EXPANDED LECTURE NOTES FOR THE COURSE

MATH 131AH: HONORS ANALYSIS

WINTER QUARTER 2026

BY MAREK BISKUP

CONTENTS

1. Propositional logic 1
2. Set theory 6
3. Relations and functions 11
4. The naturals 17
5. Arithmetic of the naturals 23
6. Integers and rationals 27
7. Ordered fields 31
8. Algebraic deficiencies of rationals 35
9. Supremum and infimum 41
10. The reals via Dedekind cuts 46
11. Properties of the reals 54
12. Cardinality and countability 60
13. Uncountable sets and beyond 67
14. Metric space convergence 74
15. Basic topology 82
16. Sequences and point-set topology 88
17. Completeness 94
18. Contraction maps and completion 100
19. Sequential compactness 109
20. Compactness in topology 115
21. Limsup and liminf 121
22. Infinite series 128
23. Absolute vs conditional convergence 134

Note: Preliminary version, comments welcome!

1. PROPOSITIONAL LOGIC

Mathematics depends on precise statements for which the common (English) language is often too inaccurate and at times even plainly ambiguous. For this reason, mathematical statements are often cast in the dry language (and formalism) of propositional logic. We thus start by reviewing the important concepts from propositional logic that we will need throughout the rest of the course.

1.1 Logical propositions and operations.

A basic building block is a *proposition* which is simply a phrase or a statement (expressed using the *language* and *objects* of the theory) that can be assigned a TRUE/FALSE value. Examples of propositions is the statement $1 < 0$ or the statement

$$x^2 + 2x + 1 = 0 \text{ has no solution over } \mathbb{R}. \tag{1.1}$$

We will denote propositions P or Q in the sequel.

There are natural operations on propositions that produce new propositions. These correspond (roughly) to English words

$$\text{NOT, AND, OR, IMPLIES and IS EQUIVALENT TO} \tag{1.2}$$

For propositions P, Q these are denoted by

$$\neg P,\ P \wedge Q,\ P \vee Q,\ P \Rightarrow Q \text{ and } P \Leftrightarrow Q, \tag{1.3}$$

and are referred to as *negation*, logical *conjunction*, logical *disjunction*, *implication* and *equivalence*, respectively. To define these new propositions precisely, we give a list of their TRUE/FALSE values depending on those of P and Q. This results in

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$
TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE

which we refer to as the *truth table* for these operations.

Two points are worthy of a note. First, the OR represented by logical disjunction $\vee$ is not what is usually meant by "or" in idiomatic English. Indeed, when a host offers her guest a "beer or wine," she presumably means exactly one of the two while the logical OR — and likely also the thirsty guest — enjoys the possibility of taking both drinks concurrently. This arises because we often mistake OR for *exclusive-OR*, a.k.a. XOR, which is defined as

$$P \text{ XOR } Q := (P \wedge \neg Q) \vee (\neg P \wedge Q) \tag{1.4}$$

The logical OR is sometimes translated into English by the phrase and/or but that requires disambiguation with logical AND.

Another point to note is the somewhat unintuitive convention that if P is FALSE, then $P \Rightarrow Q$ is TRUE regardless of the truth value of Q. This subtle point shows the utmost importance of soundness of logical reasoning: Indeed, it implies that the introduction of

a single FALSE statement into the language or theory permits a logical inference of any statement and, consequently, the collapse of the whole theory.

Caution is at times needed to properly interpret a logical statement written in English as a statement in propositional logic. The most common situation arises with phrases "if" and "only if". These are meant to be read as follows

$$\text{“}P \text{ only if } Q\text{”} \text{ means } P \Rightarrow Q \tag{1.5}$$

while

$$\text{“}P \text{ if } Q\text{”} \text{ means } Q \Rightarrow P \tag{1.6}$$

One can easily err when converting the English phrase into the logical proposition so we should pay extra attention to these in what follows.

The above phrases commonly occur jointly in statements of form "P if and only if Q" which we take to mean $P \Leftrightarrow Q$. To give an example of the above, the phrase "x is positive only if x equals it absolute value" translates into

$$x > 0 \Rightarrow x = |x| \tag{1.7}$$

while "x is positive if x equals its absolute value" translates into

$$x = |x| \Rightarrow x > 0 \tag{1.8}$$

Note that the former statement is TRUE for all real-valued x, while the latter is FALSE for at least one x (namely, $x = 0$).

1.2 Primitive operations and equivalent forms.

Not all of the above operations are primitive; in fact, it would suffice to have just NOT and, for instance, OR. Indeed, with "$\neg$" and "$\vee$" defined by the corresponding columns in the above truth table, we could then set

$$\begin{aligned} P \wedge Q &:= \neg(\neg P \vee \neg Q) \\ P \Rightarrow Q &:= \neg P \vee Q \end{aligned} \tag{1.9}$$

and, with $\wedge$ and $\Rightarrow$ reduced to the primitives as above, let

$$P \Leftrightarrow Q := (P \Rightarrow Q) \wedge (Q \Rightarrow P) \tag{1.10}$$

where we employ the convention that the NOT operation is always taken first, unless parentheses get in the way.

We also note that, here and henceforth,

$$\text{“:=” means “defined as”} \tag{1.11}$$

where the latter (unlike for ordinary equality) requires only that the object on the right-hand side is defined; the object on the left is then identified with that on the right (and so an ordinary equality henceforth applies).

That the first line (1.9) gives the desired concept is verified with the help of the fact that NOT is an *involution*,

$$\neg(\neg P) \Leftrightarrow P \tag{1.12}$$

and that the *de Morgan formulas*

$$\begin{aligned} \neg(P \wedge Q) &\Leftrightarrow (\neg P \vee \neg Q) \\ \neg(P \vee Q) &\Leftrightarrow (\neg P \wedge \neg Q) \end{aligned} \tag{1.13}$$

hold true. The second line in (1.9) is then deduced from the following lemma:

Lemma 1.1 (Proof by contradiction) *For any logical propositions P and Q,*

$$(P \Rightarrow Q) \;\Leftrightarrow\; \neg(P \wedge \neg Q) \tag{1.14}$$

Proof. This can be checked directly by working out the truth tables. Alternatively, note that $P \Rightarrow Q$ is FALSE only in one case; namely, when P is TRUE and Q is FALSE. This is also the only situation when $P \wedge \neg Q$ is TRUE. □

The title of the lemma refers to the fact that (1.14) is the logical basis for proofs by contradiction. These demonstrate the implication "P implies Q" to be TRUE by way of assuming P along with the negation of Q and using these and a sequence of logical inferences to deduce an absurd (meaning: demonstrably FALSE) conclusion. That being said, the assumption that Q is FALSE is often redundant in this argument. In such cases, one rather relies on the following proof strategy:

Lemma 1.2 (Proof by contrapositive) *For any logical propositions P and Q,*

$$(P \Rightarrow Q) \;\Leftrightarrow\; (\neg Q \Rightarrow \neg P) \tag{1.15}$$

Proof. We again readily check that both sides are FALSE only for the situation in which P is TRUE and Q is FALSE. □

While the unary operation NOT must be included among the primitives, the reliance on OR as the primitive binary operation is a matter of choice. Indeed, by de Morgan formulas, AND could be used instead and, in light of (1.9), IMPLIES could be used as well. Note that although 8 distinct binary operations exist on propositions P and Q, they can all be written as strings of logical disjunctions of logical conjunctions involving P, $\neg P$, Q and $\neg Q$ and are thus reducible to NOT and OR above. That being said, the reduction to primitives is important mainly for theory building, and perhaps implementation on a computer, so we will keep using the full set above in the sequel.

1.3 Predicates and quatifiers.

Most of mathematics deals with propositions that depend on parameters. Such propositions are called *predicates* or *Boolean functions*. The TRUE/FALSE value of the predicate $P(x)$ is generally not determined until x is specified. In order to turn predicates into proper propositions, we thus need to specify, or curb, x somewhat. This is done using the following *quantifiers*

$$\begin{aligned} \forall &:= \text{ for all} \\ \exists &:= \text{ there exists} \end{aligned} \tag{1.16}$$

With the help of these we define propositions

$$\forall x\colon P(x) \;:=\; P(x) \text{ is TRUE for every choice of } x \tag{1.17}$$

where, alternatively, the phrase on the right means that the Boolean function $x \mapsto P(x)$ has no FALSE in its range. Similarly,

$$\exists x\colon P(x) \;:=\; P(x) \text{ is TRUE for at least one choice of } x \tag{1.18}$$

which in the language of Boolean functions means $x \mapsto P(x)$ has TRUE in its range. Another word used in the above context is *satisfiability*. The phrase $\exists x\colon P(x)$ means that $P(x)$ is satisfiable while $\forall x\colon P(x)$ means that $\neg P(x)$ is NOT satisfiable.

We may often need to further restrict the x for which the quantifier applies. This requires the language of set theory which, for the sake of easier explanation, we assume the reader to be familiar with. Given a set E, we write

$$\forall x \in E\colon P(x) \;:=\; \forall x\colon\; x \in E \Rightarrow P(x) \tag{1.19}$$

and

$$\exists x \in E\colon P(x) \;:=\; \exists x\colon\; x \in E \wedge P(x) \tag{1.20}$$

Examples of these are the predicates (depending on a set $E \subseteq \mathbb{R}$)

$$\forall x \in E\colon x^2 > 0 \tag{1.21}$$

which is TRUE if $0 \notin E$ and FALSE otherwise and

$$\exists x \in E\colon x^2 > 0 \tag{1.22}$$

which is FALSE if $E \subseteq \{0\}$ and TRUE otherwise.

Having both quantifiers around is again done mainly for comfort of expression as one can be reduced to the other. Indeed, we have:

Lemma 1.3 *For any predicate $P(x)$,*

$$\begin{aligned} \neg\big(\forall x\colon P(x)\big) &\;\Leftrightarrow\; \exists x\colon \neg P(x) \\ \neg\big(\exists x\colon P(x)\big) &\;\Leftrightarrow\; \forall x\colon \neg P(x) \end{aligned} \tag{1.23}$$

Proof. We will only prove the first equivalence employing the formalism of Boolean functions. Indeed, $\forall x\colon P(x)$ means that the range of the map $x \mapsto P(x)$ is the singleton $\{\text{TRUE}\}$. Since $P(x)$ takes only TRUE/FALSE values, the negation of $\forall x\colon P(x)$ means that the range of $x \mapsto P(x)$ contains FALSE. But this is the same as saying that the range of $x \mapsto \neg P(x)$ contains TRUE, which is then identified with $\exists x\colon \neg P(x)$. □

Formulas can depend on more than just one parameter some of which may need to be quantified. We use the convention that while quantifiers are read left to right, they are applied right to left. Indeed, given a formula $P(x, y)$ in two variables, we have

$$\forall x\, \exists y\colon P(x,y) \;:=\; \forall x\colon \Big(\exists y\colon P(x,y)\Big) \tag{1.24}$$

and, similarly,

$$\forall x\, \forall y\colon P(x,y) \;:=\; \forall x\colon \Big(\forall y\colon P(x,y)\Big) \tag{1.25}$$

As with most other operations in mathematics, the order quantifiers are applied matters for the result. We in fact have:

Lemma 1.4 *For any predicate $P(x, y)$,*

$$\forall x \forall y\colon\; P(x,y) \;\Leftrightarrow\; \forall y \forall x\colon\; P(x,y) \tag{1.26}$$

and

$$\exists x \exists y\colon\; P(x,y) \;\Leftrightarrow\; \exists y \exists x\colon\; P(x,y) \tag{1.27}$$

and

$$\exists x \forall y \colon\ P(x,y) \ \Rightarrow\ \forall y \exists x \colon\ P(x,y) \tag{1.28}$$

We leave the easy proof (based on similar arguments as those in the proof of Lemma 1.3) to the reader. Note that for mixed quantifiers, only one implication is claimed above because the other does not hold in general. For instance,

$$\forall m \in \mathbb{N}\, \exists n \in \mathbb{N} \colon\ m = n \tag{1.29}$$

is a TRUE statement while the one with the quantifiers reversed is not.

Parentheses have to be used whenever possible ambiguity arises. The position of the parentheses can of course change the logical content of the statement. For instance,

$$\forall n \in \mathbb{N} \colon \Big(\exists m \in \mathbb{N} \colon (m = n \Rightarrow 0 = 1) \Big) \tag{1.30}$$

is a TRUE statement (because choosing $m := n + 1$ makes $m = n$ FALSE and thus $m = n \Rightarrow 0 = 1$ TRUE) yet

$$\forall n \in \mathbb{N} \colon \Big((\exists m \in \mathbb{N} \colon m = n) \Rightarrow 0 = 1 \Big) \tag{1.31}$$

as well as

$$\Big(\forall n \in \mathbb{N}\, \exists m \in \mathbb{N} \colon m = n \Big) \Rightarrow 0 = 1 \tag{1.32}$$

are both FALSE. (Here $\mathbb{N}$ are the naturals with the usual interpretation of 0 and 1.)

2. SET THEORY

Every since about mid 1800s mathematicians realized that sets provide a useful tool to express mathematical statements and arguments in their proofs. We will now review some elementary facts from this theory.

2.1 Naive set theory.

We start with the basic setting of so called naive set theory which is is an early version of set theory in which all "sensible" operations on sets are allowed. We need two ingredients:

- The basic building blocks of naive set theory are *sets*, to be denoted by capital letters A, B, etc. Sets are basically "containers" collecting other objects which, at least in pure set theory, are themselves sets. For any two sets A, B, we thus assume the existence of a (logical) proposition $A \in B$, whose TRUE value designates that A *belongs to* or *is an element of* B. We will write $A \notin B$ for $\neg(A \in B)$.
- In order to be able to form sets from other sets, we put forward a basic assumption, termed the *Comprehension Principle*, which states that, for any predicate $P(X)$ whose parameter is a set,
$$\{X\colon P(X)\} \text{ is a set} \tag{2.1}$$
(Technically, this postulates existence of a unique set A such that $\forall X\colon X \in A \Leftrightarrow P(X)$. We then denote this A as $\{X\colon P(X)\}$.)

Using the Comprehension Principle we can construct many objects we are used to from prior experience with set theory. For instance, we can define the *empty set* by
$$\varnothing := \{X\colon \text{FALSE}\} \tag{2.2}$$
where "FALSE" stands for a logical proposition that takes only FALSE value. For a set A, we can define its *complement* by
$$A^{\mathrm{c}} := \{X\colon X \notin A\}. \tag{2.3}$$
There are also a number of familiar operations on pairs of sets. Indeed, given sets A, B,
$$\begin{aligned} A \cup B &:= \{X\colon X \in A \vee X \in Y\} \\ A \cap B &:= \{X\colon X \in A \wedge X \in Y\} \\ A \smallsetminus B &:= \{X\colon X \in A \wedge X \notin B\} \\ A \triangle B &:= (A \smallsetminus B) \cup (B \smallsetminus A), \end{aligned} \tag{2.4}$$
define their *union*, *intersection*, *set difference* and *symmetric difference*, respectively.

The Comprehension Principle is rather strong to give us far more than the above. For instance, for each set A we can define the *singleton* $\{A\}$ containing just A by
$$\{A\} := \{X\colon X = A\} \tag{2.5}$$
Here we used the equality sign "=" in the meaning of *sameness* or *identity*. (This symbol is sometimes introduced as part of the setup of the theory; if not, then we define it as $A = B \;:=\; \forall X\colon X \in A \Leftrightarrow X \in B$.) Similarly, we can *pair* two sets A and B into
$$\{A, B\} := \{X\colon X = A \vee X = B\}. \tag{2.6}$$

Another useful object is the set of all subsets of A, termed the *power set*, defined by

$$\mathcal{P}(A) := \{X\colon X \subseteq A\}, \tag{2.7}$$

where we made use of the binary relation *A is a subset of B* with the definition

$$A \subseteq B \;:=\; (\forall X \in A\colon X \in B). \tag{2.8}$$

(This is just a shorthand induced by the relation $\in$.)

A rather important consequence of the Comprehension Principle is that it implies existence of infinite sets. Indeed, first generalize the first line in (2.4) to

$$\bigcup A := \big\{X\colon (\exists B \in A\colon X \in B)\big\} \tag{2.9}$$

for the union of all sets contained in A. (Note that $A \cup B = \bigcup\{A, B\}$.) The set

$$I := \Big\{X\colon \bigcup X \subseteq X\Big\} \tag{2.10}$$

is then closed under the operation $X \mapsto X \cup \{X\}$ (meaning: $\forall X \in I\colon X \cup \{X\} \in I$) because

$$\bigcup\big(X \cup \{X\}\big) = X \cup \bigcup X \underset{X\in I}{=} X \subseteq X \cup \{X\}, \tag{2.11}$$

where we used various general facts $\cup$ and $\subseteq$ whose proof we leave to the reader.

Noting that $\varnothing \in I$ because $\bigcup \varnothing = \varnothing \subseteq \varnothing$, we now recursively check that I is *infinite* according to the following definition: A set B is said to be infinite if there is an injective map $f\colon B \to B$ such that $\mathrm{Ran}(f) := \{f(X)\colon X \in B\}$ is a proper subset of B (meaning: $\mathrm{Ran}(f) \subseteq B$ yet $\mathrm{Ran}(f) \neq B$). For I in (2.10), this is witnessed by the map

$$f(X) := X \cup \{X\} \tag{2.12}$$

which is injective because $X \cup \{X\} = Y \cup \{Y\}$ is reduced to $X = Y$ (for $X, Y \in I$) by taking the union (see the second equality in (2.11)), but is not onto because f does not have $\varnothing$ in its range. (The concepts of "injective", "onto", etc have yet to be introduced so this explanation is mainly for those already in command of these terms.) The intuition of the above construction is that I contains all elements in the set

$$\Big\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}, \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}, \dots\Big\} \tag{2.13}$$

which is "obviously" infinite.

Unfortunately, in the late 1800s it became gradually apparent that the naive set theory contains inconsistencies. These started at large infinite sets but, ultimately, surfaced in the following elementary mind-boggling paradox discovered by B. Russell in 1901:

Theorem 2.1 (Russell's antinomy) *In naive set theory,*

$$\{X\colon X \notin X\} \tag{2.14}$$

is not a set. In particular, the comprehension principle is inconsistent.

Proof. Suppose, by way to contradiction, $A := \{X\colon X \notin X\}$ is a set. Then either $A \in A$ is TRUE or $A \notin A$ is TRUE. But if $A \in A$ is TRUE then the predicate defining A forces $A \notin A$ while if $A \notin A$ is TRUE then $A \in A$ because otherwise A would otherwise be included in (2.14). Either possibility leads to a contradiction and so A is not a set. □

2.2 Zermelo's axiomatic.

It is fair to say that Russell's observation rattled the foundations of mathematics of that day. A number of solutions was gradually proposed some of which permeate various treatments of set theory till today. Here we will follow the solution proposed by E. Zermelo in 1908 which, ultimately, leads to the so called *ZFC set theory* used prevalently throughout analysis.

An important novelty of Zermelo's approach was that one should rely on *axiomatic* formulation of set theory modeled, in some way, on a similar approach to classical Euclidean geometry. To dispense with Russell's paradox, Zermelo proposed to restrict the Comprehension Principle to "containers" that are already sets:

- **Separation axiom:**
$$\forall B \text{ set}: \ \{X \in B \colon P(X)\} \text{ is a set} \tag{2.15}$$

Here we purposefully deviate from our earlier convention that tells us write the set (2.15) as $\{X \colon X \in B \wedge P(X)\}$. Since the point of writing $X \in B$ is to enforce the "separation," we will adhere to this practice throughout the course. Russell's argument then gives:

Lemma 2.2 *Let B be a set and let $A := \{X \in B \colon X \notin X\}$. Then $B \notin A \ \wedge \ A \notin A \ \wedge \ A \notin B$.*

Proof. Assuming $B \in A$ we get $B \in B$ AND $B \notin B$, a contradiction. So we must have $B \notin A$ as claimed. Similarly, $A \in A$ implies $A \notin A$, a contradiction. So $A \notin A$ holds as well. But $A \notin A$ then forces $A \notin B$ because otherwise we would have $A \in A$, a contradiction. □

Unfortunately, with the Comprehension Principle gone as stated, we lose the ability to perform many of the above constructions of sets — specifically, $\varnothing$ (as there could no sets at all), $A \cup B$ (as there could be no sets subsuming both A and B), $\{A\}$ (as there could be no set containing A) and, for similar reasons, $\mathcal{P}(A)$ and I. Further axioms are thus needed. We will now state Zermelo's axioms in bullet-point format:

- **Axiom of Extensionality:**
$$\forall A, B \ : \ A = B \Leftrightarrow \ \Big(\forall X \colon \ X \in A \Leftrightarrow X \in B\Big). \tag{2.16}$$
This axiom ensures that a set is uniquely determined by its elements.
- **Empty set axiom:**
$$\exists \varnothing \ \forall X \colon \ X \notin \varnothing \tag{2.17}$$
Thanks to Axiom of Extensionality, $\varnothing$ is the unique set with this property.
- **Pairset axiom:**
$$\forall X \ \forall Y \ \exists A \ \forall Z \colon \ Z \in A \ \Leftrightarrow \ \big(X = Z \vee Y = Z\big). \tag{2.18}$$
Again, the resulting set A is determined uniquely so we henceforth denote it $\{X, Y\}$ when X and Y are different and $\{X\}$ when $X = Y$.
- **Axiom of Union:**
$$\forall A \ \exists B \ \forall X \colon \ X \in B \Leftrightarrow \Big(\exists C \in A \colon X \in C\Big) \tag{2.19}$$
This is a bit hard to parse at first sight. Here A is a set of sets and B is the union of all elements in the elements of A. We use the notation $\bigcup A$ for B.

- **Powerset axiom:**
$$\forall A\, \exists B\, \forall X:\ X \subseteq A \Leftrightarrow X \in B \tag{2.20}$$
Here B is thus the powerset $\mathcal{P}(A)$, i.e., the set of all subsets, of A. Again, the powerset is unique by the Axiom of Extensionality.
- **Axiom of infinity:**
$$\exists I : \varnothing \in I \wedge \Big(\forall X\colon\ X \in I \Rightarrow \{X\} \in I\Big) \tag{2.21}$$
The notation $\{X\}$ is meaningful by the Pairset Axiom.

Here are some remarks on the above. First, as noted earlier, (2.16) requires the notion of identity — represented by the equality sign — to be part of the setup of the theory. Otherwise (2.16) can be read as a definition of "$=$" sign:
$$A = B \ :=\ \Big(\forall X\colon\ X \in A \Leftrightarrow X \in B\Big) \tag{2.22}$$
meaning that two sets are said to be equal when they have exactly the same elements. Yet another way to characterize equality is via:

Lemma 2.3 *We have*
$$\forall A, B\colon \quad A = B \ \Leftrightarrow\ A \subseteq B \wedge B \subseteq A \tag{2.23}$$

Proof. We first note that, given two predicates $P(x)$ and $Q(x)$ depending on same parameter x, we have
$$\big(\forall x\colon P(x)\big) \wedge \big(\forall x\colon Q(x)\big) \ \Leftrightarrow\ \forall x\colon P(x) \wedge Q(x) \tag{2.24}$$
To verify this, note that $\forall x\colon P(x)$ means that $\mathrm{Ran}(P)$ is $\{\mathrm{TRUE}\}$, and similarly $\forall x\colon Q(x)$ means $\mathrm{Ran}(Q) = \{\mathrm{TRUE}\}$. The proposition on the left thus equivalent to both ranges being equal $\{\mathrm{TRUE}\}$, which is equivalent to that on the right. To get (2.23) from this, we apply this to $P(X) := X \in A \Rightarrow X \in B$ and $Q(X) := X \in B \Rightarrow X \in A$ and invoke (2.8). □

Second, the set-theoretical notation $\bigcup A$ for the "union of all sets in A" is often substituted by other, more intuitive, expressions such as
$$\bigcup\{X\colon X \in A\} \quad \text{or even} \quad \bigcup_{X \in A} X \tag{2.25}$$
With the union postulated to exist, we can define general *intersection* by
$$\bigcap A := \Big\{X \in \bigcup A\colon (\forall C \in A\colon X \in C)\Big\}. \tag{2.26}$$
with similar alternative notations as in (2.25). Similarly, while we will keep writing $\mathcal{P}(A)$ for the power set of A in this course, in practice we often write $\{X\colon X \subseteq A\}$ in blatant violation of the Separation Axiom.

While the Emptyset Axiom ensures existence of a set, the Axiom of Infinity ensures existence of an infinite set. (Infinite sets would not exist otherwise in our universe; for instance, if our universe of sets is the collection of finite subsets of the naturals.) The Axiom of Infinity in particular ensures existence of a set that contains
$$\Big\{\varnothing, \{\varnothing\}, \{\{\varnothing\}\}, \{\{\{\varnothing\}\}\}, \dots\Big\} \tag{2.27}$$
as a subset. This axiom is phrased already with its important applications — namely, the construction of the naturals, which will arise from the set in (2.27) — in mind.

Zermelo's system includes one additional axiom, namely, the *Axiom of Choice* which we will get to in the next section. Another axiom — called *Axiom of Replacement* — was added later to the system by A. Fraenkel and, independently, T. Skolem leading to (what is now called) ZFC axiomatic — where the acronym stands for "Zermelo-Fraenkel with Axiom of Choice." The latter axiom is somewhat intricate to explain and we will not state it explicitly. As we shall see, we can actually get quite far into real analysis without needing either of them.

The Comprehension Principle can be retained in the axiomatic system by introducing the notion of a *class*. This is just another way to describe a "collection of sets with a given property" except that it is generally too large to be automatically called a set. By definition, every set is a class but the class is a set only if it is contained in a set. Russell's paradox then shows only that there are classes that are not sets.

Further details (and inspiration for the above presentation) can be found in Yannis Moschovakis' *Notes on Set Theory*, which is a wonderful advanced-undergraduate introduction to set theory.

3. Relations and functions

Once the axioms of set theory are in place, we can review some elementary albeit very useful constructions that these axioms enable. To make out notations easier, we will henceforth abandon our convention that all sets be written using the capital letters.

3.1 Cartesian product.

Given two sets A and B, a natural object to consider is the set of pairs (x, y) with the first taken from A and the second from B. This is formalized using the notion of their *Cartesian product*,

$$A \times B := \big\{(x, y) \in \mathcal{P}(\mathcal{P}(A \cup B))\colon x \in A,\, y \in B\big\}, \tag{3.1}$$

in which (x, y) denotes an *ordered pair* that, in set theory, is defined as

$$(x, y) := \big\{\{x\}, \{x, y\}\big\}, \tag{3.2}$$

(This is sometimes called the *Kuratowski pair*.) Here the set on the right exists by Pairset Axiom. An exercise in the use of Axiom of Extensionality shows:

Lemma 3.1 *Let A and B be sets. Then*

$$\forall x, \tilde{x} \in A\, \forall y, \tilde{y} \in B\colon \quad (x, y) = (\tilde{x}, \tilde{y}) \;\Leftrightarrow\; (x = \tilde{x} \;\wedge\; y = \tilde{y}). \tag{3.3}$$

In particular, the pair identifies its entries uniquely. The Cartesian product $A \times B$ is a set by Axiom of Separation.

The construction of the Cartesian product can naturally be iterated to construct the Cartesian product of three, four, etc sets. The problem is that order of operation matters, at least as far as the above definition is concerned. To demonstrate this, given three sets A, B and C and abbreviating $D := \mathcal{P}(\mathcal{P}(\mathcal{P}(\bigcup\{A, B, C\})))$, the above gives

$$A \times (B \times C) := \Big\{\big\{\{x\}, \{\{y\}, \{y, z\}\}\big\} \in D\colon x \in A \;\wedge\; y \in B \;\wedge\; z \in C\Big\} \tag{3.4}$$

while

$$\begin{aligned}&(A \times B) \times C\\ &\quad := \Big\{\big\{\{\{x\}, \{\{x\}, \{x, y\}\}\}, \{\{\{x\}, \{\{x\}, \{x, y\}\}\}, z\}\big\} \in D\colon x \in A \;\wedge\; y \in B \;\wedge\; z \in C\Big\}\end{aligned} \tag{3.5}$$

Yet, both sets should intuitively give the set of all triplets (x, y, z) and thus describe the same object modulo identification. Using terms that we will only introduce a bit later, this identification requires proving that

$$\big\{\{x\}, \{\{y\}, \{y, z\}\}\big\} \;\mapsto\; \big\{\{\{x\}, \{\{x\}, \{x, y\}\}\}, \{\{\{x\}, \{\{x\}, \{x, y\}\}\}, z\}\big\}. \tag{3.6}$$

defines a bijection of (3.4) onto (3.5); thus showing that the Cartesian product is *associative*. Once this is done, we drop the parentheses and write the "result" as $A \times B \times C$.

Unfortunately, any specific use of the triple product still requires specifying which of the two sets we refer to, and the issues with the identification become only worse when more sets are involved in the product. So we will ultimately abandon this approach and come up with a unified definition of the Cartesian product of any number of sets that is

void of these complications. For that we need the concept of a function which is in turn a special case of a relation, which is, however, based on the definition in (3.1).

3.2 Relations.

With the Cartesian product of two sets in hand, we can put forward:

Definition 3.2 *Given sets A and B, a* relation *on A and B is a subset $R \subseteq A \times B$. We say that $x \in A$ is in relation to $y \in B$, with the notation xRy, if the pair (x, y) lies in R, i.e.,*

$$xRy \;:=\; (x, y) \in R. \tag{3.7}$$

If $B = A$, we say that $R \subseteq A \times A$ is a relation on A.

An example of a relation is the set-inclusion $\subseteq$ on the power set $\mathcal{P}(A)$ of a set A. To describe this (and similar) relations formally, we give:

Definition 3.3 *Let A be a set and $R \subseteq A \times A$ a relation on A. We say that R is*

- reflexive, *if* $\forall x \in A\colon\ xRx$,
- antisymmetric, *if* $\forall x, y \in A\colon\ xRy \wedge yRx \Rightarrow x = y$, *and*
- transitive, *if* $\forall x, y, z \in A\colon\ xRy \wedge yRz \Rightarrow xRz$.

A relation which is reflexive, antisymmetric and transitive is called a partial order.

We leave it to homework for the reader to prove:

Lemma 3.4 *For any set A, the subset relation $\subseteq$ on $\mathcal{P}(A)$ is a partial order.*

Perhaps more familiar example of a relation that is reflexive, antisymmetric and transitive is the inequality $\leqslant$ on the number sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ (but not $\mathbb{C}$). As we will see when we construct these number sets from the foundations of set theory, the inequality relation $\leqslant$ will in fact be induced by $\subseteq$ applied to suitable sets.

In other to define our next frequent example of a relation, we put forward:

Definition 3.5 *A relation R on a set A is said to be* symmetric *if*

$$\forall x, y \in A\colon\ xRy \;\Leftrightarrow\; yRx. \tag{3.8}$$

Note that a symmetric relation need not be reflexive because we do not require xRx to actually hold. Nor are all pairs required to be in relation; all what symmetry says that if xRy then also yRx and *vice versa*. Also note that not being symmetric does not mean being antisymmetric, and *vice versa*. The role of symmetry is seen from:

Definition 3.6 *A relation $\sim$ on a set A that is reflexive, symmetric and transitive is called* equivalence. *For each $x \in A$, the set*

$$[x] := \{y \in A\colon y \sim x\} \tag{3.9}$$

is said to be the equivalence class of x.

The proof of the following lemma has been relegated to homework:

Lemma 3.7 *Let $\sim$ be an equivalence relation on a set A. Then*

$$\forall x, y \in A\colon [x] \cap [y] \neq \varnothing \;\Rightarrow\; [x] = [y] \tag{3.10}$$

This means that two equivalence classes are either disjoint or equal. Any element $z \in [x]$ is called a *representative* of $[x]$. In particular, x is a representative of $[x]$.

An example of an equivalence relation is the *equality* $=$, which is reflexive, symmetric and transitive. This is a very fine version of equivalence because (by Axiom of Extensionality) each equivalence class contains exactly one element; i.e., $\forall x \in A\colon [x] = \{x\}$. To give a more representative example, consider the following:

Lemma 3.8 *Writing $\mathbb{Z}$ for the set of integers and denoting by "$\cdot$" the standard operation of multiplication on $\mathbb{Z}$, let*

$$A := \big\{(m,n) \in \mathbb{Z} \times \mathbb{Z}\colon n \neq 0\big\} \tag{3.11}$$

and let $\sim$ be the relation on A defined as

$$(m,n) \sim (\widetilde{m},\widetilde{n}) \;:=\; m \cdot \widetilde{n} = \widetilde{m} \cdot n \tag{3.12}$$

Then $\sim$ is an equivalence.

Proof. Symmetry and reflexivity are checked readily; all that needs a bit of work is transitivity. Assume $(m,n) \sim (\widetilde{m},\widetilde{n})$ and $(\widetilde{m},\widetilde{n}) \sim (\widehat{m},\widehat{n})$. This means

$$m \cdot \widetilde{n} = \widetilde{m} \cdot n \;\;\wedge\;\; \widetilde{m} \cdot \widehat{n} = \widehat{m} \cdot \widetilde{n}. \tag{3.13}$$

Multiplying the first equality by $\widehat{n}$ results in

$$m \cdot \widetilde{n} \cdot \widehat{n} = \widetilde{m} \cdot n \cdot \widehat{n} = \widetilde{m} \cdot \widehat{n} \cdot n = \widehat{m} \cdot \widetilde{n} \cdot n \tag{3.14}$$

where where we used commutativity and associativity of multiplication throughout and invoked the second equality in (3.13) in the last step. Since $\widetilde{n} \neq 0$, the fact that

$$\forall k,l,m \in \mathbb{Z}\colon \quad (k \neq 0 \;\wedge\; k \cdot l = k \cdot m) \;\Rightarrow\; l = m \tag{3.15}$$

implies $m \cdot \widehat{n} = \widehat{m} \cdot n$ meaning that $(m,n) \sim (\widehat{m},\widehat{n})$. □

The punchline of this example is that the equivalence class $[(m,n)]$ then contains all pairs $(\widetilde{m},\widetilde{n})$ such that, informally (because division is not defined on $\mathbb{Z}$), obey $\frac{\widetilde{m}}{\widetilde{n}} = \frac{m}{n}$. The set of equivalence classes thus provides a construction of the set of rationals $\mathbb{Q}$.

3.3 Functions.

We now move to a concept central to analysis:

Definition 3.9 *Let A and B be sets. A relation $F \subseteq A \times B$ satisfying*

$$\forall x \in A\, \forall y,z \in B\colon xFy \wedge xFz \Rightarrow y = z \tag{3.16}$$

is called a function. *We will use notation $F(x)$ for the unique $y \in B$ such that xFy.*

Moving to the convention that functions can be denoted by lowercase letters, we will naturally think of a function f as an assignment of a value in B to a value in A, with the notation $f\colon A \to B$. The relation $f \subseteq A \times B$ corresponding to a function f is then the *graph of f*. Not every $x \in A$ may appear as the first member of a pair in relation R; those that do are collected in the *domain* of R, i.e.,

$$\mathrm{Dom}(R) := \big\{x \in A\colon (\exists y \in B\colon xRy)\big\} \tag{3.17}$$

Similarly, not every $y \in B$ is in relation with some $x \in A$; those that are form the *range* of R denoted as

$$\mathrm{Ran}(R) := \big\{y \in B \colon (\exists x \in A \colon xRy)\big\}. \tag{3.18}$$

We will write $\mathrm{Dom}(f)$, resp., $\mathrm{Ran}(f)$ for the domain, resp., range of the relation that is a function f. We remark that one often writes $f \colon A \to B$ without necessarily requiring that $\mathrm{Dom}(f) = A$.

If $R \subseteq A \times B$ and $S \subseteq B \times C$ are two relations, then their *composition* RS is the relation defined by

$$\forall x \in A \ \forall y \in C \colon \quad x(RS)y \ := \ \exists z \in B \colon xRz \ \wedge zSy. \tag{3.19}$$

If R and S are (graphs of) functions r and s, then their composition is also a function but (beware!) for functions the composition is written in reverse order; namely,

$$RS \text{ is the graph of } s \circ r \tag{3.20}$$

This is because functions "act" on the variable to the right while relations "act" on the variable to the left. Note also that for any transitive relation R we have $RR = R$.

The *inverse* R^{-1} of a relation R is the subset of $B \times A$ defined by

$$\forall x \in A \, \forall y \in B \colon \quad yR^{-1}x \ := \ xRy \tag{3.21}$$

Note that RR^{-1} is an identity relation on $\mathrm{Dom}(R)$ (which is a subset of A) and $R^{-1}R$ is an identity relation on $\mathrm{Ran}(R)$ (which is a subset of B). If F is (the graph of) a function f, then F^{-1} is (the graph of) a function if and only if

$$\forall x, \tilde{x} \in A \, \forall y \in B \colon \big(y = f(x) \ \wedge \ y = f(y)\big) \ \Rightarrow \ x = \tilde{x} \tag{3.22}$$

The inverse function is then denoted by f^{-1}.

Given a function $f \colon A \to B$, for each $C \subseteq A$ we define the *image* $f(C)$ of C by

$$f(C) := \big\{y \in B \colon (\exists x \in C \cap \mathrm{Dom}(f) \colon y = f(x))\big\} \tag{3.23}$$

sometimes written simply as $\{f(x) \colon x \in C\}$ ignoring that $f(x)$ may not be defined for all $x \in C$. The image of the domain is then the range, $\mathrm{Ran}(f) = f(\mathrm{Dom}(f))$. Similarly, for all $D \subseteq B$ we define the *preimage* $f^{-1}(D)$ of D by

$$f^{-1}(D) := \big\{x \in A \colon x \in \mathrm{Dom}(f) \ \wedge \ f(x) \in D\big\} \tag{3.24}$$

which we at times write as $\{x \colon f(x) \in D\}$ ignoring domain restrictions. The preimage of the range is the domain, $f^{-1}(\mathrm{Ran}(f)) = \mathrm{Dom}(f)$. We caution the reader that the use of f^{-1} in the preimage map does not require the inverse f^{-1} of f to exist. Notwithstanding, if f^{-1} does exist, then $f^{-1}(D)$ is the image of D under f^{-1}.

While functions in analysis are generally of interest for their analytic properties, in the rest of mathematics functions are primarily used to identify sets with one another. For this we need the following concepts:

Definition 3.10 *A function $f \colon A \to B$ is*

- injective *if* $\mathrm{Dom}(f) = A$ *and* $\forall x, y \in A \colon f(x) = f(y) \Rightarrow x = y$
- surjective *if* $\mathrm{Ran}(f) = B$ *meaning* $\forall y \in B \exists x \in A \colon f(x) = y$,
- bijective *if it is both injective and surjective.*

Exhibiting a bijection between two sets puts these in *one-to-one* or *bijective correspondence*, which are simply different ways to talk about a bijection. For instance, the map (3.6) defines a bijection $f\colon (A \times B) \times C \to A \times (B \times C)$ that permits us to write this as $A \times B \times C$. A very fruitful use of above concepts is in the following insightful definition due to G. Cantor:

Definition 3.11 (Equinumerosity) *Sets A and B are said to be* equinumerous*, or are of the same* cardinality*, if there exists a bijection $f\colon A \to B$.*

Note that given a set of sets, equinumerosity is another example of equivalence relation. Indeed, reflexivity is provided by the identity map, symmetry by the inverse map (which uses that the inverse of a bijection is a bijection) and transitivity by composed maps (which uses that a composition of two bijections is a bijection). Related to this is:

Definition 3.12 *A set A is said to be* Dedekind infinite *if there is an injection $f\colon A \to A$ such that* $\mathrm{Ran}(f) \neq A$.

This is one way to define the notion of an infinite set, albeit one that is not generally used in mathematics. We will return to these concepts when we discuss the question of cardinality in more detail.

3.4 General Cartesian products.

Relying on the concept of a function, the notion of the Cartesian product can be generalized to products of arbitrary collections of sets. Such collections is typically written as $\{A_\alpha\colon \alpha \in I\}$, where α represents the *index* of A_α and I denotes the *index set*. While this concept is quite intuitive, the reader may wonder how to phrase it in the formal language of the set theory. This comes in:

Definition 3.13 (Collections of sets) *Given sets I and A, a* collection $\{A_\alpha\colon \alpha \in I\}$ of subsets of A indexed by I *is the set* $\mathrm{Ran}(\phi)$ *for a map $\phi\colon I \to \mathcal{P}(A)$ such that* $\mathrm{Dom}(h) = I$*. Under this map, the sets in the collection are identified by $A_\alpha := \phi(\alpha)$.*

We note that, formally, $\phi \in \mathcal{P}(I \times \mathcal{P}(A))$ and so $\mathrm{Ran}(\phi)$ is indeed a set. (This is why we need all A_α's be subsets of one set.) We then put forward:

Definition 3.14 (General Cartesian product) *Given a set I and a collection $\{A_\alpha\colon \alpha \in I\}$ of sets (all of which have to be subsets of a given set) indexed by I,*

$$\bigtimes_{\alpha\in I} A_\alpha := \left\{ f \in \mathcal{P}\Big(I \times \bigcup_{\alpha\in I} A_\alpha\Big)\colon \textit{function} \wedge \mathrm{Dom}(f) = I \wedge (\forall \alpha \in I\colon f(\alpha) \in A_\alpha) \right\} \quad (3.25)$$

is the Cartesian product of the sets in $\{A_\alpha\colon \alpha \in I\}$.

The notation is easier to parse once we note that a function $f\colon I \to \bigcup_{\alpha\in I} A_\alpha$ is formally a subset of $I \times \bigcup_{\alpha\in I} A_\alpha$, which (by Unionset and Powerset axioms and our earlier construction of the Cartesian product) means that (3.25) is a set.

To see that (3.25) subsumes our earlier definition of the Cartesian product, note that a function f defined on two-point set $\{0, 1\}$ is determined by the pair of values $(f(0), f(1))$. The set of all functions with $f(0) \in A$ and $f(1) \in B$ is thus in a bijective correspondence with the set of all pairs (a, b) with $a \in A$ and $b \in B$. This identifies $A \times B$ with the set all

functions $f\colon \{0,1\} \to A \cup B$ satisfying $f(0) \in A$ and $f(1) \in B$, and thus the above general Cartesian product for $I := \{0,1\}$, $A_0 := A$ and $A_1 := B$. We will henceforth not make a distinction between the two ways to define the Cartesian product of two sets.

A special case (mainly notation) of above definition is:

Definition 3.15 *Let I and A be sets. Then*

$$\bigtimes_{\alpha\in I} A := \big\{f \in \mathcal{P}(I \times A)\colon \ \textit{function}\big\} \tag{3.26}$$

is the Cartesian power *denoted, in short, by A^I.*

For instance, the set of all real-valued sequences form the set $\mathbb{R}^{\mathbb{N}}$ while $\mathbb{R}^{\mathbb{R}}$ denotes the set of all real-valued functions of one real-valued variable.

The Cartesian product of two non-empty sets is non-empty (which is witnessed by a pair (x,y) such that $x \in A$ and $y \in B$), and the same applies to Cartesian products three, four, etc sets. However, a perplexing issue arises once I is infinite where this argument can no longer be made because we have no way to string an infinite number of such statements together. In the ZFC theory, this is resolved by imposing yet another axiom:

- **Axiom of Choice:** For each nonempty set I and all collections $\{A_\alpha\colon \alpha \in I\}$ of sets satisfying $\forall \alpha \in I\colon A_\alpha \neq \varnothing$, there exists a function $f\colon I \to \bigcup_{\alpha\in I} A_\alpha$ such that
$$\mathrm{Dom}(f) = I \ \wedge\ \forall \alpha \in I\colon f(\alpha) \in A_\alpha \tag{3.27}$$
In short,
$$\forall I\, \forall \{A_\alpha\colon \alpha \in I\}\colon \ \Big(I \neq \varnothing \ \wedge\ \big(\forall \alpha \in I\colon A_\alpha \neq \varnothing\big)\Big) \ \Rightarrow\ \bigtimes_{\alpha\in I} A_\alpha \neq \varnothing \tag{3.28}$$

The name arises from the observation that a function $f \in \times_{\alpha\in I} A_\alpha$ gives us a *choice*, simultaneously for all $\alpha \in I$, of a representative $f(\alpha) \in A_\alpha$ provided, of course, $A_\alpha \neq \varnothing$. In sets with some underlying structure, this can sometimes be guaranteed constructively but that will not work in general which is why such an axiom is needed.

We remark that in sets with some underlying structure a "distinguished element" can often be "picked" constructively, but this is not possible in general. The role of the Axiom of Choice itself is to supply such a "distinguished element" in full generality. Notwithstanding, mathematicians find the Axiom of Choice generally less acceptable than the rest of Zermelo's axioms, and so it is a good practice (to which we will adhere) to caution the reader whenever it is invoked.

4. THE NATURALS

We are now sufficiently acquainted with set theoretical foundations to move to the definition of natural numbers. First we observe that the intuitive definition

$$\mathbb{N} := \{0, 1, 2, \dots\} \tag{4.1}$$

is not proper as there is no clear meaning to the dots. This has been recognized by G. Peano who put forward the following axiomatic definition:

Definition 4.1 (Peano axioms) *A triplet* $(\mathbb{N}, 0, S)$ *is said to be a* system of naturals *if the following five axioms hold:*

(P1) $\mathbb{N}$ *is a set and* $0 \in \mathbb{N}$,
(P2) S *is a function* $S\colon \mathbb{N} \to \mathbb{N}$ *with* $\mathrm{Dom}(S) = \mathbb{N}$,
(P3) $\forall n \in \mathbb{N}\colon S(n) \neq 0$,
(P4) $\forall n, m \in \mathbb{N}\colon S(n) = S(m) \Rightarrow n = m$,
(P5) $\forall A \subseteq \mathbb{N}\colon 0 \in A \wedge S(A) \subseteq A \Rightarrow A = \mathbb{N}$.

The first two axioms basically identify what the objects in $(\mathbb{N}, 0, S)$ are so the real power rests with Axioms P3-P5. The element 0 is called the *zero element* while S is called the *successor function* and elements of its range are called *successors*. Axiom P3 tells us that 0 is not a successor while P4 imposes that the successor function is injective. These two properties ensure that $\mathbb{N}$ is not too small (and, in particular, that $\mathbb{N}$ is infinite) by ruling out, e.g., the set $\{0, 1, 2, \dots, 10\}$ with the successor function acting cyclically.

The most powerful axiom of all is P5, often referred to as the *Induction principle*, which ensures that $\mathbb{N}$ is not too large and guarantees many other useful facts. One of its elementary consequences is that 0 is the only element that is not a successor:

Lemma 4.2 *For any system of naturals* $(\mathbb{N}, 0, S)$*, we have* $S(\mathbb{N}) = \mathbb{N} \setminus \{0\}$.

Proof. Let $A := S(\mathbb{N}) \cup \{0\}$. Then $0 \in A$ and, using $A \subseteq \mathbb{N}$

$$S(A) \subseteq S(\mathbb{N}) \subseteq A \tag{4.2}$$

By P5, we have $A = \mathbb{N}$. As $0 \notin S(\mathbb{N})$ by P3, we have $S(\mathbb{N}) = \mathbb{N} \setminus \{0\}$. □

Another consequence is the ability to use proofs by induction when one verifies a statement depending on a natural first for zero and then proves the statement for n implies the statement for $n + 1$. That this is enough is the content of:

Lemma 4.3 (Proof by induction) *Let* $(\mathbb{N}, 0, S)$ *be a system of naturals and* $\{P_n\colon n \in \mathbb{N}\}$ *be (logical) propositions indexed thereby. Suppose that*

(1) (Induction basis) P_0 *is* TRUE*, and*
(2) (Induction step) $\forall n \in \mathbb{N}\colon P_n \Rightarrow P_{S(n)}$.

Then $\{n \in \mathbb{N}\colon P_n\} = \mathbb{N}$*, i.e.,* P_n *is* TRUE *for all* $n \in \mathbb{N}$.

Proof. Let $A := \{n \in \mathbb{N}\colon P_n\}$. By (1) we have $0 \in A$ and by (2) we have $\forall n \in \mathbb{N}\colon n \in A \Rightarrow S(n) \in A$, i.e., $S(A) \subseteq A$. By P5, $A = \mathbb{N}$ as claimed. □

The main task of this section is to show that the naturals exist. This comes in:

Theorem 4.4 (Existence of the naturals) *There is at least one system of naturals.*

Before we delve into the proof, let us make a historical note: In late 1800s a number of "proofs" of existence were put forward which all turned out to be flawed. One reason for this is that an axiomatic set theory can be cast *without* requiring existence of infinite sets (this is so called "Finite set theory"). Such a theory could not accommodate the naturals as these are necessarily infinite (in whatever meaning of this we take). We will thus have to use Axiom of infinity somewhere in the proof.

Proof of Theorem 4.4. Axiom of Infinity guarantees the existence of a set I such that

$$\varnothing \in I \ \wedge \ \forall X \in I \colon \{X\} \in I \tag{4.3}$$

With the choices $0 := \varnothing$ and $S(X) := \{X\}$ this set would satisfy P1-P4 of Peano axioms but it is too large to obey P5 as there are many subsets thereof (corresponding, in a related construction, to *limit ordinals*) that are closed under S. We will thus define the naturals as the smallest set that contains $\varnothing$ and is closed under S.

Consider a collection of all such sets

$$K := \Big\{ J \subseteq I \colon \varnothing \in J \wedge (\forall X \colon X \in J \Rightarrow \{X\} \in J) \Big\} \tag{4.4}$$

which exists thanks to Powerset and Separation Axioms. We then claim that

$$\mathbb{N} := \bigcap K \tag{4.5}$$

obeys $\mathbb{N} \in K$. For this note that $X \in \mathbb{N}$ is equivalent to $\forall J \in K \colon X \in J$ which then implies $\forall J \in K \colon \{X\} \in J$ and thus yields $\{X\} \in \mathbb{N}$. Similarly, $\varnothing \in \mathbb{N}$ because $\forall J \in K \colon \varnothing \in J$.

Next we define

$$0 := \varnothing \ \wedge \ \forall X \in \mathbb{N} \colon \ S(X) := \{X\} \tag{4.6}$$

and proceed to check that $(\mathbb{N}, 0, S)$ is a system of naturals. First we check Peano axioms P1-P4: From $\mathbb{N} \in K$ we have that $\mathbb{N}$ is a set with $\varnothing \in \mathbb{N}$, proving P1. For the same reason, S defined above is a function $\mathbb{N} \to \mathbb{N}$ with $\mathrm{Dom}(S) = \mathbb{N}$, proving P2. Axiom of Extensionality ensures that $\{X\} = \{Y\}$ implies $X = Y$ thus showing that S is injective, proving P4. The same axiom shows that $\varnothing$ is not a set in the range of S, proving P3.

It remains to prove the Induction Principle P5. For that let $A \subseteq \mathbb{N}$ be such that $\varnothing \in A$ and $S(A) \subseteq A$. This is readily checked to imply $A \in K$ and so $\mathbb{N} \subseteq A$ by (4.5). Lemma 2.3 now gives $A = \mathbb{N}$, proving P5. □

Our next task is to prove uniqueness of the naturals (up to natural *isomorphism*). This will hinge on a result that we will find useful later:

Theorem 4.5 (Recursion principle) *Given a system of naturals $(\mathbb{N}, 0, S)$, a set E and a function $\mathfrak{h} \colon E \to E$ with $\mathrm{Dom}(\mathfrak{h}) = E$ we have:*

$$\forall a \in E \ \exists \{X_n \colon n \in \mathbb{N}\} \subseteq E \colon \ X_0 = a \wedge \Big(\forall n \in \mathbb{N} \colon X_{S(n)} = \mathfrak{h}(X_n) \Big). \tag{4.7}$$

Moreover, the collection $\{X_n \colon n \in \mathbb{N}\}$ satisfying (4.7) is unique — meaning that if $\{X'_n \colon n \in \mathbb{N}\}$ obeys the same recursions, then $\forall n \in \mathbb{N} \colon X_n = X'_n$.

The purpose of the above is to give a rigorous meaning to the informal recursive definition whose first couple of steps are written as

$$\begin{aligned} X_0 &:= a \\ X_1 &:= \mathfrak{h}(a) && \text{where} \quad 1 := S(0) \\ X_2 &:= \mathfrak{h}(\mathfrak{h}((a)) && \text{where} \quad 2 := S(1) \\ X_3 &:= \mathfrak{h}(\mathfrak{h}(\mathfrak{h}((a))) && \text{where} \quad 3 := S(2) \\ &\vdots \quad \ddots \qquad \ddots && \qquad\qquad\qquad \ddots \end{aligned} \tag{4.8}$$

Although this sounds very plausible, the technical problem with this "construction" are the dots. Indeed, the procedure at best defines X_n "up to" any given natural n but defining that for all n simultaneously requires infinitely many iterations which cannot be formalized along the lines above.

The actual proof will avoid these ambiguities by careful use of set theory. The idea that we will consider all possible functions $f\colon \mathbb{N} \to E$ whose domain is a (finite or infinite) string of naturals such that $f(S(n)) = \mathfrak{h}(f(n))$ for all $n \in \mathrm{Dom}(f)$. Then we take the union of the graphs of these functions and prove that this is the graph of a function whose domain is all of $\mathbb{N}$.

While the idea is simple, its formal execusion is rather lengthy and may appear impenetrable for those new to the subject or untrained in logical reasoning. Readers that feel overwhelmed by what is to come may consider skipping to the statement of Theorem 4.7. That being said, all readers should understand the statement of Theorem 4.5 as it will be used repeatedly throughout the course.

Proof of Theorem 4.5. Recall that a function $f\colon \mathbb{N} \to E$ is technically a relation, and thus a subset of $\mathbb{N} \times E$. With this in mind we set

$$\mathcal{F} := \left\{ f \subseteq \mathbb{N} \times E \colon \begin{array}{l} f \text{ is a function } \wedge\ 0 \in \mathrm{Dom}(f) \wedge f(0) = a \\ \wedge \left(\begin{array}{l} \forall n \in \mathbb{N} \colon S(n) \in \mathrm{Dom}(f) \\ \qquad \Rightarrow \big(n \in \mathrm{Dom}(f) \wedge f(S(n)) = \mathfrak{h}(f(n)) \big) \end{array} \right) \end{array} \right\}. \tag{4.9}$$

In words, $\mathcal{F}$ is the set of relations that are functions from $\mathbb{N}$ to E whose domain contains 0, contains the predecessor of all non-zero elements in its domain, take value a at 0 and assign value $\mathfrak{h}(f(n))$ to the successor of n. We now note:

Step 1: $\{(0,a)\} \in \mathcal{F}$ and so $\mathcal{F} \neq \varnothing$

Proof. Let $f = \{(0,a)\}$. Then f is (the graph of) a function with

$$\mathrm{Dom}(f) := \{0\} \quad \text{and} \quad f(0) := a. \tag{4.10}$$

It follows that $f \in \mathcal{F}$ and so $\mathcal{F} \neq \varnothing$. □

Step 2: $\forall f, g \in \mathcal{F}\ \forall n \in \mathbb{N} \colon n \in \mathrm{Dom}(f) \cap \mathrm{Dom}(g) \Rightarrow f(n) = g(n)$

Proof. Pick $f, g \in \mathcal{F}$ and let

$$A := \Big\{ n \in \mathbb{N} \colon n \in \mathrm{Dom}(f) \cap \mathrm{Dom}(g) \Rightarrow f(n) = g(n) \Big\} \tag{4.11}$$

Every function in $\mathcal{F}$ is defined at 0 and takes value a there so $0 \in A$. Now let $n \in A$ and consider $S(n)$. Using that an implication is vacuously TRUE if its premise is FALSE,

$$S(n) \notin \mathrm{Dom}(f) \cap \mathrm{Dom}(g) \Rightarrow S(n) \in A. \tag{4.12}$$

is TRUE trivially. On the other hand, by the second line in the definition of $\mathcal{F}$, the assumption that $S(n) \in \mathrm{Dom}(f) \cap \mathrm{Dom}(g)$ implies $n \in \mathrm{Dom}(f) \cap \mathrm{Dom}(g)$ and

$$f\big(S(n)\big) = \mathfrak{h}\big(f(n)\big) \ \wedge g\big(S(n)\big) = \mathfrak{h}\big(g(n)\big) \tag{4.13}$$

But from $n \in A$ we know that $f(n) = g(n)$ and so

$$S(n) \in \mathrm{Dom}(f) \cap \mathrm{Dom}(g) \ \Rightarrow \ f\big(S(n)\big) = g\big(S(n)\big). \tag{4.14}$$

It follows that $n \in A$ implies $S(n) \in A$ meaning that $S(A) \subseteq A$. By P5, we get $A = \mathbb{N}$ which is equivalent to the claim. □

Step 3: Define $\widehat{f} := \bigcup \mathcal{F}$. Then $\widehat{f} \in \mathcal{F}$.

Proof. The definition gives $\widehat{f} \subseteq \mathbb{N} \times E$. We first show that $\widehat{f}$ is (the graph of) a function. For that let $n \in \mathbb{N}$ and assume that, for some $x, y \in E$ we have $(n, x) \in \widehat{f}$ and $(n, y) \in \widehat{f}$. Then there exist $f, g \in \mathcal{F}$ such that $n \in \mathrm{Dom}(f) \cap \mathrm{Dom}(g)$ and $x = f(n)$ and $y = g(n)$. But step 2 then gives $f(n) = g(n)$ and so $x = y$. So $\widehat{f}$ is a function and, moreover,

$$\forall n \in \mathrm{Dom}(\widehat{f})\, \exists f \in \mathcal{F} \colon \ n \in \mathrm{Dom}(f) \ \wedge \ \widehat{f}(n) = f(n). \tag{4.15}$$

which will come handy in what follows.

We now have to check that $\widehat{f}$ lies in $\mathcal{F}$. Step 1 gives $0 \in \mathrm{Dom}(\widehat{f})$ and $\widehat{f}(0) = a$. Suppose now $n \in \mathbb{N}$ is such that $S(n) \in \mathrm{Dom}(\widehat{f})$. By (4.15) there is $f \in \mathcal{F}$ such that $S(n) \in \mathrm{Dom}(f)$ and $\widehat{f}(S(n))) = f(S(n))$. But $f \in \mathcal{F}$ and $S(n) \in \mathrm{Dom}(f)$ implies

$$n \in \mathrm{Dom}(f) \ \wedge \ f\big(S(n)\big) = \mathfrak{h}\big(f(n)\big) \tag{4.16}$$

and, since $\widehat{f}$ is the graph of a function that contains the graph of f, also

$$n \in \mathrm{Dom}(\widehat{f}) \ \wedge \ \widehat{f}(n) = f(n) \ \wedge \ \widehat{f}\big(S(n)\big) = \mathfrak{h}\big(\widehat{f}(n)\big) \tag{4.17}$$

This proves that $\widehat{f} \in \mathcal{F}$. □

Step 4: $\mathrm{Dom}(\widehat{f}) = \mathbb{N}$

Proof. Denote $A := \mathrm{Dom}(\widehat{f})$. Then $0 \in A$ by $\widehat{f} \in \mathcal{F}$. Next let $n \in A$ and assume, for the sake of contradiction, $S(n) \notin A$. By (4.15), there is $f \in \mathcal{F}$ with $n \in \mathrm{Dom}(f)$ and $S(n) \notin \mathrm{Dom}(f)$. Now consider the function $g \colon \mathbb{N} \to E$ with domain $\mathrm{Dom}(g) := \mathrm{Dom}(f) \cup \{S(n)\}$ and values given by

$$g(m) := \begin{cases} f(m), & \text{if } m \in \mathrm{Dom}(f), \\ \mathfrak{h}\big(f(n)\big), & \text{if } m = S(n). \end{cases} \tag{4.18}$$

We claim $g \in \mathcal{F}$. Clearly, g is a function with $0 \in \mathrm{Dom}(g)$ and $g(0) = a$. Next let $m \in \mathbb{N}$ be such $S(m) \in \mathrm{Dom}(g)$. Two alternatives are then possible. First, we may have $S(m) \in \mathrm{Dom}(f)$, which by $f \in \mathcal{F}$ forces $m \in \mathrm{Dom}(f)$ and

$$g\big(S(m)\big) = f\big(S(m)\big) = \mathfrak{h}\big(f(m)\big) = \mathfrak{h}\big(g(m)\big) \tag{4.19}$$

where the first equality is by $S(m) \in \mathrm{Dom}(f)$, the second by $f \in \mathcal{F}$ and the third by $m \in \mathrm{Dom}(f)$. Second, we may have $S(m) = S(n)$ which by the injectivity of S forces $m = n$ and so we get

$$g\big(S(m)\big) = g\big(S(n)\big) = \mathfrak{h}\big(f(n)\big) = \mathfrak{h}\big(g(n)\big) \tag{4.20}$$

where the first equality is by $S(m) = S(n)$, the second by definition of $g(S(n))$ and the third by $n \in \mathrm{Dom}(f)$ and the fact that $g(n) = f(n)$. But $g \in \mathcal{F}$ implies $S(n) \in \mathrm{Dom}(\widehat{f}) = A$, a contradiction. It follows that $S(A) \subseteq A$ and, by P5, $A = \mathbb{N}$. □

With the above in hand we are ready to complete the proof: The function $\widehat{f}\colon \mathbb{N} \to E$ with $\mathrm{Dom}(\widehat{f}) = \mathbb{N}$ constructed above obeys $\widehat{f}(0) = a$ and $\widehat{f}(S(n)) = \mathfrak{h}(\widehat{f}(n))$. Setting $X_n := \widehat{f}(n)$ for $n \in \mathbb{N}$ thus proves (4.7). To show uniqueness, let $\{X'_n\colon n \in \mathbb{N}\}$ be another such a family. Set $A := \{n \in \mathbb{N}\colon X_n = X'_n\}$. Then $0 \in A$ because $X_0 = a = X'_0$ and if $n \in A$, then $X_n = X'_n$ implies $X_{S(n)} = \mathfrak{h}(X_n) = \mathfrak{h}(X'_n) = X'_{S(n)}$ and so $S(n) \in A$, thus showing $S(A) \subseteq A$. Hence $A = \mathbb{N}$ by P5. □

Remark 4.6 We note that, writing E as $\mathbb{N} \times E'$ for a set E' and letting $h\colon \mathbb{N} \times E' \to \mathbb{N} \times E'$ be the function $(n, x) \mapsto (S(n), h_n(x))$ for a given collection $\{\mathfrak{h}_n\colon n \in \mathbb{N}\}$ of functions $\mathfrak{h}_n\colon E' \to E'$, Theorem 4.5 accommodates for the situation that $\{X_n\colon n \in \mathbb{N}\}$ obeys

$$X_0 = a \ \wedge\ \big(\forall n \in \mathbb{N}\colon X_{S(n)} = \mathfrak{h}_n(X_n)\big) \tag{4.21}$$

This allows for the "recursive rule" to depend explicitly on the order of iteration.

We are now in a position to state and prove uniqueness of the naturals:

Theorem 4.7 (Uniqueness of the naturals) *Let $(\mathbb{N}, 0, S)$ and $(\mathbb{N}', 0', S')$ be two systems of naturals. Then there is a bijection $\phi\colon \mathbb{N} \to \mathbb{N}'$ such that*

$$\phi(0) = 0' \ \wedge\ \phi \circ S = S' \circ \phi. \tag{4.22}$$

Proof. Using Theorem 4.5 with the choices $E := \mathbb{N}'$, $a := 0'$ and $h := S'$ produces a collection $\{X_n\colon n \in \mathbb{N}\}$ with $X_0 = 0'$ and $\forall n \in \mathbb{N}\colon X_{S(n)} = S'(X_n)$. Setting $\phi(n) := X_n$ for each $n \in \mathbb{N}$ defines a function ϕ with domain $\mathrm{Dom}(\phi) = \mathbb{N}$ and properties (4.22). It remains to show that this (and, in fact, any such) function is a bijection.

We start by proving that ϕ is surjective. Let $A := \mathrm{Ran}(\phi)$. Then $0' \in A$ by the first part of (4.22) while the second part thereof implies

$$S'(A) = S' \circ \phi(\mathbb{N}) = \phi \circ S(\mathbb{N}) \subseteq \phi(\mathbb{N}) = A. \tag{4.23}$$

By P5 for the system $(\mathbb{N}', 0', S')$ we have $A = \mathbb{N}'$ thus showing that ϕ is onto.

Next we show that ϕ is injective. Consider the set

$$A := \Big\{n \in \mathbb{N}\colon \big(\forall m \in \mathbb{N}\colon \phi(m) = \phi(n) \Rightarrow m = n\big)\Big\} \tag{4.24}$$

The aim is to prove that $A = \mathbb{N}$. First note that if $\phi(m) = 0'$, then $m = 0$ for otherwise Lemma 4.2 gives $m = S(k)$ for some $k \in \mathbb{N}$ and

$$0' = \phi(m) = \phi \circ S(k) = S' \circ \phi(k) \in \mathrm{Ran}(S') \tag{4.25}$$

in contradiction with P3 for the system $(\mathbb{N}', 0', S')$. Since $0' = \phi(0)$ and since the above holds for all $m \in \mathbb{N}$, it follows that $0 \in A$.

Next assume that $n \in A$ and let $m \in \mathbb{N}$ be such that $\phi(S(n))) = \phi(m)$. Then the previous argument shows $m \neq 0$ and so $m \in \mathrm{Ran}(S)$, by Lemma 4.2. This means that $m = S(k)$ for some $k \in \mathbb{N}$ and $\phi(S(n)) = \phi(m)$ then rewrites into

$$S' \circ \phi(n) = S' \circ \phi(k) \tag{4.26}$$

The injectivity of S' forced by P4 for the system $(\mathbb{N}', 0', S')$ then gives $\phi(n) = \phi(k)$ which by $n \in A$ forces $n = k$ and so $S(n) = S(k) = m$. As m was arbitrary, we conclude that $S(n) \in A$ thus showing $S(A) \subseteq A$. By P5 for system $(\mathbb{N}, 0, S)$ we get $A = \mathbb{N}$ and so ϕ is indeed injective as claimed. □

Note that the above theory treats natural numbers in the abstract sense and, in particular, without reference to a specific "number system" or labeling convention. In light of our prior hands-on experience with the naturals, this may seem clumsy at first but is indispensable if we want to prove all familiar properties of standard number systems from axioms of set theory (rather than postulating them as axioms instead, as is done in many real-analysis textbooks).

5. Arithmetic of the naturals

In order to bring the abstract treatment of the naturals closer to our intuition, we will now define the basic operations of *addition*, *multiplication*, *powers* etc on the naturals and prove the standard relations between them.

5.1 Addition.

We will spend most of the time on addition as other operations are handled analogously. Pick $m \in \mathbb{N}$ and invoke the recursion principle in Theorem 4.5 for the choice $E := \mathbb{N}$, $a := m$ and $\mathfrak{h} := S$ to define $\{X_n : n \in \mathbb{N}\}$ such that

$$X_0 = m \quad \text{and} \quad \forall n \in \mathbb{N} : X_{S(n)} = S(X_n). \tag{5.1}$$

Then we denote

$$m + n := X_n. \tag{5.2}$$

As consequence of the construction (5.1) we get a symbol $m + n$ satisfying

(1) $\forall m \in \mathbb{N} : m + 0 = m$, and
(2) $\forall m, n \in \mathbb{N} : m + S(n) = S(m + n)$.

From these observation we now derive further facts about addition relying, predominantly, on the Induction Principle.

We will now prove that the operation $m, n \mapsto m + n$ is commutative. We begin by:

Lemma 5.1 $\forall m \in \mathbb{N} : 0 + m = m$

Proof. Let P_m denote the logical proposition $0 + m = m$. Then P_0 is TRUE because (1) above implies $0 + 0 = 0$. Next assume P_m holds for some $m \in \mathbb{N}$. Then

$$0 + S(m) \overset{(2)}{=} S(0 + m) \overset{P_m}{=} S(m). \tag{5.3}$$

It follows that $P_m \Rightarrow P_{S(m)}$. By the Induction Lemma, $\{m \in \mathbb{N} : 0 + m = m\} = \mathbb{N}$. □

Next we need:

Lemma 5.2 $\forall m, n \in \mathbb{N} : m + S(n) = S(m) + n$

Proof. Fix $m \in \mathbb{N}$ and let P_n be the statement $m + S(n) = S(m) + n$. Since

$$m + S(0) \overset{(2)}{=} S(m + 0) \overset{(1)}{=} S(m) \overset{(1)}{=} S(n) + 0 \tag{5.4}$$

we get that P_0 is TRUE. Next assume that P_n is TRUE for some $n \in \mathbb{N}$. Then

$$m + S(S(n)) \overset{(2)}{=} S(m + S(n)) \overset{P_n}{=} S(S(m) + n) \overset{(2)}{=} S(m) + S(n) \tag{5.5}$$

implying $P_{S(n)}$. Hence, $\forall n \in \mathbb{N} : P_n \Rightarrow P_{S(n)}$ and, by the Induction lemma, $\{n \in \mathbb{N} : m + S(n) = S(m) + n\} = \mathbb{N}$. As this holds for all $m \in \mathbb{N}$, we are done. □

Hence we finally conclude:

Proposition 5.3 (Commutativity of addition)

$$\forall m, n \in \mathbb{N} : \ m + n = n + m \tag{5.6}$$

Proof. Let Q_m be the statement $\forall n \in \mathbb{N}\colon m + n = n + m$. Then Q_0 is TRUE by (1) and Lemma 5.1. Assume now that Q_m is TRUE. Then for any $n \in \mathbb{N}$,

$$S(m) + n \overset{\text{Lemma 5.2}}{=} m + S(n) \overset{(2)}{=} S(m+n) \overset{Q_m}{=} S(n+m) \overset{(2)}{=} n + S(m). \tag{5.7}$$

It follows that $Q_m \Rightarrow Q_{S(m)}$. By induction, $\{m \in \mathbb{N}\colon Q_m\} = \mathbb{N}$. □

Similarly we also prove that the operation $m, n \mapsto m + n$ is associative:

Proposition 5.4 (Associativity of addition)

$$\forall m, n, k \in \mathbb{N}\colon \ m + (n + k) = (m + n) + k \tag{5.8}$$

Proof. Left as a homework exercise. Commutativity should not be needed. □

5.2 Ordering of the naturals.

A useful property of addition is that it acts injectively:

Lemma 5.5 $\forall m, n, k \in \mathbb{N}\colon m + n = m + k \Rightarrow n = k$

Proof. Let P_m be the statement $\forall n, k \in \mathbb{N}\colon m + n = m + k \Rightarrow n = k$. By (1) and Lemma 5.1, P_0 is TRUE. Now assume P_m to be TRUE for some $m \in \mathbb{N}$ and let $n, k \in \mathbb{N}$ be such that

$$S(m) + n = S(m) + k \tag{5.9}$$

then Lemma 5.2 implies $S(m) + n = m + S(n)$ and $S(m) + k = k + S(m)$ and so

$$m + S(n) \overset{\text{Lemma 5.2}}{=} S(m) + n = S(m) + k \overset{\text{Lemma 5.2}}{=} m + S(k) \tag{5.10}$$

thus implying $S(n) = S(k)$ via P_m. But S is injective by P4 and so we get $n = k$. Hence $P_m \Rightarrow P_{S(m)}$ and, by induction, P_m is TRUE for all $m \in \mathbb{N}$. □

This property implies that, given $m \in \mathbb{N}$, for each $n \in \mathbb{N}$ the equation $n = m + s$ has *at most one* solution for s in $\mathbb{N}$. We can formally describe the pairs $(m, n) \in \mathbb{N} \times \mathbb{N}$ for which the solution exists by way of the *less than or equal* relation $\leqslant$ defined by

$$m \leqslant n \quad \Leftrightarrow \quad \exists s \in \mathbb{N}\colon n = m + s. \tag{5.11}$$

Here are some properties of this relation:

Lemma 5.6 *The relation $\leqslant$ is reflexive, antisymmetric and transitive.*

Proof. Reflexivity is immediate from (1) and transitivity follows from the associativity of addition. So the main point to check is antisymmetry. For that assume $m, n \in \mathbb{N}$ are such that $m \leqslant n \wedge n \leqslant m$. Then there are $r, s \in \mathbb{N}$ such that $m = n + s \wedge n = m + r$. Putting these together and invoking the associativity of addition, we get $n = n + (s + r)$. Lemma 5.5 and (1) then force $s + r = 0$. By P3 and (2) above, r cannot be a successor and so $r = 0$. Then also $s = 0$ whereby we conclude

$$\forall m, n \in \mathbb{N}\colon \ m \leqslant n \wedge n \leqslant m \Rightarrow m = n \tag{5.12}$$

meaning that $\leqslant$ is antisymmetric. □

It easy to check that the following properties of $\leqslant$ are true:

Lemma 5.7 *We have*

$$\forall n \in \mathbb{N}\colon\ 0 \leqslant n \tag{5.13}$$

$$\forall n \in \mathbb{N}\colon\ n \leqslant S(n) \tag{5.14}$$

and

$$\forall m, n \in \mathbb{N}\colon m \leqslant n \Rightarrow S(m) \leqslant S(n) \tag{5.15}$$

Proof. Left to a homework exercise. □

An important point of the relation $\leqslant$ is that it is *connex*, meaning that every pair of naturals are ordered one or the other way. This is usually phrased by saying that $\leqslant$ is a *total ordering* in the following sense:

Lemma 5.8 (Total-ordering of $\mathbb{N}$)

$$\forall m, n \in \mathbb{N}\colon \quad m \leqslant n \vee n \leqslant m \tag{5.16}$$

Proof. Let P_m be the statement $\forall n \in \mathbb{N}\colon m \leqslant n \vee n \leqslant m$. Then P_0 is TRUE by (5.13) so assume that P_m holds for some $m \in \mathbb{N}$ and let $n \in \mathbb{N}$. If $n \leqslant m$ or $n = m$ then (5.14) and transitivity imply $n \leqslant S(m)$. In the opposite case we must have $m \leqslant n$ (as P_m was assumed to hold) and $m \neq n$. The definition (5.11) and Lemma 4.2 then show existence of an $r \in \mathbb{N}$ such that

$$n = m + S(r) \overset{\text{Lemma 5.2}}{=} S(m) + r \tag{5.17}$$

proving $S(m) \leqslant n$ and thus also $P_m \Rightarrow P_{S(m)}$. Hence, P_m is TRUE for all $m \in \mathbb{N}$. □

Note that we can re-state Lemma 5.8 as saying that at least one of $n = m + r$ or $m = n + r$ has a solution for r in the naturals.

5.3 Multiplication, powers, factorial.

Moving to a definition of multiplication, pick $m \in \mathbb{N}$ and use Theorem 4.5 with the choices $E := \mathbb{N}$, $\mathfrak{h}(r) := r + m$ and $a := 0$ to construct $\{X_n\colon n \in \mathbb{N}\}$ such that

$$X_0 = 0 \quad \wedge \quad \forall n \in \mathbb{N}\colon X_{S(n)} = X_n + m. \tag{5.18}$$

We will write $n \cdot m$ for X_n and thus get

$$0 \cdot m = 0 \quad \wedge \quad \forall n \in \mathbb{N}\colon S(n) \cdot m = n \cdot m + m \tag{5.19}$$

We also define the *unity* in $\mathbb{N}$ by

$$1 := S(0) \tag{5.20}$$

and observe that

$$S(n) = S(n + 0) = n + S(0) = n + 1 \tag{5.21}$$

which will eventually allow us to drop the notation using the successor function and write it as "plus one" instead. The following properties are then checked:

Proposition 5.9 (Properties of multiplication on $\mathbb{N}$) *We have:*

(1) (Commutative law) $\forall m, n \in \mathbb{N}\colon m \cdot n = n \cdot m$,
(2) (Associative law) $\forall m, n, k \in \mathbb{N}\colon (m \cdot n) \cdot k = m \cdot (n \cdot k)$,
(3) (Distributive law) $\forall m, n, k \in \mathbb{N}\colon (n + k) \cdot m = (n \cdot m) + (k \cdot m)$
(4) (Zero and unity) $\forall m \in \mathbb{N}\colon 0 \cdot m = 0 \wedge 1 \cdot m = m$
(5) (Injectivity) $\forall m, n, k \in \mathbb{N}\colon k \neq 0 \wedge k \cdot m = k \cdot n \Rightarrow m = n$

Proof. A somewhat tedious but doable exercise that we leave to the reader. □

Multiplication also behaves nicely around the total ordering relation:

Lemma 5.10 *We have*

$$\forall m, n, r \in \mathbb{N}\colon\ m \leqslant n \ \Rightarrow\ r \cdot m \leqslant r \cdot n \tag{5.22}$$

Proof. Left to homework exercise. □

With multiplication in place, we can now define natural *powers*. Here we pick $m \in \mathbb{N}$ and use Theorem 4.5 to construct $\{m^n : n \in \mathbb{N}\}$ satisfying

$$m^0 = 1 \quad \wedge \quad \forall n \in \mathbb{N}\colon m^{S(n)} = m \cdot m^n. \tag{5.23}$$

Note that this entails $m^0 = 1$ (even for $m = 0$) while $0^n = 0$ for $n \neq 0$. Similarly, $1^n = 1$ for all $n \in \mathbb{N}$. The following properties will again be of relevance:

Lemma 5.11 (Powers) *Let $m \in \mathbb{N} \setminus \{0\}$. Then*

(1) $\forall r, s \in \mathbb{N}\colon m^{r+s} = m^r \cdot m^s$,
(2) $\forall r, s \in \mathbb{N}\colon m^{r \cdot s} = (m^r)^s$.

Proof. Proved readily by induction. □

A related construction permits us to construct the *factorial* of n, with notation $n!$, by imposing

$$0! = 1 \quad \text{and} \quad \forall n \in \mathbb{N}\colon\ S(n)! = S(n) \cdot n! \tag{5.24}$$

By (5.21), the statement in the second part can be written as $(n+1)! = (n+1) \cdot n!$, which is the recursive form of the informal expression $n! = n \cdot (n-1) \cdot \cdots \cdot 1$.

Factorials appear frequently in combinatorial arguments (indeed, $n!$ is the number of permutations of n elements) but also appears in analytic expressions (thanks to, for instance, Taylor's theorem).

6. INTEGERS AND RATIONALS

Having discussed a full set-theoretical construction of the naturals, we will now take a somewhat lighter approach to the construction of the integers and rationals. For more abstract approaches we refer the reader to standard textbooks in set theory.

6.1 The integers.

Recall that, for $m, n \in \mathbb{N}$ the relation $m \leqslant n$ is equivalent to the existence of a natural $r \in \mathbb{N}$ such that $n = m + r$. We will call that r the *difference* between m and n with the notation $r = n - m$. Unfortunately, the existence of the symbol $n - m$ is restricted to the pairs $(m, n) \in \mathbb{N}$ with $m \leqslant n$. To eliminate this restriction, we enlarge the naturals (4.1) into the set of the *integers* that, informally, takes the form

$$\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}. \tag{6.1}$$

This is again too vague due to the usage of the dots and so we have to proceed more formally. First introduce a relation $\overset{+}{\sim}$ on $\mathbb{N} \times \mathbb{N}$ by setting

$$(m, n) \overset{+}{\sim} (m', n') := \ m + n' = n + m'. \tag{6.2}$$

for any $m, n, m', n' \in \mathbb{N}$. We then observe:

Lemma 6.1 *$\overset{+}{\sim}$ is an equivalence relation on $\mathbb{N} \times \mathbb{N}$.*

Proof. Reflexivity and symmetry are immediate from the definition so we just have to prove transitivity. Suppose $m, n, m', n' \in \mathbb{N}$ are such that

$$(m, n) \overset{+}{\sim} (m', n') \ \wedge \ (m', n') \overset{+}{\sim} (m'', n''). \tag{6.3}$$

Then we have

(1) $m + n' = m' + n$, and
(2) $m' + n'' = m'' + n'$.

Using the associative and commutative law of the addition, from here we get

$$m + n'' + n' \overset{(1)}{=} m' + n + n'' \overset{(2)}{=} m'' + n + n'. \tag{6.4}$$

Lemma 5.5 then gives $m + n'' = m'' + n$ and so $(m, n) \overset{+}{\sim} (m'', n'')$. □

Roughly speaking, the equivalence $(m, n) \overset{+}{\sim} (m', n')$ means that $m - n$ represents the same integer as $m' - n'$. To factor out all possible representations of one integer as the difference of two naturals, we recall the concept of a equivalence class $[x]$ represented by element x defined in (3.9) and set

$$\mathbb{Z} := \big\{[(m, n)] \colon m, n \in \mathbb{N}\big\}. \tag{6.5}$$

We will call the elements of $\mathbb{Z}$ *integers*. Informally, the equivalence class $[(m, n)]$ represents an integer "$m - n$" with "positive part" m and "negative part" n; the equivalence ensures that adding any natural to both parts simultaneously does not change the result. This in fact characterizes equivalent pairs:

Lemma 6.2 *For each $m, n, m', n' \in \mathbb{N}$:*

$$\big((m, n) \overset{+}{\sim} (m', n') \ \wedge \ m \leqslant m'\big) \ \Leftrightarrow \ \big(\exists k \in \mathbb{N} \colon m' = m + k \ \wedge \ n' = n + k\big) \tag{6.6}$$

Proof. The direction $\Leftarrow$ was noted before the statement. For $\Rightarrow$ we use that $m' \leqslant m$ implies $m = m' + k$ for some $k \in \mathbb{N}$. The equivalence $(m, n) \overset{+}{\sim} (m', n')$ then implies

$$m + n' = m' + n = m + k + n \tag{6.7}$$

The injectivity of addition (Lemma 5.5) allows us to "cancel" m on both sides, thus showing $n' = n + k$ as desired. □

With the integers defined, we define the unary operation of (taking a) negative as well as the binary operations of addition, subtraction and multiplication by

$$\begin{aligned} -[(m, n)] &:= [(n, m)] \\ [(m, n)] + [(m', n')] &:= [(m + m', n + n')], \\ [(m, n)] - [(m', n')] &:= [(m + n', n + m')], \\ [(m, n)] \cdot [(m', n')] &:= [(m \cdot m' + n \cdot n', m \cdot n' + m' \cdot n)]. \end{aligned} \tag{6.8}$$

These will not be meaningful (as operations on equivalence classes) until we check:

Lemma 6.3 *The integers on the right-hand side of* (6.8) *are the same for any choice of the representatives of* $[(m, n)]$ *and* $[(m', n')]$.

Proof. We will only deal with multiplication as that is the hardest case of all. Suppose $(\widetilde{m}, \widetilde{n}) \in [(m, n)]$. Then also $(m, n) \in [(\widetilde{m}, \widetilde{n})]$. The fact that $\leqslant$ is a total ordering implies that either $m \leqslant \widetilde{m}$ or $\widetilde{m} \leqslant m$. By symmetry, we may thus assume that $m \leqslant \widetilde{m}$. Lemma 6.2 then gives $k \in \mathbb{N}$ such that $\widetilde{m} = m + k$ and $\widetilde{n} = n + k$. The laws of addition and multiplication on $\mathbb{N}$ then give

$$\begin{aligned} &\Big(\widetilde{m} \cdot m' + \widetilde{n} \cdot n', \widetilde{m} \cdot n' + m \cdot \widetilde{n}\Big) \\ &\qquad = \Big(m \cdot m' + n \cdot n' + k \cdot (m' + n'),\ m \cdot n' + m' \cdot n + k \cdot (m' + n')\Big) \\ &\qquad\qquad \overset{+}{\sim} \big(m \cdot m' + n \cdot n', m \cdot n' + m' \cdot n\big) \end{aligned} \tag{6.9}$$

Hence, the equivalence class on the right of the third line of (6.8) is independent of the choice of the representative of $[(m, n)]$. The other cases are handled analogously. □

It it interesting to note that that the set

$$\big\{[(m, 0)] \colon m \in \mathbb{N}\big\} \tag{6.10}$$

is in a bijective correspondence with $\mathbb{N}$ and the addition and multiplication defined above then matches the two operations on $\mathbb{N}$. The naturals $\mathbb{N}$ are thus *naturally embedded* into (our model of) $\mathbb{Z}$. A key novelty of the integers is that subtraction acts as the inverse to addition. This is seen from

$$\begin{aligned} [(m', n')] + \big([(m, n)] - [(m', n')]\big) &= [(m', n')] + [(m + n', n + m')] \\ &= [(m + m' + n', n + m' + n')] = [(m, n)]. \end{aligned} \tag{6.11}$$

The element $[(0, 0)]$ is a zero element under addition and $[(1, 0)]$ is the unit element under multiplication. Abandoning the cumbersome notation of equivalence classes, similarly we also verify all other properties listed in:

Lemma 6.4 *The commutative, associative and distributive laws hold for addition and multiplication on $\mathbb{Z}$. Moreover, we have:*

(1) *(Zero element)* $\exists 0 \in \mathbb{Z}\, \forall a \in \mathbb{Z}\colon a + 0 = a$
(2) *(Additive inverse)* $\forall a \in \mathbb{Z}\, \exists(-a) \in \mathbb{Z}\colon a + (-a) = 0$
(3) *(Unit element)* $\exists 1 \in \mathbb{Z}\, \forall a \in \mathbb{Z}\colon 1 \cdot a = a$
(4) *(Injectivity of multiplication)* $\forall a, b, c \in \mathbb{Z}\colon a \cdot b = a \cdot c \ \wedge\ a \neq 0 \ \Rightarrow\ b = c$

We leave the proof of this lemma to the reader. Note that both 0 and the additive inverse are necessarily unique. Indeed, if 0 and $0'$ are two zero elements, then (1) and the commutative law gives $0 = 0 + 0' = 0' + 0 = 0'$, and if x and y are two versions of $-a$, then (2) and the commutative/associative laws yield $x = (a + y) + x = (a + x) + y = y$. For similar reasons, also the unit element under multiplication is unique.

The stated properties show that $\mathbb{Z}$ has the structure of a commutative ring. Note however, that the properties do not necessarily characterize $\mathbb{Z}$ — indeed, the rationals will satisfy these as well. We will eventually articulate conditions under which the rationals are unique but we will not attempt to do this for the integers.

The integers also admit a natural extension of the ordering relation $\leqslant$ via

$$[(m, n)] \leqslant [(m', n')] \quad := \quad m + n' \leqslant m' + n \tag{6.12}$$

Here the independence of the choice of a representative is quite apparent as well as the fact that the restriction of this relation to $\{[(m, 0)]\colon m \in \mathbb{N}\}$ reproduces $\leqslant$ on $\mathbb{N}$. The total ordering of $\mathbb{N}$ by $\leqslant$ implies the total ordering of $\mathbb{Z}$ by $\leqslant$ as well.

6.2 The rationals.

As a consequence of expanding $\mathbb{N}$ to $\mathbb{Z}$, the equation $n = m + r$ now can be solved for $r \in \mathbb{Z}$ for all $m, n \in \mathbb{Z}$. This is, however, not true for the equation $n = m \cdot r$. This motivates us to further expand $\mathbb{Z}$. We will proceed just as before. Indeed, we start by introducing a relation on $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$ by

$$(p, q) \mathrel{\dot\sim} (p', q') \ := \ p \cdot q' = p' \cdot q \tag{6.13}$$

for all $(p, q), (p', q') \in \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$. We had to exclude zero from the allowed values in the second component in order to ensure, via part (4) of Lemma 6.4, that

$$(p, q) \mathrel{\dot\sim} (p', q) \ \Leftrightarrow\ p = p'. \tag{6.14}$$

Lemma 3.8 then shows that $\dot\sim$ is an equivalence relation on $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$. This permits us to set

$$\mathbb{Q} := \Big\{[(p, q)]\colon (p, q) \in \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})\Big\} \tag{6.15}$$

Informally, the pair (p, q) stands for the fraction $\frac{p}{q}$; the equivalence relation then ensures that for any $m \in \mathbb{Z} \setminus \{0\}$, the rational $\frac{pm}{qm}$ coincides with $\frac{p}{q}$.

As before, we introduce the operations of addition, subtraction, multiplication on $\mathbb{Q}$ via the formulas

$$\begin{aligned} [(p, q)] + [(p', q')] &:= \big[(p \cdot q' + q \cdot p', q \cdot q')\big] \\ [(p, q)] - [(p', q')] &:= \big[(p \cdot q' - q \cdot p', q \cdot q')\big] \\ [(p, q)] \cdot [(p', q')] &:= \big[(p \cdot p', q \cdot q')\big] \end{aligned} \tag{6.16}$$

and, assuming $[(p',q')] \neq 0$ which entails $p' \neq 0$, also the operation of *division* via

$$[(p,q)] \div [(p',q')] := \big[(p \cdot q', q \cdot p')\big]. \tag{6.17}$$

In all of these we need to check:

Lemma 6.5 *For all $(p,q),(\tilde{p},\tilde{q}) \in \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$,*

$$(p,q) \stackrel{\cdot}{\sim} (\tilde{p},\tilde{q}) \quad \Leftrightarrow \quad \exists k,\ell \in \mathbb{Z} \setminus \{0\}\colon\ (p \cdot \ell, q \cdot \ell) = (\tilde{p} \cdot k, \tilde{q} \cdot k) \tag{6.18}$$

Consequently, the rationals on the right-hand side of (6.16–6.17) are the same for any choice of the representatives of $[(p,q)]$ and $[(p',q')]$.

The operation of division is relative to multiplication as subtraction is to addition: Indeed, assuming $[(p',q')] \neq 0$, we have

$$[(p',q')] \cdot \Big([(p,q)] \div [(p',q')]\Big) = [(p,q)] \tag{6.19}$$

Abandoning the equivalence class notation, we then check:

Theorem 6.6 *The operations of addition and multiplication defined via (6.16–6.17) satisfy the commutative, associative and distributive laws. In addition, we have:*

(1) *(Zero element)* $\forall a \in \mathbb{Q}\colon\ 0 + a = a \wedge 0 \cdot a = 0$,
(2) *(Unit element)* $\forall a \in \mathbb{Q}\colon\ 1 \cdot a = a$,
(3) *(Additive inverse)* $\forall a \in \mathbb{Q}\ \exists(-a) \in \mathbb{Q}\colon\ a + (-a) = 0$,
(4) *(Multiplicative inverse)* $\forall a \in \mathbb{Q} \setminus \{0\}\ \exists a^{-1} \in \mathbb{Q}\colon\ a \cdot a^{-1} = 1$.

As it turns out, the above is what gives $\mathbb{Q}$ an algebraic structure called a *field*. (We will give a definition of what it means for a set to be a field in the next lecture.) In our construction, the inverse elements in (3) and (4) are supplied by

$$-[(p,q)] = [(-p,q)] \quad \text{and} \quad [(p,q)]^{-1} = [(q,p)] \tag{6.20}$$

while, as noted above, $0 := [(0,1)]$ and $1 := [(1,1)]$. Note also that the integers (and thus naturals) are represented inside $\mathbb{Q}$ by

$$\big\{[(p,1)]\colon\ p \in \mathbb{Z}\big\} \tag{6.21}$$

The rationals admit also an extension of the relation $\leqslant$ by setting:

$$[(p,q)] \leqslant [(p',q')] \quad := \quad \begin{cases} p \cdot q' \leqslant p' \cdot q, & \text{if } q \cdot q' \geqslant 0, \\ p' \cdot q \leqslant p \cdot q', & \text{else,} \end{cases} \tag{6.22}$$

We again readily check that the relation does not depend on the representatives of $[(p,q)]$ and $[(p',q')]$ and that it faithfully reproduces the corresponding relation on $\mathbb{Z}$. With $\leqslant$ in place, the rationals have the structure of an *ordered field* (again, to be defined next).

An inquisitive reader may wonder whether other constructions of the integers and rationals exist that would give us intrinsically different objects from those constructed above. The answer to this is negative: the set of the rationals, viewed as an ordered field, is determined uniquely up to an isomorphism — i.e., a bijection reproducing all stated structures. We will discuss this in more detail in the next lecture.

7. ORDERED FIELDS

In this lecture we define the notion of a field and ordered field, which will allow us to give an axiomatic definition of rationals. We then show that rationals are unique up to a bijection that preserves the ordered field structure which we will later to (similarly) axiomatize the reals and prove their uniqueness. Such statements are necessary to ensure that there is only one real analysis one can build out of Zermelo's axioms.

7.1 Axioms of ordered fields.

We start by a definition that is standard in algebra:

Definition 7.1 (A field) *A set F with binary operations $+$ and $\cdot$ and two distinct distinguished elements 0 and 1 is a* field *if*

- (F1) *the operations of addition and multiplication obey the commutative and associative laws (each of them separately) as well as the distributive law,*
- (F2) *(Zero element)* $\forall a \in F\colon\ 0 + a = a \wedge 0 \cdot a = 0$,
- (F3) *(Unit element)* $\forall a \in F\colon\ 1 \cdot a = a$,
- (F4) *(Additive inverse)* $\forall a \in F\ \exists(-a) \in F\colon\ a + (-a) = 0$,
- (F5) *(Multiplicative inverse)* $\forall a \in F \setminus \{0\}\ \exists a^{-1} \in F\colon\ a \cdot a^{-1} = 1$.

We will write $(F, +, 0, \cdot, 1)$ to denote the field F along with all its important attributes.

One can check that, in every field, the following "usual" facts hold:

- The zero and unit elements are unique. Indeed, assuming 0 and $0'$ are both zero elements, then $0 = 0 + 0' = 0'$. The proof for the unit element is similar.
- The additive and multiplicative inverses are unique. Indeed, focusing on the multiplicative inverse, if $\tilde{a}^{-1}$ and a^{-1} are both inverses to $a \neq 0$, then the associative law for multiplication shows $\tilde{a}^{-1} = \tilde{a}^{-1} \cdot (a \cdot a^{-1}) = (\tilde{a}^{-1} \cdot a) \cdot a^{-1} = a^{-1}$.
- The operations of addition and multiplication by non-zero number are injective. Indeed, we have

$$\forall a, b, c \in F\colon\ \ a + b = a + c \Rightarrow b = c \tag{7.1}$$

  by adding the additive inverse to both sides and using the associative. Using the multiplicative inverse instead we get injectivity for multiplication by non-zero number,

$$\forall a, b, c \in F\colon\ \ (a \neq 0\ \wedge a \cdot b = a \cdot c) \Rightarrow b = c \tag{7.2}$$

- The product of any two non-zero elements is non-zero,

$$\forall a, b \in F\colon (a \neq 0 \wedge b \neq 0) \Rightarrow a \cdot b \neq 0 \tag{7.3}$$

  because injectivity of multiplication turns $a \cdot b = 0 = a \cdot 0$ and $a \neq 0$ into $b = 0$. On the other hand, multiplying any number by zero results in zero,

$$\forall a \in F\colon a \cdot 0 = 0 \tag{7.4}$$

  due to the fact that $a \cdot 0 = b$ along with commutative and distributive laws and the definition of zero and unity implies $a + b = a \cdot 1 + a \cdot 0 = a \cdot (1 + 0) = a \cdot 1 = a$ which by injectivity of addition gives $b = 0$.

- The negative sign can be written arbitrarily in the product: For all $a, b \in F$ we have $(-a) \cdot b = -(a \cdot b) = a \cdot (-b)$ and thus $(-a) \cdot (-a) = a \cdot a$. Similarly, we infer $(-a)^{-1} = -a^{-1}$ for all $a \neq 0$.

Theorem 6.6 shows that $\mathbb{Q}$ endowed with operations (6.16–6.17) is a field. There are other natural examples of fields. For instance, there are many finite fields. (We have yet to define the term "finite" but in these examples this will be clear intuitively.) The simplest non-trivial example is $F_2 := \{0, 1\}$, where addition is defined by

$$0 + 0 := 0, \quad 0 + 1 = 1 + 0 := 1, \quad 1 + 1 := 0 \tag{7.5}$$

and multiplication by

$$0 \cdot 0 := 0, \quad 0 \cdot 1 = 1 \cdot 0 := 0, \quad 1 \cdot 1 := 1 \tag{7.6}$$

In the absence of other elements, F_2 is clearly a field. (Note that, in this field, $-1 = 1$.) This example generalizes for any p prime to $F_p := \mathbb{Z}/(p\mathbb{Z})$ which can be identified with the set $\{0, 1, \ldots, p-1\}$ and the operations of addition and multiplication as in $\mathbb{Z}$ except taken modulo p. Note that F_p is a field *only if* p is a prime; indeed, otherwise there are two non-zero elements — namely, the divisors of p — that multiply to zero.

The need to rule out the example (7.5–7.6) as well as F_p for p prime from consideration (as good sets of numbers for real analysis) motivates us restrict the concept of a field further by requiring the existence of an ordering relation:

Definition 7.2 (Ordered field) *We say that a field $(F, +, 0, \cdot, 1)$ is an* ordered field *if F admits a binary relation $\leqslant$ that*

(O1) *is a total ordering, i.e., is reflexive, antisymmetric and transitive and is connex in the sense that $\forall a, b \in F\colon\ a \leqslant b \vee b \leqslant a$,*

(O2) *is preserved by addition, i.e.,*

$$\forall a, b, c \in F\colon\ a \leqslant b \ \Rightarrow\ a + c \leqslant b + c \tag{7.7}$$

(O3) *is preserved by multiplication by non-negative numbers, i.e.,*

$$\forall a, b, c \in F\colon \big(a \leqslant b \wedge 0 \leqslant c\big) \ \Rightarrow\ a \cdot c \leqslant b \cdot c. \tag{7.8}$$

We will write $(F, +, 0, \cdot, 1, \leqslant)$ for an ordered field with the ordering relation $\leqslant$.

The properties O1-O3 directly imply:

Lemma 7.3 $0 \leqslant 1$ *holds in any ordered field $(F, +, 0, \cdot, 1, \leqslant)$.*

Proof. Assume, on the way to a contradiction, that $1 < 0$. Then (O2) shows $0 < -1$ and, since -1 is non-negative, (O3) gives $0 = 0 \cdot (-1) \leqslant (-1) \cdot (-1) = 1$, in contradiction with the assumption. Hence $0 \leqslant 1$ by the fact that $\leqslant$ is a total order; see (O1). □

We now list further properties which the reader will find completely standard but which, to get a fully rigorous treatment, now have to be verified from the definition of an ordered field:

Lemma 7.4 *Let $(F, +, 0, \cdot, 1, \leqslant)$ be an ordered field. Then*

(1) $\forall a, b \in F\colon\ 0 \leqslant b \Leftrightarrow a \leqslant a + b$

(2) $\forall a, b \in F\colon\ a \leqslant b \Rightarrow -b \leqslant -a$

(3) $\forall a, b \in F\colon (0 < a\ \wedge a \leqslant b) \Rightarrow b^{-1} \leqslant a^{-1}$

Proof. Left to a homework exercise. □

7.2 Axiomatizing the rationals.

In order to get to the axiomatic definition of the rationals, we make the following important observation:

Lemma 7.5 *Let $(F, +, 0, \cdot, 1, \leqslant)$ be an ordered field. Set*

$$\mathbb{N}_F := \bigcap \big\{ A \subseteq F \colon 0 \in A \wedge (\forall x \in A \colon x + 1 \in A) \big\} \tag{7.9}$$

and let $S_F(x) := x + 1$. Then $(\mathbb{N}_F, 0, S_F)$ is a system of naturals.

Proof. We need to verify the Peano axioms P1-P5. First note that the set $A := F$ contributes to the intersection, which is thus non-empty. It follows that $\mathbb{N}_F$ is a set which, since every A in (7.9) contains 0, obeys $0 \in \mathbb{N}_F$, thus proving P1. Since every set on the rigth of (7.9) is closed under S_F, we also have that S_F is a map $\mathbb{N}_F \to \mathbb{N}_F$, proving P2. The map S_F is injective (even on F) because $x + 1 = y + 1$ implies $x = y$ by (7.1), proving P4. As to the range of S_F omitting zero, here we note that the set $\{x \in F \colon 0 \leqslant x\}$ of non-negative elements contributes on the right of (7.9). As $0 < 1$ by Lemma 7.3 and $x \geqslant 0$ implies $x + 1 \geqslant 1 > 0$, we have $0 \notin S_F(\mathbb{N}_F)$, proving P3.

It remains to prove the Induction Axiom P5. For this, let $A \subseteq \mathbb{N}_F$ such that $0 \in A$ and $S_F(A) \subseteq A$. But then A appears among the sets on the right of (7.9) and so $\mathbb{N}_F \subseteq A$. Hence $A = \mathbb{N}_F$ proving P5 as well. □

We will refer to $\mathbb{N}_F$ as the *naturals of F*. We leave it to the reader to prove:

Lemma 7.6 *The operations of addition and multiplication as well as the binary relation $\leqslant$ defined on $\mathbb{N}_F$ using S_F coincide with those in F. Moreover, 1 in F equals $S_F(0)$.*

We are now able to axiomatize the rationals:

Definition 7.7 *A system of rationals is an ordered field $(F, +, 0, \cdot, 1, \leqslant)$ such that*

$$\forall x \in F \; \exists m, n, r \in \mathbb{N}_F \colon \; r \neq 0 \wedge x = r^{-1} \cdot (m - n) \tag{7.10}$$

where $\mathbb{N}_F$ are the naturals of F.

As we have shown above, there is at least one systems of rationals. (The ordering relation is defined in (6.22). Checking that these satisfy O1-O3 in Definition 7.2 is left to the reader.) It remains to prove that the rationals are, in fact, unique:

Theorem 7.8 (Uniqueness of the rationals) *Let $(F, +, 0, \cdot, 1, \leqslant)$ and $(\widetilde{F}, \widetilde{+}, \widetilde{0}, \widetilde{\cdot}, \widetilde{1}, \widetilde{\leqslant})$ be two systems of rationals. Then there exists a bijection $\phi \colon F \to \widetilde{F}$ such that*

(1) $\forall a, b \in F \colon \phi(a + b) = \phi(a) \mathbin{\widetilde{+}} \phi(b)$,
(2) $\forall a, b \in F \colon \phi(a \cdot b) = \phi(a) \mathbin{\widetilde{\cdot}} \phi(b)$
(3) $\phi(0) = \widetilde{0} \;\wedge\; \phi(1) = \widetilde{1}$,
(4) $\forall a, b \in F \colon a \leqslant b \Rightarrow \phi(a) \mathbin{\widetilde{\leqslant}} \phi(b)$.

In particular, a system of rationals is unique up to an isomorphism.

Proof (main steps). We only give the main steps leaving the details to the reader. The uniqueness of the naturals (see Theorem 4.7 and Lemma 7.5) imply the existence of a bijection $\phi\colon \mathbb{N}_F \to \mathbb{N}_{\widetilde{F}}$ with

$$\phi(0) = \widetilde{0} \ \wedge \ \phi \circ S_F = S_{\widetilde{F}} \circ \phi \tag{7.11}$$

In particular, we have

$$\phi(1) = \phi \circ S_F(0) = S_{\widetilde{F}} \circ \phi(0) = S_{\widetilde{F}}(\widetilde{0}) = \widetilde{1} \tag{7.12}$$

proving (3) above. In light of (7.11) and Lemma 7.6 we now readily check that ϕ takes the operation of addition $+$ to $\widetilde{+}$; i.e.,

$$\forall m, n \in \mathbb{N}_F\colon \phi(m+n) = \phi(m) \mathbin{\widetilde{+}} \phi(n) \tag{7.13}$$

Then same applies to multiplication,

$$\forall m, n \in \mathbb{N}_F\colon \phi(m \cdot n) = \phi(m) \mathbin{\widetilde{\cdot}} \phi(n) \tag{7.14}$$

(Both of these statements are readily proved by induction.)

Next we extend ϕ to F as follows: If $x = r^{-1} \cdot (m-n)$ for some $m, n, r \in \mathbb{N}_F$ with $r \neq 0$, then we set

$$\phi(x) := \phi(r)^{-1} \mathbin{\widetilde{\cdot}} \big(\phi(m) \ \widetilde{+} \ (-\phi(n))\big) \tag{7.15}$$

To see that the right-hand side does not depend on m, n and r we used to represent x, assume that x can also be written as $x = \hat{r}^{-1} \cdot (\hat{m} - \hat{n})$. Then

$$r \cdot (\hat{m} - \hat{n}) = \hat{r} \cdot (m-n) \tag{7.16}$$

which, we note, is nothing other than the statement of equivalence relation (6.13). As this expression involves only naturals, (7.13–7.14) give

$$\phi(\hat{r}) \mathbin{\widetilde{\cdot}} \big(\phi(m) \ \widetilde{+} \ (-\phi(n))\big) = \phi(r) \mathbin{\widetilde{\cdot}} \big(\phi(\hat{m}) \ \widetilde{+} \ (-\phi(\hat{n}))\big) \tag{7.17}$$

which then translates into

$$\phi(\hat{r})^{-1} \mathbin{\widetilde{\cdot}} \big(\phi(\hat{m}) \ \widetilde{+} \ (-\phi(\hat{n}))\big) = \phi(r)^{-1} \mathbin{\widetilde{\cdot}} \big(\phi(m) \ \widetilde{+} \ (-\phi(n))\big) \tag{7.18}$$

proving that (7.15) is the same for (m, n, r) as for $(\hat{m}, \hat{n}, \hat{r})$. In particular, ϕ is a function $F \to \widetilde{F}$ with $\mathrm{Dom}(\phi) = F$.

We now reuse the previous argument to prove that ϕ is injective. Indeed, suppose $x = r^{-1} \cdot (m-n)$ and $y = \hat{r}^{-1} \cdot (\hat{m} - \hat{n})$ are such that $\phi(x) = \phi(y)$. This means that (7.18) holds. But then also (7.17) holds and, using that ϕ is invertible, (7.13–7.14) implies (7.16), proving that $x = y$. For surjectvity pick $z \in \widetilde{F}$ and use (7.10) to write it as $z = \tilde{r}^{-1} \cdot (\tilde{m} - \tilde{n})$ for some $\tilde{m}, \tilde{n}, \tilde{r} \in \mathbb{N}_{\widetilde{F}}$. The fact that ϕ is surjective implies existence of $m, n, r \in \mathbb{N}_F$ so that $\phi(r) = \tilde{r}$, $\phi(m) = \tilde{m}$ and $\phi(n) = \tilde{n}$. This gives z the representation on the right of (7.15) thus showing that $z \in \mathrm{Ran}(\phi)$. We conclude that ϕ is a bijection.

Using (7.13–7.14) and (7.15) we now readily check that ϕ takes addition in F to addition in $\widetilde{F}$ and multiplication in F to multiplication in $\widetilde{F}$, completing the proof of properties (1-3) in the statement. In order to prove (4), by additivity of ϕ it suffices to focus on the case $a = 0$. Here we note that, for $x \geqslant 0$, we can take $m \geqslant 0$, $r > 0$ and $n = 0$ in (7.15). The fact that property (4) holds for the naturals then implies $\phi(x) \geqslant 0$ as desired. □

Having proved the uniqueness of the rationals, a natural question whether there are in fact other ordered fields than rationals. We will answer this in the next two lectures.

8. Algebraic deficiencies of rationals

At first sight, the rationals appear to have most of the algebraic properties needed for daily operations with numbers. Indeed, they allow for addition, multiplication as well as the inverse operations of subtraction and division (by non-zero numbers). However, once other natural operations are introduced, problems arise.

8.1 The need for irrationals.

We start by recalling that (natural) powers are defined recursively in any ordered field $(F, +, 0, \cdot, 1, \leqslant)$ so that

$$\forall b \in F\colon \quad b^0 = 1 \ \wedge \ \big(\forall n \in \mathbb{N}_F\colon b^{n+1} = b \cdot b^n\big) \tag{8.1}$$

With these in hand we can ask for solutions to polynomial equations such as

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0 \tag{8.2}$$

for some natural $n \in \mathbb{N}_F$ and *coefficients* $a_0, \ldots, a_n \in F$, where we began to adopt the convention that the multiplication sign $\cdot$ can be omitted when no confusion arises. The simplest non-trivial case of (8.2) is arguably the quadratic equation

$$x^2 = a. \tag{8.3}$$

By the Pythagorean theorem, a positive x solving this equation gives the length of the hypothenuse in the right triangle whose legs-squared add up to a. (This remains relevant to present day: Builders use the right triangle with sides of lengths 3, 4 and 5 to check that walls meet in the corner at the right angle.)

Since the square of any number is non-negative (prove this from axioms of the field!), (8.3) has no solution for $a < 0$. However, a solution clearly exists for some positive $a \in \mathbb{Q}$, e.g., $a = 4$ or $a = \frac{4}{9}$. Unfortunately, as was noted already by ancient Greeks, there are also positive $a \in \mathbb{Q}$ for which (8.3) admits no rational solution:

Lemma 8.1 (Euclid) $\quad \forall x \in \mathbb{Q}\colon x^2 \neq 2$

Proof. Suppose, on the way to a contradiction, that there is $x \in \mathbb{Q}$ with $x^2 = 2$. Since x is rational, we have $x = p/q$ for some non-zero $p, q \in \mathbb{Z}$ (note that the square of zero is zero). Note that by multiplying p or q or both by -1, we may achieve this with $p, q > 0$. A key point is that we can take p and q so that

$$\text{not both } p \text{ and } q \text{ are event} \tag{8.4}$$

This is usually argued as "obvious" as with p and q even we can divide out a factor of 2 from each of them and, repeating, achieve that at least one is odd. However, this is not an argument that works in proper set theory; instead, we take

$$q := \min\big\{\tilde{q} \in \mathbb{N} \smallsetminus \{0\}\colon (\exists \tilde{p} \in \mathbb{Z}\colon \tilde{p} = x \cdot \tilde{q})\big\} \tag{8.5}$$

Note that the assumption $x \in \mathbb{Q}$ ensures that the set is non-empty; the minimum exists by an argument we will show in Lemma 9.10. With q defined as above, $p := x \cdot q$ and note that (8.4) holds for otherwise q would not be minimal.

From $x^2 = 2$ we then get $(p/q)^2 = 2$ and so $p^2 = 2q^2$. Since the square of an odd number is odd (prove this!), we get

$$p \text{ is even.} \tag{8.6}$$

But then there exists $r \in \mathbb{Z} \setminus \{0\}$ such that $p = 2r$ and so $q^2 = 2r$. Hence we also get

$$q \text{ is even.} \tag{8.7}$$

But this contradicts our assumption that q is a minimal element in (8.5) because $q/2$ is a member of that set as well. Hence $x^2 = 2$ has no solution in $\mathbb{Q}$. □

8.2 Testing rationality.

Lemma 8.1 readily generalizes to $x^n = p$ having no rational solution for any $p \in \mathbb{N}$ prime and any $n \in \mathbb{N}$ different from 0 and 1. Here

$$p \text{ is a prime } := \ p \in \mathbb{N} \setminus \{0,1\} \ \wedge \ \big(\forall q \in \mathbb{N} \setminus \{0,1\} \colon q|p \Rightarrow q = p\big) \tag{8.8}$$

where

$$m|n \ := \ (\exists k \in \mathbb{N} \colon \ n = k \cdot m) \tag{8.9}$$

The mathematicians of middle ages (and even ancient Greeks, who knew of the right triangle with legs of unit length having the hypothenuse of length that is not a rational number) were quite aware of this problem. A solution was proposed based using the concept of a *radical*. The idea is to introduce new elements into the existing number system that are defined as a solution of the polynomial equation $x^n = a$ for some natural n and some a that is already in the number system.

For instance, $\sqrt{2}$ is defined to be the positive number that solves the equation $x^2 = 2$, while $\sqrt[5]{7}$ is the number that solves $x^5 = 7$. The process can be iterated, which means that once $\sqrt{2}$ is already in our number system, we define $\sqrt{2+\sqrt{2}}$ to be a positive solution to the equation $x^2 = 2 + \sqrt{2}$ which then resolves into $(x^2-2)^2 = 2$ and thus

$$x^4 - 4x^2 + 2 = 0 \tag{8.10}$$

A natural question is then: Which expressions involving radicals are rational and which are not? Some insight into this question is offered by:

Theorem 8.2 (Rational root test) *Suppose that $x \in \mathbb{Q}$ solves*

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0, \tag{8.11}$$

where $a_0, \ldots, a_n \in \mathbb{Z}$ and $a_0, a_n \neq 0$. Then

$$\exists p, q \in \mathbb{Z} \setminus \{0\} \colon \quad x = \frac{p}{q} \ \wedge \ \gcd(p,q) = 1 \ \wedge \ p|a_0 \ \wedge \ q|a_n \tag{8.12}$$

In words, any rational solution can be written as the ratio $\frac{p}{q}$ of two non-zero integers with no non-trivial common divisors such that p divides a_0 and q divides a_n.

Here, to interpret (8.12), we recall that the *greatest common divisor* $\gcd(m,n)$ of m and n is the largest natural that divides both m and n. (The existence of this number will again be shown via Lemma 9.10.)

Proof of Theorem 8.2. Write x as p/q where $p, q \in \mathbb{Z}$ with $q > 0$ and $\gcd(p,q) = 1$. (Again, this is achieved by taking a representation with smallest positive q.) Since $a_0 \neq 0$ we

have $x \neq 0$ and so $p \neq 0$. Substituting $x = p/q$ into (8.11) and multiplying the whole equation by q^n then yields

$$a_n p^n = -q\big[a_{n-1}p^{n-1} + a_{n-2}p^{n-2}q + \cdots + a_1 pq^{n-2} + a_0 q^{n-1}\big] \tag{8.13}$$

The square bracket is an integer and so q divides $a_n p^n$. But the fact that $\gcd(p,q) = 1$ then forces $q|a_n$ as claimed. The proof that $p|a_0$ is similar and thus omitted. □

To demonstrate the use of Theorem 8.2 on an example, the conclusion says that any rational solution to (8.10) is an integer that divides 2, which leaves $-2, -1, 1, 2$ as only possible candidates. As none of these solves (8.10), we conclude $\sqrt{2+\sqrt{2}} \notin \mathbb{Q}$.

The reader should be aware that not all expressions involving radicals are necessarily non-rational: Obviously, $\sqrt{4}$ is rational but so is $\sqrt{7+2\sqrt{3}} - \sqrt{3}$ because

$$\sqrt{7+2\sqrt{3}} - \sqrt{3} = \sqrt{(2+\sqrt{3})^2} - \sqrt{3} = 2 + \sqrt{3} - \sqrt{3} = 2 \tag{8.14}$$

While Theorem 8.2 is typically used to rule out rational roots, it in fact outputs a finite set of numbers as possible candidates for rational roots and so, if one of these solves (8.11), it gives us a rational root if one exists. However, using this for expressions as in (8.14) is not practical as the main point there is to show that the expression simplifies.

8.3 Field extensions.

A formal way to add the radical $\sqrt{2}$ to our "valid" set of numbers (which so far are only the rationals) is by introducing the set

$$F := \big\{a + b\sqrt{2} \colon a, b \in \mathbb{Q}\big\} \tag{8.15}$$

We will now define addition on F by

$$(a + b\sqrt{2}) + (\tilde{a} + \tilde{b}\sqrt{2}) := (a + \tilde{a}) + (b + \tilde{b})\sqrt{2} \tag{8.16}$$

and multiplication by

$$(a + b\sqrt{2}) \cdot (\tilde{a} + \tilde{b}\sqrt{2}) := \big(a \cdot \tilde{a} + 2 \cdot b \cdot \tilde{b}\big) + (a \cdot \tilde{b} + \tilde{a} \cdot b)\sqrt{2}. \tag{8.17}$$

Writing $0 + 0\sqrt{2}$ for the zero element and $1 + 0\sqrt{2}$ for the unit element, we then check that the inverse to $a + b\sqrt{2}$ under addition is

$$-(a + b\sqrt{2}) = -a + (-b)\sqrt{2} \tag{8.18}$$

while that under multiplication is

$$(a + b\sqrt{2})^{-1} = \frac{a}{a^2 - 2b^2} + \frac{-b}{a^2 - 2b^2}\sqrt{2} \tag{8.19}$$

where we noted that $\forall a, b \in \mathbb{Q} \colon\ a^2 - 2b^2 \neq 0$ by Lemma 8.1. Here, for convenience of expression, we wrote rationals as fractions instead of invoking inverse elements. Lemma 8.1 also guarantees that the representation of any element of (8.15) using two rationals a and b is unique. We thus conclude that (8.15) is a *field extension* of $\mathbb{Q}$ by element $\sqrt{2}$. With some additional work — tantamount to writing $(\tilde{a} - a) \geqslant (b - \tilde{b})\sqrt{2}$ without the square root — we can even give (8.15) the structure of an ordered field.

Note that (8.15) can be thought of as a linear vector space over the field $\mathbb{Q}$ with basis $\{1, \sqrt{2}\}$. Thanks to this vector space being also a field, the extension procedure can be

iterated and further radicals gradually added. This leads to the concept a solution of polynomial equations "in radicals" to be defined next:

Definition 8.3 (Solution in radicals) *Let $P(x)$ be a polynomial in x with rational coefficients. We say that the equation $P(x) = 0$ admits a solution* in radicals *if there exists $m \in \mathbb{N}$ and fields $F_0, F_1, \dots, F_m$ such that:*

(1) $F_0 = \mathbb{Q}$
(2) $\forall i = 0, \dots, m-1\ \exists z \in F_{i+1}\ \exists a \in F_i\ \exists k \in \mathbb{N}$:

$$z^k = a \ \wedge\ F_{i+1} = \{b_0 + b_1 z + \cdots + b_{k-1} z^{k-1} \colon b_0, \dots, b_{k-1} \in F_i\} \tag{8.20}$$

(3) $\exists x \in F_m \colon P(x) = 0$

(We are not specifying how addition and multiplication acts on each F_i as this is part of the fact that F_i is a field.)

The reason why all powers less than k-th are listed on the right of (8.20) is that only then the set is closed under multiplication. Additional conditions are needed to ensure that the set in (8.20) is a field (e.g., if $z^j \in F_i$ for some $j < k$ then we may have $b_0 + b_1 z + \cdots + b_{k-1} z^{k-1} = 0$ without all $b_0, \dots, b_{k-1}$ vanishing) but the statement does not care about these as we assume that F_{i+1} is a field to begin with.

Translating Definition 8.3 into more laymen terms, we are trying to find a sequence of symbols of the form $\sqrt[k]{a}$ — that is, solutions to equations of the kind $x^k = a$ — where a is expressed as a polynomial in all symbols obtained thus far, so that, when this process terminates, we are able to write a solution of the polynomial equation $P(x) = 0$ of interest. Or, put even more simply, a solution in radicals is that which uses only a finite number of additions, multiplications (which includes subtractions and divisions, of course) and taking roots of any degree.

To demonstrate how the field extension works in practice, pick $a, b \in \mathbb{Q}$ and consider the quadratic equation

$$x^2 + ax + b = 0 \tag{8.21}$$

This is solved by "completing the square" which rewrites the equation as

$$\Big(x - \frac{a}{2}\Big)^2 = \frac{a^2}{4} - b \tag{8.22}$$

Assuming the right-hand side to be non-negative, we introduce the radical $\sqrt{a^2/4 - b}$ and solve the equation by $x = a/2 \pm \sqrt{a^2/4 + b}$. In particular, (8.21) can be solved in the field extension $F := \{p + q\sqrt{a^2/4 - b} \colon p, q \in \mathbb{Q}\}$ of $\mathbb{Q}$.

Similarly, the cubic equation

$$x^3 + ax^2 + bx + c = 0 \tag{8.23}$$

is turned by the substitutions

$$y := x + \frac{a}{3}, \quad \tilde{b} := b - \frac{a^2}{3}, \quad \tilde{c} := c - \frac{a^2}{27} + \frac{a}{3}\Big(b - \frac{a^2}{3}\Big) \tag{8.24}$$

into

$$y^3 + \tilde{b}y + \tilde{c} = 0 \tag{8.25}$$

If $\tilde{b} = 0$, then we solve this by introducing the radical $\sqrt[3]{\tilde{c}}$ and its second power $(\sqrt[3]{\tilde{c}})^2$ into our field. If instead $\tilde{b} > 0$, we introduce the radical $\sqrt{\tilde{b}/3}$ and use the substitution $y := z\sqrt{\tilde{b}/3}$ to rewrite the above as

$$z^3 + 3z + d = 0 \quad \text{where} \quad d := \frac{9\tilde{c}}{\tilde{b}}\sqrt{\tilde{b}/3} \tag{8.26}$$

Employing the usual notation for fractional powers (which we have not yet quite defined but should be familiar with), this is now solved by the substitution

$$z := u^{1/3} - u^{-1/3} \tag{8.27}$$

which converts the cubic to $u - u^{-1} + d = 0$; i.e., the *quadratic* equation

$$u^2 + du - 1 = 0 \tag{8.28}$$

We solve this readily by introducing the radical $\sqrt{1 + d^2/4}$ to our number system, which gives a solution in the form $u = -d/2 + \sqrt{1 + d^2/4}$. (We are for simplicity ignoring the second root.) Returning to z, we then introduce the radicals

$$\sqrt[3]{-d/2 + \sqrt{1 + d^2/4}} \quad \text{and} \quad \left(\sqrt[3]{-d/2 + \sqrt{1 + d^2/4}}\right)^2 \tag{8.29}$$

into our number system to write the solution in the form

$$z = \sqrt[3]{-d/2 + \sqrt{1 + d^2/4}} + \frac{1}{-d/2 + \sqrt{1 + d^2/4}}\left(\sqrt[3]{-d/2 + \sqrt{1 + d^2/4}}\right)^2 \tag{8.30}$$

(Note that the sign of d does not matter as cube-roots exist regardless of the sign.) Reversing the earlier substitutions, this now yields a solution of (8.23) "in radicals." The case of $\tilde{b} < 1$ is handled very similarly, except for working with $\sqrt{-\tilde{b}/3}$ and some negative signs popping up in calculations down the line.

8.4 Failure of the "radicals" approach.

Besides the quadratic equation (the case $n = 2$ in (8.11)) solved above, a solution in radicals turns out to be possible for the *cubic* equation (the case $n = 3$ in (8.11)) and the *quartic* equation — the case $n = 4$ in (8.11)) thanks to the classical solutions due to L. Ferrari (quartic equation, solved in 1540) and G. Cardano (cubic equation, solved in 1545). Notice that we constructed a solution of a general cubic above.

Unfortunately, this is where the program of solving polynomial equations in radicals runs into a roadblock. Indeed, as shown by P. Ruffini in 1799 (in a somewhat controversial 100-page paper) and N.H. Abel in 1824 (in a 6-page paper), no solution in radicals exists for the quintic

$$x^5 - x - 1 = 0. \tag{8.31}$$

(Being a quintic, this equation does have at least one real root.) In 1830, E. Galois developed tools to determine whether a given polynomial equation admits a solution in radicals, thus founding what is now called Galois theory.

With the process of gradually adding radicals to rationals failing to describe the solutions to even some basic polynomial equations, we may try to take a more abstract

approach and consider simply all numbers that (quite loosely) solve *some* polynomial equation (8.11) for *some* non-trivial integer coefficients. (We will need to define the reals first to make this precise.) Such numbers are called *algebraic*. Unfortunately, as it turns out, even these are not sufficient to give us important numbers such as π, or the Euler number e that are fundamental for analysis. It follows (and this is the punchline of this section) that we will have to approach the reals using different means than just algebra. This is what we will do next.

9. SUPREMUM AND INFIMUM

The algebraic deficiencies described above seem to be related to the fact that the rational axis contains "punctures" a.k.a. as "holes." Filling some of these "holes" with radicals helps somewhat but (as we explained towards the end of the last section) "holes" remain even if all roots of all polynomial equations with integer coefficients are added in. As it turns out, the presence of the "holes" is closely related to another deficiency of the rationals, this time related to the total ordering relation $\leqslant$. We will start discussing the relevant concepts at the general level and then specialize to rationals.

9.1 Definition of sup/inf and basic properties.

Consider a set E with an ordering $\leqslant$; i.e., a reflexive, antisymmetric and transitive relation. We will sometimes refer to $(E, \leqslant)$ as a *poset*; i.e., a commonly used shorthand for a partially-ordered set. We do not assume that every pair of elements from E is ordered; so $\leqslant$ can be just a partial order. We then introduce the following concepts:

Definition 9.1 (Upper/lower bound) *Let $(E, \leqslant)$ be a poset. Given a set $A \subseteq E$, an element $x \in E$ is*

- *an* upper bound on A *if* $\forall y \in A\colon\ y \leqslant x$,
- *a* lower bound on A *if* $\forall y \in A\colon\ x \leqslant y$.

If A admits an upper bound, we say that it is bounded above *while if it admits a lower bound, we say that it is* bounded below. *If it admits both, then we say that A is* bounded.

Notice that the definition of an upper (or lower) bound entails two consequences: First, every element of A compares to x and, second, the comparison is as stated. Here are some examples:

- Given a set F, let E be the powerset $\mathcal{P}(F)$ ordered by the set inclusion,
$$\forall A, B \in \mathcal{P}(F)\colon\ A \leqslant B \ :=\ A \subseteq B \tag{9.1}$$
For any set $A \subseteq \mathcal{P}(F)$ of subsets of F, including the case when A is empty, the element $x := F$ is an upper bound on A and $x := \varnothing$ is a lower bound on A. (Hence, $x := F$ is the *maximal element* of $\mathcal{P}(F)$ and $x := \varnothing$ is the *minimal element*.)
- Consider the set of pairs $E := \mathbb{Q} \times \mathbb{Q}$ and define the *lexicographic order* on E via
$$(x, y) \leqslant (\tilde{x}, \tilde{y}) \ :=\ x < \tilde{x} \vee \big(x = \tilde{x} \wedge y \leqslant \tilde{y}\big) \tag{9.2}$$
The set
$$A := \{(x, y) \in \mathbb{Q} \times \mathbb{Q}\colon\ 0 \leqslant x, y \leqslant 1\} \tag{9.3}$$
then admits lower bounds $(-1, -1)$, $(-1, 0)$ and even $(0, 0)$ and upper bounds $(2, 2)$, $(1, 2)$ and even $(1, 1)$. On the other hand, the set
$$A := \{(x, y) \in \mathbb{Q} \times \mathbb{Q}\colon\ x + y = 0\} \tag{9.4}$$
admits no upper bound and no lower bound in E.

If a set A admits an upper bound, a natural next question is whether one can find the most efficient upper bound. We take this to mean the *least* upper bound which is the one that compares to and is less than all the other upper bounds. This, and the corresponding concept for the lower bound, is the content of:

Definition 9.2 (Supremum and infimum) *Given a set $A \subseteq E$, an element $x \in E$ is*

- *the* supremum of A *if x is the least upper bound of A, i.e.,*

$$\big(\forall y \in A\colon y \leqslant x\big) \wedge \Big(\forall z \in E\colon \big(\forall y \in A\colon y \leqslant z\big) \Rightarrow x \leqslant z\Big) \tag{9.5}$$

- *the* infimum of A *if it is the greatest lower bound of A, i.e.,*

$$\big(\forall y \in A\colon x \leqslant y\big) \wedge \Big(\forall z \in E\colon \big(\forall y \in A\colon z \leqslant y\big) \Rightarrow z \leqslant x\Big) \tag{9.6}$$

We note that, just as upper/lower bounds, a supremum/infimum may not exist. Still, the use of the definite article in the definition is justified by:

Lemma 9.3 *For every set A, there is at most one supremum and at most one infimum.*

Proof. Suppose $x \in E$ and $\tilde{x} \in E$ are both suprema of A. Then both x and $\tilde{x}$ are upper bounds on A. The fact that x is a supremum of A then forces $x \leqslant \tilde{x}$ while the fact that $\tilde{x}$ is a supremum of A forces $\tilde{x} \leqslant x$. The antisymmetry of $\leqslant$ then forces $\tilde{x} = x$. The argument for the infimum is analogous and so we omit it. □

We will henceforth write $\sup(A)$ for the supremum of A and $\inf(A)$ for the infimum of A, whenever these elements exist. To give an example where the existence can be guaranteed, we note:

Lemma 9.4 *For the setting of $E := \mathcal{P}(F)$ (with F a non-empty set) ordered by set inclusion $\subseteq$,*

$$\forall A \subseteq \mathcal{P}(F)\colon\ A \neq \varnothing \ \Rightarrow\ \Big(\sup(A) = \bigcup A \ \wedge\ \inf(A) = \bigcap A\Big) \tag{9.7}$$

In addition, we have $\sup(\varnothing) = \varnothing$ and $\inf(\varnothing) = F$.

Proof. Left to homework. □

The trivial inclusion $\bigcap A \subseteq \bigcup A$ implies that $\inf(A)$ is less or equal than $\sup(A)$ for A nonempty. This in fact holds very generally:

Lemma 9.5 *Let $(E, \leqslant)$ be a poset. Then for any $A \subseteq E$ admitting both $\sup(A)$ and $\inf(A)$,*

$$A \neq \varnothing \ \Rightarrow\ \inf(A) \leqslant \sup(A) \tag{9.8}$$

Proof. Since $A \neq \varnothing$, there exists $a \in A$. But then $a \leqslant \sup(A)$ by the fact that the supremum is an upper bound and $\inf(A) \leqslant a$ by the fact that infimum is a lower bound. The transitivity of $\leqslant$ then gives the claim. □

The above shows that, for non-empty set, the infimum and supremum are ordered intuitively. However, as demonstrated by the case of $A = \varnothing$ in Lemma 9.4 this fails for $A = \varnothing$. (We will see another example of this when we talk about extended reals.) Another useful fact is:

Lemma 9.6 *Let $(E, \leqslant)$ be a poset. Then for all non-empty sets $A, B \subseteq E$ such that $\inf(B)$ as well as $\sup(A)$ exist,*

$$\Big(\forall a \in A\, \forall b \in B\colon a \leqslant b\Big) \ \Leftrightarrow\ \sup(A) \leqslant \inf(B) \tag{9.9}$$

Proof. The implication $\Leftarrow$ is obtained from $a \leqslant \sup(A)$ for each $a \in A$ and $\inf(B) \leqslant b$ for each $b \in B$, so we will focus on $\Rightarrow$ in (9.9). The premise says that every $a \in A$ is a lower bound on B. Since B admits an infimum, i.e., the greatest lower bound, we thus get

$$\forall a \in A\colon\ a \leqslant \inf(B) \tag{9.10}$$

But this means that $\inf(B)$ is an upper bound on A. Since A admits a supremum, i.e., the least upper bound, we must thus have $\sup(A) \leqslant \inf(B)$ as claimed. □

The notions of suprema and infima appear symmetric. In ordered fields, the symmetry is quite perfect and explicit:

Lemma 9.7 *Let $(F, +, 0, \cdot, 1, \leqslant)$ be a complete ordered field. Then any $A \subseteq F$ that admits a lower bound admits an infimum and, in fact,*

$$\inf(A) = -\sup(-A) \tag{9.11}$$

where

$$-A := \{a \in F\colon\ -a \in A\} \tag{9.12}$$

admits an upper bound and thus a supremum.

Proof. The claim hinges on the following fact

$$\forall x \in F\colon\ x \in F \text{ is a lower bound on } A \ \Leftrightarrow\ -x \text{ is an upper bound on } -A \tag{9.13}$$

whose proof we leave to the reader. Using that the field is complete and A admits a lower bound, (9.13) implies that $\sup(-A)$ exists and $-\sup(-A)$ is a lower bound on A. Moreover, if x is another lower bound on A, then (9.13) and the definition of supremum forces $\sup(A) \leqslant -x$ which rewrites into $x \leqslant -\sup(-A)$. Hence, $-\sup(-A)$ is the greatest lower bound on A, and thus the infimum of A. □

9.2 Consequences for the rationals.

Returning back to the example of the ordered field of rationals, we note that some bounded sets of rationals do admit supremum and infimum, e.g.,

$$\sup(\{x \in \mathbb{Q}\colon x^2 \leqslant 4\}) = 2 \quad \wedge \quad \inf(\{x \in \mathbb{Q}\colon x^2 < 4\}) = -2. \tag{9.14}$$

But there are also sets that fail this. For instance, the set $\mathbb{Q}$ admits no supremum and $\varnothing$ admits no infimum simply because the former admits no upper bound and the latter no lower bound. However, even that is not the main obstruction:

Lemma 9.8 *The set $A := \{a \in \mathbb{Q}\colon\ a < 0 \vee a^2 < 2\}$ admits an upper bound in $\mathbb{Q}$ yet no supremum (in $\mathbb{Q}$).*

The proof needs the following elementary observation:

Lemma 9.9 (Archimedean property of $\mathbb{Q}$) $\forall a \in \mathbb{Q}\colon\ a > 0 \Rightarrow (\exists n \in \mathbb{N}\colon an > 1)$

Proof. Let $a \in \mathbb{Q}$ obey $a > 0$. Then $a = p/q$ for some $p, q \in \mathbb{N} \setminus \{0\}$. Let $n := q + 1$. Then

$$anq = p(q+1) = pq + p \geqslant q + 1 > q, \tag{9.15}$$

where the inequality used that $p \geqslant 1$ and $q \geqslant 0$. Multiplying both sides by q^{-1}, which is positive and thus preserves the strict inequality, we get $an > 1$ as desired. □

We are now ready to give:

Proof of Lemma 9.8. Note that $1 \in A$ and $A \subseteq \{a \in \mathbb{Q} \colon a < 2\}$ and so A is non-empty and bounded from above by 2. Suppose for the sake of contradiction, that A admits a supremum $c \in \mathbb{Q}$. Then $1 \leqslant c$ and $c \leqslant 2$. Noting that, for any natural $n \geqslant 1$ we have

$$\left(c + \frac{1}{n}\right)^2 = c^2 + \frac{2c}{n} + \frac{1}{n^2} \leqslant c^2 + \frac{5}{n} \tag{9.16}$$

the inequality $c + \frac{1}{n} > c > 0$ along with the fact that c is the supremum of A forces $c + \frac{1}{n} \notin A$ implying $(c + \frac{1}{n})^2 > 2$ and thus $c^2 + \frac{5}{n} > 2$. This rules out that $c^2 < 2$ because that would imply $\frac{5}{n} > 2 - c^2$ for all natural $n \geqslant 1$, in contradiction with Lemma 9.9.

We thus have $c^2 \geqslant 2$ which by Lemma 8.1 forces $c^2 > 2$. But then

$$\left(c - \frac{1}{n}\right)^2 = c^2 - \frac{2c}{n} + \frac{1}{n^2} \geqslant c^2 - \frac{4}{n} \tag{9.17}$$

shows that $(c - \frac{1}{n})^2 > 2$ because the opposite inequality would give $c^2 - 2 \leqslant \frac{4}{n}$ for all natural $n \geqslant 1$ again contradicting (in light of $c^2 > 2$) Lemma 9.9. Since $c - \frac{1}{n} \geqslant 0$, we conclude that $c - \frac{1}{n}$ is an upper bound on A that is strictly smaller than c, contradicting that c is the least upper bound. Hence, A admits no infimum in $\mathbb{Q}$. □

Of course, once $\sqrt{2}$ has been added to $\mathbb{Q}$, the set A in the previous proof will admit supremum with $\sup(A) = \sqrt{2}$. The absence of the infimum/supremum in $\mathbb{Q}$ is thus reduced to the absence of a rational solution to $x^2 = 2$, or the existence of "hole" at the point $\sqrt{2}$ in the rational line.

9.3 Consequences for the naturals.

Before we move over to the reals, let us point out some facts that the concepts of suprema and infima imply for the naturals:

Lemma 9.10 *Consider the naturals $\mathbb{N}$ ordered by the relation $\leqslant$. Then*

$$\forall A \subseteq \mathbb{N} \colon \quad A \neq \varnothing \;\Rightarrow\; \Big(\inf(A) \text{ exists } \wedge\; \inf(A) \in A\Big) \tag{9.18}$$

Proof. Write $(\mathbb{N}, 0, S)$ for the system of naturals and let $A \subseteq \mathbb{N}$ be non-empty. Define

$$B := \big\{b \in \mathbb{N} \colon (\forall a \in A \colon b \leqslant a)\big\} \tag{9.19}$$

to be the set of lower bounds on A. Our goal is to show that $A \cap B \neq \varnothing$.

Since $0 \leqslant m$ for all $m \in \mathbb{N}$, we have $0 \in B$. Recall that the fact that $b \in B$ is a lower bound on A means that for each $a \in A$ there is $s \in \mathbb{N}$ such that $a = b + s$. If $b \notin A$, then we have $s \neq 0$ and, since $1 \leqslant m$ for all $m \in \mathbb{N} \setminus \{0\}$, conclude that $b + 1 \leqslant a$ for all $a \in A$ and so $b + 1 \in B$. Writing this using the successor map, this yields

$$A \cap B = \varnothing \;\Rightarrow\; S(B) \subseteq B \tag{9.20}$$

But the conclusion and Peano's axiom (P5) force $B = \mathbb{N}$, which is FALSE due to the fact that A contains an element a for which $b := a + 1$ is not a member of B.

Since $A \cap B = \varnothing$ leads to a FALSE conclusion, we have proved $A \cap B \neq \varnothing$. This means there exists $c \in A \cap B$. This c is a lower bound on A such that every $b \in B$ obeys $b \leqslant c$. It follows that c is the infimum of A and $\inf(A) \in A$, as claimed. □

Concerning the supremum, we in turn get:

Lemma 9.11 *Consider the naturals* $\mathbb{N}$ *ordered by the relation* $\leqslant$*. Then*

$$\forall A \subseteq \mathbb{N}\colon \quad A \text{ bounded } \Rightarrow \Big(\sup(A) \text{ exists } \wedge \Big(A \neq \varnothing \Rightarrow \sup(A) \in A \Big) \Big) \tag{9.21}$$

Proof. Let $A \subseteq \mathbb{N}$ be bounded and let $B := \{b \in \mathbb{N}\colon (\forall a \in A\colon a \leqslant b)\}$ be the set of its upper bounds. The assumption that A is bounded means that $B \neq \varnothing$. Lemma 9.10 then implies that $c := \inf(B)$ exists with $c \in B$. Since c is upper bound on A and a lower bound on B, it is the smallest of all upper bounds and so $c = \sup(A)$.

Now assume that $A \neq \varnothing$. If $c \notin A$, then $a + 1 \leqslant c$ for all $a \in A$ and, since this forces $c \neq 0$, there exists $c' \in \mathbb{N}$ such that $c = c' + 1$. But then $a \leqslant c'$ for all $a \in A$, contradicting that c is the least upper bound. Hence $c \in A$ after all. □

The statement of Lemma 9.10 can even be elevated further to get:

Lemma 9.12 *There exists a function* $f\colon \mathcal{P}(\mathbb{N}) \smallsetminus \{\varnothing\} \to \mathbb{N}$ *such that*

$$\mathrm{Dom}(f) = \mathcal{P}(\mathbb{N}) \smallsetminus \{\varnothing\} \ \wedge \ \Big(\forall A \in \mathcal{P}(\mathbb{N}) \smallsetminus \{\varnothing\}\colon \ f(A) = \inf(A) \Big) \tag{9.22}$$

In particular, the Axiom of Choice holds for subsets of $\mathbb{N}$*.*

Proof. Define

$$F := \big\{ (A, b) \in \mathcal{P}(\mathbb{N}) \times \mathbb{N}\colon b \in A \ \wedge \ (\forall a \in A\colon b \leqslant a) \big\} \tag{9.23}$$

We claim that F is the graph of a function f with above properties. Indeed, to see that F is a function, suppose $(A, b) \in F$ and $(A, b') \in F$. Then $b, b' \in A$ and so the second condition in (9.23) gives $b \leqslant b'$ and $b' \leqslant b$, implying $b = b'$. Lemma 9.10 ensures that $\inf(A)$ exists and belongs to A whenever $A \neq \varnothing$. This means that $(A, \inf(A)) \in F$ for all $A \neq \varnothing$ and so the domain on f is $\mathcal{P}(\mathbb{N}) \smallsetminus \{\varnothing\}$ and $f(A) = \inf(A)$ for all A in the domain. □

In the situation when either the infimum or the supremum of a set belongs to this set, we sometimes refer to them using different names:

Definition 9.13 *If* A *is a set such that* $\inf(A)$ *exists and* $\inf(A) \in A$ *we call* $\inf(A)$ *the* minimum *of* A*, with notation* $\min(A)$*. Similarly, if* A *admits* $\sup(A)$ *which belongs to* A*, we call* $\sup(A)$ *the* maximum *of* A*, with notation* $\max(A)$*.*

The reader should interpret these properly in other contexts. For instance, the *maximum of a function* is the supremum of all function values that, in addition, is achieved at some argument. This is not to be confused with the *maximizer*, which is a value of the argument (there could be more than one) where the function achieves its (necessarily unique) maximum.

10. The reals via Dedekind cuts

We now move towards the axiomatic definition (and construction) of the set of real numbers. Not all details will be spelled out; we refer to the aforementioned textbook by Yannis Moschovakis for more details.

10.1 Dedekind cuts.

As noted above, the rationals lack the property that some (even simple) bounded sets fail to admit a supremum and an infimum. In order to formalize this better, we start with the following concept:

Definition 10.1 *An ordered field $(F, +, 0, \cdot, 1, \leqslant)$ is said to be* complete *if every non-empty subset thereof that admits an upper bound admits a supremum.*

The reader might wonder why we are not saying anything about infima of sets that admit a lower bound. This is because for ordered fields we have $\inf(A) = -\sup(-A)$; see Lemma 9.7. We now axiomatize the real numbers as follows:

Definition 10.2 (Real numbers) *A system of reals is any complete ordered field.*

A natural question is whether any such object exists. This is the content of:

Theorem 10.3 (Dedekind 1872, Cantor 1872) *There exists at least one system of reals.*

In order to prove Theorem 10.3, we will follow a construction that represents the real numbers as "half-infinite intervals" of rationals. Since all systems of rationals are isomorphic, we pick one and denote it $(\mathbb{Q}, +, 0, \cdot, 1, \leqslant)$. We then put forward:

Definition 10.4 (Dedekind cut) *We say that a set $A \subseteq \mathbb{Q}$ is a (Dedekind) cut if*

(C1) $A \neq \varnothing \ \wedge \mathbb{Q} \smallsetminus A \neq \varnothing$,
(C2) $\forall a, b \in \mathbb{Q}\colon\ a \in A \ \wedge\ b \leqslant a \Rightarrow b \in A$,
(C3) $\forall a \in A\ \exists b \in A\colon\ a < b$.

We write

$$\mathbb{R} := \big\{A \subseteq \mathbb{Q}\colon \text{C1-C3 hold}\big\} \tag{10.1}$$

for the set of all cuts.

These defining properties C1-C3 may be verbalized as follows: C1 means that the pair $(A, \mathbb{Q} \smallsetminus A)$ forms a non-trivial partition of $\mathbb{Q}$ while C2 means that A is an interval which, by C3 contains no largest element. It is easy to check that

$$\forall a \in \mathbb{Q}\colon\ \{b \in \mathbb{Q}\colon b < a\} \in \mathbb{R} \tag{10.2}$$

and

$$\{b \in \mathbb{Q}\colon b < 0 \ \vee\, b^2 < 2\} \in \mathbb{R} \tag{10.3}$$

In fact, as formalized in the next lemma, all cuts look like like this:

Lemma 10.5 *For all $A \in \mathbb{R}$,*

(1) $\forall a \in \mathbb{Q} \smallsetminus A\ \forall b \in \mathbb{Q}\colon\ a \leqslant b \Rightarrow b \in \mathbb{Q} \smallsetminus A$,
(2) $\forall a \in A\ \forall b \in \mathbb{Q} \smallsetminus A\colon\ a < b$

In addition, we have

$$\forall A \in \mathbb{R}\colon\ \big\{b - a\colon a \in A \wedge b \in \mathbb{Q} \setminus A\big\} = \{c \in \mathbb{Q}\colon c > 0\} \tag{10.4}$$

Proof. Properties (1) and (2) follow directly from C1-C3. For (10.4), $\subseteq$ follows from (2). For the opposite inclusion, assume that the set on the left misses a point $c \in \mathbb{Q}$ with $c > 0$. For all $a \in A$ we then have $a + c \notin \mathbb{Q} \setminus A$ and thus $a + c \in A$. Induction then shows that $a + cn \in A$ for all $n \in \mathbb{N}$. But then $A = \mathbb{Q}$ for otherwise there is $x \in \mathbb{Q} \setminus A$ which, by C2, must obey $a + cn \leqslant x$ for all $n \in \mathbb{N}$ which is impossible by the Archimedean property. But $A = \mathbb{Q}$ is impossible by C1 either and so no such c exists to begin with. □

10.2 Ordering relation for Dedekind cuts.

In order to turn $\mathbb{R}$ into a system of reals, we need to define addition $\oplus$, multiplication $\odot$ and an ordering relation $\preccurlyeq$ so that $\mathbb{R}$ becomes an ordered field. We start by the ordering relation. Define

$$\forall A, B \in \mathbb{R}\colon\ A \preccurlyeq B := A \subseteq B \tag{10.5}$$

We then have:

Lemma 10.6 *$\preccurlyeq$ is a total ordering of $\mathbb{R}$.*

Proof. That $\preccurlyeq$ is a partial order follows from that being true about $\subseteq$; see Lemma 3.4. It remains show that the relation is connex; meaning that

$$\forall A, B \in \mathbb{R}\colon A \preccurlyeq B \vee B \preccurlyeq A \tag{10.6}$$

Assume for contradiction that this fails for some $A, B \in \mathbb{R}$. Then both $A \setminus B \neq \varnothing$ and $B \setminus A \neq \varnothing$ hold which implies existence of $a \in A \setminus B$ and $b \in B \setminus A$. The total ordering of $\mathbb{Q}$ gives that one of $a = b$, $a < b$ or $b < a$ are TRUE. Equality is ruled out directly from $a \notin B$ and $b \in B$. If $a < b$ is TRUE, then $b \in B$ and C2 forces $a \in B$, a contradiction. The case $b < a$ is handled by symmetry. Hence, $\preccurlyeq$ is connex and the ordering is total. □

A key fact that makes the Dedekind construction work is the fact that the union of any non-empty set of cuts contained in a given cut is a cut:

Lemma 10.7 *We have*

$$\forall B \in \mathbb{R}\, \forall C \subseteq \mathbb{R}\colon\ \Big(C \neq \varnothing\ \wedge\ \forall A \in C\colon A \subseteq B\Big)\ \Rightarrow\ \bigcup C \in \mathbb{R} \tag{10.7}$$

Proof. We need to verify C1-C3 for $\bigcup C$. Starting with C1, note that $\bigcup C \subseteq B$ and so, by C1 for B, we have $\mathbb{Q} \setminus \bigcup C \neq \varnothing$. Since $C \neq \varnothing$ there is $A \in C$ and so $A \subseteq \bigcup C$. Hence, by C1 for A we get $\bigcup C \neq \varnothing$ showing that $\bigcup C$ obeys C1.

To prove C2 for $\bigcup C$ let $a \in \bigcup C$ and let $b \leqslant a$. Then there is $A \in C$ such that $a \in A$ and, by C2 for A, we have $b \in A$ and thus $b \in \bigcup C$, showing that $\bigcup C$ obeys C2. Similarly, if $a \in \bigcup C$, then for some $A \in C$ we have $a \in A$ and, by C3 for A, there is $b \in A$ with $a < b$. But then $b \in \bigcup C$ showing that $\bigcup C$ obeys C3 as well. □

Using this we now conclude that the ordering $\preccurlyeq$ has the least upper bound property, which is why we do the construction in the first place:

Lemma 10.8 *The following holds with supremum relative to the ordering* (10.5)*:*

$$\forall C \subseteq \mathbb{R}\colon \Big(C \neq \varnothing \ \wedge\ \big(\exists B \in \mathbb{R} \forall A \in C\colon A \subseteq B\big)\Big) \\ \Rightarrow\ \sup(C) \text{ exists } \wedge\ \sup(C) = \bigcup C \tag{10.8}$$

In words, each non-empty $C \subseteq \mathbb{R}$ that admits an upper bound admits a supremum.

Proof. Lemma 10.7 ensures that $\bigcup C \in \mathbb{R}$. Next we note that $A \subseteq \bigcup C$ for all $A \in C$ and so $\bigcup C$ is an upper bound on C. Finally, if D is another such upper bound, then $A \subseteq D$ holds for all $A \in C$ and so $\bigcup C \subseteq D$. This shows that $\bigcup C$ is the least upper bound on C and is thus the supremum of C. □

The reader should note that we have not used any of the field properties of the rationals. As a consequence, the above holds for any non-empty totally-ordered set.

10.3 Addition for cuts.

Next we move to the operation of addition on $\mathbb{R}$, to be denoted by symbol $\oplus$ in order to prevent confusion with the operation of addition in $\mathbb{Q}$. Given any sets $A, B \subseteq \mathbb{Q}$, define the following sets of rationals

$$\begin{aligned} A \oplus B &:= \big\{a + b\colon a \in A \wedge b \in B\big\} \\ \underline{0} &:= \{a \in \mathbb{Q}\colon a < 0\} \\ \ominus A &:= \big\{a \in \mathbb{Q}\colon (\exists b \in \mathbb{Q} \smallsetminus A\colon a + b < 0)\big\} \end{aligned} \tag{10.9}$$

The first point to check is that, for A and B cuts, the sets $A \oplus B$ and $\ominus A$ are also cuts. (For $\underline{0}$ this was done in (10.2).) We start with the former:

Lemma 10.9 $\forall A, B \in \mathbb{R}\colon A \oplus B \in \mathbb{R}$

Proof. From $A, B \neq \varnothing$ we get $A \oplus B \neq \varnothing$. On the other hand, if $a' \in \mathbb{Q} \smallsetminus A$ and $b' \in \mathbb{Q} \smallsetminus B$, then $a < a'$ for all $a \in A$ and $b < b'$ for all $b \in B$ by Lemma 10.5. It follows that $a' + b' \notin A \oplus B$ and so $A \oplus B \neq \mathbb{Q}$. We have proved C1 for $A \oplus B$.

For C2, if $a + b \in A \oplus B$ and $c \in \mathbb{Q}$ obeys $c \leqslant a + b$, then $a' := c - b$ obeys $a' \leqslant a$. Hence $a' \in A$ by C2 for A and so $c = a' + b \in A \oplus B$, proving C2 for $A \oplus B$. Concerning C3, if $a + b \in A \oplus B$ for some $a \in A$ and $b \in B$ then C2 for A implies existence of $a' \in A$ with $a < a'$. This gives $a' + b \in A \oplus B$ with $a + b < a' + b$, proving C3 for $A \oplus B$ as well. □

For Dedekind cuts representing rationals, the operation $\oplus$ acts as desired

$$\forall a, b \in \mathbb{Q}\colon\ \{a' \in \mathbb{Q}\colon a' < a\} \oplus \{b' \in \mathbb{Q}\colon b' < b\} = \{c \in \mathbb{Q}\colon c < a + b\} \tag{10.10}$$

whose proof we leave to the reader. The same applies to $\ominus$,

$$\forall c \in \mathbb{Q}\colon\ \ominus \{a \in Q\colon a < c\} = \{a \in \mathbb{Q}\colon a < -c\} \tag{10.11}$$

This is seen from the fact that the set on the left-hand side consists of all $a' \in Q$ such that $\forall b \geqslant c\colon a' + b < 0$, which is equivalent to $a' + c < 0$; i.e., $a' < -c$. For general cuts A we just verify that $\ominus A$ is a cut:

Lemma 10.10 $\forall A \in \mathbb{R}\colon \ominus A \in \mathbb{R}$

Proof. Let $A \in \mathbb{R}$. First note that, if $c \in A$ then for all $b \in \mathbb{Q} \setminus A$, Lemma 10.5(2) gives $b - c > 0$ showing that $-c \notin \ominus A$. This is worthy of noting separately,

$$\forall c \in \mathbb{Q}\colon\ c \in A \ \Rightarrow\ -c \notin \ominus A \tag{10.12}$$

From $A \neq \varnothing$, we then get $\ominus A \neq \mathbb{Q}$. Next, if $b \in \mathbb{Q} \setminus A$ and $d \in \mathbb{Q}$ obeys $d > b$, then $b + (-d) < 0$ and so $-d \in \ominus A$, thus showing C1 for $\ominus A$. Conditions C2 and C3 are then checked directly from the definition: if $a \in \mathbb{Q}$ is such that $a + b < 0$ for some $b \in \mathbb{Q} \setminus A$, then $a' + b < 0$ for all $a' \leqslant a$ and, by the Archimedean property, there is $a'' > a$ such that $a'' + b < 0$ still holds. It follows that $\ominus A \in \mathbb{R}$ as desired. □

A key point now is to verify that $\ominus A$ is the inverse element to A under $\oplus$:

Lemma 10.11 $\forall A \in \mathbb{R}\colon\ A \oplus (\ominus A) = \underline{0}$

Proof. We start by proving the inclusion $\subseteq$. For this note let $a \in A$ and $a' \in \ominus A$. Then there exists $b \in \mathbb{Q} \setminus A$ such that $a' + b < 0$. But this means that $a + a' < a - b < 0$, by Lemma 10.5(2). Hence $A \oplus (\ominus A) \subseteq \underline{0}$.

For the opposite inclusion $\supseteq$, let $c > 0$. Then (10.4) ensures existence of $a \in A$ and $b \in \mathbb{Q} \setminus A$ such that $b - a = c/2$. But then $b + (-a - c) = -c/2 < 0$ and so $-a - c \in \ominus A$. Hence $-c = a + (-a - c) \in A \oplus (\ominus A)$, thus proving $\underline{0} \subseteq A \oplus (\ominus A)$. □

Using these lemmas we conclude:

Corollary 10.12 *$\oplus$ is a commutative and associative binary operation (of addition) on $\mathbb{R}$ with $\underline{0}$ being the zero element, $\ominus A$ being the inverse element to A and $\ominus(\ominus A) = A$ valid for all $A \in \mathbb{R}$. In short, $(\mathbb{R}, \oplus)$ is a commutative group with unit element $\underline{0}$.*

Proof. Commutativity and associativity is checked from the definition of $A \oplus B$. That $A \oplus \underline{0} = A$ for each $A \in \mathbb{R}$ is checked directly from the definition of the cut. The inverse element property was proved in Lemma 10.10. Abbreviating $B := \ominus A$, the commutativity and associativity of addition along with Lemma 10.11 show

$$A = A \oplus (B \oplus (\ominus B)) = (A \oplus B) \oplus (\ominus B) = \ominus B \tag{10.13}$$

proving $\ominus(\ominus A) = A$ as desired. □

It remains to link addition to the ordering relation:

Lemma 10.13 $\forall A, B, C \in \mathbb{R}\colon\ A \leqslant B \Rightarrow A \oplus C \leqslant B \oplus C$

Proof. Let $A, B \in \mathbb{R}$. Then $A \leqslant B$ means $A \subseteq B$. For each $C \in \mathbb{R}$ the definition of addition gives $A \oplus C \subseteq B \oplus C$ and so $A \oplus C \leqslant B \oplus C$. □

Note that thanks to Lemma 10.11, this shows that $A \leqslant B$ is equivalent to $\ominus B \leqslant \ominus A$. (Note that by (10.11) we also get $\ominus\underline{0} = \underline{0}$.)

10.4 Multiplication for cuts.

The multiplication between cuts is defined similarly, albeit in two stages. Writing $A \prec B$ for $A \leqslant B \wedge A \neq B$ and abbreviating $\mathbb{R}^- := \{A \in \mathbb{R}\colon\ A \prec \underline{0}\}$, for $A, B \in \mathbb{R}^-$ we set

$$A \odot B := \ominus\big\{-a \cdot b\colon\ a \in A \wedge b \in B\big\} \tag{10.14}$$

Before we proceed to other cases, we check that the resulting object is a cut.

Lemma 10.14 $\forall A, B \in \mathbb{R}^- \colon A \odot B \in \mathbb{R}$

Proof. Given $A, B \in \mathbb{R}^-$, it suffices to show that $C := \{-a \cdot b \colon a \in A \wedge b \in B\}$ obeys $C \in \mathbb{R}$ as then $\ominus C \in \mathbb{R}$ by Lemma 10.10. Clearly $C \neq \varnothing$ because there is at least one $a \in A$ and one $b \in B$. Since $-a \cdot b < 0$ for all $a \in A$ and $b \in B$ we also have $C \neq \mathbb{Q}$, proving C1. For C2-C3, let $c \in C$ and write it as $c = -a \cdot b$ for $a \in A$ and $b \in B$. Then $c' < 0$ can be written as $c' = -a' \cdot b$ for $a' := (-b)^{-1} \cdot c'$. Since $(-b)^{-1} > 0$, for $c' \leqslant c$ we get $a' \leqslant a$ which by $a' \in A$ imply $c' \in C$ proving C2 for $A \odot B$. The existence of some $a' \in A$ with $a < a' < 0$ in turn gives existence of $c' > c$ with $c' \in C$, proving C3. □

We now complete the definition of $\odot$ by setting

$$A \odot B := \begin{cases} (\ominus A) \odot (\ominus B) & \text{if } \underline{0} \prec A \wedge \underline{0} \prec B, \\ \ominus\big((\ominus A) \odot B\big) & \text{if } \underline{0} \prec A \wedge B \prec \underline{0} \\ \ominus\big(A \odot (\ominus B)\big) & \text{if } A \prec \underline{0} \wedge \underline{0} \prec B \\ \underline{0} & \text{if } A = \underline{0} \vee B = \underline{0} \end{cases} \tag{10.15}$$

and observe that these ensure

$$\forall A, B \in \mathbb{R} \colon (\ominus A) \odot B = A \odot (\ominus B) = \ominus(A \odot B) \tag{10.16}$$

The operation $\odot$ acts on cuts representing rationals as expected,

$$\forall a, b \in \mathbb{Q} \colon \ \{a' \in \mathbb{Q} \colon a' < a\} \odot \{b' \in \mathbb{Q} \colon b' < b\} = \{c \in \mathbb{Q} \colon c < a \cdot b\} \tag{10.17}$$

which is checked directly for $a, b < 0$ and extended to the other cases with the help of (10.11). We also directly check that $A \odot \underline{0} = \underline{0}$.

To complete the definitions we need following objects:

$$\underline{1} := \{a \in \mathbb{Q} \colon a < 1\} \tag{10.18}$$

and, denoting $\mathbb{Q}^- := \{a \in \mathbb{Q} \colon a < 0\}$, setting

$$A^{-1} := \begin{cases} \big\{b \in \mathbb{Q}^- \colon (\exists a \in \mathbb{Q} \setminus A \colon a \cdot b > 1)\big\}, & \text{if } A \prec \underline{0} \\ \ominus(\ominus A)^{-1}, & \text{if } 0 \prec A \end{cases} \tag{10.19}$$

which for cuts representing rationals again acts as expected,

$$\forall a \in \mathbb{Q} \setminus \{0\} \colon \ \{b \in \mathbb{Q} \colon b < a\}^{-1} = \{b \in \mathbb{Q} \colon b < a^{-1}\} \tag{10.20}$$

Following similar arguments as for addition, we check:

Lemma 10.15 $\forall A \in \mathbb{R} \colon A \neq \underline{0} \Rightarrow A^{-1} \in \mathbb{R}$

Proof. It suffices to prove the claim for $A \in \mathbb{R}^-$. The fact that there is $a \in \mathbb{Q} \setminus A$ with $a < 0$ with the help of the Archimedean property implies existence of $n \in \mathbb{N}$ with $-n \cdot a > 1$. Hence, $A^{-1} \neq \varnothing$. Since $A^{-1} \subseteq \mathbb{Q}^-$, we get C1.

For C2-C3, let $b \in A^{-1}$. If $a \cdot b > 1$ for some $a \in \mathbb{Q} \setminus A$, then $b < 0$ forces $a < 0$. For any $b' \leqslant b$ we then have $a \cdot b' \geqslant a \cdot b > 1$, proving $b' \in A^{-1}$ and thus C2. Taking $c := (a \cdot b + 1)/2$ gives $1 < c < a \cdot b$ which for $b' := a^{-1} \cdot c$ shows $a \cdot b' = c > 1$ and so $b' \in A^{-1}$. Since $a^{-1} < 0$ implies $b = a^{-1} \cdot (a \cdot b) < a^{-1} \cdot c = b'$, we get C3 as well. □

Lemma 10.16 $\forall A \in \mathbb{R} \colon A \neq \underline{0} \Rightarrow A \odot A^{-1} = \underline{1}$

Proof. Again, by (10.16), we just deal with $A \in \mathbb{R}^-$. Denote $C := \{-a \cdot b\colon\ a \in A \wedge b \in A^{-1}\}$ and observe that $b \in A^{-1}$ implies existence of $a' \in \mathbb{Q} \setminus A$ such that $a' \cdot b > 1$. The fact that $b < 0$ and $a < a'$ for all $a \in A$ give $a \cdot b > a' \cdot b > 1$, i.e., $\forall a \in A \forall b \in A^{-1}\colon\ -a \cdot b < -1$. This shows $C \subseteq \ominus\underline{1}$.

For $\supseteq$, suppose that $-c \notin C$ for some $1 < c < 3/2$. Then we have $\forall a \in A\colon a^{-1} \cdot c \notin A^{-1}$ which means $\forall a \in A \forall b \in \mathbb{Q} \setminus A\colon a^{-1} \cdot c \cdot b \leqslant 1$. As $a < 0$, the latter inequality is equivalent to $b \cdot c \geqslant a$ which is the same as $(-b)(c-1) \leqslant b - a$. Pick $b' < 0$ with $b' \in \mathbb{Q} \setminus A$. By (10.4), there are $a \in A$ and $b \in \mathbb{Q} \setminus A$ with $b - a = (-b'/2)(c-1)$. Note that by $c - 1 \leqslant 1/2$ and $a \leqslant b'$ this entails $b \leqslant (-b'/2)(c-1) + b' \leqslant 3b'/4$ and so

$$\frac{3}{4}(-b')(c-1) \leqslant (-b)(c-1) \leqslant b - a = \frac{1}{2}(-b')(c-1) \tag{10.21}$$

which is (in light of $c - 1 > 0$ and $-b' > 0$) is FALSE. Hence no such c exists. As C is a cut, we get $\ominus\underline{1} \subseteq C$, proving the claim. □

Lemma 10.17 *$\odot$ is commutative, associative and distributive about $\oplus$. In addition, $\underline{1}$ is the unit element and A^{-1} is the inverse element for each $A \in \mathbb{R}$ with $A \neq \underline{0}$. In short,*

$$(\mathbb{R}, \oplus, \underline{0}, \odot, \underline{1}) \quad \text{is a field} \tag{10.22}$$

Proof. The commutativity is clear from the definition. For associativity, let $A, B \in \mathbb{R}^-$ and note that (as shown in the proof of Lemma 10.14), $C := \{a \cdot b\colon a \in A \wedge\ b \in B\}$ equals $\ominus(A \odot B)$. For $D \in \mathbb{R}^-$ the fact that $C < \underline{0}$ gives

$$(A \odot B) \odot D = \ominus(C \odot D) = \ominus\big\{(a \cdot b) \cdot d\colon\ a \in A\ \wedge\ b \in B\ \wedge\ d \in D\big\} \tag{10.23}$$

Invoking $(a \cdot b) \cdot d = a \cdot (b \cdot d)$, the same argument equates this with $A \odot (B \odot D)$. The case of general $A, B, C \in \mathbb{R}^-$ then follow via (10.16).

The distributive law is proved similarly. First we ask the reader to check the fact that if the distributive law holds for any negative elements, it holds for all elements. So we will only focus on $A, B, C \in \mathbb{R}^-$. Then

$$(A \oplus B) \odot C = \ominus\big\{-(a+b) \cdot c\colon\ a \in A\ \wedge\ b \in B\ \wedge\ c \in C\big\} \tag{10.24}$$

Using that $(a+b) \cdot c = a \cdot c + b \cdot c$ we write the set on the right as $\ominus(A \odot C) \oplus (\ominus(B \odot C))$ and so the whole expression equals $(A \odot C) \oplus (B \odot C)$. The claims about the multiplicative inverse follow from Lemmas 10.15-10.16. □

As our final step, we check the behavior of the ordering under multiplication:

Lemma 10.18 $\forall A, B, C \in \mathbb{R}\colon\ (A \leqslant B \wedge \underline{0} \leqslant C) \Rightarrow A \odot C \leqslant B \odot C.$

Proof. Lemma 10.13 and the distributive law allow us to restrict to $A, B \in \mathbb{R}^-$. The case of $C = \underline{0}$ is trivial so we may assume $C > \underline{0}$. Then $\ominus C \in \mathbb{R}^-$ and, invoking $A \leqslant B$, we get

$$A \odot C = \big\{-a \cdot c\colon a \in A\ \wedge\ c \in \ominus C\big\} \subseteq \big\{-b \cdot c\colon b \in B\ \wedge\ c \in \ominus C\big\} = B \odot \ominus C \tag{10.25}$$

thus proving the claim. □

We are now ready to give:

Proof of Theorem 10.3. Combining of Lemma 10.6, Corollary 10.12, Lemma 10.13 and Lemmas 10.14-10.18, $(\mathbb{R}, \oplus, \underline{0}, \odot, \underline{1}, \preccurlyeq)$ is an ordered field. The ordering is complete by Lemma 10.8 and so $(\mathbb{R}, \oplus, \underline{0}, \odot, \underline{1}, \preccurlyeq)$ is a system of reals. □

Another way to construct a system of reals follows an argument of G. Cantor which is in fact more general than the problem itself. The argument goes by interpreting $\mathbb{Q}$ as a metric space endowed with the metric $\varrho(a, b) := |a - b|$. The set of reals is the identified with equivalence classes of Cauchy sequences in $(\mathbb{Q}, \varrho)$. We will return to this argument when we discuss metric spaces in detail.

10.5 Uniqueness.

Having established existence, we now move to the statement and proof of uniqueness.

Theorem 10.19 (Uniqueness) *Let $(\mathbb{R}, \oplus, \underline{0}, \odot, \underline{1}, \preccurlyeq)$ be as above. For any complete ordered field $(F, +, 0, \cdot, 1, \leqslant)$, there is a bijection $\phi\colon \mathbb{R} \to F$ which is an order-preserving (field) isomorphism. (The latter means that ϕ obeys properties (1-4) in Theorem 7.8).*

Proof (main steps). Let $(F, +, 0, \cdot, 1, \leqslant)$ be complete ordered field. First we identify elements of F with Dedekind cuts. For this let $\mathbb{N}_F$ be the naturals of F (see (7.9)) and write

$$\mathbb{Q}_F := \left\{ r^{-1} \cdot (m - n) \colon m, n, r \in \mathbb{N}_F \wedge r \neq 0 \right\} \tag{10.26}$$

for the *rationals of F*. Following the proof of Theorem 10.3, we now construct a complete ordered field $(\mathbb{R}_F, \oplus_F, 0_F, \odot_F, 1_F, \preccurlyeq_F)$ of Dedekind cuts based on rationals $\mathbb{Q}_F$.

Next observe that, since F has the least upper bound property the cuts in $\mathbb{R}_F$ do admit a universal representation:

Lemma 10.20 *We have:*

$$\forall A \in \mathbb{R}_F\colon \ \ \sup(A) \text{ exists } \wedge\ A = \left\{ a \in \mathbb{Q}_F \colon a < \sup(A) \right\} \tag{10.27}$$

In particular, $\sup\colon \mathbb{R}_F \to F$ *is a bijection.*

Now check that, for all $A, B \in \mathbb{R}_F$,

$$\begin{aligned} \sup(A \oplus_F B) &= \sup(A) + \sup(B) \\ \sup(A \odot_F B) &= \sup(A) \cdot \sup(B) \\ \sup(\ominus_F A) &= -\sup(A) \end{aligned} \tag{10.28}$$

and, if $A \neq 0_F$, also

$$\sup(A^{-1}) = \sup(A)^{-1} \tag{10.29}$$

Noting that $\sup(0_F) = 0$ and $\sup(1_F) = 1$, and that $A \preccurlyeq_F B$ is equivalent to $\sup(A) \leqslant \sup(B)$, the map $\sup\colon \mathbb{R}_F \to \mathbb{R}$ is an order preserving isomorphism.

We now recall that Theorem 7.8 asserts the existence of a bijection $\psi\colon \mathbb{Q} \to \mathbb{Q}_F$ which is an order-preserving isomorphism. The image map associated with ψ acting as

$$\psi(A) := \{\psi(x) \in F \colon x \in A\} \tag{10.30}$$

then maps cuts over $\mathbb{Q}$ to those over $\mathbb{Q}_F$ and thus defines a bijection $\psi\colon \mathbb{R} \to \mathbb{R}_F$. Since all arithmetic operations on cuts are defined the same way in $\mathbb{R}$ as in $\mathbb{R}_F$, we get that

that ψ is an order-preserving isomorphism. Setting

$$\phi := \sup \circ \psi \tag{10.31}$$

we get the map in the claim. □

A take-away message of this section is that the reals exist as a complete ordered field and they are unique up to an order-preserving field-isomorphism. The latter implies that there is only one real analysis one can build out of Zermelo's axioms.

11. PROPERTIES OF THE REALS

We proceed to discuss some standard consequences of the construction of the reals. Among these are the definition of some basic functions; namely, roots, exponentials and logs. We also mention extensions of the reals to other fields; namely, complex numbers and hyperreals. Throughout we assume that a complete ordered field $(\mathbb{R}, +, 0, \cdot, 1, \leqslant)$ is given and use $\mathbb{N}$, resp., $\mathbb{Q}$ for the associated sets of naturals, resp., rationals in $\mathbb{R}$.

11.1 Archimedean property and density of (ir)rationals.

As our first consequence of the completeness of the reals, we extend the Archimedean property of the rationals to the reals. Note that the proofs of these are very different.

Lemma 11.1 (Archimedean property of $\mathbb{R}$) $\quad \forall x \in \mathbb{R} \colon x > 0 \Rightarrow (\exists n \in \mathbb{N} \colon x \cdot n > 1)$

Proof. Let $x > 0$ and suppose that $x \cdot n \leqslant 1$ for all $n \in \mathbb{N}$. By the properties of the field, this means that $\mathbb{N}$ is bounded by x^{-1} and so $\sup(\mathbb{N})$ exists by the least upper bound property. Then

$$\exists n \in \mathbb{N} \colon n > \sup(\mathbb{N}) - 1 \tag{11.1}$$

because otherwise $\mathbb{N}$ is bounded by $\sup(\mathbb{N}) - 1$ and so $\sup(\mathbb{N})$ is not the least upper bound. However, with n such that the statement in (11.1) holds, we have $\sup(\mathbb{N}) < n + 1$ which, as $n + 1 \in \mathbb{N}$, shows that $\sup(\mathbb{N})$ is not an upper bound, a contradiction. □

The Archimedean property serves as a useful tool in proofs. One of its useful consequence is the fact that, if y is smaller than x plus any "arbitrarily small" positive rational, then, necessarily, $y \leqslant x$:

Corollary 11.2 *We have*

$$\forall x, y \in \mathbb{R} \colon \left(\forall n \in \mathbb{N} \colon x \leqslant y + \frac{1}{n+1} \right) \Rightarrow x \leqslant y \tag{11.2}$$

Proof. We will prove the contrapositive. Suppose $x > y$. By the Archimedean property, there exists $m \in \mathbb{N}$ such that $m(x - y) > 1$. This forces $m \neq 0$ and so $m = n + 1$ for some $n \in \mathbb{N}$. The inequality $m(x - y) > 1$ then reads as $x > y + \frac{1}{n+1}$ contradicting the clause on the left of $\Rightarrow$. □

A closely related consequence is the fact that rationals are spread "densely" in $\mathbb{R}$:

Lemma 11.3 (Density of rationals in $\mathbb{R}$)

$$\forall x, y \in \mathbb{R} \colon \; x < y \; \Rightarrow \; (\exists a \in \mathbb{Q} \colon x < a \wedge a < y) \tag{11.3}$$

Proof. We may suppose that $y > 0$ for otherwise we just replace y by $-x$ and x by $-y$ and then apply a sign change at the very end. The intuitive idea of the proof is simple: Multiplying x and y by a large natural n, we can arrange that $ny - nx > 1$. Since consecutive naturals are just a unit apart, this forces a natural lie between ny and nx.

A formal proof is a tiny bit more complicated. First, the Archimedean property tells us that there is $n \in \mathbb{N}$ such that $n(y - x) > 1$. As $yn > 0$, the Archimedean property also tells us that $k((yn)^{-1} - 0) > 1$ for some $k \in \mathbb{N}$. This translates into $A := \{k \in \mathbb{N} \colon k \geqslant yn\}$ being non-empty and so, by Lemma 9.10, $m := \inf(A)$ exists and obeys $m \in A$. But the

latter forces $m-1<ny$ while the above shows

$$m-1\geqslant ny-1>ny-n(y-x)=nx \tag{11.4}$$

implying $m-1>nx$. Since $m\geqslant yn$ by $m\in A$ and $n>0$, dividing the above by n we get

$$x<\frac{m-1}{n}<\frac{m}{n}\leqslant y \tag{11.5}$$

proving that $x<\frac{m-1}{n}<y$. Noting that $\frac{m-1}{n}\in\mathbb{Q}$, the claim is proved. □

The same actually applies to *irrationals*, which are those reals that are not rational. However, the proof (which we leave to a homework exercise) can be done solely based on the algebraic properties of the reals:

Lemma 11.4 (Density of irrationals in $\mathbb{R}$)

$$\forall x,y\in\mathbb{R}\colon\ x<y\ \Rightarrow\ (\exists a\in\mathbb{R}\smallsetminus\mathbb{Q}\colon x<a\wedge a<y) \tag{11.6}$$

It should be noted that the notion of "being dense" will be given another meaning once we discuss topological aspects of the reals. These considerations also drive Cantor's proof of existence of the reals.

11.2 Arbitrary roots.

Our motivation for considering the reals was to fix the algebraic issues we had with the rationals; namely, the fact that polynomial equations $x^n=a$ generally do not admit solutions over the rationals. To see that this is fixed in the reals, we prove:

Theorem 11.5 (Arbitrary roots) *We have*

$$\forall n\in\mathbb{N}\smallsetminus\{0\}\,\forall a\in\mathbb{R}\colon a\geqslant 0\ \Rightarrow\ \exists x\in\mathbb{R}\colon x\geqslant 0\ \wedge\ x^n=a \tag{11.7}$$

and

$$\forall n\in\mathbb{N}\smallsetminus\{0\}\forall x,y\in\mathbb{R}\colon x\geqslant 0\ \wedge\ y\geqslant 0\ \wedge x^n=y^n\ \Rightarrow\ x=y \tag{11.8}$$

or, to put this in words, for each real $a\geqslant 0$ and each natural $n\geqslant 1$ there exists a unique real $x\geqslant 0$ such that $x^n=a$.

The proof will use the following identity which is checked (not by induction but) by invoking the distributive law and elementary manipulations with sums:

$$\forall x,y\in\mathbb{R}\,\forall n\in\mathbb{N}\smallsetminus\{0\}\colon\ y^n-x^n=(y-x)\sum_{k=0}^{n-1}x^k y^{n-k-1} \tag{11.9}$$

Here and henceforth, the symbol $\sum_{k=0}^{m}c_k$ is constructed recursively so that

$$\sum_{k=0}^{0}c_k=c_0\quad\wedge\quad\forall m\in\mathbb{N}\colon\ \sum_{k=0}^{m+1}c_k=c_{m+1}+\sum_{k=0}^{m}c_k \tag{11.10}$$

With this we are ready for:

Proof of Theorem 11.5. We start with the existence part (11.7). Fix a real $a\geqslant 0$ and a natural $n\geqslant 1$ and let

$$A:=\{y\geqslant 0\colon y^n\leqslant a\} \tag{11.11}$$

Then $0 \in A$, so $A \neq \varnothing$. Also the fact that $1 + a \geqslant 1$ gives $(1+a)^n \geqslant 1 + a > a$ and, since $1 + a < x$ implies $a < (1+a)^n \leqslant x^n$, the number $1 + a$ bounds all elements of A from above. As A is non-empty and admits an upper bound, its supremum $\sup(A)$ exists. We will now show that $\sup(A)^n = a$.

The fact that $\sup(A)$ is the least upper bound on A means that, for each $m \in \mathbb{N}$, $\sup(A) - \frac{1}{m+1}$ is no-longer an upper bound. Hence, for each $m \in \mathbb{N}$ there exists $y \in A$ with $y > \sup(A) - \frac{1}{m+1}$ yet $y \leqslant \sup(A)$. For such m and y the identity (11.9) yields

$$\begin{aligned} \sup(A)^n - a \leqslant \sup(A)^n - y^n &= (\sup(A) - y) \sum_{k=0}^{n-1} \sup(A)^k y^{n-k-1} \\ &\leqslant (\sup(A) - y) n \sup(A)^{n-1} \end{aligned} \tag{11.12}$$

where $y^n \leqslant a$ and $y \leqslant \sup(A)$ were used for the inequalities. Invoking $\sup(A) - y \leqslant \frac{1}{m+1}$ this translates into

$$\forall m \in \mathbb{N} \colon (m+1)(\sup(A)^n - a) \leqslant n \sup(A)^{n-1} \tag{11.13}$$

If $\sup(A) > 0$ we can divide the inequality by the right-hand side. The Archimedean principle then forces

$$\sup(A)^n \leqslant a \tag{11.14}$$

which is then checked to be true directly if $\sup(A) = 0$ as well.

To prove that equality holds, note that $\sup(A) + \frac{1}{m+1} \in A$ would imply that $\sup(A)$ is not an upper bound on A. So we must have

$$\forall m \in \mathbb{N} \colon \Big(\sup(A) + \frac{1}{m+1}\Big)^n > a \tag{11.15}$$

But then (11.9) tells us that, for all $m \in \mathbb{N}$,

$$\begin{aligned} a - \sup(A)^n &< \Big(\sup(A) + \frac{1}{m+1}\Big)^n - \sup(A)^n \\ &= \frac{1}{m+1} \sum_{k=0}^{n-1} \Big(\sup(A) + \frac{1}{m+1}\Big)^k \sup(A)^{n-k-1} \\ &\leqslant \frac{1}{m+1} n \big(1 + \sup(A)\big)^{n-1} \end{aligned} \tag{11.16}$$

which translates into

$$\forall m \in \mathbb{N} \colon \ (m+1)\big(a - \sup(A)^n\big) \leqslant n\big(1 + \sup(A)\big)^{n-1} \tag{11.17}$$

The Archimedean principle gives that $a - \sup(A)^n \leqslant 0$; i.e., $\sup(A)^n \geqslant a$. Combining with (11.14) we get $\sup(A)^n = a$, thus proving (11.7).

The uniqueness of the solution to $x^n = a$ comes from the fact that if y is another real, then $x < y$ implies $x^n < y^n$ while $y < x$ implies $y^n < x^n$. □

We will henceforth use notations

$$\sqrt[n]{a} \quad \text{or} \quad a^{1/n} \tag{11.18}$$

for the unique non-negative solution of $x^n = a$ for $a \geqslant 0$. Writing

$$\mathbb{R}^+ := \{a \in \mathbb{R} \colon 0 \leqslant a\} \tag{11.19}$$

the proof of Theorem 11.5 in fact gives:

Corollary 11.6 (n-th root function) *For each $n \in \mathbb{N} \smallsetminus \{0\}$ there exists a unique function $f\colon \mathbb{R}^+ \to \mathbb{R}^+$ such that*

$$\mathrm{Dom}(f) = \mathbb{R}^+ \ \wedge\ \forall a \in \mathbb{R}^+\colon f(a) = a^{1/n} \tag{11.20}$$

Proof. Define the binary relation

$$F := \Big\{(a,b) \in \mathbb{R}^+ \times \mathbb{R}^+ \colon b = \sup\{y \geqslant 0 \colon y^2 \leqslant a\}\Big\} \tag{11.21}$$

where $b = \sup\{y \geqslant 0 \colon y^2 \leqslant a\}$ abbreviates the logical sentence

$$\begin{aligned}\Big(\forall x \in \mathbb{R}^+ \colon x^n \leqslant a \ \Rightarrow\ x \leqslant b\Big)& \\ \wedge\ \forall y \in \mathbb{R}^+ \colon \Big(\forall x \in \mathbb{R}^+ \colon x^n \leqslant a \ \Rightarrow\ x \leqslant y\Big) \ \Rightarrow\ b \leqslant y&\end{aligned} \tag{11.22}$$

The definition ensures that F is a graph of a function f which, by the proof of Theorem 11.5, is defined for all $a \geqslant 0$ and obeys $f(a)^n = a$. □

11.3 Exponentials and logs.

Recall that integer powers are defined from natural powers via $a^{-n} := (a^{-1})^n$ whenever $a > 0$ and $n \in \mathbb{N}$. The natural roots then allow us to define symbols of the form $(a^p)^{1/q}$ and $(a^{1/q})^p$ for any $p, q \in \mathbb{Z}$ with $q \neq 0$. Not surprisingly, these quantities only depend on the value of p/q. This leads to:

Theorem 11.7 (Arbitrary powers) *For all $a > 0$ and all $p, q, p', q' \in \mathbb{Z}$ with $q, q' \neq 0$,*

$$(a^p)^{1/q} = (a^{1/q})^p \tag{11.23}$$

and, assuming that $\frac{p}{q} = \frac{p'}{q'}$, also

$$(a^p)^{1/q} = (a^{p'})^{1/q'} \tag{11.24}$$

Denoting by $a^{p/q}$ the common value of these quantities, for each $x \in \mathbb{R}$ we then set

$$a^x := \begin{cases} \sup\{a^z \in \mathbb{R} \colon z \in \mathbb{Q} \ \wedge\ z \leqslant x\}, & \text{if } a > 1, \\ \inf\{a^z \in \mathbb{R} \colon z \in \mathbb{Q} \ \wedge\ z \leqslant x\}, & \text{if } a < 1, \end{cases} \tag{11.25}$$

and put $1^x := 1$. We then have

$$\forall a > 0\, \forall x, y \in \mathbb{R} \colon a^{x+y} = a^x \cdot a^y \ \wedge\ a^{x \cdot y} = (a^x)^y = (a^y)^x \tag{11.26}$$

Also, $\forall a > 0\, \forall x \in \mathbb{R} \colon a^x > 0$.

We relegate the proof to a homework exercise. As for the roots, we get:

Corollary 11.8 (Exponential function of base a) *For each $a > 0$ there exists a unique function $f\colon \mathbb{R} \to \mathbb{R}^+$ with $\mathrm{Dom}(f) = \mathbb{R}$ such that*

$$\forall x \in \mathbb{R} \colon\ f(x) = a^x \tag{11.27}$$

where a^x is as in (11.25).

Proof (idea). We proceed as in Corollary 11.6 to define the graph of f; all needs to be done is to write the suprema/infima in (11.25) in terms of logical propositions. We leave the details to the reader. □

As it turns out, the exponential function $x \mapsto a^x$ with any base $a \neq 0, 1$ is strictly monotone and thus injective:

Lemma 11.9 *Let $a > 0$ obey $a \neq 1$. Then*

$$\forall x, y \in \mathbb{R}\colon\ x < y \Rightarrow \begin{cases} a^x < a^y, & \text{if } a > 1 \\ a^y < a^x, & \text{if } a < 1 \end{cases} \tag{11.28}$$

In particular, $x \mapsto a^x$ is injective.

Another fact is that a^x cannot "skip" values:

Lemma 11.10 *For all $a > 0$ with $a \neq 1$ and all $y \in \mathbb{R}$:*

$$y > 0 \Rightarrow \exists x \in \mathbb{R}\colon a^x = y \tag{11.29}$$

In particular, $\operatorname{Ran}(x \mapsto a^x) = \{y \in \mathbb{R}\colon y > 0\}$.

These lemmas, whose proof we relegate to homework, imply that the exponential function with base $a \neq 0, 1$ is a bijection of $\mathbb{R}$ onto the set of strictly positive reals. This allows us to conclude:

Theorem 11.11 (Logarithm) *Let $a > 0$ obey $a \neq 1$. Then there exists a unique function* $\log_a\colon \mathbb{R} \to \mathbb{R}_+$ *such that* $\operatorname{Dom}(\log_a) = \{x \in \mathbb{R}\colon x > 0\}$ *and*

$$\forall x > 0\colon\ a^{\log_a(x)} = x \tag{11.30}$$

and

$$\forall y \in \mathbb{R}\colon\ \log_a(a^y) = y \tag{11.31}$$

We call $\log_a$ the *logarithm of base a*. The properties in (11.26) then readily give:

Lemma 11.12 *Let $a > 0$ be such that $a \neq 1$. Then*

$$\forall x, y > 0\colon\ \log_a(x \cdot y) = \log_a(x) + \log_a(y) \tag{11.32}$$

and

$$\forall x > 0\, \forall y \in \mathbb{R}\colon\ \log_a(x^y) = y \log_a(x) \tag{11.33}$$

We leave the above claims to a homework exercise. We also note that both the exponential and logarithm functions can and will later be constructed again using methods of calculus. Naturally, these will give us the same objects as above.

11.4 Beyond the reals.

As we noted after the proof of Theorem 10.19, the fact that there is only one system of reals modulo order-preserving isomorphism implies that there is only one real analysis one can build out of Zermelo's axioms. Still, the reader might wonder whether other natural fields exist that are of significance for analysis.

Starting with a somewhat esoteric topic, we first note that non-Archimedean extensions of $\mathbb{R}$ exist that are ordered fields; e.g., the so called *hyperreal numbers* or *surreal*

numbers. The main feature of these is that, besides $\mathbb{R}$, they include infinitesimals (i.e., numbers those whose absolute value is smaller than any positive number) and also infinities (which are numbers that are arbitrarily large). The infinitesimals pile up arbitrarily close on both sides of every real which causes these fields fail the least upper bound property (for otherwise they would be the same as $\mathbb{R}$) which has its disadvantages. But we can then talk about "infinitesimal increments" and other things that in $\mathbb{R}$ have to be dealt with via approximations. A variant of analysis, called the *non-standard analysis*, is based on these fields instead of $\mathbb{R}$.

A very important extension is that to *complex numbers* discovered by G. Cardano in his solution of the cubics. Formally, the complex numbers are defined as a linear vector space over $\mathbb{R}$ with basis $\{1, \mathrm{i}\}$, i.e.,

$$\mathbb{C} := \{x + \mathrm{i}y \colon x, y \in \mathbb{R}\} \tag{11.34}$$

with the multiplication extended to the *imaginary number* i via the rule

$$\mathrm{i} \cdot \mathrm{i} = -1 \tag{11.35}$$

Explicitly, we have

$$(x + \mathrm{i}y) \cdot (\tilde{x} + \mathrm{i}\tilde{y}) = (x \cdot \tilde{x} - y \cdot \tilde{y}) + \mathrm{i}(x \cdot \tilde{y} + \tilde{x} \cdot y) \tag{11.36}$$

It is readily checked that this operation is commutative and associative and distributes around addition with number $1 = 1 + \mathrm{i}0$ acting as the unit element. With the multiplicative inverse defined as

$$(x + \mathrm{i}y)^{-1} := \frac{x - \mathrm{i}y}{x^2 + y^2} \tag{11.37}$$

which is meaningful once $x^2 + y^2 \neq 0$ (which means $x + \mathrm{i}y \neq 0$), the object $(\mathbb{C}, +, 0, \cdot, 1)$ becomes a field.

The field of complex numbers is an extension of $\mathbb{R}$ with the property — called the *Fundamental Theorem of Algebra* — that all polynomials, even those with coefficients in $\mathbb{C}$, have a root in $\mathbb{C}$ and, consequently, thus factor completely into a product of monomials. This means that $\mathbb{C}$ is *algebraically closed*. A deficiency of $\mathbb{C}$ over $\mathbb{R}$ is that (as you have been asked to show in homework) it does not admit an ordering that would make it an ordered field (in the sense of Definition 7.2).

Another extension of the reals are the *Euclidean spaces* $\mathbb{R}^n$, which are simply sets of n-tuples of real numbers. These are all linear vector spaces over $\mathbb{R}$ (with addition and scalar multiplication defined coordinate-wise). Unfortunately, with the exception of $\mathbb{R}^2$, which is isomorphic to $\mathbb{C}$, defining multiplication to make these fields is impossible. Multiplication can be defined in $\mathbb{R}^4$, although its no longer commutative, which leads to the notion of *quaternions*. In $\mathbb{R}^8$ one can define multiplication as well, but it is neither commutative nor associative, which leads to the notion of *octonions*.

12. CARDINALITY AND COUNTABILITY

Before we start discussing actual real analysis, we have one more topic to talk about: the meaning of the "size" of a set. While the notion of a "size" is really the subject of so called measure theory, its limited version in set theory is extremely important for the foundations as well.

12.1 Finite and infinite sets.

One way to interpret the loose notion of the "size" of a set is as the number of its elements. In daily life this is something that would be decided by counting which is nothing else but a labeling of the elements of the set by naturals. To make this mathematically precise, we need to introduce the concept of the set of "first n naturals"

$$[0, n) := \{k \in \mathbb{N} \colon k < n\} \tag{12.1}$$

where $k < n := k \leqslant n \wedge k \neq n$ for $\leqslant$ defined as $m \leqslant n := \exists k \in \mathbb{N} \colon n = m + k$. With this, we first divide all sets into two basic categories:

Definition 12.1 *A set A is said to be*

- finite *if there exist $n \in \mathbb{N}$ and a bijection $f \colon [0, n) \to A$, and*
- infinite *if it is not finite.*

Note that, since $[0, 0) = \varnothing$, a bijective map $f \colon [0, 0) \to \varnothing$ exists trivially and $\varnothing$ is thus finite. Concerning non-empty finite sets, recall a lemma from homework:

Lemma 12.2 *Let $m, n \in \mathbb{N}$ and let $f \colon [0, n) \to [0, m)$ be a function. Then*

(1) f injective $\Rightarrow\ n \leqslant m$
(2) f surjective $\Rightarrow\ m \leqslant n$
(3) f bijective $\Rightarrow\ m = n$

Noting that the inverse of a bijection as well as the composition of two bijections are bijections, part (3) shows that for each set A there is at most one $n \in \mathbb{N}$ for which a bijection $f \colon [0, n) \to A$ exists. It is thus meaningful to state:

Definition 12.3 *Let A be a finite set. The unique $n \in \mathbb{N}$ for which there is a bijection $f \colon [0, n) \to A$ is called the* cardinality *of A with notation $|A|$ (or sometimes $\#A$).*

We leave to the reader to verify that the concept of finite set is closed under finite unions and subset relation:

Lemma 12.4 *For any sets A and B, we have:*

(1) A finite $\wedge\ B \subseteq A\ \Rightarrow\ B$ finite
(2) A finite $\wedge\ B$ finite $\Leftrightarrow A \cup B$ finite
(3) A finite $\wedge\ B$ finite $\Leftrightarrow A \times B$ finite

These also translate into inequalities for cardinality:

Lemma 12.5 *For any finite sets A and B:*

$$B \subseteq A\ \Rightarrow\ |B| \leqslant |A| \tag{12.2}$$

and

$$|A \cup B| \leqslant |A| + |B| \tag{12.3}$$

and

$$|A \times B| = |A| \cdot |B| \tag{12.4}$$

Thus, intuitively, union translates (as a bound) into addition and Cartesian product into multiplication. The inequality in (12.3) arises from possible overcounting. Indeed, equality holds if (and only if) A and B are disjoint.

Returning to the concept of finite and infinite sets, we recall that earlier (specifically, in Definition 3.12) we put forward the notion of a set A being *Dedekind infinite* to denote the situation that there exists an injection $f\colon A \to A$ with $\mathrm{Ran}(f) \neq A$. This concept draws on (what is known as) *Galileo's paradox* which points out an apparent contradiction between the fact that only some naturals are squares and so there seems to be more naturals than squares and the existence of the map $f\colon \mathbb{N} \to \mathbb{N}$ defined by $f(n) := n^2$ that puts all squares with all naturals in one-to-one correspondence.

The notion of being Dedekind infinite is attractive for its intrinsic nature; indeed, it does not require anything else than the set itself. The negation of this concept is being *Dedekind finite* which means that every injection of the set into itself is necessarily a surjection. Lemma 12.2 then shows

$$A \text{ finite } \Rightarrow \ A \text{ Dedekind finite} \tag{12.5}$$

and, by contrapositive,

$$A \text{ Dedekind infinite } \Rightarrow \ A \text{ infinite} \tag{12.6}$$

Proving the converse to these implications seems intuitive as well; indeed, for (12.6) we keep listing elements of A to produce a sequence $\{x_n\colon n \in \mathbb{N}\}$ of distinct members. Then define $f(x_n) := x_{n+1}$ for $n \in \mathbb{N}$ and $f(x) := x$ for x not a member of the sequence $\{x_n\colon n \in \mathbb{N}\}$ to get an injection of A into itself which misses x_0.

Unfortunately, this is where a precise argument runs into a problem: the need to repeatedly choose an element from a set cannot be formalized without the use of Axiom of Choice. This is not a mere technicality; indeed, there are models of set theory with Zermelo's axioms but without Axiom of Choice in which the converse to (12.6) fails. For this reason we abandon all use of the notion of Dedekind-infinite sets and stick henceforth with the notions put foward in Definition 12.1.

12.2 Countable sets.

Having elucidated how to interpret the "size" of finite sets, let us move to refining the notion of infinite sets. In analogy with Definition 12.1, here we put:

Definition 12.6 *An infinite set A is said to be*

- countable *if there exists a bijection $f\colon \mathbb{N} \to A$, and*
- uncountable *if it is not countable.*

Informally, a set is countable if it can be enumerated into a sequence. While this notion is defined for infinite sets above, it is commonly extended to include finite sets as well. This is not true about Rudin's book, which talks about being *at most countable* in this case.

In other texts, the terms *denumerable* or *countably infinite* are used to denote "infinite and countable."

We will focus on countable sets first and derive some basic facts about them. Our first observation is that, just as for finite sets, countability is inherited by the subsets:

Lemma 12.7 *Let A be a countable set. Then*

$$\forall B \subseteq A\colon \quad B \text{ infinite } \Rightarrow \ B \text{ countable} \tag{12.7}$$

In particular, every subset of a countable set is either finite or countable.

Proof. Since B is infinite, $B \subseteq A$ implies that A is infinite. With A being countable, there exists a bijection $f\colon \mathbb{N} \to A$. We can then define $\widetilde{B} := f^{-1}(B)$ which, effectively, reduces to prove to subsets of $\mathbb{N}$.

At the informal level, our argument will use the ordering of $\mathbb{N}$ to enumerate all elements of $\widetilde{B}$ into a sequence. The formal construction requires the use of the recursion principle, which allows us to define sets $\{B_k\colon k \in \mathbb{N}\}$ such that

$$B_0 = \widetilde{B} \ \wedge \ \forall k \in \mathbb{N}\colon B_{k+1} = \begin{cases} \varnothing & \text{if } B_k = \varnothing \\ B_k \smallsetminus \{\inf(B_k)\} & \text{if } B_k \neq \varnothing, \end{cases} \tag{12.8}$$

where we recalled that, as shown in Lemma 9.10, each non-empty set of naturals has an infimum which is then contained therein.

We now claim that

$$\forall k \in \mathbb{N}\colon \ B_k \text{ infinite} \tag{12.9}$$

We prove this by induction: The case $k = 0$ is checked from $B_0 = \widetilde{B}$ being infinite. For the induction step, assume that B_k is infinite for some $k \in \mathbb{N}$. Then $B_k \neq \varnothing$ and so $B_k = B_{k+1} \cup \{\inf(B_k)\}$. If B_{k+1} were finite, then so would B_k by Lemma 12.4, which gives that B_{k+1} is infinite as well.

Next we claim

$$\forall k \in \mathbb{N}\colon \quad \inf(B_k) + 1 \leqslant \inf(B_{k+1}) \ \wedge \ k \leqslant \inf(B_k) \tag{12.10}$$

The first part is proved directly from $\inf(B_k) \notin B_{k+1}$ which forces $\inf(B_k) < \inf(B_{k+1})$. The second part is proved by induction; indeed, $0 \leqslant \inf(B_0)$ trivially and, if $k \leqslant \inf(B_k)$ then the first part gives $k + 1 \leqslant \inf(B_k) + 1 \leqslant \inf(B_{k+1})$, proving the induction step.

With the above in hand, we define a function $h\colon \mathbb{N} \to \mathbb{N}$ so that

$$\forall k \in \mathbb{N}\colon \quad h(k) = \inf(B_k) \tag{12.11}$$

(Note that B_k is the value at k of the function $\mathbb{N} \to \mathcal{P}(\mathbb{N})$ that defines $\{B_k\colon k \in \mathbb{N}\}$ and the infimum map $\mathcal{P}(\mathbb{N}) \smallsetminus \{\varnothing\} \to \mathbb{N}$ constructed in Lemma 9.12.) We claim

$$\forall k, \ell \in \mathbb{N}\colon k < \ell \ \Rightarrow \ h(k) < h(\ell) \tag{12.12}$$

which we prove by induction on ℓ with the induction step supplied by

$$h(\ell + 1) = \inf(B_{\ell+1}) \geqslant \inf(B_\ell) + 1 = h(\ell) + 1 > h(\ell) \tag{12.13}$$

where the middle inequality follows from (12.10). In particular, from (12.12) we immediately get that h is injective.

It remains to show that h is surjective. Let $n \in \widetilde{B}$ and set

$$k(n) := \inf\{k \in \mathbb{N} \colon n \notin B_k\}. \tag{12.14}$$

where the infimum is well defined because the set under the infimum is non-empty thanks to $n \notin B_{n+1}$ proved from $\inf(B_{n+1}) \geqslant n+1$ stated in (12.10). Now observe that $k(n) > 0$ because $n \in B_0 = \widetilde{B}$ and so $k(n) - 1 \in \mathbb{N}$. In addition, $n \notin B_{k(n)}$ but $n \in B_{k(n)-1}$ for otherwise $k(n)$ is not the infimum. Hence

$$n \in B_{k(n)-1} \smallsetminus B_{k(n)} = \{\inf(B_{k(n)-1})\} \tag{12.15}$$

proving that $n = \inf(B_{k(n)-1}) = h(k(n)-1)$. This gives $n \in \mathrm{Ran}(h)$ thus showing that $\mathrm{Ran}(h) = \widetilde{B}$ and h is surjective.

To complete the proof of the claim, consider the map $f \circ h$ and check that it maps $\mathbb{N}$ bijectively onto B, as desired. □

The statement and proof can be summed up as follows. First put forward a very useful notion:

Definition 12.8 (Sequence) *Let A be a set. An A-valued* sequence $\{x_n\}_{n\in\mathbb{N}}$ *(or a sequence taking values in A) is a function $f \colon \mathbb{N} \to A$ with $\mathrm{Dom}(f) = \mathbb{N}$ and such that $\forall n \in \mathbb{N} \colon x_n = f(n)$. We reserve the notation $\{x_n \colon n \in \mathbb{N}\}$ for $\mathrm{Ran}(f)$.*

The above proof then shows:

Corollary 12.9 *Let A be an infinite countable set. For each infinite set $B \subseteq A$ there exists a sequence $\{x_n\}_{n\in\mathbb{N}}$ such that $B = \{x_n \colon n \in \mathbb{N}\}$. (We say that B is* enumerated *or* exhausted *by $\{x_n\}_{n\in\mathbb{N}}$.) If $A = \mathbb{N}$, then the sequence $\{x_n\}_{n\in\mathbb{N}}$ can be taken to be strictly increasing in the sense $\forall n \in \mathbb{N} \colon x_n < x_{n+1}$.*

Another observation that we will find useful is as follows:

Corollary 12.10 *For any set B:*

$$B \text{ finite or countable} \Leftrightarrow \exists f \colon B \to \mathbb{N} \colon \text{ injection}. \tag{12.16}$$

Proof. The direction $\Rightarrow$ follows directly from Definitions 12.1 and 12.6. For $\Leftarrow$ we just need to address the case of B infinite. Here the injection f puts B into bijective correspondence with $f(B) = \mathrm{Ran}(f)$ which, being a subset of $\mathbb{N}$, is countable by Lemma 12.7. Thus B is countable as well. □

We will now proceed to note that, similarly as for finite sets, countability is preserved by certain natural operations on sets. We begin by the Cartesian product:

Lemma 12.11 *Let A and B be countable sets. Then so is $A \times B$.*

Proof. Let $f \colon A \to \mathbb{N}$ and $g \colon A \to \mathbb{N}$ be injections ensured by A and B being countable. Then $h(a,b) := (f(a), g(b))$ defines an injection $h \colon A \times B \to \mathbb{N} \times \mathbb{N}$. (We leave checking that h is an injection to the reader.) It thus suffice so show that there is an injection $\phi \colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ because the map $\phi \circ h$ then gives us an injection $A \times B \to \mathbb{N}$, proving the claim via Corollary 12.10.

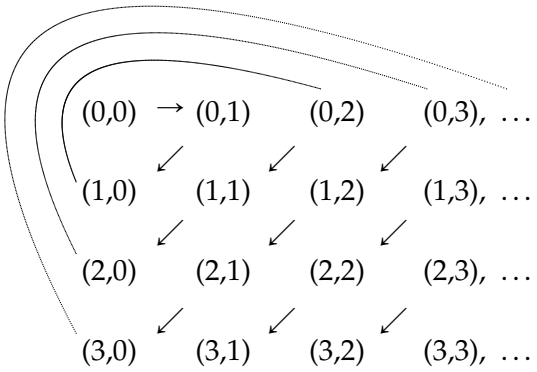

FIG 1. The path sweeping through $\mathbb{N} \times \mathbb{N}$ increasingly with respect to the ordering relation (12.17). The resulting bijection $\phi\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ is given explicitly in (12.19).

We will choose the injection that labels pairs of integers by the ordering relation

$$(m,n) \preceq (\tilde{m},\tilde{n}) \;:=\; m+n < \tilde{m}+\tilde{n} \;\vee\; \Big(m+n = \tilde{m}+\tilde{n} \;\wedge\; m \leqslant \tilde{m}\Big) \tag{12.17}$$

which, in light of $\leqslant$ being a total ordering of $\mathbb{N}$, is a total order of $\mathbb{N} \times \mathbb{N}$. As is easy to check, this injection is given explicitly by

$$\phi(m,n) := \frac{1}{2}(m+n)(m+n+1) + m, \tag{12.18}$$

where the first term on the right is a natural by the fact that the product of two consecutive naturals is even.

To see that ϕ is an injection, it suffices to show that it is strictly monotone. Indeed, if $k := m+n < \tilde{k} := \tilde{m}+\tilde{n}$, then $k+1 \leqslant \tilde{k}$ and so

$$\begin{aligned} \phi(m,n) &= \frac{1}{2}k(k+1) + m \leqslant \frac{1}{2}k(k+1) + k = \frac{1}{2}k(k+3) \\ &\leqslant \frac{1}{2}(\tilde{k}-1)(\tilde{k}+2) = \frac{1}{2}\tilde{k}(\tilde{k}+1) - 1 < \frac{1}{2}\tilde{k}(\tilde{k}+1) + \tilde{m} = \phi(\tilde{m},\tilde{n}) \end{aligned} \tag{12.19}$$

If $m+n = \tilde{m}+\tilde{n}$ and $m < \tilde{m}$, then $\phi(m,n) < \phi(\tilde{m},\tilde{n})$ directly from the definition. □

Corollary 12.12 *The set $\mathbb{Q}$ of all rationals is infinite and countable.*

Proof. Since $\mathbb{N}$ is infinite by Lemma 12.2, $\mathbb{N} \subseteq \mathbb{Q}$ and Lemma 12.4 force that $\mathbb{Q}$ is infinite as well. For the second part we write each rational a in the unique form $\frac{p}{q}$, where q is minimal positive. This is achieved by defining, for all $a \in \mathbb{Q}$,

$$\begin{aligned} q(a) &:= \inf\big\{q' \in \mathbb{N}\colon q' > 0 \wedge (\exists p' \in \mathbb{Z}\colon aq' = p')\big\} \\ p(a) &:= a \cdot q(a) \end{aligned} \tag{12.20}$$

(Lemma 9.10 makes this well defined and ensures that $p \in \mathbb{Z}$.) Then

$$f(a) := \begin{cases} \big(2p(a), q(a)\big) & \text{if } a \geqslant 0 \\ \big(1-2p(a), q(a)\big) & \text{if } a < 0 \end{cases} \tag{12.21}$$

then defines an function $f\colon \mathbb{Q} \to \mathbb{N} \times \mathbb{N}$ which is injective because the first component determines the sign of a via parity and then, using the appropriate alternative above, also $q(a)$ and $p(a)$ and thus $a = p(q)/q(a)$. Since Lemma 12.11 ensures that $\mathbb{N} \times \mathbb{N}$ is countable, so is $\mathbb{Q}$. □

Using Lemma 12.11 we also readily check:

Lemma 12.13 *For all infinite sets A and B,*

$$A, B \text{ countable} \;\Rightarrow\; A \cup B \text{ countable} \tag{12.22}$$

Proof. The assumption that A and B are infinite countable tells us that there exist bijections $f\colon A \to \mathbb{N}$ and $g\colon B \to \mathbb{N}$. Lemma 12.11 also gives us a bijection $h\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$. Define $\phi\colon A \cup B \to \mathbb{N} \times \mathbb{N}$ with $\mathrm{Dom}(\phi) = A \cup B$ by

$$\phi(x) := \begin{cases} \big(f(x)+1, 0\big), & \text{if } x \in A \smallsetminus B \\ \big(0, g(x)+1\big), & \text{if } x \in B \smallsetminus A \\ \big(f(x)+1, g(x)+1\big), & \text{if } x \in A \cap B \end{cases} \tag{12.23}$$

The definition ensures that ϕ is an injection and so $h \circ \phi$ is an injection $A \cup B \to \mathbb{N}$. The claim now follows from Corollary 12.10. □

We note that the same conclusion applies to any finite union of countable sets. For infinite unions, we have to be somewhat more careful:

Lemma 12.14 (AC) *For all sets $\{A_n\colon n \in \mathbb{N}\}$,*

$$\Big(\forall n \in \mathbb{N}\colon A_n \text{ finite or countable}\Big) \;\Rightarrow\; \bigcup_{n\in\mathbb{N}} A_n \text{ finite or countable} \tag{12.24}$$

In short, assuming AC, a countable union of countable sets is countable.

Proof. Abbreviate $A := \bigcup_{n\in\mathbb{N}} A_n$. Assuming each A_n to be countable, the Axiom of Choice ensures existence of a collection $\{f_n\colon n \in \mathbb{N}\}$ of functions $f_n\colon A \to \mathbb{N}$ such that f_n is bijective for each $n \in \mathbb{N}$. (This amounts to existence of a function $f\colon \mathbb{N} \to N^A$ such that for all $n \in \mathbb{N}$, the function $f_n(\cdot) := f(n, \cdot)$ is an bijection $A_n \to \mathbb{N}$. Since a choice is made here, the Axiom of Choice must be invoked.) For each $a \in \bigcup_{n\in\mathbb{N}} A_n$, set

$$n(a) := \inf\{k \in \mathbb{N}\colon a \in A_k\} \tag{12.25}$$

which exists by Lemma 9.10 and the fact that $a \in \bigcup_{n\in\mathbb{N}} A_n$ implies $\exists k \in \mathbb{N}\colon a \in A_k$. Then define $h\colon \bigcup_{n\in\mathbb{N}} A_n \to \mathbb{N} \times \mathbb{N}$ by

$$h(a) := \big(n(a), f_{n(a)}(a)\big). \tag{12.26}$$

Note that if $n(a) = n(b)$ and $f_{n(a)}(a) = f_{n(b)}(b)$, then the fact that $f_{n(a)}$ is an injection forces $a = b$. It follows that h is an injection of $\bigcup_{n\in\mathbb{N}} A_n$ into $\mathbb{N} \times \mathbb{N}$ and so $\bigcup_{n\in\mathbb{N}} A_n$ is countable by Lemma 12.11. □

To explain the need for Axiom of Choice, note that assuming only the existence of bijections $f\colon \mathbb{N} \to A$ and $g\colon \mathbb{N} \to B$, Lemma 12.13 concluded the existence of a bijection $\mathbb{N} \to A \cup B$ with no choice required whatsoever. However, once an infinite number of sets are involved, we run into a technical obstruction that

$$\forall n \in \mathbb{N}\colon \{\, f \in A_n^{\mathbb{N}}\colon \text{ bijection}\} \neq \varnothing \tag{12.27}$$

does not generally imply

$$\bigtimes_{n\in\mathbb{N}} \{ f \in A_n^{\mathbb{N}} \colon \text{ bijection} \} \neq \varnothing \tag{12.28}$$

unless the Axiom of Choice is assumed. That being said, the use of the Axiom of Choice can be avoided once the sets A_n have some *a priori* structure; e.g., are naturally ordered such as subsets of the naturals or finite subsets of $\mathbb{R}$. This is demonstrated on:

Lemma 12.15 (Cantor) *The set*

$$\left\{ x \in \mathbb{R} \colon \left(\exists n \in \mathbb{N}\, \exists a_0, \dots, a_n \in \mathbb{Z} \colon a_n \neq 0 \ \wedge\ a_n x^n + \cdots + a_1 x + a_0 = 0 \right) \right\} \tag{12.29}$$

of (so called) algebraic numbers is countable.

We leave a proof of this fact to a homework exercise. The existence of non-algebraic numbers was first shown by Dedekind who communicated the proof in a letter to Cantor (the existence of which Cantor withheld in his main article on this topic). Cantor's statement above is more general as it shows that, in fact, most (in the sense of cardinality) real numbers are non-algebraic.

We note in passing that the set of algebraic numbers is actually preserved by operations of addition, multiplication and taking additive a multiplicative inverses and so is a field. It is the smallest field one needs to solve all polynomial equations with rational, or even algebraic, coefficients.

13. Uncountable sets and beyond

Here we continue discussing the notion of the "size" of the set focusing on infinite sets. This will allow us to take a quick peek into this highly developed subject in set theory.

13.1 Uncountable sets.

The notion of countability would not be very useful if all sets were finite or countable. That this is not the case is the content of:

Theorem 13.1 (Cantor 1891) *There exists no surjection* $\mathbb{N} \to \{0,1\}^{\mathbb{N}}$. *In particular,*

$$\{0,1\}^{\mathbb{N}} \text{ is uncountable} \tag{13.1}$$

Proof. Recall that $\{0,1\}^{\mathbb{N}}$ is the set of all functions $f\colon \mathbb{N} \to \{0,1\}$ with $\mathrm{Dom}(f) = \mathbb{N}$. Let $\phi\colon \mathbb{N} \to \{0,1\}^{\mathbb{N}}$ be a function. Abbreviating $f_n := \phi(n)$, define $h\colon \mathbb{N} \to \{0,1\}$ so that

$$\forall k \in \mathbb{N}\colon \quad h(k) = 1 - f_k(k) \tag{13.2}$$

Then $h \in \{0,1\}^{\mathbb{N}}$ with $\mathrm{Dom}(h) = \mathbb{N}$ so $h \in \{0,1\}^{\mathbb{N}}$. Yet h is distinct from each f_k as it differs from f_k at k and so $h \notin \mathrm{Ran}(\phi)$. In particular ϕ is not a surjection and since ϕ was arbitrary, $\{0,1\}^{\mathbb{N}}$ is uncountable. □

The idea to arrange elements into a two-dimensional array and then produce another element by picking or changing the diagonal terms is often referred to as the *Cantor diagonal argument*. The above statement plus a bit of arithmetic shows:

Corollary 13.2 *The interval* $[0,1] := \{x \in \mathbb{R}\colon 0 \leqslant x \leqslant 1\}$ *and thus also* $\mathbb{R}$ *are uncountable.*

Proof. The proof uses the notion of infinite series that we have not covered systematically. However, all we need is the following definition:

Definition 13.3 *For any* $I \neq \varnothing$ *and any* $\{x_\alpha\colon \alpha \in I\} \subseteq [0,\infty) := \{x \in \mathbb{R}\colon 0 \leqslant x\}$, *if there is* $C \in [0,\infty)$ *such that*

$$\forall F \subseteq I\colon \ F \text{ finite } \Rightarrow \sum_{\alpha\in F} x_\alpha \leqslant C \tag{13.3}$$

we set

$$\sum_{\alpha\in I} x_\alpha := \sup\Big\{\sum_{\alpha\in F} x_\alpha\colon \ F \subseteq I \text{ finite}\Big\} \tag{13.4}$$

(The supremum exists by the least upper bound property of the reals.)

Now pick $\sigma \in \{0,1\}^{\mathbb{N}}$ and note that (11.9) gives $(1-q)\sum_{k=0}^{n} q^k = 1 - q^{n+1}$ for all $q \in \mathbb{R}$ and all $n \in \mathbb{N}$. This shows

$$0 \leqslant \sum_{k=0}^{n} \frac{2\sigma_k}{3^{k+1}} \leqslant \frac{2}{3}\sum_{k=0}^{n} 3^{-k} = \frac{2}{3}\frac{1-3^{-n+1}}{1-3^{-1}} \leqslant 1 \tag{13.5}$$

Hence we can set

$$f(\sigma) := \sum_{k\in\mathbb{N}} \frac{2\sigma_k}{3^{k+1}} \tag{13.6}$$

in the sense of Definition 13.3. This gives us a map $f\colon \{0,1\}^{\mathbb{N}} \to [0,1]$.

Remark 13.4 Note that $f(\sigma)$ is the real number whose representation in base-3 expansion takes the from $0.\eta_1\eta_2\eta_3\ldots$, where $\eta_i := 2\sigma_{i-1}$. In particular, $\mathrm{Ran}(f)$ is the set of all such numbers that can be written using 0's and 2's only, with no 1's allowed. Restricting to such numbers removes the degeneracy that all these expansions suffer from (e.g., $0.02222222\ldots$ is the same number as $0.10000000\ldots$) which would ruin injectivity of the map; see below. We will return to this later when we discuss *Cantor's ternary set*.

Next let $\sigma, \sigma' \in \{0,1\}^{\mathbb{N}}$ be such that $\sigma_k = \sigma'_k$ for all $k = 0, \ldots, n-1$ and $\sigma_n = 0$ while $\sigma'_n = 1$. Then the fact that $\sup(b + A) = b + \sup(A)$ for $b + A := \{b + a \colon a \in A\}$ whenever $\sup(A)$ exists shows

$$\begin{aligned} f(\sigma) &= \sum_{k=0}^{n-1} \frac{2\sigma_k}{3^{k+1}} + \sum_{k\in[n+1,\infty)} \frac{2\sigma_k}{3^{k+1}} \\ &\leqslant \sum_{k=0}^{n-1} \frac{2\sigma_k}{3^{k+1}} + \frac{1}{3^{n+1}} \\ &< \sum_{k=0}^{n-1} \frac{2\sigma_k}{3^{k+1}} + \frac{2}{3^{n+1}} = \sum_{k=0}^{n} \frac{2\sigma'_k}{3^{k+1}} \leqslant f(\sigma') \end{aligned} \tag{13.7}$$

thus showing that f is injective. If $[0,1]$ or $\mathbb{R}$ were countable, then Lemma 12.7 would imply that the image $f(\{0,1\}^{\mathbb{N}})$ is countable. But f is a bijection of $\{0,1\}^{\mathbb{N}}$ onto its image, so that would give that $\{0,1\}^{\mathbb{N}}$ is countable, in contradiction with Theorem 13.1. □

The take-away message here is that there are just many more reals than rationals. From Theorem 13.1 and Lemma 12.15 we also conclude:

Corollary 13.5 *There exists a real which is not algebraic.*

In fact, the argument shows most reals are not algebraic. Non-algebraic real or complex numbers are sometimes called *transcendental*.

We note that, prior to Cantor, Corollary 13.5 was proved by Liouville using the concept of *Liouville numbers* which are those $x \in \mathbb{R}$ such that for each $n \in \mathbb{N}$ there is a rational of the form p/q with $q > 1$ for which $0 < |x - p/q| < q^{-n}$. As it turns out, such numbers exist and are all transcendental.

13.2 Non-equinumerosity of any set with its power set.

Cantor continued to develop the notion of cardinality further to include even larger sets than reals. We already encountered his relation $\simeq$ of *equinumerosity* in Definition 3.11 defined for any two sets A and B by

$$A \simeq B \;:=\; \exists f \colon A \to B \text{ bijection} \tag{13.8}$$

This is readily checked to be an equivalence relation on sets (technically, we have to talk about the class of all sets at this point or restrict to subsets of a given set). However, unlike for finite sets, we then define the cardinality of A somewhat differently:

Definition 13.6 *For any set A, the* cardinality class *of A is the equivalence class $[A]$ under the equinumerosity relation.*

Thus, the cardinality class of the interval $[0,n) := \{k \in \mathbb{N} : k < n\}$ includes all sets with exactly n elements. By Lemma 12.7, all infinite countable sets lie in $[\mathbb{N}]$ which (by Theorem 13.1 and Corollary 13.2) includes neither $\{0,1\}^{\mathbb{N}}$ nor $\mathbb{R}$. As each $\sigma \in \{0,1\}^{\mathbb{N}}$ is in one-to-one correspondence with the subset $A := \{n \in \mathbb{N} : \sigma(n) = 1\}$ of the naturals, also $\mathcal{P}(\mathbb{N})$ is uncountable and so

$$\mathcal{P}(\mathbb{N}) \notin [\mathbb{N}] \tag{13.9}$$

or, equivalently, $\mathbb{N} \notin [\mathcal{P}(\mathbb{N})]$. Another profound discovery made by Cantor is that this holds for all sets:

Theorem 13.7 (Cantor 1891) *For any set A, there is no surjection $f : A \to \mathcal{P}(A)$ and so*

$$A \notin \big[\mathcal{P}(A)\big] \tag{13.10}$$

Proof. The proof is based on a variation of the Cantor diagonal argument albeit in the form that is more reminiscent of Russell's antinomy (see Theorem 2.1). Indeed, assume $f : A \to \mathcal{P}(A)$ to be a surjection. Define

$$B := \big\{x \in A : x \notin f(x)\big\} \tag{13.11}$$

Then $B \in \mathcal{P}(A)$ and, since f is a surjection, there is $b \in A$ such that $f(b) = B$. But then $b \in B$ implies $b \notin f(b) = B$ while $b \notin B = f(b)$ implies $b \in B$. As at least one of $b \in B$ or $b \notin B$ must be TRUE — remember that $b \notin B$ is a shorthand for $\neg(b \in B)$ — we arrive at a contradiction and so no such f can exist after all. □

13.3 Size comparisons.

Theorem 13.7 shows that applying the powerset to a given set keeps producing sets of different cardinality. Intuitively, $\mathcal{P}(A)$ is larger than A but to formulate that precisely, we need a tool to compare sizes of non-equinumerous sets. This is furnished by the binary relation $\lesssim$ of *size comparison* defined as

$$A \lesssim B \;:=\; \exists f : A \to B \text{ injection} \tag{13.12}$$

This relation has many reasonable properties. Indeed, using the identity map we immediately verify the intuitive fact

$$A \subseteq B \Rightarrow A \lesssim B \tag{13.13}$$

We also check that it extends to equivalence classes as

$$A \simeq \widetilde{A} \;\wedge\; A \lesssim B \;\Rightarrow\; \widetilde{A} \lesssim B \tag{13.14}$$

The relation $\lesssim$ is also readily checked to be reflexive and transitive, but is not an ordering as antisymmetry cannot hold in general. Indeed, $A \lesssim B$ and $B \lesssim A$ do not imply that A equals B as these could be different sets (take $A := \mathbb{N}$ and $B := \mathbb{Q}$, for instance). However, the following natural alternative to equality does hold:

Theorem 13.8 (Cantor-Bernstein/Schröder-Bernstein) *Let A and B be sets. Then*

$$A \lesssim B \;\wedge\; B \lesssim A \;\Rightarrow\; A \simeq B \tag{13.15}$$

In particular, $\lesssim$ induces a partial order on equinumerosity classes.

Proof. The assumptions $A \lesssim B \wedge B \lesssim A$ imply existence of an injection $f\colon A \to B$ and an injection $g\colon B \to A$. The map $g \circ f$ then images A injectively into A. Let $\{(g \circ f)^n \colon n \in \mathbb{N}\}$ be functions with domain A constructed recursively so that, for all $x \in A$, we have $(g \circ f)^0(x) = x$ and $\forall n \in \mathbb{N}\colon (g \circ f)^{n+1}(x) = (g \circ f)((g \circ f)^n(x))$. Consider the set

$$C := \big\{x \in A\colon (\forall n \in \mathbb{N}\colon x \in \mathrm{Ran}((g \circ f)^n)\big\} \tag{13.16}$$

of points on which $g \circ f$ can be inverted any number of times and let

$$\begin{aligned} D &:= \bigcup_{n\in\mathbb{N}} (g \circ f)^n\big(A \setminus g(B)\big) \\ E &:= \bigcup_{m\in\mathbb{N}} (g \circ f)^m\big(g(B) \setminus g \circ f(A)\big) \end{aligned} \tag{13.17}$$

A key part of the argument is to show that the sets C, D and E are disjoint and their union is all of A; i.e., they form a disjoint partition of A.

For all $x \in A \setminus C$ set

$$n(x) := \sup\big\{n \in \mathbb{N}\colon x \in \mathrm{Ran}((g \circ f)^n)\big\} \tag{13.18}$$

Note that $n(x) = n$ means that $x = (g \circ f)^n(z)$ for some $z \in A$ but $x = (g \circ f)^{n+1}(z')$ for all $z' \in A$. For each $n \in \mathbb{N}$, the injectivity of $g \circ f$ then gives

$$\begin{aligned} &(g \circ f)^n\big(A \setminus g(B)\big) \cup (g \circ f)^n\big(g(B) \setminus g \circ f(A)\big) \\ &\qquad = (g \circ f)^n\Big(\big(A \setminus g(B)\big) \cup \big(g(B) \setminus g \circ f(A)\big)\Big) \\ &\qquad = (g \circ f)^n\big(A \setminus g \circ f(A)\big) \\ &\qquad = (g \circ f)^n(A) \setminus (g \circ f)^{n+1}(A) \\ &\qquad = \big\{x \in A\colon n(x) = n\big\} \end{aligned} \tag{13.19}$$

The latter set is disjoint from C by an earlier argument, so we have

$$D \cap C = \varnothing \ \wedge\ E \cap C = \varnothing \tag{13.20}$$

Since the sets $\{x \in A\colon n(x) = n\}$ and $\{x \in A\colon n(x) = m\}$ are disjoint for $n \neq m$, the computation (13.19) also proves disjointness of the n-th term in the union defining D and the m-th term in the union defining E, whenever $n \neq m$. For the terms with $m = n$ we in turn observe

$$\begin{aligned} &(g \circ f)^n\big(A \setminus g(B)\big) \cap (g \circ f)^n\big(g(B) \setminus g \circ f(A)\big) \\ &\qquad = (g \circ f)^n\Big(\big(A \setminus g(B)\big) \cap \big(g(B) \setminus g \circ f(A)\big)\Big) = \varnothing \end{aligned} \tag{13.21}$$

thus showing that

$$D \cap E = \varnothing \tag{13.22}$$

as well. One final use of the calculation (13.19) along with the fact that $n(x) \in \mathbb{N}$ for all $x \in A \setminus C$ shows

$$C \cup D \cup E = C \cup \bigcup_{n\in\mathbb{N}} \big\{x \in A\colon n(x) = n\big\} = A \tag{13.23}$$

proving that the sets C, D and E form a partition of A.

Moving to the proof of the stated claim, observe that

$$E \subseteq \bigcup_{n\in\mathbb{N}} (g\circ f)^n\big(g(B)\big) = g\Big(\bigcup_{n\in\mathbb{N}} (f\circ g)^n(B)\Big) \tag{13.24}$$

implying that $E \subseteq \mathrm{Ran}(g)$. This means that g^{-1} is well defined everywhere on E which permits us to define $h\colon A \to B$ with $\mathrm{Dom}(h) = A$ by

$$h(x) := \begin{cases} f(x), & \text{if } x \in C \cup D \\ g^{-1}(x), & \text{if } x \in E \end{cases} \tag{13.25}$$

We claim that $g \circ h$ is a bijection of A onto $g(B)$. Indeed, in both alternatives defining h the map $g \circ f$ is an injection which acts as the identity on E while on $C \cup D$ it effectively increases $n(\cdot)$ by one, and so it images C onto C and D onto $g \circ f(D)$. Using the disjointness of the sets constituting the unions in (13.17) along with (13.23) we get

$$g\circ f(D) \cup C = C \cup \bigcup_{n\in\mathbb{N}\smallsetminus\{0\}} (g\circ f)^n\big(A \smallsetminus g(B)\big) = A \smallsetminus \Big(E \cup \big(A \smallsetminus g(B)\big)\Big) \tag{13.26}$$

and, in particular, $g \circ h(E) \cap g \circ h(C \cup D) = \varnothing$, proving that $g \circ h$ is an injection on all of A. The observation (13.26) then also gives that

$$\mathrm{Ran}(h) = g\circ f(D) \cup C \cup E = A \smallsetminus \big(A \smallsetminus g(B)\big) = g(B) \tag{13.27}$$

Since g^{-1} is well defined and bijective on $g(B)$, it follows that h is a bijection of A onto B, thus proving the claim. □

Theorem 13.8 has a rather colorful history: The result was first stated without proof by Cantor in 1887 with Dedekind finding a proof the same year albeit without publishing it (this proof was discovered by Zermelo in 1908). Then in 1897 Bernstein and Schröder found a proof independently although that of Schröder was later shown to be incorrect. Etc. (See the wiki page for more details.)

The exercises in the textbook delineate Zermelo's proof which is aimed at the statement that $A \subseteq F \subseteq B$ and $A \simeq B$ imply $A \simeq F \simeq B$.

13.4 Towers of equinumerosity classes.

The punchline of Theorem 13.8 is that, in order to prove equinumerosity of two sets, it suffices to demonstrate an injection from one to the other and *vice versa*. To see how this works in practice, note that the proof of Corollary 13.2 demonstrated an injection $\{0,1\}^{\mathbb{N}} \to [0,1]$. From $\mathcal{P}(\mathbb{N}) \simeq \{0,1\}^{\mathbb{N}}$ and $[0,1] \subseteq \mathbb{R}$ we get $\mathcal{P}(\mathbb{N}) \lesssim \mathbb{R}$. On the other hand, lifting the bijection between $\mathbb{N}$ and $\mathbb{Q}$ to their power sets shows $\mathcal{P}(\mathbb{N}) \simeq \mathcal{P}(\mathbb{Q})$ while the concept of Dedekind cut gives an injective embedding of $\mathbb{R}$ into $\mathcal{P}(\mathbb{Q})$ thus proving $\mathbb{R} \lesssim \mathcal{P}(\mathbb{N})$. Theorem 13.8 then concludes

$$\mathcal{P}(\mathbb{N}) \simeq \{0,1\}^{\mathbb{N}} \simeq [0,1] \simeq (0,1) \simeq \mathbb{R} \tag{13.28}$$

A special name is reserved for the equivalence class of all these sets:

Definition 13.9 *The equivalence class of* $\mathbb{R}$ *under equinumerosity relation is called the* continuum. *Any set in* $[\mathbb{R}]$ *is said to have* cardinality of the continuum.

Similar arguments as used above in fact show that

$$\mathbb{R} \times \mathbb{R} \lesssim \mathbb{R} \tag{13.29}$$

and so, using also the trivial embedding $\mathbb{R} \lesssim \mathbb{R} \times \mathbb{R}$, Theorem 13.8 and induction give

$$\forall n \in \mathbb{N} \colon n \geqslant 1 \;\Rightarrow\; \mathbb{R}^n \simeq \mathbb{R} \tag{13.30}$$

meaning that all Euclidean spaces are of cardinality of the continuum. (We leave details to homework exercise.)

Pushing this further, the fact that $\mathbb{R} \simeq \{0,1\}^{\mathbb{N}}$ and $\mathbb{N} \times \mathbb{N} \simeq \mathbb{N}$ imply

$$\mathbb{R}^{\mathbb{N}} \simeq \{0,1\}^{\mathbb{N}\times\mathbb{N}} \simeq \{0,1\}^{\mathbb{N}} \simeq \mathbb{R} \tag{13.31}$$

so even the space of all real-valued sequences is of cardinality of the continuum. However, by Theorem 13.7, the space $\mathbb{R}^{\mathbb{R}}$ of all functions $\mathbb{R} \to \mathbb{R}$ and, by $\mathbb{N} \times \mathbb{R} \simeq \mathbb{R}$ (prove this!), also the space of just zero-one valued functions on $\mathbb{R}$ are strictly larger:

$$\mathbb{R} \lesssim \mathcal{P}(\mathbb{R}) \simeq \{0,1\}^{\mathbb{R}} \simeq \{0,1\}^{\mathbb{N}\times\mathbb{R}} \simeq \mathbb{R}^{\mathbb{R}} \tag{13.32}$$

where we used

$$A \lesssim B := A \lesssim B \wedge \neg(A \simeq B) \tag{13.33}$$

We again leave checking the details to the reader.

An attentive reader will notice that the various "orders of infinity" encountered above (namely, the naturals $\mathbb{N}$, the reals $\mathbb{R}$ in (13.28) and $\mathbb{R}$-valued functions on $\mathbb{R}$ in (13.32)) can be constructed from $\mathbb{N}$ by powerset operation. This suggests that we push this further: Using Theorem 13.7 along with the natural injection $x \mapsto \{x\}$ of every set in its power set demonstrates an infinite sequence

$$\mathbb{N} \lesssim \mathcal{P}(\mathbb{N}) \simeq \mathbb{R} \lesssim \mathcal{P}(\mathcal{P}(\mathbb{N})) \lesssim \mathcal{P}(\mathcal{P}(\mathbb{N})) \lesssim \mathcal{P}\big(\mathcal{P}(\mathcal{P}(\mathbb{N}))\big) \lesssim \dots \tag{13.34}$$

of infinite sets with distinct cardinalities. A natural question is whether other "orders of infinity" might exist as well. An immediate answer to this is in the affirmative: The union of the infinite sequence of iterated powersets of $\mathbb{N}$ above (which can be defined precisely via the Recursion principle) is not equinumerous to any member thereof. We have thus discovered yet another way (besides powerset) to produce "a set larger than that before" by taking the union of an infinite family of all "previous" infinities. (A formal construction requires the notion of a *limit ordinal*.)

That being said, more important for analysis (which is concerned mostly with the first two or three terms in the above sequence) is the question whether some "orders of infinity" are perhaps missing from (13.34) because the powerset construction "jumped" over them. This already concerns the first and second term and is the basis of:

Continuum hypothesis: (Cantor 1878) *Every infinite subset of the reals is either countable or of cardinality of the continuum. In short,*

$$\forall A \subseteq \mathbb{R} \colon\; A \text{ infinite } \;\Rightarrow\; \big(A \simeq \mathbb{N} \;\vee\; A \simeq \mathbb{R}\big) \tag{13.35}$$

This "hypothesis" has been a subject of much debate in the first half of the 20th century for it seemed to have demonstrable consequences for what could/could not be done in mathematics. A striking resolution was presented by K. Gödel in 1940 and P. Cohen in 1963 who showed that the Continuum hypothesis can be neither disproved (Gödel) nor proved (Cohen) in Zermelo's theory (assuming the latter is free of contradictions). In

other words, adding the Continuum hypothesis to Zermelo's axioms creates one mathematical universe, adding its negation creates another and yet another universe is produced by leaving the TRUE/FALSE value of (13.35) undecided.

For this reason, just as with the Axiom of choice, mathematicians are careful to regard results that rely on the Continuum hypothesis as conditional on this axiom to be added and often mark that by adding the acronym "CH" to the statement of the theorem. ("AC" is used to mark the use of the Axiom of choice.)

14. METRIC SPACE CONVERGENCE

We are ready at last to commence the discussion of topics that should be familiar from calculus (which can be thought of as a non-technical, or practical-use oriented, version of analysis). We start with limits of real-valued sequences.

14.1 Convergence of real-valued sequences.

The concept of *convergence* is fundamental for analysis. We will first discuss it in the context of convergence of sequences. Recall that the notion of a sequence was introduced already in Definition 12.8 where we defined an A-valued sequence $\{x_n\}_{n\in\mathbb{N}}$ to be a function $\mathbb{N} \to A$ with $\mathrm{Dom}(f) = \mathbb{N}$ whose value at n is x_n. We will largely suppress this technical interpretation in what follows and think of a sequence intuitively as a line of objects indexed by the naturals.

Consider the following example of a sequence $\{a_n\}_{n\in\mathbb{N}}$ taking values in rationals which is defined recursively as

$$a_0 := 1 \ \wedge\ \forall n \in \mathbb{N}\colon\ a_{n+1} := 3 - \frac{1}{a_n} \tag{14.1}$$

It is easy to evaluate a couple of first terms,

$$a_0 = 1,\ a_1 = 2,\ a_3 = 2.5,\ a_4 = 2.6,\ a_5 = 2.615\ldots \tag{14.2}$$

It appears that the values rise with the rising index, albeit not above (and even to) the value 3. And, indeed, we easily prove:

Lemma 14.1 $\forall n \in \mathbb{N}\colon\ 1 \leqslant a_n \leqslant 3 \ \wedge\ a_n < a_{n+1}$

Proof. Let $P_n := 1 \leqslant a_n \leqslant 3 \ \wedge\ a_n < a_{n+1}$. Noting that $1 = a_0 < a_1 = 2 \leqslant 3$ we get that P_0 is TRUE. Assuming P_n, we have $\frac{1}{a_{n+1}} < \frac{1}{a_n}$ and so

$$a_{n+2} = 3 - \frac{1}{a_{n+1}} > 3 - \frac{1}{a_n} = a_{n+1} \tag{14.3}$$

Since $\frac{1}{a_n} \geqslant 0$, last equality also gives $a_{n+1} \leqslant 3$ (in fact, as $\frac{1}{a_n} \geqslant \frac{1}{3}$ we even have $a_{n+1} \leqslant 2.666\ldots$) while the fact that $a_n \geqslant 1$ implies $1 \leqslant a_n \leqslant a_{n+1}$. Hence $P_n \Rightarrow P_{n+1}$ and so the claim holds by induction. □

We have thus verified that $\{a_n\}_{n\in\mathbb{N}}$ conforms to:

Definition 14.2 *Let $(A, \leqslant)$ be a partially ordered set. A sequence $\{x_n\}_{n\in\mathbb{N}}$ taking values in A is then said to be*

- non-decreasing *if* $\forall n \in \mathbb{N}\colon\ x_n \leqslant x_{n+1}$ *and* strictly increasing *if* $\forall n \in \mathbb{N}\colon\ x_n < x_{n+1}$
- non-increasing *if* $\forall n \in \mathbb{N}\colon\ x_{n+1} \leqslant x_n$ *and* strictly decreasing *if* $\forall n \in \mathbb{N}\colon\ x_{n+1} < x_n$

Such sequences are generally referred to as monotone.

We remark that the words *increasing*, resp., *decreasing* are used as colloquial equivalents of non-decreasing, resp., non-increasing, but the uncertainty which of "$\leqslant$" or "$<$" is meant makes them less desirable when precision is of concern.

Another observation that may be gleaned from (14.2) is that the values of the sequence are getting closer and closer together, and perhaps even approach a "limit" point. Here are the precise definitions of these intuitive terms:

Definition 14.3 (Cauchy sequence in $\mathbb{R}$) *A real-valued sequence $\{x_n\}_{n\in\mathbb{N}}$ is said to be Cauchy if*

$$\forall k \in \mathbb{N}\, \exists n_0 \in \mathbb{N}\, \forall m, n \in \mathbb{N}\colon\ n, m \geqslant n_0 \ \Rightarrow\ |x_m - x_n| < \frac{1}{k+1} \tag{14.4}$$

Definition 14.4 (Limit of $\mathbb{R}$-valued sequence) *An real-valued sequence $\{x_n\}_{n\in\mathbb{N}}$ is said to have a limit, or converges, if*

$$\exists L \in \mathbb{R}\, \forall k \in \mathbb{N}\, \exists n_0 \in \mathbb{N}\, \forall m, n \in \mathbb{N}\colon\ n \geqslant n_0 \ \Rightarrow\ |x_n - L| < \frac{1}{k+1} \tag{14.5}$$

Any such L is then called a limit of $\{x_n\}_{n\in\mathbb{N}}$. We abbreviate (14.5) *as $x_n \to L$.*

We remark that in the literature (14.5) is typically stated as $\forall \epsilon > 0 \ldots |x_n - L| < \epsilon$; i.e., with $\frac{1}{k+1}$ replaced by ϵ. However, we do not need to check all positive $\epsilon > 0$ for this to be TRUE; it suffices to check this for a sequence of upper bounds that puts values arbitrarily close to zero. (The above uses sequence $\{\frac{1}{k+1}\}_{k\in\mathbb{N}}$ for these upper bounds.)

We now readily prove:

Lemma 14.5 *The sequence $\{a_n\}_{n\in\mathbb{N}}$ from* (14.1) *is Cauchy.*

Proof. Let $n \in \mathbb{N}$. A calculation shows

$$a_{n+2} - a_{n+1} = \left(3 - \frac{1}{a_{n+1}}\right) - \left(3 - \frac{1}{a_n}\right) = \frac{1}{a_n} - \frac{1}{a_{n+1}} = \frac{a_{n+1} - a_n}{a_n a_{n+1}} \tag{14.6}$$

Taking absolute values and noting that $a_n \geqslant 1$ but $a_{n+1} \geqslant a_1 = 2$ gives

$$|a_{n+2} - a_{n+1}| \leqslant \frac{|a_{n+1} - a_n|}{a_n a_{n+1}} \leqslant \frac{1}{2}|a_{n+1} - a_n| \tag{14.7}$$

We then use induction to verify that, for all $n \in \mathbb{N}$,

$$|a_{n+1} - a_n| \leqslant 2^{-n}|a_1 - a_0| = 2^{-n} \tag{14.8}$$

and then, for all $m, n \geqslant N$,

$$|a_m - a_n| \leqslant 2^{-N+1} \tag{14.9}$$

Since the right-hand side is decreasing in N, it then suffices to show that, for each $k \in \mathbb{N}$ there is $N \in \mathbb{N}$ such that $2^{-N+1} \leqslant \frac{1}{k+1}$. Since $\forall n \in \mathbb{N}\colon n + 1 \leqslant 2^n$ as verified again by induction, it suffices to choose $N := k + 1$. □

A very similar argument also gives:

Lemma 14.6 *The sequence $\{a_n\}_{n\in\mathbb{N}}$ from* (14.1) *is convergent with limit $L := \frac{3+\sqrt{5}}{2}$.*

Proof. We start by explaining where the precise value of L comes from. If we already know that the sequence is convergent, both a_n and a_{n+1} are close to L for all n large. Replacing these values by L in the recursive formula produces the equation

$$L = 3 - \frac{1}{L} \tag{14.10}$$

This is a quadratic equation $L^2 - 3L + 1 = 0$ with solutions $\frac{3\pm\sqrt{5}}{2}$. However, the solution with the negative sign is less than 1 and so cannot be a limit. So we get $\frac{3+\sqrt{5}}{2}$.

We now convert this argument to a proof that L is a limit of $\{a_n\}_{n\in\mathbb{N}}$. Suppose that L satisfies $L = 3 - \frac{1}{L}$ and $L \geqslant 2$. (This is satisfied for $\frac{3+\sqrt{5}}{2}$ but not $\frac{3-\sqrt{5}}{2}$.) Then

$$L - a_{n+1} = \left(3 - \frac{1}{L}\right) - \left(3 - \frac{1}{a_n}\right) = \frac{1}{a_n} - \frac{1}{L} = \frac{L - a_n}{a_n L} \tag{14.11}$$

Taking absolute value and using that $L \geqslant 2$ and $a_n \geqslant 1$ gives

$$|L - a_{n+1}| \leqslant \frac{1}{a_n L}\,|L - a_n| \leqslant \frac{1}{2}|L - a_n| \tag{14.12}$$

We then again readily prove by induction that, for all $n \in \mathbb{N}$,

$$|L - a_n| \leqslant \frac{1}{2^n}|L - a_0| = \frac{1}{2^n}|L - 1| \tag{14.13}$$

In order to complete the proof, we need a lemma whose proof (based on Archimedean property) we leave to the reader:

Lemma 14.7 $\forall x, y \in \mathbb{R}\colon\ x, y > 0 \ \wedge\ x < 1 \ \Rightarrow\ \exists n \in \mathbb{N}\colon\ x^n < y$

Indeed, given any $k \in \mathbb{N}$ and setting $y := \frac{1}{k+1}|L-1|^{-1}$ and $x := \frac{1}{2}$, this lemma gives us $n_0 \in \mathbb{N}$ such that the right-hand side of (14.13) is less than $\frac{1}{k+1}$ for all $n \geqslant n_0$. □

Some remarks are in order. First, note that while $\{a_n\}_{n\in\mathbb{N}}$ is $\mathbb{Q}$-valued, the limit L is not rational. It thus follows that the sequence is Cauchy even as as $\mathbb{Q}$-valued sequence, but it is not convergent in $\mathbb{Q}$. This is because being Cauchy is an *intrinsic* property of the sequence while being convergent depends also on the ambient space in which the sequence is immersed. We will return to this question more systematically later.

Second, the above procedure for controlling limit behavior of recursively defined sequences works quite generally. Indeed, if $f\colon \mathbb{R} \to \mathbb{R}$ is a given function, we can define the sequence by $\forall n \in \mathbb{N}\colon\ a_{n+1} := f(a_n)$ starting from the initial value a_0. The limit L, if such exist, will typically be a solution to the equation $L = f(L)$, meaning that L is a *fixed point* of f. Another example of this appears as a homework exercise.

14.2 Metric spaces.

Having digested the above notions in the context of real-valued sequences, we now turn to their generalizations beyond the reals. Here we note that what made Definitions 14.3 and 14.4 work was that in $\mathbb{R}$ we have a natural notion of *closeness*. Indeed, we say that x and y are close if $|x - y|$ is small. This motivates the following concept:

Definition 14.8 (Metric space) *A* metric space *is a pair (X, ϱ), where X is a set and $\varrho\colon\ X \times X \to \mathbb{R}$ is a function that obeys:*

(1) *(positivity)* $\forall x, y \in X\colon\ \varrho(x, y) \geqslant 0 \ \wedge\ (\varrho(x, y) = 0 \Leftrightarrow x = y)$
(2) *(symmetry)* $\forall x, y \in X\colon\ \varrho(x, y) = \varrho(y, x)$
(3) *(triangle inequality)* $\forall x, y, z \in X\colon\ \varrho(x, y) \leqslant \varrho(x, z) + \varrho(z, y)$

We call any such ϱ a metric *on X.*

The triangle inequality expresses the intuitive fact that the passage from x to y via z will be at least as long as the shortest possible way. Metric thus axiomatizes the intuitive notion of *distance*, with both words used synonymously in practice.

Using the definition of absolute value it is fairly easy to check that

$$\varrho(x,y) := |x-y| \tag{14.14}$$

defines a metric on $\mathbb{R}$ (and also on any subset thereof; in particular, on $\mathbb{Q}$). However, as asked to show in the homework, also

$$\varrho'(x,y) := \Big|\frac{x}{1+|x|} - \frac{y}{1+|y|}\Big| \tag{14.15}$$

is a metric on $\mathbb{R}$.

The metric (14.14) finds a number of possible generalizations in $\mathbb{R}^n$. Arguably the most natural of these is the *Euclidean metric* which, for points $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ is given by

$$\varrho_2(x,y) := \Big(\sum_{i=1}^{n}(x_i-y_i)^2\Big)^{1/2} \tag{14.16}$$

This is linked to the one-dimensional case by the fact that $\varrho_2(x,y)$ is the length of a straight line segment between x and y as measured by the metric (14.14). The squares and square root appear as a result of the Pythagorean theorem.

However, the Euclidean metric is not the only metric on $\mathbb{R}^n$ that is linked to (14.14). One other such metric is the *∞-metric*

$$\varrho_\infty(x,y) := \max_{i=1,\dots,n} |x_i - y_i| \tag{14.17}$$

which correspond to the largest difference of the coordinates of the two points. Another possible metric is the 1-distance

$$\rho_1(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{14.18}$$

All these metrics on $\mathbb{R}^n$ happen to be special elements of a one-parameter family of *p-metrics*. These are indexed by a real-valued parameter $p \in [1,\infty]$ and, for p finite, they are given by

$$\varrho_p(x,y) := \Big(\sum_{i=1}^{d} |x_i - y_i|^p\Big)^{1/p} \tag{14.19}$$

To see that the ∞-metric belongs to this family, we check that $\varrho_p(x,y)$ converges to $\varrho_\infty(x,y)$ as $p \to \infty$. (This requires concepts we have not yet talked about in detail.)

The $p = \infty$ case is sometimes referred to as a *box metric* as $\varrho_\infty(x,y)$ is the side of the smallest (square-shaped) box that fits both x and y. The case $p = 1$ is sometimes referred to as the *Manhattan distance*, or *taxicab distance*, as it corresponds to total distance traveled via a square grid aligned with coordinate axes. While (14.19) make sense for all $p > 0$, the triangle inequality holds only for $p \geqslant 1$.

In order to prove that ϱ_p are actually metrics (for $p \in [1,\infty]$) we use the fact that ϱ_p actually arises from a norm. The latter is a concept defined for all vector spaces:

Definition 14.9 (Norm) *Given a vector space V over a field F such that $F = \mathbb{R} \vee F = \mathbb{C}$, a function $\|\cdot\| \colon V \to \mathbb{R}$ is a* norm *if it satisfies the following requirements:*

(1) *(positivity)* $\forall u \in V \colon\ 0 \leqslant \|u\| \wedge (\|u\| = 0 \Leftrightarrow u = 0)$
(2) *(homogeneity)* $\forall u \in V\ \forall \lambda \in F \colon\ \|\lambda u\| = |\lambda|\|u\|$
(3) *(triangle inequality)* $\forall u, v \in V \colon\ \|u + v\| \leqslant \|u\| + \|v\|$.

A vector space V endowed with a norm $\|\cdot\|$ is called a normed space.

We then check:

Lemma 14.10 *Let $\|\cdot\|$ be a norm on vector space V. Then $\varrho \colon V \times V \to \mathbb{R}$ defined by*

$$\varrho(x, y) := \|x - y\| \tag{14.20}$$

is a metric on V.

Proof. The positivity of ϱ follows from the positivity of $\|\cdot\|$ while homogeneity of the latter implies symmetry via $\|x - y\| = \|(-1)(y - x)\| = \|y - x\|$. The triangle inequality for the norm along with the rewrite $x - y = (x - z) + (z - y)$ then gives the triangle inequality for ϱ. □

Noting that $\varrho_p(x, y) = \|x - y\|_p$, for $\|\cdot\|_p$ defined as in the next claim, all we need to do is to prove:

Proposition 14.11 (p-norms on $\mathbb{R}^n$) *Let $p \in \mathbb{R}$ obey $p \geqslant 1$. For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ let*

$$\|x\|_p := \Big(\sum_{i=1}^{n} |x_i|^p \Big)^{1/p} \tag{14.21}$$

Then $\|\cdot\|_p$ is a norm on $\mathbb{R}^n$.

Proof for $p = 1, 2, \infty$. The positivity and homogeneity of $\|\cdot\|_p$ are checked readily, so the main point is to prove the triangle inequality. For the 1-norm this comes from the triangle inequality for the absolute value:

$$\|x + y\|_1 = \sum_{i=1}^{n} |x_i + y_i| \leqslant \sum_{i=1}^{n} \big(|x_i| + |y_i| \big) = \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = \|x\|_1 + \|y\|_1 \tag{14.22}$$

For the ∞-norm this is even easier

$$\begin{aligned} \|x + y\|_\infty = \max_{i=1,\ldots,n} |x_i + y_i| &\leqslant \max_{i=1,\ldots,n} \big(|x_i| + |y_i| \big) \\ &\leqslant \max_{i=1,\ldots,n} |x_i| + \max_{i=1,\ldots,n} |y_i| = \|x\|_\infty + \|y\|_\infty \end{aligned} \tag{14.23}$$

For the 2-norm, we have to work a bit harder. First we write

$$\|x + y\|_2^2 = \sum_{i=1}^{n} (x_i + y_i)^2 + \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \tag{14.24}$$

We will need to bound the middle sum, which we accomplish by considering a similar rewrite but for the combination $x - \lambda y$ where λ is a generic scalar. This gives

$$0 \leqslant \|x - \lambda y\|_2^2 = \sum_{i=1}^{n} x_i^2 - 2\lambda \sum_{i=1}^{n} x_i y_i + \lambda^2 \sum_{i=1}^{n} y_i^2 \tag{14.25}$$

Viewing the right-hand side as a quadratic polynomial in λ, The fact that the left-hand side is non-negative means that this polynomial must have at most one real root. This in turn forces the discriminant to be non-positive which yields

$$4\Big(\sum_{i=1}^{n} x_i y_i\Big)^2 - 4\Big(\sum_{i=1}^{n} x_i^2\Big)\Big(\sum_{i=1}^{n} y_i^2\Big) \leqslant 0 \tag{14.26}$$

This now quickly translates into the so-called *Cauchy-Schwarz inequality*

$$\Big|\sum_{i=1}^{n} x_i y_i\Big| \leqslant \Big(\sum_{i=1}^{n} x_i^2\Big)^{1/2}\Big(\sum_{i=1}^{n} y_i^2\Big)^{1/2} \tag{14.27}$$

Using this for the middle term on the right of (14.24) we then get

$$\|x+y\|_2^2 \leqslant \left(\Big(\sum_{i=1}^{n} x_i^2\Big)^{1/2} + \Big(\sum_{i=1}^{n} y_i^2\Big)^{1/2}\right)^2 = \big(\|x\|_2 + \|y\|_2\big)^2 \tag{14.28}$$

Taking square roots we get the triangle inequality for $\|\cdot\|_2$. □

We will not give a proof for general $p \geqslant 1$ but only comment that its most difficult part (the triangle inequality) is the so called *Minkowski inequality* which is itself proved from the extension of the Cauchy-Schwarz inequality called the *Hölder inequality*

$$\Big|\sum_{i=1}^{n} x_i y_i\Big| \leqslant \Big(\sum_{i=1}^{n} x_i^p\Big)^{1/p}\Big(\sum_{i=1}^{n} y_i^q\Big)^{1/q} \tag{14.29}$$

where q is the number such that $\frac{1}{p} + \frac{1}{q} = 1$. This inequality is itself an extension of the AMGM (Arithmetic mean geometric mean) inequality.

As our last example of a metric we introduce the concept of *discrete metric* which is defined for any non-empty set A by

$$\varrho(x,y) := \begin{cases} 0, & \text{if } x = y, \\ 1, & \text{if } x \neq y, \end{cases} \tag{14.30}$$

This metric is not very useful in practice but is very good for theory building as it provides an easy test case for various facts about metric spaces.

14.3 Sequences in metric spaces.

We are now ready to go back to sequences taking values in a metric space and generalize the notions introduced in Definitions 14.3–14.4 as follows:

Definition 14.12 (Cauchy and convergent sequences) *Let (X, ϱ) be a metric space. We say that an X-valued sequence $\{x_n\}_{n\in\mathbb{N}}$ is*

- Cauchy *if* $\forall \epsilon > 0\, \exists n_0 \geqslant 0\, \forall n, m \geqslant n_0 \colon\ \varrho(x_n, x_m) < \epsilon$
- convergent *if* $\exists z \in X\, \forall \epsilon > 0\, \exists n_0 \geqslant 0\, \forall n \geqslant n_0 \colon\ \varrho(x_n, z) < \epsilon$

We call any such z a limit of $\{x_n\}_{n\in\mathbb{N}}$ and write $x_n \to z$ in this case.

We now observe a couple of general facts:

Lemma 14.13 (Uniqueness of the limit) *Any sequence has at most one limit. More precisely, for any metric space (X, ϱ) and any X-valued sequence $\{x_n\}_{n\in\mathbb{N}}$,*

$$\forall z, \tilde{z} \in X\colon\ x_n \to z \ \wedge\ x_n \to \tilde{z} \ \Rightarrow\ z = \tilde{z} \tag{14.31}$$

Proof. Suppose that $z, \tilde{z} \in X$ be such that $x_n \to z$ and $x_n \to \tilde{z}$. Given $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $n \geqslant n_0 \Rightarrow \varrho(x_n, z) < \epsilon/2$ and there exists $\tilde{n}_0 \in \mathbb{N}$ such that $n \geqslant n_0 \Rightarrow \varrho(x_n, z) < \epsilon/2$. Then for any $n \geqslant \max\{n_0, \tilde{n}_0\}$,

$$\varrho(z, \tilde{z}) \leqslant \varrho(z, x_n) + \varrho(\tilde{z}, x_n) < \epsilon/2 + \epsilon/2 = \epsilon \tag{14.32}$$

Since this holds for all $\epsilon > 0$, we must have $\varrho(z, \tilde{z}) = 0$ forcing $z = \tilde{z}$. □

The notation $\lim_{n\to\infty} x_n$ is often used to denote the (unique) limit of the sequence $\{x_n\}_{n\in\mathbb{N}}$. Note that the existence of the limit is implicit in this notation — we would not write this if the limit did not exist.

The two notions from Definition 14.12 are closely related:

Lemma 14.14 (Convergent implies Cauchy) *If $x_n \to x$ then $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy.*

Proof. Fix $\epsilon > 0$ and let $n_0 \in \mathbb{N}$ be such that for all $n \geqslant n_0$ we have $\varrho(x_n, x) < \epsilon/2$. By the triangle inequality we then get

$$\forall m, n \geqslant n_0\colon\ \varrho(x_m, x_n) \leqslant \varrho(x_m, x) + \varrho(x, x_n) < \epsilon/2 + \epsilon/2 = \epsilon. \tag{14.33}$$

This yields (2.4) and so $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy. □

For spaces with discrete metric we get even a full characterization of convergent and Cauchy sequences:

Lemma 14.15 (Cauchy/convergent sequences in discrete metric) *Consider any set X endowed with the discrete metric ϱ. Then for any sequence $\{x_n\}_{n\in\mathbb{N}}$ of points in X,*

$$\{x_n\}_{n\in\mathbb{N}} \text{ is Cauchy} \quad\Leftrightarrow\quad \exists n \in \mathbb{N}\ \forall m \geqslant n\colon\ x_m = x_n. \tag{14.34}$$

Every Cauchy sequence is thus eventually constant and all Cauchy sequences converge.

Proof. Since ϱ takes values 0 and 1, letting $\epsilon := 1/2$ in (2.4) forces $\varrho(x_n, x_m) = 0$ once $m, n \geqslant n_0$. The positivity axiom for d then shows that $x_n = x_m$ for all $m, n \geqslant n_0$. Eventually constant sequences are always convergent. □

The last example is of course very special; the notion of being Cauchy is actually weaker than being convergent. The sequence from (14.1) was Cauchy in $\mathbb{Q}$ but not convergent in $\mathbb{Q}$. The notions also depend on which metric we consider. Indeed, the $\mathbb{R}$-valued sequence $x_n := n$ is not Cauchy under the Euclidean metric (14.14) but it is Cauchy yet not convergent under the metric (14.15). We will spend considerable time discussing these aspects further.

As our final note, we observe that while the sequence $x_n := (-1)^n$ is neither convergent nor even Cauchy, restricting to even numbered indices gives us a constant, and thus a Cauchy and even convergent sequence. This naturally leads to:

Definition 14.16 *Given a sequence $\{x_n\}_{n\in\mathbb{N}}$ its* subsequence *is any sequence of the form $\{x_{n_k}\}_{k\in\mathbb{N}}$ where $\{n_k\}_{k\in\mathbb{N}}$ is a strictly increasing sequence taking values in $\mathbb{N}$.*

Informally, a subsequence arises via a selection of some values in the line-up of the whole sequence. This selection is usually driven by the desire to refine possible behaviors as the index becomes large. The following lemma was left to homework:

Lemma 14.17 *Let $\{x_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence. If $\{x_n\}_{n\in\mathbb{N}}$ admits a convergent subsequence, $x_{n_k} \to z$, then $\{x_n\}_{n\in\mathbb{N}}$ is convergent and $x_n \to z$.*

It follows that a non-convergent Cauchy sequence admits no convergent subsequence. Another criterion for convergent sequences is given in:

Lemma 14.18 (Urysohn's subsequence principle) *Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence in a metric space (X, ϱ) and let $z \in X$. Then the following are equivalent:*

(1) $x_n \to z$
(2) *every subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ contains a subsubsequence $\{x_{n_{k_j}}\}_{j\in\mathbb{N}}$ such that $x_{n_{k_j}} \to z$*

Proof. The implication (1) $\Rightarrow$ (2) is easy, so we focus on (2) $\Rightarrow$ (1). We will prove the contrapositive. Suppose that $\{x_n\}_{n\in\mathbb{N}}$ does not converge to z. Then there exists $\epsilon > 0$ such that

$$\forall n \in \mathbb{N}\, \exists m \in \mathbb{N}\colon\ m > n \ \wedge\ \varrho(x_m, z) \geqslant \epsilon \tag{14.35}$$

The recursion principle allows us to construct an increasing sequence $\{n_k\}_{k\in\mathbb{N}}$ of naturals such that

$$n_0 = 0 \ \wedge\ \forall k \in \mathbb{N}\colon n_{k+1} = \inf\big\{m \in \mathbb{N}\colon m > n_k \ \wedge\ \varrho(x_m, z) \geqslant \epsilon\big\} \tag{14.36}$$

where the set under the infimum is non-empty thanks to (14.35). But then $\varrho(n_k, z) \geqslant \epsilon$ for all $k \geqslant 0$ and so there is no subsubsequence $\{n_{k_j}\}_{j\in\mathbb{N}}$ of $\{n_k\}_{k\in\mathbb{N}}$ such that $x_{n_j} \to z$. □

While the double use of subsequences in the second part of the lemma may seem a bit convoluted, the lemma offers a tool to decide whether an abstract notion of convergence can possibly arise from a metric. (One use of this comes with the concept almost-everywhere convergence in measure theory for which (2) holds whenever we have convergence in measure yet (1) can still be false.)

15. BASIC TOPOLOGY

Having discussed the notion of Cauchy and convergent sequences, we now turn to the following natural questions:

- In what metric spaces or subsets thereof do all Cauchy sequences have a limit?
- What sets in a given metric space contain the limit of all convergent sequences contained therein.
- In what sets or spaces do sequences admit convergent subsequences.

These questions will ultimately be answered by the words *complete*, *closed* and *compact*, respectively. Here we develop the necessarily tools starting with metric spaces and then move to their generalizations using notions from *topology*.

15.1 Open balls and open sets.

We start with the basic definition:

Definition 15.1 (Open ball) *Let (X, ϱ) be a metric space. Given an $x \in X$ and a real number $r > 0$, the* open ball $B(x, r)$ *of radius $r > 0$ centered at $x \in X$ is the set*

$$B(x, r) := \big\{ y \in X \colon \varrho(x, y) < r \big\}. \tag{15.1}$$

Note that we have $x \in B(x, r)$ for all $r > 0$. We do not consider open balls for radii $r \leqslant 0$ as these are empty and thus not very interesting.

Before we start using the notion of open ball, it is instructive to check what the open balls look like in some of the basic examples of the metric spaces:

- *discrete metric*: As the metric takes only values 0 and 1, here we get

$$B(x, r) = \begin{cases} \{x\}, & \text{if } 0 < r \leqslant 1, \\ X, & \text{if } r > 1. \end{cases} \tag{15.2}$$

  In short, the ball is either a single point (namely, the center) or the whole space.
- *The real line*: Using the Euclidean metric $\varrho(x, y) := |x - y|$, we have

$$B(x, r) = (x - r, x + r) \tag{15.3}$$

  so Euclidean balls in $\mathbb{R}$ are simply open intervals. The same is true for the metric ϱ' from (14.15) although there the interval is no longer centered at x and no longer of (Euclidean) length $2r$ (verify this precisely!).
- *Euclidean space*: Consider the normed space $(\mathbb{R}^d, \|\cdot\|_p)$ for $p \in [1, \infty]$ and denote by $B_p(x, r)$ the open ball in $\mathbb{R}^d$ with respect to the norm-metric ϱ_p derived from $\|\cdot\|_p$; see (14.19) and (14.21). For $p = 2$, $B_2(x, r)$ is the usual, perfectly round, Euclidean ball. However, when $p = \infty$, we have

$$B_\infty(x, r) = \bigtimes_{i=1}^{d} (x_i - r, x_i + r) \tag{15.4}$$

  meaning that the ball in the ϱ_∞-metric is an open cube centered at x. (In this sense the ∞-metric is a more natural extension of (14.14) because, just as $\mathbb{R}^d$ is the Cartesian product of $\mathbb{R}$'s, the ∞-ball in $\mathbb{R}^d$ is the Cartesian product of (15.3).

For $p = 1$ the ball $B_1(x, r)$ takes a shape of a diamond centered at x. As p increases above 1, the corners of the diamond become rounded to become the Euclidean ball at $p = 2$. As p increases further, the p-ball gradually morphs to a cube.

With the notion of the open ball in hand, we now put forward:

Definition 15.2 (Open and closed sets) *Let (X, ϱ) be a metric space. A set $A \subseteq X$ is said to be* open *if*

$$\forall x \in A\, \exists r > 0 \colon B(x, r) \subseteq A, \tag{15.5}$$

i.e., if along with every point the set contains an open ball centered at that point. A set $A \subseteq X$ is said to be closed *if $X \setminus A$ is open.*

We remark that that latter already introduces closed sets using the method typical for topology. The text book uses a definition based on the concept of a *limit point* which we will show to be equivalent in the next section.

Here are some basic examples:

Lemma 15.3 *Every singleton is closed, i.e., $\forall x \in X \colon \{x\}$ is closed.*

Proof. This is equivalent to saying that, $\forall x \in X \colon X \setminus \{x\}$ is open. To prove that, let $y \in X \setminus \{x\}$. Since $y \neq x$, we have $\varrho(x, y) > 0$ so if we let $r := \varrho(x, y)$, then $r > 0$. But then $\varrho(x, y) = r$ and so $x \notin B(y, r)$ meaning that $B(y, r) \subseteq X \setminus \{x\}$. As this is true for every $y \in X \setminus \{x\}$, the set $X \setminus \{x\}$ is open and its complement $\{x\}$ is closed. □

Lemma 15.4 *Every open ball is open, i.e., $\forall x \in X\, \forall r > 0 \colon B(x, r)$ is open.*

Proof. This is small variation on the previous proof. Let $y \in B(x, r)$. Then $\varrho(x, y) < r$ and so $\epsilon := r - \varrho(x, y) > 0$. Let $z \in B(y, \epsilon)$. The triangle inequality then implies

$$\begin{aligned} \varrho(x, z) &\leqslant \varrho(x, y) + \varrho(y, z) \\ &< \varrho(x, y) + \epsilon = \varrho(x, y) + [r - \varrho(x, y)] = r \end{aligned} \tag{15.6}$$

and so $z \in B(x, r)$. It follows that $B(y, \epsilon) \subseteq B(x, r)$ and so $B(x, r)$ is open. □

Lemma 15.5 *Let ϱ be a discrete metric on X. Then every subset of X is open and closed.*

Proof. Let $A \subseteq X$ be arbitrary. Then for all $x \in A$ we have $B(x, 1/2) = \{x\} \subseteq A$ by (15.2) and so A is open. As this holds for all $A \subseteq X$, we get that $X \setminus A$ is open and so A is also closed, as claimed. □

The example of the discrete metric may be misleading it that it might make the reader believe that most (or even all) sets are either open or closed. But this is far from the truth in general; indeed, being open or closed is a very special property and most sets in general metric space are neither open nor closed. For instance, the interval $(0, 1]$ is neither open nor closed in $\mathbb{R}$ with respect to the usual metric. So, please beware that

$$A \text{ is NOT open} \quad \not\Rightarrow \quad A \text{ is closed} \tag{15.7}$$

and

$$A \text{ is open} \quad \not\Rightarrow \quad A \text{ is NOT closed} \tag{15.8}$$

In other words, the notions of open and closed sets are neither exhaustive (as other sets than these may exist) nor exclusive (as there could be sets that are both open and closed).

Lemma 15.4 naturally guides us to:

Definition 15.6 *Let (X, ϱ) be a metric space. Given $x \in X$ and $r \in \mathbb{R}$ with $r \geqslant 0$, the set*

$$\{y \in X \colon \varrho(y, x) \leqslant r\} \tag{15.9}$$

is called the closed ball *of radius r centered at x.*

The reader should check that the closed ball is indeed closed, justifying its name.

15.2 Topology.

We will now characterize the open sets in a metric space as follows:

Lemma 15.7 *Let (X, ϱ) be a metric space and set $\mathcal{T} := \{O \subseteq X \colon \text{open}\}$. Then*

(1) $\varnothing, X \in \mathcal{T}$
(2) $\forall A \subseteq \mathcal{T} \colon \bigcup A \in \mathcal{T}$
(3) $\forall A \subseteq \mathcal{T} \colon A \text{ finite } \Rightarrow \bigcap A \in \mathcal{T}$

In words, the set of open sets in a metric space (X, ϱ) contains $\varnothing$ and X and is closed under arbitrary unions and finite intersections.

Proof. (1) Since $\varnothing$ contains no points, it is trivially open (there is no x for which it would have to contain an open ball centered at x). Similarly, X is open as it by definition contains all open balls.

(2) Let A be a collection of open sets and let $x \in \bigcup A$. Then there is $O \in A$ such that $x \in O$. But O is open and so there is an $r > 0$ such that $B(x, r) \subseteq O$. It follows that

$$B(x, r) \subseteq O \subseteq \bigcup A \tag{15.10}$$

and so $\bigcup A$ is open.

(3) Let $A \subseteq \mathcal{T}$ be finite. This means that there is $n \in \mathbb{N}$ and a map that assigns each natural $k = 0, \ldots, n$ to a set $O_k \in A$ such that $A = \{O_k \colon k = 0, \ldots, n\}$. If $x \in \bigcap A$ then for each $k = 0, \ldots, n$ we have $x \in O_k$ and, since O_k is open, there is $r_k > 0$ such that $B(x, r_k) \subseteq O_k$. (Define $r_k := \sup\{r \in (0, 1] \colon B(x, r) \subseteq O_k\}$.) Now set

$$r := \min_{k \leqslant n} r_k \tag{15.11}$$

and note that, being the minimum of a finite number of positive reals, $r > 0$. As $r \leqslant r_k$, this shows

$$\forall k = 0, \ldots, n \colon B(x, r) \subseteq B(x, r_k) \subseteq O_k \tag{15.12}$$

But then $B(x, r) \subseteq \bigcap_{k=0}^{n} O_k = \bigcap A$ and so $\bigcap A$ is open. □

The properties of open sets in the previous lemma can be abstractized as:

Definition 15.8 (Topology) *Let X be a set. A collection $\mathcal{T} \subseteq \mathcal{P}(X)$ is said to be a* topology *on X if $\varnothing, X \in \mathcal{T}$ and $\mathcal{T}$ is closed under arbitrary unions and finite intersections.*

When a topology $\mathcal{T}$ is given, we refer to sets in $\mathcal{T}$ as *open* and, in accord with our earlier definition, call complements of open sets *closed*. The closed sets then obey:

Lemma 15.9 *The set of closed sets corresponding to a topology on X contains $\varnothing$ and X and is closed under arbitrary intersections and finite unions.*

Proof. With the help of de Morgan law

$$\forall A \subseteq \mathcal{P}(X): \quad X \smallsetminus \bigcap A = \bigcup \{X \smallsetminus C : C \in A\} \tag{15.13}$$

this follows directly from the corresponding properties of open sets. □

Every non-empty set X supports two topologies: first, the *coarsest* topology $\{\varnothing, X\}$ and the *finest* or *discrete* topology $\mathcal{P}(X)$. The name of the latter arises from the fact that this topology comes from a metric — namely, the discrete metric, thanks to Lemma 15.5 — and is thus *metrizable*. The coarsest topology does not come from a metric unless X is a singleton. (This is because all singletons are closed in every metric space.)

We now introduce the following concepts:

Definition 15.10 (Interior and closure) *Let $A \subseteq X$. The* interior of A *is then the set*

$$\operatorname{int}(A) := \bigcup \{O \subseteq X : \text{ open } \wedge\ O \subseteq A\}, \tag{15.14}$$

namely, the union of all open sets contained in A. The closure of A *is then the set*

$$\overline{A} := \bigcap \{C \subseteq X : \text{ closed } \wedge\ A \subseteq C\}, \tag{15.15}$$

namely, the intersections of all closed sets containing A.

Note that each $A \subseteq X$ contains at least one open set (namely, $\varnothing$) and is contained in at least one closed set (namely, X). Lemmas 15.5-15.9 then readily show

$$\forall A \subseteq X: \ \operatorname{int}(A) \text{ is open } \wedge\ \overline{A} \text{ is closed.} \tag{15.16}$$

We remark that other notations may be encountered in the literature for the interior (e.g., A°) and the closure (e.g., cl(A)). In addition, we have the following facts:

Lemma 15.11 *For each $A \subseteq X$,*

(1) $\operatorname{int}(A) \subseteq A \subseteq \overline{A}$,
(2) *A is open* $\Leftrightarrow A = \operatorname{int}(A)$,
(3) *A is closed* $\Leftrightarrow A = \overline{A}$.

Proof. (1) is the consequence of (15.14) and (15.15). In light of (15.16) we only have to prove $\Rightarrow$ in (2-3). But this again follows from (15.14) and (15.15): if A is open, then A is part of the collection in (15.14) and so $A \subseteq \operatorname{int}(A)$. By (1) we get $A = \operatorname{int}(A)$, proving $\Rightarrow$ in (2). If A is in turn closed, then A belongs to the collection in (15.15) and so $\overline{A} \subseteq A$. By (1) we then get $A = \overline{A}$, proving $\Rightarrow$ in (3). □

The interior of A is thus the largest open set contained in A while the closure of A is the smallest closed set containing A. Specializing to balls in a metric space (X, ϱ), for all $x \in X$ and $r \geqslant 0$ we thus have

$$\overline{B(x,r)} \subseteq \{y \in X : \varrho(x,y) \leqslant r\} \tag{15.17}$$

meaning that the closure of the open ball is contained in the closed ball. The inclusion is always strict when $r = 0$ but for the discrete metric it is strict even for $r = 1$.

Part (1) of Lemma 15.11 leads to another very natural concept:

Definition 15.12 (Topological boundary) *For each $A \subseteq X$, the set*

$$\partial A := \overline{A} \setminus \operatorname{int}(A) \tag{15.18}$$

is the (topological) boundary of A.

Note that writing

$$\overline{A} \setminus \operatorname{int}(A) = \overline{A} \cap (X \setminus \operatorname{int}(A)) \tag{15.19}$$

shows that

$$\forall A \subseteq X\colon\ \partial A \text{ is closed.} \tag{15.20}$$

Also note that while A may not be disjoint from ∂A, we always have

$$\operatorname{int}(A) \cap \partial A = \varnothing. \tag{15.21}$$

An intuitive image of the boundary of A is the "curve or surface enclosing A" but while this is true for nice subsets of the Euclidean space, it may fail miserably in general.

For instance, the boundary of the set can be the set itself (e.g., for $\mathbb{N}$ regarded as a subset of $\mathbb{R}$ with the Euclidean metric we have $\partial\mathbb{N} = \mathbb{N}$) or even much larger than that (e.g., for $\mathbb{Q} \subseteq \mathbb{R}$ where $\partial\mathbb{Q} = \mathbb{R}$). The boundary can also be empty, e.g., for the finest topology (which, as argued above, corresponds to the discrete metric of X) we have $\partial A = 0$ for each $A \subseteq X$. The boundary is also empty in spaces with multiple connected components; e.g., X being the union two disjoint open sets O_1 and O_2 for which $\partial O_i = \varnothing$.

Additional properties of interior and closure under the set inclusion and complementation are stated in:

Lemma 15.13 *For each $A, B \subseteq X$:*

$$A \subseteq B \quad \Rightarrow \quad \operatorname{int}(A) \subseteq \operatorname{int}(B) \quad \wedge \quad \overline{A} \subseteq \overline{B}. \tag{15.22}$$

Moreover, for each $A \subseteq X$,

$$X \setminus \operatorname{int}(A) = \overline{X \setminus A} \quad \wedge \quad X \setminus \overline{A} = \operatorname{int}(X \setminus A). \tag{15.23}$$

In particular, we have

$$\partial A = \overline{A} \cap \overline{X \setminus A} = \partial(X \setminus A). \tag{15.24}$$

We leave the proof of this lemma to an exercise. There is (quite naturally) no relation between the boundaries ∂A and ∂B whether A is a subset of B or not.

Describing the whole class of open sets in a given metric space is usually hard if not impossible. One example where it can be done is the real line with the usual metric:

Theorem 15.14 *Consider the metric space $(\mathbb{R}, d)$ where $\varrho(x, y) := |x - y|$. Then*

$$\forall A \subseteq \mathbb{R}\colon \qquad A \text{ open} \quad \Leftrightarrow \quad \begin{cases} \exists\{I_n\colon n \in \mathbb{N}\} \subseteq \mathcal{P}(\mathbb{R})\colon \\ \forall n \in \mathbb{N}\colon\ I_n = \varnothing \vee I_n = \text{ open interval} \\ \wedge\ \forall m, n \in \mathbb{N}\colon\ I_n \cap I_m \neq \varnothing \ \Rightarrow\ m = n \\ \wedge\ A = \bigcup\limits_{n \in \mathbb{N}} I_n \end{cases} \tag{15.25}$$

In words, every open set in $\mathbb{R}$ is a finite or countable union of disjoint open intervals.

We leave the proof of this interesting theorem to homework. The statement is very special to the one-dimensional Euclidean space. The open sets in $\mathbb{R}^d$ for $d \geqslant 2$ are much harder to characterize.

We finish with a cautionary remark that not all interesting properties of metric spaces can be relegated to topology. Indeed, it is easy to check that any subset of $\mathbb{R}$ that is open for the Euclidean metric is open for the metric (14.15), and *vice versa*. This is not too surprising in itself until we realize that, as shown in a homework exercise, these two metrics have *different* sets of Cauchy sequences. It follows that, while the topological point of view of metric spaces is useful in many ways, it is not good for studying the relation between Cauchy and convergent sequences. In particular, the notion of completeness to be introduced later is tied to the metric structure rather than topology.

16. Sequences and point-set topology

The previous section defines a number of concepts having to do with the topology (i.e., study of open and closed sets) in a metric space. (The phrase "point-set topology" is used here as our focus is on the points and sets thereof rather than the abstract notions of open and closed sets.) We will now link these to the notion based on open balls in the underlying metric, and then also to convergent sequences.

16.1 Point classification.

We being by introducing a classification of points relative to a given set using tools and objects from metric space theory — specifically, open balls:

Definition 16.1 *Let (X, ρ) be a metric space, $A \subseteq X$ and $x \in X$. We say that x is*

- *an adherent point of A if $\forall r > 0 \colon B(x,r) \cap A \neq \varnothing$*
- *a boundary point of A if $\forall r > 0 \colon B(x,r) \cap A \neq \varnothing \wedge B(x,r) \cap (X \setminus A) \neq \varnothing$*
- *an interior point of A if $\exists r > 0 \colon B(x,r) \subseteq A$*
- *an exterior point of A if $\exists r > 0 \colon B(x,r) \cap A = \varnothing$*
- *a limit point of A if $\forall r > 0 \colon (B(x,r) \cap A) \setminus \{x\} \neq \varnothing$*
- *an isolated point of A if $\exists r > 0 \colon B(x,r) \cap A = \{x\}$*

As it turns out, most of these are just different words for notions we already introduced using the notions from topology:

Lemma 16.2 *Let $A \subseteq X$. Then*

(1) $\{x \in X \colon \text{adherent point of } A\} = \overline{A}$
(2) $\{x \in X \colon \text{boundary point of } A\} = \partial A$
(3) $\{x \in X \colon \text{interior point of } A\} = \text{int}(A)$
(4) $\{x \in X \colon \text{exterior point of } A\} = \text{ext}(A) := \text{int}(X \setminus A) = X \setminus \overline{A}$

Moreover,

$$\overline{A} = \{x \in X \colon \text{limit point of } A\} \cup \{x \in X \colon \text{isolated point of } A\} \tag{16.1}$$

where the two sets in the union are disjoint.

Proof. (1) Let x be an adherent point of A and let C be a closed set with $A \subseteq C$. Then $x \notin X \setminus C$ for otherwise, by the fact that $X \setminus C$ is open, there would be $r > 0$ with $B(x,r) \subseteq X \setminus C$ implying $B(x,r) \cap A \subseteq B(x,r) \cap C = \varnothing$, a contradiction with x being adherent. Taking $C := \overline{A}$ we get $x \in \overline{A}$, proving

$$\{x \in X \colon \text{adherent point of } A\} \subseteq \overline{A}. \tag{16.2}$$

For the other inclusion, denote

$$O := X \setminus \{x \in X \colon \text{adherent point of } A\}. \tag{16.3}$$

Then for every $x \in O$, there is $r > 0$ such that $B(x,r) \cap A = \varnothing$ (otherwise x would be adherent). Since $B(x,r)$ is open, every point therein is separated by an open ball from A and so $B(x,r)$ contains no adherent points of A. This means that $B(x,r) \subseteq O$ and so O is open. Thus

$$\{x \in X \colon \text{adherent point of } A\} \text{ is closed.} \tag{16.4}$$

The fact that $x \in B(x,r)$ for all $r > 0$ shows that all the points in A are automatically adherent, and so we get $A \subseteq \{x \in X\colon \text{adherent point of } A\}$. Since the closure of A is the smallest closed set containing A, this yields

$$\overline{A} \subseteq \{x \in X\colon \text{adherent point of } A\} \tag{16.5}$$

Along with (16.2), this proves (1).

(2) By inspecting the definition of a boundary and adherent point, we readily check that a point is a boundary point if and only if it is adherent to A and to $X \setminus A$. Using (1) we thus get

$$\{x \in X\colon \text{boundary point of } A\} = \overline{A} \cap \overline{X \setminus A} \tag{16.6}$$

The claim now follows from (15.24).

(3-4) The definition of an interior point readily implies that

$$\{x \in X\colon \text{interior point of } A\} = X \setminus \{x \in X\colon \text{adherent point of } X \setminus A\} \tag{16.7}$$

As $\operatorname{int}(A) = X \setminus \overline{X \setminus A}$ by (15.23), the claim (3) follows from (1). The claim (4) is (3) applied to the complement of A.

It remains to prove (16.1). Note that, for all x, we have

$$\begin{aligned} &\forall r > 0\colon B(x,r) \cap A \neq \varnothing \\ &\quad \Leftrightarrow \big(\exists r > 0\colon B(x,r) \cap A = \{x\}\big) \\ &\qquad \vee \Big(\big(\forall r > 0\colon B(x,r) \cap A \neq \varnothing\big) \wedge \neg\big(\exists r > 0\colon B(x,r) \cap A = \{x\}\big)\Big) \end{aligned} \tag{16.8}$$

Using rules for negation of quantified clauses, the last line can be converted to

$$\forall r > 0\colon \big(B(x,r) \cap A \neq \varnothing \ \wedge\ B(x,r) \cap A \neq \{x\}\big) \tag{16.9}$$

This is equivalent to $\forall r > 0\colon (B(x,r) \setminus \{x\}) \cap A \neq \varnothing$, thus proving the claim. The fact that the decomposition in (16.1) is into disjoint sets is checked similarly. □

It is clear from the definition that an isolated point of A always belongs to A. However, the last argument in the previous proof allows us to characterize isolated and limits points further:

Lemma 16.3 *For all $A \subseteq X$ and all $x \in X$:*

$$\big(\exists r > 0\colon\ A \cap B(x,r) \text{ non-empty finite }\big) \ \Leftrightarrow\ x \text{ is isolated point of } A \tag{16.10}$$

In particular,

$$\forall x \in X\colon\ x \text{ is a limit point of } A \ \Leftrightarrow\ \forall r > 0\colon A \cap B(x,r) \text{ is infinite} \tag{16.11}$$

In short, each open ball centered at a limit point of A contains infinitely many points of A.

Proof. Let $x \in A$. If $A \cap B(x,r)$ is finite for some $r > 0$, then there is $n \in \mathbb{N}$ and a bijection $f\colon [0,n) \to A$. Denote $\forall k \in [0,n)\colon\ x_k := f(k)$ and set $r' := r$ if $n = 0$ and $r' := \min\{\rho(x,x_k)\colon k \in [0,n) \wedge x_k \neq x\}$ if $n > 0$. Then, as is readily checked, $A \cap B(x,r') = \{x\}$ and so x is an isolated point. The converse direction follows immediately from the definition of an isolated point. □

Corollary 16.4 *If $A \subseteq X$ is finite, then all points of A are isolated.*

Proof. Since singletons are closed and finite unions of closed sets are closed, if A is finite then it is closed. In particular, all points of A are adherent. Thanks to finiteness of A, the characterization (16.11) rules out limit points, so by (16.1), all points of A are isolated. (One can also prove this directly by setting $r := \min\{\varrho(x,y)\colon x,y \in A \wedge x \neq y\}$ and noting that then $\forall x \in A\colon A \cap B(x,r) = \{x\}$.) □

16.2 Sequential characterization of closedness.

Let us check out a few examples demonstrating the above notions. In all of these we take $X := \mathbb{R}$ with ϱ being the Euclidean metric.

- $A := \{\frac{1}{n+1}\colon n \in \mathbb{N}\}$. Here each point of A is isolated but $\overline{A} = A \cup \{0\}$ and 0 is a limit point of A. Every point of A lies in ∂A and $\operatorname{int}(A) = \varnothing$.
- $A := \{(-1)^n \frac{n}{n+1}\colon n \in \mathbb{N}\}$. Here, again, each point of A is isolated and $\overline{A} = A \cup \{+1,-1\}$ and the points $+1$ and -1 are limit points.
- $A := \{\frac{1}{n+1} + \frac{\sqrt{2}}{m+1}\colon n,m \in \mathbb{N} \wedge n \leqslant m\}$. Here $\{\frac{1}{n+1}\colon n \in \mathbb{N}\} \cup \{0\}$ are all the limit points while the remaining points of A are isolated.
- $A := \{n\sqrt{2} \bmod 1\colon n \in \mathbb{N}\}$. Here $A \subseteq [0,1]$ and $\overline{A} = [0,1]$. Every point of $[0,1]$ is a limit point of A. Still $\operatorname{int}(A) = \varnothing$.

We leave the proofs of the above claims to the reader.

Notice that in the examples we often relied on convergence of sequences from the set. As it turns out, this gives us another way to think of closed sets and closures:

Theorem 16.5 (AC) *Let (X,ϱ) be a metric space. Then for all sets $A \subseteq X$:*

$$A \text{ is closed} \quad \Leftrightarrow \quad \Big(\forall \{x_n\}_{n\in\mathbb{N}} \in A^{\mathbb{N}}\, \forall x \in X\colon x_n \to x \Rightarrow x \in A\Big) \tag{16.12}$$

In words, a set $A \subseteq X$ is closed if and only if all convergent sequences from A have a limit in A.

Proof. Let us start with the proof of $\Rightarrow$. Suppose $A \subseteq X$ is closed and let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence from A such that $x_n \to x$. If $x \notin A$ then x lies in $X \setminus A$ which is open and so there is $r > 0$ such that $B(x,r) \cap A = \varnothing$. But $x_n \to x$ means that $x_n \in B(x,r)$ when n is sufficiently large in contradiction with $x_n \in A$. Summarizing, $x \notin A$ implies $\neg(x_n \to x)$ which is equivalent to $x_n \to x \Rightarrow x \in A$, proving $\Rightarrow$ in (16.12).

Let us now consider the implication $\Leftarrow$ which we will again prove by proving the contrapositive. Suppose A is NOT closed. Then $X \setminus A$ is NOT open and so

$$\exists x \in X \setminus A\, \forall r > 0\colon B(x,r) \cap A \neq \varnothing. \tag{16.13}$$

This means that there exists $x \notin A$ that is adherent to A. Specializing (16.13) to r in the set $\{2^{-n}\colon n \in \mathbb{N}\}$, the Axiom of Choice yields

$$\bigtimes_{n\in\mathbb{N}} A \cap \big(B(x,2^{-n}) \setminus \{x\}\big) \neq \varnothing \tag{16.14}$$

meaning that there exists $f\colon \mathbb{N} \to A$ with $\forall n \in \mathbb{N}\colon f(n) \in B(x,2^{-n}) \setminus \{x\}$. Setting $x_n := f(n)$, we have $\rho(x,x_n) < 2^{-n}$ and so $x_n \to x$. Summarizing, assuming that A is NOT closed we showed that there exists an A-valued sequence $\{x_n\}_{n\in\mathbb{N}}$ and $x \in X$ such that $x_n \to x \wedge \neg(x \in A)$. This proves $\Leftarrow$ in (16.12) as desired. □

We used "AC" in the label of the theorem to mark that the proof required the Axiom of Choice. This is necessary when no additional structure is assumed about (X, ϱ) and A. However, in most spaces that we encounter in practice (e.g., when X is separable, see Definition 16.7 below) the choice of x_n can be performed constructively and a reference to the Axiom of Choice is no longer required.

Corollary 16.6 (AC)(Density of a set in its closure) *Let $A \subseteq X$. Then*

$$\forall x \in X\colon \quad x \in \overline{A} \;\Leftrightarrow\; \exists \{x_n\}_{n\in\mathbb{N}} \in A^{\mathbb{N}}\colon\ x_n \to x. \tag{16.15}$$

Proof. For $\Rightarrow$ use that, by Lemma 16.2, $\overline{A}$ is the set of adherent points and apply the argument after (16.13). For $\Leftarrow$ note that if $x_n \to x$ for some $\{x_n\}_{n\in\mathbb{N}} \in A^{\mathbb{N}}$, then x is an adherent point of A and so $x \in \overline{A}$, again by Lemma 16.2. □

The reason why we used the word "density" to label the property in Corollary 16.6 arises from:

Definition 16.7 *Let $A, B \subseteq X$. We say that a set B is* dense *in A if $A \subseteq \overline{B}$.*

Thus we also get:

Corollary 16.8 (AC) *Let $A, B \subseteq X$. Then B is dense in A if and only if for each $x \in A$ there exists a B-valued sequence $\{x_n\}_{n\in\mathbb{N}}$ such that $x_n \to x$.*

Typically, we will apply this to $A = X$ or $B \subseteq A$ (or both). A standard example of a dense subset of $\mathbb{R}$ (with the Euclidean metric) is $\mathbb{Q}$, although $\mathbb{R} \setminus \mathbb{Q}$ does as well. However, the former set demonstrates an important property of the reals:

Definition 16.9 (Separability) *We say that a metric space (X, ϱ) is* separable *if it contains a countable dense set, i.e., if*

$$\exists A \subseteq X\colon A \text{ countable } \wedge\ \overline{A} = X. \tag{16.16}$$

(This is one example where we do allow a finite A to be regarded as countable.)

The reals are thus separable. A homework exercise asks to show that the same applies to n-dimensional Euclidean spaces $\mathbb{R}^n$ under any norm metric. This extends even to some, but not all, infinite-dimensional generalizations thereof. For instance, $X := [0,1]^{\mathbb{N}}$ of $[0,1]$-valued sequences endowed with the metric

$$\rho\big(\{x_n\}_{n\in\mathbb{N}}, \{y_n\}_{n\in\mathbb{N}}\big) := \sup\big\{|x_n - y_n|\colon n \in \mathbb{N}\big\} \tag{16.17}$$

is NOT separable as X contains the uncountable set $\{0,1\}^{\mathbb{N}}$ whose all points are isolated. We leave a proof of non-separability to homework. (An easier non-example is any uncountable set endowed with the discrete metric.)

16.3 Relative notions.

Given a metric space (X, ϱ), associated with each (non-empty) $Y \subseteq X$ is a natural metric space (Y, ϱ_Y) where ϱ_Y is simply the restriction of ϱ to pairs of points from Y. We call ϱ_y the *induced metric*. Now define the following *relative* topological concepts:

Definition 16.10 (Relative open/closed set) *Let $Y \subseteq X$ be as above. We say that $A \subseteq Y$ is* relatively open *if A is open in (Y, ϱ_Y), and* relatively closed *if A is closed in (Y, ϱ_Y).*

In order to give an example, consider the following setting: $X := \mathbb{R}$ endowed with the standard Euclidean metric, $Y := \mathbb{Q}$. Then $A := [\sqrt{2}, \sqrt{3}] \cap \mathbb{Q}$ is relatively closed and relatively open in $\mathbb{Q}$ while it is neither open nor closed in $\mathbb{R}$. Same applies to the set $A := (\sqrt{2}, \sqrt{3}] \cap \mathbb{Q}$ as well as to $A := (\sqrt{2}, \sqrt{3}) \cap \mathbb{Q}$. This (of course) has to do with the fact that $\sqrt{2}, \sqrt{3} \notin \mathbb{Q}$; indeed, the set $A := [0,1) \cap \mathbb{Q}$ is neither relatively open nor relatively closed while $A := [0, \sqrt{2}) \cap \mathbb{Q}$ is relatively closed but not relatively open.

Another example to consider is $Y := [0,2)$. Then $A := [0,1)$ is relatively open and $A := [1,2)$ is relatively closed. These facts can be verified directly or by invoking the following general characterization of relatively open and closed sets:

Theorem 16.11 *Let (X, ϱ) be a metric space and let $Y \subseteq X$ be nonempty. Then*

$$\forall A \subseteq Y\colon \quad A \text{ relatively open } \Leftrightarrow \Big(\exists O \subseteq X\colon\ O \text{ open } \wedge\ A = O \cap Y\Big) \tag{16.18}$$

Similarly, we get

$$\forall A \subseteq Y\colon \quad A \text{ relatively closed } \Leftrightarrow \Big(\exists C \subseteq X\colon\ C \text{ closed } \wedge\ A = C \cap Y\Big) \tag{16.19}$$

In short, a set is relatively open/closed if and only if it is a restriction of an open/closed set.

Proof. For the purpose of this proof, let $B_Y(x,r) := \{y \in Y\colon \varrho(x,y) < r\}$ denote the open ball (Y, ϱ_Y) and let $B_X(x,r) := \{y \in X\colon \varrho(x,y) < r\}$ be the open ball in (X, ϱ). Note that

$$\forall x \in Y\ \forall r > 0\colon\ B_Y(x,r) = Y \cap B_X(x,r) \tag{16.20}$$

as is directly checked from the definition.

Let us start with "$\Rightarrow$" in (16.18). If $A \subseteq Y$ is a relatively open set, then it is open in (Y, ϱ_Y) which means that

$$\forall x \in A\ \exists r_x > 0\colon\ B_Y(x, r_x) \subseteq A. \tag{16.21}$$

(This r_x can be picked constructively; e.g., as $r_x := \frac{1}{2} \sup\{r \in (0,1]\colon B(x,r) \subseteq A\}$.) Set

$$O := \bigcup_{x \in A} B_X(x, r_x). \tag{16.22}$$

Since $B_X(x, r_x)$ is open in (X, ϱ) (see Lemma 15.4) and the union of a family of open sets is open, we have

$$O \text{ is open in } (X, \varrho). \tag{16.23}$$

Since $x \in B_X(x, r_x)$, we have $A \subseteq O$ and so

$$A \subseteq O \cap Y. \tag{16.24}$$

For the opposite inclusion note

$$O \cap Y = \bigcup_{x \in A} B_X(x, r_x) \cap Y = \bigcup_{x \in A} B_Y(x, r_x) \subseteq A \tag{16.25}$$

where we used (16.1) and then (16.21) at the very end. Combining (16.24) and (16.25) we get "$\Rightarrow$" in (16.18).

In order to prove "$\Leftarrow$" in (16.18), let $O \subseteq X$ be open with $A = O \cap Y$. Then for each $x \in A$ we have $x \in O$ and so there is $r > 0$ such that $B_X(x,r) \subseteq O$. But then (16.20) shows $B_Y(x,r) = B_X(x,r) \cap Y \subseteq O \cap Y = A$ proving that A is relatiely open.

The proof of (16.19) is handled by complementation. Indeed,

$$A \text{ relatively closed } \Leftrightarrow \ Y \smallsetminus A \text{ relatively open.} \tag{16.26}$$

By (5.1) the statement on the right-hand side is equivalent to the existence of $O \subseteq X$ open in (X, ϱ) such that $Y \smallsetminus A = O \cap Y$. But that is in turn equivalent to

$$A = Y \smallsetminus (O \cap Y) = Y \cap (X \smallsetminus O) \tag{16.27}$$

which is the right-hand side of (16.19) because $X \smallsetminus O$ is closed in (X, ϱ). □

We finish by noting that in topology, the relative notions are *defined* by the right-hand sides of (5.1–16.19). The relative topology on Y is then the projection of the topology on X by way of intersecting all sets by Y.

17. COMPLETENESS

Having discussed the connections between sequences and topology, we now turn back to sequences alone and examine the following basic question: In what metric spaces do all Cauchy sequences converge? We first answer this question for the reals endowed with the Euclidean metric and then treat general metric spaces.

17.1 Completeness of the reals.

As noted earlier, a special name is reserved for the metric spaces for which the above question is answered affirmatively:

Definition 17.1 (Completeness) *We say that a metric space (X, ϱ) is* complete *if every Cauchy sequence is convergent, i.e., if*

$$\forall \{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}\colon\ \{x_n\}_{n\in\mathbb{N}} \text{ Cauchy } \Rightarrow\ \exists x \in X\colon x_n \to x. \tag{17.1}$$

If the space is not complete, then we say it is incomplete.

Note that our earlier use of the term "complete" concerned the validity of the least-upper bound property in ordered fields. This is no loss in light of:

Theorem 17.2 *The metric space $(\mathbb{R}, \varrho)$, where $\varrho(x, y) := |x - y|$, is complete.*

The proof will require some facts about convergence of sequences which we will be useful throughout the rest of the course. The first lemma works for all metric spaces and is based on the following concept:

Definition 17.3 *Let (X, ϱ) be a metric space. A set $A \subseteq X$ is said to be* bounded *if it is contained in an open ball, i.e.,*

$$\exists x \in X\, \exists r > 0\colon\ A \subseteq B(x, r) \tag{17.2}$$

If a set is not bounded, then we call it unbounded.

We now observe:

Lemma 17.4 *Let (X, ϱ) be a metric space. Then*

$$\forall \{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}\colon\ \{x_n\}_{n\in\mathbb{N}} \text{ Cauchy } \Rightarrow\ \{x_n\colon n \in \mathbb{N}\} \text{ bounded} \tag{17.3}$$

Proof. Let $\{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ be a Cauchy sequence. Then (choosing $\epsilon := 1$ in Definition 14.12) there exists $n_0 \in \mathbb{N}$ such that

$$\forall n \geqslant n_0\colon\ \varrho(x_n, x_{n_0}) < 1 \tag{17.4}$$

Fix any $x \in X$. The triangle inequality then implies $\varrho(x, x_n) \leqslant \varrho(x, x_{n_0}) + 1$ for all $n \geqslant n_0$ and so we have

$$\forall n \in \mathbb{N}\colon\ \varrho(x, x_n) \leqslant \max\bigl\{\varrho(x, x_k)\colon k \in \mathbb{N} \ \wedge\ k \leqslant n_0\bigr\} + 1 \tag{17.5}$$

Denoting the number on the right as $\tilde{r}$, we thus have $\forall n \in \mathbb{N}\colon x_n \in B(x, \tilde{r} + 1)$, proving (17.3) with $r := \tilde{r} + 1$. □

Focusing now on sequences of reals, the next lemma calls upon the notions of "non-decreasing" and "strictly increasing" sequences from Definition 14.2:

Lemma 17.5 *Let $(A, \leqslant)$ be a totally ordered set. For each sequence $\{x_n\}_{n\in\mathbb{N}} \in A^{\mathbb{N}}$ there exists a strictly increasing sequence $\{n_k\}_{k\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ such that*

$$\forall k \in \mathbb{N}\colon x_{n_k} \leqslant x_{n_{k+1}} \ \vee \ \forall k \in \mathbb{N}\colon x_{n_{k+1}} < x_{n_k} \tag{17.6}$$

In words, each sequence in a totally ordered set contains a subsequence that is either non-decreasing or strictly decreasing.

Proof. Given $\{x_n\}_{n\in\mathbb{N}} \in A^{\mathbb{N}}$ let $J := \{n \in \mathbb{N}\colon (\forall j > n\colon x_j < x_n)\}$, where we set $a < b := a \leqslant b \wedge a \neq b$. If J is finite (empty or non-empty), then $\sup(J)$ exists (and equals 0 when $J = \varnothing$) and belongs to $\mathbb{N}$. Then we recursively construct $\{n_k\}_{k\in\mathbb{N}}$ so hat

$$n_0 = \sup(J) + 1 \ \wedge \ \forall k \in \mathbb{N}\colon n_{k+1} = \inf\{j > n_k\colon x_j \geqslant x_{n_k}\}, \tag{17.7}$$

where we note that $n_k \geqslant n_0$ by construction and so $\{j > n_k\colon x_j \geqslant x_{n_k}\} \neq \varnothing$ by the fact that $n_k \notin J$ as implied by $n_k \geqslant n_0 > \sup(J)$. Since n_{k+1} belongs to the set under infimum, we get $x_{n_{k+1}} \geqslant x_{n_k}$ for all $k \in \mathbb{N}$ thus proving the first alternative in (17.6).

If on the other hand J is infinite, then we construct $\{n_k\}_{k\in\mathbb{N}}$ so that

$$n_0 := \inf(J) \ \wedge \ \forall k \in \mathbb{N}\colon \ n_{k+1} = \inf\{j > n_k\colon x_j < x_{n_k}\} \tag{17.8}$$

where the infimum exists and belongs to the set on the right by Lemma 9.10 because that set is infinite for each $k \in \mathbb{N}$. This now readily gives $n_{k+1} > n_k$ and $x_{n_{k+1}} < x_{n_k}$ for all $k \in \mathbb{N}$, proving the second alternative in (17.6). □

Next we call on another important fact about monotone sequences of reals:

Lemma 17.6 (Bounded monotone sequence of reals converge) *Let $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be non-decreasing and bounded from above, i.e.,*

$$\exists c \in \mathbb{R}\, \forall n \in \mathbb{N}\colon \ x_n \leqslant x_{n+1} \ \wedge \ x_n \leqslant c \tag{17.9}$$

Then $\{x_n\}_{n\in\mathbb{N}}$ is convergent and, in fact,

$$\lim_{n\to\infty} x_n = \sup\{x_n\colon n \in \mathbb{N}\} \tag{17.10}$$

If $\{x_n\}_{n\in\mathbb{N}}$ is instead non-increasing (and bounded), then $\lim_{n\to\infty} x_n = \inf\{x_k\colon k \in \mathbb{N}\}$.

Proof. The assumptions (along with the least-upper bound property of $\mathbb{R}$) ensure that the supremum exists. Denote the supremum by L and let $\epsilon \in \mathbb{R}$ obey $\epsilon > 0$. Then $L - \epsilon$ is not an upper bound and so $\exists n_0 \in \mathbb{N}\colon L - \epsilon < x_{n_0}$. But then the monotonicity claim in (17.9) guarantees

$$\forall n \geqslant n_0\colon \ L - \epsilon < x_{n_0} \leqslant x_n \leqslant L < L + \epsilon \tag{17.11}$$

Noting that the extreme ends of these inequalities imply $|x_n - L| < \epsilon$, we have verified (14.5) for all $\epsilon > 0$ and thus proved (17.10). □

With the above lemmas in hand, we have proved a classical result discovered in 1817 by B. Bolzano in his proof of the Intermediate Value Theorem and flagged some 50 years later by K. Weierstrass as a result of independent interest:

Theorem 17.7 (Bolzano-Weierstrass theorem) *Every bounded sequence of reals contains a convergent subsequence.*

Proof. Let $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be a bounded sequence. By Lemma 17.5, there exists strictly increasing sequence $\{n_k\}_{k\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ such that the subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ is monotone. Being still bounded, this subsequence converges by Lemma 17.6. □

With these in hand, we are ready to give:

Proof of Theorem 17.2. Let $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be a Cauchy sequence. The sequence is then bounded by Lemma 17.4 and so it contains a convergent subsequence by Theorem 17.7. To conclude the claim, we thus need:

Lemma 17.8 *Let (X, ϱ) be a metric space and $\{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ a Cauchy sequence. If $\{x_n\}_{n\in\mathbb{N}}$ contains a convergent subsequence, then $\{x_n\}_{n\in\mathbb{N}}$ is itself convergent.*

The proof of this lemma is a homework exercise. □

The above demonstrates that the completeness property of the reals as an ordered field is essential for the completeness in the sense of metric spaces. The converse is actually true as well:

Lemma 17.9 *Let F be an ordered subfield of $\mathbb{R}$ which we regard as a metric space (F, ϱ) for the Euclidean metric $\varrho(x, y) = |x - y|$. Then*

$$(F, \varrho) \text{ complete} \quad \Leftrightarrow \quad F \text{ has least upper bound property} \tag{17.12}$$

In particular, no proper ordered subfield of $\mathbb{R}$ is complete in the Euclidean metric.

Proof. For "$\Leftarrow$" in (17.12), recall that every ordered field with least upper bound property is isomorphic to the reals. That $(\mathbb{R}, \varrho)$ is complete as a metric space was shown in Theorem 17.2. The implication "$\Rightarrow$" is left to a homework exercise. □

We remark that, in the previous lemma, the restriction to a subfield of the reals is necessary for the metric to take values in $\mathbb{R}$. (As noted in Section 10.5, there are ordered fields larger than $\mathbb{R}$ but in these the absolute value is not generally $\mathbb{R}$-valued.)

We also note that the arguments underpinning Theorem 17.2 depend crucially on the choice of the metric. And, indeed, as noted earlier, all convergent sequences for the reals with the Euclidean metric ϱ also converge in the metric ϱ' in (14.15), but the latter also admits $x_n := n$ as a non-convergent Cauchy sequence. So while $(\mathbb{R}, \varrho)$ is complete, $(\mathbb{R}, \varrho')$ is not. As noted in homework, this holds regardless of the fact that both metrics induce the same topology.

17.2 Completeness of Euclidean spaces.

There are many complete metric spaces. For instance, Lemma 14.15 shows that every discrete metric space is complete. Our interest is of course in metric space that are pertinent to analysis so our next step is the completeness of the Euclidean spaces.

Theorem 17.10 *Let $d \geqslant 1$ be a natural and let ϱ be a norm-metric on $\mathbb{R}^d$; i.e., $\varrho(x, y) := \|x - y\|$ for $\|\cdot\|$ a norm on $\mathbb{R}^d$. Then $(\mathbb{R}^d, \varrho)$ is complete.*

We start by a useful fact from linear algebra which is proved by analysis.

Proposition 17.11 *Let $d \geqslant 1$ be a natural and let $\|\cdot\|$ be a norm on on $\mathbb{R}^d$. Then*

$$\exists c, C > 0 \, \forall x \in \mathbb{R}^d : \; c\|x\|_\infty \leqslant \|x\| \leqslant C\|x\|_\infty \tag{17.13}$$

In short, all norms on $\mathbb{R}^d$ are comparable.

Proof. The upper bound in (17.13) is immediate: Writing $x = (x_1, \dots, x_d) = \sum_{i=1}^d x_k e_i$, where $e_1, \dots, e_d$ are the coordinate vectors in $\mathbb{R}^d$, the triangle inequality shows

$$\|x\| \leqslant \sum_{i=1}^n |x_i| \|e_1\| \leqslant \Big(\sum_{i=1}^n \|e_i\| \Big) \|x\|_\infty \tag{17.14}$$

so the upper bound in (17.13) holds with $C := \sum_{i=1}^n \|e_i\|$.

We will prove the lower bound with

$$c := \inf\{\|x\| \colon x \in \mathbb{R}^d \wedge \|x\|_\infty = 1\} \tag{17.15}$$

where the infimum exists because the set of reals on the right-hand side is non-empty and bounded below by zero. Our aim is to show that $c > 0$.

We start by noting that the existence of a "minimizing sequence" $\{x^{(n)}\}_{n\in\mathbb{N}} \in (\mathbb{R}^d)^{\mathbb{N}}$ such that

$$\big(\forall n \in \mathbb{N} \colon \|x^{(n)}\|_\infty = 1\big) \wedge \|x^{(n)}\| \to c \tag{17.16}$$

(There is no need to invoke the Axiom of Choice since $(\mathbb{R}^d, \rho_\infty)$ is separable.) Writing $x^{(n)} = (x_1^{(n)}, \dots, x_d^{(n)})$, the condition $\|x^{(n)}\|_\infty = 1$ shows that that the coordinate sequences $\{x_i^{(n)}\}_{n\in\mathbb{N}}$ are all bounded, i.e.,

$$\forall n \in \mathbb{N}\, \forall i = 1, \dots, d \colon |x_i^{(n)}| \leqslant 1 \tag{17.17}$$

Invoking the Bolzano-Weierstrass theorem, there exists a strictly increasing subsequence $\{n_k^{(1)}\}_{k\in\mathbb{N}}$ such that $\{x_i^{(n_k^{(1)})}\}_{k\in\mathbb{N}}$ is convergent. By induction we then prove that, for each $m = 2, \dots, d$, there exists a strictly increasing sequence $\{n_k^{(m)}\}_{k\in\mathbb{N}}$ which is a subsequence of $\{n_k^{(m-1)}\}_{k\in\mathbb{N}}$ such that $\{x_i^{(n_k^{(m)})}\}_{k\in\mathbb{N}}$ is convergent for all $i = 1, \dots, m$.

Now define $\hat{n}_k := n_k^{(d)}$. Then $\{\hat{n}_k\}_{k\in\mathbb{N}}$ is strictly increasing and $\{x_i^{(\hat{n}_k)}\}_{k\in\mathbb{N}}$ is convergent for each $i = 1, \dots, d$. This means we can define

$$\hat{x}_i := \lim_{k\to\infty} x_i^{(\hat{n}_k)} \tag{17.18}$$

and set $\hat{x} := (\hat{x}_1, \dots, \hat{x}_d)$. We now readily check that

$$\|x^{(\hat{n}_k)} - \hat{x}\|_\infty = \max\big\{|x_i^{(\hat{n}_k)} - \hat{x}_i| \colon i = 1, \dots, d\big\} \underset{k\to\infty}{\longrightarrow} 0 \tag{17.19}$$

which in light of (17.14) implies

$$\|x^{(\hat{n}_k)} - \hat{x}\| \underset{k\to\infty}{\longrightarrow} 0 \tag{17.20}$$

But the triangle inequality for the ∞-norm shows

$$\|x^{(\hat{n}_k)}\|_\infty - \|x^{(\hat{n}_k)} - \hat{x}\|_\infty \leqslant \|\hat{x}\|_\infty \leqslant \|x^{(\hat{n}_k)}\|_\infty + \|x^{(\hat{n}_k)} - \hat{x}\|_\infty \tag{17.21}$$

which via (17.17) and (17.19) yields $\|\hat{x}\|_\infty = 1$, and a similar argument for $\|\cdot\|$ gives

$$\|x^{(\hat{n}_k)}\| - \|x^{(\hat{n}_k)} - \hat{x}\| \leqslant \|\hat{x}\| \leqslant \|x^{(\hat{n}_k)}\| + \|x^{(\hat{n}_k)} - \hat{x}\| \tag{17.22}$$

implying $\|\hat{x}\| = c$ by (17.17) and (17.20). This rules out that $c = 0$ because that would force $\hat{x} = 0$ and thus $\|\hat{x}\|_\infty = 0$.

Having proved that $c > 0$ we now note that, since for any $x \neq 0$, the vector $z := \frac{1}{\|x\|_\infty} x$ obeys $\|z\|_\infty = 1$, we have

$$\forall x \in \mathbb{R}^d \smallsetminus \{0\}: \quad \|x\| = \|x\|_\infty \Big\| \frac{1}{\|x\|_\infty} x \Big\| \geqslant c\|x\|_\infty \tag{17.23}$$

proving the lower bound in (17.13). (For $x = 0$ this bound holds trivially.) □

As part of the previous proof, we have established the following facts:

Corollary 17.12 *All norm-metrics on $\mathbb{R}^d$ have the same Cauchy sequences and the same convergent sequences.*

Proof. That a sequence that is Cauchy (or convergent) in $\|\cdot\|$-metric is Cauchy (or convergent) in $\|\cdot\|_\infty$-metric and *vice versa* follows directly from (17.13). □

Corollary 17.13 *A sequence in $\mathbb{R}^d$ converges in any norm metric if and only if each coordinate thereof converges in the reals endowed with the Euclidean norm.*

Proof. The above proof shows this for the $\|\cdot\|_\infty$-metric; the extension to other norm metric then comes from (17.13). □

This now readily gives:

Proof of Theorem 17.10. If $\{x_n\}_{n\in\mathbb{N}} \in (\mathbb{R}^d)^{\mathbb{N}}$ is Cauchy in a norm metric ϱ, then it is Cauchy in $\|\cdot\|_\infty$-metric by Corollary 17.12. The argument following (17.19) then shows that the coordinate sequences are Cauchy in the reals endowed with the Euclidean metric. By Theorem 17.2, the coordinates converge and, by Corollaries 17.12 and 17.13, so does $\{x_n\}_{n\in\mathbb{N}}$ in $(\mathbb{R}^d, \varrho)$. □

We also record another fact proved above:

Corollary 17.14 (Bolzano-Weierstrass theorem in $\mathbb{R}^d$) *Every bounded sequence in $\mathbb{R}^d$ endowed with a norm-metric contains a convergent subsequence.*

Proof. This follows from Corollary 17.13 and Theorem 17.7. □

We emphasize that all of the above developments apply solely to norm-metrics; just as for the metric ϱ' on $\mathbb{R}$ not being complete, it is easy to come up with a metric ϱ'' on $\mathbb{R}^d$ that is not complete.

Another remark concerns the proof of Proposition 17.11. A reader might wonder how come that, when the upper bound is proved by essentially algebraic means, the lower bound requires analysis. To see that this is necessary, observe that if instead of $\mathbb{R}^2$ we work in the vector space $\mathbb{Q}^2$ over the field $\mathbb{Q}$ and a map $x \mapsto \|x\|$ such that

$$\forall x = (x_1, x_2) \in \mathbb{Q}^2: \quad \|x\| = |x_1 + x_2\sqrt{2}| \tag{17.24}$$

It is easy to check that this is a norm, where the only subtle point is strict positivity which again boils down to the fact that there are no rationals $a, b \in \mathbb{Q}$ such that $a + b\sqrt{2} = 0$. Yet, for this norm, c in (17.15) vanishes and the lower bound in (17.13) fails for this norm because there exists a sequence $\{b_n\}_{n\in\mathbb{N}} \in \mathbb{Q}^{\mathbb{N}}$ such that $|b_n| < 1$ and $b_n \to 1/\sqrt{2}$ for which $x_n := (1, b_n)$ obeys $\|x_n\| \to 0$ while $\|x_n\|_\infty = 1$. The comparability of the norms is thus subtly tied to the completeness of $\mathbb{R}$ and $\mathbb{R}^d$ for all $d \geqslant 1$.

We also note that while the conclusion of Proposition 17.11 extends to all finite-dimensional vector spaces (as these are isomorphic with $\mathbb{R}^d$ for d being their dimension), the conclusion *fails* in infinite-dimensional generalizations thereof. For instance, if we consider the space $\mathbb{R}^{\mathbb{N}}$ of real-valued sequences, we can try to consider the generalization of the norm from (14.21)

$$\|x\|_p := \left(\sum_{n\in\mathbb{N}} |x_n|^p\right)^{1/p} \tag{17.25}$$

where the infinite sum is interpreted via Definition 13.3. The *Minkowski inequality* shows that these are norms whenever $p \in [1,\infty)$. However, the problem is that the expression is not finite for all elements of $\mathbb{R}^{\mathbb{N}}$ but rather only for a subset thereof defined by

$$\ell^p(\mathbb{N}) := \left\{\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \colon \sum_{n\in\mathbb{N}} |x_n|^p < \infty\right\} \tag{17.26}$$

Then $\|\cdot\|_p$ is a norm on $\ell^p(\mathbb{N})$ and, as it turns out, the space is complete in the associated metric. Yet, while $p < q$ implies $\ell^p(\mathbb{N}) \subseteq \ell^q(\mathbb{N})$, the inclusion is always strict as seen, e.g., by taking $x = \{x_n\}_{n\in\mathbb{N}}$ with $x_n := (n+1)^{-\frac{2}{p+q}}$ which belongs to $\ell^q(\mathbb{N})$ but not to $\ell^p(\mathbb{N})$. Consequently, the norms $\|\cdot\|_p$ and $\|\cdot\|_q$ are not comparable either.

18. CONTRACTION MAPS AND COMPLETION

Here we continue discussing completeness albeit now for general metric spaces. For spaces that are not complete, we introduce the notion of their completion. As it turns out, this will give us yet another construction of the reals.

18.1 Completness and its consequences.

Complete metric spaces have a number of attractive properties that makes working with them more convenient. We start by making some general observations about complete spaces. The first one relates completeness to closedness:

Lemma 18.1 (AC)(Inheritance to closed subsets) *Let (X, ϱ) be a complete metric space and, given $A \subseteq X$, let ϱ_A be the metric induced on A. Then for all non-empty $A \subseteq X$,*

$$(A, \varrho_A) \text{ complete} \quad \Leftrightarrow \quad A \text{ closed.} \tag{18.1}$$

Proof. Let $\{x_n\}_{n\in\mathbb{N}} \in A^{\mathbb{N}}$. Then $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy in (X, ϱ) is equivalent to $\{x_n\}_{n\in\mathbb{N}}$ being Cauchy in (A, ϱ_A) so all Cauchy sequences in (A, ϱ_A) converge to some point in X. By Theorem 16.5 (which requires AC), this point is in A for all such sequences if and only if A is closed. □

Another type of inheritance concerns Cartesian products. Here we note that if (X, ϱ_X) and (Y, ϱ_Y) be metric spaces, then $\varrho\colon X \times Y \to \mathbb{R}$ defined by

$$\varrho_\infty\big((x, y), (\tilde{x}, \tilde{y})\big) = \max\big\{\varrho_X(x, \tilde{x}), \varrho_Y(y, \tilde{y})\big\} \tag{18.2}$$

is a metric on $X \times Y$. We write the infinity symbol because the ∞-norm on $\mathbb{R}^2$ is used implicitly to combine the two metrics into one. If instead another norm (e.g., the Euclidean norm) was used, we would get another metric, which by Proposition 17.11 turns out to be equivalent to ϱ_∞ according to the following definition:

Definition 18.2 (Equivalent metrics) *Let ϱ and ϱ' be two metrics on X. We say that ϱ and ϱ' are* equivalent *if*

$$\exists c, C > 0\, \forall x, y \in X\colon \;\; c\varrho(x, y) \leqslant \varrho'(x, y) \leqslant C\varrho(x, y) \tag{18.3}$$

We leave it to the reader to check:

Lemma 18.3 *Equivalent metrics have the same Cauchy and convergent sequences, as well as the same induced topologies.*

We then put forward:

Definition 18.4 *Let (X, ϱ_X) and (Y, ϱ_Y) be metric spaces. The associated* product metric space *is the space $(X \times Y, \varrho)$ where ϱ is any metric equivalent to ϱ_∞ in* (18.2).

We then have:

Lemma 18.5 (Inheritance under Cartesian products) *If (X, ϱ_X) and (Y, ϱ_Y) are complete, then so is the product metric space $(X \times Y, \varrho)$, for any metric ϱ that is equivalent to ϱ_∞ in* (18.2).

Proof. Let ϱ be a metric on $X \times Y$ that obeys $c\varrho(\cdot,\cdot) \leqslant \varrho_\infty(\cdot,\cdot) \leqslant C\varrho(\cdot,\cdot)$. Let $\{(x_n, y_n)\}_{n\in\mathbb{N}}$ be a Cauchy sequence in $(X \times Y, \varrho_\infty)$. Since

$$\varrho_X(x_n, x_m) \leqslant \varrho_\infty\big((x_n, y_n), (x_m, y_m)\big) \leqslant C\varrho\big((x_n, y_n), (x_m, y_m)\big) \tag{18.4}$$

also $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy in X and, by the same argument, $\{y_n\}_{n\in\mathbb{N}}$ is Cauchy in Y. The assumed completeness implies existence of $x \in X$ and $y \in Y$ such that $x_n \to x$ and $y_n \to y$. Then $\varrho_X(x_n, x) \to 0$ and $\varrho_Y(y_n, y) \to 0$, which implies $\varrho_\infty((x_n, y_n), (x, y)) \to 0$ and thus also $\varrho((x_n, y_n), (x, y)) \to 0$. Hence $(x_n, y_n) \to (x, y)$ and $(X \times Y, \varrho)$ is thus complete. □

Having noted that completeness inherits nicely downward and upward, we now move to one important practical consequence of completeness, which is the fact that maps that contract distances admit a fixed point. We start with:

Definition 18.6 (Contraction map) *Let (X, ϱ) be a metric space. A map $\phi\colon X \to X$ (with $\mathrm{Dom}(\phi) = X$) is a* contraction *if*

$$\exists c \in \mathbb{R}\colon \quad 0 \leqslant c < 1 \ \wedge\ \forall x, y \in X\colon\ \varrho\big(\phi(x), \phi(y)\big) \leqslant c\varrho(x, y) \tag{18.5}$$

The fact that $c < 1$ is crucial for this notion. That being said, we warn the reader that the terminology is broken because linear operators (i.e., linear maps of linear spaces) are called contractions if the above holds with $c = 1$. Using the above definition, we have:

Theorem 18.7 (Banach's contraction principle) *Let (X, ϱ) be a complete metric space and let $\phi\colon X \to X$ be a contraction map as in* (18.5)*. Then*

$$\exists x \in X\colon\ \phi(x) = x \tag{18.6}$$

meaning that ϕ admits a fixed point. Moreover, the fixed point is unique,

$$\forall x, y \in X\colon\ \big(\phi(x) = x \wedge \phi(y) = y\big) \ \Rightarrow\ x = y \tag{18.7}$$

In words, a contraction on a complete metric space has a unique fixed point.

Proof. Let $c \in \mathbb{R}$ and $\phi\colon X \to X$ be a contraction such that (18.5) holds. Pick $x \in X$ and use recursion to construct $\{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ so that

$$x_0 = x \ \wedge\ \forall n \in \mathbb{N}\colon\ x_{n+1} = \phi(x_n) \tag{18.8}$$

We claim that

$$\forall n \in \mathbb{N}\colon\ \varrho(x_n, x_{n+1}) \leqslant c^n \varrho(x_0, x_1). \tag{18.9}$$

This is proved by induction: Let P_n be the statement after the quantifier. Then P_0 holds trivially because $c^0 = 1$ and if P_n, then the contraction property (18.5) implies

$$\varrho(x_{n+1}, x_{n+2}) = \varrho\big(\phi(x_n), \phi(x_{n+1})\big) \leqslant c\varrho(x_n, x_{n+1}) \overset{P_n}{\leqslant} c \cdot c^n \varrho(x_0, x_1) \tag{18.10}$$

showing $P_n \Rightarrow P_{n+1}$ with the help of $c^{n+1} = c \cdot c^n$. Hereby we get (18.9) via Lemma 4.3.

Next we upgrade (18.9) into

$$\forall n, m \in \mathbb{N}\colon\ n \leqslant m \ \Rightarrow\ \varrho(x_n, x_m) \leqslant \frac{c^n - c^m}{1 - c}\varrho(x_0, x_1). \tag{18.11}$$

We again prove this by induction, this time on m. Let P_m be the logical sentence that the inequality holds for all $n \in \mathbb{N}$ satisfying $n \leqslant m$. The base case P_0 is checked immediately,

because then the only non-trivial value is $n = 0$ for which the distance on the left vanishes while the right-hand side is non-negative because $c^m \leqslant c^n$ thanks to $n \leqslant m$. If P_m is TRUE, then for any $n \leqslant m$ (18.9) gives

$$\begin{aligned} \varrho(x_n, x_{m+1}) &\leqslant \varrho(x_n, x_m) + \varrho(x_m, x_{m+1}) \\ &\leqslant \frac{c^n - c^m}{1-c}\varrho(x_0, x_1) + c^m\varrho(x_0, x_1) = \frac{c^n - c^{m+1}}{1-c}\varrho(x_0, x_1). \end{aligned} \tag{18.12}$$

As for case $n = m + 1$ the clause P_{m+1} holds trivially, we get $P_m \Rightarrow P_{m+1}$ and so (18.11) is TRUE as stated by Lemma 4.3.

Dropping the c^m term from the numerator of (18.11) shows

$$\forall n, m \in \mathbb{N}\colon\ n \leqslant m \ \Rightarrow\ \varrho(x_n, x_m) \leqslant \frac{\varrho(x_1, x_0)}{1-c}c^n \tag{18.13}$$

Lemma 14.7 then gives

$$\forall \epsilon > 0\, \exists n_0 \in \mathbb{N}\, \forall n \geqslant n_0\colon\ \frac{\varrho(x_1, x_0)}{1-c}c^n < \epsilon \tag{18.14}$$

proving that $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy. By the assumed completeness of (X, ϱ), there is $x \in X$ such that $x_n \to x$. Then

$$\begin{aligned} \varrho\big(\phi(x), x\big) &\leqslant \varrho\big(\phi(x), x_{n+1}\big) + \varrho(x, x_{n+1}) \\ &= \varrho\big(\phi(x), \phi(x_n)\big) + \varrho(x, x_{n+1}) \leqslant c\varrho(x, x_n) + \varrho(x, x_{n+1}) \end{aligned} \tag{18.15}$$

and since both terms on the right tend to zero, we get $\varrho(\phi(x), x) = 0$ implying $\phi(x) = x$ as desired. The fixed point is unique because if x and y are both fixed points, then $\varrho(x, y) = \varrho(\phi(x), \phi(y)) \leqslant c\varrho(x, y)$ which forces $\varrho(x, y) = 0$ and thus $x = y$. □

We remark that completeness is absolutely crucial for the fixed point to exists. This is seen by taking $X := \mathbb{Q}^+ = \{a \in \mathbb{Q}\colon a \geqslant 0\}$ with Euclidean metric and letting $x_0 := 1$ and $\phi(x) := \sqrt{3 + x}$. The iterates defined by $x_{n+1} = \phi(x_n)$ then converge to $\frac{1+\sqrt{13}}{2}$ which does not lie in $\mathbb{Q}$. The same argument shows that the map ϕ has no fixed points in $\mathbb{Q}$.

Another name for Theorem 18.7 is *Banach's fixed point theorem*. While we will not give applications of the above theorem at this time, we note that Theorem 18.7 finds many practical uses most of which, however, are phrased using terms (such as the space of continuous functions) that we do not yet have the tools to discuss here.

Note also that the proof actually suggests an algorithm for constructing the fixed point: Iterate the map successively starting from an arbitrary point. This is in fact how this method is often used in practice; for instance, when constructing solutions of differential equations by way of so called *Picard iterations*.

18.2 Intrinsic closedness.

Let us now move to more abstract aspects of completness. As noted in Lemma 18.1, completeness is somehow analogous to (sequential characterization of) closedness, albeit with convergent sequences replaced by Cauchy sequences. We will now expound on this connection further. We need:

Definition 18.8 (Isometry) *Let (X, ϱ_X) and (Y, ϱ_Y) be metric spaces. We say that the map $\phi\colon X \to Y$ is an* isometry *if*

$$\forall x, y \in X\colon \quad \varrho_Y\big(\phi(x), \phi(y)\big) = \varrho_X(x, y). \tag{18.16}$$

If $\mathrm{Dom}(\phi) = X$ we say that ϕ is an isometric embedding of X into Y.

Unlike contractions that shrink distances, isometries preserve them. An example of an isometry is a translation of $\mathbb{R}$ by a fixed amount or rotation of $\mathbb{R}^2$ by a fixed angle (around any given point). The use of the word "embedding" refers to the fact that the isometric embedding allows us to realize X as a subset of Y. The distance-preserving property implies:

Lemma 18.9 *Any isometry is automatically injective on its domain.*

Proof. Let $\phi\colon X \to Y$ be an isometry and suppose that $x, y \in \mathrm{Dom}(\phi)$ are such that $\phi(y) = \phi(x)$. Then (18.16) implies $\varrho_X(x, y) = 0$ which by the separation axiom for the metric gives $x = y$. □

Not all isometries are necessarily onto, of course. For instance, $(\mathbb{R}, \varrho)$ with $\varrho(x, y) := |x - y|$ embeds isometrically into $(\mathbb{R}^d, \varrho_2)$ yet the embedding is not surjective. We thus introduce another qualifier:

Definition 18.10 *An isometric isomorphism (a.k.a. bijective isometry) is an isometry which is onto.*

The reason for the word isomorphis is that two spaces related by an isometric isomorphism are indistinguishable as far as their metric properties are concerned. Relating two metric spaces by an isomorphism is thus saying that they are basically the same. We now characterize complete metric spaces by closedness of their isometric embedding in other complete spaces:

Theorem 18.11 (AC)(Intrinsic closedness of complete spaces) *Let (X, ϱ) be a metric space. Then the following are equivalent:*

(1) *(X, ϱ) is complete*
(2) *$\forall (Y, \varrho')$ complete $\forall \phi\colon X \to Y$ isometric embedding: $\phi(X)$ is closed in (Y, ϱ')*

In words, a space is complete if and only if its embedding into any complete space is closed.

Proof of (1) $\Rightarrow$ (2). Assume that (X, ϱ) and (Y, ϱ') are complete and let $\phi\colon X \to Y$ be an isometric embedding. Consider a sequence $\{y_n\}_{n\in\mathbb{N}} \in \phi(X)^{\mathbb{N}}$ such that $y_n \to y$, for some $y \in Y$. Then there is $\{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ such that $\phi(x_n) = y_n$ for each $n \in \mathbb{N}$. The assumed convergence implies that $\{y_n\}_{n\in\mathbb{N}}$ is Cauchy and, using that ϕ is an isometry we readily check that so is $\{x_n\}_{n\in\mathbb{N}}$. By completeness of (X, ϱ) there exists $x \in X$ such that $x_n \to x$. The isometry property now shows that $y_n = \phi(x_n) \to \phi(x)$. The uniqueness of the limit then gives $y = \phi(x)$, implying $y \in \phi(X)$. By Theorem 16.5, $\phi(X)$ is closed. □

The term *intrinsic* has been used to emphasize that a complete space embeds isometrically into any complete space (for which such an embedding exists) as a closed set. That (1) and (2) are equivalent means that for incomplete spaces this fails for *all* isometric embeddings (not just one) into any complete space.

The proof of the opposite implication is harder as it requires the introduction (and construction of an instance) of the following concept:

Definition 18.12 (Completion) *Let (X, ϱ) be a metric space. A* completion of X *is any metric space $(\overline{X}, \overline{\varrho})$ such that*

(1) *$(\overline{X}, \overline{\varrho})$ is complete, and*
(2) *$\exists \phi\colon X \to \overline{X}$ isometric embedding : $\overline{\phi(X)} = \overline{X}$.*

Here, in (2), the closure of $\phi(X)$ is in the metric space $(\overline{X}, \overline{\varrho})$.

A few remarks are in order:

- The notation $\overline{X}$ has, *a priori*, nothing to do with closure; it is just a notation for the completion. However, in light of (2), it is in fact a sort of a closure as (2) says that there is an embedding of X is which the closure of X is all of $\overline{X}$.
- Condition (2) is a minimality condition. Indeed, we already noted that $(\mathbb{Q}, \varrho)$ can be embedded into $(\mathbb{R}, \varrho)$, which is complete and is in fact the closure of $(\mathbb{Q}, \varrho)$, but also into $(\mathbb{R}^d, \varrho_p)$ for any $d \geqslant 1$ and any $p \in [1, \infty]$. We would not want to regard the latter spaces as the completion of $(\mathbb{Q}, \varrho)$.
- Using an earlier definition, condition in (2) means that $\phi(X)$ is dense in $\overline{X}$.

We now claim:

Theorem 18.13 *For each metric space there is at least one completion.*

Leaving the proof to the next subsection, we note that this is enough to give the proof of the opposite implication in Theorem 18.11:

Proof of (2) $\Rightarrow$ (1) *in Theorem 18.11.* Consider the complete space $(Y, \varrho') := (\overline{X}, \overline{\varrho})$ and let ϕ be the isometric embedding of X into $\overline{X}$. If $\phi(X)$ is closed then $\phi(X) = \overline{X}$ and ϕ is thus onto. Then X is isometric to a complete space and so it is thus complete. □

18.3 Existence of a completion.

We now move to the proof of Theorem 18.13. The argument builds on Cantor's 1878 proof of existence of a system reals which was based on the fact that one way to think of a real number as a Cauchy sequence of rationals.

While such a representation is quite natural, a number of conceptual problems arise in its rigorous implementation. The first one is that many sequences of rationals converge to the same real number. We thus somehow need to find a way to identify the sequences with the same limit as one. This will be done by grouping them into equivalence classes under a suitable equivalence relation. Another problem is that, before the reals are actually constructed, some Cauchy sequences of rationals may not converge because there is no limit point for them to converge to. Instead of convergent sequences, we should thus rather focus on Cauchy sequences.

We now move to implementing this strategy in the context of a general metric space (X, ϱ). We start with by grouping Cauchy sequences in equivalence classes. Given two Cauchy sequences $\{x_n\}_{n\in\mathbb{N}}, \{y_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$, we set

$$\{x_n\}_{n\in\mathbb{N}} \sim \{y_n\}_{n\in\mathbb{N}} := \lim_{n\to\infty} \varrho(x_n, y_n) = 0 \tag{18.17}$$

We leave it to the reader to check that this is a reflexive, symmetric and transitive relation on the set of Cauchy sequences, and thus is an equivalence relation. The equivalence class of associated with a Cauchy sequence $\{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ is then

$$\big[\{x_n\}_{n\in\mathbb{N}}\big] := \Big\{\{y_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}} \colon \lim_{n\to\infty} \varrho(x_n, y_n) = 0\Big\} \tag{18.18}$$

We leave it to the reader to check the easy consequences of above definitions:

Lemma 18.14 *For any Cauchy sequence $\{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$,*

(1) $\{x_n\}_{n\in\mathbb{N}} \in \big[\{x_n\}_{n\in\mathbb{N}}\big]$,
(2) $\forall \{y_n\}_{n\in\mathbb{N}} \in \big[\{x_n\}_{n\in\mathbb{N}}\big] \colon \{y_n\}_{n\in\mathbb{N}}$ *is Cauchy,*
(3) $\forall \{y_n\}_{n\in\mathbb{N}}, \{\tilde{y}_n\}_{n\in\mathbb{N}} \in \big[\{x_n\}_{n\in\mathbb{N}}\big] \colon \varrho(y_n, \tilde{y}_n) \to 0$ *and so*

$$\forall \{y_n\}_{n\in\mathbb{N}} \in \big[\{x_n\}_{n\in\mathbb{N}}\big] \colon \big[\{y_n\}_{n\in\mathbb{N}}\big] = \big[\{x_n\}_{n\in\mathbb{N}}\big] \tag{18.19}$$

With these in hand, we set

$$\overline{X} := \Big\{[\{x_n\}_{n\in\mathbb{N}}] \in X^{\mathbb{N}} \colon \{x_n\}_{n\in\mathbb{N}} \text{ Cauchy}\Big\} \tag{18.20}$$

Our next goal is to define a metric on $\overline{X}$. To this end, we recall a lemma based on an exercise from homeowork:

Lemma 18.15 *For any two Cauchy sequences $\{x_n\}_{n\in\mathbb{N}}, \{\tilde{x}_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$,*

$$\{\varrho(x_n, \tilde{x}_n)\}_{n\in\mathbb{N}} \text{ is Cauchy} \tag{18.21}$$

and so

$$\lim_{n\to\infty} \varrho(x_n, \tilde{x}_n) \text{ exists in } \mathbb{R} \tag{18.22}$$

Moreover, for all $\{y_n\}_{n\in\mathbb{N}} \in \big[\{x_n\}_{n\in\mathbb{N}}\big]$ and all $\{\tilde{y}_n\}_{n\in\mathbb{N}} \in \big[\{\tilde{x}_n\}_{n\in\mathbb{N}}\big]$,

$$\lim_{n\to\infty} \varrho(y_n, \tilde{y}_n) = \lim_{n\to\infty} \varrho(x_n, \tilde{x}_n) \tag{18.23}$$

and so the limit depends only on the equivalence classes of the sequences.

Proof. In order to get (18.21), let $n, m \in \mathbb{N}$ and note that

$$\big|\varrho(x_m, \tilde{x}_m) - \varrho(x_n, \tilde{x}_n)\big| \leqslant \varrho(x_m, x_n) + \varrho(\tilde{x}_m, \tilde{x}_n) \tag{18.24}$$

Thanks to the Cauchy property of the sequences, given $\epsilon > 0$ there is $n_0 \in \mathbb{N}$ such that both terms on the right-hand side are smaller than $\epsilon/2$ once $n, m \geqslant n_0$. It follows that $\{\varrho(x_n, \tilde{x}_n)\}_{n\in\mathbb{N}}$ is Cauchy. The completeness of $\mathbb{R}$ proved in Theorem 17.2 then gives (18.22). The argument for (18.23) is based on a similar inequality as (18.24) and observation (3) in Lemma 18.14 so we leave it to the reader. □

We can now define $\overline{\varrho} \colon \overline{X} \times \overline{X} \to \mathbb{R}$ by

$$\overline{\varrho}\big([\{x_n\}_{n\in\mathbb{N}}], [\{\tilde{x}_n\}_{n\in\mathbb{N}}]\big) := \lim_{n\to\infty} \varrho(x_n, \tilde{x}_n) \tag{18.25}$$

where, by (18.23), the limit is independent of the representatives. We then quickly check:

Lemma 18.16 *$\overline{\varrho}$ is a metric on $\overline{X}$.*

Proof. The symmetry and non-negativity are immediate from the corresponding properties of ϱ and so is the triangle inequality. The definition (18.25) along with (18.18) ensure

$$\overline{\varrho}([\{x_n\}_{n\in\mathbb{N}}],[\{\tilde{x}_n\}_{n\in\mathbb{N}}]) = 0 \quad \Rightarrow \{\tilde{x}_n\}_{n\in\mathbb{N}} \in [\{x_n\}_{n\in\mathbb{N}}] \tag{18.26}$$

and so $[\{x_n\}_{n\in\mathbb{N}}] = [\{\tilde{x}_n\}_{n\in\mathbb{N}}]$ by Lemma 18.14(3). □

We are now ready to give:

Proof of Theorem 18.13. Let $(\overline{X},\overline{\varrho})$ be as above and let $\phi\colon X \to \overline{X}$ be defined by

$$\phi(x) := [\{x\}_{n\in\mathbb{N}}] \tag{18.27}$$

where $\{x\}_{n\in\mathbb{N}}$ denotes the constant sequence whose all terms are equal to x. (This sequence is trivially Cauchy.) The proof now splits into three claims:

Claim 1: ϕ is an isometry. This is immediate from

$$\overline{\varrho}\big([\{x\}_{n\in\mathbb{N}}],[\{\tilde{x}\}_{n\in\mathbb{N}}]\big) = \lim_{n\to\infty} \varrho(x,\tilde{x}) = \varrho(x,\tilde{x}). \tag{18.28}$$

Claim 2: $\phi(X)$ is dense in $\overline{X}$: Pick any $[\{x_n\}_{n\in\mathbb{N}}] \in \overline{X}$. Then

$$\lim_{m\to\infty} \overline{\varrho}\big(\phi(x_m),[\{x_n\}_{n\in\mathbb{N}}]\big) = \lim_{n\to\infty} \varrho(x_m,x_n) = 0 \tag{18.29}$$

where the last conclusion follows from the fact that $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy. This implies $\phi(x_m) \to [\{x_n\}_{n\in\mathbb{N}}]$ in $(\overline{X},\overline{\varrho})$ and so $[\{x_n\}_{n\in\mathbb{N}}]$ is an adherent point of $\phi(X)$. As this is true for any $[\{x_n\}_{n\in\mathbb{N}}] \in \overline{X}$, the closure of $\phi(X)$ is all of $\overline{X}$.

Claim 3: $(\overline{X},\overline{\varrho})$ is complete: Consider a sequence $\{[\{x_n^{(m)}\}_{n\in\mathbb{N}}]\}_{m\in\mathbb{N}}$ (indexed by m) of elements in $\overline{X}$ and assume that it is Cauchy in metric $\overline{\varrho}$. As $\phi(X)$ is already known to be dense in $\overline{X}$, for each $m \in \mathbb{N}$ there is $y_m \in X$ such that

$$\overline{\varrho}\big([\{x_n^{(m)}\}_{n\in\mathbb{N}}],\phi(y_m)\big) \leqslant \frac{1}{m+1} \tag{18.30}$$

where, we racall, $\phi(y_m)$ is the equivalence class of Cauchy sequences represented by a constant sequence equal to y_m. Since ϕ is an isometry, the triangle inequality and (18.30) show

$$\forall m,k \in \mathbb{N}\colon \quad \varrho(y_m,y_k) = \overline{\varrho}\big(\phi(y_m),\phi(y_k)\big) \leqslant \frac{1}{m+1} + \frac{1}{k+1} \tag{18.31}$$

and so $\{y_n\}_{n\in\mathbb{N}}$ is Cauchy. This means that $[\{y_n\}_{n\in\mathbb{N}}]$ is an element of $\overline{X}$. Taking a limit $k \to \infty$ in (18.31) then shows

$$\forall m \in \mathbb{N}\colon \quad \overline{\varrho}\big(\phi(y_m),[\{y_n\}_{n\in\mathbb{N}}]\big) \leqslant \frac{1}{m+1}. \tag{18.32}$$

Combining (18.30) and (18.32) using the triangle inequality now shows

$$\forall m \in \mathbb{N}\colon \quad \overline{\varrho}\big([\{x_n^{(m)}\}_{n\in\mathbb{N}}],[\{y_n\}_{n\in\mathbb{N}}]\big) \leqslant \frac{2}{m+1} \tag{18.33}$$

and so $[\{x_n^{(m)}\}_{n\in\mathbb{N}}] \to [\{y_n\}_{n\in\mathbb{N}}]$ in $(\overline{X},\overline{\varrho})$. This proves that $(\overline{X},\overline{\varrho})$ is complete. □

We remark that parts of the above proof can indeed be used (as Cantor did in his paper from 1878) to construct a system of reals out of a system of rationals. Indeed, specialize to $X := \mathbb{Q}$ and $\varrho(a,b) := |a-b|$ (which takes rational values) and observe that the notion of being Cauchy can be defined using rationals alone (see, e.g., Definition 14.3). Then

let $\mathbb{R}$ as the set of classes of equivalence of Cauchy sequences. The map (18.27) then gives us an injection $\mathbb{Q} \to \mathbb{R}$. We now define the algebraic operations on $\mathbb{R}$ as follows

$$\begin{aligned} [\{a_n\}_{n\in\mathbb{N}}] + [\{b_n\}_{n\in\mathbb{N}}] &:= [\{a_n + b_n\}_{n\in\mathbb{N}}] \\ [\{a_n\}_{n\in\mathbb{N}}] \cdot [\{b_n\}_{n\in\mathbb{N}}] &:= [\{a_n \cdot b_n\}_{n\in\mathbb{N}}] \end{aligned} \tag{18.34}$$

(which requires showing that $\{a_n + b_n\}_{n\in\mathbb{N}}$ and $\{a_n \cdot b_n\}_{n\in\mathbb{N}}$ are Cauchy if $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are). Along with $\underline{0} := \{0\}_{n\in\mathbb{N}}$, $\underline{1} := \{1\}_{n\in\mathbb{N}}$ and the relation

$$[\{a_n\}_{n\in\mathbb{N}}] \leqslant [\{b_n\}_{n\in\mathbb{N}}] \;:=\; \forall k \in \mathbb{N}\colon \left\{ n \in \mathbb{N}\colon\; a_n \leqslant b_n + \frac{1}{k+1} \right\} \text{ is infinite} \tag{18.35}$$

we then check that $(\mathbb{R}, +, \underline{0}, \cdot, \underline{1}, \leqslant)$ is an ordered field. A variant of the argument in Claim 3 above then shows that this ordered field is in fact complete (in the sense of obeying the supremum axiom) and is thus a system of reals. (Of course, all arguments there have to be phrased using rationals only.)

18.4 Uniqueness of the completion.

As our final item we also address the uniqueness of the completion. Of course, this can only be true modulo an isometric bijection:

Theorem 18.17 (Uniqueness up to an isomorphism) *If $(\overline{X}_1, \overline{\varrho}_1)$ and $(\overline{X}_2, \overline{\varrho}_2)$ are two completions of a metric space (X, ϱ), then there is bijection $\overline{\psi}\colon \overline{X}_1 \to \overline{X}_2$ which is an isometry.*

Proof. The definition of a closure ensures existence of the isometries $\phi_i\colon X \to \overline{X}_i$, $i = 1, 2$, such that the closure of $\phi_i(X)$ in $\overline{X}_i$ is all of $\overline{X}_i$. By Lemma 18.9 these maps are injective and so we may define $\psi\colon \phi_1(X) \to \phi_2(X)$ by

$$\psi(x) := \phi_2 \circ \phi_1^{-1}(x). \tag{18.36}$$

Then the fact that both ϕ_1 and ϕ_2 are isometries imply, for all $x, y \in \phi_1(X)$,

$$\begin{aligned} \varrho_2\bigl(\psi(x), \psi(y)\bigr) &= \varrho_2\bigl(\phi_2 \circ \phi_1^{-1}(y), \phi_2 \circ \phi_1^{-1}(x)\bigr) \\ &= \varrho(\phi_1^{-1}(y), \phi^{-1}(x) = \varrho_1(y, x) \end{aligned} \tag{18.37}$$

and so ψ is an isometry of $\phi_1(X)$ onto $\phi_2(X)$.

For ease of presentation we will now proceed using an argument that relies on the Axiom of Choice, although this can be avoided (see Remark 18.18 after the proof): Consider now any $x \in \overline{X}_1$. Since the closure of $\phi_1(X)$ is $\overline{X}_1$, Corollary 16.6 (which is where the AC is used) implies existence of $\{x_n\}_{n\in\mathbb{N}} \in \phi_1(X)^{\mathbb{N}}$ such that $x_n \to x$. Using that ψ is an isometry, we readily check the following facts:

$$x \in \phi_1(X) \;\Rightarrow\; \psi(x_n) \to \psi(x) \tag{18.38}$$

and if $x \notin \phi_1(X)$, then $\psi(x_n)$ is Cauchy and, by completeness of $\overline{X}_2$, convergent. We may thus define

$$\overline{\psi}(x) := \lim_{n\to\infty} \psi(x_n) \tag{18.39}$$

A technical caveat is this definition seems to depend on the choice of the sequence convergent to x. However, using that ψ is an isometry one readily checks that any sequence that converges to x will lead to the same value of the limit of $\psi(x_n)$.

Noting that $\overline{\psi}$ is defined on all of $\overline{X}_1$, all that remains to prove two claims:

Claim 1: $\overline{\psi}$ is an isometry. Note that if $\{x_n\}_{n\in\mathbb{N}}, \{\tilde{x}_n\}_{n\in\mathbb{N}} \in \phi_1(X)^{\mathbb{N}}$ are such that $x_n \to x$ and $\tilde{x}_n \to \tilde{x}$, then the triangle inequality and the fact that ψ is an isometry on $\phi_1(X)$ shows

$$\Big|\varrho_2\big(\overline{\psi}(x), \overline{\psi}(\tilde{x})\big) - \varrho_1(x, \tilde{x})\Big| \leqslant \varrho_2\big(\overline{\psi}(x), \psi(x_n)\big) + \varrho_2\big(\psi(\tilde{x}_n), \overline{\psi}(\tilde{x})\big) \tag{18.40}$$

where the right-hand side tends to zero by (18.39). Hence we get

$$\varrho_2\big(\overline{\psi}(x), \overline{\psi}(\tilde{x})\big) = \varrho_1(x, \tilde{x}) \tag{18.41}$$

and so $\overline{\psi}$ is an isometry as desired.

Claim 2: $\overline{\psi}$ is onto. Let $y \in \overline{X}_2$ and note that, since the closure of $\phi_2(X)$ is all of $\overline{X}_2$, Corollary 16.6 ensures the existence of $\{y_n\}_{n\in\mathbb{N}} \subseteq \phi_2(X)$ such that $y_n \to y$. Since ψ is onto $\phi_2(X)$, for each $n \in \mathbb{N}$ there is $x_n \in X$ be such that $\psi(x_n) = y_n$ and the fact that ψ is an isometry implies that $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy. As $\overline{X}_1$ is complete, there is $x \in \overline{X}_1$ such that $x_n \to x$. But the aforementioned independence of (18.39) on the sequence approaching x, we have that $\psi(x_n) \to \overline{\psi}(x)$ which implies $y = \overline{\psi}(x)$. The map $\overline{\psi}$ is thus onto. □

Remark 18.18 Here is a trick that allows us to construct $\overline{\psi}$ without the use of the Axiom of Choice: Consider the set

$$G := \big\{(x, \psi(x)) \in X_1 \times X_2 \colon x \in \phi_1(X)\big\} \tag{18.42}$$

This is a subset of the product metric space $(\overline{X}_1 \times \overline{X}_2, \rho_\infty)$ where $\rho_\infty((x_1, x_2), (y_1, y_2)) := \max\{\rho_1(x_1, y_1), \rho_2(x_2, y_2)\}$ so we may consider its closure $\overline{G}$ in $\overline{X}_1 \times \overline{X}_2$.

We now check (without use of sequences but instead relying on the fact that all points in $\overline{G}$ are adherent to G) that $\overline{G}$ is a graph of a function $\overline{\psi}$ satisfying Claim 1 and 2 above. (Details of this are left to the reader.)

Theorems 18.13 and 18.17 can thus be considered as an variation on the proof of the existence and uniqueness of the reals. However, unlike Dedekind's approach, the advantage of the metric-space based approach is its seamless extension to other contexts, and in particular, to linear vector spaces of infinite dimension. This is quite appreciated in the subject of mathematics called functional analysis that deals with such spaces systematically.

19. SEQUENTIAL COMPACTNESS

As part of our proof of completeness of the reals we proved the *Bolzano-Weierstrass theorem* which states that every bounded sequence of the reals contains a convergent subsequence. The aim of this section is to investigate how this concept generalizes to other metric spaces.

19.1 Definition and necessary conditions.

We start by stating the desired property formally. First, given a sequence $\{n_k\}_{k\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$, let us henceforth use the shorthand

$$n_k \to \infty \quad := \quad \forall m \in \mathbb{N}\colon \{k \in \mathbb{N}\colon n_k \leqslant m\} \text{ is finite} \tag{19.1}$$

whose meaning is reasonably self-explanatory. We then introduce:

Definition 19.1 (Sequential compactness) *Let (X, ϱ) be a metric space. A set $A \subseteq X$ is said to be* sequentially compact *if every sequence from A contains a subsequence convergent to a point in A, i.e.,*

$$\forall \{x_n\}_{n\in\mathbb{N}} \in A^{\mathbb{N}}\, \exists \{n_k\}_{k\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}\, \exists x \in A\colon n_k \to \infty \,\wedge\, x_{n_k} \to x. \tag{19.2}$$

A metric space (X, ϱ) is sequentially compact if the above holds for $A := X$.

We can check that $A \subseteq X$ is sequentially compact if and only if (A, ϱ_A), where ρ_A is the restriction of ρ to A, is a sequentially compact metric space. This means that, for many statements, we can and will focus directly on $A := X$.

Here is a simple example of a compact space:

Lemma 19.2 (Finite sets are compact) *Any (X, ϱ) with X finite is sequentially compact.*

Proof. Let $\{x_n\}_{n\in\mathbb{N}}$ be any sequence from X. Writing $z_1, \dots, z_m$ for the points in X, set

$$\forall k = 1, \dots, m\colon \quad I_k := \big\{n \in \mathbb{N}\colon x_n = z_k\big\}. \tag{19.3}$$

As $\bigcup_{k=1}^m I_k = \mathbb{N}$, there exists $k = 1, \dots, m$ such that I_k is infinite. Let $\{n_j\}_{j\in\mathbb{N}}$ enumerate I_k; i.e., set $n_0 := \inf(I_k)$ and $\forall j \in \mathbb{N}\colon n_{j+1} := \inf\{n \in I_k\colon n > n_j\}$. Then $n_j \geqslant j$ and $x_{n_j} = z_k$ for all $j \in \mathbb{N}$ and so $n_j \to \infty$ and $x_{n_j} \to z_k$ as desired. □

While the previous proof may seems special to the setting of finite sets, the key argument there — which is a version of the "pigeon-hole principle" — is that the union of a finite number of sets is infinite only if one of the sets is infinite. This argument will drive the proofs characterizing sequentially compact sets in $\mathbb{R}^d$ and linking sequential compactness to total boundedness. In this sense, sequentially compact spaces are the closest relatives of finite ones.

We will now observe two properties that are implied by, and are thus *necessary* for, sequential compactness. We start with:

Lemma 19.3 (AC)(Compactness implies boundedness) *Let (X, ϱ) be a metric space. Then*

$$\forall A \subseteq X\colon \quad A \text{ sequentially compact } \Rightarrow \; A \text{ bounded} \tag{19.4}$$

Here we recall that a set $A \subseteq X$ is bounded if $\exists x \in X\, \exists r > 0\colon A \subseteq B(x, r)$.

Proof. We will prove the contrapositive. The Axiom of Choice will have to be invoked. Suppose that A is NOT bounded. Then

$$\forall x \in X \, \forall n \in \mathbb{N}: \;\; A \setminus B(x,n) \neq \varnothing \tag{19.5}$$

Given any $x \in X$, the AC thus yields $\times_{n\in\mathbb{N}} A \setminus B(x,n) \neq \varnothing$ meaning that there exists $f: \mathbb{N} \to A$ such that $f(n) \in A \setminus B(x,n)$ for each $n \in \mathbb{N}$. Writing $x_n := f(n)$, we have $\varrho(x,x_n) \geqslant n$ and so $\{\varrho(x,x_n)\}_{n\in\mathbb{N}}$ is unbounded. Such a sequence $\{x_n\}_{n\in\mathbb{N}}$ cannot contain a convergent (or even Cauchy) subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ because, by Lemma 17.4, that would require that $\{\varrho(x,x_{n_k})\}_{k\in\mathbb{N}}$ converges and is thus bounded. Hence, A NOT bounded implies that A is NOT sequentially compact, proving (19.4). □

Lemma 19.4 (AC)(Compactness implies closedness) *Let (X,ϱ) be a metric space. Then*

$$\forall A \subseteq X: \quad A \text{ sequentially compact } \Rightarrow \; A \text{ closed} \tag{19.6}$$

Proof. We again prove the contrapositive. Theorem 16.5 (with the AC) implies

$$\neg(A \;\text{ closed}) \;\Rightarrow\; \exists \{x_n\}_{n\in\mathbb{N}} \in A^{\mathbb{N}} \, \exists x \in X: \, x_n \to x \;\wedge\; x \notin A. \tag{19.7}$$

But any subsequence of $\{x_n\}_{n\in\mathbb{N}}$ will then converge to x and so A NOT closed implies A NOT sequentially compact. □

From the previous lemmas we conclude that boundedness and closedness are necessary conditions for sequential compactness. As it turns out, for $X := \mathbb{R}$ or $\mathbb{R}^d$ endowed with the norm metric, these two conditions are also *sufficient*, thus completely characterizing sequentially compact subsets of the Euclidean space. For reasons to be explained later, this characterization is referred to by the names of E. Heine and E. Borel (with others just as deserving to be included) although for $\mathbb{R}^d$ it is not more than a mere restatement of the Bolzano-Weierstrass Theorem.

Theorem 19.5 (AC)(Heine-Borel property of $\mathbb{R}^d$) *Let $d \geqslant 1$ be a natural and consider the metric space $(\mathbb{R}^d,\varrho)$ where ϱ is a norm-metric on $\mathbb{R}^d$. Then*

$$\forall A \subseteq \mathbb{R}^d: \; A \text{ sequentially compact } \Leftrightarrow \; A \text{ closed and bounded} \tag{19.8}$$

The AC is used only for the direction $\Rightarrow$.

Proof. We have already proved $\Rightarrow$ (with the help of AC) in the above lemmas, so let us focus on $\Leftarrow$. This was already stated in Corollary 17.14 which requires completeness of $(\mathbb{R}^d,\varrho)$ proved in Theorem 17.10. □

The previous proof relied on the completeness of the underlying space. This is no loss in light of completeness being another necessary condition for compactness:

Lemma 19.6 (Compactness implies completeness) *Let (X,ϱ) be a metric space. Then*

$$(X,\varrho) \text{ sequentially compact } \Rightarrow \; (X,\varrho) \text{ complete.} \tag{19.9}$$

Proof. Let $\{x_n\}_{n\in\mathbb{N}}$ be Cauchy. If $\{x_n\}_{n\in\mathbb{N}}$ had a convergent subsequence, say $x_{n_j} \to x$, then the Cauchy property would ensure $x_n \to x$ (HW problem) and so every Cauchy sequence would be convergent. Thus compactness implies completeness. □

We remark that, even with completeness in place, a key additional ingredient of the proof was the finite-dimensionality of $\mathbb{R}^d$. This is essential as seen in the following example: Consider the set of bounded sequences

$$\ell^\infty(\mathbb{N}) := \Big\{ \{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \colon \text{bounded} \Big\} \tag{19.10}$$

endowed with the metric ϱ_∞ associated with the norm

$$\big\|\{x_n\}_{n\in\mathbb{N}}\big\|_\infty := \sup_{n\in\mathbb{N}} |x_n| \tag{19.11}$$

where the supremum on the right abbreviates $\sup\{|x_n| \colon n \in \mathbb{N}\}$. (That this is a norm is checked just as for the ∞-norm on $\mathbb{R}^d$.) The space $(\ell^\infty(\mathbb{N}), \varrho_\infty)$ is also easily checked to be complete. However, it fails to have the Heine-Borel property since the closed unit ball

$$B'(0,1) := \Big\{ \{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} \colon \sup_{n\in\mathbb{N}} |x_n| \leqslant 1 \Big\} \tag{19.12}$$

which is also bounded, contains sequences $x^{(k)} := \{x_n^{(k)}\}_{n\in\mathbb{N}}$ defined as

$$x_n^{(k)} := \begin{cases} 1, & \text{if } n = k, \\ 0, & \text{else,} \end{cases} \tag{19.13}$$

whose terms are all unit distance apart, $\|x^{(k)} - x^{(\ell)}\|_\infty = 1$ when $k \neq \ell$. Such a sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ cannot contain convergent, or even Cauchy, subsequences.

The culprit here is really the infinite dimension. In the above sequences, the coordinate sequences converge (and this can be always arranged for subsequences by the so called *Cantor diagonal argument* that we will elaborate on later) but, unlike in finite dimension, that alone is no longer sufficient to ensure the convergence in the ∞-norm metric. (Indeed, the latter requires that coordinates converge *uniformly*.)

The problem is actually the same regardless of the choice of the norm-metric: A theorem in functional analysis asserts that closed norm-metric balls in complete linear vector spaces are sequentially compact if and only if the space is of finite dimension.

19.2 Total boundedness.

In light of the previous counterexample, the question is as follows: Given a general metric space (X, ϱ), what conditions do we need to add to (already necessary) boundedness and completeness to infer sequential compactness? For this we need:

Definition 19.7 (Total boundedness) *We say that a set $A \subseteq X$ is* totally bounded *if*

$$\forall r > 0\, \exists n \in \mathbb{N}\ \exists x_0, \dots, x_n \in A \colon \quad A \subseteq \bigcup_{i=0}^{n} B(x_i, r) \tag{19.14}$$

The space (X, ϱ) is totally bounded if this applies to $A := X$.

Total boundedness implies boundedness, indeed, (19.14) shows that

$$A \subseteq B(x_0, r') \quad \text{for} \quad r' := r + (n+1)\, \max\big\{\varrho(x_0, x_j) \colon i = 0, \dots, n\big\} \tag{19.15}$$

but the converse is generally false. Also note that we actually do not need to require (19.14) for all $r > 0$; it suffices to require this for a sequence of r's tending to zero. We can

also check that the total boundedness is inherited to relative topologies; indeed, if (X, ϱ) is totally bounded, so is every subset $A \subseteq X$ (prove this!). In particular, we only need to prove statements about total boundedness of the whole space.

The reason why we introduce total boundedness is that it is another necessary condition for sequential compactness:

Lemma 19.8 (AC)(Compactness implies total boundedness) *For each metric space* (X, ϱ),

$$(X, \varrho) \text{ sequentially compact } \Rightarrow \ (X, \varrho) \text{ totally bounded} \tag{19.16}$$

Proof. We will again aim to prove the contrapositive. Suppose that (X, ϱ) is NOT totally bounded. Then

$$\exists r > 0 \ \forall n \in \mathbb{N} \ \forall x_0, \dots, x_n \in X\colon \quad X \setminus \bigcup_{i=0}^{n} B(x_i, r) \neq \varnothing \tag{19.17}$$

Using the Axiom of Choice, we may thus choose a sequence $\{z_k\}_{k\in\mathbb{N}}$ such that

$$x_0 \in X \ \wedge \ \forall k \in \mathbb{N}\colon x_{k+1} \in X \setminus \bigcup_{i=0}^{k} B(x_i, r) \tag{19.18}$$

Now note that $\varrho(x_i, x_{n+1}) \geqslant r$ for all $i = 0, \dots, n$ and so we have

$$\forall m, n \in \mathbb{N}\colon \quad m \neq n \ \Rightarrow \ \varrho(x_m, x_n) \geqslant r \tag{19.19}$$

This again implies that $\{x_n\}_{n\in\mathbb{N}}$ contains no convergent subsequence and so (X, ϱ) is NOT sequentially compact. □

The total boundedness offers a way to approximate the metric space (X, ϱ) by a finite metric space — namely, the space $\{x_0, \dots, x_n\}$ consisting of the centers of the r-balls $B(x_0, r), \dots, B(x_n, r)$ that cover X. This allows us to capitalize on the proof of sequential compactness in finite spaces and, particularly, on the argument underlying the characterization of sequential compactness in Euclidean spaces:

Theorem 19.9 (AC) *Let* (X, ϱ) *be a metric space. Then*

$$(X, \varrho) \text{ sequentially compact } \Leftrightarrow \ (X, \varrho) \text{ complete and totally bounded} \tag{19.20}$$

In particular, if (X, ϱ) *is complete then*

$$\forall A \subseteq X\colon \ A \text{ sequentially compact } \Leftrightarrow \ A \text{ closed and totally bounded} \tag{19.21}$$

Proof. The proof of $\Rightarrow$ reduces to the above lemmas, so we just need to prove $\Leftarrow$. First observe that, using the total boundedness of (X, ϱ), for each $k \in \mathbb{N}$ there is $m_k \in \mathbb{N}$ and the points $z_0^{(k)}, \dots, z_{m_k}^{(k)}$ such that

$$\bigcup_{i=0}^{m_k} B\big(z_i^{(k)}, 2^{-k}\big) = X \tag{19.22}$$

(As a choice of $z_0^{(k)}, \dots, z_{m_k}^{(k)}$ was made, we had to invoke the AC.) Fix $\{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ and proceed by iterating the argument in the proof of Lemma 19.2 to define a sequence $\{I_k\}_{k\in\mathbb{N}}$ of infinite subsets of $\mathbb{N}$ and a sequence $\{B_k\}_{k\in\mathbb{N}}$ of open balls in X recursively as

follows: For $k = 0$ we just note that, since X is bounded, there are $r > 0$ and $z \in X$ such that $X = B(z_0, r)$. Then set

$$I_0 := \mathbb{N} \quad \text{and} \quad B_0 := B(z_0, r_0) \tag{19.23}$$

Proceeding recursively, assume now that for some $k \in \mathbb{N}$ the infinite sets $I_0, \dots, I_k \subseteq \mathbb{N}$ and open balls $B_0, \dots, B_k$ have already been defined. For each $i = 1, \dots, m_{n+1}$ set

$$J_k^{(i)} := \Big\{ j \in I_k \colon x_j \in B\big(z_i^{(k+1)}, 2^{-(k+1)}\big) \Big\}. \tag{19.24}$$

and, noting that (19.22) implies

$$\bigcup_{i=0}^{m_{k+1}} J_k^{(i)} = I_k \tag{19.25}$$

observe that the "pigeon-hole principle" forces at least one of the $J_k^{(i)}$'s to be infinite. This means that we can set

$$i_{k+1} := \min\big\{ i \in \{1, \dots, m_{k+1}\} \colon J_k^{(i)} \text{ infinite} \big\} \tag{19.26}$$

and let

$$I_{k+1} := J_k^{(i_{k+1})} \ \wedge \ B_{k+1} := B\big(z_{i_{k+1}}^{(k+1)}, 2^{-(k+1)}\big) \tag{19.27}$$

Since I_{k+1} is infinite, the recursive definition can proceed for all $k \in \mathbb{N}$.

The construction of the above objects ensures

$$\forall k \in \mathbb{N} \colon \quad I_k \text{ infinite } \wedge \ I_{k+1} \subseteq I_k \ \wedge \ \forall n \in I_k \colon x_n \in B_k \tag{19.28}$$

The fact that B_k is an open ball of radius 2^{-k} along with the triangle inequality gives

$$\forall k \in \mathbb{N} \, \forall m, n \in I_k \colon \varrho(x_n, x_m) < 2 \cdot 2^{-k} = 2^{1-k} \tag{19.29}$$

We now call upon the following elementary but useful fact:

Lemma 19.10 (Cantor's diagonal argument) *Let $\{I_k\}_{k\in\mathbb{N}} \in \mathcal{P}(\mathbb{N})^{\mathbb{N}}$ be such that*

$$\forall k \in \mathbb{N} \colon I_k \text{ infinite } \wedge \ I_{k+1} \subseteq I_k \tag{19.30}$$

Then $\{n_k\}_{k\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ constructed recursively so that

$$n_0 = \inf(I_0) \ \wedge \forall k \in \mathbb{N} \colon \ n_{k+1} = \inf\{n \in I_{k+1} \colon n > n_k\} \tag{19.31}$$

obeys

$$\forall k \in \mathbb{N} \colon \ n_k \in I_k \tag{19.32}$$

Proof. This follows from Lemma 9.10 and the fact that the set on the right of (19.31) is infinite, and thus non-empty, for each $k \in \mathbb{N}$. □

Moving back to the proof of Theorem 19.9, let $\{n_k\}_{k\in\mathbb{N}}$ be as in (19.31). Then (19.29), (19.32) and $n_k \geqslant k$ give

$$\forall j, k \in \mathbb{N} \colon j \leqslant k \ \Rightarrow \ \varrho(x_{n_j}, x_{n_k}) \leqslant 2^{-n_k+1} \leqslant 2^{1-k} \tag{19.33}$$

showing that $\{x_{n_k}\}_{k\in\mathbb{N}}$ is Cauchy. Since (X, ϱ) is assumed complete, $\{x_{n_k}\}_{k\in\mathbb{N}}$ is convergent and so (X, ϱ) is sequentially compact. □

Remark 19.11 To explain the phrase "diagonal argument" we note the following version of the statement of Lemma 19.10: Given sets $\{I_k\}_{k\in\mathbb{N}}$ satisfying (19.30), enumerate each I_k into a sequence $\{n_j^{(k)}\}_{j\in\mathbb{N}}$. The condition $I_{k+1} \subseteq I_k$ then means that $\{n_j^{(k+1)}\}_{j\in\mathbb{N}}$ is a subsequence of $\{n_j^{(k)}\}_{j\in\mathbb{N}}$. Taking the so-called "diagonal sequence"

$$\hat{n}_k := n_k^{(k)} \tag{19.34}$$

then gives us $\{\hat{n}_k\}_{k\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ which (except for a few initial elements) is a subsequence of all of these sequences and obeys $\hat{n}_k \in I_k$ for each $k \in \mathbb{N}$.

It can be checked that a totally bounded space has a compact completion. Metric spaces that have a compact completion are called *precompact*. (In topological spaces that are not metric, the notion of a completion is meaningless, but one then says that a set is precompact if its closure is compact.) The above shows that being precompact is equivalent to being totally bounded. A direct way to define a precompact set is by saying that every sequence drawn from the set has a Cauchy subsequence.

20. COMPACTNESS IN TOPOLOGY

In the previous section, we defined the notion of sequential compactness by asking that every sequence contain a convergent subsequence. Here we will discuss the consequences of compactness for the open sets, a.k.a. topology and then explain how compactness arises in topological spaces.

20.1 Cantor's intersection property.

A classical result of Cantor says that the reals are uncountable. We showed this in Theorem 13.1 using a diagonal argument. This theorem was dated 1891, but Cantor first proved the result already nearly 20 years earlier by an argument that relies, in its nature, on sequential compactness. Here is his theorem again:

Theorem 20.1 (Cantor 1874) $[0,1] := \{x \in \mathbb{R} \colon 0 \leqslant x \leqslant 1\}$ *is not countable.*

Proof. Suppose, for the sake of contradiction, that there is a sequence $\{x_n\}_{n\in\mathbb{N}}$ of real numbers such that $[0,1] = \{x_n \colon n \in \mathbb{N}\}$. We will now construct two auxiliary sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ satisfying

$$\forall n \in \mathbb{N} \colon\ 0 \leqslant a_n < b_n \leqslant 1 \tag{20.1}$$

as follows: Set $a_0 := 0$ and $b_0 := 1$ and, assuming a_n and b_n have been defined so that (20.1) holds, define a_{n+1} and b_{n+1} recursively so that

$$\forall n \in \mathbb{N} \colon \begin{cases} x_n \leqslant \dfrac{a_n + b_n}{2} \ \Rightarrow\ a_{n+1} = \dfrac{a_n + 2b_n}{3} \ \wedge\ b_{n+1} = b_n \\ \dfrac{a_n + b_n}{2} < x_n \ \Rightarrow\ a_{n+1} = a_n \ \wedge\ b_{n+1} = \dfrac{2a_n + b_n}{3} \end{cases} \tag{20.2}$$

We now readily check that $\{a_n\}_{n\in\mathbb{N}}$ is non-decreasing and $\{b_n\}_{n\in\mathbb{N}}$ is non-increasing. The monotonicity upgrades (20.1) into

$$\forall n, m \in \mathbb{N} \colon n \leqslant m \Rightarrow a_n \leqslant b_m \tag{20.3}$$

and so each a_n is a lower bound on $\{b_m \colon m \in \mathbb{N}\}$ and each b_m is an upper bound on $\{a_n \colon n \in \mathbb{N}\}$. Denoting

$$a := \sup\{a_n \colon n \in \mathbb{N}\} \ \wedge\ b := \inf\{b_m \colon m \in \mathbb{N}\} \tag{20.4}$$

we thus have $0 \leqslant a \leqslant b \leqslant 1$ by an exercise in an earlier homework. (Alternatively, we can use Lemma 17.6 to show that $a = \lim_{n\to\infty} a_n$ and $b = \lim_{n\to\infty} b_n$ and then infer the inequality from (20.1).) But (20.1–20.2) ensure

$$\forall n \in \mathbb{N} \colon x_n \notin [a_{n+1}, b_{n+1}] \tag{20.5}$$

and, since $[a,b] \subseteq [a_n, b_n]$ for all $n \in \mathbb{N}$,

$$\forall n \in \mathbb{N} \colon x_n \notin [a,b] \tag{20.6}$$

In particular, the number a, which lies in $[0,1]$, is not a member of $\{x_n\}_{n\in\mathbb{N}}$ in contradiction with the assumption that this sequence lists all points in $[0,1]$. □

The previous proof clearly uses the same idea as our proof of the Bolzano-Weierstrass theorem and could be deduced from sequential compactness of $[0,1]$. Notwithstanding,

we can also recast the key argument in the previous proof as follows: Denote

$$C_n := [a_n, b_n] \tag{20.7}$$

Then the monotonicities of $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ give

$$\forall n \in \mathbb{N}\colon\ C_{n+1} \subseteq C_n \tag{20.8}$$

and so $\{C_n\}_{n\in\mathbb{N}}$ is a sequence of nested closed non-empty subintervals of $[0,1]$. The proof then hinges on the fact that these properties imply

$$\bigcap_{n\in\mathbb{N}} C_n \neq \varnothing \tag{20.9}$$

As it turns out, the underlying argument (suitably generalized to closed sets) applies to all sequentially compact spaces:

Theorem 20.2 (AC)(Cantor's intersection property) *A metric space (X,ϱ) is sequentially compact if and only if every nested sequence of non-empty close subsets has a non-empty intersection, i.e., for all $\{C_n\}_{n\in\mathbb{N}} \in \mathcal{P}(X)^{\mathbb{N}}$ we have*

$$\Big(\forall n \in \mathbb{N}\colon C_n \text{ closed } \wedge\ C_n \neq \varnothing\ \wedge\ C_{n+1} \subseteq C_n\Big)\ \Rightarrow\ \bigcap_{n\in\mathbb{N}} C_n \neq \varnothing \tag{20.10}$$

Proof of necessity of (20.10). Assume that (X,ϱ) is sequentially compact and let $\{C_n\}_{n\in\mathbb{N}}$ be a sequence of non-empty closed sets with $C_{n+1} \subseteq C_n$ for each $n \in \mathbb{N}$. Since $C_n \neq \varnothing$, we may pick (using the Axiom of Choice) $x_n \in C_n$ for each $n \in \mathbb{N}$. The compactness of X ensures existence of convergent subsequence, $x_{n_k} \to x$. Since $n_k \geqslant k$, for each $n \in \mathbb{N}$ we have $x_{n_k} \in C_n$ as soon as $k \geqslant n$ and so, since C_n is closed and thus contains the limits of all convergent sequences, $x \in C_n$ for all $n \in \mathbb{N}$. It follows that $x \in \bigcap_{n\in\mathbb{N}} C_n$ and so the intersection is indeed non-empty. □

Proof of sufficiency of (20.10). For the converse let us now assume that, for each sequence $\{C_n\}_{n\in\mathbb{N}} \in \mathcal{P}(X)^{\mathbb{N}}$, we have (20.10). Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence from X. Then

$$C_n := \overline{\{x_m\colon m \geqslant n\}} \tag{20.11}$$

are closed (by definition) and non-empty, because $x_n \in C_n$ for each $n \in \mathbb{N}$. Since $A \subseteq B$ implies $\overline{A} \subseteq \overline{B}$, we also have $C_{n+1} \subseteq C_n$ for all $n \in \mathbb{N}$. By (20.10), $\bigcap_{n\in\mathbb{N}} C_n \neq \varnothing$.

Let $x \in \bigcap_{n\in\mathbb{N}} C_n$. By the fact that the closure of the set coincides with the set of the adherent points (see Lemma 16.2), we have

$$\forall r > 0\, \forall n \in \mathbb{N}\colon\ B(x,r) \cap \{x_m\colon m \geqslant n\} \neq \varnothing \tag{20.12}$$

We claim that this also gives

$$\forall r > 0\, \forall k \in \mathbb{N}\colon\ \big\{n \geqslant k\colon x_n \in B(x,r)\big\} \text{ is infinite} \tag{20.13}$$

because if the set is finite for some k, then taking k' the largest element, the set is empty for $k \geqslant k' + 1$, contradicting (20.12). The sets $I_k := \{n \geqslant k\colon x_n \in B(x,2^{-k})\}$ then obey

$$\forall k \in \mathbb{N}\colon\ I_k \text{ infinite } \wedge\ I_{k+1} \subseteq I_k \tag{20.14}$$

Defining $\{n_k\}_{k\in\mathbb{N}}$ from $\{I_k\}_{k\in\mathbb{N}}$ as in Lemma 19.10, we get

$$\forall k \in \mathbb{N}\colon \varrho(x_{n_k}, x) < 2^{-k} \tag{20.15}$$

and so $x_{n_k} \to x$. The sequence $\{x_n\}_{n\in\mathbb{N}}$ thus contains a convergent subsequence and (X, ϱ) is sequentially compact as claimed. □

20.2 Compactness via open covers.

The Cantor intersection property has the following equivalent formulation:

Theorem 20.3 (AC)(Countable open cover property) *A metric space (X, ϱ) has the Cantor intersection property* (20.10) *or, equivalently, is sequentially compact if and only if for any sequence $\{O_n\}_{n\in\mathbb{N}}$ of open subsets of X,*

$$\bigcup_{n\in\mathbb{N}} O_n = X \;\Rightarrow\; \exists n \in \mathbb{N}\colon \bigcup_{k=0}^{n} O_k = X \tag{20.16}$$

i.e., if and only if every countable open cover contains a finite subcover.

Proof. Given $\{O_n\}_{n\in\mathbb{N}}$ be a sequence of open subsets of X,

$$C_n := X \smallsetminus \bigcup_{k=0}^{n} O_k \tag{20.17}$$

defines a sequence of nested closed subsets of X. If, in addition, $\{O_n\}_{n\in\mathbb{N}}$ is a cover of X — i.e., $\bigcup_{n\in N} O_n = X$ — then $\bigcap_{n\in\mathbb{N}} C_n = \varnothing$ while $X = \bigcup_{k=0}^{n} O_k$ implies $C_n = \varnothing$.

Conversely, if $\{C_n\}_{n\in\mathbb{N}}$ is a family of nested closed sets, then $O_n := X \smallsetminus C_n$ are open with $\bigcup_{n\in\mathbb{N}} O_n = X$ as soon as $\bigcap_{n\in\mathbb{N}} C_n = \varnothing$. Moreover, $C_n = \varnothing$ implies $X = O_n$. It follows that (20.16) is equivalent to the statement that, for any $\{C_n\}_{n\in\mathbb{N}} \in \mathcal{P}(X)^{\mathbb{N}}$:

$$\Big(\forall n \in \mathbb{N}\colon C_n \text{ closed } \wedge\; C_{n+1} \subseteq C_n\Big) \wedge \bigcap_{n\in\mathbb{N}} C_n = \varnothing \;\Rightarrow\; \exists n \in \mathbb{N}\colon C_n = \varnothing \tag{20.18}$$

This is the contrapositive to (20.10). □

The property from the previous theorem can further be generalized as follows:

Definition 20.4 (Compactness in topology) *A topological space — i.e., a set X with a class of open sets satisfying the standard axioms — is said to be* compact *if for any set $\{O_\alpha\colon \alpha \in I\}$ of open subsets of X,*

$$\bigcup_{\alpha\in I} O_\alpha = X \;\Rightarrow\; \exists F \subseteq I\colon F \text{ finite } \wedge \bigcup_{\alpha\in F} O_\alpha = X \tag{20.19}$$

(Note the absence of the adjective "sequentially".)

The difference compared to Theorem 20.3 is that here we are asking the open cover property to hold for arbitrary covers by open sets, not just countable ones. This makes a difference in general — and constitutes the distinction between *compactness* and *countable compactness* in generally topology — but not for metric spaces. To explain this, recall the notion of separability from Definition 16.9. We then note:

Lemma 20.5 (AC) *Any totally bounded metric space (X, ϱ) is separable.*

Proof. The total boundedness implies

$$\forall k \in \mathbb{N}\, \exists m_k \in \mathbb{N}\, \exists z_1^{(k)}, \dots, z_{m_k}^{(k)} \in X\colon \bigcup_{i=0}^{m_k} B(z_i^{(k)}, 2^{-k}) = X. \tag{20.20}$$

Let

$$A := \bigcup_{k\in\mathbb{N}} \{z_1^{(k)}, \dots, z_{m_k}^{(k)}\} \tag{20.21}$$

Then, being a countable union of finite sets, A is countable by Lemma 12.14. Moreover, for each $x \in X$ and each $n \in \mathbb{N}$, there is $z \in A$ — namely, $z \in \{z_1^{(n)}, \dots, z_{m_n}^{(n)}\}$ — with $\varrho(x,z) < 2^{-n}$. It follows that $\overline{A} = X$ and so X is separable. □

The fact that the distinction between general open covers and countable open covers makes no difference for metric spaces is then a consequence of:

Lemma 20.6 (AC)(Lindelöf's lemma) *Let (X, ϱ) be separable. Then any open cover of X contains a countable subcover, i.e., any class $\{O_\alpha\colon \alpha \in I\}$ of open sets,*

$$\bigcup_{\alpha\in I} O_\alpha = X \;\Rightarrow\; \exists J \subseteq I\colon\ J \text{ countable} \wedge \bigcup_{\alpha\in J} O_\alpha = X \tag{20.22}$$

Proof. Let $\{O_\alpha\colon \alpha \in I\}$ be an open cover of X and let $A \subseteq X$ be a countable dense subset. Then $A = \{x_n\colon n \in \mathbb{N}\}$ for some sequence $\{x_n\}_{n\in\mathbb{N}} \in X^{\mathbb{N}}$. For each $n \in \mathbb{N}$, we now choose the largest ball of radius of the form 2^{-k} that fits into at least one of O_α's. This amounts to setting r to

$$m(n) := \inf\big\{k \in \mathbb{N}\colon (\exists \alpha \in I\colon B(x_n, 2^{-k}) \subseteq O_\alpha)\big\}, \tag{20.23}$$

where the set under infimum is non-empty because, since $\{O_\alpha\colon \alpha \in I\}$ is a cover, the set $\{\alpha \in I\colon x \in O_\alpha\}$ is non-empty and the fact that O_α is open shows that $x_n \in O_\alpha$ implies $B(x_n, 2^{-k}) \subseteq O_\alpha$ for $k \in \mathbb{N}$ sufficiently large.

Assuming the Axiom of Choice, we now pick

$$\alpha_n \in \big\{\alpha \in I\colon B(x_n, 2^{-m(n)}) \subseteq O_\alpha\big\} \tag{20.24}$$

for each $n \in \mathbb{N}$ and claim that

$$\bigcup_{n\in\mathbb{N}} O_{\alpha_n} = X. \tag{20.25}$$

To prove this, let $x \in X$. Since $\{O_\alpha\colon \alpha \in I\}$ is an open cover of X, there exist $\alpha \in I$ and $k \in \mathbb{N}$ such that $B(x, 2^{-k}) \subseteq O_\alpha$. The fact that A is dense in X in turn implies that there exists $n \in \mathbb{N}$ such that $x_n \in B(x, 2^{-k-1})$. But then

$$B(x_n, 2^{-k-1}) \subseteq B(x, 2^{-k}) \subseteq O_\alpha \tag{20.26}$$

and so $m(n) \leqslant k+1$. This in turn implies

$$x \in B(x_n, 2^{-k-1}) \subseteq B(x_n, 2^{-m(n)}) \subseteq O_{\alpha_n} \tag{20.27}$$

Hence, every $x \in X$ satisfies $x \in \bigcup_{n\in\mathbb{N}} O_{\alpha_n}$ and so we get (20.25). □

We now have all the ingredients needed for:

Theorem 20.7 (AC) *For metric spaces, sequential compactness is equivalent to compactness.*

Proof. Since the open cover property implies the countable open cover property as a special case, the "if" part of Theorem 20.3 shows that compactness implies sequential compactness. For the converse direction, a sequentially compact metric space is separable by Lemmas 19.8 and 20.5 and so, by Lemma 20.6, any open cover can be reduced to a countable subcover. The "only if" part of Theorem 20.3 then ensures that this subcover contains a finite sub-subcover, proving compactness. □

We note that Lindelöf's lemma extends even beyond metric spaces; namely, to the spaces where the topology admits a countable base — these are called *second-countable* spaces. In general, the topological spaces for which the conclusion of Lemma 20.6 holds are called *Lindelöf spaces*. Second countability is sufficient but not necessary for being Lindelöf. The argument used in the proof of Lemma 20.6 can be used to prove the characterization of open subsets of $\mathbb{R}$; cf Theorem 15.14 which is sometimes also called Lindelöf's lemma adding prefix "generalized" to the version in Lemma 20.6.

20.3 Consequences for cardinality.

The notions of compactness (and completeness) are interestingly linked with certain cardinality considerations for metric spaces. We saw one of these in Theorem 20.1 and Lemma 20.5 but other similar connections exist. As these go beyond the scope of these lectures, we will be very brief.

We start with a definition that is already familiar from homework:

Definition 20.8 (Perfect set) *A subset A of a metric space is said to be* perfect *if it is closed and has no isolated points.*

There are many examples of perfect sets; e.g., any closed subinterval of $\mathbb{R}$ or the *Cantor ternary set*, which is the image of $\{0,1\}^{\mathbb{N}}$ under the map f from (13.6). The latter example of the Cantor ternary set is actually very typical:

Theorem 20.9 *Let (X,ϱ) be a complete metric space and $A \subseteq X$ a perfect set. Then there is an injection $f\colon \{0,1\}^{\mathbb{N}} \to A$.*

Proof (main idea). We present only the main idea. Let $x \in A$. Since A is perfect, x is not isolated and so a ball of radius 1 contains at least two points in A distinct from x, say x_0 and x_1. Letting $r_0 := \frac{1}{3}\min\{\varrho(x,x_0),\varrho(x,x_1),\varrho(x_0,x_1)\}$, the closed balls $B'(x_0,r)$ and $B'(x_1,r)$ then also contain two points each, say $x_{00}, x_{01} \in B'(x_0,r)\setminus\{x_0\}$ and $x_{10}, x_{11} \in B'(x_1,r)\setminus\{x_1\}$. Proceeding recursively, at level $n \in \mathbb{N}$ of the recursion, we have defined a distinct point $x_{\sigma_0\ldots\sigma_n} \in A$ for each $\sigma_0,\ldots,\sigma_n \in \{0,1\}$ with all these points separated by at least distance r_n. Then we set r_{n+1} to be $1/3$ of the minimum distance between all the points defined so far and then, in each closed ball $B'(x_{\sigma_0\ldots\sigma_n}, r_{n+1})$, we pick two points $x_{\sigma_0\ldots\sigma_n 0}$ and $x_{\sigma_0\ldots\sigma_n 1}$ distinct from $x_{\sigma_0\ldots\sigma_n}$.

Since $\varrho(x_{\sigma_0\ldots\sigma_{n+1}}, x_{\sigma_0\ldots\sigma_n}) \leqslant r_{n+1}$ and $r_n \to 0$ exponentially fast, we have

$$\forall \sigma = (\sigma_0,\sigma_1,\ldots) \in \{0,1\}^{\mathbb{N}}\colon\ f(\sigma) := \lim_{n\to\infty} x_{\sigma_0\ldots\sigma_n} \text{ exists} \tag{20.28}$$

with $f(\sigma) = f(\sigma')$ only if $\sigma = \sigma'$ thanks to the use of closed balls and the fact that the balls identified at level n are disjoint from one another. This is the desired injection. □

As a consequence of this we get:

Corollary 20.10 *A perfect set A has always at least the cardinality of the continuum. If the underlying metric space is separable, then A is of the cardinality of the continuum.*

Proof. Since $\{0,1\}^{\mathbb{N}}$ has the cardinality of the continuum by (13.31), any sets that embeds it injectively has at least that cardinality. On the other hand, any point x in a separable metric space is a limit of a subsequence of the dense sequence of points and so can be identified with a subset of $\mathbb{N}$. As $\mathcal{P}(\mathbb{N})$ is equinumerous to $\{0,1\}^{\mathbb{N}}$, the space is of the same cardinality as $\{0,1\}^{\mathbb{N}}$, which is that of the continuum. □

The conclusion can be pushed further: The *Cantor-Bendixon theorem* says that every closed subset of a complete separable metric space decomposes uniquely into the union of a perfect set and a countable set.

Note that, unless we assume the Continuum hypothesis, being at least of cardinality of the continuum is generally more restrictive than just being uncountable (a proof of any perfect subset of $\mathbb{R}$ being uncountable is given in the textbook). However, this is actually not relevant here because, by *Kuratowski's theorem*, every infinite complete and separable metric space is either countable or of the cardinality of the continuum. In short, the Continuum hypothesis actually does hold as a theorem in the set of complete separable metric spaces. (Such spaces are called *Polish*, due to may of these ideas being developed by Polish mathematicians in 1920-30s.)

Interested readers can find further analysis of these questions in textbooks on descriptive set theory as well as book on general topology.

21. LIMSUP AND LIMINF

In the remaining lectures of this quarter, we will return to the subjects that are more familiar from calculus. Our first step is to finish some aspects of convergence in $\mathbb{R}$ that have been overshadowed by our treatment of metric spaces. The first step towards this goal is to introduce the notions of upper and lower limits of a sequence.

21.1 The extended reals.

The least-upper bound axiom of the reals guarantees that every non-empty set A of reals that admits an upper bound admits a supremum. The two needed attributes of A are actually closely related; indeed, thanks to the field properties of the reals, the supremum of A is the infimum of the set $B := \{x \in \mathbb{R} \colon (\forall a \in A \colon a \leqslant x)\}$ of its upper bounds which is non-empty because A was assumed to have an upper bound and admits a lower bound by the fact that A was assumed to be non-empty. Still, having to check non-emptyness and existence of an upper bound each time we say "supremum" will be impractical. For this reason, we extend the reals as follows:

Definition 21.1 *The set of* extended reals $\overline{\mathbb{R}}$ *is defined as*

$$\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\} \tag{21.1}$$

where $+\infty$ *and* $-\infty$ *are elements called* positive *and* negative infinity *that obey*

$$+\infty \notin \mathbb{R} \ \wedge \ -\infty \notin \mathbb{R} \ \wedge \ +\infty \neq -\infty \tag{21.2}$$

The ordering relation $\leqslant$ *on* $\mathbb{R}$ *is then extended to* $\overline{\mathbb{R}}$ *so that*

$$-\infty \leqslant \infty \ \wedge \ \big(\forall a \in \mathbb{R} \colon \ -\infty \leqslant a \wedge a \leqslant \infty\big) \tag{21.3}$$

It is readily checked that $\leqslant$ is a total ordering of $\overline{\mathbb{R}}$ with $+\infty$ being the maximal element and $-\infty$ being the minimal element. As a consequence, every subset $A \subseteq \overline{\mathbb{R}}$ now admits at least one upper bound and at least one lower bound. The sets $A \subseteq \mathbb{R}$ that admit no upper bound in $\mathbb{R}$ then have their supremum equal to positive infinity,

$$\forall A \subseteq \mathbb{R} \colon \neg\big(\exists x \in \mathbb{R}\, \forall a \in A \colon a \leqslant x\big) \Rightarrow \sup(A) = +\infty \tag{21.4}$$

while the sets without a lower bound in $\mathbb{R}$ have their infimum equal to negative infinity,

$$\forall A \subseteq \mathbb{R} \colon \neg\big(\exists x \in \mathbb{R}\, \forall a \in A \colon x \leqslant a\big) \Rightarrow \sup(A) = -\infty \tag{21.5}$$

Note that, although (21.4–21.5) do not include the case when $A = \varnothing$, because there each $x \in \mathbb{R}$ (in fact, each $x \in \overline{\mathbb{R}}$) is an upper bound as well as a lower bound, we still get

$$\sup(\varnothing) = -\infty \ \wedge \ \inf(\varnothing) = +\infty \tag{21.6}$$

by the minimality, resp., maximality of $-\infty$, resp., $+\infty$ in $\overline{\mathbb{R}}$. Sets that do contain $+\infty$ have only $+\infty$ as an upper bound, and so the supremum equals $+\infty$ for these. Similarly for the infimum of sets that contain $-\infty$.

In summary, we have proved:

Lemma 21.2 *Every* $A \subseteq \overline{\mathbb{R}}$ *admits a supremum and an infimum in* $\overline{\mathbb{R}}$.

The introduction of the two infinities to $\mathbb{R}$ is very convenient for the ordering and it preserves most of the intuitive properties we usually associate with these concepts in the reals. For instance we have

$$\forall A, B \subseteq \overline{\mathbb{R}}\colon A \subseteq B \;\Rightarrow\; \Big(\inf(B) \leqslant \inf(A) \;\wedge\; \sup(A) \leqslant \sup(B)\Big) \tag{21.7}$$

and

$$\forall A \subseteq \overline{\mathbb{R}}\colon A \neq \varnothing \;\Rightarrow\; \inf(A) \leqslant \sup(A) \tag{21.8}$$

with the warning that the conclusion actually fails for A empty, due to (21.6).

Unfortunately, the situation is more complicated once algebraic operations with infinities are needed. Many standard operations remain defined; for instance,

$$\forall a \in \mathbb{R}\colon a + (+\infty) = +\infty \;\wedge\; a + (-\infty) = -\infty \tag{21.9}$$

and

$$\forall a \in \mathbb{R}\colon a > 0 \Rightarrow a \cdot (\pm\infty) = \pm\infty \tag{21.10}$$

and

$$\forall a \in \mathbb{R}\colon a < 0 \Rightarrow a \cdot (\pm\infty) = \mp\infty \tag{21.11}$$

where, by convention, we either read only the top signs or only the bottom signs from all $\pm$ and $\mp$ on the same line. We also define

$$(+\infty) + (+\infty) := +\infty \;\wedge\; (-\infty) + (-\infty) := -\infty \tag{21.12}$$

which show $-(+\infty) = -\infty$ and $-(-\infty) = +\infty$, and

$$(+\infty) \cdot (+\infty) := +\infty \;\wedge\; (+\infty) \cdot (-\infty) := -\infty \tag{21.13}$$

If need arises, we might at times also stipulate that

$$(\pm\infty)^{-1} := 0 \tag{21.14}$$

but this is not in the sense of the inverse element under multiplication. However, $\overline{\mathbb{R}}$ is no longer a field because expressions

$$+\infty + (-\infty), \quad -\infty + (+\infty), \quad 0 \cdot (\pm\infty) \tag{21.15}$$

are left *undefined*. In any case, the reader is warned to perform all algebraic operations involving the two infinities with extreme caution as errors are made easily.

21.2 Upper and lower limits.

Having extended supremum and infimum to all subsets of extended reals, we will now apply these concepts to sequences. Given a sequence $\{a_n\}_{n\in\mathbb{N}}$ of extended reals, for each $n \in \mathbb{N}$ we define the symbols

$$\sup_{m \geqslant n} a_m := \sup\big\{a_m\colon m \in \mathbb{N} \wedge n \leqslant m\big\} \tag{21.16}$$

and

$$\inf_{m \geqslant n} a_m := \inf\big\{a_m\colon m \in \mathbb{N} \wedge n \leqslant m\big\} \tag{21.17}$$

From $\{a_n\}_{n\in\mathbb{N}}$ we have thus generated the sequences of its suprema and infima,

$$\Big\{\sup_{m \geqslant n} a_m\Big\}_{n\in\mathbb{N}} \quad \text{and} \quad \Big\{\inf_{m \geqslant n} a_m\Big\}_{n\in\mathbb{N}} \tag{21.18}$$

that are non-increasing and non-decreasing, respectively, and will thus converge provided they are bounded. This is leads to:

Definition 21.3 (Limsup and liminf) *Given a sequence* $\{a_n\}_{n\in\mathbb{N}} \in \overline{\mathbb{R}}^{\mathbb{N}}$*, we define its* limes superior*, a.k.a. upper limit or limsup, by*

$$\limsup_{n\to\infty} a_n := \inf_{n\geqslant 0} \sup_{m\geqslant n} a_m \tag{21.19}$$

and its limes inferior*, a.k.a. lower limit or liminf, by*

$$\liminf_{n\to\infty} a_n := \sup_{n\geqslant 0} \inf_{m\geqslant n} a_m \tag{21.20}$$

Both upper and lower limits generally take values in $\overline{\mathbb{R}}$ even if $\{a_n\}_{n\in\mathbb{N}}$ is $\mathbb{R}$-valued. Note also that, by the monotonicity of the sequences (21.18) $n \geqslant 0$ in (21.19–21.20) could be replaced by $n \geqslant k$ for any $k \in \mathbb{N}$, and so the quantities depend only on the asymptotic properties of $\{a_n\}_{n\in\mathbb{N}}$ (meaning that changing any finite number of elements will not affect the upper and lower limits). The quantities are also naturally ordered:

Lemma 21.4 *For any* $\{a_n\}_{n\in\mathbb{N}} \in \overline{\mathbb{R}}^{\mathbb{N}}$*,*

$$\liminf_{n\to\infty} a_n \leqslant \limsup_{n\to\infty} a_n \tag{21.21}$$

Proof. We claim that

$$\forall n, k \in \mathbb{N}\colon \inf_{m\geqslant k} a_m \leqslant \sup_{m\geqslant n} a_m \tag{21.22}$$

To prove this we first note that the conclusion of (21.22) is TRUE if $n = k$ by (21.8). Now if $k \leqslant n$, we use this to get

$$\inf_{m\geqslant k} a_m \leqslant \inf_{m\geqslant n} a_m \leqslant \sup_{m\geqslant n} a_m \tag{21.23}$$

which holds by (21.7) because the set of indices involved in the first infimum is larger than that in the second infimum, while for $n \leqslant k$ we use

$$\inf_{m\geqslant k} a_m \leqslant \sup_{m\geqslant k} a_m \leqslant \sup_{m\geqslant n} a_m \tag{21.24}$$

where the same argument gives the bound between the two suprema. Since $\leqslant$ is a total ordering of $\mathbb{N}$, we have proved (21.22) in all cases.

From (21.22) we get that $\{\sup_{m\geqslant n} a_m\}_{n\in\mathbb{N}}$ are all upper bounds on $\{\inf_{m\geqslant n} a_m\}_{n\in\mathbb{N}}$. Lemma 9.6 along with (21.7) then gives

$$\begin{aligned} \sup_{n\geqslant 0} \inf_{m\geqslant n} a_m &= \sup\Big\{ \inf_{m\geqslant n} a_m \colon n \in \mathbb{N} \Big\} \\ &\leqslant \inf\Big\{ \sup_{m\geqslant n} a_m \colon n \in \mathbb{N} \Big\} = \inf_{n\geqslant 0} \sup_{m\geqslant n} a_m \end{aligned} \tag{21.25}$$

thus proving the desired inequality. □

21.3 Connection with convergence.

The sequences (21.18) squeeze the terms of the sequence $\{a_n\}_{n\in\mathbb{N}}$ between them, and the further we go along the sequence, the more squeezed that they get. It thus appears that

equality holding in (21.21) must correspond to the sequence having a limit. This is true, albeit under the additional assumption of boundedness:

Theorem 21.5 *Let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of reals. Then*

$$\lim_{n\to\infty} a_n \text{ exists } \Leftrightarrow \{a_n\}_{n\in\mathbb{N}} \text{ bounded } \wedge \liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n. \tag{21.26}$$

Moreover, when both sides are TRUE, then the limit on the left equals the common value of limsup and liminf on the right.

Proof of $\Rightarrow$ in (21.26). Suppose that $\{a_n\}_{n\in\mathbb{N}}$ has a limit and let us call the limit L. As sequential convergence in $\mathbb{R}$ arises from a metric, Lemmas 17.4 shows that $\{a_n\}_{n\in\mathbb{N}}$ is bounded, so it suffices to prove equality of limsup and liminf. Here we note that convergence to a limit means that for all $k \in \mathbb{N}$ there is $n_0 \in \mathbb{N}$ such that

$$\forall n \geqslant n_0 \colon \ |a_n - L| < \frac{1}{k+1} \tag{21.27}$$

This is rewritten as

$$\forall n \geqslant n_0 \colon \ L - \frac{1}{k+1} < a_n < L + \frac{1}{k+1} \tag{21.28}$$

Using the definitions (21.19–21.20) and Lemma 21.4 it follows that for all $n \geqslant n_0$,

$$L - \frac{1}{k+1} < \inf_{m\geqslant n} a_m \leqslant \liminf_{n\to\infty} a_n \leqslant \limsup_{n\to\infty} a_n \leqslant \sup_{m\geqslant n} a_m < L + \frac{1}{k+1} \tag{21.29}$$

But both limsup and liminf are finite by boundedness of $\{a_n\}_{n\in\mathbb{N}}$ and so can subtract one of the other to get

$$0 \leqslant \limsup_{n\to\infty} a_n - \liminf_{n\to\infty} a_n < \frac{2}{k+1} \tag{21.30}$$

By the Archimedean property of the reals (see Theorem 11.1), the only non-negative real number that is less than $\frac{2}{k+1}$ for all $k \in \mathbb{N}$ is zero and so

$$\limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n \tag{21.31}$$

as claimed on the right-hand side of (21.26). □

Proof of $\Leftarrow$ in (21.26). The argument is similar, albeit somewhat easier. Suppose $\{a_n\}_{n\in\mathbb{N}}$ is bounded and (21.31) holds. The common value L of the latter quantities is then $\mathbb{R}$-valued. Fix $k \in \mathbb{N}$. Then

$$\exists n_0 \in \mathbb{N} \colon \ \sup_{m\geqslant n_0} a_m < L + \frac{1}{k+1} \tag{21.32}$$

for otherwise $L + \frac{1}{k+1}$ would be a better lower bound on the supremum sequence and, similarly,

$$\exists \tilde{n}_0 \in \mathbb{N} \colon \ \inf_{m\geqslant \tilde{n}_0} a_m > L - \frac{1}{k+1} \tag{21.33}$$

But then

$$\forall m \geqslant \max\{n_0, \tilde{n}_0\} \colon \ L - \frac{1}{k+1} < a_m < L + \frac{1}{k+1} \tag{21.34}$$

Rewriting the inequalities on the right as $|a_m - L| < \frac{1}{k+1}$, we have proved that L is the limit of $\{a_n\}_{n\in\mathbb{N}}$. □

Definition 21.6 (Improper limit) *We say that a sequence $\{a_n\}_{n\in\mathbb{N}}$ of reals has an improper limit if*

$$\{a_n\}_{n\in\mathbb{N}} \text{ is unbounded} \quad \wedge \quad \liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n \tag{21.35}$$

Under (21.35), the common value of limsup and liminf is then necessarily $+\infty$ or $-\infty$ and the sequence $\{a_n\}_{n\in\mathbb{N}}$ is bounded either from above or from below (but not both). This permits us to extend the notation so that:

$$\lim_{n\to\infty} a_n = +\infty \quad \text{if } \{a_n\}_{n\in\mathbb{N}} \text{ is bounded from below and (21.35) holds} \tag{21.36}$$

and

$$\lim_{n\to\infty} a_n = -\infty \quad \text{if } \{a_n\}_{n\in\mathbb{N}} \text{ is bounded from above and (21.35) holds} \tag{21.37}$$

In this case we will at times say that the limit exists in $\overline{\mathbb{R}}$.

Improper limits do not conform to the definition of the limit in $\mathbb{R}$, which would imply that the sequence is Cauchy and bounded (under Euclidean metric), both of which fail for improper limits. However, they do become proper limits once we endow $\overline{\mathbb{R}}$ with a different metric, e.g.,

$$\tilde{\varrho}(x,y) := \Big|\frac{x}{1+|x|} - \frac{y}{1+|y|}\Big| \tag{21.38}$$

where we set $\frac{\pm\infty}{1+|\pm\infty|} := \pm 1$. (As shown in a homework assignment, in this metric $\overline{\mathbb{R}}$ is a completion of $\mathbb{R}$.) This makes saying that the limit exists in $\overline{\mathbb{R}}$ consistent with our earlier metric-space based definitions.

21.4 Manipulations with limits.

In order to conclude our general discussion of limits of real-valued sequences, we now recall some standard "rules" for computing with such limits:

Lemma 21.7 (Sum, Product and Quotient Rules) *Suppose $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are two sequences such that the limits*

$$A := \lim_{n\to\infty} a_n \quad \wedge \quad B := \lim_{n\to\infty} b_n \tag{21.39}$$

exist in $\overline{\mathbb{R}}$. Then:

(1) $\lim_{n\to\infty} (a_n + b_n)$ *exists and equals $A + B$,*
(2) *for any $c \in \mathbb{R}$, $\lim_{n\to\infty} ca_n$ exists and equals cA,*
(3) $\lim_{n\to\infty} a_n \cdot b_n$ *exists and equals $A \cdot B$,*
(4) *if $b_n \neq 0$ for all $n \in \mathbb{N}$ AND $B \neq 0$, then also*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{A}{B} \tag{21.40}$$

provided the expressions on the right are meaningful.

Proof of (4). To demonstrate the type of arguments one needs to use, let us prove part (4) for the case when A and B are finite. Since $b_n, B \neq 0$ we may write

$$\frac{a_n}{b_n} - \frac{A}{B} = \frac{Ba_n - Ab_n}{Bb_n} = \frac{B(a_n - A) - A(b_n - B)}{Bb_n} \tag{21.41}$$

The fact that $b_n \to B \neq 0$ ensures that, for some $n_0 \in \mathbb{N}$,

$$n \geqslant n_0 \;\Rightarrow\; |b_n| \geqslant |B|/2 \tag{21.42}$$

Using the triangle inequality and homogeneity of $|\cdot|$, for $n \geqslant n_0$ we then get

$$\left|\frac{a_n}{b_n} - \frac{A}{B}\right| \leqslant \frac{|a_n - A|}{|b_n|} + \frac{|A|}{|B|}\frac{|b_n - B|}{|b_n|} \leqslant 2\frac{|B| + |A|}{|B|^2}\big(|a_n - A| + |b_n - B|\big) \tag{21.43}$$

Now, thanks to $a_n \to A$ and $b_n \to B$, given $\epsilon > 0$ we can find $n_1 \in \mathbb{N}$ so that

$$\forall n \geqslant n_1 \colon |a_n - A| < \frac{1}{4}\frac{|B|^2}{|A| + |B| + 1}\,\epsilon \;\wedge\; |b_n - B| < \frac{1}{4}\frac{|B|^2}{|A| + |B| + 1}\,\epsilon \tag{21.44}$$

Hereby we conclude that

$$\forall n \geqslant \max\{n_0, n_1\} \colon \left|\frac{a_n}{b_n} - \frac{A}{B}\right| < \epsilon/2 + \epsilon/2 = \epsilon \tag{21.45}$$

proving the desired convergence. □

We leave the proof of remaining parts to the reader (do it!) while noting that, for $A, B \in \mathbb{R}$, the requirement that the right-hand side be meaningful is trivial except in (21.40), where we need to assume $B \neq 0$. Once one or both of A and B are infinite, we have to exclude expressions of the form (21.15).

We also get the very popular tool for proving existence of a limit:

Lemma 21.8 (Squeeze Theorem) *Suppose $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ and $\{c_n\}_{n\in\mathbb{N}}$ are $\overline{\mathbb{R}}$-valued sequences such that*

$$\forall n \in \mathbb{N} \colon\; b_n \leqslant a_n \leqslant c_n \tag{21.46}$$

If the limits in

$$L := \lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n \tag{21.47}$$

exist in $\overline{\mathbb{R}}$. Then

$$\lim_{n\to\infty} a_n = L \tag{21.48}$$

We leave proofs of these facts to a homework exercise. Further "rules" about "plugging the value" and those dealing with "indeterminate expressions" will be demonstrated once we have introduced the concept of continuous and differentiable functions.

There are also some "special" sequences whose limit is good to remember as these naturally come up in estimates. We state these in:

Lemma 21.9 *We have*

$$\forall q > 0 \colon \quad \lim_{n\to\infty} \frac{1}{n^q} = 0 \tag{21.49}$$

and

$$\forall q \in \mathbb{R}\; \forall b \in (0,1) \colon \quad \lim_{n\to\infty} n^q b^{-n} = 0 \tag{21.50}$$

Proof. Starting with (21.49), the sequence is non-increasing and bounded below by one so the limit exists. Calling the limit L, we have $0 \leqslant L$. If we had $L > 0$ then $L \leqslant n^{-q}$ for each $n \geqslant 1$ and so $\sup(\mathbb{N} \smallsetminus \{0\}) \leqslant L^{-1/q}$, which we showed to be FALSE in the proof of the Archimedean property of the reals (see Lemma 11.1). Hence $L = 0$ and (21.49) holds.

For (21.50) we denote $a_n := n^q b^n$ and observe that then

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \left(\frac{n+1}{n}\right)^q b = b \tag{21.51}$$

due to the fact that $1 \leqslant \frac{n+1}{n} = 1 + \frac{1}{n} \to 1$ by (21.49) and the Squeeze Theorem and so, by the product rule for the limits and the Squeeze Theorem, $1 \leqslant (\frac{n+1}{n})^q \leqslant (\frac{n+1}{n})^{\lceil q \rceil} \to 1$. The existence of the limit (21.51) now implies that

$$\forall p \in (b, 1)\, \exists n_0 \in \mathbb{N}\, \forall n \geqslant n_0 \colon \frac{a_{n+1}}{a_n} \leqslant p \tag{21.52}$$

By induction, this is checked to imply

$$\forall n \geqslant n_0 \colon a_n \leqslant (a_{n_0} p^{-n_0}) p^n \tag{21.53}$$

and so we are down to proving $\lim_{n\to\infty} p^n = 0$. The limit exists because $\{p^n\}_{n\in\mathbb{N}}$ is non-increasing and non-negative. Calling the limit x, the fact that $p^n \to x$ and $pp^n = p^{n+1} \to x$ implies $px = p$. Since $p \neq 1$, this forces $x = 0$. □

The important message to be learned from (21.50) is that exponential (a.k.a. geometric) decay is always stronger than any kind of polynomial growth.

22. INFINITE SERIES

We will now proceed discussing an interesting application of the concept of limit of real-valued sequences to infinite series.

22.1 Definition and examples.

Let us first settle on some notation. Given a sequence $\{a_n\}_{n\in\mathbb{N}}$ of real numbers, for each $n \in \mathbb{N}$ we can recursively define the symbol $\sum_{k=0}^{n} a_k$ by:

$$\sum_{k=0}^{0} a_k := a_0 \ \wedge \ \Big(\forall n \in \mathbb{N}\colon \ \sum_{k=0}^{n+1} a_k := a_{n+1} + \sum_{k=0}^{n} a_k\Big). \tag{22.1}$$

For the resulting sequence $\left\{\sum_{k=0}^{n} a_k\right\}_{n\in\mathbb{N}}$ of *partial sums* we then impose:

Definition 22.1 (Infinite series) *Given a sequence $\{a_n\}_{n\in\mathbb{N}}$ of reals, the infinite series $\sum_{k=0}^{\infty} a_k$ is said to be* convergent *(or* converges*) if $\lim_{n\to\infty} \sum_{k=0}^{n} a_k$ exists in $\mathbb{R}$. We then use the symbol of infinite series to denote the limit, i.e.,*

$$\sum_{k=0}^{\infty} a_k := \lim_{n\to\infty} \sum_{k=0}^{n} a_k. \tag{22.2}$$

If the limit does not exist (in $\mathbb{R}$), we say that the infinite series is divergent *(or* diverges*). In this case, the symbol of infinite series remains formal (i.e., without a numerical value).*

As we are basing our labeling on the naturals, we will typically "start" the summations at $n = 0$. However, other initial values of the summation come up as well with the definitions adapted accodingly.

There are very few examples for which the series is computable. One of these is the *geometric series*

$$\sum_{n=0}^{\infty} q^n \tag{22.3}$$

which (at this point) is a formal expression depending on parameter q called the *quotient*. Here we get:

Lemma 22.2 (Geometric series) *For each $q \in \mathbb{R}$ with $|q| < 1$, the geometric series* (22.3) *is convergent with*

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1-q}. \tag{22.4}$$

For $q \in \mathbb{R}$ with $|q| \geqslant 1$ the series is divergent.

Proof. Denote $s_n := \sum_{k=0}^{n} q^k$. Then $s_n + q^{n+1} = s_{n+1} = 1 + qs_n$, which gives $(1-q)s_n = 1 - q^{n+1}$. When $q = 1$ this equation contains no information but otherwise we get

$$\forall q \neq 1 \, \forall n \in \mathbb{N}\colon \quad \sum_{k=0}^{n} q^k = \frac{1-q^{n+1}}{1-q}\,. \tag{22.5}$$

For q with $|q| < 1$, we have $q^{n+1} \to 0$ and so the infinite series converges with the limit as in (22.4). On the other hand, for q with $|q| > 1$ as well as $q = -1$, the sequence $\{q^{n+1}\}_{n\in\mathbb{N}}$ does not converge and nor does the infinite series. The same applies to $q = 1$ (which was excluded from (22.5)) where $\sum_{k=0}^{n} q^k$ equals $n + 1$ that diverges as $n \to \infty$ as well. □

We note that the simplicity of the criterion for convergence of the geometric series is so simple, and the limit being readily computable, puts the geometric series at the center of many estimates and computations involving infinite series.

Building on the geometric series, our second example concerns the vary familiar expression of a real number using a decimal expansions. While quite intuitive and ubiquitous, the precise meaning of this expansion cannot be explained without the notion of the limit or, more accurately, infinite series. In order to state everything precisely, recall the symbol $\lfloor x \rfloor$ for lower-integer rounding of x defined by

$$\lfloor x \rfloor := \sup\{n \in \mathbb{Z} \colon n \leqslant x\}. \tag{22.6}$$

We then get:

Lemma 22.3 (Expansion of the reals) *Let $L \in \mathbb{N}$ be such that $L \geqslant 2$. For each $x \in [0, 1)$, define the sequence $\{x_n\}_{n\in\mathbb{N}}$ recursively by*

$$x_0 := x \ \wedge \ \Big(\forall n \in \mathbb{N} \colon x_{n+1} := Lx_n - \lfloor Lx_n \rfloor\Big) \tag{22.7}$$

and set $a_n := \lfloor Lx_n \rfloor$. Then $a_n \in \{0, 1, \ldots, L-1\}$ for all $n \in \mathbb{N}$ and

$$x = \sum_{n=0}^{\infty} \frac{a_n}{L^{n+1}} \tag{22.8}$$

where the series on the right is convergent.

Proof. Notice that, since $z - \lfloor z \rfloor \in [0, 1)$ for each $z \in \mathbb{R}$, we have $x_n \in [0, 1)$ for each $n \in \mathbb{N}$. Also note that $a_n \in \{0, 1, \ldots, L-1\}$ for each $n \in \mathbb{N}$. We claim

$$\forall n \in \mathbb{N} \colon \quad x = \frac{x_{n+1}}{L^{n+1}} + \sum_{k=0}^{n} \frac{a_k}{L^{k+1}} \tag{22.9}$$

This is checked readily for $n = 0$ and then proved by induction using the fact that

$$x_n = \frac{x_{n+1}}{L} + \frac{a_n}{L}. \tag{22.10}$$

(We leave the details to the reader.) Since $x_{n+1} \in [0, 1)$, from (22.9) we get

$$x - \frac{1}{L^{n+1}} \leqslant \sum_{k=0}^{n} \frac{a_k}{L^{k+1}} \leqslant x. \tag{22.11}$$

As $L > 1$, the left-hand side converges to x. By Lemma 14.7 and the Squeeze Theorem (Lemma 21.8), so do the partial sums in the middle. □

The construction in (22.7) can be easily visualized with the help of the long-division algorithm: the a_n's are the digits extracted in progressive divisions by L and x_n's are the

corresponding remainders. The number $x \in [0,1)$ can then be represented by a sequence of digits from $\{0,\dots,L-1\}$ written as

$$0.a_0a_1a_2\dots \tag{22.12}$$

All $x \in \mathbb{R}$ can written this way by adding the integer $\lfloor x \rfloor$ to the expression representing the number $x - \lfloor x \rfloor$.

We note that the construction (22.7) never outputs a sequence $\{a_n\}_{n\in\mathbb{N}}$ that ends with an infinite run of $(L-1)$'s. For instance, for base-10 expansions ($L := 10$), the number $0.099999\dots$ will thus never arise; instead, we get $0.1000\dots$ right away in the first step of the long division. (Prove this!) The map $x \mapsto \{a_n\}_{n\in\mathbb{N}}$ taking $[0,1)$ into $\{0,\dots,L-1\}^{\mathbb{N}}$ is thus not onto and $\{a_n\}_{n\in\mathbb{N}} \mapsto x$ defined by (22.8) is not injective. But the defect is not too serious as it concerns only a countable set (not even all rationals).

Another interesting aspect of decimal expansions is the subject of:

Lemma 22.4 *Let $x \in [0,1)$ and the sequence $\{a_n\}_{n\in\mathbb{N}} \in \{0,\dots,L-1\}^{\mathbb{N}}$ be as in Lemma 22.3. The following are equivalent:*

(1) *x is rational, $x \in \mathbb{Q}$,*
(2) *$\{a_n\}_{n\in\mathbb{N}}$ is eventually periodic, i.e.,*

$$\exists n_0 \in \mathbb{N}\, \exists p \in \mathbb{N}\, \forall n \in \mathbb{N}\colon\ p > 1 \ \wedge\ (n \geqslant n_0 \ \Rightarrow\ a_{n+p} = a_n) \tag{22.13}$$

Thus, the number $0.21\overline{345}$ is rational but $0.101001000100001000001\dots$ is not.

The above expansion is not the only way to represent reals by sequences of naturals. Another such representation for irrationals is the *continued-fraction expansion*

$$x = \cfrac{1}{a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \dots}}} \tag{22.14}$$

which is for $x \in (0,1) \setminus \mathbb{Q}$ is obtained by first constructing recursively a sequence $\{x_n\}_{n\in\mathbb{N}}$ of numbers in $(0,1)$ such that

$$x_0 = x \ \wedge\ \forall n \in \mathbb{N}\colon\ x_{n+1} = 1/x_n - \lfloor 1/x_n \rfloor \tag{22.15}$$

and then setting $a_n := \lfloor 1/x_n \rfloor$ for each $n \in \mathbb{N}$. (The restriction to $x \notin \mathbb{Q}$ ensures that $x_n \neq 0$ and $a_n \geqslant 1$ for all $n \in \mathbb{N}$.) While this may appear somewhat similar to the decimal expansions, there is no apparent connection to infinite series.

22.2 Criteria for convergence.

As noted above, most infinite series are not explicitly computable. (One standard exception is the series $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ which can be computed using partial-fraction expansion. Do it!) Therefore, in order to determine whether a series converges one has to resort to various general criteria. We will now discuss a few of these, starting with:

Lemma 22.5 (Necessary conditions for convergence) *Let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of reals such that $\sum_{n=0}^{\infty} a_n$ converges. Then*

$$\lim_{n\to\infty} a_n = 0. \tag{22.16}$$

Proof. Let $s_n := \sum_{k=0}^{n} a_k$. Then $a_n = s_n - s_{n-1}$. Under the assumption of convergence, the limit $L := \lim_{n\to\infty} s_n$ exists and equals the value of the infinite series. From the addition/subtraction rule for limits, we then get

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} (s_n - s_{n-1}) = \lim_{n\to\infty} s_n - \lim_{n\to\infty} s_{n-1} = L - L = 0. \tag{22.17}$$

This is the desired claim. □

We warn the reader that this is a *necessary* condition for convergence. Such conditions are typically used to rule out convergence, rather than to prove it. A *sufficient* condition for convergence is provided in:

Lemma 22.6 (Boundedness suffices for positive coefficients) *Suppose $\{a_n\}_{n\in\mathbb{N}}$ is such that $a_n \geqslant 0$ for each $n \in \mathbb{N}$. Then*

$$\sum_{n=0}^{\infty} a_n \text{ converges} \quad \Leftrightarrow \quad \left\{\sum_{k=0}^{n} a_k\right\}_{n\in\mathbb{N}} \text{ is bounded.} \tag{22.18}$$

Proof. The positivity requirement ensures that the sequence on the right of (22.18) is non-decreasing. Non-decreasing sequences converge if and only if they are bounded. □

Somewhat more useful is:

Lemma 22.7 (Comparison test) *Suppose $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$, $\{c_n\}_{n\in\mathbb{N}}$ are sequences with*

$$\forall n \in \mathbb{N}: \quad 0 \leqslant b_n \leqslant a_n \leqslant c_n. \tag{22.19}$$

Then

$$\sum_{n=0}^{\infty} c_n \text{ converges} \quad \Rightarrow \quad \sum_{n=0}^{\infty} a_n \text{ converges} \tag{22.20}$$

and

$$\sum_{n=0}^{\infty} b_n \text{ diverges} \quad \Rightarrow \quad \sum_{n=0}^{\infty} a_n \text{ diverges} \tag{22.21}$$

Proof. From (22.19) we have

$$\forall n \in \mathbb{N}: \quad \sum_{k=0}^{n} b_k \leqslant \sum_{k=0}^{n} a_k \leqslant \sum_{k=0}^{n} c_k. \tag{22.22}$$

The partial sums of series with non-negative coefficients form non-decreasing sequences which converge if and only if they are bounded. This readily yields (22.20–22.21). □

Using the comparison criterion, we readily conclude that the infinite series (22.8) converges for any choice of $\{a_n\}_{n\in\mathbb{N}}$ satisfying $a_n \in \{0, 1, \ldots, L-1\}$. Another use of the Comparison Test produces:

Lemma 22.8 (Harmonic series) $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n}$ *diverges.*

Proof. The idea of the proof is to bound the sequence

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}, \frac{1}{9}, \ldots \tag{22.23}$$

from below by the sequence

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}, \dots \tag{22.24}$$

and then notice that each block with the same denominator adds up to $1/2$.

Formally, this is done as follows: First note that for each $n \in \mathbb{N}$ with $n \geqslant 1$ there is exactly one $k \in \mathbb{N}$ such that $2^k \leqslant n < 2^{k+1}$. Then

$$2^k \leqslant n < 2^{k+1} \quad \Rightarrow \quad \frac{1}{n} > \frac{1}{2^{k+1}} \tag{22.25}$$

and so for all $m \geqslant 1$,

$$\sum_{n=1}^{2^m-1} \frac{1}{n} = \sum_{k=0}^{m-1} \sum_{n=2^k}^{2^{k+1}-1} \frac{1}{n} \geqslant \sum_{k=0}^{m-1} \sum_{n=2^k}^{2^{k+1}-1} \frac{1}{2^{k+1}} = \sum_{k=0}^{m-1} 2^k \frac{1}{2^{k+1}} = \frac{1}{2} m \tag{22.26}$$

The right hand side diverges as $m \to \infty$, which means that the sequence of partial sums for the harmonic contains a diverging subsequence, and is thus diverging itself. □

The same type of reasoning then also gives:

Lemma 22.9 *For each $p > 1$, the series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges.*

Proof. We follow a similar reasoning as in the previous lemma, but now aiming to prove convergence. Indeed, for any $m \geqslant 1$,

$$\sum_{n=1}^{2^m-1} \frac{1}{n^p} = \sum_{k=0}^{m-1} \sum_{n=2^k}^{2^{k+1}-1} \frac{1}{n^p} \leqslant \sum_{k=0}^{\infty} (2^{1-p})^k \tag{22.27}$$

The geometric series on the right converges because its quotient 2^{1-p} has absolute value less than one, due to $p > 1$. □

Note that, despite the sequence of coefficients decaying only *polynomially*, in both cases we ended up comparing the series to the geometric series (where *exponential* decay/growth is of concern). As noted earlier, this is a very common approach — and, usually, the first one to try — as it is guided by the fact that the geometric series has a simple convergence criterion and/or is explicitly computable.

We also note that the situation around the harmonic series can be further refined using similar methods. Indeed, we thus show that

$$\sum_{n=2}^{\infty} \frac{1}{n \log n} \text{ diverges} \quad \text{yet} \quad \forall p > 1 \colon \ \sum_{n=2}^{\infty} \frac{1}{n (\log n)^p} \text{ converges} \tag{22.28}$$

The case on the left can be further refined by adding $\log \log n$ terms, etc.

The next criterion is based on the equivalence of convergence and being Cauchy:

Lemma 22.10 (Cauchy criterion) *Let $\{a_n\}_{n \in \mathbb{N}}$ be a sequence of reals. Then*

$$\sum_{n=0}^{\infty} a_n \text{ converges} \quad \Leftrightarrow \quad \forall \epsilon > 0 \, \exists n_0 \in \mathbb{N} \, \forall m \geqslant n \geqslant n_0 \colon \ \left| \sum_{k=n}^{m} a_k \right| < \epsilon. \tag{22.29}$$

Proof. The convergence of $\sum_{n=0}^{\infty} a_n$ is defined by the existence of the limit of $\{\sum_{k=0}^{n} a_k\}_{n\in\mathbb{N}}$. This is equivalent to the sequence of partial sums being Cauchy. As

$$\sum_{k=0}^{m} a_k - \sum_{k=0}^{n-1} a_k = \sum_{k=n}^{m} a_k, \tag{22.30}$$

that is in turn equivalent to the condition on the right of (22.29). □

As a consequence of this we get:

Corollary 22.11 (Finitary changes are irrelevant for convergence) *If $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are sequences such that*

$$\{n \in \mathbb{N} : a_n \neq b_n\} \text{ is finite} \tag{22.31}$$

then

$$\sum_{n=0}^{\infty} a_n \text{ converges} \quad \Leftrightarrow \quad \sum_{n=0}^{\infty} b_n \text{ converges.} \tag{22.32}$$

Proof. The Cauchy criterion is not affected by changing the underlying sequence on a naturals bounded by some n' since we can always take n_0 larger than n'. □

Another consequence is:

Corollary 22.12 (Decaying tail) *Let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of reals. Then*

$$\sum_{n=0}^{\infty} a_n \text{ converges } \Rightarrow \left(\forall n \in \mathbb{N} : \sum_{k=n}^{\infty} a_k \text{ converges}\right) \wedge \lim_{n\to\infty} \sum_{k=n}^{\infty} a_k = 0. \tag{22.33}$$

Proof. Suppose $\sum_{n=0}^{\infty} a_n$ converges. Changing the first n terms of the underlying sequence to zero, Corollary 22.11 shows that $\sum_{k=n}^{\infty} a_k$ converges for each $n \in \mathbb{N}$. Then $|\sum_{k=n}^{m} a_k| \leqslant \epsilon$ for all $m \geqslant n$ implies $|\sum_{k=n}^{\infty} a_k| \leqslant \epsilon$. The gives the limit in (22.33). □

We will give other criteria for convergence when we discuss the notions of absolute and conditional convergence in the next lecture.

23. Absolute vs conditional convergence

Here we refine the concept of convergent series into absolutely convergent and conditionally convergent series. The former notion will later be appreciated once we discuss power series in the next quarter.

23.1 Absolute convergence.

Although the convergence of infinite series reduces to the notion of convergence sequences, the fact that we are writing the relevant sequence as a sum brings up the following natural questions: Can the sum of infinitely many numbers be performed in any order? And what if some of the terms are subtracted instead of being added? Such considerations naturally guide us towards the following concept:

Definition 23.1 (Absolute convergence) *We say that the infinite series $\sum_{n=0}^{\infty} a_n$ converges absolutely if $\sum_{n=0}^{\infty} |a_n|$ converges (in $\mathbb{R}$).*

We note that an infinite series with non-negative entries converges if and only if the sequence of partial sums is bounded. So absolute convergence is often stated as

$$\sum_{n=0}^{\infty} |a_n| < \infty. \tag{23.1}$$

The reader may also wonder why the term "convergence" is made part of the definition of "absolute convergence" as it refers to convergence of a different infinite series. That this is fine is the content of:

Lemma 23.2 *If an infinite series converges absolutely, then it converges.*

Proof. By the Cauchy criterion (Lemma 22.10), the convergence of the series $\sum_{n=0}^{\infty} |a_n|$ is equivalent to

$$\forall \epsilon > 0 \ \exists n_0 \in \mathbb{N} \ \forall n, m \in \mathbb{N}: \quad n \geqslant m \geqslant n_0 \ \Rightarrow \ \sum_{k=m}^{n} |a_k| < \epsilon. \tag{23.2}$$

The triangle inequality for absolute value gives

$$\Big| \sum_{k=m}^{n} a_k \Big| \leqslant \sum_{k=m}^{n} |a_k|, \tag{23.3}$$

and so (23.2) implies

$$\forall \epsilon > 0 \ \exists n_0 \in \mathbb{N} \ \forall n, m \in \mathbb{N}: \quad n \geqslant m \geqslant n_0 \ \Rightarrow \ \Big| \sum_{k=m}^{n} a_k \Big| < \epsilon. \tag{23.4}$$

By the Cauchy criterion again, $\sum_{n=0}^{\infty} a_n$ converges. □

A lot of properties of finite sums extends to infinite series as well; for instance:

Lemma 23.3 (Triangle inequality for infinite series) *Suppose $\sum_{n=0}^{\infty} a_n$ is absolutely convergent. Then*

$$\Big|\sum_{n=0}^{\infty} a_n\Big| \leqslant \sum_{n=0}^{\infty} |a_n| \tag{23.5}$$

We leave the proof of this easy lemma to homework. Note that the expression is meaningful although not very informative even without absolute convergence (we get $+\infty$ on the right-hand side); of course, we then still have to assume that $\sum_{k=0}^{\infty} a_k$ converges.

Moving forward on one of our questions above, we now note:

Theorem 23.4 *An absolutely convergent infinite series can be summed in any order with the same result. More precisely, if $\{a_n\}_{n\in\mathbb{N}}$ is a sequence of reals such that* (23.1) *holds, then for every bijection $\phi\colon \mathbb{N} \to \mathbb{N}$,*

$$\lim_{n\to\infty} \sum_{k=0}^{n} a_{\phi(k)} = \sum_{k=0}^{\infty} a_k. \tag{23.6}$$

Proof. Let $\phi\colon \mathbb{N} \to \mathbb{N}$ be a bijection. Fix $\epsilon > 0$. Since the series $\sum_{k=0}^{n} a_n$ converges absolutely, the Cauchy criterion gives $n_0 \geqslant 1$ such that

$$\forall n, m \in \mathbb{N}\colon \quad m \geqslant n \geqslant n_0 \;\Rightarrow\; \sum_{k=n}^{m} |a_k| < \epsilon. \tag{23.7}$$

Define $m_0 \in \mathbb{N}$ by

$$m_0 := \inf\big\{m \geqslant 0\colon \phi([0, n_0)) \subseteq [0, m)\big\} \tag{23.8}$$

The fact that ϕ is injective then forces $n_0 \leqslant m_0$ and we have $\phi([0, n_0)) \subseteq [0, m_0)$. Using that ϕ is bijective we get that, for each $m \geqslant m_0$ (which implies $m \geqslant n_0$), the terms $a_0, \ldots, a_{n_0-1}$ appear in both sums in

$$\sum_{k=0}^{m} a_{\phi(k)} - \sum_{k=0}^{m} a_k \tag{23.9}$$

and thus cancel out from the expression, while the terms $a_{n_0}, \ldots, a_{m_1}$ appear at most twice there. Using also (23.7) it follows that

$$\forall m \geqslant m_0\colon \quad \Big|\sum_{k=0}^{m} a_{\phi(k)} - \sum_{k=0}^{m} a_k\Big| \leqslant \sum_{k=n_0}^{m} 2|a_k| < 2\epsilon. \tag{23.10}$$

As $\{\sum_{k=0}^{m} a_k\}_{m\in\mathbb{N}}$ converges to $\sum_{k=0}^{\infty} a_k$, from (23.3) and (23.7) we get

$$\forall m \geqslant m_0\colon \quad \Big|\sum_{k=0}^{m} a_k - \sum_{k=0}^{\infty} a_k\Big| \leqslant \epsilon. \tag{23.11}$$

Using the triangle inequality we conclude

$$\forall m \geqslant m_0\colon \quad \Big|\sum_{k=0}^{m} a_{\phi(k)} - \sum_{k=0}^{\infty} a_k\Big| < 2\epsilon + \epsilon = 3\epsilon. \tag{23.12}$$

Since ϵ was arbitrary, this proves that $\sum_{k=0}^{\infty} a_{\phi(k)}$ converges to $\sum_{k=0}^{\infty} a_k$. □

The fact that the order of summation does not matter for absolutely convergent series underlies the proof that various standard manipulations with finite sums apply to infinite series. One of the useful manipulations concerns the product of two infinite series:

Theorem 23.5 (Merten's theorem for Cauchy product) *Let $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ be sequences of reals such that $\sum_{n=0}^{\infty} a_n$ is convergent and $\sum_{n=0}^{\infty} b_n$ is absolutely convergent. For each $n \in \mathbb{N}$, set*

$$c_n := \sum_{k=0}^{n} a_k b_{n-k} \tag{23.13}$$

The series $\sum_{n=0}^{\infty} c_n$ is then convergent as well and

$$\sum_{n=0}^{\infty} c_n = \Big(\sum_{n=0}^{\infty} a_n\Big)\Big(\sum_{n=0}^{\infty} b_n\Big) \tag{23.14}$$

If both $\sum_{n=0}^{\infty} a_n$ and $\sum_{n=0}^{\infty} b_n$ converge absolutely, then so does $\sum_{n=0}^{\infty} c_n$.

Proof. Assume that $\sum_{n=0}^{\infty} b_n$ converges absolutely and $\sum_{n=0}^{\infty} a_n$ converges. Let $n \in \mathbb{N}$. A simple rearrangement of the sums shows

$$\sum_{k=0}^{n} c_k = \sum_{k=0}^{n} b_k \sum_{j=0}^{n-k} a_j \tag{23.15}$$

Hereby we get

$$\Big(\sum_{k=0}^{n} b_k\Big)\Big(\sum_{j=0}^{n} a_j\Big) - \sum_{k=0}^{n} c_k = \sum_{k=1}^{n} b_k \sum_{j=n-k+1}^{n} a_j \tag{23.16}$$

Since $\sum_{n=0}^{\infty} a_n$ and $\sum_{n=0}^{\infty} |b_n|$ converge, given $\epsilon > 0$, the Cauchy criterion gives existence of $k_0 \in \mathbb{N}$ such that

$$\forall m \geqslant n \geqslant k_0 \colon \ \Big|\sum_{k=n}^{m} a_k\Big| < \epsilon \ \wedge \ \sum_{k=m}^{n} |b_k| < \epsilon \tag{23.17}$$

The convergence also implies that

$$a := \sup_{m\geqslant n\geqslant 0}\Big|\sum_{k=n}^{m} a_k\Big| < \infty \ \wedge \ b := \sum_{k=0}^{\infty} |b_k| < \infty \tag{23.18}$$

Assuming $n \geqslant 2n_0$, we then have

$$\begin{aligned}\Big|\Big(\sum_{k=0}^{n} b_k\Big)\Big(\sum_{j=0}^{n} a_j\Big) - \sum_{k=0}^{n} c_k\Big| &\leqslant \sum_{k=0}^{n} |b_k|\Big|\sum_{j=n-k+1}^{n} a_j\Big| \\ &\leqslant \Big(\sum_{k=0}^{\lfloor n/2\rfloor} |b_k|\Big)\Big|\sum_{j=n_0}^{n} a_j\Big| + \sum_{k=\lfloor n/2\rfloor}^{n} |b_k|\Big|\sum_{j=n-k+1}^{n} a_j\Big| \leqslant b\epsilon + \epsilon a = \epsilon(a+b)\end{aligned} \tag{23.19}$$

Using $\sum_{k=0}^{n} a_k \to A := \sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{n} b_k \to B := \sum_{k=0}^{\infty} b_k$, this shows

$$AB - \epsilon(a+b) \leqslant \liminf_{n\to\infty} \sum_{k=0}^{n} c_k \leqslant \limsup_{n\to\infty} \sum_{k=0}^{n} c_k \leqslant AB + \epsilon(a+b) \tag{23.20}$$

Since ϵ is arbitrary positive, this shows equality of the *limes superior* and *limes inferior* and, consequently, $\sum_{k=0}^{n} c_k \to AB$. For the class clause note that $|c_n| \leqslant \sum_{k=0}^{n} |a_k||b_{n-k}|$ so the claim follows from (23.14) with a_n replaced by $|a_n|$ and b_n by $|b_n|$.

The last clause is proved by repeating the arguments with $|a_n|$ and $|b_n|$ instead of a_n and b_n (although a shorter and more direct argument is possible). □

Remark 23.6 The previous proof is considerably easier when both series converge absolutely. Indeed, (23.16) gives

$$\begin{aligned}\Big|\Big(\sum_{k=0}^{n} b_k\Big)\Big(\sum_{j=0}^{n} a_j\Big) - \sum_{k=0}^{n} c_k\Big| &\leqslant \sum_{k=0}^{n} |b_k| \sum_{j=n-k+1}^{n} |a_j| \\ &\leqslant \Big(\sum_{k=0}^{n} |b_k|\Big)\Big(\sum_{j=\lfloor n/2\rfloor}^{n} |a_j|\Big) + \Big(\sum_{k=\lfloor n/2\rfloor}^{n} |b_k|\Big)\Big(\sum_{j=0}^{n} |a_j|\Big)\end{aligned} \tag{23.21}$$

and the right-hand side then tends to zero as $n \to \infty$ by the "decaying tail" property of convergence series; cf Corollary 22.12. The same argument applies for series of absolute values, which by the inequality

$$|c_n| \leqslant \sum_{k=0}^{n} |a_k||b_{n-k}| \tag{23.22}$$

also shows absolute convergence of $\sum_{k=0}^{\infty} c_k$.

23.2 Conditional convergence.

The reliance on absolute convergence in above statements is not merely a convenience of proofs. In order to demonstrate that, introduce the following concept:

Definition 23.7 (Conditional convergence) *We say that the infinite series $\sum_{n=0}^{\infty} a_n$ converges conditionally if*

$$\sum_{n=0}^{\infty} a_n \text{ converges} \quad \wedge \quad \sum_{n=0}^{\infty} |a_n| \text{ diverges} \tag{23.23}$$

We have thus separated convergent series into those that are absolutely convergent and those that are (only) conditionally convergent. An example of a conditionally convergent series is

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \tag{23.24}$$

Since the harmonic series diverges (see Lemma 22.8), this series definitely fails to converge absolutely. But it converges conditionally, thanks to even numbered partial sums converging in light of

$$\sum_{k=1}^{2n} \frac{(-1)^{n-1}}{k} = \sum_{k=1}^{n} \Big(\frac{1}{2k-1} - \frac{1}{2k}\Big) = \sum_{k=1}^{n} \frac{1}{(2k-1)2k} \leqslant \sum_{k=1}^{n} \frac{1}{4k^2} \tag{23.25}$$

where the series on the right converges by Lemma 22.9, and thanks to

$$\Big|\sum_{k=1}^{2n-1}\frac{(-1)^{n-1}}{k}-\sum_{k=1}^{2n}\frac{(-1)^{n-1}}{k}\Big|\leqslant\frac{1}{2n}\underset{n\to\infty}{\longrightarrow}0 \tag{23.26}$$

which shows that the odd-numbered partial sums converge to the same limit as the even-numbered ones. This example is actually a special case of a general fact:

Lemma 23.8 (Alternating series) *Let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence such that*

$$\forall n\in\mathbb{N}\colon 0\leqslant a_n\ \wedge\ a_{n+1}\leqslant a_n\ \wedge\ \lim_{n\to\infty}a_n=0 \tag{23.27}$$

Then the alternating series $\sum_{n=0}^{\infty}(-1)^n a_n$ converges.

Proof. Define the even and odd partial sums by

$$e_n:=\sum_{k=0}^{2n}(-1)^k a_k\quad\text{and}\quad o_n:=\sum_{k=0}^{2n+1}(-1)^k a_k \tag{23.28}$$

Now observe that the assumptions (23.27) give that, for all $n\in\mathbb{N}$

$$\begin{aligned}e_{n+1}&=e_n+a_{2n+2}-a_{2n+1}\geqslant e_n\\ o_{n+1}&=o_n-a_{2n+3}+a_{2n+1}\leqslant o_n\end{aligned} \tag{23.29}$$

as well as

$$o_n=e_n-a_{2n+1}\leqslant e_n \tag{23.30}$$

and, by $|o_n-e_n|=a_{2n+1}$,

$$\lim_{n\to\infty}(e_n-o_n)=0 \tag{23.31}$$

Hereby we get that $\{o_n\}_{n\in\mathbb{N}}$ is non-increasing and bounded below by any element of $\{o_n\}_{n\in\mathbb{N}}$ and, similarly, $\{e_n\}_{n\in\mathbb{N}}$ is non-decreasing and bounded from below by any element of $\{o_n\}_{n\in\mathbb{N}}$. It follows that both sequences converge and, using (23.31), their limits coincide. This now implies the claim. □

We leave the easy proof of this lemma to homework while noting that although the conditions on $\{a_n\}_{n\in\mathbb{N}}$ require that $a_n\to 0$, which we know to be necessary for convergence of $\sum_{n=0}^{\infty}(-1)^n a_n$, apart from positivity and monotonicity they do not require anything else. Thus, there are many examples where $\sum_{n=0}^{\infty}a_n$ diverges while $\sum_{n=0}^{\infty}(-1)^n a_n$ converges (albeit, by definition, only conditionally).

A similar idea underlies an example which shows that we *cannot* apply the Cauchy product formula (23.14) to series neither of which converge absolutely. Indeed, taking $a_n=b_n:=(-1)^n/\sqrt{n}$ for $n=1$ and $a_0=b_0=0$ in (23.13) shows $c_0=0$ and, for $n\geqslant 1$,

$$c_n=(-1)^n\sum_{k=1}^{n-1}\frac{1}{\sqrt{k(n-k)}} \tag{23.32}$$

Since at least one of k or $n-k$ is at least $\lfloor n/2\rfloor$, hereby we get

$$|c_n|\geqslant\frac{1}{\sqrt{n/2}}\sum_{k=1}^{\lfloor n/2\rfloor}k^{-1/2} \tag{23.33}$$

where (by a similar reasoning underlying the proof of Lemma 22.9) the sum is at least a constant times $\sqrt{n}$. Hence c_n does not tend to zero as $n \to \infty$ and so $\sum_{k=1}^{n} c_k$ fails to converge by Lemma 22.5.

As our last counterexample, we show that not even Theorem 23.4 holds for conditionally convergent sequences. In fact, we have:

Theorem 23.9 (Riemann's rearrangement theorem) *Suppose $\sum_{n=0}^{\infty} a_n$ converges conditionally (and thus not absolutely). Then for each $x \in \mathbb{R}$ there is a bijection $\phi\colon \mathbb{N} \to \mathbb{N}$ such that*

$$\lim_{n\to\infty} \sum_{k=0}^{n} a_{\phi(k)} = x. \tag{23.34}$$

In short, conditionally convergent series can be rearranged to converge to any real number.

The proof hinges on the following observation:

Lemma 23.10 *Suppose $\sum_{n=0}^{\infty} a_n$ converges conditionally (and thus not absolutely). Define*

$$a_n^+ := \max\{a_n, 0\} \quad \textit{and} \quad a_n^- := \max\{-a_n, 0\}. \tag{23.35}$$

Then

$$\sum_{n=0}^{\infty} a_n^+ \text{ diverges} \quad \wedge \quad \sum_{n=0}^{\infty} a_n^- \text{ diverges}. \tag{23.36}$$

Proof. Note that $a_n = a_n^+ - a_n^-$ while $|a_n| = a_n^+ + a_n^-$. The lack of absolute convergence means that at least one of the series in (23.36) diverges. Since

$$\sum_{n=0}^{\infty} a_n^+ = \sum_{n=0}^{\infty} a_n^- + \sum_{n=0}^{\infty} a_n \tag{23.37}$$

where the series on the right converges, once one of the series in (23.36) diverges, so must the other. □

Proof of Theorem 23.9. Pick $x \in \mathbb{R}$. The main idea is quite simple: We will start listing the non-negative terms of $\{a_n\}_{n\in\mathbb{N}}$ in the given order until their sum first exceeds x. Then we start listing the negative terms of $\{a_n\}_{n\in\mathbb{N}}$ (starting from the first one) until the sum of all terms so far first drops again under x. Then we start listing the positive terms again, and then the negative terms, etc until all terms have been listed. (That we never fail to reach x is the consequence of (23.36).) Since x is overshot by at most $|a_n|$, for a_n being the last term added, the fact that $a_n \to 0$ as implied by convergence of $\sum_{k=0}^{\infty} a_k$ then shows that the partial sums of thus rearranged series tend to x, as desired.

The formal construction of the bijection ϕ requires introduction of there auxiliary sequences $\{n_k\}_{k\in\mathbb{N}}$, $\{m_k\}_{k\in\mathbb{N}}$ and $\{s_k\}_{k\in\mathbb{N}}$. These are defined recursively by

$$n_0 := 0 \ \wedge \ m_0 := 0 \ \wedge \ s_0 := a_0 \ \wedge \ \phi(0) := 0 \tag{23.38}$$

and, for all $k \in \mathbb{N}$,

$$s_k \leqslant x \ \Rightarrow \ \begin{cases} n_{k+1} := \inf\{n > n_k\colon a_n \geqslant 0\} \ \wedge \ m_{k+1} := m_k \\ s_{k+1} := s_k + a_{n_k} \ \wedge \ \phi(k+1) := n_{k+1} \end{cases} \tag{23.39}$$

and

$$x < s_k \Rightarrow \begin{cases} n_{k+1} := n_k \wedge m_{k+1} := \inf\{m > m_k \colon a_m < 0\} \\ s_{k+1} := s_k + a_{m_{k+1}} \wedge \phi(k+1) := m_{k+1} \end{cases} \tag{23.40}$$

Here we note that (23.36) implies that $\{a_n\}_{n\in\mathbb{N}}$ has infinitely many non-negative terms and infinitely many negative terms, and so the infima in (23.39–23.40) are well defined. Both sequences are non-decreasing (we will show that they both tend to infinity in the proof of surjectivity of ϕ).

It remains to check that ϕ is a bijection and that (23.34) holds. For injectivity note that $\phi(k) = \phi(j)$ implies that either $a_{\phi(k)}$ and $a_{\phi(j)}$ are both non-negative and so $\phi(k) = n_k \wedge \phi(j) = n_j$ by (23.39), or $a_{\phi(k)}$ and $a_{\phi(j)}$ are both negative and so $\phi(k) = m_k \wedge \phi(j) = m_j$ by (23.40). But $n_k = n_j$ with $k < j$ implies that $a_{n_j} < 0 \leqslant a_{n_k}$ by (23.39), while $m_k = m_j$ with $k < j$ implies $a_{m_j} \geqslant 0 > a_{m_k}$ by (23.40), a contradiction. We conclude that $\phi(k) = \phi(j)$ implies $k = j$ and so ϕ is injective.

To prove that ϕ is surjective, assume $\mathrm{Ran}(\phi) \neq \mathbb{N}$ and let $n := \inf \mathrm{Ran}(\phi)$. If $a_n \geqslant 0$, then the fact that $\{n_k\}_{k\in\mathbb{N}}$ is non-decreasing implies that it is bounded by n. This means that the alternative (23.40) occurs from some k on, showing that

$$x \leqslant s_k + \sum_{j=k+1}^{\infty} a_{m_j} = s_k - \sum_{j=m_k+1} a_j^- \tag{23.41}$$

in contradiction with the second part of (23.36). The case $a_n < 0$ is handled similarly and so we omit it.

Finally, to show that the partial sums converge, let $\epsilon > 0$ and, noting that $a_n \to 0$ by the fact that the series $\sum_{n=0}^{\infty} a_n$ converges, let $q_0 \in \mathbb{N}$ be such that $\forall q \geqslant q_0 \colon |a_q| < \epsilon$. Set

$$k_0 := \inf\{k \geqslant 1 \colon n_k \geqslant q_0 \wedge m_k \geqslant q_0 \wedge s_k > x > s_{k+1}\} \tag{23.42}$$

This is the first index such that the sequence $\{s_k\}_{k\in\mathbb{N}}$ "crossed" level x downward and all a_k's past this index are already smaller than ϵ. The construction then ensures

$$\forall k \geqslant k_0 \colon x - \epsilon \leqslant s_k = \sum_{j=0}^{k} a_{\phi(j)} \leqslant x + \epsilon \tag{23.43}$$

and so we get the desired claim. □

Remark 23.11 An inspection of the above proof reveals that the assumption of convergence of $\sum_{n=0}^{\infty} a_n$ is not really used. All we need is that $a_n \to 0$ and that (23.36) hold.

23.3 Tests for absolute convergence.

We finish this section by listing some criteria for proving absolute convergence known, very likely, already from Calculus. The first one is:

Lemma 23.12 (Comparison test) *Suppose $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are sequences such that*

$$\Big(\forall n \in \mathbb{N} \colon \ |a_n| \leqslant b_n\Big) \quad \wedge \quad \sum_{n=0}^{\infty} b_n < \infty \tag{23.44}$$

Then $\sum_{n=0}^{\infty} a_n$ converges absolutely.

Proof. This follows from Lemma 22.7 with a_n replaced by $|a_n|$ (and $b_n := 0$). □

A first try at the dominating sequence is the geometric progression. This leads to two limit criteria well-known called the Ratio and Root Test in calculus (albeit generalized by invoking *limes superior* instead of a plain limit). Let us start with:

Lemma 23.13 (Ratio test, convergence part) *Suppose $\{a_n\}_{n\in\mathbb{N}}$ is a sequence such that*

$$\forall n \in \mathbb{N}\colon\ a_n \neq 0 \tag{23.45}$$

and

$$\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1 \tag{23.46}$$

Then $\sum_{n=0}^{\infty} a_n$ converges absolutely.

Proof. The properties of *limes superior* implies

$$\exists n_0 \in \mathbb{N}\colon \quad q := \sup_{n\geqslant n_0}\left|\frac{a_{n+1}}{a_n}\right| < 1 \tag{23.47}$$

By induction we then infer

$$\forall n \geqslant n_0\colon \quad |a_n| \leqslant q^{n-n_0}|a_{n_0}| = q^n\left(q^{-n_0}|a_{n_0}|\right) \tag{23.48}$$

Since $|a_n| = q^n(q^{-n}|a_n|)$, hence we get

$$\forall n \in \mathbb{N}\colon \quad |a_n| \leqslant q^n \max_{k=0,\ldots,n_0}\left(q^{-k}|a_k|\right) \tag{23.49}$$

Denoting the term on the right-hand side by c_n, the fact that $q < 1$ ensures that $\sum_{n=0}^{\infty} c_n$ converges. By Lemma 23.12, the series $\sum_{n=0}^{\infty} a_n$ converges absolutely. □

The Ratio Test is inconvenient for two reasons: First, we need to require that $a_n \neq 0$. Second, the failure of the condition (23.46) does not signify divergence of the series. Indeed, for a sequence $\{a_n\}_{n\in\mathbb{N}}$ such that

$$\forall n \in \mathbb{N}\colon \quad a_{2n} = \left(\frac{1}{3}\right)^{2n} \quad \wedge \quad a_{2n+1} = \left(\frac{1}{2}\right)^{2n} \tag{23.50}$$

the series $\sum_{n=0}^{\infty} a_n$ converges absolutely yet the *limes superior* in (23.46) is infinite. This can be mended by requiring that the limit exists or by using *limes inferior* instead:

Lemma 23.14 (Ratio test, divergence part) *Suppose $\{a_n\}_{n\in\mathbb{N}}$ is a sequence such that*

$$\forall n \in \mathbb{N}\colon\ a_n \neq 0 \tag{23.51}$$

and

$$\liminf_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| > 1 \tag{23.52}$$

Then $\sum_{n=0}^{\infty} a_n$ diverges.

We leave the proof of this lemma to the reader. Instead we move to:

Lemma 23.15 (Root test) *Suppose $\{a_n\}_{n\in\mathbb{N}}$ is a sequence of reals. Then*

$$\limsup_{n\to\infty} \sqrt[n]{|a_n|} < 1 \quad \Rightarrow \quad \sum_{n=0}^{\infty} a_n \ \text{ converges absolutely} \tag{23.53}$$

while

$$\limsup_{n\to\infty} \sqrt[n]{|a_n|} > 1 \quad \Rightarrow \quad \sum_{n=0}^{\infty} a_n \text{ diverges} \tag{23.54}$$

Proof. As for the Ratio Test, we again dominate the series $\sum_{n=0}^{\infty} a_n$ by a geometric series. Assume first $\limsup_{n\to\infty} \sqrt[n]{|a_n|} < 1$. Then

$$\exists n_0 \in \mathbb{N}: \quad q := \sup_{n\geqslant n_0} \sqrt[n]{|a_n|} < 1 \tag{23.55}$$

It follows that $|a_n| \leqslant q^n$ for all $n \geqslant n_0$ and thus

$$\forall n \geqslant n_0: \quad |a_n| \leqslant q^n \max_{k=0,\dots,n_0} \left(q^{-k}|a_k|\right). \tag{23.56}$$

As $q < 0$, Lemma 23.12, the series $\sum_{n=0}^{\infty} a_n$ converges absolutely.

Next let us assume $\limsup_{n\to\infty} \sqrt[n]{|a_n|} > 1$. Then there exists a strictly increasing sequence $\{n_k\}_{k\in\mathbb{N}}$ such that

$$\forall k \in \mathbb{N}: \quad \sqrt[n_k]{|a_{n_k}|} \geqslant 1 \tag{23.57}$$

This means that the necessary condition $a_n \to 0$ from Lemma 22.5 for convergence of $\sum_{n=0}^{\infty} a_n$ fails, proving that the series diverges. □

We remark that if the ratio test applies, then so does the root test. The root test is particularly useful for power series, i.e., series of the form $\sum_{n=0}^{\infty} a_n x^n$, where x is a real or complex variable. We will discuss these next quarter after we have covered absolute convergence.

Both ratio and root tests are based on comparison to the geometric series. Neither test is exhaustive because no conclusion is made when the *limes superior* in (23.53–23.54) equals one. In this case a more elaborate comparison (usually, to a polynomially decaying series) is made or some other analytic tools have to be invoked to decide convergence or divergence. There are also necessary and sufficient conditions for convergence (e.g., one bearing the name of Kummer) but we will not go over these here.