_2 \in X \setminus U_{j_1}$ and an index $j_2 \in J$ such that $\mathcal{B}_\lambda(x_2) \subset U_{j_2}$. We continue like this inductively: with $x_1, \ldots, x_n \in X$ and $j_1, \ldots, j_n \in J$ chosen, since $U_{j_1}, \ldots, U_{j_n}$ do not suffice to cover X, we can choose $x_{n+1} \in X \setminus (U_{j_1} \cup \cdots \cup U_{j_n})$ and an index $j_{n+1} \in J$ so that $\mathcal{B}_\lambda(x_{n+1}) \subset U_{j_{n+1}}$. This recursive construction yields a sequence $(x_n)_{n\in\mathbb{N}}$ such that all pairwise distances between its members are bounded from below by λ: if $k, \ell \in \mathbb{N}$ with $\ell > k$, then $x_\ell \notin U_{j_k} \supset \mathcal{B}_\lambda(x_k)$ by construction, so in particular $\mathsf{d}(x_k, x_\ell) \geq \lambda$. But this implies that no subsequence of $(x_n)_{n\in\mathbb{N}}$ can be Cauchy, and in particular no subsequence can be convergent. This is a contradiction with the assumed sequential compactness of X. The proof is thus complete. $\square$

It follows from Theorems XI.15 and XI.23 that an open cover of a closed and bounded subset of $\mathbb{R}^d$ has a finite subcover — in this form the compactness theorem of subsets of the Euclidean space $\mathbb{R}^d$ is known as the Heine-Borel theorem.

[7] Indeed, in particular the open balls $\mathcal{B}_{2^{-n}}(z)$ of radii $r_n = 2^{-n}$, for $n \in \mathbb{N}$, are neighborhoods of z, and we can construct a subsequence recursively as follows. First, since $\mathcal{B}_{2^{-1}}(z)$ contains members of the sequence, we can let $\varphi(1) \in \mathbb{N}$ be an index such that $x_{\varphi(1)} \in \mathcal{B}_{2^{-1}}(z)$. Then inductively once $\varphi(1) < \cdots < \varphi(n)$ have been chosen, we use the fact that $\mathcal{B}_{2^{-n-1}}(z)$ contains infinitely many members of the sequence, and therefore it contains some member with larger index than $\varphi(n)$. We can therefore choose an index $\varphi(n+1) \in \mathbb{N}$ such that $\varphi(n+1) > \varphi(n)$ and $x_{\varphi(n+1)} \in \mathcal{B}_{2^{-n-1}}(z)$. With such choices we have $0 \leq \mathsf{d}(z, x_{\varphi(n)}) \leq 2^{-n} \to 0$ so by the squeeze theorem the subsequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ converges to z.

[8] Neighborhoods $U_z \subset X$ are by definition open, and any point z is covered at least by the neighborhood U_z of it.

[9] Note that $X \neq \emptyset$ since otherwise already the empty collection covers it, and we assumed that no finite subcollection of $(U_j)_{j\in J}$ covers X. Therefore picking some $x_1 \in X$ is indeed possible, or we have already reached the desired contradiction.

The following exercise illustrates that it is occasionally more convenient to work with covering compactness than sequential compactness.

Exercise XI.11 (The union of two compact sets is compact).
Prove that for any two compact subsets $A, B \subset X$ of a metric space X, the union $A \cup B$ is compact, ...

(a) ...directly from the definition of (sequential) compactness;
(b) ...by using the characterization of compactness by open covers.

Hint: *The proofs should begin as follows:*
a) Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $A \cup B$. ...
b) Let $(U_j)_{j\in J}$ be an open cover of $A \cup B$. ...

XI.5. Uniform continuity

Let (X, d_X) and (Y, d_Y) be two metric spaces.

Definition XI.24 (Uniform continuity).
A function $f : X \to Y$ is said to be **uniformly continuous** if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that $\mathsf{d}_Y\big(f(x_1), f(x_2)\big) < \varepsilon$ whenever $\mathsf{d}_X(x_1, x_2) < \delta$.

Remark XI.25 (Order of quantifiers).
Definition XI.24 (uniform continuity) differs from Definition VI.1 (continuity) "only" in the order of quantifiers:

- the continuity of $f : X \to Y$ means

$$\forall x_1 \in X \quad \forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall x_2 \in X \quad \mathsf{d}_X(x_1, x_2) < \delta \implies \mathsf{d}_Y\big(f(x_1), f(x_2)\big) < \varepsilon;$$

- the uniform continuity of $f : X \to Y$ means

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall x_1 \in X \quad \forall x_2 \in X \quad \mathsf{d}_X(x_1, x_2) < \delta \implies \mathsf{d}_Y\big(f(x_1), f(x_2)\big) < \varepsilon.$$

The crucial difference is that in the former case, the promised δ may depend on both x_1 and ε, whereas in the latter case δ may only depend on ε — one choice must work "uniformly" over all locations x_1 in the space X.

The proof of the following is easy once you have internalized the difference in Remark XI.25.

Lemma XI.26 (Uniform continuity implies continuity).
If $f : X \to Y$ is uniformly continuous, then it is continuous.

Exercise (✓) XI.12 (Proof of Lemma XI.26).
Prove Lemma XI.26.

The converse implication to Lemma XI.26 does not hold in general.

Example XI.27 (Continuity does not imply uniform continuity).
Consider the function

$$f : \mathbb{R} \to \mathbb{R} \qquad \text{given by} \qquad f(x) = x^2 \;\text{ for } x \in \mathbb{R}.$$

As a polynomial function, f is continuous (Corollary III.12). But f is *not* uniformly continuous, as we will see below. Indeed, assume, by contradiction, that f were uniformly continuous. Then consider for example $\varepsilon = 1$. By the (hypothetical) uniform continuity of f, there would exist a $\delta > 0$ such that $|f(x_2) - f(x_1)| < 1$ whenever $x_1, x_2 \in \mathbb{R}$ are such that $|x_2 - x_1| < \delta$. Now for example $x_1 = \frac{1}{\delta}$ and $x_2 = x_1 + \frac{\delta}{2} = \frac{1}{\delta} + \frac{\delta}{2}$ have the proximity $|x_2 - x_1| = \frac{\delta}{2} < \delta$ of the two arguments, so we should have $|f(x_2) - f(x_1)| < 1$. However, the difference of the values can be estimated by

$$f(x_2) - f(x_1) \;=\; \int_{x_1}^{x_2} f'(x)\,\mathrm{d}x \;=\; \int_{1/\delta}^{1/\delta+\delta/2} 2x\,\mathrm{d}x \;\geq\; \frac{\delta}{2}\,2/\delta \;=\; 1,$$

where we used $f'(x) = 2x \geq 2/\delta$ when $x \in [x_1, x_2] = [1/\delta,\ 1/\delta + \delta/2]$. Therefore we get a lower bound on the distance of the values, $|f(x_2) - f(x_1)| \geq 1$, contradicting the promised $|f(x_2) - f(x_1)| < 1$. The impossibility to provide a suitable δ for $\varepsilon = 1$ (for example) shows that f is not uniformly continuous.

This example shows in particular that not all continuous functions are uniformly continuous.

If the domain X is compact, however, continuity does imply uniform continuity.

Theorem XI.28 (Continuous functions on compacts are uniformly continuous)**.** *If X is a compact metric space and $f : X \to Y$ is a continuous function, then f is uniformly continuous.*

Proof. Assume that X is compact and $f : X \to Y$ is continuous. We will prove uniform continuity.

Let $\varepsilon > 0$. Observe that for any $y \in Y$, the subset

$$f^{-1}[\mathcal{B}_{\varepsilon/2}(y)] \;\subset\; X$$

is open, as the preimage of an open ball under the continuous function f. Any $x \in X$ belongs to a set of the above form: at least taking $y = f(x)$ we obviously have $x \in f^{-1}[\mathcal{B}_{\varepsilon/2}(y)]$. Therefore the collection

$$\big(f^{-1}[\mathcal{B}_{\varepsilon/2}(y)]\big)_{y \in Y}$$

is an open cover of X. Since X is compact, from Lebesgue number lemma (Lemma XI.20) we get that there exists a $\lambda > 0$ such that for all $x \in X$ there exists some $y \in Y$ such that $\mathcal{B}_\lambda(x) \subset f^{-1}[\mathcal{B}_{\varepsilon/2}(y)]$. We will check that the choice $\delta = \lambda$ works in the uniform continuity condition. So let $x_1, x_2 \in X$ be such that $\mathsf{d}_X(x_1, x_2) < \delta = \lambda$. By the above, there then exists some $y' \in Y$ such that $\mathcal{B}_\lambda(x_1) \subset f^{-1}[\mathcal{B}_{\varepsilon/2}(y')]$. Now we have $x_1, x_2 \in \mathcal{B}_\lambda(x_1)$, so we get $x_1, x_2 \in f^{-1}[\mathcal{B}_{\varepsilon/2}(y')]$, which means $f(x_1), f(x_2) \in \mathcal{B}_{\varepsilon/2}(y')$, or in other words, $\mathsf{d}_Y(f(x_1), y') < \varepsilon/2$ and $\mathsf{d}_Y(f(x_2), y') < \varepsilon/2$. It remains to note that by triangle inequality we then get

$$\mathsf{d}_Y\big(f(x_1), f(x_2)\big) \;\leq\; \underbrace{\mathsf{d}_Y\big(f(x_1), y'\big)}_{<\varepsilon/2} + \underbrace{\mathsf{d}_Y\big(y', f(x_2)\big)}_{<\varepsilon/2} \;<\; \varepsilon.$$

This proves the uniform continuity of f. ∎

♡ Application to Riemann-integrability

Since a closed interval $[a, b] \subset \mathbb{R}$ is compact, any continuous function $f : [a, b] \to \mathbb{R}$ is uniformly continuous by Theorem XI.28. This fact is at the heart of the proof that every continuous function f on a closed interval is Riemann-integrable, i.e., that the integral

$$\int_a^b f(t)\,\mathrm{d}t$$

can be meaningfully defined!

The proof of Riemann-integrability of a uniformly continuous function is not difficult starting from the characterization of Riemann-integrability given in Theorem B.7.

Lecture XII

Connectedness

Connectedness is a topological notion that has a very geometric flavor. The intuition about it is very easy: a space is connected if it doesn't consist of two (or more) "separate" pieces — see Definition XII.1 for the precise formulation of this.

There is a related, stronger notion of path-connectedness. The underlying idea is possibly even easier: a space is path-connected if between any two points in the space there exists a continuous path. We will see that in general path-connectedness implies connectedness but not vice versa. For open subsets of Euclidean spaces, the two are nevertheless equivalent. Also in some other sufficiently nice geometric contexts (such as manifolds and Lie groups, for example), connectedness and path-connectedness are equivalent, and path-connectedness is then used extensively in particular for integration along paths (contour integrals in complex analysis, 1-forms in differential geometry, etc.).

Connectedness, path-connectedness, and a related notion of simple connectedness can also be seen as small glimpses of algebraic topology; they concern global properties of (topological) spaces that would be encoded more systematically (and more generally) by homotopy and homology theories.

XII.1. Definition and examples of connectedness

It is actually easier to define the notion of *disconnectedness*, and then define *connectedness* by its negation.

Definition XII.1 (Disconnected and connected spaces).
A metric space X is **disconnected** if there exists two subsets $D_1, D_2 \subset X$ for which the following hold:

(i) $D_1 \cup D_2 = X$;
(ii) $D_1 \cap D_2 = \emptyset$;
(iii) $D_1 \neq \emptyset$ and $D_2 \neq \emptyset$;
(iv) both $D_1, D_2 \subset X$ are open subsets of X.

A metric space X is **connected** if it is not disconnected.

The following example of "breaking the real line in the middle" illustrates the notion of disconnectedness.

Example XII.2 (Punctured real line is disconnected).
Let $X = \mathbb{R} \setminus \{0\}$, with the metric inherited as a subset of $\mathbb{R}$. We will verify that X is disconnected.

Indeed, the positive real axis $D_1 = (0, \infty)$ and the negative real axis $D_2 = (-\infty, 0)$ are two subsets of $X = \mathbb{R} \setminus \{0\}$ such that (i): their union is the whole space X, (ii): they are disjoint, (iii): both of them are non-empty, and (iv): both of them are open subsets of X.

The previous example of real line without the origin actually had just one way of breaking into two separate pieces to show disconnectedness. In the following examples there are infinitely many ways to break into separate pieces.

Example XII.3 (The set of rational numbers is disconnected)**.**
The set $\mathbb{Q}$ of rational numbers is disconnected: it is easy to see that for example the sets $D_1 = (-\infty, \sqrt{2}) \cap \mathbb{Q}$ and $D_2 = (\sqrt{2}, \infty) \cap \mathbb{Q}$ satisfy the conditions in Definition XII.1.

Instead of $\sqrt{2}$, we could have chosen to "break" at any other irrational number, e.g., at π.

Example XII.4 (The set of irrational numbers is disconnected)**.**
The set $\mathbb{R} \setminus \mathbb{Q}$ of irrational numbers is disconnected: it is easy to see that for example the sets $D_1 = (-\infty, \frac{361}{42}) \setminus \mathbb{Q}$ and $D_2 = (\frac{361}{42}, \infty) \setminus \mathbb{Q}$ satisfy the conditions in Definition XII.1.

Instead of $\frac{361}{42}$, we could have chosen to "break" at any other rational number, e.g., at 0.

At this point, let us content ourselves with giving just trivial examples of connected spaces.

Example XII.5 (Spaces with fewer than two points are connected)**.**
If a space X has fewer than two points, it obviously cannot be disconnected: Definition XII.1 requires two disjoint non-empty "pieces" D_1 and D_2, and consequently at least two distinct points in X.

For this reason the empty set $\emptyset$ and all singletons $\{a\}$ are trivial examples of connected sets.[1]

Although we only provided trivial examples of connected spaces, one may anticipate that for example the real line $\mathbb{R}$ and intervals on it are connected. We will verify this later. Before that, however, we introduce some tools for working with the definition.

Theorem XII.6 (Characterizations of disconnectedness)**.**
Let X be a metric space. Then the following are equivalent:

(a) *X is disconnected;*
(b) *$X = C_1 \cup C_2$, where the subsets $C_1, C_2 \subset X$ are disjoint ($C_1 \cap C_2 = \emptyset$), closed, and non-empty ($C_1 \neq \emptyset$, $C_2 \neq \emptyset$);*
(c) *There exists a subset $D \subset X$ which is both open and closed (clopen), and neither empty nor equal to the whole space $\emptyset \neq D \neq X$;*
(d) *There exists a continuous surjection $f \colon X \to \{0, 1\}$.*

Proof. We show that (a) is equivalent with each of the other conditions by showing implications in both directions.

A key observation throughout is that when two sets $A_1, A_2 \subset X$ are disjoint ($A_1 \cap A_2 = \emptyset$) and cover the whole space ($A_1 \cup A_2 = X$), then they are the complements of each other, $A_1 = X \setminus A_2$ and $A_2 = X \setminus A_1$.

[1] The empty set and singletons have unique possible choices of metrics: in the empty set $\emptyset$ we don't have any points to define distances for, and in a singleton space $\{a\}$ the only distance to define is the distance from the single point a to itself, which has to be 0 by (M-0). Not terribly interesting. . .

proof of (a) ⇒ (b): Suppose that X is disconnected. Then there exist two subsets $D_1, D_2 \subset X$ satisfying the properties (i)–(iv) in Definition XII.1. Now let $C_1 = X \setminus D_2$ and $C_2 = X \setminus D_1$. Since D_2 and D_1 are open by (iv), and since C_1 and C_2 are their complements, the sets C_1 and C_2 are by definition closed. Using the observation that D_1 and D_2 are complements of each other, we furthermore get $C_1 = X \setminus D_2 = D_1$ and $C_2 = X \setminus D_1 = D_2$. Therefore these sets are non-empty, $C_1 = D_1 \neq \emptyset$ and $C_2 = D_2 \neq \emptyset$, they cover the space, $C_1 \cup C_2 = D_1 \cup D_2 = X$, and they are disjoint, $C_1 \cap C_2 = D_1 \cap D_2 = \emptyset$. This proves that they have the properties required in (b).

proof of (b) ⇒ (a): The idea of this implication is similar. We leave the details as an exercise.

proof of (a) ⇒ (c): The idea is again similar. Assuming disconnectedness, we obtain two non-empty, open, disjoint sets D_1 and D_2, which cover X. We can take $D = D_1$, and as before we observe that its complement is $X \setminus D = D_2$. Both D and its complement are non-empty, by (ii), so $\emptyset \neq D \neq X$. Also $D = D_1$ is open by (iv) and closed (since its complement $X \setminus D = D_2$ is open). Thus D satisfies the properties required in (c).

proof of (c) ⇒ (a): Once more, the idea is similar, and we leave the details as an exercise.

proof of (a) ⇒ (d): Suppose that X is disconnected. Then there exist two subsets $D_1, D_2 \subset X$ satisfying the properties (i)–(iv) in Definition XII.1. Since the sets cover the whole space and are disjoint, the formula

$$f\colon\ X \to \{0,1\} \qquad\qquad f(x) = \begin{cases} 1 & \text{if } x \in D_1 \\ 0 & \text{if } x \in D_2. \end{cases}$$

gives rise to a well-defined function $f\colon X \to \{0,1\}$.

Since $D_1 \neq \emptyset$, there exists a point $x_1 \in D_1 \subset X$ such that $f(x_1) = 1$, and since $D_2 \neq \emptyset$, there exists a point $x_2 \in D_2 \subset X$ such that $f(x_2) = 0$. This shows that f is surjective, with the two element set $\{0,1\}$ as its codomain.

To check the continuity of f, we use the characterization of continuity in Theorem VII.22: it suffices to show that the preimage of any open subset $U \subset \{0,1\}$ is open in X. There are only four possible subsets of $\{0,1\}$, namely $\emptyset$, $\{0\}$, $\{1\}$, and $\{0,1\}$. From the definition of f we easily see that their preimages are $\emptyset$, D_2, D_1, and X, respectively. Each of these subsets of X is open, so the continuity of f is established. This proves (d).

proof of (d) ⇒ (a): Assume (d), i.e., that there exists a continuous surjection $f\colon X \to \{0,1\}$. Define $D_1 = f^{-1}[\{1\}] \subset X$ and $D_2 = f^{-1}[\{0\}] \subset X$. By surjectivity, both D_1 and D_2 must be non-empty, and they are clearly the complements of each other, since the only possible values of f are 0 and 1. Also, observing that the singletons $\{1\} \subset \{0,1\}$ and $\{0\} \subset \{0,1\}$ are open in the discrete space $\{0,1\}$, from continuity of f we get that their preimages D_1 and D_2 are open. This shows that D_1 and D_2 have the properties required in the definition of disconnectedness. □

Connected components

A gluing lemma

One way of verifying connectedness is to exhibit connected subsets "glued together" at one common point.

Lemma XII.7 (Connectedness by gluing)**.**

Let X be a metric space and $(E_j)_{j\in J}$ a collection of connected subsets $E_j \subset X$, which have at least one common point, i.e., $\bigcap_{j\in J} E_j \neq \emptyset$. Then the union $\bigcup_{j\in J} E_j$ is also connected.

Proof. Suppose, by contrary, that the union $\bigcup_{j\in J} E_j$ is disconnected. By the characterizing property (d) in Theorem XII.6, there then exists a continuous surjection

$$f\colon \bigcup_{j\in J} E_j \to \{0,1\}\,.$$

By assumption, there exists some common point $p \in \bigcap_{j\in J} E_j$ of the sets E_j. Without loss of generality, we may assume $f(p) = 0$ (otherwise consider the continuous surjection $x \mapsto 1 - f(x)$ instead). By surjectivity, there exists also some point $q \in \bigcup_{j\in J} E_j$ such that $f(q) = 1$. By the definition of union, we have $q \in E_j$ for some $j \in J$. Since $p, q \in E_j$ and f attains the values 0 and 1 at these points, respectively, we have that the restriction $f|_{E_j}\colon E_j \to \{0,1\}$ is surjective. As the restriction of the continuous function f, this function $f|_{E_j}$ is also continuous. But the existence of such a continuous surjection $f|_{E_j}\colon E_j \to \{0,1\}$ means, by Theorem XII.6, that E_j is disconnected, which contradicts the assumption of its connectedness.

The contradiction proves that $\bigcup_{j\in J} E_j$ had to be connected. □

Definition of connected components

Definition XII.8 (Connected components)**.**
For $x \in X$, let $C_x \subset X$ be the union of all connected subsets $E \subset X$ that contain x,

$$C_x = \bigcup_{\substack{E\subset X \text{ s.t.}\\ x\in E\\ E \text{ connected}}} E. \tag{XII.1}$$

This set C_x is called the **connected component** of x in X.

Remark XII.9 (Connected components are connected)**.**
Since each set E in the union (XII.1) is connected and the sets contain at least the common point x, it follows from Lemma XII.7 that C_x is connected.

For $x, y \in X$, let us denote

$$x \sim y \qquad \text{if} \qquad y \in C_x, \tag{XII.2}$$

(and $x \not\sim y$ if $y \notin C_x$). We next check that this defines an equivalence relation in X.

Lemma XII.10 (An equivalence relation from connected components)**.**
The relation $\sim$ defined by (XII.2) *satisfies*

(R) *"reflexivity": for any $x \in X$, we have $x \sim x$;*
(S) *"symmetricity": for any $x, y \in X$, we have $x \sim y$ if and only if $y \sim x$;*
(T) *"transitivity": for any $x, y \in X$, if $x \sim y$ and $y \sim z$, then $x \sim z$.*

Proof. We prove the three properties separately.

proof of (R): Let $x \in X$. The singleton $\{x\} \subset X$ is connected by Example XII.5, so it appears in the union (XII.1) defining C_x. Therefore we have $\{x\} \subset C_x$, that is, $x \in C_x$.

proof of (S): Let $x, y \in X$. Let us first show that if $x \sim y$, then also $y \sim x$. So assume $x \sim y$, i.e., that $y \in C_x$. Observe that C_x is connected by Remark XII.9, and $x \in C_x$ by (R) above. Therefore the set $E = C_x$ itself appears in the union (XII.1). By connectedness of C_x and the assumption $y \in C_x$, we have that C_x also appears in the union defining C_y. Therefore we have $C_x \subset C_y$. This implies in particular $x \in C_x \subset C_y$ and we get $y \sim x$.

Exchanging the roles of x and y above, we similarly see that $y \sim x$ implies $x \sim y$, showing the if and only if form of (S). We note furthermore that if $x \sim y$, we thus in fact get both inclusions $C_x \subset C_y$ and $C_y \subset C_x$, and therefore the connected components agree, $C_x = C_y$.

proof of (T): Let $x, y, z \in X$. Assume $x \sim y$ and $y \sim z$, i.e., $y \in C_x$ and $z \in C_y$. As in the proof of (S), we deduce $C_x = C_y$ and $C_y = C_z$. In particular we see that $z \in C_z = C_x$, and we get $x \sim z$, proving transitivity. □

The equivalence classes for the equivalence relation $\sim$ are called the **connected components** of the space X.[2] The connected components form a partition of the space X to maximally large connected subsets.[3]

Remark XII.11 (Connectedness means having a single connected component)**.**
If a space X is connected, then from (XII.1) it is obvious that $C_x = X$ for every $x \in X$. So a connected space has just one connected component: the whole space X.

Conversely, if a space X has only one connected component, then this connected component is the whole space X (there is only one equivalence class in the partition). Since connected components are connected (why?), the whole space must then be connected.

As an example, it is intuitively clear that the punctured real line $\mathbb{R} \setminus \{0\}$ of Example XII.2 has two connected components: the negative half $(-\infty, 0)$ and the positive half $(0, +\infty)$. This intuitive idea was in fact how one in practice discovers the proof of disconnectedness of $\mathbb{R} \setminus \{0\}$. The disconnectedness proof did not, however, require us to show that the two halves are themselves connected. Making that part of the intuition precise requires some more tools, and we now turn to those.

Connected subsets of the real line

It turns out that, apart from trivial cases[4], the only connected subsets of the real line are intervals — and conversely, all intervals on the real axis are connected subsets. Here, by interval we mean open intervals, closed intervals, or half-open intervals, and we allow also unbounded intervals: examples of intervals include $[-1, 1]$, $[0, 2\pi)$, $(-\infty, 0]$, and the real line $\mathbb{R} = (-\infty, +\infty)$ itself.

Theorem XII.12 (Connected subsets of the real line)**.**
A subset $A \subset \mathbb{R}$ of the real line with $\#A \geq 2$ is connected if and only if A is an interval.

Proof. We prove the "if" and "only if" implications separately.

proof of "if": We first prove that a closed interval $[a, b]$ is connected; the case of a general interval will be reduced to this.

So let $a < b$ and consider the closed interval $[a, b] \subset \mathbb{R}$. We use a proof by contradiction to show that $[a, b]$ is connected. So suppose instead that $[a, b]$ is disconnected. By

[2] Note that the terminology is in line with Definition XII.8 in the following sense: each equivalence class $C \subset X$ of the relation $\sim$ is of the form $C = C_x$ for any $x \in X$. In other words, each connected component of the space (in the sense of the definition here) is the connected component of any of its members (in the sense of Definition XII.8).

[3] A small exercise: How to show that any connected component $C \subset X$ is necessarily connected?

[4] The trivial cases are the empty set $\emptyset \subset \mathbb{R}$ and all singletons $\{x\} \subset \mathbb{R}$, see Example XII.5.

characterization (b) of Theorem XII.6, there then exist two closed subsets $C_1, C_2 \subset [a, b]$, which are disjoint, $C_1 \cap C_2 = \emptyset$, non-empty, $C_1, C_2 \neq \emptyset$, and for which $C_1 \cup C_2 = [a, b]$. The closed interval $[a, b]$ is compact (Example XI.3) and $C_1, C_2 \subset [a, b]$ are closed subsets, also C_1 and C_2 are compact by Theorem XI.8. By Exercise XI.9 there then exists points $c_1 \in C_1$ and $c_2 \in C_2$ so that $\mathrm{dist}(C_1, C_2) = \mathsf{d}_{\mathbb{R}}(c_1, c_2)$. By disjointness we have $c_1 \neq c_2$, and consequently $|c_2 - c_1| = \mathsf{d}_{\mathbb{R}}(c_1, c_2) > 0$. Now consider $z = \frac{c_1+c_2}{2}$. We have

$$\mathsf{d}_{\mathbb{R}}(c_1, z) \;=\; |z - c_1| \;=\; \left|\frac{c_2 - c_1}{2}\right| \;=\; \frac{1}{2}\,\mathsf{d}_{\mathbb{R}}(c_1, c_2) \;=\; \frac{1}{2}\,\mathrm{dist}(C_1, C_2),$$

which implies $z \notin C_2$, because for any $x_2 \in C_2$ we would have

$$\mathsf{d}_{\mathbb{R}}(c_1, x_2) \;\geq\; \mathrm{dist}(C_1, C_2) \;>\; \frac{1}{2}\,\mathrm{dist}(C_1, C_2).$$

A similar calculation of $\mathsf{d}_{\mathbb{R}}(c_2, z)$ shows that $z \notin C_1$. Together these show that $z \notin C_1 \cup C_2 = [a, b]$, which is a contradiction, because z is the midpoint of the two points $c_1 \in C_1 \subset [a, b]$ and $c_2 \in C_2 \subset [a, b]$ on the closed interval $[a, b]$. This contradiction proves the connectedness of $[a, b]$.

Let us then consider the case of a general interval $A \subset \mathbb{R}$ with $\#A \geq 2$. Pick a point $z \in A$ (possible since $A \neq \emptyset$). For any $x \in A$, let E_x be the closed interval with endpoints x and z, specifically

$$E_x \;=\; \begin{cases} [x, z] & \text{if } x < z \\ \{z\} & \text{if } x = z \\ [z, x] & \text{if } x > z. \end{cases}$$

Then for every $x \in A$ we have that E_x is connected (as a closed interval, by the first part of the proof), and $z \in E_x$ (by construction). Applying the gluing as in Lemma XII.7, we get that $\bigcup_{x \in A} E_x$ is connected. We now claim that $A = \bigcup_{x \in A} E_x$, which allows us to conclude that the interval A is connected. Clearly we have $A \subset \bigcup_{x \in A} E_x$ (since any $x \in A$ belongs to at least E_x) and on the other hand $\bigcup_{x \in A} E_x \subset A$ since $E_x \subset A$ for every x (the interval A contains the closed subinterval between its two points x and z). The combination of these inclusions shows the desired equality $A = \bigcup_{x \in A} E_x$.

proof of "only if": Assume that $A \subset \mathbb{R}$ is connected and $\#A \geq 2$. Let $a = \inf A$ and $b = \sup A$. To show that $A \subset \mathbb{R}$ is an interval, it suffices to show that $(a, b) \subset A$; then whether a and b belong to A only affects the type of the interval (open, closed, or half-open), but not the fact that it is an interval.[5]

We use a proof by contradiction to show that $(a, b) \subset A$, so assume instead that $(a, b) \not\subset A$. Then there exists some $z \in (a, b)$ such that $z \notin A$. But now $D_1 = (-\infty, z) \cap A$ and $D_2 = (z, \infty) \cap A$ are open (by Theorem VII.9), non-empty (since $\inf A = a < z < b = \sup A$) disjoint, and we have $A = D_1 \cup D_2$. This by definition shows that A is disconnected, which contradicts the assumption that A is connected. Therefore we actually have that $(a, b) \subset A$, and the proof is complete. □

XII.2. Preservation of connectedness under continuous mappings

Theorem XII.13 (Connectedness is preserved under continuous mappings)**.** *Let X and Y be metric spaces and $f : X \to Y$ a continuous function. If $A \subset X$ is connected, then the image $f[A] \subset Y$ is also connected.*

Proof. We prove this by the contrapositive. So suppose that $f[A] \subset Y$ is disconnected. By characterization (d) of Theorem XII.6, there then exists a continuous surjection $g : f[A] \to \{0, 1\}$.

[5]The cases $a = -\infty$ and $b = +\infty$ are possible by the convention regarding the infimum of a set not bounded from below and supremum of a set not bounded from above. The reasoning is, however, exactly the same also in these cases.

The restriction $f|_A : A \to Y$ can be viewed, by restriction of the codomain, as a function $A \to f[A]$, which is by construction surjective. The continuity is not affected by this restriction of the codomain (convince yourself about this!), so $f|_A : A \to f[A]$ is continuous.[6] Now the composition of the continuous surjections $f|_A : A \to f[A]$ and $g : f[A] \to \{0,1\}$ is a continuous surjection $g \circ f : A \to \{0,1\}$ (see Exercise I.10 and Theorem VII.29). By characterization (d) of Theorem XII.6, the set A is then disconnected. This finishes the proof by contrapositive. □

Exercise XII.1 (Cartesian products of disconnected spaces)**.**
Let X and Y be two non-empty metric spaces. Show that if either X or Y is disconnected, then the Cartesian product $X \times Y$ is also disconnected.

XII.3. Path-connectedness

Paths

Let X be a metric space. A **path** in X is a continuous function

$$\gamma \colon [a,b] \to X$$

from a closed interval $[a,b] \subset \mathbb{R}$ to the space X. We often use the unit interval $[0,1]$ in this context[7], but for some purposes it is convenient to allow an arbitrary closed interval here.

If you think of the argument $t \in [a,b]$ of a path $\gamma \colon [a,b] \to X$ as "time", then it is natural to view $\gamma(a) \in X$ as the *starting point* of the path, and $\gamma(b) \in X$ as its *end point*, and generally $\gamma(t) \in X$ as the point along the path at time $t \in [a,b]$.

From Theorems XII.13 and XII.12 it follows that the image $\gamma[[a,b]] \subset X$ of a path γ is a connected subset of the space X.

Let us give some examples of paths.

Example XII.14 (A path tracing the unit circle)**.**
The formula

$$\gamma \colon [0,2\pi] \to \mathbb{R}^2 \qquad \gamma(t) = \big(\cos(t),\ \sin(t)\big)$$

defines a path in the Euclidean plane $\mathbb{R}^2$, which traces the unit circle in the counterclockwise direction.[8] It starts and ends at the same point, $\gamma(0) = \gamma(2\pi) = (1,0) \in \mathbb{R}^2$.

Example XII.15 (Line segments in normed spaces)**.**
Let V be a normed space and $u, v \in \mathsf{V}$ two points. The formula

$$\gamma \colon [0,1] \to \mathsf{V} \qquad \gamma(t) = u + t\,(v-u)$$

defines a path in V, from u along a straight line to v. An easy calculation gives that γ is $\|v-u\|$-Lipschitz, verifying the continuity of γ, in particular.

[6] We are mildly abusing the notation here, as we use the same name for the function after changing the codomain.

[7] Using only the unit interval in the definition of a path would in fact be a totally benign change. Indeed, for other choices $a < b$, a path $\gamma \colon [a,b] \to X$ naturally gives rise to a "reparametrized path" $\widetilde{\gamma} \colon [0,1] \to X$ by the formula $\widetilde{\gamma}(t) = \gamma\big(a + t(b-a)\big)$, which is also continuous as the composition of the continuous function γ with the first degree polynomial $t \mapsto a + t(b-a)$ (whose continuity follows from Corollary III.12).

[8] The continuity of this $\gamma \colon [0,2\pi] \to \mathbb{R}^2$ follows (by Exercise VI.10) from the continuity of the component functions, which are trigonometric functions (Lemma IX.10).

The image $\gamma[[0,1]] \subset \mathsf{V}$ of this path is called the **line segment** from u to v and denoted also by $[u,v] \subset \mathsf{V}$.[9]

Convexity

Note that in general metric spaces there is no such thing as a straight line segment, but in normed spaces the vector addition and scalar multiplication made it possible to define the line segment path in Example XII.15. The notion of straight line segments makes the following concept meaningful: A subset $A \subset \mathsf{V}$ in a normed space V is **convex** if for all $x, y \in A$ we have $[x,y] \subset A$.

Example XII.16 (Balls in normed spaces are convex).
Let V be a normed space, $w \in \mathsf{V}$ a point, and $r > 0$. We claim that the open ball $\mathcal{B}_r(w)$ and the closed ball $\overline{\mathcal{B}}_r(w)$ are convex subsets of V.

Let us consider the case $w = 0 \in \mathsf{V}$ and the closed ball $\overline{\mathcal{B}}_r(0)$; we leave the details of the other cases as an exercise. So let $u, v \in \overline{\mathcal{B}}_r(0)$. By definition of the closed ball, this means that $\|u\| = \mathsf{d}(0,u) \leq r$ and $\|v\| = \mathsf{d}(0,v) \leq r$. A general point in the line segment $[u,v] \subset \mathsf{V}$ is of the form $u + t(v-u)$ for some $t \in [0,1]$. Using the triangle inequality we get

$$\begin{aligned}\|u + t(v-u)\| \;=\; \big\|t\,v + (1-t)\,u\big\| \;&\leq\; \|t\,v\| \,+\, \|(1-t)\,u\| \\ &=\; |t|\,\|v\| \,+\, |1-t|\,\|u\| \\ &\leq\; t\,r + (1-t)\,r \\ &=\; r.\end{aligned}$$

This shows $u + t(v-u) \in \overline{\mathcal{B}}_r(0)$, proving convexity of $\overline{\mathcal{B}}_r(0)$.

Exercise XII.2 (Completing the proof of convexity of balls).
Complete the details for the remaining cases in Example XII.16 to show that all closed balls and open balls in a normed space are convex.

Exercise XII.3 (Rectangles are convex).
Show that a rectangle $[a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$ is convex.

Path-connectedness

Using paths, we get a very intuitive notion related to connectedness.

Definition XII.17 (Path connectedness).
A metric space X is **path-connected** if for all $x, y \in X$ there exists a path $\gamma\colon [a,b] \to X$ in X such that $\gamma(a) = x$ and $\gamma(b) = y$.

Example XII.18 (Normed spaces are path-connected).
Let V be a normed space. For any two points $u, v \in \mathsf{V}$, the straight line segment of Example XII.15 is a path γ from $\gamma(0) = u$ to $\gamma(1) = v$. The existence of such paths shows that any normed space is path-connected.

In particular the Euclidean spaces $\mathbb{R}^d$ are path-connected.

[9] In the normed space $(\mathbb{R}, |\cdot|)$, line segments are indeed just closed intervals, so this notation is reasonable.

Also for example the space $(\mathcal{C}([a,b]), \|\cdot\|_\infty)$ is path-connected, and so are all other examples of normed spaces we have seen (and also those that we have not seen...[10]).

Example XII.19 (Convex subsets of normed spaces are path-connected).
Let V be a normed space and $A \subset \mathsf{V}$ a convex subset. By definition of convexity, for any two points $u, v \in \mathsf{V}$, the straight line segment of Example XII.15 is a path γ from $\gamma(0) = u$ to $\gamma(1) = v$ in A. The existence of such paths shows the path-connectedness of convex sets.

In particular with Example XII.16 and Exercise XII.2 we get that all open and closed balls in normed spaces are path-connected.

Theorem XII.20 (Path connectedness implies connectedness).
If a metric space X is path-connected, then it is connected.

Proof. Assume that X is path-connected. We use a proof by contradiction to show the connectedness of X. So assume that X is instead disconnected. By characterization (d) of Theorem XII.6, there then exists a continuous surjection $f\colon X \to \{0,1\}$. By surjectivity there exist points $x, y \in A$ such that $f(x) = 0$ and $f(y) = 1$. By the assumed path-connectedness there exists a continuous $\gamma\colon [a,b] \to X$ such that $\gamma(a) = x$ and $\gamma(b) = y$. Consider now the composition $f \circ \gamma\colon [a,b] \to \{0,1\}$. As a composition of continuous functions $f \circ \gamma$ is continuous (Theorem VII.29). Since $(f \circ \gamma)(a) = f(\gamma(a)) = f(x) = 0$ and $(f \circ \gamma)(b) = f(\gamma(b)) = f(y) = 1$, it is also surjective onto $\{0,1\}$. But the existence of a continuous surjection $f \circ \gamma\colon [a,b] \to \{0,1\}$ would imply, by characterization (d) of Theorem XII.6, that $[a,b]$ is disconnected, contradicting Theorem XII.12. This contradiction proves that a path-connected space X must be connected. $\square$

Applying Theorem XII.20 to the path-connected spaces in Examples XII.18 and XII.19 immediately yields the following.

Corollary XII.21 (Normed spaces and their convex subsets are connected).
Let V be a normed space. Then V is connected. Moreover, any convex subset $A \subset \mathsf{V}$ is connected.

Combined with Example XII.16, the corollary above implies that open and closed balls in normed spaces are connected. In general metric spaces, one cannot obtain the same conclusion.

Exercise XII.4 (Balls may be disconnected).
Give an example of a metric space (X, d) and an open ball $\mathcal{B}_r(z) \subset X$ which is not connected.

While path-connectedness implies connectedness by Theorem XII.20, the converse implication is not generally true. To find a counterexample, it is useful to first prove that closure preserves connectedness.

Exercise XII.5 (The closure of a connected set is connected).
Let X be a metric space and $A \subset X$ a connected subset. Prove that then the closure $\overline{A}$ is also connected.

Example XII.22 (The topologist's sine curve).
Let $f\colon (0,\infty) \to \mathbb{R}$ be the function given by $f(x) = \sin(1/x)$ for $x \in (0,\infty)$. Consider the

[10] To give an example of a normed space we have not seen, let us ... Oh wait! But then we will have seen it! So let us leave that example for a subsequent course or your own creativity.

graph of this function, i.e., the subset

$$A = \{(x, y) \in \mathbb{R}^2 \mid y = f(x)\}$$

in the plane $\mathbb{R}^2$.

(a) Prove that $A \subset \mathbb{R}^2$ is path-connected.
(b) Use Exercise XII.5 to show that that $\overline{A} \subset \mathbb{R}^2$ is connected.
(c) Prove that $\overline{A} \subset \mathbb{R}^2$ is not path-connected.
Hint: *What can you say about the two coordinates of points that belong to $\overline{A} \setminus A$? What could you say about the two coordinates along a continuous path which would connect a point of $\overline{A} \setminus A$ to a point of A in $\overline{A}$? Is it possible?*

While connectedness does not in general imply path-connectedness, the situation is better with open subsets of Euclidean spaces. Connected open sets in Euclidean spaces are often called domains.

Theorem XII.23 (Connected open sets in Euclidean spaces are path connected)**.** *Let $U \subset \mathbb{R}^d$ be an open subset of the Euclidean space $\mathbb{R}^d$. Then U is connected if and only if it is path-connected.*

Proof. The "if" part follows readily from Theorem XII.20. We leave the "only if" part as Exercise XII.6. □

Exercise XII.6 (Connectedness by broken lines)**.**
A *broken line* in $\mathbb{R}^n$ through points $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell$ is the union of line segments

$$M = [\vec{x}_1, \vec{x}_2] \cup [\vec{x}_2, \vec{x}_3] \cup \cdots \cup [\vec{x}_{\ell-1}, \vec{x}_\ell] \subset \mathbb{R}^n.$$

A broken line M of this form is said to connect the points $\vec{x}_1$ and $\vec{x}_\ell$.

Prove the following: If $D \subset \mathbb{R}^n$ is open and connected, then for all points $\vec{x}, \vec{y} \in D$ there exists a broken line $M \subset D$ connecting the points $\vec{x}$ and $\vec{y}$. Using this, complete the proof of Theorem XII.23.
Hint: *Fix some point in D. Consider the set $U \subset D$ of points that can be connected to it by a broken line in D. Show that both U and $D \setminus U$ are open. Conclude using the assumed connectedness of D.*

♡ Simple connectedness

A notion of "holes" in topological spaces can also be defined in terms of paths. For example, the plane $\mathbb{R}^2$ and the punctured plane $\mathbb{R}^2 \setminus \{\vec{0}\}$ are both connected and even path-connected, but there is clearly something topologically different about them: the latter has a "hole" at the origin. It can be shown, indeed, that the plane and the punctured plane are not homeomorphic. The standard way to prove this is by showing that the former is simply connected ("has no holes") whereas the latter is not.

We briefly introduce the notion of simple connectedness here for two reasons. First, it shows how to formulate the intuitive notion of holes in topological spaces in a mathematically precise way. Second, simple connectedness is one of the most rudimentary notions in algebraic topology, so it provides at least a glimpse of an important viewpoint to topology, which has so far been entirely missing in this introductory course of point-set topology.

For simple connectedness, one considers not just any paths, but “loops”: paths that return to their starting point. Moreover, the paths are considered up to continuous deformation (homotopy), i.e., two paths in a space that can be continuously deformed to each other are considered (homotopy) equivalent. Roughly speaking, a “hole” is detected by finding a loop around the hole, i.e., a loop that can not be continuously deformed to a single point.

By the precise definition, a **loop**[11] in X is a continuous function

$$\gamma\colon \mathbf{S}^1 \to X$$

from the unit circle $\mathbf{S}^1 = \left\{ \vec{u} \in \mathbb{R}^2 \;\middle|\; \mathsf{d}(\vec{0}, \vec{u}) = 1 \right\}$ to the space X. Note that the definition is entirely similar to that of paths, except the “time parametrization” is not by a closed interval but by the unit circle (one periodically returns to the starting point when traveling along a loop).

A continuous deformation of a loop to a constant loop can be viewed as follows. A continuous function

$$\Gamma\colon \overline{\mathcal{B}}_1(\vec{0}) \to X$$

from the closed unit disc $\overline{\mathcal{B}}_1(\vec{0}) \subset \mathbb{R}^2$ to the space X gives rise to a parametrized family of loops γ_s, $s \in [0,1]$, via

$$\gamma_s\colon \mathbf{S}^1 \to X \qquad\qquad \gamma_s(\vec{u}) = \Gamma(s\vec{u}) \quad \text{ for } \vec{u} \in \mathbf{S}^1.$$

By continuity of Γ, the loops γ_s depend continuously on the deformation parameter $s \in [0,1]$. At deformation parameter $s = 1$, we have the loop $\gamma_1 = \Gamma|_{\mathbf{S}^1} \colon \mathbf{S}^1 \to X$ obtained by restricting Γ to the unit circle $\mathbf{S}^1 \subset \overline{\mathcal{B}}_1(\vec{0})$. At deformation parameter $s = 0$, we have the loop γ_0, $\vec{u} \mapsto \Gamma(\vec{0})$, which just stays constant at the point $\Gamma(\vec{0}) \in X$.

A loop $\gamma\colon \mathbf{S}^1 \to X$ is said to be **contractible** (or **null homotopic**) if there exists a continuous function $\Gamma\colon \overline{\mathcal{B}}_1(\vec{0}) \to X$ such that $\Gamma|_{\mathbf{S}^1} = \gamma$. The interpretation is that a continuous deformation to a point is possible.

Definition XII.24 (Simple connectedness)**.**
A space X is **simply connected** if every loop in X is contractible.

Recall the idea that a hole could be detected by a loop around it, which cannot be continuously shrunk to a single point within the space — simple connectedness should be interpreted as saying that no such holes exist; that every loop can be shrunk.

Example XII.25 (Convex sets are simply connected)**.**
Let $A \subset \mathsf{V}$ be a convex subset of a normed space V, and let $\gamma\colon \mathbf{S}^1 \to A$ be a loop in A. Fix an arbitrary $\vec{u}_0 \in \mathbf{S}^1$, for example $\vec{u}_0 = (1,0)$. The formula

$$\Gamma(s\vec{u}) = (1-s)\gamma(\vec{u}_0) + s\gamma(\vec{u}) \qquad \text{for } s \in [0,1],\ \vec{u} \in \mathbf{S}^1,$$

defines a function $\Gamma\colon \overline{\mathcal{B}}_1(\vec{0}) \to A$ (the right-hand side is a convex combination of points on the path γ in A, thus remains in the convex set A), which is moreover continuous (we leave the detailed proof as an exercise). By construction, $\Gamma|_{\mathbf{S}^1} = \gamma$, so this shows that the loop γ

[11] It is common to use the term “closed path” instead of “loop”, but we avoid this terminology in order to avoid confusion with the unrelated and even more standard notion of closed sets.

in A is contractible. Since this holds for an arbitrary loop in A, we conclude that the convex set A is simply connected.

Example XII.26 (Annulus).
Let $0 < r < R$. Consider the annulus

$$\mathbf{A}_{r,R} = \left\{ \vec{v} \in \mathbb{R}^2 \;\middle|\; r < \|v\| < R \right\}.$$

For any $\rho \in (r, R)$, one can define a loop $\gamma : \mathbf{S}^1 \to \mathbf{A}_{r,R}$ by the formula $\gamma(\vec{u}) = \rho\vec{u}$. This loop γ winds around the annulus, and it can be shown that γ is not contractible (proving its non-contractibility is not entirely trivial, however!). Therefore the annulus is not simply connected.

Example XII.27 (Torus).
Consider the torus surface $\mathbf{S}^1 \times \mathbf{S}^1$ (see Figure V.3(c)). It is intuitively clear that for any $\vec{w}_1 \in \mathbf{S}^1$, a loop of the form $\mathbf{S}^1 \ni \vec{u} \mapsto (\vec{w}_1, \vec{u}) \in \mathbf{S}^1 \times \mathbf{S}^1$ winds around the torus, and is not contractible (although proving this is again non-trivial). Likewise, for any $\vec{w}_2 \in \mathbf{S}^1$, a loop of the form $\vec{u} \mapsto (\vec{u}, \vec{w}_2)$ winds around the torus differently, and is also not contractible. If you think of the torus as the surface of a donut, then one of these loops winds around the empty hole of the donut, whereas the other winds around the dough-filled "hole". Since such non-contractible loops exist, the torus is not simply connected.

A key topological idea is that contractibility of loops is preserved under continuous maps.

Lemma XII.28 (A continuously mapped contractible loop is contractible).
Suppose that X and Y are topological spaces, $f \colon X \to Y$ is a continuous function, and $\gamma \colon \mathbf{S}^1 \to X$ is a contractible loop in X. Then $f \circ \gamma \colon \mathbf{S}^1 \to Y$ is a contractible loop in Y.

Proof. Since γ is contractible, there exists a continuous function $\Gamma \colon \overline{\mathcal{B}}_1(\vec{0}) \to X$ such that $\Gamma|_{\mathbf{S}^1} = \gamma$. Then the function

$$f \circ \Gamma : \overline{\mathcal{B}}_1(\vec{0}) \to X$$

is continuous as the composition of the continuous functions Γ and f, and moreover its restriction to the unit circle is $(f \circ \Gamma)|_{\mathbf{S}^1} = f \circ \gamma$. This shows that $f \circ \gamma$ is contractible. □

Theorem XII.29 (Simple connectedness is a topological property).
Suppose that X and Y are metric spaces which are homeomorphic, $X \approx Y$. Then X is simply connected if and only if Y is simply connected.

Proof. Fix a homeomorphism$f \colon X \to Y$ from X to Y.

Note that the situation is symmetric under exchanging the two spaces, so it suffices to prove one implication and the other is obtained by exchanging the roles of the spaces (and considering the homeomorphism $f^{-1} \colon Y \to X$). We will prove that if Y is not simply connected, then X is not simply connected either.

So assume that Y is not simply connected, i.e., that there exists a non-contractible loop $\eta \colon \mathbf{S}^1 \to Y$ in Y. Assume now, by contradition, that X would nevertheless be simply connected. Consider the loop $\gamma = f^{-1} \circ \eta \colon \mathbf{S}^1 \to X$ in X obtained as its image of η under the inverse $f^{-1} \colon Y \to X$ of the homomorphism f. Having assumed that X is simply connected, we get that γ is contractible. By Lemma XII.28, mapping this loop byt the continuous f, we obtain a contractible loop $f \circ \gamma = f \circ f^{-1} \circ \eta = \eta$. But η was non-contractible, yielding a contradiction. The asserted implication is thus proven, and the reverse direction is symmetric. □

In particular, if one proves that the unit circle loop $\gamma\colon \mathbf{S}^1 \to \mathbb{R}^2 \setminus \{\vec{0}\}$ in the punctured plane $\mathbb{R}^2 \setminus \{\vec{0}\}$ is not contractible, then one immediately obtains as a corollary the non-homeomorphism result $\mathbb{R}^2 \setminus \{\vec{0}\} \not\approx \mathbb{R}^2$, since the full plane $\mathbb{R}^2$ is simply connected (this follows from Example XII.25).

One reason why these developments naturally lead towards algebraic topology is that suitable equivalence classes of loops (with a fixed starting and end point) under an equivalence accounting for similar deformation (homotopy) form a *group*[12] called the *fundamental group of the space*. Slightly generalizing Theorem XII.29, one could show that homeomorphic spaces have isomorphic fundamental groups. Simple connectedness, or "no holes", corresponds to the fundamental group being a trivial group. More generally, the (isomorphism class of the) fundamental group can be seen as "the right way to count holes" of a topological space.

[12] Groups are one of the most important algebraic structures. We refer to courses in abstract algebra for a proper mathematical treatment and more details.

Appendix A

Further topics in set theory

A.1. ♡ Cardinalities of sets

An observation about finite sets

Let us denote the number of elements of a set A by $\#A$.

Example A.1 (Number of elements in a few example sets)**.**

- For the empty set, we have $\#\emptyset = 0$.
- For a singleton, we have $\#\{a\} = 1$.
- We have $\#\{0, 1, 2, \ldots, 8, 9\} = 10$.
- Let $c_0, c_1, \ldots, c_{d-1}, c_d \in \mathbb{R}$ with $c_d \neq 0$, so that $p(x) = c_d\, x^d + c_{d-1}\, x^{d-1} + \cdots + c_1\, x + c_0$ is a polynomial of degree $d \in \mathbb{N}_0$. Then for the set of real zeroes of p we have
$$\# \left\{ x \in \mathbb{R} \;\middle|\; p(x) = 0 \right\} \;\leq\; d.$$

If a set A has infinitely many different elements, we denote $\#A = \infty$.

Example A.2 (Some infinite sets)**.**

The sets of natural numbers, integers, rational numbers, and real numbers are infinite
$$\#\mathbb{N} = \infty, \qquad \#\mathbb{Z} = \infty, \qquad \#\mathbb{Q} = \infty, \qquad \#\mathbb{R} = \infty.$$

Also nontrivial intervals are infinite, for example $\#(-\pi, \pi] = \infty$.

Moreover, as we will verify later, a non-empty open interval $(a, b) \subset \mathbb{R}$ contains infinitely many rational numbers and infinitely many irrational numbers
$$\#\Big((a, b) \cap \mathbb{Q}\Big) \;=\; \infty, \qquad \#\Big((a, b) \setminus \mathbb{Q}\Big) \;=\; \infty.$$

A set A is called **finite** if $\#A \in \{0, 1, 2, \ldots\}$, and **infinite** if $\#A = \infty$. It turns out that among infinite sets, some are nevertheless larger than others. The notion of *cardinality* captures this. To motivate the definition, let us nevertheless begin by some observations about sizes of finite sets.

Observation A.3 (Comparing sizes of sets using surjective functions)**.**

If A, B are two finite sets and $B \neq \emptyset$, then the following are equivalent:[1]

- $\#A \geq \#B$
- there exists a surjective function $A \to B$.[2]

[1] For the empty set $B = \emptyset$, the comparison of sizes has to be handled separately (the empty set is smaller than any other set), since from a nonempty set $A \neq \emptyset$ there does not exist any functions $A \to \emptyset$ — let alone surjective functions.

[2] The idea is the following. Suppose that A is the set of all students of this course, and B is the set of all exercise groups. Every student is assigned to exactly one exercise group, so that the assignment defines a function $A \to B$. The function is surjective if every exercise group has at least one student. The gist of this observation is that in such a case we can conclude that there are

Cardinality comparison and equal cardinalities

In view of the above observation, it appears meaningful to consider a set A at least as large as a set $B \neq \emptyset$ if there exists a surjective function $A \to B$. In this case we denote $A \succeq B$ — or interchangeably $B \preceq A$. This is the comparison of **cardinalities** of sets. We may observe the following properties, which match with our intuition about sizes of sets:

- If $A \succeq B$ and $B \succeq C$, then $A \succeq C$.[3]
- If $B \subset A$ is a nonempty subset, then $A \succeq B$.[4]

Example A.4 (Cardinality comparisons of some infinite sets).
Because of the subset relations $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$, we have the cardinality comparisons $\mathbb{N} \preceq \mathbb{Z} \preceq \mathbb{Q} \preceq \mathbb{R}$.

We say that two nonempty sets A and B have **equal cardinalities** if $A \succeq B$ and $A \preceq B$. By Observation A.3, (nonempty) finite sets have equal cardinalities if and only if they have the same number of elements. Let us then look at some examples with infinite sets.

Example A.5 (Equal cardinality for a set and its proper subsets).
Consider the set $\mathbb{N}$ of natural numbers, and the set $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ of nonnegative integers,

$$\begin{aligned} \mathbb{N} &= \{1, 2, 3, 4, \ldots\} \\ \mathbb{N}_0 &= \{0, 1, 2, 3, 4 \ldots\} . \end{aligned}$$

Because of the subset relation $\mathbb{N} \subset \mathbb{N}_0$, we have $\mathbb{N} \preceq \mathbb{N}_0$. On the other hand, we also have a surjective function $f \colon \mathbb{N} \to \mathbb{N}_0$ given by $f(n) = n - 1$ for $n \in \mathbb{N}$, so that $\mathbb{N} \succeq \mathbb{N}_0$ holds, too. Therefore the sets $\mathbb{N}$ and $\mathbb{N}_0$ have equal cardinalities.

In particular an infinite set can have a proper subset with equal cardinality: the set $\mathbb{N}_0$ has one extra element compared to $\mathbb{N}$ (namely zero), but this does not affect the “size” of the set; these two infinite sets have equal cardinalities.

Example A.6 (Another equal cardinality for a set and its proper subsets).
Consider the set $\mathbb{N}$ of natural numbers and the set $A = \{2, 4, 6, \ldots\}$ of even natural numbers. Because of the subset relation $A \subset \mathbb{N}$, we have $A \preceq \mathbb{N}$. On the other hand, we also have a surjective function $f \colon A \to \mathbb{N}$ given by $f(m) = \frac{1}{2}m$ for $m \in A$, so that $A \succeq \mathbb{N}$ holds, too. Therefore the sets $\mathbb{N}$ and A have equal cardinalities.

The set $\mathbb{N}_0$ has infinitely many extra elements compared to A (namely the odd natural numbers), but this does not affect the “size” of the set; these two infinite sets have equal cardinalities.

---

at least as many students as there are exercise classes (and conversely: an assignment that leaves no exercise group empty exists if there are at least as many students as there are exercise classes).

[3]Recall Exercise I.10: if $f \colon A \to B$ and $g \colon B \to C$ are surjective, then also the composition $g \circ f \colon A \to C$ is surjective.

[4]Indeed, choose some element $b_0 \in B$ of the nonempty set B, and define a function f from A to the subset $B \subset A$ by

$$f(a) = \begin{cases} a & \text{if } a \in B \subset A \\ b_0 & \text{if } a \in A \setminus B. \end{cases}$$

Then f is surjective.

Exercise A.1 (The sets of natural numbers and integers have equal cardinalities).
Prove that we have $\mathbb{N} \preceq \mathbb{Z}$ and $\mathbb{N} \succeq \mathbb{Z}$.
Hint: *One direction is obvious; for the other, construct some surjectve function* $\mathbb{N} \to \mathbb{Z}$.

Countable and uncountable infinite sets

These examples show that many common infinite sets have the same cardinality as the set of natural numbers. This motivates the following definition.

Definition A.7 (Countable and uncountable infinite sets).
An infinite set A is **countably infinite** if $A \preceq \mathbb{N}$, i.e., if there exists a surjective function $f : \mathbb{N} \to A$; otherwise A is **uncountably infinite**.

Remark A.8 (Enumeration of elements of a countably infinite set).
If A is a countably infinite set, then it is possible to list or enumerate all elements of A as follows. If $f : \mathbb{N} \to A$ is a surjective function, then we may form a sequence

$$(a_1, a_2, a_3, \ldots)$$

with $a_n = f(n)$ for $n \in \mathbb{N}$. By surjectivity, each element of A appears at least once in this sequence.

It is not possible to enumerate elements of uncountably infinite sets in this way!

In Examples A.5 and A.6 and Exercise A.1 we have seen some examples of countably infinite sets. The most important example of an uncountably infinite set is the set of real numbers.

Theorem A.9 (The set of real numbers is uncountably infinite).
The set $\mathbb{R}$ of real numbers is uncountably infinite.

The proof is given in the next subsection.

By contrast, and perhaps surprisingly, the set of rational numbers is in fact countable. In this sense the set $\mathbb{Q}$ of rational numbers is much smaller than the set $\mathbb{R}$ of real numbers.

Theorem A.10 (The set of rational numbers is countably infinite).
The set $\mathbb{Q}$ of rational numbers is countably infinite.

Countability of the set of rational numbers

The key to the proof of Theorem A.10 is the following frequently useful lemma.

Lemma A.11 (The Cartesian product of two countable sets is countable).
Suppose that A and B are two countably infinite sets. Then their Cartesian product $A \times B$ is also countably infinite.

It is still clearest to prove this lemma by first explicitly addressing the following special case.

Lemma A.12 (The set of pairs of natural numbers is countable)**.** *The set* $\mathbb{N} \times \mathbb{N} = \{(n,m) \mid n \in \mathbb{N},\ m \in \mathbb{N}\}$ *of pairs of natural numbers is countably infinite.*

Proof. We must show that there exists a surjective function $t\colon \mathbb{N} \to \mathbb{N} \times \mathbb{N}$. This can be done in various ways; the enumeration illustrated in Figure A.1 is easy to visualize, but let us choose a different construction. Namely, the function $t\colon \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ given by

$$t(n) = \begin{cases} (k,\ell) & \text{if } n = 2^{k-1}\,3^{\ell-1} \text{ for } k,\ell \in \mathbb{N} \\ (1,1) & \text{if } n \text{ contains other prime factors besides 2 and 3} \end{cases}$$

is well-defined because of the unique prime factorization of natural numbers, and it is surjective, since any $(k,\ell) \in \mathbb{N} \times \mathbb{N}$ is obtained as a value at $n = 2^{k-1}\,3^{\ell-1} \in \mathbb{N}$. □

Proof of Lemma A.11. Assume that A and B are countable. To prove countability of $A \times B$, we must exhibit a surjective function $\mathbb{N} \to A \times B$.

By countability of A and B, there exists surjective functions

$$g_1\colon\ \mathbb{N} \to A \qquad\qquad g_2\colon\ \mathbb{N} \to B.$$

We may combine them into a function $g\colon \mathbb{N} \times \mathbb{N} \to A \times B$ defined by

$$g(n,m) = \big(g_1(n),\, g_2(m)\big),$$

which is clearly surjective.[5] From Lemma A.12 we get the existence of a surjective function $t\colon \mathbb{N} \to \mathbb{N} \times \mathbb{N}$. Now the composition $g \circ t\colon \mathbb{N} \to A \times B$,

$$\begin{tikzcd} \mathbb{N} \arrow[r, "t"] \arrow[rr, bend right, "g\circ t"'] & \mathbb{N} \times \mathbb{N} \arrow[r, "g"] & A \times B, \end{tikzcd}$$

is surjective. Countability of $A \times B$ follows. □

Proof of Theorem A.10. To show countability of the set $\mathbb{Q}$ of rational numbers, we must exhibit a surjective function $\mathbb{N} \to \mathbb{Q}$.

By definition, rational numbers $q \in \mathbb{Q}$ are of the form $q = \frac{m}{n}$, with $m \in \mathbb{Z}$ and $n \in \mathbb{N}$. In other words, the function

$$r\colon \mathbb{Z} \times \mathbb{N} \to \mathbb{Q} \qquad \text{given by} \qquad r(m,n) = \frac{m}{n} \qquad \text{for } (m,n) \in \mathbb{Z} \times \mathbb{N}$$

is surjective.

The set $\mathbb{Z}$ of integers is countable by Exercise A.1, and the set $\mathbb{N}$ of natural numbers is obviously countable. By Lemma A.11 the Cartesian product $\mathbb{Z} \times \mathbb{N}$ is therefore countable, so there exists a surjective function

$$s\colon \mathbb{N} \to \mathbb{Z} \times \mathbb{N}.$$

Now the composition $r \circ s\colon \mathbb{N} \to \mathbb{Q}$,

$$\begin{tikzcd} \mathbb{N} \arrow[r, "s"] \arrow[rr, bend right, "r\circ s"'] & \mathbb{Z} \times \mathbb{N} \arrow[r, "r"] & \mathbb{Q}, \end{tikzcd}$$

is surjective. Countability of $\mathbb{Q}$ follows. □

---

[5] Indeed, suppose that $(a,b) \in A \times B$. By surjectivity of $g_1\colon \mathbb{N} \to A$ there exists an $n \in \mathbb{N}$ such that $g_1(n) = a$ and by surjectivity of $g_2\colon \mathbb{N} \to B$ there exists an $m \in \mathbb{N}$ such that $g_1(m) = b$. Therefore we have $g(n,m) = \big(g_1(n),\, g_2(m)\big) = (a,b)$, showing surjectivity of g.

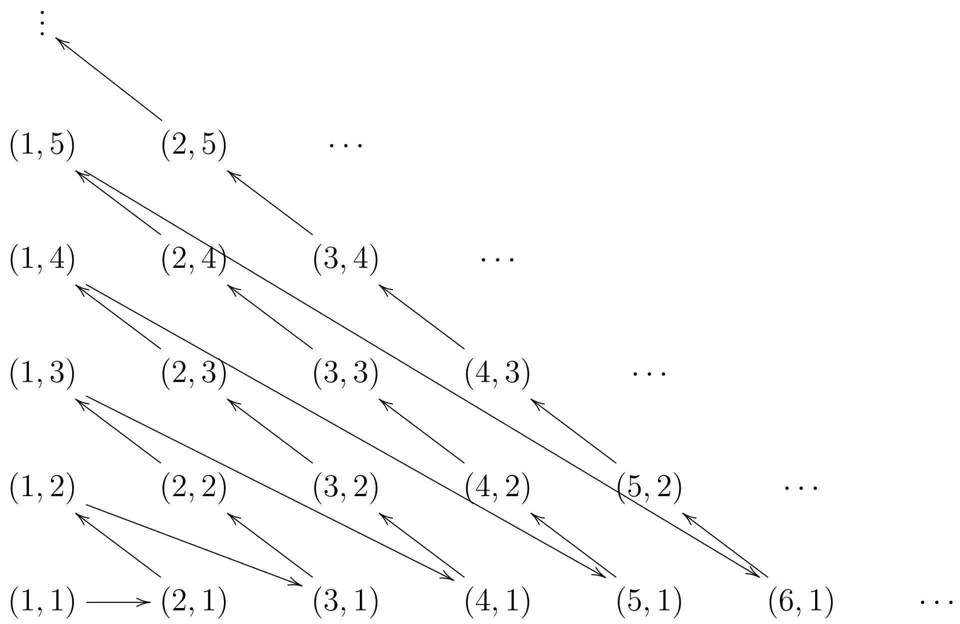


Figure A.1. The arrows indicate an enumeration of $\mathbb{N} \times \mathbb{N}$:

$$\big((1,1),\ (2,1),\ (1,2),\ (3,1),\ (2,2),\ (1,3),\ (4,1),\ (3,2),\ (2,3),\ (1,4),\ \ldots\big).$$

The idea is to observe that for pairs $(n,m) \in \mathbb{N} \times \mathbb{N}$, the sum $n+m$ takes values $2, 3, 4, \ldots$, and for each value of the sum it is straightforward to list the finitely many pairs (n,m) in increasing order of the corresponding m (for example). This enumeration gives another construction of a surjective function $\mathbb{N} \to \mathbb{N} \times \mathbb{N}$.

♡ *Uncountability of the set of real numbers*

Let us now provide the proof that the set $\mathbb{R}$ of real numbers is uncountably infinite. The argument is known as Cantor's diagonal argument. Here we use it to show the impossibility of enumerating the real numbers (even on the unit interval), but variants of the diagonal argument are also used frequently constructively; we will even see some examples later in this course.

Proof of Theorem A.9. In order to prove that $\mathbb{R}$ is not countable, it suffices to prove that the subset $[0,1) \subset \mathbb{R}$ is not countable.

We do the proof by contradiction: we assume that $[0,1)$ is countable, and derive a contradiction from this assumption. If $[0,1)$ were countable, then there would exist a surjective function $f \colon \mathbb{N} \to [0,1)$. For each $n \in \mathbb{N}$, we could write the decimal expansion of the number $f(n) \in [0,1)$:

$$f(n) = 0.\, d_1^{(n)}\, d_2^{(n)}\, d_3^{(n)} \ldots \qquad \text{or more precisely} \qquad f(n) = \sum_{j=1}^{\infty} d_j^{(n)}\, 10^{-j},$$

where $d_j^{(n)} \in \{0, 1, 2, \ldots, 8, 9\}$ for every $j \in \mathbb{N}$. In order to show that f could not have been surjective, we now examine these decimal expansions

$$\begin{aligned}
f(1) &= 0.\, \underline{d}_1^{(1)}\, d_2^{(1)}\, d_3^{(1)}\, d_4^{(1)} \ldots \\
f(2) &= 0.\, d_1^{(2)}\, \underline{d}_2^{(2)}\, d_3^{(2)}\, d_4^{(2)} \ldots \\
f(3) &= 0.\, d_1^{(3)}\, d_2^{(3)}\, \underline{d}_3^{(3)}\, d_4^{(3)} \ldots \\
f(4) &= 0.\, d_1^{(4)}\, d_2^{(4)}\, d_3^{(4)}\, \underline{d}_4^{(4)} \ldots \\
\vdots &= \qquad \ddots
\end{aligned}$$

in particular along the diagonal marked by the underlined digits $\underline{d}_k^{(k)}$, $k \in \mathbb{N}$. Let us construct a number $x \in [0, 1)$ whose first decimal digit is different from the first digit of $f(1)$, second digit is different from the second digit of $f(2)$, etc. Specifically, we choose the k:th digit as

$$d_k := \begin{cases} 8 & \text{if } d_k^{(k)} = 3, \\ 3 & \text{if } d_k^{(k)} \neq 3, \end{cases}$$

which guarantees that $d_k \neq d_k^{(k)}$ for each $k \in \mathbb{N}$. Now consider the number

$$x = 0.\, d_1 d_2 d_3 \ldots \qquad \text{or more precisely} \qquad x = \sum_{j=1}^{\infty} d_j\, 10^{-j}.$$

We have $0 \leq x < 1$, (the strict second inequality uses the fact that 9 does not appear as a digit in x, so x is not $0.9999\ldots = 1$). It is straightforward to check that for every $k \in \mathbb{N}$, we have $x \neq f(k)$, since the k:th digits of x and $f(k)$ are different. But the existence of such an $x \in [0, 1)$ contradicts the surjectivity of f, and finishes the proof. $\square$

Appendix B

Further topics about real numbers

B.1. ♡ Field axioms and order axioms of the real numbers

Here we take a somewhat closer look into the *field axioms* and *order axioms* of the real numbers. The properties addressed in these axioms are undoubtedly more familiar than those in the completeness axiom. A closer look at nevertheless offers an instructive perspective especially into the similarities and differences of the real numbers to other fields used in mathematics (rational numbers, complex numbers, algebraic number fields, finite fields, ...).

♡ Field axioms

The field axioms concern two operations of calculation with numbers: addition and multiplication, and we of course use the standard notational convention that multiplication takes precedence over addition.

The field axioms of real numbers state that the set $\mathbb{R}$ equipped with the operations $+$ and $\cdot$ is a field:

$$\begin{aligned}
&\text{Commutativity of addition:} \\
&\forall x, y \in \mathbb{R}: \quad x + y = y + x \qquad \text{(B.1)} \\
&\text{Associativity of addition:} \\
&\forall x, y, z \in \mathbb{R}: \quad x + (y + z) = (x + y) + z \qquad \text{(B.2)} \\
&\text{Neutral element for addition:} \\
&\exists 0 \in \mathbb{R}: \quad \forall x \in \mathbb{R}: \quad x + 0 = x \qquad \text{(B.3)} \\
&\text{Opposite elements:} \\
&\forall x \in \mathbb{R}: \quad \exists (-x) \in \mathbb{R}: \quad x + (-x) = 0 \qquad \text{(B.4)} \\
&\text{Commutativity of multiplication:} \\
&\forall x, y \in \mathbb{R}: \quad x \cdot y = y \cdot x \qquad \text{(B.5)} \\
&\text{Associativity of multiplication:} \\
&\forall x, y, z \in \mathbb{R}: \quad x \cdot (y \cdot z) = (x \cdot y) \cdot z \qquad \text{(B.6)} \\
&\text{Neutral element of multiplication:} \\
&\exists 1 \in \mathbb{R} \setminus \{0\}: \quad \forall x \in \mathbb{R}: \quad x \cdot 1 = x \qquad \text{(B.7)} \\
&\text{Inverse elements:} \\
&\forall x \in \mathbb{R} \setminus \{0\}: \quad \exists x^{-1} \in \mathbb{R}: \quad x \cdot x^{-1} = 1 \qquad \text{(B.8)} \\
&\text{Distributivity of addition over multiplication:} \\
&\forall x, y, z \in \mathbb{R}: \quad x \cdot (y + z) = x \cdot y + x \cdot z. \qquad \text{(B.9)}
\end{aligned}$$

Hopefully the reader finds these uncontroversial enough to be admitted as axioms.

Note that the axioms explicitly state the existence of two real numbers, 0 ("zero", the neutral element for addition) and 1 ("one", the neutral element for multiplication), and they require that these two are different, $1 \neq 0$ (since $1 \in \mathbb{R} \setminus \{0\}$). Moreover, zero and one are uniquely determined by the properties required of them in the axioms.

Example B.1 (Uniqueness of zero)**.**

Claim: If both $0 \in \mathbb{R}$ and $0' \in \mathbb{R}$ satisfy the property of being neutral elements for addition, then we have $0 = 0'$.

Proof: Suppose that $x + 0 = x$ and $x + 0' = x$ for all $x \in \mathbb{R}$. Then we get

$$0 \overset{\text{(B.3)'}}{=} 0 + 0' \overset{\text{(B.1)}}{=} 0' + 0 \overset{\text{(B.3)}}{=} 0' ,$$

where we first used the neutral element property of $0'$, then commutativity of addition, and finally the neutral element property of 0. □

Exercise (✓) B.1 (Uniqueness of one)**.**

Prove from the field axioms of real numbers that if both $1 \in \mathbb{R}$ and $1' \in \mathbb{R}$ satisfy the property of being neutral elements for multiplication, then we have $1 = 1'$.

Exercise B.2 (Uniqueness of opposite elements and inverse elemens)**.**

Prove that for any $x \in \mathbb{R}$, the opposite element $(-x)$ is unique, and that for any $x \in \mathbb{R} \setminus \{0\}$, the inverse element x^{-1} is unique.

The following example indicates how yet a few other familiar facts about the real numbers are proved starting from the field axioms.

Example B.2 (Natural numbers in $\mathbb{R}$)**.**

Since by the field axiom B.7 there exists a real number $1 \in \mathbb{R}$, we can use addition to define a new number $2 := 1+1 \in \mathbb{R}$. Similarly we define $3 := 2+1$, $4 := 3+1$, $5 := 4+1$, $6 := 5+1$, etc. These satisfy the usual properties. As an example, let us verify the following.

Claim: We have $2 + 2 = 4$.

Proof: Calculate, using the definition of 2, the associativity of addition, the definition of 3, and the definition of 4:

$$2 + 2 \overset{\text{(def of 2)}}{=} 2 + (1 + 1) \overset{\text{(B.2)}}{=} (2 + 1) + 1 \overset{\text{(def of 3)}}{=} 3 + 1 \overset{\text{(def of 4)}}{=} 4 .$$

This proves the claim. □

Note, however, that we did not yet prove that the numbers $1 \in \mathbb{R}$, $2 \in \mathbb{R}$, $3 \in \mathbb{R}$, $4 \in \mathbb{R}$, …defined above are *different* elements of $\mathbb{R}$. This indeed does not follow from the field axioms alone, the order axioms are needed as well. The ideas needed should become clear in Example B.4.

Exercise B.3 (Other exciting properties of natural numbers)**.**

Prove some other exciting properties of these numbers, for example $2 + 3 = 5$ and $2 \cdot 3 = 6$.

Example B.3 (Yet another consequence of the field axioms)**.**

We next verify a familiar property of real numbers which involves this number $2 := 1 + 1$ and an arbitrary real number x.

Claim: For any $x \in \mathbb{R}$ we have $x + x = 2x$.

Proof: Using the neutral element for multiplication (twice), distributivity, and definition of 2, we get the following expression for $x + x$:

$$x + x \overset{\text{(B.7)}}{=} 1 \cdot x + 1 \cdot x \overset{\text{(B.9)}}{=} (1 + 1) \cdot x \overset{\text{(def of 2)}}{=} 2 \cdot x.$$

The claim follows. □

Exercise B.4 (Multiplication by zero)**.**
Prove that for any $x \in \mathbb{R}$ we have $0 \cdot x = 0$.

Ok, you get the point! The familiar facts about the real numbers that you knew since kindergarten can be deduced from the axioms. In order that we get somewhere in this course, we will not actually require you to provide detailed proofs of totally commonplace statements such as $0 \cdot x = 0$ or $2 \cdot 2 = 4$ or $(-1)^2 = 1$ or $(x - y)^2 = x^2 - 2xy + y^2$ — as long as you realize that *in principle* they should be obtained as logical consequences of the axioms (and familiar notational conventions and definitions).[1]

Other fields

There are many other fields besides $\mathbb{R}$, and they all satisfy the field axioms. In this course we do not use other fields than $\mathbb{R}$ and $\mathbb{Q}$, but for perspective we mention a few prominent examples of fields:

- the field of rational numbers $\mathbb{Q}$;
- the field $\mathbb{Q}(\sqrt{5})$ of rational numbers adjoined with a square root of 5;
- the field of complex numbers $\mathbb{C}$;
- the finite field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ of p elements, where p is a prime number;
- the finite field $\mathbb{F}_{p^k}$ of p^k elements, where p is a prime number and $k \in \mathbb{N}$;
- the field $\mathbb{Q}_p$ of p-adic numbers, where p is a prime number;
- the field $\mathbb{K}(q)$ of rational functions in a single variable q over another field $\mathbb{K}$;
- $\cdots$

In particular all (familiar) consequences of the field axioms hold for these as well: for example if $z, w \in \mathbb{C}$ are complex numbers, then $(z + w)^2 = z^2 + 2zw + w^2$, where $2 := 1 + 1 \in \mathbb{C}$. Note, however, that for instance in the two-element field $\mathbb{F}_2$, we have $2 = 0$. So you should not try to derive the property $2 \neq 0$ of real numbers from the field axioms (you would fail, it is not a logical consequence of them!).

[1]Nowadays there are, by the way, formally verified proofs of a vast amount of basic statements about real numbers and even metric space topology; certainly including all statements in this section. What this means is that the statements have been phrased in an entirely formal language, proofs have also been given in the formal language, and a computer has verified that the proofs adhere to the rules of logic and actually demonstrate the statements. See, for example, `https://leanprover-community.github.io/`. I slightly doubt, however, that all of that makes for much more pleasant reading than these lecture notes… (If you think otherwise, I will accept your formally verified solutions to the exam problems!) Moreover, producing the fully formal proofs is so tedious that already in some parts of this undergraduate course we actually get deeper into mathematics than where hundreds of mathematicians and computer scientists have gotten in the past decades with formally verified mathematics. A noble pursuit, nevertheless!

♡ Order axioms

Since there are other fields besides just the real numbers, the real numbers must have some further specific properties that distinguish them. Having an order relation $<$ turns out to be quite special already. To work with the order relation $<$, we adopt the usual additional notations such as: $y > x$ means $x < y$, $x \leq y$ means that either $x < y$ or $x = y$, etc.

The order axioms of the real number field are:

$$\begin{aligned}
&\text{Alternatives:} \\
&\forall x, y \in \mathbb{R} : \quad \text{one and only one of the three relations} \\
&\qquad\qquad x < y,\ x = y,\ x > y \text{ is true} && \text{(B.10)} \\
&\text{Transitivity:} \\
&\forall x, y, z \in \mathbb{R} : \quad x < y \text{ and } y < z \implies x < z && \text{(B.11)} \\
&\text{Compatibility with addition:} \\
&\forall x, y, z \in \mathbb{R} : \quad x < y \implies x + z < y + z && \text{(B.12)} \\
&\text{Compatibility with multiplication:} \\
&\forall x, y \in \mathbb{R} : \quad x > 0 \text{ and } y > 0 \implies x \cdot y > 0 && \text{(B.13)}
\end{aligned}$$

From these and the field axioms together, one can derive consequences.

Example B.4 (Some consequences of the order axioms)**.**

Claim: We have $0 < 1$.

Proof: Since we know that $0 \neq 1$, there are two mutually exclusive alternatives left: $0 < 1$ or $0 > 1$. Let us prove that the second one is impossible. So assume that $0 > 1$. By adding -1 and using compatibility with addition, we get (after simplifying by neutral element of addition and opposite element properties) that $-1 > 0$. Then by compatibility with multiplication we get $(-1)(-1) > 0$. But from the field axioms one can show that $(-1)(-1) = 1$, so this simplifies to $1 > 0$. In this case we have both $1 > 0$ and $0 > 1$, which is not allowed by the alternatives. So we have to discard the possibility that $0 > 1$, and we conclude that $0 < 1$ holds. □

As the example of the field of two elements shows, the property $2 \neq 0$ of real numbers cannot be derived from the field axioms alone. Having also the order axioms, we can now prove it.

Claim: We have $2 \neq 0$ in $\mathbb{R}$.

Proof: We already know that $1 > 0$. Adding 1 and using compatibility with addition, we find $2 > 1$. Transitivity then implies $2 > 0$. This rules out the possibility that $2 = 0$ by the axiom that exactly one of the alternatives $2 > 0$, $2 = 0$, and $2 < 0$, is true. □

Exercise B.5 (Product of two negative numbers)**.**

Prove that the product of any two negative real numbers is positive.

Ok, you get the point again. All the familiar properties you already knew in kindergarten are logical consequences of the axioms.

Note that the rational number field $\mathbb{Q}$ also satisfies the order axioms. Therefore, since $\mathbb{R} \neq \mathbb{Q}$ after all, there must be yet something else that distinguishes the real numbers. It is going to be the most profound of the axioms of real numbers.

B.2. ♡ Equivalence of the formulations of the completeness axiom

Recall that in Lecture II, any of the following statements was called the **completeness axiom** of $\mathbb{R}$:

(C1): Every non-empty subset $A \subset \mathbb{R}$ which is bounded from above has a least upper bound $\sup A \in \mathbb{R}$.

(C2): Every increasing real-number sequence $(a_n)_{n\in\mathbb{N}}$ which is bounded from above has a limit $\lim_{n\to\infty} a_n \in \mathbb{R}$.

(C3): Every collection $(I_n)_{n\in\mathbb{N}}$ of closed intervals $I_n \subset \mathbb{R}$, which is nested in the sense that $I_{n+1} \subset I_n$ for every $n \in \mathbb{N}$, has a nonempty intersection

$$\bigcap_{n\in\mathbb{N}} I_n \neq \emptyset.$$

Let us now prove the equivalence of (C1), (C2), and (C3). We prove separately the implications

$$\text{(C1)} \implies \text{(C2)}, \qquad \text{(C2)} \implies \text{(C3)}, \qquad \text{(C3)} \implies \text{(C1)}.$$

The equivalence of all three follows by combining these implications.

Proof of (C1) ⇒ (C2). Assume (C1). Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers which is increasing and bounded from above. We must prove that this sequence converges.

Since the sequence $(a_n)_{n\in\mathbb{N}}$ is bounded from above, the set

$$A = \{a_n \mid n \in \mathbb{N}\}$$

of its values is a subset in $\mathbb{R}$ which is bounded from above. By assumption (C1), then, this subset has a least upper bound $t_0 := \sup\{a_n \mid n \in \mathbb{N}\} \in \mathbb{R}$. We will prove that the sequence $(a_n)_{n\in\mathbb{N}}$ converges to t_0.

Let $\varepsilon > 0$. Since $t_0 - \varepsilon$ is not an upper bound for $\{a_n \mid n \in \mathbb{N}\}$, there exists some $n_\varepsilon \in \mathbb{N}$ such that $a_{n_\varepsilon} > t_0 - \varepsilon$. Since the sequence is increasing, for all $n \geq n_\varepsilon$ we must then also have $a_n > t_0 - \varepsilon$. On the other hand, since t_0 is an upper bound for $\{a_n \mid n \in \mathbb{N}\}$, we have $a_n \leq t_0$ for all $n \in \mathbb{N}$. For $n \geq n_\varepsilon$ we have thus obtained

$$t_0 - \varepsilon \ < \ a_n \ \leq \ t_0 \ < \ t_0 + \varepsilon,$$

which implies $|a_n - t_0| < \varepsilon$. Since such an $n_\varepsilon \in \mathbb{N}$ was found for an arbitrary $\varepsilon > 0$, we have by definition of limits shown that $\lim_{n\to\infty} a_n = t_0$.

Property (C2) is thus established. □

Proof of (C2) ⇒ (C3). Assume (C2). Let $(I_n)_{n\in\mathbb{N}}$ be a collection of closed intervals $I_n = [a_n, b_n] \subset \mathbb{R}$ such that $I_{n+1} \subset I_n$ for every $n \in \mathbb{N}$. We must prove that $\bigcap_{n\in\mathbb{N}} I_n \neq \emptyset$.

The nesting condition

$$[a_{n+1}, b_{n+1}] \ = \ I_{n+1} \ \subset \ I_n \ = \ [a_n, b_n]$$

simply amounts to the following inequalities among the endpoints of the intervals:

$$a_n \ \leq \ a_{n+1} \ \leq \ b_{n+1} \ \leq \ b_n.$$

Therefore the sequence $(a_n)_{n\in\mathbb{N}}$ of the left endpoints of the intervals is increasing. This sequence is also bounded from above, because for any $n \in \mathbb{N}$ we have $a_n \leq b_n \leq b_1$ (the second inequality follows from the fact that the sequence of right endpoints is decreasing). By assumption (C2), the sequence $(a_n)_{n\in\mathbb{N}}$ therefore has a limit $\alpha = \lim_{n\to\infty} a_n$. We will show that $\alpha \in \bigcap_{n\in\mathbb{N}} I_n$, and it will follow that the intersection is nonempty.

To show that $\alpha \in \bigcap_{m\in\mathbb{N}} I_m$, by definition of intersection we must show that $\alpha \in I_m$ for all $m \in \mathbb{N}$. So fix $m \in \mathbb{N}$. Since $I_m = [a_m, b_m]$, showing $\alpha \in I_m$ amounts to

$$a_m \overset{?}{\leq} \alpha \overset{?}{\leq} b_m.$$

The first inequality above is clear: for $n \geq m$ we have $a_n \geq a_m$ (the left endpoints form an increasing sequence), so by preservation of bounds (Corollary II.15) we have $\alpha = \lim_{n\to\infty} a_n \geq a_m$. To prove the other inequality, we argue by contradiction. Suppose that it is not true, i.e., that $\alpha > b_m$. By definition of the limit $\alpha = \lim_{n\to\infty} a_n$, corresponding to the positive number $\varepsilon = \alpha - b_m > 0$, there exists an $n_\varepsilon \in \mathbb{N}$ so that for $n \geq n_\varepsilon$ we have $|a_n - \alpha| < \varepsilon = \alpha - b_m$, which implies

$$a_n > \alpha - \varepsilon = \alpha - (\alpha - b_m) = b_m.$$

In particular for $n \geq \max\{n_\varepsilon, m\}$, since the sequence of right endpoints is decreasing, we would get $a_n > b_m \geq b_n$. This, however, is a contradiction with the ordering $a_n \leq b_n$ of the left and right endpoints of $I_n = [a_n, b_n]$. With this contradiction, we have also established the other claimed inequality. We have thus concluded $\alpha \in I_m$ for all $m \in \mathbb{N}$, which shows that the intersection $\bigcap_{m\in\mathbb{N}} I_m$ is not empty, establishing (C3). $\square$

Proof of (C3) $\Rightarrow$ (C1). Assume (C3). Let $A \subset \mathbb{R}$ be a nonempty subset which is bounded from above. We must prove that A has a least upper bound.

Since A is bounded from above, it has an upper bound, and since A is moreover nonempty, it is possible to choose an upper bound $b_0 \in \mathbb{R}$ so that the number $a_0 = b_0 - 1$ is not an upper bound for A.

Consider the midpoint $c_0 = \frac{a_0+b_0}{2}$. If c_0 is an upper bound for A, we set $a_1 = a_0$ and $b_1 = c_0$. Otherwise we set $a_1 = c_0$ and $b_1 = b_0$. With such choices, we are guaranteed that b_1 is an upper bound for A, while a_1 is not.

Continue inductively. When b_n is an upper bound for A and a_n is not, consider the midpoint $c_n = \frac{a_n+b_n}{2}$. Depending on whether c_n is an upper bound for A or not, set $a_{n+1} = a_n$ and $b_{n+1} = c_n$, or $a_{n+1} = c_n$ and $b_{n+1} = b_n$. By construction we get a decreasing sequence $(b_n)_{n\in\mathbb{N}}$ of upper bounds for A, and an increasing sequence $(a_n)_{n\in\mathbb{N}}$ of numbers that are not upper bounds for A, and moreover $a_n < b_n$ for all $n \in \mathbb{N}$. The distances are halved at every step, so $b_n - a_n = 2^{-n}$.

From the above properties we get that the closed intervals $I_n = [a_n, b_n]$ for $n \in \mathbb{N}$ are nested, $I_{n+1} \subset I_n$ for all $n \in \mathbb{N}$. By the assumed property (C3), the intersection $\bigcap_{n\in\mathbb{N}} I_n$ is then nonempty. On the other hand, the intersection cannot contain two distinct points: if $x, y \in \bigcap_{n\in\mathbb{N}} I_n$, then for all $n \in \mathbb{N}$ we have $x, y \in I_n = [a_n, b_n]$, which implies that $|x - y| \leq b_n - a_n = 2^{-n} \to 0$, so $|x - y| = 0$ and $x = y$. We conclude that the intersection contains exactly one point ξ, i.e., it is a singleton $\bigcap_{n\in\mathbb{N}} I_n = \{\xi\}$. From the same estimate we also get that

$$\xi - 2^{-n} \ \leq\ a_n \ \leq\ \xi \ \leq\ b_n \ \leq\ \xi + 2^{-n}.$$

By the squeeze theorem (Lemma II.17), this implies $\lim_{n\to\infty} a_n = \xi$ and $\lim_{n\to\infty} b_n = \xi$.

Our goal was to show that the set A has a least upper bound. We will show that ξ is it.

Let us first verify that ξ is an upper bound for A. So let $x \in A$. By construction each b_n is an upper bound for A, so $b_n \geq x$ for all $n \in \mathbb{N}$. Since we have $\xi = \lim_{n\to\infty} b_n$, from the preservation of bounds (Corollary II.15) we get that $\xi \geq x$. This was true for an arbitrary element $x \in A$, so indeed ξ is an upper bound for A.

Let us then show that no smaller number $\xi' < \xi$ is an upper bound for A. Since $\lim_{n\to\infty} a_n = \xi$, for any $\xi' < \xi$ we have $a_n > \xi'$ for large enough n (take $\varepsilon = \xi - \xi' > 0$ in the definition of limit), and since a_n was not an upper bound for A, the smaller number $\xi' < a_n$ can not be either.

This proves that $\sup A = \xi$ exists, establishing (C1). $\square$

$C_0 = [0,1]$

$C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$

$C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$

C_3 : union of 8 intervals of length $\frac{1}{27}$

C_4 : union of 16 intervals of length $\frac{1}{81}$

C_5 : union of 32 intervals of length $\frac{1}{243}$

$\vdots$

FIGURE B.1. The Cantor set $C = \bigcap_{n\in\mathbb{N}} C_n \subset [0,1]$ is constructed iteratively so that from stage n approximation C_n one removes the middle thirds of the remaining intervals to obtain stage $n+1$ approximation C_{n+1}.

B.3. ♡ Cantor set

The Cantor set, or Cantor's $\frac{1}{3}$-set, is a subset $C \subset [0,1]$, which has interesting properties from topological, set-theoretic, and measure theoretic points of view.

Informally, the Cantor set C is constructed by taking the unit interval $[0,1]$, removing its middle third, then removing the middle third of both of the remaining thirds, and successively always removing the middle third of each remaining interval. Figure B.1 illustrates this iterative construction.

More formally, the Cantor set C is defined, using a sequence $(C_n)_{n\in\mathbb{N}}$ of nested approximating sets, as the intersection

$$C = \bigcap_{n\in\mathbb{N}} C_n, \tag{B.14}$$

where the sets C_n, $n \in \mathbb{N}$, are as follows. It makes sense to think of the unit interval $C_0 = [0,1]$ as the zeroth stage approximation. The first stage approximation is the set $C_1 = \big[0, \frac{1}{3}\big] \cup \big[\frac{2}{3}, 1\big]$; a union of 2 closed intervals of length $\frac{1}{3}$. The second stage approximation is $C_2 = \big[0, \frac{1}{9}\big] \cup \big[\frac{2}{9}, \frac{1}{3}\big] \cup \big[\frac{2}{3}, \frac{7}{9}\big] \cup \big[\frac{8}{9}, 1\big]$; a union of 4 closed intervals of length $\frac{1}{9}$. The n:th stage approximation is the union of 2^n closed intervals of length 3^{-n}, specifically

$$C_n = \bigcup_{b_1,\ldots,b_n\in\{0,1\}} I^{(n)}_{b_1,\ldots,b_n} \tag{B.15}$$

where

$$I^{(n)}_{b_1,\ldots,b_n} = \Big[2\sum_{j=1}^{n} b_j\, 3^{-j}\,,\;\; 2\sum_{j=1}^{n} b_j\, 3^{-j} + 3^{-n}\Big]. \tag{B.16}$$

The interval (B.16) is exactly obtained by keeping the left or right third, corresponding to $b_n = 0$ and $b_n = 1$, respectively, of the $n-1$:st stage interval $I^{(n-1)}_{b_1,\ldots,b_{n-1}}$. The middle third of $I^{(n-1)}_{b_1,\ldots,b_{n-1}}$ has been removed in stage n. In particular, we have the nesting property $C_n \subset C_{n-1}$ for all $n \in \mathbb{N}$, i.e.,

$$C_0 \supset C_1 \supset C_2 \supset C_3 \supset \cdots .$$

Let us now make some observations regarding the set C defined as the intersection (B.14).

The Cantor set is nonempty

The first observation is that the set C is non-empty,

$$C \neq \emptyset.$$

There are various ways to see this; the easiest is to just note that $0 \in C_n$ and $1 \in C_n$ for every $n \in \mathbb{N}$, so in the intersection we at least have $0 \in C$ and $1 \in C$. In other words, at least the two endpoints of the unit interval which constitutes the zeroth stage approximation C_0 belong to the Cantor set C. By an essentially similar reasoning one can see that all of the 2^{n+1} endpoints of the 2^n intervals of stage n remain in the intersection — we leave the detailed verification of this as an exercise. When all different stages $n \in \mathbb{N}$ are considered, such approximating interval endpoints give countably infinitely many different points in C — we again leave the details as an exercise, because it turns out that there are in fact vastly more points in C, and they admit a unified description.

The Cantor set is uncountably infinite

The unified description of points of C is by infinite binary sequences $(b_1, b_2, b_3, \ldots)$.

First, suppose that an infinite binary sequence $(b_n)_{n\in\mathbb{N}}$ is given, with $b_n \in \{0,1\}$ for $n \in \mathbb{N}$. Note the nesting of the intervals,

$$I^{(1)}_{b_1} \supset I^{(2)}_{b_1,b_2} \supset I^{(3)}_{b_1,b_2,b_3} \supset \cdots \supset I^{(n)}_{b_1,\ldots,b_n} \supset I^{(n+1)}_{b_1,\ldots,b_n,b_{n+1}} \supset \cdots$$

By the formulation (C3) of the completeness axiom, the intersection of these nested closed intervals is nonempty, and since the lengths 3^{-n} of the intervals are tending to 0 as $n \to \infty$, it is not difficult to see that the intersection cannot contain more than one point. Therefore there exists some $x \in \mathbb{R}$ (which of course depends on the chosen binary sequence $(b_n)_{n\in\mathbb{N}}$) such that

$$\bigcap_{n\in\mathbb{N}} I^{(n)}_{b_1,\ldots,b_n} = \{x\} .$$

Since $x \in I^{(n)}_{b_1,\ldots,b_n} \subset C_n$ for every $n \in \mathbb{N}$, we then have also

$$x \in \bigcap_{n\in\mathbb{N}} C_n = C.$$

This shows that to any infinite binary sequence $(b_n)_{n\in\mathbb{N}}$, we can associate a point x in the Cantor set C. We will observe also that if $(b_n)_{n\in\mathbb{N}}$ and $(b'_n)_{n\in\mathbb{N}}$ are two different binary sequences, then the corresponding points x and x' are different. Indeed then for some $k \in \mathbb{N}$ we have $b_k \neq b'_k$, and the corresponding k:th stage intervals are disjoint, $I^{(k)}_{b_1,\ldots,b_k} \cap I^{(k)}_{b'_1,\ldots,b'_k} = \emptyset$. But since the corresponding points belong to these disjoint intervals, $x \in I^{(k)}_{b_1,\ldots,b_k}$ and $x' \in I^{(k)}_{b'_1,\ldots,b'_k}$, we must have $x \neq x'$.

Exercise (♯) B.6 (The point in Cantor set corresponding to a binary sequence)**.**
Show that the point x, corresponding to the binary sequence $(b_n)_{n\in\mathbb{N}}$, is given by the formula

$$x \;=\; 2\sum_{j=1}^{\infty} b_j\, 3^{-j}. \tag{B.17}$$

Use this to give a different proof that different binary sequences yield different points in the Cantor set.

Remark B.5 (Describing the Cantor set by ternary numbers)**.**
Consider the analogue of decimal expansions $\sum_{j=1}^{\infty} d_j\, 10^{-j}$ where $d_1, d_2, \ldots \in \{0,1,\ldots,8,9\}$ (see Example II.33), but in base three instead of base ten. These are numbers of the form $\sum_{j=1}^{\infty} t_j\, 3^{-j}$ where the "digits" are $t_1, t_2, \ldots \in \{0,1,2\}$.

From the expression (B.17) we see that the points in the Cantor set correspond to base three "decimal expansions" (ternary expansions) which only ever use digits 0 and 2; never 1.

Conversely, we will show that each point of the Cantor set corresponds to some binary sequence as above. Suppose that $x \in C$. By the definition of intersection, we then have $x \in C_n$ for every $n \in \mathbb{N}$. Since C_n is the union of disjoint intervals (B.15), there must exist exactly one choice of $b_1^{(n)}, \ldots, b_n^{(n)} \in \{0,1\}$ such that $x \in I^{(n)}_{b_1^{(n)},\ldots,b_n^{(n)}}$. Let us now define an infinite sequence $(b_n)_{n\in\mathbb{N}}$ by setting $b_n = b_n^{(n)}$ where $b_n^{(n)}$ came from the above choice of the n:th stage interval. But in fact $b_n^{(m)}$ will not depend on which stage $m \geq n$ we were considering[2]; we have $b_n^{(m)} = b_n$ for all $m \geq n$. Therefore we have $x \in I^{(n)}_{b_1,\ldots,b_n}$ for all $n \in \mathbb{N}$, i.e., $x \in \bigcap_{n\in\mathbb{N}} I^{(n)}_{b_1,\ldots,b_n}$. But this intersection of nested intervals is exactly the singleton that we used to define the point in C which corresponds to the binary sequence $(b_n)_{n\in\mathbb{N}}$, so we have now shown that any point $x \in C$ is of this form.

On measure theoretic properties of the Cantor set

While this is not a course in measure theory, it is interesting to note that the Cantor set C, which is set-theoretically large in that it is uncountably infinite, is nevertheless measure theoretically as small as possible: it has *Lebesgue measure zero*. The Lebesgue measure on $\mathbb{R}$ is the good mathematical generalization of length. It is not difficult to convince oneself that the "length" of the Cantor set C should indeed be zero, by the following argument. For any $n \in \mathbb{N}$, we have $C \subset C_n$, i.e., The Cantor set C is contained in the union of 2^n intervals of length 3^{-n} each. Therefore it is reasonable to say that its length is at most $2^n\, 3^{-n} = \left(\frac{2}{3}\right)^n$. But $n \in \mathbb{N}$ was arbitrary here, and as $n \to \infty$, this upper bound on the length is tending to 0.

---

[2]Suppose $m \geq n$. The numbers $b_1^{(n)}, \ldots, b_n^{(n)} \in \{0,1\}$ are chosen so that $x \in I^{(n)}_{b_1^{(n)},\ldots,b_n^{(n)}}$. Likewise, the numbers $b_1^{(m)}, \ldots, b_m^{(m)} \in \{0,1\}$ are chosen so that $x \in I^{(m)}_{b_1^{(m)},\ldots,b_m^{(m)}}$. But we have the nesting property: an m:th stage interval is contained in exactly one n:th stage interval

$$I^{(m)}_{b_1^{(m)},\ldots,b_m^{(m)}} \;\subset\; I^{(n)}_{b_1^{(m)},\ldots,b_n^{(m)}}.$$

By disjointness of the n:th stage intervals, we must have $b_1^{(m)} = b_1^{(n\;)}, \ldots, b_n^{(m)} = b_n^{(n)}$.

Another measure theoretic point of view to the Cantor set is its fractal dimension. In arbitrary metric spaces one can define Hausdorff measures of arbitrary dimensions $d \in [0, +\infty)$, and the Hausdorff dimension of a set is defined as the infimum of dimensions d such that the d-dimensional Hausdorff measure vanishes. The set C has Hausdorff dimension $\frac{\log 2}{\log 3} \approx 0.6309$, which is strictly less than the dimension 1 of any nonempty open subset of the real line, but strictly larger than the dimension 0 of countable sets such as $\mathbb{Q}$.

On topological properties of the Cantor set

Without entering details, we mention a few interesting topological properties of the Cantor set $C \subset \mathbb{R}$. These require notions from later lectures, and are perhaps best treated as additional exercises at the suitable point in the course.

After Lecture XI you can do the following.

Exercise B.7 (The Cantor set is compact)**.**
Prove that the Cantor set C is compact.

After the definitions in Lecture VII, it is already possible to do the following, although with concepts from Lecture XI it becomes much easier.

Exercise ($\sharp$) B.8 (The Cantor set is nowhere dense)**.**
Prove that the Cantor set C has the property that any nonempty open set $V \subset \mathbb{R}$ contains a nonempty open set $U \subset V \subset \mathbb{R}$ such that $U \cap C = \emptyset$.

The set C is in this topological sense small; smaller even than the countable but dense set $\mathbb{Q}$ of rational numbers, which intersects every nonempty open set $U \subset \mathbb{R}$ (as can be seen from Theorem II.21).

After Lecture XII you can do the following.

Exercise ($\sharp$) B.9 (The Cantor set is totally disconnected)**.**
Prove that all connected components of the Cantor set C are singletons.

B.4. ♡ Applications of supremum and infimum

The notions of supremum and infimum may seem a bit abstract. Indirectly, via the different formulation (C2) of the completeness axiom, we saw that supremum is important for instance for the existence of real numbers with given decimal expansions (Example II.33). But it is good to realize that supremum is also quite directly used in many common constructions in analysis. We describe a few examples here, albeit without any intention to elaborate on the full details.

♡ Arc length of a curve

One relatively simple, intuitive, and commonly used application of supremum is the length of a curve; also called the *arc length*. Here we describe the definition in the context of curves in a metric space (X, d), which is introduced Lecture V. Also the notion of continuity from Lecture VI is briefly invoked.

It is also possible to treat this as an example without the general definitions, however. The reader who so prefers may consider curves in the familiar d-dimensional space $X = \mathbb{R}^d$, and interpret the distances between two points $\vec{x} = (x_1, \ldots, x_d)$ and $\vec{y} = (y_1, \ldots, y_d)$ in $\mathbb{R}^d$ as given by the usual formula

$$\mathsf{d}(\vec{x}, \vec{y}) = \sqrt{(y_1 - x_1)^2 + \cdots + (y_d - x_d)^2}.$$

A (parametrized) curve in X is a continuous function from a closed interval $[a, b]$ to the space X,

$$\gamma\colon [a, b] \to X.$$

For a given subdivision

$$a = t_0 < t_1 < t_2 < \cdots < t_n = b$$

of the interval $[a, b]$, it is natural to interpret the expression

$$\sum_{j=1}^{n} \mathsf{d}\big(\gamma(t_{j-1}), \gamma(t_j)\big) \;=\; \mathsf{d}\big(\gamma(t_0), \gamma(t_1)\big) + \cdots + \mathsf{d}\big(\gamma(t_{n-1}), \gamma(t_n)\big)$$

as an approximation of the length of the curve γ. Under refinement of the subdivision, the approximate length never decreases.[3] Therefore the natural notion of the length of the curve, as obtained in the limit of refinements of subdivisions, is the supremum of all the approximating expressions: the **length** of the curve γ is *defined* as

$$L(\gamma) \;:=\; \sup \left\{ \sum_{j=1}^{n} \mathsf{d}\big(\gamma(t_{j-1}), \gamma(t_j)\big) \;\middle|\; n \in \mathbb{N},\; a = t_0 < t_1 < \cdots < t_n = b \right\}. \quad \text{(B.18)}$$

In $\mathbb{R}^d$, for differentiable (or at least piecewise differentiable) curves γ, one can show that this definition of length coincides with the integral $L(\gamma) = \int_a^b \|\gamma'(t)\| \, \mathrm{d}t$, but the integral is in fact more complicated to define precisely (see below) and it exists less generally than (B.18): it cannot be defined in general metric spaces, and even in $\mathbb{R}^d$ it requires differentiability assumptions, which depend on the chosen parametrization, etc. The staightforward definition (B.18) using supremum is clearly good!

♡ Infinite sums with non-negative terms

If we have an arbitrary collection $(s_i)_{i \in I}$ of nonnegative terms $s_i \geq 0$, then there is a quick and rather simple way to give a meaning to the sum

$$\sum_{i \in I} s_i.$$

This is rather remarkable, since we can allow the index set I to be arbitrarily large, meaning that the number of terms in the sum can be finite, countably infinite, or even uncountably infinite.

[3] This is a simple consequence of the triangle inequality (M-Δ) of the metric d.

Of course a partial sum that includes only finitely many terms $s_{i_1}, \dots, s_{i_k}$, with $i_1, \dots, i_k \in I$ distinct, is trivially defined as

$$s_{i_1} + \cdots + s_{i_k}.$$

The idea is that the other terms omitted are non-negative, so the "true sum" must be at least as large. In other words the "true sum" should be an upper bound for the set of all finite partial sums. The definition

$$\sum_{i\in I} s_i \;:=\; \sup\left\{ s_{i_1} + \cdots + s_{i_k} \;\middle|\; k \in \mathbb{N}_0,\ i_1, \dots, i_k \in I \text{ distinct} \right\}$$

declares the "true sum" as the least upper bound for the set of all finite partial sums — i.e., it requires that the gap in between the finite partial sums and the "true sum" can be made arbitrarily small by including enough terms.

The set of all finite partial sums is nonempty (we always have the possibility of including $k = 0$ terms in the partial sum), but it may or may nor be bounded, so we obviously have to allow the possibility $\sum_{i\in I} s_i = +\infty$ (in keeping with the convention of supremum of unbounded sets) if arbitrarily large finite partial sums exist.

This definition also has the advantage that it is clear that the sum does not depend on the "order of elements" (in fact we did not even require the index set I to be ordered!).

But to be fair, this definition crucially relies on the terms being nonnegative. It does not admit easy generalizations to infinite sums with terms of both signs.[4]

Exercise B.10 (♡ Uncountably many strictly positive terms yields infinite sum)**.**
Let $(s_i)_{i\in I}$ be a collection of nonnegative terms $s_i \geq 0$. Consider the sum $\sum_{i\in I} s_i$.

(a) Suppose that $\sum_{i\in I} s_i < +\infty$. Show that for any $m \in \mathbb{N}$, there can only exist finitely many indices $i \in I$ such that $s_i \geq \frac{1}{m}$.

(b) Suppose again that $\sum_{i\in I} s_i < +\infty$. Show that there can only exist countably many indices $i \in I$ such that $s_i > 0$.
Hint: *Consider all different $m \in \mathbb{N}$ in (a). Recall that countable unions of finite sets are countable.*

(c) Now prove that if there are uncountably infinitely many indices i corresponding to a strictly positive term $s_i > 0$, then we have $\sum_{i\in I} s_i = +\infty$.
Hint: *Contrapositive of (b).*

♡ Riemann integration

Let

$$f \colon [a, b] \to \mathbb{R}$$

be a real valued function defined on a closed interval $[a, b]$. Assume also that f is bounded in the sense that there exists some $M \in \mathbb{R}$ such that $|f(x)| \leq M$ for all $x \in [a, b]$.

Consider a finite subdivision of the interval $[a, b]$, consisting of points

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

[4]Somewhat more general sums (and integrals) are studied in measure theory.

On the subdivision interval $[x_{j-1}, x_j]$ for $j \in \{1, \ldots, n\}$, the values of the function have the greatest lower bound and the least upper bound

$$s_j = \inf \Big\{ f(x) \;\Big|\; x_{j-1} \leq x \leq x_j \Big\} \qquad \text{and} \qquad t_j = \sup \Big\{ f(x) \;\Big|\; x_{j-1} \leq x \leq x_j \Big\}.$$

Using these, we define the lower and upper Riemann sums

$$S = \sum_{j=1}^{n} s_j \left(x_j - x_{j-1} \right) \qquad \text{and} \qquad T = \sum_{j=1}^{n} t_j \left(x_j - x_{j-1} \right)$$

associated with the subdivision $a = x_0 < x_1 < x_2 < \cdots < x_n = b$.

It is not particularly difficult to show that the lower and upper Riemann sums satisfy the following:

- If a subdivision $a = x'_0 < x'_1 < x'_2 < \cdots < x'_m = b$ is a refinement of the above subdivision (i.e. $\{x_0, x_1, \ldots, x_n\} \subset \{x'_0, x'_1, \ldots, x'_m\}$) and S' and T' are its associated lower and upper Riemann sums, then $S \leq S'$ and $T' \leq T$.
- If $\tilde{S}$ and $\tilde{T}$ are the lower and upper Riemann sums associated with any subdivision, then $\tilde{S} \leq T$ (and symmetrically $S \leq \tilde{T}$).

Exercise B.11 (♡ Proof of the properties of upper and lower Riemann sums).
Prove the two claims above.

Since increasingly fine subdivisions lead to larger lower Riemann sums, it is natural to define the lower integral $I_-(f)$ as the least upper bound for the set of all lower Riemann sums for all subdivisions; or explicitly

$$I_-(f) := \sup \left\{ \sum_{j=1}^{n} \Big(\inf_{x \in [x_{j-1}, x_j]} f(x) \Big) \left(x_j - x_{j-1} \right) \;\middle|\; a = x_0 < x_1 < \ldots < x_n = b \right\}.$$

Similarly, since increasingly fine subdivisions lead to smaller upper Riemann sums, it is natural to define the upper integral $I_+(f)$ as the greatest lower bound for the set of all upper Riemann sums for all subdivisions; or explicitly

$$I_+(f) := \inf \left\{ \sum_{j=1}^{n} \Big(\sup_{x \in [x_{j-1}, x_j]} f(x) \Big) \left(x_j - x_{j-1} \right) \;\middle|\; a = x_0 < x_1 < \ldots < x_n = b \right\}.$$

From the earlier observations one straightforwardly gets $I_-(f) \leq I_+(f)$.

Definition B.6 (Riemann integral).
A function $f \colon [a, b] \to \mathbb{R}$ is said to be **(Riemann) integrable**, if $I_-(f) = I_+(f)$. In this case we define its **(Riemann) integral** as

$$\int_a^b f(x) \, \mathrm{d}x \; := \; I_-(f) \; = \; I_+(f).$$

From characterizations of supremum and infimum, it is quite straightforweard to prove that

Theorem B.7 (Characterization of Riemann integrability).
A function $f \colon [a, b] \to \mathbb{R}$ is Riemann integrable if and only if for every $\varepsilon > 0$

there exists a subdivision such that the associated upper and lower Riemann sums satisfy $T - S < \varepsilon$*.*

Exercise B.12 (♡ Proof of Theorem B.7)**.**
Prove the above theorem.

To properly define the familiar integral, therefore, we end up using the notions of supremum and infimum — in fact many times.

Appendix C

Further topics about continuity

C.1. ♡ Other quantitative notions of continuity

Besides Lipschitz-continuity, there are also other quite frequently used quantitative notions of continuity of functions between metric spaces. Here we briefly introduce a few that you are likely to encounter in subsequent mathematics courses.

Let (X, d_X) and (Y, d_Y) be two metric spaces.

Hölder continuity

Definition C.1 (Hölder property)**.**
Let $\alpha > 0$ and $M \geq 0$ be constants. A function $f\colon X \to Y$ is said to be **(M, α)-Hölder** if for all $x_1, x_2 \in X$ we have

$$\mathsf{d}_Y\big(f(x_1), f(x_2)\big) \;\leq\; M\,\big(\mathsf{d}_X(x_1, x_2)\big)^{\alpha}. \tag{C.1}$$

Note that the case $\alpha = 1$ reduces to the Lipschitz property, so the Hölder property is a generalization of the Lipschitz property.

The parameter α above is often called the *Hölder exponent*, and the parameter M the *Hölder constant*. Like the Lipschitz property, the Hölder property is also a sufficient condition for continuity. The proof is a good exercise.

Exercise C.1 (Hölder property implies continuity)**.**
Suppose that $f\colon X \to Y$ is (M, α)-Hölder for some $M \geq 0$ and $\alpha > 0$. Show that f is continuous.

Correspondingly, functions satisfying the Hölder property of Definition C.1 are said to be *Hölder continuous*. In this context it is still common to still specify the Hölder exponent α and say that a function is α-Hölder-continuous, while the constant M is typically left unspecified. One reason for this is that when the codomain Y is $\mathbb{R}$ (or more generally a normed space), then the set of α-Hölder-continuous functions is a vector space.

Exercise C.2 (The vector space of α-Hölder continuous functions)**.**
Let $\alpha > 0$. Show that the subset

$$\mathcal{C}^{0,\alpha}(X) = \Big\{ f\colon X \to \mathbb{R} \;\Big|\; f \text{ is } (M, \alpha)\text{-Hölder for some } M \geq 0 \Big\} \;\subset\; \mathbb{R}^X$$

is a vector subspace of the space $\mathbb{R}^X$ of all functions $X \to \mathbb{R}$.

In favorable situations it is possible to equip the space $\mathcal{C}^{0,\alpha}(X)$ of α-Hölder-continuous functions $X \to \mathbb{R}$ with a norm which takes into account the Hölder constant. We refer to subsequent analysis courses for details.

Modulus of continuity

UNDER CONSTRUCTION!

Bilipschitz functions

UNDER CONSTRUCTION!

Appendix D

The starting point of general topology

D.1. ♡ Axioms and examples of topological spaces

The general definition of a topological space is the following.

Definition D.1 (Topology).
Let X be a set. A collection $\mathscr{T}$ of subsets of X is said to be a **topology** on X, if:

(i) the empty set and the whole space belong to the collection, i.e., $\emptyset \in \mathscr{T}$ and $X \in \mathscr{T}$;
(ii) the collection is stable under arbitrary unions, i.e., if $(V_j)_{j \in J}$ is an indexed collection of sets such that $V_j \in \mathscr{T}$ for all $j \in J$, then also $\bigcup_{j \in J} V_j \in \mathscr{T}$;
(iii) the collection is stable finite intersections, i.e., if $V_1, \ldots, V_n \in \mathscr{T}$, then also $V_1 \cap \cdots \cap V_n \in \mathscr{T}$.

A set X equipped with a topology $\mathscr{T}$ on it is called a **topological space**.

To concisely and unambiguously mention both the set X and the topology $\mathscr{T}$ on it, we may refer to the *pair* $(X, \mathscr{T})$ as a topological space.

Example D.2 (The topology induced by a metric).
Let (X, d) be a metric space and

$$\mathscr{T} = \left\{ U \subset X \;\middle|\; U \text{ is an open set} \right\}.$$

Then $\mathscr{T}$ satisfies the properties in Definition D.1: property (i) was seen in Examples VII.3 and VII.4, property (ii) in Theorem VII.6(a), and property (iii) in Theorem VII.6(b). The topology $\mathscr{T}$ defined in such a way is said to be the *topology induced by the metric* d.

In a general topological space $(X, \mathscr{T})$, subsets $U \subset X$ which belong to the topology, $U \in \mathscr{T}$, are also called **open sets**.

Remark D.3 (All metric spaces are topological spaces).
By Example D.2, every metric space becomes a topological space (with the topology induced by the metric).

As a consequence, the familiar Euclidean spaces, or more generally all inner product spaces and normed spaces become topological spaces (they have metrics induced by their norms).

But not every topology comes from a metric — as a simple example, consider the following.

Example D.4 (The trivial topology).
Let X be any set. The collection $\mathscr{T} = \{\emptyset, X\}$ consisting of only the empty set and the whole

space satisfies the properties in Definition D.1, and is therefore a topology on X. It is called the **trivial topology** on X.

Exercise (✓) D.1 (The trivial topology is not metrizable).
Assume that X is a set with more than one point. Show that there does not exist a metric on X whose induced topology is the trivial topology.

Another easy example of a topology is the following.

Example D.5 (The discrete topology).
Let X be any set. The collection $\mathscr{T}$ consisting of all subsets of X obviously satisfies the properties in Definition D.1, and is therefore a topology on X. It is called the **discrete topology** on X.

We encountered discrete topology first in Example VII.8. That example shows that the discrete topology is the topology induced by the $0/1$-metric on X.

Definition D.6 (Subspace topology).
Let $(X, \mathscr{T})$ be a topological space and $X' \subset X$ a subset. The subspace topology on X' is the collection

$$\mathscr{T}' = \left\{ U' \subset X' \;\middle|\; \exists U \in \mathscr{T} : \; U' = U \cap X' \right\}. \tag{D.1}$$

Exercise D.2 (The subspace topology is a topology).
Prove that $\mathscr{T}'$ in (D.1) is indeed a topology on X', i.e., that it satisfies the three properties in Definition D.1.

In metric spaces, we considered how a metric is inherited to a subspace by (V.8). In Theorem VII.9 we then found that the open sets in the subspace are then exactly of the form (D.1) (this is both a consistency property of the definitions of subspace metric and subspace topology, and in fact a key motivation for the latter definition).

D.2. ♡ Notions and results in topological spaces

In our study of metric spaces, we were able to characterize many properties purely in terms of open sets. These characterizations can then serve as definitions of the properties in general topological spaces (they are consistent with the earlier usage by virtue of having been logically equivalent in the metric space generality to the definitions which made reference to the metrics).

Continuity

A prime example is that of *continuity*.

Definition D.7 (Continuity of functions between topological spaces).
Let $(X, \mathscr{T}_X)$ and $(Y, \mathscr{T}_Y)$ be two topological spaces. A function $f \colon X \to Y$ is **continuous** if for every $V \in \mathscr{T}_Y$ we have $f^{-1}[V] \in \mathscr{T}_X$.

Preimages of open sets being open was exactly the characterization of continuity we found in metric space generality in Theorem VII.22.

Those continuity proofs that only made use of this characterization (and made no reference to the metrics) also carry over verbatim! You may for example observe the following:

- A constant function from one topological space to another is continuous. (The argument in Example VII.24 works in topological spaces as such.)
- The identity function on a topological space is a continuous function. (The argument in Example VII.25 works in topological spaces as such.)
- The composition $g \circ f$ of continuous functions $f \colon X \to Y$ and $g \colon Y \to Z$ between topological spaces X, Y, Z is continuous.
  (The proof of Theorem VII.29 works in topological spaces as such.)

Exercise D.3 (Inclusion of a subspace to a topological space is continuous).
Let X be a topological space and $X' \subset X$ a subset. Recall that X' also becomes a topological space by Definition D.6. Show that the inclusion function $\iota \colon X' \to X$ is continuous.

Homeomorphism

Having defined continuous functions between topological spaces, the notion of homeomorphisms remains basically unchanged from Definition VII.30:

Definition D.8 (Homeomorphism).
Let $(X, \mathscr{T}_X)$ and $(Y, \mathscr{T}_Y)$ be topological spaces. A function $f \colon X \to Y$ is called a **homeomorphism** if it is bijective and both $f \colon X \to Y$ and its inverse $f^{-1} \colon Y \to X$ are continuous.

If a homeomorphism $f \colon X \to Y$ exists, the (topological) spaces X and Y are said to be **homeomorphic**, and we denote $X \approx Y$.

One easily verifies that Lemma VII.31 goes through unchanged[1], so:

- Homeomorphism is an equivalence relation among topological spaces.

Product topology

UNDER CONSTRUCTION!

Properties specific to metric spaces

Not everything from the present course generalizes to topological spaces, however. As quick indications and warnings, let us mention:

[1]Besides basic set theory, the only relevant observations needed in its proof are the continuity of the identity function and the continuity of compositions of continuous functions, which remain valid in the generality of topological spaces.

- Limits of sequences are not necessarily unique in topological spaces![2]
- Continuity cannot in general be characterized by limits of sequences.[3]
- Sequential compactness is not in general equivalent to covering compactness; neither implies the other in the full generality of topological spaces.
- Neither completeness nor uniform continuity can even be defined in general topological spaces: both genuinely depend on the metric rather than just the topology.[4]

[2]Even in general topology, one mostly cares about Hausdorff topological spaces, and in those, limits are still unique.

[3]A countability property for generation of neighborhoods of points is sufficient to rescue this fact.

[4]There is, however, a less commonly used notion of uniform spaces, which is more general than metric spaces but less general than topological spaces, and in which these make sense without a metric.

Appendix E

Applications of function sequences

♡ Power series

Power series are a very common form of function series. They play a particularly prominent role in complex analysis, but they are also used heavily in other branches of mathematics. Let us take a brief look at what the tools we now have immediately say about power series, focusing on the real-valued case for concreteness. Let us also focus on power series expanded around the origin, to simplify the notation slightly.

By a power series with coefficients $c_0, c_1, c_2, \ldots \in \mathbb{R}$ we mean the function series

$$f(x) \; = \; \sum_{k=0}^{\infty} c_k \, x^k. \tag{E.1}$$

In this case the partial sums are polynomials

$$f_n(x) \; = \; \sum_{k=0}^{n} c_k \, x^k \; = \; c_n \, x^n + c_{n-1} \, x^{n-1} + \cdots + c_1 \, x + c_0,$$

so in particular they are continuous functions (on whichever subset of the real line we choose to consider as their domain).

To start addressing the convergence of (E.1), an important quantity depending on the coefficients $(c_k)_{k \in \mathbb{N}_0}$ is the *radius of convergence* R given by[1]

$$R = \frac{1}{\limsup_n \; |c_n|^{1/n}} \tag{E.2}$$

(with interpretation $R = 0$ if the limsup in the denominator is $+\infty$, and $R = +\infty$ if the limsup in the denominator is 0). Suppose now that $r \geq 0$ is strictly less than the radius of convergence, $r < R$, and use the closed interval $[-r, r]$ as the domain in which the series (E.1) is considered.[2] Choose $R_0 \in (r, R)$. From (E.2) one can then straightforwardly deduce that there exists a $C > 0$ such that for all $k \in \mathbb{N}_0$ we have $|c_k| \leq C \, R_0^{-k}$. We will apply Weierstrass' M-test with $M_k = C \, (r/R_0)^k$, so let us check the two assumptions. First of all, for $x \in [-r, r]$ and $k \in \mathbb{N}_0$ we indeed have $|c_k \, x^k| \leq |c_k| \, r^k \leq C \, R_0^{-k} \, r^k = M_k$. Also, the required property $\sum_{k=0}^{\infty} M_k < +\infty$ is simply the convergence of a geometric series with ratio $r/R_0 < 1$. The Weierstrass' M-test thus applies: the power series (E.1) converges uniformly on the interval $[-r, r]$, and defines a continuous function $f \colon [-r, r] \to \mathbb{R}$.

[1] Here, for a sequence (a_n) of nonnegative real numbers, the *limit superior* $\limsup_n a_n$ means $\lim_{n \to \infty} \sup \{ a_m \,|\, m \geq n \}$. If the sets $\{ a_m \,|\, m \geq n \}$ are unbounded for all n, or if the sequence of their supremums is unbounded, then it is a natural convention that the limit superior equals the symbol $+\infty$.

[2] If $R = 0$, of course no such r exists! Indeed when the radius of convergence is zero, we are not going to make any convergence statements. More interesting conclusions will be obtained in the case $R > 0$, in which choosing $0 \leq r < R$ is indeed possible.

Above, we allowed any $r < R$, and we got the uniform convergence and therefore also pointwise convergence of the power series (E.1) on the closed interval $[-r, r]$. The union of these allowed closed intervals is $\bigcup_{0 \leq r < R} [-r, r] = (-R, R)$, an open interval. Pointwise convergence of the power series (E.1) holds on $(-R, R)$, too.[3] Also continuity of still holds on the open interval $(-R, R)$.[4] However, on the open interval $(-R, R)$, the convergence of the power series (E.1) is not necessarily uniform.[5]

A very important example of a power series is the exponential function

$$e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k,$$

where the radius of convergence is $R = +\infty$ (easy calculation). From the above we conclude that the series is uniformly convergent on any closed interval of the form $[-r, r]$, pointwise convergent on the whole real line $(-\infty, +\infty)$, and that $x \mapsto e^x$ is continuous on the whole real line $(-\infty, +\infty)$. Similarly, these simple first results on power series yields that the trigonometric functions

$$\sin(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1} \quad \text{and} \quad \cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}, \tag{E.3}$$

are uniformly convergent on any closed interval of the form $[-r, r]$, pointwise convergent on the whole real line, and continuous on the whole real line.

With similar techniques, in analysis one would proceed to prove for example that the function $f\colon (-R, R) \to \mathbb{R}$ determined by the power series (E.1) is infinitely differentiable, and its derivatives are the power series obtained by differentiating the series termwise.

It is trivial to adapt these considerations to power series expanded around a different point $x_0 \in \mathbb{R}$, i.e., series of the form $f(x) = \sum_{k=0}^{\infty} c_k \, (x - x_0)^k$. Likewise, in the case of complex coefficients $c_0, c_1, c_2, \ldots \in \mathbb{C}$ and complex-valued functions of a complex variable, one fixes $z_0 \in \mathbb{C}$ and considers the power series $f(z) = \sum_{k=0}^{\infty} c_k \, (z - z_0)^k$, with no essential changes to the above: we get uniform convergence in all closed balls $\overline{\mathcal{B}}_r(z_0) \subset \mathbb{C}$ of radii $r < R$, and pointwise convergence and differentiability in the open ball $\mathcal{B}_R(z_0) \subset \mathbb{C}$.

---

[3] Easy exercise: Why?

[4] Another easy exercise: Why?

[5] Exercise: Find a counterexample. Hint: *The geometric series* $\sum_{k=0}^{\infty} x^k$ *is a power series. What is its radius of convergence? Does it converge uniformly for* x *on the open interval* $(-1, 1)$*?*

Index

References

Bibliography

[Car00] N. L. Carothers. *Real analysis*. Cambridge University Press, 2000.

[Hal74] Paul R. Halmos. *Naive set theory*. Undergraduate Texts in Mathematics. Springer, 1974.

[Rud76] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill, 1976.

[Väi99] Jussi Väisälä. *Topologia I*. Limes ry., 1999.