Analysis: Foundations and Selected Topics

Work in progress, last updated: May 29, 2026

387 pages, 21 figures

Joris Roos and Andreas Seeger

Department of Mathematics and Statistics, University of Massachusetts Lowell, Lowell, MA, USA
Email address: joris_roos@uml.edu

Department of Mathematics, University of Wisconsin-Madison, Madison, WI, USA
Email address: seeger@math.wisc.edu

Contents

Preface 5

Chapter 1. Fundamentals 7
1.1. Logic and sets 7
1.2. Numbers 17
1.3. Sequences 22
1.4. Continuity 31
1.5. Further exercises 35
1.6. Appendix: Complete ordered fields 36

Chapter 2. Single variable calculus 51
2.1. Integration 51
2.2. Differentiation 58
2.3. Fundamental theorem of calculus 63
2.4. Taylor's theorem 65
2.5. Transcendental functions 66
2.6. Improper integrals 71
2.7. Complex numbers 75
2.8. Uniform convergence 77
2.9. Further exercises 82

Chapter 3. Sums and series 91
3.1. Convergence of series 91
3.2. Power series 107
3.3. Infinite products 117
3.4. Nested radicals 121
3.5. The Euler–Maclaurin formula 125
3.6. Further exercises 133

Chapter 4. Metric spaces 143
4.1. Topology 144
4.2. The contraction principle 151
4.3. Compactness 153
4.4. Densely defined functions 172
4.5. Completion 174
4.6. Further exercises 175

Chapter 5. Linear operators 181
5.1. Norms 181
5.2. Bounded linear operators 183
5.3. Finite dimensional spaces 186
5.4. The uniform boundedness principle 188
5.5. ℓ^p spaces 192

5.6. Inner product spaces 196
5.7. Hahn–Banach theorem 202
5.8. Further exercises 210

Chapter 6. Differential calculus 215
6.1. Derivatives 215
6.2. Inverse function theorem 225
6.3. Derivatives of higher order 235
6.4. Extreme values and optimization 245
6.5. Ordinary differential equations 254
6.6. Further exercises 263
6.7. Appendix: Higher order differentials 271

Chapter 7. Approximation of functions 279
7.1. Weierstrass' theorem 280
7.2. The Stone-Weierstrass theorem 282
7.3. Orthogonality 286
7.4. Trigonometric polynomials 295
7.5. More on convergence of Fourier series 307
7.6. Best uniform approximation by polynomials 315
7.7. Chebyshev polynomials 318
7.8. Interpolation of functions 321
7.9. Bernstein and Markov inequalities 339
7.10. The Müntz approximation theorem 341
7.11. Further exercises 347

Chapter 8. Baire's theorem and consequences 359
8.1. The Baire category theorem 359
8.2. Nowhere differentiable continuous functions 361
8.3. Sets of continuity 362
8.4. Baire functions 363
8.5. The Banach-Steinhaus theorem 364
8.6. Open mappings 365
8.7. Further exercises 369

Hints for selected exercises 373

Index 377

Bibliography 385

Preface

This book provides an introduction to real analysis from first principles aimed at motivated undergraduate and beginning graduate students. The book contains the standard material of typical first and second courses in real analysis, as well as a number of selected topics providing many additional examples beyond the typical content of introductory analysis courses. For example, Chapter 7 could form the basis of a first course on approximation theory. Chapters 5 and 8 could be used as an introduction to functional analysis.

We omit the Lebesgue theory of measure and integration in this book. Instead we focus on including a large number of optional interesting topics that illlustrate the core material. The elementary Riemann integral is sufficient for all topics we discuss.

The large number of exercises should make the text suitable for self-study. Most of the book has no prerequisites beyond a thorough high school math curriculum, except that in parts of Chapters 5-8 we assume some familiarity with basic concepts from linear algebra.

The book is an outgrowth of various lecture notes originally written for Math 522 – a second course in analysis at the University of Wisconsin–Madison. It has since significantly expanded in scope and now contains much more material than can be covered in one or two semesters of lectures. Here are rough suggestions for typical one-semester courses that could be taught using the book:

- Analysis I: Ch. 1.1-1.4, Ch. 2.1-2.8, 3.1-3.2.
- Analysis II: Ch. 2.8, 3.1-3.2, 4.1-4.3, 5.1-5.3, 5.6.1, 6.1-6.3, 6.5, 7.1-7.3, 7.4.1, 8.1-8.3.
- Introduction to functional analysis: Ch. 5 and 8 with selected applications from Ch. 6, 7.
- Approximation theory: Ch. 5.6 and 7 with review of other background material as needed.

Acknowledgements. We thank many former students of several iterations of Math 522 for pointing out typos and inaccuracies. We also thank David Beltran, Xianghong Gong, Lars Niedorf, Ben Seeger, Betsy Stovall, Brian Street, Sarah Tammen, Juan Tolosa, Hung Tran, and Billy Wehring for valuable comments. J.R. was partially supported by NSF grant DMS-2154835, and A.S. was partially supported by NSF grant DMS-2348797.

CHAPTER 1

Fundamentals

1.1. Logic and sets

There are essentially two different things that we do in mathematics: we formulate *definitions* that precisely define mathematical objects, and we prove *theorems* – precise logical statements about these objects. Every definition must build on previous definitions so that unraveling the hierarchy of definitions inevitably leads to certain "primitive" objects. The existence and properties of these primitive objects must be postulated in the form of *axioms*, statements that we accept to be true without proof. Several alternative approaches for laying the foundations of mathematics have been proposed.

The historically most widely accepted primitive is that of a *set*. The naive "definition" of a set is that it is a collection of distinct elements drawn from a certain *"universe of discourse"* consisting of all objects we are interested in within a given context (such as "natural numbers", or "vector spaces", or "continuous functions"). A set A consisting of those elements x that have a property $P(x)$ is denoted

$$A = \{x \,:\, P(x)\}.$$

The "universe" from which x is drawn may itself be a set; it is usually clear from context and is often omitted from notation. If x is an element of A, we also write $x \in A$ and if not, $x \notin A$. It is important to recall that sets do not feature multiplicity of elements, i.e. the set notations $\{a\}$ and $\{a, a\}$ define the same set.

We should be aware that the notion of set we have just introduced is unfortunately not very sound. Inconsistencies arise as soon as we allow elements to be sets themselves. For example, *Russell's paradox* involves the "set" of all sets that are not elements of themselves, i.e.

$$A = \{x \,:\, x \notin x\}$$

Is this set an element of itself, i.e. is $A \in A$? Of course this question has no meaning. Any good theory of sets should not allow such logical abominations to be called sets – but then what exactly should the rules governing the formation of sets be? This is a very subtle problem and *set theory* is its own field of mathematics beyond the scope of this text. The interested reader should consult e.g. Halmos [**29**].

Luckily, for the purpose of this book, and indeed for much of modern analysis, we can get away with the naive notion of a set.

1.1.1. Proofs. For the purpose of this book, a *proof* is an *informal* description that explains why a statement is true by deriving it systematically from previously proved statements and axioms. The finer points of how such descriptions should be written are best learned from examples of which there are many in this book. This notion of proof as a text in natural language unfortunately means that the correctness and especially the completeness of a "proof" are social constructs and can be subject to bias and interpretation and generally involve much more ambiguity than one might ideally hope for. However, solace may be found in knowing that it is in principle possible to remove all such ambiguity by *formalizing* a proof (the interested reader should learn about Lean's *mathlib* project).

1.1.2. Propositional and predicate logic. In this book, as in most of modern mathematics, we use the principles of classical logic. A fundamental logical principle is *tertium non datur* (Latin for "no

third is given") meaning that a logical proposition must be either true or false. Given two propositions A, B one may use *logical connectives* to form new propositions:

- "A and B" is true exactly when A and B are both true
- "A or B" is true exactly when at least one of A, B is true
- "not A" is true exactly when A is not true (false).

Other logical connectives can be expressed in terms of these basic ones. In fact, "or" can be expressed in terms of "not" and "and": "A or B" is true exactly when "not (not A and not B)" is true. Two more logical connectives will be important: The proposition "A implies B" is true exactly when "(not A) or B" is true. In particular, if A is false, then "A implies B" is true independently of the truth value of B. This is also called *ex falso quodlibet* (Latin for "from falsehood everything (follows)"). Finally, the proposition "A is equivalent to B" (or "A if and only if B") means that A is true exactly when B is true, in other words, when "(A implies B) and (B implies A)". A *predicate* $P(x)$ is a proposition that assigns true or false to each object x from the universe of discourse. They can be combined with *quantifiers* to form new propositions: "for all x, $P(x)$" is true exactly when $P(x)$ is true for every x and "exists x s.t. $P(x)$" is true exactly when there exists an x such that $P(x)$ is true. Symbols like $\wedge, \vee, \forall, \exists, \iff$ *etc.* are often used in place of words to express logical propositions, but we prefer to use natural language.

1.1.3. Set operations. The *empty set* $\varnothing = \{\}$ is the only set without elements. The *union* of two sets A, B is denoted $A \cup B$ and consists of the elements that are contained in A or B (or both). The *intersection* of two sets A, B is denoted $A \cap B$ and consists of the elements contained in both A and B.

The *complement* of a set A is denoted $A^{\complement}$ and consists of the elements not contained in A – this is taken with respect to the universe of discourse, which is clear from context. The *difference* between two sets A, B is denoted $A \backslash B$ and consists of the elements of A that are not in B. Unions and intersections are commutative and associative and obey *de Morgan's law*:

$$(A \cup B)^{\complement} = A^{\complement} \cap B^{\complement}.$$

An *ordered pair* of two elements a, b is defined as the set $(a, b) = \{\{a\}, \{a, b\}\}$. This and further set-theoretic definitions serve only to remind the reader that really every object in our text can be interpreted as a set. However, for the purpose of doing real analysis, the set-theoretic interpretations are best forgotten once understood. The *(Cartesian) product* of two sets A, B is denoted $A \times B$ and defined by

$$A \times B = \{(a, b) \,:\, a \in A \text{ and } b \in B\}.$$

A set A is a *subset* of a set B if every element of A is also an element of B, written $A \subset B$ (the notation $A \subseteq B$ is also common). If we wish to exclude the case that $A = B$ we write $A \subsetneq B$. The *power set* of a set A is denoted $\mathcal{P}(A)$ and consists of all subsets of A. Note that the empty set is a subset of every set.

1.1.4. Natural numbers and induction. The *natural numbers* can be defined recursively as sets by setting

$$0 = \varnothing, \quad n + 1 = n \cup \{n\}.$$

For example, $3 = \{0, 1, 2\} = \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}$. Thus, 0 is a natural number, and every natural number n has a *successor* which is denoted $n + 1$ and is also a natural number. The *principle of induction* is both an important mechanism for defining a family of objects indexed by the natural numbers and a technique for proving a family of statements $P(n)$ depending on a natural number n: given that $P(0)$ holds and the statement '$P(n)$ implies $P(n + 1)$' holds, one may conclude that $P(n)$ holds for all natural numbers n. Whether or not zero is considered part of the natural numbers depends on context and preference. The definition of natural numbers as sets makes it particularly convenient

to define their ordering: $n \leqslant m$ holds if and only if n is a subset of m. In this book, the sets of natural numbers with and without 0 are denoted by

$$\mathbb{N}_0 = \{0, 1, 2, 3, 4, \dots\}, \qquad \mathbb{N} = \{1, 2, 3, 4, \dots\}.$$

The arithmetic operations $n+m$ and $n \cdot m$ for all natural numbers n, m can also be defined inductively.

1.1. Example (Addition of natural numbers). Let n, m be natural numbers. For each fixed n we define the natural number $n + m$ by induction on m:

- If $m = 0$, we define $n + m = n$.
- If m is not 0, then m is the successor of a natural number m' (in symbols, $m = m' + 1$). Given that $n + m'$ has been defined already, we define $n + m$ as the successor of $n + m'$, i.e.

$$n + m = (n + m') + 1.$$

One may then prove that the laws of commutativity $n+m = m+n$ and associativity $(n+m)+r = n + (m + r)$ hold for all natural numbers n, m, r. The proofs are, of course, by induction.

The reader is invited to similarly define multiplication and exponentiation of natural numbers and prove the familiar laws of arithmetic. Once we have defined multiplication, we may define factors and prime numbers: for natural numbers n, d we say d is a *factor* of n if there exists a natural number k so that $n = k \cdot d$ (we also say that d *divides* n and write $d|n$). A natural number is called *prime* if it is not equal to 1 and has no factors other than 1 and itself. Two natural numbers are called *coprime* if they share no factors other than 1.

Notation for sums and products. If $a_1, a_2, \dots, a_n$ are finitely many numbers (for now natural numbers, but later also rational, real and complex numbers), then we denote their sum by the symbol

$$\sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n$$

and their product by the symbol

$$\prod_{k=1}^{n} a_k = a_1 \cdot a_2 \cdots a_n.$$

(The formal definitions are again by the principle of induction.)

1.2. Example. Let us prove that for all $n \in \mathbb{N}$

$$P(n): \quad \sum_{k=1}^{n} (2k-1) = n^2$$

For $n = 1$, we must prove $P(1)$, which amounts to checking that $(2 \cdot 1 - 1) = 1$. It remains to show that $P(n)$ implies $P(n+1)$, i.e. the relation $\sum_{k=1}^{n+1}(2k-1) = (n+1)^2$. For this step we may use $P(n)$ and split the sum to get

$$\sum_{k=1}^{n+1} (2k-1) = \sum_{k=1}^{n} (2k-1) + (2(n+1)-1) = n^2 + (2(n+1)-1).$$

The right hand side equals $n^2 + 2n + 1 = (n+1)^2$ and so we have shown that '$P(n)$ implies $P(n+1)$ for all $n \geqslant 1$'. Thus since $P(1)$ holds and '$P(n)$ implies $P(n+1)$ holds for all $n \in \mathbb{N}$', we infer from the principle of induction that $P(n)$ holds for all $n \in \mathbb{N}$.

1.3. Exercise. Prove by induction the following statements for all $n \in \mathbb{N}$:

(i) $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$
(ii) $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$
(iii) $\sum_{k=1}^{n} k^3 = \left(\frac{n(n+1)}{2}\right)^2$

(iv) $n \leqslant 2^{n-1}$
(v) $n^2 < 2^n$ if $n \geqslant 5$
(vi) If $x \neq 1$, then $\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}$.

1.1.5. Relations and orders. A *relation* R on a set X is a subset of $X \times X$. If $(x, y) \in R$, we also write xRy and R may often be a symbol other than a letter (such as $=, <, \leqslant, \preceq, \equiv, \sim$ etc.). A relation is called

- *reflexive* if $(x, x) \in R$
- *symmetric* if $(x, y) \in R$ implies $(y, x) \in R$
- *anti-symmetric* if $(x, y) \in R$ and $x \neq y$ imply $(y, x) \notin R$
- *transitive* if $(x, y) \in R$ and $(y, z) \in R$ imply $(x, z) \in R$

(for all $x, y, z \in X$, respectively). A reflexive, symmetric and transitive relation is called an *equivalence relation.* This should be viewed as a generalization of equality. Indeed, equality (denoted by '=') is an example of an equivalence relation, albeit not a very interesting one. Given an equivalence relation $\sim$ on a set X and an element $x \in X$, one can consider the *equivalence class* of x,

$$[x] = \{y \in X \,:\, x \sim y\}.$$

By reflexivity, every element is contained in its own equivalence class. One also checks that two equivalence classes are either equal or disjoint. Thus, the equivalence classes partition a set X into disjoint subsets. The set of equivalence classes is denoted $X/\sim$.

1.4. Example. Let d be a non-zero natural number. We say that natural numbers x and y are *equivalent modulo d,* denoted

$$x \equiv y \pmod{d}$$

if x and y have the same remainder upon division by d. This partitions the natural numbers into d equivalence classes $[0], [1], \ldots, [d-1]$, each corresponding to one of the d possible remainders.

Another important kind of relation is an order, generalizing the notion of $\leqslant$ for natural numbers. There are partial and total orders. A reflexive, anti-symmetric and transitive relation is called a *partial order.* It is usually denoted by a symbol such as $\preceq$. Note, from the definition of antisymmetry, that if $a \preceq b$ and $b \preceq a$ both hold, we must have $a = b$.

A partial order on a set X is a *total order* if $x \preceq y$ or $y \preceq x$ holds for every $x, y \in X$. A set with a total order is called *totally ordered.*

1.5. Examples.

(i) The '$\leqslant$' relation on natural numbers is a total order.
(ii) The 'divides' relation (written '|') on natural numbers is a partial order. It is not a total order, since not every two natural numbers divide each other.
(iii) The subset relation '$\subset$' is a partial order that is not a total order, since not every two sets are contained in each other.

1.1.6. Functions. Let A and B be nonempty sets. A *function* (or synonymously, *map,* or *mapping*) denoted

$$f : A \to B$$

is a relation $f \subset A \times B$ such that for every $a \in A$ there exists exactly one $b \in B$ so that $(a, b) \in f$. In that case we say that f *maps a to b* and write $f(a) = b$, i.e. b can be denoted by the symbol $f(a)$. One also writes

$$f : a \longmapsto b$$

(note the difference in the arrow used in $f : A \to B$). The set of all maps $A \to B$ is also denoted by B^A (to motivate this notation consider finite sets: if A has n elements and B has m elements, then

there are m^n possible maps $A \to B$). If $f(a) = b$, we call b the *image of a under f* and a the *preimage of b under f*. The set of preimages of a value $b \in B$ is denoted

$$f^{-1}(b) = \{a \in A \,:\, f(a) = b\}.$$

Similarly, if $B' \subset B$ one writes

$$f^{-1}(B') = \{a \in A \,:\, f(a) \in B'\}.$$

Note that $f^{-1}(b)$ may be empty or contain more than one element. If $f^{-1}(b)$ consists of at most one element for every $b \in B$ one calls f *injective* or *one-to-one*. In this case one also writes $f^{-1}(b) = a$ instead of $f^{-1}(b) = \{a\}$ if $f(a) = b$. The set A is called the *domain* of f and B is called the *codomain* of f. Note that B may be different from the set

$$f(A) = \{f(a) \,:\, a \in A\} \subset B$$

which is called the *image* of A under f (or the *range* of f). If $f(A) = B$ one also says that f is *surjective* or *onto*. A map that is injective and surjective is called *bijective*. If f is bijective, then it has a well-defined *inverse* denoted f^{-1} which is the function $f^{-1} : B \to A$ so that $f^{-1}(b) = a$ if and only if $f(a) = b$ for all $a \in A$ and $b \in B$. One then also verifies that f^{-1} is bijective and that $(f^{-1})^{-1} = f$. Every one-to-one map $f : A \to B$ can be made into a bijective map when reconsidered as mapping to its image: $f : A \to f(A)$. The *composition* of two functions $f : A \to B$, $g : B \to C$ is the function $g \circ f : A \to C$ given by

$$(g \circ f)(a) = g(f(a)).$$

Composition of functions is associative, but not commutative. The *identity map* on A is the function $\mathrm{id}_A : A \to A$ with $\mathrm{id}_A(a) = a$ for all $a \in A$. We say that $g : B \to A$ is a *left inverse* of $f : A \to B$ (and f a *right inverse* of g) if

$$g \circ f = \mathrm{id}_A.$$

If g is a left inverse, then g is automatically surjective (since for every $a \in A$ the element $f(a) \in B$ is mapped to a by g). Similarly, if f is a right inverse, f must be injective (since $f(a) = f(b)$ implies $a = g(f(a)) = g(f(b)) = b$).

1.6. Exercise. Show that f is injective if and only if it has a left inverse.

It is also true that f is surjective if and only if it has a right inverse, but this is more subtle (see Example 1.7 below).

Indexed families. Let Ω be a set (whose elements may also be sets). An *indexed family* with *index set* I is a map

$$I \to \Omega, \quad i \longmapsto x_i.$$

It is usually denoted $(x_i)_{i \in I}$. In the special case that I is the set of natural numbers one calls this a *sequence* (also see §1.3 below). If I is the set of the natural numbers from 1 to n, then the indexed family is called an *(ordered) n-tuple* and denoted $(x_1, \ldots, x_n)$.

When $\mathcal{A} = (A_i)_{i \in I}$ is a family of sets we define its *union* as the set of all elements that are contained in at least one of the sets A_i. We use the notation

$$\bigcup \mathcal{A} = \bigcup_{i \in I} A_i.$$

Similarly the *intersection* of $\mathcal{A}$ is the set of elements contained in each of the sets A_i, denoted

$$\bigcap \mathcal{A} = \bigcap_{i \in I} A_i.$$

The *Cartesian product* of a family $\mathcal{A}$ is the set of all indexed families $(a_i)_{i \in I}$ so that $a_i \in A_i$ for all $i \in I$ and is denoted

$$\prod \mathcal{A} = \prod_{i \in I} A_i = \{x : I \to \bigcup_{i \in I} A_i \,:\, x(i) \in A_i \quad \text{for all } i \in I\}.$$

Will will also write $x : I \to \bigcup_{i\in I} A_i$ with $x(i) \in A_i$ as a tuple $x = (x_i)_{i\in I} \in \prod \mathcal{A}$. These definitions will often be used when $\mathcal{A}$ is an n-tuple or a sequence, but occasionally also when the index set I is uncountable.

1.1.7. Axiom of choice. Given a nonempty family of nonempty sets $(A_i)_{i\in I}$, can one choose a family of elements $(a_i)_{i\in I}$ such that $a_i \in A_i$ for every $i \in I$?

If I consists of the natural numbers 1 through n, it is clear how to do this: begin with A_1 and choose an arbitrary element (which exists since A_1 is nonempty), continue with A_2 and so on until done. Since I is finite, this constitutes an algorithm that terminates after finitely many steps and produces the desired choice of $(a_i)_{i\in I}$.

Further, if I is the set of all natural numbers (or any countably infinite set), one may proceed in the same manner – with the unfortunate caveat that the procedure will now never terminate, so this argument is inadequate to demonstrate existence of the full family of choices $(a_i)_{i\in I}$. Making matters worse, what if I is uncountable? There is no sensible way to adapt the procedure.

The *axiom of choice* postulates the existence of such a choice for every nonempty index set I, including very large ones. Note from the definition of Cartesian products that a succinct way of stating the axiom of choice is that a nonempty Cartesian product of nonempty sets shall be nonempty.

If I is a nonempty index set and $\{X_\alpha\}_{\alpha\in I}$ is a collection of nonempty sets, then an associated *choice function* is a function $f : I \to \bigcup_\alpha X_\alpha$ such that $f(\alpha) \in X_\alpha$ for all $\alpha \in I$. The axiom of choice postulates the existence of such a choice function.

There is a weaker version of the axiom of choice restricted to the case of countable index sets I (such as the natural numbers) known as the axiom of *countable choice.*

A remarkable feature of the axiom of choice is that while it appears at first glance to be compatible with intuition, it has certain very counter-intuitive consequences especially when applied in situations where the index set I is not countable. An example is the existence of certain "absurd" subsets of the real numbers called non-measurable sets (see e.g. Stein–Shakarchi [**76**]). On the other hand, the axiom of choice often appears in natural ways and has important consequences throughout mathematics that cannot be proved without it. Let us illustrate this.

1.7. Example. Let $f : A \to B$ be a map between sets. Recall that f has a left inverse if and only if f is injective (Exercise 1.6). One also sees directly from the definitions that if f has a right inverse, then f is surjective. What about the converse? Suppose $f : A \to B$ is surjective and B is nonempty. We need to construct $g : B \to A$ so that

$$f(g(b)) = b \quad \text{for all } b \in B.$$

Since f is surjective, the preimage $f^{-1}(b)$ is nonempty for each $b \in B$. Thus by the axiom of choice, the product

$$\prod_{b\in B} f^{-1}(b)$$

is also nonempty. An element of this product can be viewed as a function $g : B \to A$ such that for every $b \in B$, $g(b) \in f^{-1}(b)$, in other words, $f(g(b)) = b$.

There are many statements that are equivalent to the axiom of choice (i.e. either one can be proved from the other). We will present one of them. We need two important definitions regarding partial orders.

Let A be a set with a partial order $\preceq$. A subset $C \subset A$ is called a *chain* if it is totally ordered by $\preceq$. Let $\tilde{A} \subset A$. An element $M \in A$ such that $x \preceq M$ for all $x \in \tilde{A}$ is called an *upper bound* for $\tilde{A}$. An element $L \in A$ such that $L \preceq x$ for all $x \in \tilde{A}$ is called a *lower bound* for $\tilde{A}$.

An element $a \in A$ is called *maximal* if for every $a' \in A$ with $a \preceq a'$ we have $a' = a$. This generally does not imply that a is an upper bound of A. In particular, a partially ordered set may have multiple

maximal elements (or none). However, if $\preceq$ is a total order, then maximal elements are also upper bounds and there can be at most one maximal element.

The axiom of choice is often used in the form of the following theorem, which is referred to as *Zorn's lemma*. The reader may skip its proof for now. For some applications of Zorn's lemma in this text, see Example 1.9, Exercise 1.10, Theorem 5.86, Theorem 7.44, and Example 8.35.

1.8. THEOREM (Zorn's lemma). *Suppose a partially ordered set A has the property that every chain has an upper bound in A. Then A has a maximal element.*

PROOF. The proof is by contradiction. Assume that A does not have a maximal element. This means that every chain C has a *strict upper bound*, i.e. there is $u \in A$ such that $x \precneqq u$ for all $x \in C$. By the axiom of choice there is a choice function f defined on the set of chains which assigns to every chain in A a strict upper bound $f(C) \in A$.

For a nonempty chain C, and $x \in C$ we define the chain

$$\mathcal{I}(C, x) = \{y \in C : y \precneqq x\},$$

which is labeled the *C-initial segment below x*. We say that a subset $C \subset A$ is *conforming to f* if the following two conditions are satisfied:

(a) C is a chain in A such that every nonempty subset of C has a minimal element.
(b) For every $x \in C$ we have $x = f(\mathcal{I}(C, x))$, i.e. the choice function picks x when applied to the C-initial segment below x.

As a first example for the existence of nonempty chains conforming to f we observe that $\varnothing$ is a chain, and let $x = f(\varnothing)$. Then $\mathcal{I}(\{x\}, x) = \varnothing$ and the one-element chain $C_1 = \{x\}$ is conforming to f. We may inductively construct conforming sets of higher cardinality. Namely if $C_n = \{x_1, \dots, x_n\}$ with $x_{j-1} \precneqq x_j$ for $j = 2, \dots, n$, we define $x_{n+1} = f(C_n)$, a strict upper bound for C_n, and then $C_{n+1} = \{x_1, \dots, x_{n+1}\}$ is a chain conforming to f.

Our proof relies on the following crucial observation concerning chains conforming to f.

OBSERVATION. Let C_1, C_2 be chains conforming to f. Then one of them is an initial segment of the other.

We take this for granted for a moment to arrive at the desired contradiction. We note that if C is conforming to f and $a \in C$ then for $b \precneqq a$ we either have $b \in C$ or b does not belong to any chain conforming to f. We let U be the union of all chains in A that are conforming to f and observe that U is itself conforming to f. Let $x = f(U)$, and note that $U \cup \{x\}$ is conforming to f. But by definition of U this implies $x \in U$ in contradiction to $x = f(U)$ being a *strict* upper bound to U. Thus the assumption that there is no maximal element leads to a contradiction and we obtain Zorn's lemma.

We still have to give the proof of the observation: we may assume that $C_1 \backslash C_2 \neq \varnothing$, and our objective is to show that there exists $x \in C_1$ such that $C_2 = \mathcal{I}(C_1, x)$. Indeed we will show this for x being the minimal element of $C_1 \backslash C_2$ (using defining property (a) in the definition of conforming chains).

Note $\mathcal{I}(C_1, x) \subset C_2$, in view of the minimality of x in $C_1 \backslash C_2$. We claim that $C_2 = \mathcal{I}(C_1, x)$ and prove this by contradiction. Assume $C_2 \backslash \mathcal{I}(C_1, x) \neq \varnothing$ and let y be the minimal element in $C_2 \backslash \mathcal{I}(C_1, x)$. Let z be the minimal element in $C_1 \backslash \mathcal{I}(C_2, y)$ (which is nonempty since it contains $C_1 \backslash C_2$). Since x is the minimal element in $C_1 \backslash C_2$ we have $z \preceq x$.

We claim that $\mathcal{I}(C_1, z) = \mathcal{I}(C_2, y)$. If $v \in \mathcal{I}(C_1, z)$ then $v \in C_1$, and $v \precneqq z$, and by the minimality of z in $C_1 \backslash \mathcal{I}(C_2, y)$, v must be in $\mathcal{I}(C_2, y)$. Therefore $\mathcal{I}(C_1, z) \subset \mathcal{I}(C_2, y)$.

To see the opposite inclusion, let $u \in \mathcal{I}(C_2, y)$. Then $u \in C_2$, $u \precneqq y$ and by the minimality of y in $C_2 \backslash \mathcal{I}(C_1, x)$, u must be in $\mathcal{I}(C_1, x)$, i.e. $u \in C_1$ and $u \precneqq x$. Now consider z, which belongs to $C_1 \backslash \mathcal{I}(C_2, y)$, so $z \neq u$. We do not have $z \precneqq u$ and $z \in C_1 \backslash C_2$ since this would contradict the minimality of x in $C_1 \backslash C_2$. We also do not have $z \precneqq u$ and $z \in C_2$, again since $z \notin \mathcal{I}(C_2, y)$. Hence $u \precneqq z$ and thus $u \in \mathcal{I}(C_1, z)$. Hence $\mathcal{I}(C_2, y) \subset \mathcal{I}(C_1, z)$.

We now have established the equality $\mathcal{I}(C_1, z) = \mathcal{I}(C_2, y)$ and obtain

$$z = f(\mathcal{I}(C_1, z)) = f(\mathcal{I}(C_2, y)) = y.$$

Since $z = y \in C_2$ and $x \in C_1 \backslash C_2$ we cannot have $z = x$ and therefore we conclude $z \precneqq x$, and thus $y = z \in \mathcal{I}(C_1, x)$. This contradicts the definition of y as an element of $C_2 \backslash \mathcal{I}(C_1, x)$. Thus the observation is proved. □

REMARK. Here we followed an argument by Lewin [**47**].

1.9. EXAMPLE. Zorn's lemma can also be used to prove the axiom of choice. In this sense, the two are equivalent. We give the proof which is quite typical for other applications of Zorn's lemma later in the text.

Let I be a nonempty index set. For each $\alpha \in I$ let X_α be a nonempty set. For subsets $J \subset I$ we consider choice functions $f : J \to \bigcup_{\alpha \in J} X_\alpha$ such that $f(\alpha) \in X_\alpha$ for $\alpha \in J$. We consider the set $\mathcal{P}$ of pairs (J, f) where J is a nonempty subset of I and f is an associated choice function. On $\mathcal{P}$ we define a partial order $\preceq$ by setting $(J_1, f_1) \preceq (J_2, f_2)$ if $J_1 \subset J_2$ and $f_2(\alpha) = f_1(\alpha)$ for $\alpha \in J_1$ (i.e. f_2 is an extension of f_1).

Notice that for fixed $\alpha \in I$ there exists $x_\alpha \in X_\alpha$ and setting $f(\alpha) = x_\alpha$ produces a pair $(\{\alpha\}, f)$ in $\mathcal{P}$. Each such choice is an upper bound for an empty chain. We now consider a nonempty chain of pairs $(J_\nu, f_\nu)_{\nu \in C}$. Let $J = \bigcup_\nu J_\nu$ and define $f(\alpha) = f_\nu(\alpha)$ if $\alpha \in J_\nu$. Note that $f : J \to \bigcup_{\alpha \in J} X_\alpha$ is well defined, since $f_\nu|_{J_{\nu'}} = f_{\nu'}$ if $J_{\nu'} \subset J_\nu$. The pair $(\bigcup_\nu J_\nu, f)$ is then an upper bound for our chain.

The validity of Zorn's lemma implies that the set $\mathcal{P}$ has a maximal element (J, f). We claim that $J = I$. If $I \backslash J \neq \varnothing$ there exists $\beta \in I \backslash J$. Also there exists $x_\beta \in X_\beta$ since X_β is not empty. Define $g(\alpha) = f(\alpha)$ for $\alpha \in J$ (then $g(\alpha) \in X_\alpha$) and set $g(\beta) = x_\beta$. Then $(J, f) \precneqq (J \cup \{\beta\}, g)$ in contradiction to the maximality of (J, f). Hence $J = I$ and f is the required choice function on the full index set I.

1.10. EXERCISE. Given a nonempty set X, *a well-order* on X is a total order on X with the property that every nonempty subset has a minimal element. Prove that every set has a well-order.

1.1.8. Cardinality. The cardinality of a finite set is its number of elements. But what about infinite sets? Does it make sense to say that one infinite set has "more" elements than another? As we shall see, there is a well-defined notion of cardinality for arbitrary sets.

1.11. DEFINITION. Let A, B be sets. We say that $\operatorname{card}(A) = \operatorname{card}(B)$ if there exists a bijection $f : A \to B$.

1.12. EXERCISE. Show that the relation $A \sim B$ defined as $\operatorname{card}(A) = \operatorname{card}(B)$ is an equivalence relation.

1.13. DEFINITION. Let A, B be sets. We say that $\operatorname{card}(A) \leqslant \operatorname{card}(B)$ if there is an injective map $f : A \to B$. We say that $\operatorname{card}(A) \geqslant \operatorname{card}(B)$ if there is a surjective map $g : A \to B$. If A is nonempty we say that $\operatorname{card}(A) < \operatorname{card}(B)$ if there is an injective map from A to B but not a bijective map from A to B. We also define $\operatorname{card}(\varnothing) < \operatorname{card}(A)$ and $\operatorname{card}(A) > \operatorname{card}(\varnothing)$ for all nonempty A.

1.14. PROPOSITION. $\operatorname{card}(A) \leqslant \operatorname{card}(B)$ *if and only if* $\operatorname{card}(B) \geqslant \operatorname{card}(A)$.

PROOF. If $\operatorname{card}(A) \leqslant \operatorname{card}(B)$ and $f : A \to B$ is the corresponding injective function, then we define $g(f(x)) = x$ and $g(y)$ arbitrarily if $y \notin \operatorname{range}(f)$. This is well-defined because f is injective. Then $g : B \to A$ is surjective.

Conversely, assume $\operatorname{card}(B) \geqslant \operatorname{card}(A)$ and let $g : B \to A$ be surjective. For every $x \in A$ consider $g^{-1}(\{x\}) = \{y : g(y) = x\}$. We use the axiom of choice to pick $f(x) \in g^{-1}(\{x\})$ for every $x \in A$ and the choice function $f : A \to B$ is injective. Hence $\operatorname{card}(A) \leqslant \operatorname{card}(B)$. □

The following result (referred to as the *Cantor–Schröder–Bernstein theorem*) can be very helpful in establishing that two sets have the same cardinality.

1.15. THEOREM. *Let A, B be sets and assume that there exist injective maps $f : A \to B$ and $g : B \to A$. Then there is a bijective map $h : A \to B$.*

PROOF. The function g is bijective as a map from B to its image $g(B) \subset A$. Let

$$g^{-1} : g(B) \to B$$

denote its inverse. A basic idea is to divide A into two disjoint parts A_0, A_1, such that A_0 is a suitable subset of $g(B)$ where we use g^{-1} for our definition of h, and such that on A_1 we use f for this definition. It is not at all obvious what this decomposition should be. Since we will never reach points in $B\backslash f(A)$ using f we observe that we should work with the choice $h(a) = g^{-1}(a)$ when $a \in g(B\backslash f(A))$, that is when $f(a)$ belongs to the set $f \circ g(B\backslash f(A))$.

One carries this further and considers n-fold applications of the map $f \circ g : B \to B$ which we denote by $(f \circ g)^n$. For $n = 1, 2, \ldots$, define sets $C_n \subset B$ by

$$C_n = (f \circ g)^n\big(B\backslash f(A)\big).$$

We have $C_1 = f(g(B\backslash f(A)))$ and $C_n = f(g(C_{n-1}))$ for $n \geqslant 2$. Let

$$C = \bigcup_{n=1}^{\infty} C_n$$

and set

$$A_0 = \{a \in A : f(a) \in C\},$$
$$A_1 = \{a \in A : f(a) \notin C\}.$$

This gives a partition of A, as $A = A_0 \cup A_1$ and $A_0 \cap A_1 = \varnothing$. Next set

$$B_0 = (B\backslash f(A)) \cup C,$$
$$B_1 = f(A) \cap (B\backslash C)$$

Note that $B = B_0 \cup B_1$ and $B_0 \cap B_1 = \varnothing$, so we have a partition of B.

We claim:

(i) $f|_{A_1}$ is a bijection from A_1 to B_1.
(ii) $A_0 \subset g(B_0)$ and $g^{-1}|_{A_0}$ is a bijection from A_0 to B_0.

Taking the claim for granted, it follows that

$$h : A \to B, \qquad a \mapsto h(a) = \begin{cases} g^{-1}(a) & \text{if } a \in A_0 \\ f(a) & \text{if } a \in A_1 \end{cases}$$

is a bijection from A to B.

Proof of part (i) of the claim. By definition of A_1 we see that $f|_{A_1}$ maps A_1 into B_1. $f|_{A_1}$ is injective since f is injective. It remains to check that $f|_{A_1}$ maps A_1 onto B_1. Let $b \in B_1$. That is $b = f(a)$ for some $a \in A$ and $f(a) \notin C$. Thus by definition of A_1 we have $a \in A_1$ and so $f|_{A_1} : A_1 \to B_1$ is surjective. We have established part (i) of the claim.

Proof of part (ii) of the claim. Consider $a \in A_0$ then $f(a) \in C_k$ for some $k \geqslant 1$, i.e. $f(a) = (f \circ g)^k b$ for some $b \in B\backslash f(A)$. Since f is injective this implies

$$a = \begin{cases} g(b) & \text{if } k = 1, \\ g((f \circ g)^{k-1}(b)) & \text{if } k \geqslant 2. \end{cases}$$

Since $b \in B\backslash f(A) \subset B_0$ and $(f \circ g)^{k-1}(b) \in C_{k-1} \subset B_0$ if $k \geqslant 2$ we see that $a \in g(B_0)$. We have shown that $A_0 \subset g(B_0)$ and thus $g^{-1}|_{A_0}$ maps A_0 into B_0.

Since $g^{-1} : g(B) \to B$ is a bijection and $A_0 \subset g(B)$ it is clear that $g^{-1}|_{A_0} : A_0 \to B_0$ is injective.

To see that $g^{-1}|_{A_0}$ maps A_0 onto B_0, we let $b \in B_0$, i.e. b belongs to the union of $B\backslash f(A)$ and $C = \bigcup_{n=1}^{\infty} C_n$. This means that $f(g(b))$ belongs either to $f \circ g(B\backslash f(A)) = C_1$ or to $f \circ g(C_k) = C_{k+1}$ for some $k \geqslant 1$. Hence $f(g(b)) \in C$. Therefore, by definition of A_0, we have $g(b) \in A_0$ and $b = g^{-1}(g(b))$. This shows that $g^{-1}|_{A_0} : A_0 \to B_0$ is surjective, and we have established part (ii) of the claim. □

As a consequence of Theorem 1.15 we obtain:

1.16. Corollary. *Let A and B be two sets. Then*

$$\operatorname{card}(A) = \operatorname{card}(B) \text{ if and only if } \operatorname{card}(A) \leqslant \operatorname{card}(B) \text{ and } \operatorname{card}(B) \leqslant \operatorname{card}(A)\,.$$

1.17. Theorem (Cantor's power set theorem). *Let A be a set and $\mathcal{P}(A)$ the power set. Then*

$$\operatorname{card}(A) < \operatorname{card}(\mathcal{P}(A)).$$

Proof. The map $\imath : A \to \mathcal{P}(A)$ which sends $a \in A$ to $\imath(a) = \{a\} \in \mathcal{P}(A)$ is an injective map. It remains to show that no map $f : A \to \mathcal{P}(A)$ is surjective. Let $Y = \{a \in A : a \notin f(a)\}$. We show that $Y \subset A$ is not in the range of f. Assume that $Y = f(x)$ for some $x \in A$. If $x \in Y$, then $x \notin f(x) = Y$ which gives a contradiction. If $x \notin Y$, then $x \in f(x) = Y$, again a contradiction. Hence Y is not in the range of f. □

1.18. Definition. A set A is called *countable* if $\operatorname{card}(A) \leqslant \operatorname{card}(\mathbb{N})$, i.e. if there exists an injective map from A into the natural numbers.

A set that is not countable is called *uncountable.*

All finite sets are countable. For a finite set, we may define $\operatorname{card}(A)$ as the number of elements in A. We thus write $\operatorname{card}(A) = n$ to mean that $\operatorname{card}(A) = \operatorname{card}\{1, \ldots, n\}$. For finite sets one also often writes $\#A$ for $\operatorname{card}(A)$.

1.19. Proposition. *Let A be countable and infinite; then* $\operatorname{card}(A) = \operatorname{card}(\mathbb{N})$.

Proof. By definition there is a bijection between A and a subset $B = g(A)$ of $\mathbb{N}$, where B is infinite. We construct a bijection $f : \mathbb{N} \to B$. Let $f(1)$ be the smallest element in B. With $f(1), \ldots, f(n-1)$ defined, let $f(n)$ be the smallest element in $B\backslash\{f(1), \ldots, f(n-1)\}$. $f : \mathbb{N} \to B$ is a bijection, and thus $\operatorname{card}(\mathbb{N}) = \operatorname{card}(B) = \operatorname{card}(A)$. □

1.20. Proposition. *The set $\mathbb{N} \times \mathbb{N}$ is countable.*

Proof. We construct a bijection $f : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$. We first describe this map informally. We arrange the pairs (i, j) in $\mathbb{N} \times \mathbb{N}$ into an array (or infinite matrix) with rows labeled by i and columns labeled by j. For $k \in \mathbb{N}$ the kth diagonal consists of the k pairs (i, j) with $i + j = k + 1$. Beginning with the first diagonal $(1, 1)$, we enumerate for each $k \in \mathbb{N}$ the pairs on the kth diagonal, then pass to the $k+1$st diagonal and so forth. Thus $f(1) = (1, 1)$, then $f(2) = (2, 1)$, $f(3) = (1, 2)$, now passing to the third diagonal $f(4) = (3, 1)$, $f(5) = (2, 2)$, $f(6) = (1, 3)$, etc.

More formally, we use that $\sum_{m=1}^{k-1} m = \frac{(k-1)k}{2}$ for $k \geqslant 2$ to observe that every $n \in \mathbb{N}$ can be uniquely written as $n = \frac{(k-1)k}{2} + \nu$ with $k \in \mathbb{N}$ and $\nu \in \{1, \ldots, k\}$. We then define the bijection f by

$$f(\tfrac{(k-1)k}{2} + \nu) = (k + 1 - \nu, \nu), \quad \nu = 1, \ldots, k,\ k \in \mathbb{N}.$$

□

1.21. Proposition. *For any two sets A and B, at least one of the two statements*

$$\operatorname{card}(A) \leqslant \operatorname{card}(B) \quad \text{or} \quad \operatorname{card}(B) \leqslant \operatorname{card}(A)$$

holds.

PROOF. Let $\mathcal{J}$ be the collection of all bijective maps from a subset X of A to a subset Y of B. Each $\imath \in \mathcal{J}$ corresponds to a subset of $A \times B$. Thus the inclusion relation on $A \times B$ provides a partial order on $\mathcal{J}$.

Let $\mathcal{C}$ be a chain consisting of bijective maps $\imath_\alpha : X_\alpha \to Y_\alpha$. Let $X = \cup_\alpha X_\alpha$, $Y = \cup_\alpha Y_\alpha$ and define $\imath : X \to Y$ by $\imath(x) = \imath_\alpha(x) \in Y_\alpha$ if $x \in X_\alpha$. Since $\mathcal{C}$ is totally ordered we can check that $\imath$ is well defined as a bijective map, and it is an upper bound for $\mathcal{C}$.

By Zorn's lemma there is a maximal element in $\mathcal{J}$; it is a bijective map $\jmath : X_{\max} \to Y_{\max}$ where $X_{\max} \subset A$ and $Y_{\max} \subset B$. If both are strict inclusions, we get a contradiction, because if $a_\circ \in A \backslash X_{\max}$ and $b_\circ \in B \backslash Y_{\max}$ then $\jmath$ can be extended to a bijective map $f : X_{\max} \cup \{a_\circ\} \to Y_{\max} \cup \{b_\circ\}$ by setting $f(a_\circ) = b_\circ$, contradicting the maximality of $\jmath$.

We conclude that either $X_{\max} = A$ or $Y_{\max} = B$ (or both). If $X_{\max} = A$, then the existence of the bijective map $\jmath : A \to Y_{\max} \subset B$ gives $\operatorname{card}(A) \leqslant \operatorname{card}(B)$. If $Y_{\max} = B$, then we use the inverse map $\jmath^{-1} : B \to X_{\max}$ to see that $\operatorname{card}(B) \leqslant \operatorname{card}(A)$. □

As a consequence of Corollary 1.16 and Proposition 1.21 we get

1.22. COROLLARY. *Consider the equivalence relation $A \sim B$ of A and B having the same cardinality. Denote the equivalence class with respect to $\sim$ of a set A by $[A]$. Then the equivalence classes for $\sim$ form a totally ordered set (with respect to the order $[A] \preceq [B]$ given by $\operatorname{card}(A) \leqslant \operatorname{card}(B)$ for all representatives $A \in [A]$, $B \in [B]$).*

1.23. EXERCISE (Countable union of countable sets). Prove: If the index set $\mathcal{I}$ is countable and if for every $i \in \mathcal{I}$ the set X_i is countable, then $\cup_{i \in \mathcal{I}} X_i$ is countable. *Careful:* This requires the axiom of (countable) choice.

1.24. EXERCISE (Binary sequences). Show that the set of functions $a : \mathbb{N} \to \{0, 1\}$ (infinite binary sequences) is uncountable, but the set of finite (but arbitrarily long) binary sequences

$$\{a : \{1, \dots, N\} \to \{0, 1\} \,:\, N \in \mathbb{N}\}$$

is countable.

1.2. Numbers

1.2.1. Rational numbers. The set of integers is denoted $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$. The set $\mathbb{Q}$ of *rational numbers* consists of fractions p/q (or $\frac{p}{q}$) with $p \in \mathbb{Z}$, $q \in \mathbb{N}$. A fraction $\frac{p}{q}$ is *reduced* if p and q are coprime, that is, they do not have a common prime factor. The usual addition, multiplication and ordering can be formally defined first on integers, and then on the rational numbers from the corresponding operations on integers.

The rational numbers with the usual addition, multiplication and ordering then form an *ordered field.* This means that the familiar laws of arithmetic with respect to addition and multiplication (commutativity, associativity, distributivity, additive and multiplicative inverses, arithmetic with the neutral elements zero and one) apply and are compatible with the ordering of numbers (that is, $a < b$ implies $a + c < b + c$ and if $a > 0$ and $b > 0$, then $a \cdot b > 0$). For the precise axiomatic definition of ordered fields see Definitions 1.102, 1.107 in the appendix.

The set of rational numbers has some deficiencies. For example, we may not be able to take square roots of some positive rational numbers.

1.25. THEOREM. *There is no rational number x such that $x^2 = 2$.*

PROOF. We argue by contradiction. Suppose $x \in \mathbb{Q}$ with $x^2 = 2$, and assume that x is written in lowest terms as $x = \frac{m}{n}$ (thus m and n do not have a common prime factor). Then $m^2 = 2n^2$, so m^2 is even and thus m must be even. Thus $m = 2k$ for some integer k, hence $4k^2 = 2n^2$, thus n^2 is even and therefore n must be even, too. We get a contradiction since it was assumed that m and n do not have a common prime factor. □

1.2.2. The field of real numbers. Theorem 1.25 suggests that the field of rational numbers has "gaps". We would like to fill in these gaps; for example one would like to add a number $\sqrt{2}$ satisfying $x^2 = 2$. We check below that the set $\{x \in \mathbb{Q} : x^2 \leqslant 2\}$ does not have a *least upper bound* in $\mathbb{Q}$ and the set $\{x \in \mathbb{Q} : x^2 \geqslant 2\}$ does not have a greatest lower bound in $\mathbb{Q}$.

Recall that M is an upper bound for $A \subset \mathbb{K}$ if $x \leqslant M$ for all $x \in A$ (this is a special case of the corresponding definition in general partially ordered sets). L is a lower bound for A if $L \leqslant x$ for all $x \in A$.

1.26. Definition. Let $\mathbb{K}$ be an ordered field, and $A \subset \mathbb{K}$. We say that an upper bound M for A is a *least upper bound* if for every upper bound M' we have $M \leqslant M'$. The least upper bound is also called the *supremum* of A and is denoted by $\sup A$.

Similarly we say that a lower bound L is the *greatest lower bound* if for every lower bound L' we have $L' \leqslant L$. The greatest lower bound for A is also called *infimum* and denoted by $\inf A$. A set is called *bounded above* if it has an upper bound. A set is called *bounded below* if it has a lower bound.

1.27. Observation. *Let $\mathbb{K}$ be an ordered field. $A \subset \mathbb{K}$ is bounded below if and only if $-A = \{-x : x \in A\}$ is bounded above. Moreover,*

$$\inf A = -\sup(-A).$$

1.28. Definition. An ordered field is *Dedekind complete* if it satisfies the *least upper bound axiom*:

Every nonempty set $A \subset \mathbb{K}$ bounded from above has a least upper bound.

1.29. Observation. *The ordered field $\mathbb{Q}$ is not Dedekind complete.*

Proof. Consider the set $A = \{x \in \mathbb{Q}, x \geqslant 0, x^2 < 2\}$. A is bounded above, because 2 is an upper bound. We argue by contradiction and assume that A has a least upper bound R. By Theorem 1.25 we must have $R^2 > 2$ or $R^2 < 2$. In both cases we derive a contradiction.

If $R^2 > 2$, set $r = R - h$ with $h = \frac{R^2-2}{2R}$ and we get

$$r^2 = \big(R - h\big)^2 = 2 + \frac{(R^2-2)^2}{4R^2} > 2.$$

For $x \in A$ we thus have $r^2 > x^2$, hence $r > x$, so r is an upper bound for A. But $r < R$, and we get a contradiction since R was the least upper bound for A.

If $R^2 < 2$, consider $r = R + h$ with $h = \frac{2-R^2}{2R+1}$, so that $h > 0$ and since $R \geqslant 1$ we also have $h \leqslant \frac{1}{2R+1} < 1$, hence $h^2 < h$. Then

$$r^2 = R^2 + 2Rh + h^2 < R^2 + 2Rh + h = 2.$$

Thus $r \in A$ but $r > R$, which is a contradiction since R was an upper bound for A. □

An *isomorphism between ordered fields* is a bijective map which respects the operations of addition and multiplication and preserves the order (see Definition 1.109). One usually identifies ordered fields that are isomorphic. The following theorem is a foundational result on which all of real analysis is built.

1.30. Theorem. *There exists a Dedekind complete ordered field $\mathbb{R}$ which contains the rational numbers as a subfield. Every Dedekind complete ordered field is isomorphic to $\mathbb{R}$.*

The proof of the theorem is via an explicit construction. It is not actually relevant for doing analysis, and in practice we need not care about how the real numbers are constructed. A construction and the uniqueness proof will be given in §1.6.2 and §1.6.3.

We state some consequences of the least upper bound axiom. The first theorem expresses the *Archimedean property of the real numbers*. In particular there is no real number which is larger than all the others.

1.31. Theorem. *If $a, b \in \mathbb{R}$, $a > 0$, then there exists an $n \in \mathbb{N}$ such that $na > b$.*

PROOF. We argue by contradiction and assume such an n does not exist, i.e. $na \leqslant b$ for all $n \in \mathbb{N}$. Let $E = \{na : n \in \mathbb{N}\}$ which has b as an upper bound. Thus E has a least upper bound $C = \sup E \in \mathbb{R}$. Then $C - a$ is not an upper bound for E, i.e. there exists $m \in \mathbb{N}$ such that $ma > C - a$ which implies $(m+1)a > C$ which gives a contradiction since $(m+1)a \in E$ and C was an upper bound for E. □

REMARK. The field $\mathbb{Q}$ of rational numbers also has the Archimedean property: we may assume $b = \frac{p}{q} > 0$ with positive integers p, q. Let $a \in \mathbb{Q}$, $a > 0$ then $a = \frac{j}{k}$ with positive integers j, k. Let $n > kp$ and we obtain $na = \frac{nj}{k} \geqslant \frac{n}{k} > p \geqslant \frac{p}{q} = b$.

As a consequence of the Archimedean property of $\mathbb{R}$ we also get:

1.32. THEOREM. *If $a, b \in \mathbb{R}$, $a < b$, then there exists $q \in \mathbb{Q}$ such that $a < q < b$.*

PROOF. Assume first $a \geqslant 0$. Pick $n \in \mathbb{N}$ such that $n > \frac{2}{b-a}$ i.e. $\frac{1}{n} < \frac{b-a}{2}$. Let $m \in \mathbb{N}$ be the smallest natural number such that $\frac{m}{n} > a$ and then $\frac{m}{n} \leqslant a + \frac{1}{n}$. Hence $b - \frac{m}{n} \geqslant b - a - \frac{1}{n} > \frac{b-a}{2} > 0$ and thus $a < \frac{m}{n} < b$.

Let $a < 0$ and $b_\circ = \min\{b, 0\}$. By the above we have $-b_\circ < \frac{m}{n} < -a$, for some natural numbers m, n and thus $a < \frac{-m}{n} < b_\circ \leqslant b$. □

1.2.3. The binomial formula and roots. When dealing with powers and roots, the *binomial formula* can be very helpful. We first review some operations with natural numbers that are frequently used.

1.33. DEFINITION. Let n, k be natural numbers.

(i) The number n *factorial* is defined by

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

and $0! = 1$.

(ii) If $0 \leqslant k \leqslant n$, then the *binomial coefficient* $\binom{n}{k}$ (pronounced 'n choose k') is

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

and $\binom{n}{k} = 0$ if $k > n$.

The binomial coefficient $\binom{n}{k}$ equals the number of ways one may choose k objects from a set of n distinct objects, without repetition and without regard to ordering. Here are some basic properties of binomial coefficients.

1.34. THEOREM.

(i) For $0 \leqslant k \leqslant n$,

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

(ii) For $1 \leqslant \ell \leqslant n$,

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

(iii) For $a, b \in \mathbb{R}$ the binomial formula holds:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

PROOF. Item (i) follows by inspection of the definition of binomial coefficients. For (ii) we compute

$$\binom{n}{\ell - 1} + \binom{n}{\ell} = \frac{n!}{(\ell-1)!(n-\ell)!}\Big(\frac{1}{n-\ell+1} + \frac{1}{\ell}\Big) = \frac{(n+1)!}{\ell!(n+1-\ell)!} = \binom{n+1}{\ell}.$$

For (iii) we use induction: the case $n = 1$ follows by inspection. Assuming the formula for the power n we derive the formula for the power $n + 1$:

$$(a+b)^{n+1} = (a+b)(a+b)^n = (a+b)\sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k},$$

which equals

$$\begin{aligned}\sum_{k=0}^{n}\binom{n}{k}a^{k+1}b^{n-k} + \sum_{k=0}^{n}\binom{n}{k}a^k b^{n+1-k} &= \sum_{\ell=1}^{n+1}\binom{n}{\ell-1}a^\ell b^{n+1-\ell} + \sum_{k=0}^{n}\binom{n}{k}a^k b^{n+1-k}\\ &= \binom{n}{n}a^{n+1} + \sum_{\ell=1}^{n}\left[\binom{n}{\ell-1} + \binom{n}{\ell}\right]a^\ell b^{n+1-\ell} + \binom{n}{0}b^{n+1}.\end{aligned}$$

By (ii), and using $\binom{k}{0} = \binom{k}{k} = 1$ for $k = n$ and $k = n + 1$, this is equal to

$$\binom{n+1}{n+1}a^{n+1} + \sum_{\ell=1}^{n}\binom{n+1}{\ell}a^\ell b^{n+1-\ell} + \binom{n+1}{0}b^{n+1} = \sum_{k=0}^{n+1}\binom{n+1}{k}a^k b^{n+1-k}.$$

□

Binomial coefficients are often visualized as *Pascal's triangle*, which arises by writing out the non-zero coefficients row by row for increasing n starting with $n = 0$:

$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1
\end{array}$$

1.35. Exercise. Prove

$$\sum_{k=0}^{n}\binom{n}{k} = 2^n \qquad \text{and} \qquad \sum_{\substack{0\leqslant k\leqslant n\\ k \text{ even}}}\binom{n}{k} = \sum_{\substack{1\leqslant k\leqslant n\\ k \text{ odd}}}\binom{n}{k} = 2^{n-1}.$$

1.36. Exercise. Let X be a set of *cardinality* n (i.e. a set with n elements). Let $\mathcal{P}_k(X)$ be the collection of subsets of X which have cardinality k.

(i) Show that $\mathcal{P}_k(X)$ has cardinality $\binom{n}{k}$.
(ii) Show that $\mathcal{P}(X)$ has cardinality 2^n.

We can now show that any positive real number has a unique positive nth root.

1.37. Theorem (nth roots). *For every positive real number b there exists a unique positive real number a such that $a^n = b$.*

Notation. Given $b > 0$ we write $a = b^{1/n}$ or $a = \sqrt[n]{b}$ for the unique positive solution of $a^n = b$.

Proof. For $0 < \alpha < \beta$ we have $0 < \alpha^n < \beta^n$ which shows uniqueness.

Let $E = \{t > 0 : t^n < b\}$. For $t = \frac{b}{1+b}$ we have $t^n = (\frac{b}{1+b})^n \leqslant \frac{b}{1+b} < b$ so E is not empty. Moreover, if $t > 1 + b$, then $t^n > (1 + b)^n \geqslant b$ so $t \notin E$. That means if $t \in E$, then $t \leqslant 1 + b$ and E is bounded above by $1 + b$. Therefore, it has a least upper bound, denoted by $a = \sup E$. We will prove $a^n = b$ by ruling out the possibilities $a^n < b$ and $a^n > b$.

First assume $a^n < b$. For small positive $\varepsilon < 1$ estimate

$$(a + a\varepsilon)^n = a^n + a^n \sum_{k=1}^{n} \binom{n}{k} \varepsilon^k < a^n(1 + 2^n\varepsilon) < a^n + 2^n b\varepsilon$$

Thus, if $0 < \varepsilon < \frac{b-a^n}{2^n b}$ we obtain $(a+a\varepsilon)^n < b$ which means that a could not have been an upper bound for E. So we must have $a^n \geqslant b$.

Now assume $a^n > b$. For $0 < \varepsilon < 2^{-n}$ we have

$$(a - a\varepsilon)^n = a^n + a^n \sum_{k=1}^{n} (-1)^k \binom{n}{k} \varepsilon^k > a^n - a^n \sum_{k=1}^{n} \binom{n}{k} \varepsilon^k > a^n(1 - 2^n\varepsilon)$$

Thus if $0 < \varepsilon < \frac{a^n-b}{2^n a^n}$ then $a^n(1 - 2^n\varepsilon) > b$ and thus $(a - a\varepsilon)^n > b$, and we see that a could not have been the least upper bound for E. Hence we must have $a^n \leqslant b$.

We have proved that $a^n \geqslant b$ and $a^n \leqslant b$ and it follows $a^n = b$. □

1.2.4. The extended real line. It will often be convenient to augment the real line with two additional elements 'at positive and negative infinity', denoted ∞ and $-\infty$. The set $\mathbb{R} \cup \{\pm\infty\}$ is called the *extended real line*. Addition, multiplication and ordering naturally extend to the extended real line where applicable: for example, $a < \infty$ for every $a \in \mathbb{R}$, $a + \infty = \infty$ for every $a \in \mathbb{R}$, and so on. The operations $\pm\infty \cdot 0$ and $-\infty + \infty$ are not defined (the reason for this will become clear in Exercise 1.50 and Exercise 1.51). The extended real line allows in particular to define supremum (and infimum) for every set $A \subset \mathbb{R}$: if A is nonempty and has no upper bound, then we set

$$\sup A = \infty.$$

The supremum of the empty set is defined to be $-\infty$.

The real numbers are not *algebraically complete*: this means that polynomial equations in the reals do not always admit real-valued solutions. For example, the equation $z^2 = -1$ has no solution in the real numbers. The *complex numbers* $\mathbb{C}$ are designed to remedy this issue. Properties of complex numbers will be discussed in §2.7.

1.2.5. Topology of the real line. An *interval* I is a subset of real numbers that is one of the following types:

- *open*: $(a, b) = \{x \in \mathbb{R} : a < x < b\}$
- *closed*: $[a, b] = \{x \in \mathbb{R} : a \leqslant x \leqslant b\}$
- *half-open*: $[a, b) = \{x \in \mathbb{R} : a \leqslant x < b\}$, $(a, b] = \{x \in \mathbb{R} : a < x \leqslant b\}$

Here a and b are real numbers called *endpoints* of the interval, with $a \leqslant b$. We also allow the values $a = -\infty$ or $b = \infty$, provided the respective endpoint is not included, and call such intervals *unbounded*. Thus $(a, \infty) = \{x \in \mathbb{R} : x > a\}$ and $(-\infty, b) = \{x \in \mathbb{R} : x < b\}$ are unbounded open intervals, and $[a, \infty) = \{x \in \mathbb{R} : x \geqslant a\}$ and $(-\infty, b] = \{x \in \mathbb{R} : x \leqslant b\}$ are unbounded closed intervals.

1.38. Examples. The set of real numbers $\mathbb{R}$ is the unbounded open interval $(-\infty, \infty)$. The *positive half-line* is the unbounded open interval $(0, \infty)$. The (closed) unit interval is $[0, 1]$. Intervals with $a = b$ are also called degenerate. The empty set can be viewed as a degenerate open interval with $a = b$. The *singleton set* $\{x\}$ for a real number $x \in \mathbb{R}$ can be viewed as the degenerate closed interval $[x, x]$.

We may also generalize the notions of open and closed to arbitrary subsets $E \subset \mathbb{R}$. A point $x \in E$ is called an *interior point* if there exists $\varepsilon > 0$ so that $(x - \varepsilon, x + \varepsilon) \subset E$. A subset $E \subset \mathbb{R}$ is called *open* if every $x \in E$ is an interior point. A subset $E \subset \mathbb{R}$ is called *closed* if its complement $\mathbb{R} \backslash E$ is open.

It is instructive to convince oneself that open intervals are indeed open and closed intervals are indeed closed. Open and closed sets will be discussed in much greater detail in the more general context of metric spaces in Ch. 4.

1.3. Sequences

A sequence of real numbers, typically indexed by natural numbers,

$$a_1, a_2, a_3, \dots$$

will be denoted by the symbol $(a_n)_{n=1}^{\infty}$. The indices may also start at other integers, e.g. $(a_n)_{n=0}^{\infty}$, $(a_n)_{n=-3}^{\infty}$ etc. We may also use the notations (a_n) or $(a_n)_n$ to denote sequences if the index variable, and/or starting index are clear from context. It is helpful to understand that sequences are simply real-valued functions on the natural numbers (or other index set):

$$(a_n) : \mathbb{N} \longrightarrow \mathbb{R}, \quad n \longmapsto a_n.$$

In particular, if $\mathbf{a} = (a_n)$ denotes a sequence, we could write its elements as $a_n = \mathbf{a}(n)$, the value of the function $\mathbf{a}$ at n. Addition and multiplication of sequences are defined *termwise*:

$$(a_n) + (b_n) = (a_n + b_n),$$

$$(a_n) \cdot (b_n) = (a_n \cdot b_n).$$

1.3.1. Limits.

1.39. Definition. The sequence (a_n) *converges* to the *limit* $a \in \mathbb{R}$ if for every $\varepsilon > 0$ there exists a natural number $N = N(\varepsilon)$ such that for all $n \geqslant N$,

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

In that case we write $a_n \to a$ as $n \to \infty$ or

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

To warm up to this definition, let us show that limits are uniquely determined.

1.40. Observation. *If (a_n) converges to a and to a', then $a = a'$.*

Proof. Suppose $a \neq a'$. Applying the definition of convergence of (a_n) to a with $\varepsilon = \frac{1}{10}|a - a'| > 0$ there exists $N_1 \in \mathbb{N}$ such that for all $n \geqslant N_1$:

$$|a_n - a| < \tfrac{1}{10}|a - a'|.$$

Likewise, by definition of convergence of (a_n) to a' there exists $N_2 \in \mathbb{N}$ such that for all $n \geqslant N_2$:

$$|a_n - a'| < \tfrac{1}{10}|a - a'|.$$

Thus, if $n \geqslant \max(N_1, N_2)$, the triangle inequality implies

$$|a - a'| = |a - a_n + a_n - a'| \leqslant |a - a_n| + |a_n - a'| < \tfrac{1}{10}|a - a'| + \tfrac{1}{10}|a - a'| = \tfrac{1}{5}|a - a'|.$$

Since $|a - a'| > 0$ this implies $1 < \frac{1}{5}$, a contradiction. Thus $a = a'$. □

1.41. Exercise. Prove using the definition of convergence:

(i) The sequence $a_n = 2 + 10^{50}n^{-1}$ converges to the limit $a = 2$.
(ii) For every $\alpha > 0$ we have $\lim_{n\to\infty} n^{-\alpha} = 0$.
(iii) The sequence $a_n = (-1)^n$ does not converge.

A useful fact to show convergence of a sequence of real numbers is the *squeeze theorem*:

1.42. Observation. *Suppose that $(a_n), (b_n), (c_n)$ are sequences of real numbers satisfying*

$$a_n \leqslant b_n \leqslant c_n$$

and suppose that $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} c_n$ exist and are equal to L, then also $\lim_{n\to\infty} b_n$ exists and is equal to L.

PROOF. Let $\varepsilon > 0$. Since $a_n \to L$ there exists $N_1 \in \mathbb{N}$ such that $L - \varepsilon < a_n < L + \varepsilon$ for $n \geqslant N_1$. Since $c_n \to L$ there exists $N_2 \in \mathbb{N}$ such that $L - \varepsilon < c_n < L + \varepsilon$ for $n \geqslant N_2$. Let $N = \max(N_1, N_2)$. Since $a_n \leqslant b_n \leqslant c_n$ for all $n \geqslant N$ we have $L - \varepsilon < b_n < L + \varepsilon$ for $n \geqslant N$. Since $\varepsilon > 0$ was arbitrary this proves $\lim_{n\to\infty} b_n = L$. □

It is very useful to know how limits behave under arithmetic operations.

1.43. OBSERVATION. *Assume that* $A, B \in \mathbb{R}$ *and* $\lim_{n\to\infty} a_n = A$, $\lim_{n\to\infty} b_n = B$. *Then*

(i) $\lim_{n\to\infty}(a_n + b_n) = A + B$
(ii) $\lim_{n\to\infty} a_n b_n = AB$
(iii) If $B \neq 0$ *then* $\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{A}{B}$.

PROOF. (i) Let $\varepsilon > 0$. Then there exists $N_1 \in \mathbb{N}$ such that $|a_n - A| < \varepsilon/2$ for $n \geqslant N_1$, and there exists $N_2 \in \mathbb{N}$ such that $|b_n - B| < \varepsilon/2$ for $n \geqslant N_2$. Then for $n \geqslant \max(N_1, N_2)$,

$$|a_n + b_n - (A + B)| \leqslant |a_n - A| + |b_n - B| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

and the first claim is proved since $\varepsilon > 0$ was arbitrary.

(ii) Consider the product $a_n b_n$ and write

$$a_n b_n - AB = (a_n - A)b_n + A(b_n - B).$$

Let $\varepsilon > 0$. We may assume $\varepsilon \leqslant 1$. Then there exists $N_1 \in \mathbb{N}$ such that

$$|b_n - B| \leqslant \frac{\varepsilon}{2(1 + |A|)}$$

for $n \geqslant N_1$. In particular $|b_n - B| \leqslant 1$ for $n \geqslant N_1$, hence $|b_n| \leqslant 1 + |B|$. There exists $N_2 \in \mathbb{N}$ such that

$$|a_n - A| \leqslant \frac{\varepsilon}{2(1 + |B|)}$$

for $n \geqslant N_2$. Thus if $n \geqslant \max(N_1, N_2)$, then

$$|(a_n - A)b_n + A(b_n - B)| \leqslant |a_n - A|(1 + |B|) + |A||b_n - B| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

and, since $\varepsilon > 0$ was arbitrary, the claim is proved.

(iii) Let $B \neq 0$. By (ii) it suffices to prove $\lim_{n\to\infty} b_n^{-1} = B^{-1}$. Write

$$|b_n^{-1} - B^{-1}| = \frac{|B - b_n|}{|Bb_n|}.$$

Since $b_n \to B$ there is $N_1 \in \mathbb{N}$ such that $|b_n - B| < |B|/2$ for $n \geqslant N_1$ and hence

$$|b_n| \geqslant |B|/2 \text{ for } n \geqslant N_1.$$

Fix any $\varepsilon > 0$. There is $N_2 \in \mathbb{N}$ such that

$$|b_n - B| < \frac{|B|^2}{2}\varepsilon \text{ for } n \geqslant N_2.$$

Therefore $\frac{|B-b_n|}{|Bb_n|} < \varepsilon$ for $n \geqslant \max\{N_1, N_2\}$. Again, since $\varepsilon > 0$ was arbitrary, the claim is proved. □

1.44. EXAMPLE. Let $0 \leqslant x < 1$. Then

$$\lim_{n\to\infty} x^n = 0.$$

To see this write $x = (1 + a)^{-1}$ with $a > 0$. For $n \geqslant 1$ we have $(1 + a)^n \geqslant 1 + na$ by the binomial formula. Therefore $(1 + a)^{-n} \leqslant (1 + na)^{-1}$. For $\varepsilon > 0$ and $n \geqslant 1 + \lfloor \frac{1}{a\varepsilon} \rfloor$ we get $(1 + na)^{-1} < \varepsilon$ and therefore $0 < (1 + a)^{-n} < \varepsilon$. *Note:* Here $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x.

1.45. Example. Let $A \in \mathbb{R}$. Then

$$\lim_{n\to\infty} \frac{A^n}{n!} = 0.$$

To see this we choose $M \in \mathbb{N}$ such that $M \geqslant 2|A|$. Setting $c_M = \prod_{k=1}^{M} \frac{A}{k}$ we can write, for $n > M$,

$$\Big|\frac{A^n}{n!}\Big| = \frac{|A|^M}{M!} \prod_{k=M+1}^{n} \frac{|A|}{k} \leqslant \frac{|A|^M}{M!} 2^{M+1} 2^{-n}$$

and the assertion follows since $\lim_{n\to\infty} 2^{-n} = 0$.

1.46. Example. Let $a > 0$. Then

$$\lim_{n\to\infty} a^{1/n} = 1, \qquad \lim_{n\to\infty} n^{1/n} = 1.$$

For the second limit we apply the binomial formula to $(1 + x_n)^n$ with $x_n = n^{1/n} - 1 > 0$ and get

$$n = (1 + x_n)^n \geqslant \binom{n}{2} x_n^2 = \frac{n(n-1)}{2} x_n^2$$

which yields $0 \leqslant x_n \leqslant (\frac{2}{n-1})^{1/2}$. Hence $\lim_{n\to\infty} x_n = 0$ and thus $\lim_{n\to\infty} n^{1/n} = 1$. Assuming $a \geqslant 1$ we have $1 \leqslant a^{1/n} \leqslant n^{1/n}$ for $n \geqslant a$ and by the squeeze theorem $a^{1/n} \to 1$. This also holds for $0 < a < 1$ by the rules for quotients in Theorem 1.43.

1.47. Example (Exponential growth is faster than polynomial growth). For all integers $M \geqslant 0$ and real numbers $\delta > 0$:

$$\lim_{n\to\infty} \frac{n^M}{(1+\delta)^n} = 0.$$

To prove this apply the binomial formula for $n \geqslant 2k > 2M$

$$(1+\delta)^n \geqslant \binom{n}{k} \delta^k = \frac{n(n-1)\cdots(n-k+1)\delta^k}{k!} \geqslant \Big(\frac{n}{2}\Big)^k \frac{\delta^k}{k!}$$

which implies (still for $n \geqslant 2k > 2M$)

$$0 < \frac{n^M}{(1+\delta)^n} \leqslant n^{M-k} \frac{n^k}{(1+\delta)^n} \leqslant n^{M-k} \frac{2^k k!}{\delta^k}.$$

The claim follows by the squeeze theorem since $\lim_{n\to\infty} n^{M-k} = 0$ for $k > M$.

1.48. Definition. For a sequence (a_n) of real numbers we write

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

if for every $M > 0$ there exists $N \in \mathbb{N}$ such that $a_n > M$ for all $n \geqslant N$.

We write $\lim_{n\to\infty} a_n = -\infty$ if for every $M > 0$ there exists $N \in \mathbb{N}$ such that $a_n < -M$ for all $n \geqslant N$. If the limit is ∞ or $-\infty$ we say that the limit *exists in the extended sense*, or that the sequence converges in the extended real line.

1.49. Exercise. Show:

(i) If $\lim_{n\to\infty} a_n$ exists and $\lim_{n\to\infty} b_n = \infty$, then $\lim_{n\to\infty}(a_n + b_n) = \infty$.
(ii) If $A = \lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n = \infty$, then $\lim_{n\to\infty} a_n b_n = \infty$ if $A > 0$ and $\lim_{n\to\infty} a_n b_n = -\infty$ if $A < 0$.

It is important to realize that not all arithmetic operations for finite limits can be extended to the situation of Definition 1.48:

1.50. Exercise. Let $c \in \mathbb{R} \cup \{\pm\infty\}$. Find sequences (a_n), (b_n) with

$$\lim_{n\to\infty} a_n = \infty \quad \text{and} \quad \lim_{n\to\infty} b_n = -\infty$$

such that

$$\lim_{n\to\infty} (a_n + b_n) = c.$$

1.51. Exercise. Let $c > 0$. Find sequences (a_n), (b_n) with

$$\lim_{n\to\infty} a_n = 0 \quad \text{and} \quad \lim_{n\to\infty} b_n = \infty$$

such that

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

Can it also happen that $\lim_{n\to\infty} a_n b_n$ equals 0 or ∞?

1.52. Exercise. A *polynomial of degree* $d \in \mathbb{N}_0$ is a function of the form

$$p(x) = \sum_{j=0}^{d} a_j x^j$$

with $a_d \neq 0$ (the numbers a_j are called *coefficients*). Suppose p and q are polynomials of degrees $d \geqslant 0$ and $\ell \geqslant 1$, respectively. Prove that $q(n) \neq 0$ for sufficiently large n and determine

$$\lim_{n\to\infty} \frac{p(n)}{q(n)}$$

1.3.2. Monotone sequences. A sequence of real numbers (a_n) is called

- *non-increasing* or *decreasing* if $a_n \leqslant a_m$ for all $n \geqslant m$,
- *strictly decreasing* if $a_n < a_m$ for all $n \geqslant m$.

The terms *non-decreasing, increasing, strictly increasing* are defined accordingly. Increasing or decreasing sequences are also called *monotone*.

1.53. Theorem. *An increasing sequence (a_n) is convergent to a limit in $\mathbb{R}$ if it is bounded from above, or converges in the extended sense to ∞ if it is not bounded from above.*

Proof. Let $S = \sup\{a_n : n \in \mathbb{N}\}$. If $S = \infty$, there is nothing to show. Otherwise, for every $\varepsilon > 0$ there exists an N such that for all $n \geqslant N$, $S - \varepsilon \leqslant a_n \leqslant S$, by definition of the supremum as the least upper bound. This implies that $\lim_{n\to\infty} a_n = S$. □

An analogous statement holds for decreasing sequences: if (a_n) is decreasing, then $(-a_n)$ is increasing.

1.54. Example (Euler's number). Let

$$s_n = \sum_{k=0}^{n} \frac{1}{k!} = 1 + 1 + \frac{1}{2} + \cdots + \frac{1}{n!}$$

Then $s_n = s_{n-1} + \frac{1}{n!}$ so s_n is increasing. The sequence is also bounded above. Indeed $k! = 1 \cdot 2 \cdots k \geqslant 1 \cdot 2^{k-1}$ for $k \geqslant 2$. Hence $s_n \leqslant 2 + \sum_{k=2}^{n} 2^{1-k} < 3$ and thus s_n converges by Theorem 1.53. The limit is called *Euler's number* and is denoted

$$e = \lim_{n\to\infty} s_n. \tag{1.1}$$

The above estimation shows $2 < e < 3$. The sequence (s_n) converges rapidly and can be used to compute the decimal digits $e = 2.7182818284...$ We will later see that Euler's number is irrational (see Theorem 2.58). It shows up frequently in analysis.

1.55. Theorem. *The sequence*

$$a_n = \Big(1 + \frac{1}{n}\Big)^n$$

is increasing and converges to Euler's number e.

The proof is facilitated by using

1.56. Observation (Bernoulli's inequality).

$$(1 + x)^n \geqslant 1 + nx, \qquad \text{for } x \geqslant -1, n \in \mathbb{N}.$$

Note that this is immediate from the binomial theorem for $x \geqslant 0$ (and in this range we have used it before). However, below we need it for some negative choices of x. For the proof in the full range we use induction. The case $n = 1$ holds with equality, and, assuming the inequality for some $n \geqslant 1$, we get $(1 + x)^{n+1} \geqslant (1 + nx)(1 + x) = 1 + (n + 1)x + nx^2 \geqslant 1 + (n + 1)x$.

Proof of Theorem 1.55. Compute for $n \geqslant 2$

$$\frac{a_n}{a_{n-1}} = \Big(\frac{n+1}{n}\Big)^n \Big(\frac{n-1}{n}\Big)^n \frac{n}{n-1} = \Big(1 - \frac{1}{n^2}\Big)^n \frac{n}{n-1}.$$

Using Bernoulli's inequality 1.56 for $x = -n^{-2}$ we get

$$(1 - \frac{1}{n^2})^n \geqslant 1 - \frac{1}{n} = \frac{n-1}{n}.$$

It follows that $a_n/a_{n-1} \geqslant 1$, so a_n is increasing. To show that $\lim_{n\to\infty} a_n$ exists it suffices by Theorem 1.53 to show that a_n is bounded above. By the binomial theorem

$$\Big(1 + \frac{1}{n}\Big)^n = \sum_{k=0}^{n} \binom{n}{k} n^{-k} = 1 + \sum_{k=1}^{n} \frac{1}{k!} \prod_{j=0}^{k-1} \Big(1 - \frac{j}{n}\Big) < 1 + \sum_{k=1}^{n} \frac{1}{k!},$$

which proves both that $\lim_{n\to\infty}(1 + \frac{1}{n})^n = L$ exists and that $a_n \leqslant s_n$ where $s_n = \sum_{k=0}^{n} \frac{1}{k!}$, so $L \leqslant e$. To prove the opposite inequality, fix $m \in \mathbb{N}$ and estimate for $n \geqslant m$

$$a_n = 1 + \sum_{k=1}^{n} \frac{1}{k!} \prod_{j=0}^{k-1} \Big(1 - \frac{j}{n}\Big) \geqslant 1 + \sum_{k=1}^{m} \frac{1}{k!} \prod_{j=0}^{k-1} \Big(1 - \frac{j}{n}\Big).$$

Keeping m fixed and taking the limit as $n \to \infty$ (using rules for limits in Theorem 1.43) we get the inequality $L \geqslant 1 + \sum_{k=1}^{m} \frac{1}{k!}$ for every $m \geqslant 2$. Letting $m \to \infty$ gives $L \geqslant e$. □

1.57. Exercise. Define the sequence $(a_n)_{n=0}^{\infty}$ recursively by $a_1 = 1$, $a_{n+1} = \frac{a_n}{2} + \frac{1}{a_n}$ for $n \geqslant 1$. Prove that $a_n^2 > 2$ and $a_{n+1} > a_n$ for $n \geqslant 2$. Conclude that (a_n) converges and determine the limit.

1.3.3. Subsequences.

1.58. Definition. A sequence $\mathbf{b} = (b_n)_{n=1}^{\infty}$ is called a *subsequence* of $\mathbf{a} = (a_n)$ if there is a strictly increasing function $\kappa : \mathbb{N} \to \mathbb{N}$ such that $\mathbf{b} = \mathbf{a} \circ \kappa$, i.e. $b_n = a_{\kappa(n)}$ for all $n \in \mathbb{N}$.

1.59. Example. (a_{2n}) is a subsequence of (a_n) formed by the function $\kappa(n) = 2n$.

A sequence (a_n) is called *bounded* if there exists $C \in (0, \infty)$ such that $|a_n| \leqslant C$ for all n. Convergent sequences are always bounded (why?). While bounded sequences need not be convergent, they enjoy the following remarkable property.

1.60. Theorem (Bolzano–Weierstrass). *Every bounded sequence of real numbers has a convergent subsequence.*

PROOF. Let (a_n) be a sequence such that $|a_n| \leqslant M$ for all $n \in \mathbb{N}$. We construct a strictly increasing sequence of natural numbers $\kappa(n)$ and an increasing sequence b_n with $b_1 = \inf_{k \geqslant 1} a_k$, and for $n \geqslant 2$, $b_n = \inf_{k \geqslant \kappa(n-1)+1} a_k$, and $b_n \leqslant a_{\kappa(n)} \leqslant b_n + n^{-1}$. Assuming this, since (b_n) is increasing and $|b_n| \leqslant M$, it converges by Theorem 1.53. Since $|b_n - a_{\kappa(n)}| \leqslant n^{-1}$, the sequence $a_{\kappa(n)}$ is convergent, too, with the same limit.

The construction of $\kappa(n)$ and b_n is by induction. Let $b_1 = \inf_{k \geqslant 1} a_k$, and choose $\kappa(1)$ such that $b_1 \leqslant a_{\kappa(1)} \leqslant b_1 + 1$. This is possible by definition of the infimum and the observation that $|b_1| \leqslant M$.

Suppose $\kappa(1) < \kappa(2) < \cdots < \kappa(n)$ and $b_1 \leqslant b_2 \leqslant \cdots \leqslant b_n$ have already been constructed with $b_1 = \inf_{k \geqslant 1} a_k$, $b_l = \inf_{k \geqslant \kappa(l-1)+1} a_k$ for $2 \leqslant l \leqslant n$, and $b_l \leqslant a_{\kappa(l)} \leqslant b_l + l^{-1}$for $1 \leqslant l \leqslant n$. We have $|b_l| \leqslant M$ for $1 \leqslant l \leqslant n$. Define $b_{n+1} = \inf_{k \geqslant \kappa(n)+1} a_k$, so that $|b_{n+1}| \leqslant M$. By the definition of the infimum there is an index $\kappa(n+1) \geqslant \kappa(n) + 1$ so that $b_{n+1} \leqslant a_{\kappa(n+1)} \leqslant b_{n+1} + (n+1)^{-1}$. From the definitions of b_n and b_{n+1} we get $b_{n+1} \geqslant b_n$. This establishes the induction step. □

1.61. EXERCISE. Show that if a convergent sequence (c_n) is contained in a closed interval $[a, b]$, then the limit $\lim_{n\to\infty} c_n$ is also contained in $[a, b]$.

1.3.4. Limit superior and limit inferior.

1.62. DEFINITION. For a sequence (a_n) of real numbers that is bounded above we define the *limit superior* as

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \sup\{a_k \,:\, k \geqslant n\}.$$

If (a_n) is not bounded above we set $\limsup a_n = \infty$.

Similarly one defines the *limit inferior* for sequences that are bounded below as

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} \inf\{a_k \,:\, k \geqslant n\}.$$

If (a_n) is not bounded below we set $\liminf_{n\to\infty} a_n = -\infty$.

The notations $\overline{\lim}_{n\to\infty}$ and $\underline{\lim}_{n\to\infty}$ are also in common use for the limit superior and limit inferior, respectively. Note that the definition of the limit superior and limit inferior crucially relies on the least upper bound axiom and the fact that the non-increasing sequence $(\sup\{a_k \,:\, k \geqslant n\})_n$ converges.

1.63. EXERCISE. Show that

$$\liminf_{n\to\infty} a_n = -\limsup_{n\to\infty} (-a_n).$$

1.64. EXERCISE. Let (a_n) be a non-decreasing sequence. Show that

$$\lim_{n\to\infty} a_n = \sup\{a_n \,:\, n \in \mathbb{N}\}.$$

Also formulate and show the corresponding statement for non-increasing sequences.

1.65. EXERCISE. Let $(a_{\kappa(n)})$ be a convergent subsequence of a sequence of real numbers (a_n). Show that

$$\liminf_{n\to\infty} a_n \leqslant \lim_{n\to\infty} a_{\kappa(n)} \leqslant \limsup_{n\to\infty} a_n.$$

1.66. EXERCISE. Let $(a_{\kappa(n)})$ be a subsequence of a convergent sequence (a_n). Show that the subsequence converges to the same limit as a_n:

$$\lim_{n\to\infty} a_{\kappa(n)} = \lim_{n\to\infty} a_n.$$

Also show this for the case that the sequence is real-valued and the limit is ∞ or $-\infty$.

Let (a_n) be a sequence of real numbers, and let

$$\mathcal{L} = \big\{a \in \mathbb{R} \cup \{\infty, -\infty\} : a \text{ is the (extended) limit of a subsequence of } (a_n)\big\}. \tag{1.2}$$

We call $\mathcal{L}$ the *set of subsequential limits*.

1.67. THEOREM. *Let (a_n) be a sequence of real numbers, and let $\mathcal{L}$ be the set of subsequential limits as in* (1.2). *Then* $\inf \mathcal{L}$, $\sup \mathcal{L} \in \mathcal{L}$ *and*

$$\limsup_{n\to\infty} a_n = \sup \mathcal{L}, \qquad \liminf_{n\to\infty} a_n = \inf \mathcal{L}.$$

PROOF. It is no loss of generality to only prove the statements for lim sup. Consider the case where (a_n) is bounded above and (a_n) does not converge to $-\infty$. Then $B_n = \sup\{a_k : k \geqslant n\}$ is a decreasing sequence of real numbers with $B = \lim_{n\to\infty} B_n$ finite (indeed $B = \limsup_{n\to\infty} a_n$ by definition).

We first show that $\sup \mathcal{L} \leqslant B$. Let $\varepsilon > 0$. Since B_n decreases and converges to B there exists $N \in \mathbb{N}$ such that $B_N \leqslant B + \varepsilon$. This implies that $a_k \leqslant B + \varepsilon$ for $k \geqslant N$. If (b_n) is a convergent subsequence of (a_n), with $b_n = a_{\kappa(n)}$ and $\kappa(n)$ strictly increasing, then $b_n \leqslant B + \varepsilon$ provided that $\kappa(n) \geqslant N$ (in particular when $n \geqslant N$). Therefore if $L = \lim_{n\to\infty} b_n$ we get $L \leqslant B + \varepsilon$. Since the convergent subsequence was arbitrary we obtain $\sup \mathcal{L} \leqslant B + \varepsilon$. Since $\varepsilon > 0$ was arbitrary this implies $\sup \mathcal{L} \leqslant B$.

We now show the opposite inequality ($B \leqslant \sup \mathcal{L}$). To accomplish this we inductively construct two strictly increasing sequences of integers, $(\ell(n))$, $(\kappa(n))$ such that $\ell(n) \leqslant \kappa(n)$, $\ell(n+1) \geqslant \kappa(n)$, $B_{\ell(n)} \leqslant B + 2^{-n}$ and $a_{\kappa(n)} \geqslant B_{\ell(n)} - 2^{-n}$. For $n = 1$ we choose $\ell(1)$ such that $B_{\ell(1)} \leqslant B + \frac{1}{2}$ (possible since B_n converges to B), and then choose $\kappa(1) \geqslant \ell(1)$ such that $a_{\kappa(1)} \geqslant B_{\ell(1)} - \frac{1}{2}$ (possible by the definition of $B_{\ell(1)} = \sup\{a_k : k \geqslant \ell(1)\}$). Assume that we have defined $\ell(n)$, $\kappa(n)$ for $n \leqslant N$. Then we may choose $\ell(N+1) \geqslant \kappa(N)$ such that $B_{\ell(N+1)} \leqslant B + 2^{-N-1}$ and then $\kappa(N+1) \geqslant \ell(N+1)$ so that $a_{\kappa(N+1)} \geqslant B_{\ell(N+1)} - 2^{-(N+1)}$, completing the induction step.

Now consider the subsequence $(a_{\kappa(n)})$. We have $a_{\kappa(n)} \leqslant B_{\ell(n)} \leqslant B + 2^{-n}$ and $a_{\kappa(n)} \geqslant B_{\ell(n)} - 2^{-n} \geqslant B - 2^{-n}$ which implies that $|a_{\kappa(n)} - B| \leqslant 2^{-n}$, so $(a_{\kappa(n)})$ converges to B. Hence $B \in \mathcal{L}$ and therefore $B \leqslant \sup \mathcal{L}$. Combined with the above we obtain $B = \sup \mathcal{L}$, and we have also proved that $B = \sup \mathcal{L} \in \mathcal{L}$.

If (a_n) is bounded above and $\lim_{n\to\infty} B_n = -\infty$ then all subsequences of (a_n) converge to $-\infty$ in the extended sense and we have $\mathcal{L} = \{-\infty\}$. Finally if (a_n) is not bounded above then one checks that there is a subsequence $(a_{\kappa(n)})$ with $\kappa(n) < \kappa(n+1)$ and $a_{\kappa(n)} > n$. Then $\lim_{n\to\infty} a_{\kappa(n)} = \infty$ and $\sup \mathcal{L} = \limsup_{n\to\infty} a_n = \infty$, also $\infty \in \mathcal{L}$. □

1.68. EXERCISE. Let (a_n) be a sequence and $\mathcal{L}$ be the set of subsequential limits, as in (1.2).

Prove: If (b_n) is convergent with $b_n \in \mathcal{L}$ for all $n \in \mathbb{N}$ then its limit belongs to $\mathcal{L}$.

1.69. EXERCISE.

(i) Show that there exists a sequence (a_n) of real numbers such that the set of subsequential limits is $\mathbb{R}$.

(ii) Does there exist a sequence such that the set of subsequential limits equals $\mathbb{R}\backslash\{0\}$?

1.3.5. Cauchy sequences. It is often possible to prove convergence of a sequence without explicitly determining the limit (see the examples for monotone sequences in §1.3.2). In the absence of monotonicity, the following concept is useful.

1.70. DEFINITION. The sequence (a_n) is a *Cauchy sequence* if for every $\varepsilon > 0$ there exists a natural number $N = N(\varepsilon)$ such that for all $n, m \geqslant N$,

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

1.71. OBSERVATION. *Cauchy sequences are bounded.*

PROOF. Applying the definition with $\varepsilon = 1$, there exists $N \in \mathbb{N}$ such that $|a_n - a_N| < 1$ for all $n \geqslant N$. If we set $M = \max(|a_1|, \ldots, |a_N|)$ we have $|a_n| \leqslant M + 1$ for all $n \geqslant 1$, so (a_n) is bounded. □

1.72. OBSERVATION. *Convergent sequences are Cauchy sequences.*

PROOF. Suppose (a_n) converges to a limit a. Let $\varepsilon > 0$ and $N \in \mathbb{N}$ such that $|a_n - a| < \varepsilon$ for all $n \geqslant N$. Then for $n, m \geqslant N$,

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

so (a_n) is a Cauchy sequence. □

Remarkably, the converse to this is also true. This crucially relies on the existence of least upper bounds, i.e. the completeness of the real numbers. It also gives us a useful criterion to detect convergence of sequences without knowing what their limit is.

1.73. THEOREM (Cauchy criterion). *Every Cauchy sequence of real numbers is convergent.*

PROOF. Let (a_n) be a Cauchy sequence. Define

$$a = \limsup_{n\to\infty} a_n.$$

Since (a_n) is bounded, $a < \infty$ is a real number. By definition, a is the limit of the non-increasing sequence of least upper bounds $(\sup\{a_k : k \geqslant n\})_{n=1}^{\infty}$.

Let us show that (a_n) converges to a. Let $\varepsilon > 0$. By definition of Cauchy sequences, we may choose $N = N(\varepsilon) \in \mathbb{N}$ so that $|a_n - a_m| \leqslant \varepsilon$ for all $n, m \geqslant N$. In particular, for all $n \geqslant N$

$$a_n \leqslant a_N + \varepsilon$$

This means, $a_N + \varepsilon$ is an upper bound for every set $\{a_k : k \geqslant n\}$ with $n \geqslant N$, so

$$\sup\{a_k : k \geqslant n\} \leqslant a_N + \varepsilon$$

for $n \geqslant N$. Also,

$$\sup\{a_k : k \geqslant n\} \geqslant a_n \geqslant a_N - \varepsilon$$

for $n \geqslant N$. In summary, $a_N - \varepsilon \leqslant \sup\{a_k : k \geqslant n\} \leqslant a_N + \varepsilon$ for all $n \geqslant N$, which implies

$$a_N - \varepsilon \leqslant a \leqslant a_N + \varepsilon.$$

Finally, this implies for all $n \geqslant N$ that

$$|a_n - a| \leqslant |a_n - a_N| + |a_N - a| \leqslant 2\varepsilon,$$

proving the claim. □

REMARK. One can also take the Cauchy criterion as a definition of completeness and prove the existence of least upper bounds from it. In this sense, the Cauchy criterion and the existence of least upper bounds are equivalent (see §1.6.4).

1.3.6. Limits of functions.

1.74. DEFINITION. Let I be an interval, $f : I \to \mathbb{R}$ a function and $x_* \in I$. Then we write

$$\lim_{\substack{x\to x_* \\ x\in I}} f(x) = L$$

if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $x \in I$ with $0 < |x - x_*| < \delta$ we have

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

If x_* is not an endpoint of I we also simply write $\lim_{x\to x_*} f(x) = L$.

1.75. THEOREM. *Let f be defined on an open interval containing x_*. Then $\lim_{x\to x_*} f(x) = L$ holds if and only if $\lim_{n\to\infty} f(x_n) = L$ for all sequences (x_n) with $x_n \neq x_*$ and $\lim_{n\to\infty} x_n = x_*$.*

PROOF OF THEOREM 1.75. Suppose that $\lim_{x\to x_*} f(x) = L$. Let $x_n \to x_*$ (with $x_n \neq x_*$). Let $\varepsilon > 0$. Then there exists a $\delta > 0$ such that $|f(x) - L| < \varepsilon$ for $0 < |x - x_*| < \delta$. There exists $N_\delta \in \mathbb{N}$ such that $|x_n - x_*| < \delta$ for $n \geqslant N_\delta$ and then also $|f(x_n) - L| < \varepsilon$. Since $\varepsilon > 0$ was arbitrary we get $\lim_{n\to\infty} f(x_n) = L$.

Now suppose that $f(x_n) \to L$ whenever $x_n \to x_*$, with $x_n \neq x_*$. We argue by contradiction and assume that $\lim_{x\to x_*} f(x) = L$ does not hold. Then there exists an $\varepsilon > 0$ so that there is no $\delta > 0$ with the property that $|f(x) - L| < \varepsilon$ for all $x \neq x_*$ with $|x - x_*| < \delta$. Hence for every $n \in \mathbb{N}$ there exists an $x_n \neq x_*$ with $|x_n - x_*| < 1/n$ and $|f(x_n) - L| \geqslant \varepsilon$. We have $x_n \to x_*$ but $f(x_n)$ does not converge to L which gives us a contradiction. □

If x_* is the right endpoint of I, then $\lim\limits_{\substack{x\to x_*\\ x\in I}} f(x)$ is called a *one-sided limit* (from the left). In this instance one uses the notation

$$\lim_{x\to x_*-} f(x) = L.$$

The *one-sided limit from the right*, $\lim_{x\to x_*+} f(x) = L$ is defined analogously. It is useful to extend these notations also to cases where L or x_* is $\pm\infty$. For example:

1.76. DEFINITION. We say that

$$\lim_{\substack{x\to x_*\\ x\in I}} f(x) = \infty$$

if for every C there exists $\delta > 0$ such that for every $x \in I$ with $0 < |x - x_*| < \delta$ we have $f(x) > C$.

1.77. DEFINITION. Suppose that f is defined on an interval $I = (a, \infty)$. Then we write $\lim_{x\to\infty} f(x) = L$ if for every $\varepsilon > 0$ there exists a $b > a$ such that for every $x \geqslant b$ we have $|f(x) - L| < \varepsilon$.

Many other combinations are possible:

1.78. EXERCISE. Provide appropriate definitions for the following notions of limits.
(i) $\lim_{x\to\infty} f(x) = \infty$, (ii) $\lim_{x\to a+} f(x) = -\infty$,
(iii) $\lim_{x\to-\infty} f(x) = L$, (iv) $\lim_{x\to\infty} f(x) = -\infty$.

REMARK. The resulting zoo of slight variations on the definition of a limit can be elegantly unified using the notion of *filters*, see Bourbaki [**7**].

1.3.7. Asymptotic notation. We are often interested in the *asymptotic* behavior of a sequence (a_n), i.e., behavior as $n \to \infty$. Suppose $b_n > 0$ for all n. The following notation is common:

- *Big-O:* We write
$$a_n = O(b_n)$$
to mean that there exists a constant $C \in (0, \infty)$ such that
$$|a_n| \leqslant Cb_n$$
for all natural numbers n. This is equivalent to saying that $\limsup_{n\to\infty} \frac{|a_n|}{b_n} < \infty$.
- *Little-o:* We write
$$a_n = o(b_n)$$
to mean that for every $\varepsilon > 0$ there exists N such that for all $n \geqslant N$,
$$|a_n| \leqslant \varepsilon b_n.$$
In other words, $a_n = o(b_n)$ means $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

The asymptotic notation is also used to indicate the asymptotic behavior of the functions $f(x)$ as x tends to a limit a (or $\pm\infty$). Let f be a function defined in an open interval near a and let $g > 0$ be a positive function defined on the same interval. Then

$$f(x) = O(g(x)) \quad (\text{as } x \to a)$$

means that

$$\limsup_{x\to a} \frac{|f(x)|}{g(x)} < \infty.$$

If a is a real number, then this is equivalent to saying that there exists a constant $C \in (0,\infty)$ and $\delta > 0$ such that

$$|f(x)| \leqslant C\, g(x)$$

for all x with $0 < |x-a| < \delta$. Similarly,

$$f(x) = o(g(x)) \quad (\text{as } x \to a)$$

means that

$$\lim_{x\to a} \frac{f(x)}{g(x)} = 0.$$

It is important to realize that O and o are *not* functions and

$$f(x) = O(g(x)), \quad f(x) = o(g(x))$$

are *not* equations. One can think of $O(g)$ as a placeholder for a function which may change at every occurrence of the symbol $O(g)$ but always satisfies the respective condition that it is dominated by a constant times $|g(x)|$ if $|x-a|$ is sufficiently small. When it is clear from context, one omits the phrase "as $x \to a$".

1.79. Examples.

- If $a_n = O(b_n)$ and $0 < b_n \leqslant Cc_n$ for a constant $C \in (0,\infty)$, then $a_n = O(c_n)$. In other words, if $a_n = O(b_n)$ and $b_n = O(c_n)$, then $a_n = O(c_n)$.
- A sequence (a_n) converges to 0 if and only if $a_n = o(1)$ and it is bounded if and only if $a_n = O(1)$.
- We have $\frac{(-1)^n}{5n^2-1} = O(\frac{1}{n^2})$. It is also $o(\frac{1}{n})$, but not $o(\frac{1}{n^2})$.
- If $f(x) = O(|x-a|^k)$, as $x \to a$, then $f(x) = o(|x-a|^{k-\varepsilon})$ for every $\varepsilon > 0$.

There are also other asymptotic notations in use: for example, $f(x) \sim g(x)$ as $x \to \infty$ means $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$. For example, $x^3 + x^2 \sim x^3 + x$ as $x \to \infty$.

1.4. Continuity

1.80. Definition. Let I be an interval. A function $f : I \to \mathbb{R}$ is *continuous at* $x_* \in I$ if

$$\lim_{\substack{x\to x_*,\\ x\in I}} f(x) = f(x_*).$$

If f is continuous at every point in I, then we call f *continuous (on I).*

The special case when x_* is an endpoint of I comes with extra terminology: A function f is called *left-continuous*, resp. *right-continuous* at x_* if $\lim_{x\to x_*-} f(x) = f(x_*)$, resp. $\lim_{x\to x_*+} f(x) = f(x_*)$. Continuity can be phrased in terms of sequences:

1.81. Observation. *f is continuous at x_* if and only if for every sequence (x_n) in I, $\lim_{n\to\infty} x_n = x_*$ implies $\lim_{n\to\infty} f(x_n) = f(x_*)$. This property is also called sequential continuity.*

(This follows from Theorem 1.75.) In particular, we see from the properties of limits that sums and products of continuous functions are continuous.

1.82. Exercise. Let I, J be intervals and let $f : J \to \mathbb{R}$ and $g : I \to J$ be continuous. Show that the composition $f \circ g : I \to \mathbb{R}$, $(f \circ g)(x) = f(g(x))$ is continuous.

1.83. Examples.

(i) The constant function $f(x) = 1$ and the identity function $f(x) = x$ are continuous on $\mathbb{R}$.

(ii) Since sums and products of continuous functions are continuous, all polynomials are continuous on $\mathbb{R}$.
(iii) The function $f(x) = \frac{1}{x}$ is continuous on $(0, \infty)$.
(iv) The function

$$f(x) = \begin{cases} 0, & \text{if } x < 2, \\ 1, & \text{if } x \geqslant 2 \end{cases}$$

is right-continuous at $x_* = 2$, but not left-continuous. It is continuous at every point $x_* \neq 2$.

1.84. Exercise (Thomae's function). Let $f : (0,1) \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 0 & \text{if } x \in (0,1)\backslash\mathbb{Q}, \\ \frac{1}{n} & \text{if } x = \frac{m}{n} \text{ with } \gcd(m,n) = 1. \end{cases}$$

Show that f is continuous at every $x_* \in (0,1)\backslash\mathbb{Q}$, but not continuous at $x_* \in (0,1) \cap \mathbb{Q}$.

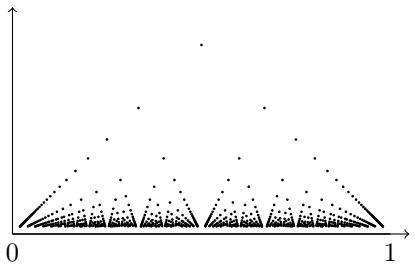


Figure 1.1. Non-zero values of Thomae's function visualized.

1.85. Exercise. Let $\mathbb{Q} = \{q_1, q_2, \dots\}$ be an enumeration of the rationals and define

$$f(x) = \sum_{n \in \mathbb{N} \,:\, q_n < x} 2^{-n}$$

Determine at exactly which points f is continuous.

1.4.1. Existence of minima and maxima.

1.86. Definition (Bounded function). A function $f : [a,b] \to \mathbb{R}$ is called *bounded* on $[a,b]$ if there exists $M > 0$ so that $|f(x)| \leqslant M$ for all $x \in [a,b]$.

We write $\sup_{x \in [a,b]} f(x)$ for $\sup\{f(x) : x \in [a,b]\}$, and $\inf_{x \in [a,b]} f(x)$ for $\inf\{f(x) : x \in [a,b]\}$. For a bounded function these quantities are finite.

We shall now use the Bolzano–Weierstrass theorem (Theorem 1.60) to prove that every continuous function on a closed bounded interval $[a,b]$ is bounded and has a maximum and a minimum point.

1.87. Theorem. *Let $f : [a,b] \to \mathbb{R}$ be continuous. Then f is a bounded function and there exist $\xi, \tilde{\xi} \in [a,b]$ so that*

$$f(\xi) = \sup_{x \in [a,b]} f(x), \quad f(\tilde{\xi}) = \inf_{x \in [a,b]} f(x).$$

Proof. Suppose

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

is not bounded above. Then there exists $x_n \in [a,b]$ such that $f(x_n) \geqslant n$. By the Bolzano–Weierstrass theorem (Theorem 1.60) there is a strictly increasing sequence of integers $(\kappa(n))$ so that $x_{\kappa(n)}$ converges to $x_* \in [a,b]$. Since f is continuous at x_* there exists a $\delta > 0$ such that for all $x \in [a,b]$ with $|x - x_*| \leqslant \delta$ we have $f(x) < f(x_*) + 1$, hence

$$f(x_{\kappa(n)}) < f(x_*) + 1$$

for sufficiently large n. This gives a contradiction.

Let $S = \sup A < \infty$. For every n we find $w_n \in [a,b]$ such that

$$S - n^{-1} \leqslant f(w_n) \leqslant S.$$

Again by the Bolzano-Weierstrass theorem there exists a subsequence $w_{k(n)}$ which converges to some $\xi \in [a,b]$. It follows by continuity of f at ξ that $f(\xi) = S$. The corresponding statements for functions bounded below follow by considering $-f$. □

REMARK. It is essential in this theorem that f is continuous on a closed interval $[a,b]$. It is not enough that the interval is bounded: for example, the function $f(x) = \frac{1}{x}$ is continuous on $(0,1)$, but is not bounded on $(0,1)$.

1.4.2. The intermediate value theorem. If f is continuous on $[a,b]$, then intuition suggests that every number between $f(a)$ and $f(b)$ should be in the range of f (give an example that this may fail for functions with a discontinuity in (a,b)). It is important to realize that this is not obvious and requires a proof.

1.88. THEOREM (Intermediate value theorem). *Let $f : [a,b] \to \mathbb{R}$ be continuous. Then for every $y \in \mathbb{R}$ with $f(a) < y < f(b)$ or $f(b) < y < f(a)$ there exists $\xi \in (a,b)$ with $f(\xi) = y$.*

PROOF. It suffices to consider the case $f(a) < y < f(b)$ (otherwise consider $-f$). Let

$$\xi = \sup\{x \in [a,b] \,:\, f(x) \leqslant y\}.$$

This is well-defined by completeness of the real numbers (a bounded nonempty set has a least upper bound). We claim that $f(\xi) = y$.

Let us first prove $\xi \in (a,b)$. Since $f(a) < y$ we see by the continuity of f at a that there exists $x > a$ with $f(x) < y$, so $\xi \geqslant x > a$. Similarly, by continuity of f at b and since $f(b) > y$ there exists $\delta > 0$ such that $f(x) > y$ for all $b - \delta \leqslant x \leqslant b$, which implies $\xi \leqslant b - \delta < b$. Next, by the definition of sup as a *least* upper bound, for every δ with $0 < \delta < \xi - a$ there exists an $x \in [\xi - \delta, \xi]$ such that $f(x) \leqslant y$. By continuity of f this implies $f(\xi) \leqslant y$. Moreover, by the definition of sup as an upper bound we have $f(x) > y$ for $\xi < x \leqslant b$, and by the continuity of f this implies $f(\xi) \geqslant y$. We have proved $f(\xi) = y$. □

1.4.3. Monotone functions. A real-valued function f on an interval I is called

- *increasing* or *non-decreasing* if $f(x_1) \leqslant f(x_2)$ for all $x_1 < x_2$ in I,
- *strictly increasing* if $f(x_1) < f(x_2)$ for all $x_1 < x_2$ in I,

The terms *non-increasing, decreasing, strictly decreasing* are defined accordingly. A function that is increasing or decreasing is called *monotone* and a function that is strictly increasing or strictly decreasing is called *strictly monotone*.

Observe that if f is strictly monotone, then it is one-to-one and thus has an inverse function f^{-1} defined on its image $f(I)$. If f is also continuous, then $f(I)$ is an interval (by the intermediate value theorem).

1.89. THEOREM. *Let $f : [a,b] \to [c,d]$ be continuous and strictly monotone with*

$$c = \min(f(a), f(b)), \quad d = \max(f(a), f(b)).$$

Then f is bijective and the inverse function

$$f^{-1} : [c,d] \to [a,b]$$

is continuous.

PROOF. Assume without loss of generality that f is strictly increasing (if f is strictly decreasing, consider $-f$). Then $c = f(a)$, $d = f(b)$. f is one-to-one because it is strictly increasing and surjective by the intermediate value theorem. It remains to show that f^{-1} is continuous, i.e. for every $y_0 \in [c, d]$ and $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|f^{-1}(y) - f^{-1}(y_0)| < \varepsilon.$$

whenever $y, y_0 \in [c, d]$ and $|y - y_0| < \delta$. Fix $y_0 \in [c, d]$ and $\varepsilon > 0$. There is a unique $x_0 \in [a, b]$ with $f(x_0) = y_0$.

Pick a positive $\varepsilon_0 \leqslant \varepsilon$ as follows: If $a \leqslant x_0 < b$ then we require $x_0 + \varepsilon_0 \leqslant b$ and if $a < x_0 \leqslant b$ we require $a \leqslant x_0 - \varepsilon_0$ (i.e. if $a < x_0 < b$ both conditions are satisfied). Note that $f(x_0 + \varepsilon_0) \leqslant d$ and $f(x_0 - \varepsilon_0) \geqslant c$. Now define

$$\delta = \begin{cases} \min(f(x_0 + \varepsilon_0) - y_0,\ y_0 - f(x_0 - \varepsilon_0)) & \text{if } c < y_0 < d \\ f(x_0 + \varepsilon_0) - y_0 & \text{if } y_0 = c \\ y_0 - f(x_0 - \varepsilon_0) & \text{if } y_0 = d \end{cases}$$

Note that $\delta > 0$ since $f(x_0) = y_0$ and f is strictly increasing. We verify $|f^{-1}(y) - f^{-1}(y_0)| < \varepsilon$, distinguishing the cases $c \leqslant y_0 < d$, $y_0 \leqslant y < y_0 + \delta$, and $c < y_0 \leqslant d$, $y_0 - \delta < y \leqslant y_0$.

First, consider $c \leqslant y_0 < d$, $y_0 \leqslant y < y_0 + \delta$. By definition of δ we have $y_0 + \delta \leqslant f(x_0 + \varepsilon_0)$. There is a unique $x \in [a, b]$ such that $f(x) = y$ and then $f(x_0) \leqslant f(x) < f(x_0 + \varepsilon_0)$. Since f is increasing we get $x_0 \leqslant x \leqslant x_0 + \varepsilon_0$, i.e. $f^{-1}(y_0) \leqslant f^{-1}(y) < f^{-1}(y_0) + \varepsilon_0$.

Second, consider $c < y_0 \leqslant d$, $y_0 - \delta < y \leqslant y_0$. By definition of δ we have $y_0 - \delta \geqslant f(x_0 - \varepsilon_0)$. Again $y = f(x)$ for a unique $x \in [a, b]$ and we get $f(x_0) \geqslant f(x) > f(x_0 - \varepsilon_0)$. Since f is strictly increasing this implies $x \in (x_0 - \varepsilon_0, x_0]$ and thus $f^{-1}(y_0) - \varepsilon_0 < f^{-1}(y) \leqslant f^{-1}(y_0)$.

Combining both cases we see that $|f^{-1}(y) - f^{-1}(y_0)| < \varepsilon_0 < \varepsilon$ for $y, y_0 \in [c, d]$ with $|y - y_0| < \delta$. □

1.4.4. Uniform continuity. The following notion will be important in the discussion of integrals in §2.1 below.

1.90. DEFINITION. Let I be an interval. A function $f : I \to \mathbb{R}$ is *uniformly continuous* if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $|f(x) - f(x')| < \varepsilon$ provided that $x, x' \in I$ and $|x - x'| < \delta$.

Compare this definition with the definition of continuity at a given point where the δ depends on ε and the given point. Here the δ is chosen uniformly over all $x \in I$. The following exercise shows that a function f can be continuous at every $x \in I$ but not uniformly continuous in I.

1.91. EXERCISE. Define $f : (0, 1] \to \mathbb{R}$ as follows. For $n \in \mathbb{N}$ the definition on $(2^{-n}, 2^{1-n}]$ is

$$f(x) = \begin{cases} 2^{n+1}x - 2 & \text{if } 2^{-n} < x \leqslant 2^{-n} + 2^{-n-1}, \\ 4 - 2^{n+1}x & \text{if } 2^{-n} + 2^{-n-1} < x \leqslant 2^{1-n}. \end{cases}$$

(i) Sketch the graph.
(ii) Show that f is continuous on $(0, 1]$ but not uniformly continuous.
(iii) Let $g : [0, 1] \to \mathbb{R}$ be a function such that $g(x) = f(x)$ for $0 < x \leqslant 1$. Show that g is not continuous at 0.

The failure of uniform continuity in this exercise is due to the large oscillation of f near 0, so that f cannot be extended to a continuous function on $[0, 1]$. The outcome is different for continuous functions on closed bounded intervals:

1.92. THEOREM. *Let $a, b \in \mathbb{R}$, $a < b$ and let $f : [a, b] \to \mathbb{R}$ be a continuous function. Then f is uniformly continuous on $[a, b]$.*

This result will be crucial in the integration theory for continuous functions. The proof below uses the special properties of an interval, and the least upper bound property. It could be skipped for now, since an alternative (independent) proof which works in a more general setup will be given in §4.3.1.

PROOF OF THEOREM 1.92. We may assume $a < b$, otherwise there is nothing to show. Let $\varepsilon > 0$. We let $\mathcal{P}(\varepsilon)$ be the set of points $c \in [a, b]$ with the property that there exists $\delta > 0$ (depending on c and ε) such that for all $x, x' \in [a, c]$ with $|x - x'| \leqslant \delta$ we have $|f(x) - f(x')| < \varepsilon$. We have $a \in \mathcal{P}(\varepsilon)$. Moreover, by definition, if $c \in \mathcal{P}(\varepsilon)$, then $[a, c] \subset \mathcal{P}(\varepsilon)$. We let $u = \sup \mathcal{P}(\varepsilon)$, then $a \leqslant u \leqslant b$.

We now show that $u = b$, and argue by contradiction. Assume that $u < b$. f is continuous at u and thus there exists $\gamma_1 > 0$ such that $\gamma_1 < b - u$ and $|f(x) - f(u)| < \varepsilon/2$ whenever $|x - u| \leqslant \gamma_1$ and $x \in [a, b]$. By the triangle inequality, we get for $x, x' \in [u - \gamma_1, u + \gamma_1] \cap [a, b]$,

$$|f(x) - f(x')| \leqslant |f(x) - f(u)| + |f(u) - f(x')| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Since u is the least upper bound for $\mathcal{P}(\varepsilon)$ there exists $u_0 \in [a, b]$ such that $u - \gamma_1/4 \leqslant u_0 \leqslant u$ and $u_0 \in \mathcal{P}(\varepsilon)$, and thus $[a, u_0] \subset \mathcal{P}(\varepsilon)$. Therefore there exists $\gamma_2 > 0$ such that $|f(x) - f(x')| < \varepsilon$ whenever $|x - x'| < \gamma_2$ and $x, x' \in [a, u_0]$. Let $\gamma = \min(\gamma_1/4, \gamma_2)$. If $a \leqslant x \leqslant u - \gamma_1/2$ and $|x - x'| \leqslant \gamma$ then $x, x' \in [a, u - \gamma_1/4] \subset [a, u_0]$ and $|x - x'| \leqslant \gamma_2$. Hence $|f(x) - f(x')| < \varepsilon$. If $x \in [u - \gamma_1/2, u + \gamma_1/2] \cap [a, b]$ and $|x - x'| \leqslant \gamma$ then $x, x' \in [u - \gamma_1, u + \gamma_1] \cap [a, b]$ and therefore, as shown above, $|f(x) - f(x')| < \varepsilon$. Consequently $[a, u + \gamma_1/2] \cap [a, b] \subset \mathcal{P}(\varepsilon)$, which is a contradiction to the hypothesis that u is an upper bound for $\mathcal{P}(\varepsilon)$. Hence we must have $u = b$.

Finally we prove that $b \in \mathcal{P}(\varepsilon)$ (by essentially repeating the argument above). Since f is continuous at b there exists $\delta_1 > 0$ such that $|f(x) - f(b)| < \varepsilon/2$ for $|x - b| \leqslant \delta_1$ and therefore, again by the argument using the triangle inequality, we get $|f(x) - f(x')| < \varepsilon$ provided that $x, x' \in [b - \delta_1, b]$. We have proved above that $b = \sup \mathcal{P}(\varepsilon)$ and thus there exists $b_0 \in [b - \delta_1/4, b) \cap \mathcal{P}(\varepsilon)$. Then also $[a, b_0] \subset \mathcal{P}(\varepsilon)$. By definition of $\mathcal{P}(\varepsilon)$ there exists $\delta_2 > 0$ such that $|f(x) - f(x')| < \varepsilon$ provided that $|x - x'| \leqslant \delta_2$ and $x, x' \in [a, b_0]$. Let $x, x' \in [a, b]$ and $|x - x'| < \delta = \min\{\delta_1/4, \delta_2\}$. We claim that $|f(x) - f(x')| < \varepsilon$. To check this we again distinguish two cases. In the first case we have $a \leqslant x \leqslant b - \delta_1/2$ then $x' \leqslant b - \delta_1/2 + \delta \leqslant b_0$ and therefore $|f(x) - f(x')| < \varepsilon$. In the second case $b - \delta_1/2 \leqslant x \leqslant b$ and if $x' \in [a, b]$ satisfies $|x - x'| \leqslant \delta$ then $|x - b| \leqslant \delta_1$ and $|x' - b| \leqslant \delta_1$. Therefore $|f(x) - f(x')| \leqslant |f(x) - f(b)| + |f(b) - f(x')| < \varepsilon/2 + \varepsilon/2 = \varepsilon$, for all $x, x' \in [a, b]$ such that $|x - x'| \leqslant \delta$. We have proved that $[a, b] = \mathcal{P}(\varepsilon)$.

Since $\varepsilon > 0$ was arbitrary we deduce that f is uniformly continuous on $[a, b]$. □

1.5. Further exercises

1.93. EXERCISE. Show that there is a bijection from $[0, 1]$ to $[0, 1]^2$.

1.94. EXERCISE. Prove that $\sqrt{6}$, $\sqrt[3]{2}$, $\sqrt{2} + \sqrt{3}$, $\sqrt{2} + \sqrt{3} + \sqrt{5}$ are irrational.

1.95. EXERCISE. Prove or disprove convergence for each of the following sequences and in case of convergence, determine the limit:

(i) $a_n = \sqrt{n^4 + \cos(n^2)} - n^2$

(ii) $a_n = n^2 + \frac{1}{2}n - \sqrt{n^4 + n^3}$

(iii) $a_n = \sum_{k=n}^{n^2} \frac{1}{k}$

(iv) $a_0 = 3$, $a_{n+1} = \frac{a_n + 6}{a_n + 2}$

(v) $a_n = \prod_{k=2}^{n} \frac{k^2 - 1}{k^2}$

1.96. EXERCISE. Define the sequence a_n by $a_0 = A$, $a_1 = B$, and $a_n = \frac{1}{2}(a_{n-2} + a_{n-1})$ for $n \geqslant 2$. Show that (a_n) converges and determine the limit.

1.97. EXERCISE. Let $a_0 \in \mathbb{R}$ and define $a_{n+1} = 2a_n(1 - a_n)$ recursively for $n \geqslant 0$. Show that for all $n \geqslant 0$,

$$1 - 2a_n = (1 - 2a_0)^{2^n}.$$

1.98. Exercise. Let $b_0 \in [0,1]$ and define $b_{n+1} = 4b_n(1-b_n)$ recursively for $n \geqslant 0$. Show that for all $n \geqslant 0$,

$$b_n = (\sin(2^n \arcsin(\sqrt{b_0})))^2$$

1.99. Exercise. Prove that the sequence $a_n = (1+\frac{1}{n})^{n+1}$ is decreasing, and that for $n \in \mathbb{N}$

$$\left(1+\frac{1}{n}\right)^n < e < \left(1+\frac{1}{n}\right)^{n+1}.$$

1.100. Exercise. Find sequences $(a_n)_{n=1}^\infty$, $(b_n)_{n=1}^\infty$ such that

$$\begin{aligned}\liminf a_n + \liminf b_n &< \liminf(a_n+b_n) < \liminf a_n + \limsup b_n \\ &< \limsup(a_n+b_n) < \limsup a_n + \limsup b_n.\end{aligned}$$

1.101. Exercise. (i) Construct a function $f : \mathbb{R} \to \mathbb{R}$ such that f is continuous at every $x \in \mathbb{Z}$ and discontinuous at every $x \notin \mathbb{Z}$.

(ii) For every interval (open, half-open or closed) $I \subset \mathbb{R}$ give an example of a function $f : \mathbb{R} \to \mathbb{R}$ such that f is continuous on I and discontinuous on $\mathbb{R}\backslash I$.

1.6. Appendix: Complete ordered fields

The main purpose of this appendix will be to prove the existence and uniqueness (up to isomorphisms) of the real numbers, i.e., to provide a proof of Theorem 1.30. We emphasize that the result of this theorem is fundamental for analysis, but it is not necessary for a real understanding of analysis to go through the sometimes tedious details of the proof. We begin with a general discussion of ordered fields and their isomorphisms in §1.6.1, prove the existence of real numbers in 1.6.2 and establish uniqueness up to isomorphisms in §1.6.3. Equivalent notions of completeness (such as the one involving convergence of Cauchy sequences) will be discussed in §1.6.4.

Remark. For an extensive treatment of the foundations of the natural, rational, real and complex numbers, see e.g. Landau [**46**]. In §1.6.1 and §1.6.2 we essentially follow Rudin [**65**].

1.6.1. Ordered fields.

1.102. Definition. A *field* $(\mathbb{K}, +, \cdot)$ is a set $\mathbb{K}$ with operations of addition and multiplication

$$\text{add} : \begin{matrix}(x,y) \mapsto x+y \\ \mathbb{K}\times\mathbb{K} \to \mathbb{K}\end{matrix}, \qquad \text{mul} : \begin{matrix}(x,y) \mapsto x\cdot y \\ \mathbb{K}\times\mathbb{K} \to \mathbb{K}\end{matrix}$$

such that the following axioms hold.

(add-comm) Addition is commutative: $x+y = y+x$.
(add-assoc) Addition is associative: $x+(y+z) = (x+y)+z$.
(add-zero) There is an element $0 \in \mathbb{K}$ such that $0+x = x$ for all $x \in \mathbb{K}$.
(add-neg) For every $x \in \mathbb{K}$ there is $-x \in \mathbb{K}$ such that $x+(-x) = 0$.
(mul-comm) Multiplication is commutative: $x \cdot y = y \cdot x$.
(mul-assoc) Multiplication is associative: $x \cdot (y \cdot z) = (x \cdot y) \cdot z$.
(mul-one) There is an element $1 \in \mathbb{K}$ such that $1 \neq 0$ and $1 \cdot x = x$ for all $x \in \mathbb{K}$.
(mul-inv) For every $x \in \mathbb{K}$ with $x \neq 0$ there exists an element x^{-1} such that $x \cdot x^{-1} = 1$.
(mul-add) Distributivity: $x \cdot (y+z) = (x \cdot y) + (x \cdot z)$ for $x, y, z \in \mathbb{K}$.

The inverse element x^{-1} for multiplication is also denoted by $\frac{1}{x}$ or $1/x$. Multiplication $x \cdot y$ will often be written as xy. We also follow the usual convention that multiplication has priority over addition, so that the distributivity law can be rewritten as $x(y+z) = xy + xz$. Subtraction is defined by $x - y = x + (-y)$.

1.103. Definition. A *commutative (or Abelian) group* $(\mathbb{G}, +)$ is a set $\mathbb{G}$ with an operation of addition $+ : \mathbb{G} \times \mathbb{G} \to \mathbb{G}$ such that axioms (add-comm), (add-assoc), (add-zero), (add-neg) hold. 0 is also called the *neutral element* and $-x$ is called the *inverse element* of $x \in \mathbb{G}$.

Various familiar rules follow from the axioms.

1.104. Lemma. *Let $\mathbb{G}$ be a commutative group, with addition $+$. Then for all $x, y, z \in \mathbb{G}$*

(i) *If $x + y = x + z$, then $y = z$.*
(ii) *If $x + y = x$, then $y = 0$.*
(iii) *If $x + y = 0$, then $y = -x$.*
(iv) *$-(-x) = x$.*

Proof. (i) First note that by the axioms $y = -x + (x + y)$ and $z = -x + (x + z)$:

$$\begin{aligned} y &= 0 + y && \text{by (add-zero)} \\ &= (x + (-x)) + y && \text{by (add-neg)} \\ &= (-x + x) + y && \text{by (add-comm)} \\ &= -x + (x + y) && \text{by (add-assoc)} \end{aligned}$$

Then $x + y = x + z$ implies $y = -x + (x + y) = -x + (x + z) = z$ and (i) is proved.

(ii) If $x + y = x$ then $x + y = x + 0$ by (add-zero) and (add-comm), so (ii) follows from (i).

(iii) If $x + y = 0$, then $x + y = x + (-x)$ by (add-neg) and (i) implies $y = -x$.

(iv) By (add-neg) and (add-comm) we have $-x + x = 0$, so by (iii) applied to $-x$ in place of x we get $x = -(-x)$. □

If $\mathbb{K}$ is a field, then the set $\mathbb{K}\backslash\{0\}$ is a commutative group with respect to the multiplication operation, with the neutral element equal to 1, and inverse elements given by x^{-1}. Thus Lemma 1.104 implies:

1.105. Lemma. *Let $\mathbb{K}$ be a field. Then for all $x, y, z \in \mathbb{K}$*

(i) *If $x \neq 0$ and $xy = xz$, then $y = z$.*
(ii) *If $x \neq 0$ and $xy = x$, then $y = 1$.*
(iii) *If $x \neq 0$ and $xy = 1$, then $y = x^{-1}$.*
(iv) *If $x \neq 0$, then $(x^{-1})^{-1} = x$.*

Combinations with the distributivity axiom (mul-add) yield

1.106. Lemma. *Let $\mathbb{K}$ be a field. Then for all $x, y \in \mathbb{K}$*

(i) *$0 \cdot x = 0$.*
(ii) *If $x \neq 0$ and $y \neq 0$, then $xy \neq 0$.*
(iii) *$(-x) \cdot y = x \cdot (-y) = -(x \cdot y)$.*
(iv) *$(-x) \cdot (-y) = x \cdot y$.*

Proof. (i) By (mul-add) and (mul-comm),

$$0 \cdot x + 0 \cdot x = x \cdot 0 + x \cdot 0 = x \cdot (0 + 0) = (0 + 0) \cdot x = 0 \cdot x$$

and by Lemma 1.104 (ii) it follows that $0 \cdot x = 0$.

(ii) Assume $x \neq 0$, $y \neq 0$. To show $xy \neq 0$ we argue by contradiction, so assume $xy = 0$. Then:

$$\begin{aligned} 1 &= x \cdot x^{-1} && \text{by (mul-inv)} \\ &= (x \cdot 1) \cdot x^{-1} && \text{by (mul-one), (mul-comm)} \\ &= (x \cdot (y \cdot y^{-1})) \cdot x^{-1} && \text{by (mul-inv)} \\ &= (x \cdot y) \cdot (y^{-1} \cdot x^{-1}) && \text{by (mul-assoc)} \\ &= 0 \cdot (y^{-1} \cdot x^{-1}) && \text{by assumption} \\ &= 0 && \text{by (i),} \end{aligned}$$

which is a contradiction to (mul-one).

(iii) By (mul-add) and (mul-comm),

$$(-x)y + xy = (-x + x) \cdot y,$$

which equals $0 \cdot y = 0$ by (add-neg), (add-comm) and (i). Similarly,

$$x(-y) + xy = x(-y + y) = x \cdot 0 = 0 \cdot x = 0.$$

Using Lemma 1.104 (iii) concludes the proof.

(iv) Apply (iii) twice:

$$(-x)(-y) = -(x(-y)) = -(-(xy)).$$

Finally, $-(-(xy)) = xy$ by Lemma 1.104 (iv). □

We now introduce the concept of order.

1.107. Definition. An *ordered field* $\mathbb{K}$ is a field $\mathbb{K}$ equipped with a total order $\preceq$ (see §1.1.5) such that

(ord-add) If $x, y, z \in \mathbb{K}$ and $y \preceq z$ then $x + y \preceq x + z$.
(ord-mul) If $x, y \in \mathbb{K}$ and $0 \preceq x$, $0 \preceq y$ then $0 \preceq x \cdot y$.

We write $a \precneqq b$ if $a \preceq b$ and $a \neq b$. We use the notation $b \succeq a$ synonymously with $a \preceq b$, and $b \succneqq a$ synonymously with $a \precneqq b$. Elements x with $0 \precneqq x$ (i.e. $0 \preceq x$ and $x \neq 0$) are called *positive,* and elements x with $x \precneqq 0$ are called *negative.*

Again, various other familiar properties for inequalities are implied by the axioms.

1.108. Lemma. *Let $\mathbb{K}$ be an ordered field. Then for all $x, y, z \in \mathbb{K}$*

(i) *$x \succeq 0$ if and only if $-x \preceq 0$.*
(ii) *If $x \succeq 0$ and $y \preceq z$, then $xy \preceq xz$.*
(iii) *If $x \preceq 0$ and $y \preceq z$, then $xy \succeq xz$.*
(iv) *If $x \neq 0$, then $x \cdot x \succeq 0$.*
(v) *$1 \succneqq 0$.*
(vi) *If $x \succeq 0$ and $y \preceq 0$, then $xy \preceq 0$.*
(vii) *If $0 \precneqq x \preceq y$, then $0 \precneqq y^{-1} \preceq x^{-1}$.*

Proof. (i) Assume first $x \succeq 0$. Using (add-comm), (add-neg), (add-zero) and (ord-add), $0 = -x + x \succeq -x + 0 = -x$ and thus $-x \preceq 0$. Similarly, if $x \preceq 0$, then $0 = -x + x \preceq -x + 0 = -x$ so that $-x \succeq 0$.

(ii) Using $z \succeq y$ and (ord-add) we get $z - y \succeq y - y = 0$. From this and $x \succeq 0$ we get $x \cdot (z - y) \succeq 0$ by (ord-mul). Using (add-zero), (add-comm), (add-neg), (mul-add) we get

$$xz = x(z - y) + xy$$

and by (ord-add) this is $\succeq 0 + xy$, which equals xy by (add-zero).

(iii) By the same reasoning as in (ii) we have $z - y \succeq 0$. By $x \preceq 0$ and (i) we have $-x \succeq 0$. By Lemma 1.106 (iii) and (ord-mul) we get

$$-(x(z - y)) = (-x)(z - y) \succeq 0$$

and thus $x(z-y) \preceq 0$. Hence $xz \preceq xy$ by the same reasoning as in (ii).

(iv) Since $\preceq$ is a total order, we have $x \succeq 0$ or $x \preceq 0$. Suppose first $x \succeq 0$. Then $x \cdot x \succeq 0$ by (ord-mul). Next suppose $x \preceq 0$. Then $-x \succeq 0$, and we get $(-x) \cdot (-x) \succeq 0$. By Lemma 1.106 (iv), $x \cdot x = (-x) \cdot (-x)$, so $x \cdot x \succeq 0$.

(v) $1 \succeq 0$ follows from (iv) and (mul-one). Also $1 \neq 0$ holds by (mul-one).

(vi) $y \preceq 0$ implies $-y \succeq 0$ by (i). Then by (ord-mul) and Lemma 1.106 (iii), $-(xy) = x(-y) \succeq 0$, so $xy \preceq 0$ by (i).

(vii) Suppose that $0 \precneqq y^{-1}$ does not hold. Then we must have $y^{-1} \preceq 0$ since $\preceq$ is a total order. Since also $y \succneqq 0$, we have from (vi) that $y \cdot y^{-1} \preceq 0$. But, also $y \cdot y^{-1} = 1 \succneqq 0$ by (mul-inv) and (v), leading to a contradiction. This proves $0 \precneqq y^{-1}$. By the same reasoning, $0 \precneqq x^{-1}$. By (ord-mul) then also $0 \preceq x^{-1}y^{-1}$. It remains to prove $y^{-1} \preceq x^{-1}$. Then by (ii) and (mul-one), (mul-inv), (mul-assoc), (mul-comm),

$$y^{-1} = x \cdot (x^{-1}y^{-1}) \preceq y \cdot (x^{-1}y^{-1}) = x^{-1}.$$

□

Sometimes we may need to consider two ordered fields and their relation.

1.109. Definition. Let $(\mathbb{F}, +, \cdot, \leqslant)$, $(\mathbb{K}, \oplus, \odot, \preceq)$ be ordered fields. A map $\varphi : \mathbb{F} \to \mathbb{K}$ is called an *isomorphism of ordered fields*, if it is bijective and satisfies

$$\begin{aligned}
&\varphi(p+q) = \varphi(p) \oplus \varphi(q)\\
&\varphi(p \cdot q) = \varphi(p) \odot \varphi(q)\\
&p \leqslant q \text{ if and only if } \varphi(p) \preceq \varphi(q)
\end{aligned}$$

for all $p, q \in \mathbb{F}$.

1.110. Theorem. *Let $\mathbb{K}$ be an ordered field. Then there is a unique isomorphism of ordered fields, $\phi : \mathbb{Q} \to \mathbb{K}_{\mathbb{Q}}$ mapping $\mathbb{Q}$ to a unique ordered subfield $\mathbb{K}_{\mathbb{Q}}$ of $\mathbb{K}$.*

A *subfield* of $\mathbb{K}$ is a subset that is a field by itself.

Proof sketch. We denote by $\oplus$ the addition in $\mathbb{K}$, by $\odot$ the multiplication, by O the zero element, by $\mathbb{1}$ the neutral element with respect to multiplication, by -x the negative of x, and by $\preceq$ the total order in $\mathbb{K}$.

We first show uniqueness but the argument will also lead to an existence proof. Let $\phi : \mathbb{Q} \to \mathbb{K}$ be an injective map satisfying the three conditions in Definition 1.109. Then $\phi(p) = \phi(p) \oplus \phi(0)$ for all $p \in \mathbb{Q}$ which implies $\phi(0) = O$. Also $\phi(p) = \phi(p) \odot \phi(1)$, which implies $\phi(1) = \mathbb{1}$. Since $\phi(n) = \phi(n-1) \oplus \mathbb{1}$, and $\phi(-n) =$ -$\phi(n)$, this defines ϕ uniquely on all $n \in \mathbb{Z}$. Also $\mathbb{1} = \phi(n\frac{1}{n}) = \phi(n) \odot \phi(\frac{1}{n})$ which uniquely determines $\phi(\frac{1}{n})$. We arrive at $\phi(\frac{m}{n}) = \phi(m) \odot (\phi(n))^{-1}$.

For the existence we argue similarly. We define $\phi : \mathbb{N} \to \mathbb{K}$ by $\phi(0) = O$, $\phi(1) = \mathbb{1}$ and $\phi(n) = \phi(n-1) + \mathbb{1}$. Define for $n \in \mathbb{N}$, $\phi(-n) =$ -$\phi(n)$. This extends ϕ to the integers. Because $\mathbb{K}$ is an ordered field we have $\phi(n-1) \preceq \phi(n)$ and $\phi(n-1) \neq \phi(n)$ which implies $\phi(n) \neq 0$ for all nonzero integers n. By an induction argument one can then show for all $m, n \in \mathbb{Z}$:

$$\begin{aligned}
\phi(m+n) &= \phi(m) \oplus \phi(n)\\
\phi(mn) &= \phi(m) \odot \phi(n)
\end{aligned}.$$

Now we extend the definition of ϕ to $\mathbb{Q}$ by setting $\phi(\frac{m}{n}) = (\phi(m)) \odot (\phi(n))^{-1}$, provided that m/n is written in lowest terms. We notice that we can drop the assumption that m/n is written in lowest terms, since we get

$$\phi\Big(\frac{m}{n}\Big) = (\phi(m)) \odot (\phi(n))^{-1}$$

for arbitrary integers m, n with $n \neq 0$. To check this, assume $\frac{m}{n} = \frac{p}{q}$ with integers p, q, so $mq = np$ and thus $\phi(m) \odot \phi(q) = \phi(n) \odot \phi(p)$ and thus $\phi(m) \odot \phi(n)^{-1} = \phi(p) \odot \phi(q)^{-1}$. Using this, one can check that for all $r, s \in \mathbb{Q}$

$$\begin{aligned}
&\phi(r+s) = \phi(r) \oplus \phi(s)\\
&\phi(rs) = \phi(r) \odot \phi(s)\\
&\phi(r) \preceq \phi(s) \text{ if and only if } r \leqslant s.
\end{aligned}$$

This shows that $\mathbb{K}_{\mathbb{Q}} = \{\phi(r) : r \in \mathbb{Q}\}$ is a subfield of $\mathbb{K}$ which is isomorphic to $\mathbb{Q}$ as an ordered field. □

1.111. Definition. For an ordered field $\mathbb{K}$ the subfield $\mathbb{K}_{\mathbb{Q}}$ is called the field of rational elements in $\mathbb{K}$.

1.112. Definition. Let $(\mathbb{K}, +, \cdot, \preceq)$ be an ordered field and let $\phi : \mathbb{Q} \to \mathbb{K}_{\mathbb{Q}}$ be the canonical isomorphism. The field $\mathbb{K}$ is *Archimedean* (or has the *Archimedean property*) if for every $a, \varepsilon \in \mathbb{K}$ with $\varepsilon \succnsim 0$, $a \succnsim 0$ there exists an $n \in \mathbb{N}$ such that $a \precnsim \phi(n)\varepsilon$.

Notice that $\phi(n)\varepsilon$ is just the sum $\varepsilon + \cdots + \varepsilon$ with n summands, each equal to ε.

1.113. Theorem. *Let $(\mathbb{K}, +, \cdot, \preceq)$ be an Archimedean ordered field. If $a, b \in \mathbb{K}$ such that $a \precnsim b$ then there exists $q \in \mathbb{K}_{\mathbb{Q}}$ such that $a \precnsim q \precnsim b$.*

Proof. The proof is essentially the same as the proof of Theorem 1.32. Let $\phi : \mathbb{Q} \to \mathbb{K}_{\mathbb{Q}}$ be the canonical isomorphism. Assume first $a \succeq 0$. Pick $n \in \mathbb{N}$ such that $\phi(2) \precnsim \phi(n)(b-a)$. Hence $\phi(\frac{1}{n}) \precnsim (b-a)\phi(\frac{1}{2})$. By the Archimedean property there is a smallest $m \in \mathbb{N}$ such that $\phi(n)a \precnsim \phi(m)$. Then

$$\phi(m) \preceq \phi(n)a + 1$$

and therefore (using $\phi(m)(\phi(n))^{-1} = \phi(\frac{m}{n})$)

$$a \precnsim \phi(\tfrac{m}{n}) \preceq a + \phi(\tfrac{1}{n}) \precnsim a + \phi(\tfrac{1}{2})(b-a) \precnsim a + b - a = b.$$

Hence

$$b - \phi(\tfrac{m}{n}) \succeq b - a - \phi(\tfrac{1}{n}) \succnsim b - a - (b-a)\phi(\tfrac{1}{2}) = (b-a)\phi(\tfrac{1}{2})$$

and thus $a \precnsim \phi(\frac{m}{n}) \precnsim b$. If $a \precnsim 0$, let $b_0 = \min\{b, 0\}$. By the above we have $-b_0 \precnsim \phi(\frac{m}{n}) \precnsim -a$, for some natural numbers m, n and thus $a \precnsim \phi(-\frac{m}{n}) \precnsim b_0 \preceq b$. □

1.6.2. Existence. We follow the construction of the real numbers due to Dedekind (1872). The idea is to identify a real number ρ with an open interval in $\mathbb{Q}$ of the form $(-\infty, \rho) \cap \mathbb{Q}$, but ρ is not initially defined. We thus begin with an axiomatic definition of such intervals.

1.114. Definition (Dedekind cuts). A *Dedekind cut,* or simply *cut*, is a set $A \subset \mathbb{Q}$ satisfying the following properties for all $p, q, r \in \mathbb{Q}$:

(De1) $A \neq \varnothing$.
(De2) $A \neq \mathbb{Q}$.
(De3) If $p \in A$, $q \in \mathbb{Q}$ and $q \leqslant p$, then $q \in A$.
(De4) If $p \in A$, then $p < r$ for some $r \in A$.

The set of Dedekind cuts is denoted by $\mathbb{R}$ and is called the set of *real numbers.*

Property (De3) describes the interval structure described above, and property (De4) says that there is no largest element in a cut. As a direct consequence of the definition we get:

1.115. Observation. *Let A be a cut. Then the following hold for every $p, q \in \mathbb{Q}$:*

(i) *If $p \in A$ and $q \notin A$, then $p < q$.*
(ii) *If $p \notin A$ and $p \leqslant q$, then $q \notin A$.*

1.116. Definition. On the set $\mathbb{R}$ of Dedekind cuts we take for the partial order $\preceq$ the inclusion relation, i.e. $A \preceq B$ if $A \subseteq B$.

1.117. Lemma.

(i) $\mathbb{R}$ *is totally ordered.*
(ii) $\mathbb{R}$ *has the least upper bound property.*

Proof. (i) Let A, B be two cuts and assume that A is not contained in B. Then there exists $q \in A \backslash B$. For any $p \in B$ we have $p < q$ since $q \notin B$. But this implies that $p \in A$, by (De3). Hence $B \subseteq A$. It follows that $\mathbb{R}$ is totally ordered.

(ii) Let $\mathcal{E} \subset \mathbb{R}$ be a nonempty set of cuts which is bounded by $\mathcal{M} \in \mathbb{R}$. We define

$$U = \bigcup_{A \in \mathcal{E}} A$$

and intend to show that U is the least upper bound to $\mathcal{E}$.

We first need to check that $U \in \mathbb{R}$ (that is to verify (De1)-(De4) for the set U). Since $\mathcal{E}$ is nonempty, there is $A_0 \in \mathcal{E}$ and $A_0 \neq \varnothing$ by (De1), moreover $A_0 \subset U$ so $U \neq \varnothing$. Since $\mathcal{E}$ is bounded above by $\mathcal{M} \in \mathbb{R}$ and $U \subset \mathcal{M} \neq \mathbb{Q}$, we see that $U \neq \mathbb{Q}$. To check (De3) let $p \in U$, $q \in \mathbb{Q}$, $q < p$; we need to prove $q \in U$. By definition of U there is $A \in \mathcal{E}$ such that $p \in A$. Since $q < p$ we also have $q \in A \subset U$, which is what we wanted to prove. For (De4) if $p \in U$, then p belongs to some $A \in \mathcal{E}$, thus by (De4) for A there exists $r \in A$ such that $p < r$, and clearly $r \in U$. This establishes that $U \in \mathbb{R}$.

By definition, $A \subseteq U$ for every $A \in \mathcal{E}$, so U is an upper bound for $\mathcal{E}$. To show that U is the least upper bound we pick $V \in \mathbb{R}$ such that $V \subsetneq U$. Then there exists $r \in U \backslash V$, and $r \in A$ for some $A \in \mathcal{E}$. Then by Observation 1.115 (i) applied to V we conclude $V \subsetneq A$ so V is not an upper bound for $\mathcal{E}$. Thus we have shown that U is the least upper bound to $\mathcal{E}$. □

We shall now turn to the axioms for addition.

1.118. Lemma. *For two cuts A, B let $A \oplus B$ be the set of all $p + q$ where $p \in A$, $q \in B$. Then $A \oplus B$ is a Dedekind cut. Let the zero element O be defined as the set of negative rational numbers (which is a Dedekind cut). Then the axioms (add-comm), (add-assoc), (add-zero), (add-neg) hold.*

Proof. (De1), (De2) Since A, B are nonempty, so is $A \oplus B$. Also there exist $p_0 \in \mathbb{Q} \backslash A$, $q_0 \in \mathbb{Q} \backslash B$ which means $p_0 + q_0 > p + q$ for all $p \in A$, $q \in B$. Hence $p_0 + q_0 \notin A \oplus B$.

(De3) Let $p \in A \oplus B$, $q \in \mathbb{Q}$ and let $q \leqslant p$; we must show that $q \in A \oplus B$. We can write $p = r + s$ for $r \in A$, $s \in B$. Let $q \in \mathbb{Q}$ with $q \leqslant p$. Then $q - s \leqslant r$ hence $q - s \in A$, so $q = q - s + s \in A \oplus B$.

(De4) Let $p \in A \oplus B$, and split $p = r + s$ with $r \in A$, $s \in B$. Then there exists $r' \in A$, $s' \in B$ with $r' > r$, $s' > s$ and we get $r' + s' \in A \oplus B$ with $r' + s' > p$.

(add-comm), (add-assoc): These follow from the definition of addition, and the fact that (add-comm), (add-assoc) hold in $\mathbb{Q}$.

(add-zero) Let $p \in A$ and let $q \in O$ be a negative rational number. Then $p + q < p$, hence $p + q \in A$ and we obtain $A \oplus O \subseteq A$. For the opposite inclusion, let $p \in A$. By (De4) there is $r > p$ with $r \in A$ and we get $p - r < 0$, hence $p - r \in O$. Now $p = r + (p - r) \in A \oplus O$, and it follows $A \subseteq A \oplus O$. Altogether $A = A \oplus O$ and (add-zero) is verified.

(add-neg) We need to define the negative of A. We claim that it is given by[1]

$$B = \text{-}A = \{p \in \mathbb{Q} : \text{There exists } r > 0 \text{ such that } -p - r \notin A\}.$$

[1]We use in this section the -A notation (and not $-A$) for the Dedekind cut "negative (A)" to emphasize that the set -A is *different* from the set $\{-a : a \in A\}$.

We begin by verifying that $B \in \mathbb{R}$. First, $B \neq \varnothing$; indeed since A is nonempty we pick $q \in \mathbb{Q}\backslash A$ and set $p = -q - 1$ then $-p - 1 = q \notin A$, so $p \in B$. Secondly $\mathbb{Q}\backslash B \neq \varnothing$ since for $q \in A$ we have $q - r \in A$ for all $r > 0$ which implies $-q \notin B$.

To check (De3) for B, let $p \in B$ and $q \in \mathbb{Q}$ with $q < p$. Then there exists $r > 0$ such that $-p - r \notin A$ and since $-q - r > -p - r$ we also have $-q - r \notin A$. Thus $q \in B$.

To check (De4) for B let $p \in B$. Then there exists $r > 0$ such that $-p - r \notin A$, i.e. $-p - \frac{r}{2} - \frac{r}{2} \notin A$. Since $\frac{r}{2} > 0$ this implies $p + \frac{r}{2} \in B$ and we have $p + \frac{r}{2} > p$, thus (De4) is verified. We have now checked that $B \in \mathbb{R}$. To complete the verification of (add-neg) we must prove that $A \oplus B = O$, and this is done by proving that $A \oplus B \subseteq O$ and $O \subseteq A \oplus B$.

Let $p \in A$, $q \in B$. Then there exists $r > 0$ such that $-q - r \notin A$, and thus $-q \notin A$. This implies $p < -q$ and hence $p + q < 0$. We have shown $A \oplus B \subseteq O$.

For the converse pick $r \in O$ (any negative rational number). Let $t = -\frac{1}{2}r$, a positive rational number. Then there is an integer n such that $nt \in A$ but $(n+1)t \notin A$. Let $s = -(n+2)t$, then $-s - t \notin A$ and thus $s \in B$. We see that $r = -2t = nt + s \in A \oplus B$ and thus we have proved $O \subseteq A \oplus B$. □

We now turn to multiplication.

1.119. Definition.

(i) Let $\mathbb{R}^+ = \{A \in \mathbb{R} : O \subsetneq A\}$ and $\mathbb{R}^- = \{A \in \mathbb{R} : A \subsetneq O\}$.
(ii) Let $\mathbb{1}$ be the Dedekind cut in $\mathbb{R}^+$ defined by $\mathbb{1} = \{r \in \mathbb{Q} : r < 1\}$.
(iii) For $A, B \in \mathbb{R}^+$ let $A \odot B$ denote

$$A \odot B = \{p \in \mathbb{Q} : p \leqslant rs \text{ for some } r \in A,\ s \in B \text{ with } r > 0, s > 0\}.$$

(iv) For any $A \in \mathbb{R}$ define $A \odot O = O$, $O \odot A = O$.
(v) For the remaining cases define

$$A \odot B = \begin{cases} (\text{-}A) \odot (\text{-}B) & \text{if } A \in \mathbb{R}^-,\ B \in \mathbb{R}^- \\ \text{-}\big((\text{-}A) \odot B\big) & \text{if } A \in \mathbb{R}^-, B \in \mathbb{R}^+ \\ \text{-}\big(A \odot (\text{-}B)\big) & \text{if } A \in \mathbb{R}^+, B \in \mathbb{R}^- \end{cases}$$

1.120. Lemma. *For $A, B \in \mathbb{R}^+$ we have $A \odot B \in \mathbb{R}^+$, and the multiplication and distributivity axioms hold for $\mathbb{R}^+$, i.e. for $A, B, C \in \mathbb{R}^+$,*

(mul-comm) $A \odot B = B \odot A$.
(mul-assoc) $A \odot (B \odot C) = (A \odot B) \odot C$.
(mul-one) $\mathbb{1} \neq O$ *and* $\mathbb{1} \odot A = A$ *for all* $A \in \mathbb{R}^+$.
(mul-inv) *For every* $A \in \mathbb{R}^+$ *there exists an element* $A^{-1} \in \mathbb{R}^+$ *such that* $A \odot A^{-1} = \mathbb{1}$.
(mul-add) *Distributivity:* $A \odot (B \oplus C) = A \odot B \oplus A \odot C$.

Remark. We use the usual convention to remove parenthesis in $(A \odot B) \oplus (A \odot C)$; i.e. multiplication takes precedence over addition.

Proof of Lemma 1.120. First we verify that $A \odot B \in \mathbb{R}^+$.

(De1) For $A, B \in \mathbb{R}^+$ we have $r \in A$ for some rational $r > 0$ and $s \in B$ for some rational s and thus $rs \in A \odot B$. Thus $A \odot B$ satisfies (De1).

(De2) Since A and B are cuts, there are $p_0 \in \mathbb{Q}\backslash A$, $q_0 \in \mathbb{Q}\backslash B$ with $p_0 > 0$, $q_0 > 0$. We have $p_0 > p$ for all $p \in A$ and $q_0 > q$ for all $q \in B$ and thus $p_0 q_0 > rs$ whenever $r \in A$, $s \in B$ are positive rational numbers. Thus $p_0 q_0 \notin A \odot B$ which means $A \odot B$ satisfies (De2).

(De3) Every nonpositive rational number is in $A \odot B$. Let $p > 0$, $p \in A \odot B$ and $p \leqslant rs$, $r \in A$, $s \in B$. Let $0 < q \leqslant p$, then $q \leqslant \frac{q}{p} rs$ and $\frac{q}{p} r \in A$, $s \in B$, so $q \in A \odot B$.

(De4) Let $p \in A \odot B$. Without loss of generality we may assume $p > 0$. Then $p \leqslant rs$ with $r \in A$, $s \in B$. There exists a rational $u > 1$ such that $ur \in A$ and thus $p < up \leqslant (ur)s$ and so $up \in A \odot B$.

This completes the verification of $A \odot B \in \mathbb{R}^+$.

(mul-comm) and (mul-assoc) are a consequence of the commutativity and associativity axioms for multiplication of rational numbers.

(mul-one) First, $\mathbb{1} \neq O$ since $\mathbb{1}$ contains all rational numbers in $(0,1)$. If $p \leqslant rs$, $r < 1$, $s \in A$, then $p \leqslant s$. This shows $\mathbb{1} \odot A \subseteq A$. Vice versa, if $p \in A$, then there is $r > p$ with $r \in A$, therefore $p = \frac{p}{r} r$ and $\frac{p}{r} \in \mathbb{1}$, $r \in A$. Hence $A \subseteq \mathbb{1} \odot A$. Altogether $\mathbb{1} \odot A = A$ for $A \in \mathbb{R}^+$.

(mul-inv) The definition of the multiplicative inverse for $A \in \mathbb{R}^+$ is

$$S = A^{-1} = \{p \in \mathbb{Q} : p > 0, \text{ and there exists } r \text{ such that } r \notin A \text{ and } r < p^{-1}\} \cup \{0\} \cup O.$$

We check that $S \in \mathbb{R}$. Clearly S is nonempty, so (De1) holds. Let $q \in A$ and $q > 0$. Then $2/q \notin S$ (since the condition $r < (2/q)^{-1} = q/2$ implies $r \in A$). Thus (De2) holds.

Let $p \in S$ and $q \leqslant p$. Then there exists $r \notin A$ such that $r \leqslant 1/p$, but $1/p \leqslant 1/q$ so also $r \leqslant 1/q$. This means $q \in S$, so (De3) is verified.

Now let $p \in S$. Then there is $q \notin A$ such that $q < p^{-1}$. Define $q_0 = q + \frac{1}{3}(p^{-1} - q)$ and $q_1 = q + \frac{2}{3}(p^{-1} - q)$. Then $q < q_0 < q_1 < \frac{1}{p}$. We have thus $q_0 \notin A$, and $q_0 < \frac{1}{p}$. With $r_0 = \frac{1}{q_1}$ we have $r_0 > p$, and $q_0 < 1/r_0$. Thus $r_0 \in S$ and we have verified (De4) for the set S.

We need to check that $A \odot S = \mathbb{1}$. If $p \in A \odot S$ then there exists $r \in A$, $s \in S$, $r > 0$, $s > 0$, $p \leqslant rs$, and by definition of S, there exists $t \notin A$ such that $t < 1/s$. Then $r \leqslant t$. Hence $p \leqslant rs \leqslant ts < \frac{1}{s} s = 1$ which implies $p \in \mathbb{1}$. We have shown $A \odot S \subset \mathbb{1}$.

For the converse direction let $p \in \mathbb{1}$ be positive, i.e. $0 < p < 1$. Let $r > 0$ and $r \in A$. There exists an $n \in \mathbb{N}$ such that $r_1 = n(1-p)r \in A$ and $r_2 = (n+1)(1-p)r \notin A$. Hence $r < r_2$. Then $r_2 - r_1 + pr_2 = (1-p)r + pr_2 < (1-p)r_2 + pr_2 = r_2 = (r_2 - r_1) + r_1$ (since $r \in A$, $r_2 \notin A$). Consequently $pr_2 < r_1$ and hence $r_2 < r_1/p$. Since $r_2 \notin A$ this means that $p/r_1 \in S$. Since $r_1 \in A$ and $p = r_1 \frac{p}{r_1}$ we get $p \in A \odot S$. This shows $\mathbb{1} \subset A \odot S$.

We have now shown $\mathbb{1} = A \odot S$, i.e. (mul-inv) holds for cuts in $\mathbb{R}^+$.

(mul-add) Inspect the definitions

$$A \odot (B \oplus C) = \{p \in \mathbb{Q} : p \leqslant r(s+t) : r \in A, s \in B, t \in C, r, s, t > 0\}$$

and

$$A \odot B \oplus A \odot C = \{p_1 + p_2 : p_1, p_2 \in \mathbb{Q},\ p_1 \leqslant r_1 s,\ p_2 \leqslant r_2 t,\ r_1, r_2, s, t > 0,\ r_1, r_2 \in A, s \in B, t \in C\}$$

and use the distributivity law for the rational numbers and the properties of cuts. □

We now drop the assumption of positivity, and our task is to check the multiplication axioms on all of $\mathbb{R}$. We frequently use

1.121. Observation. *For $P \in \mathbb{R}^+$ and $N \in \mathbb{R}^-$,*

$$N \odot P = \text{-}((\text{-}N) \odot P) = \text{-}(N \odot (\text{-}P)),$$
$$P \odot N = \text{-}(P \odot (\text{-}N)) = \text{-}((\text{-}P) \odot N).$$

This is a consequence of Definition 1.119 and $\text{-}(\text{-}E) = E$ for $E \in \mathbb{R}$.

1.122. Lemma.

(i) *For any $A, B \in \mathbb{R}$ the product $A \odot B$ is in $\mathbb{R}$.*
(ii) *For all $A, B \in \mathbb{R}$,*

$$\text{-}(A \odot B) = (\text{-}A) \odot B = A \odot (\text{-}B)$$

(iii) *For all* $A, B \in \mathbb{R}$,

$$A \odot B = (\text{-}A) \odot (\text{-}B)$$

(iv) *For all* $A, B, C \in \mathbb{R}$

$$\text{-}((A \odot B) \odot C) = ((\text{-}A) \odot B) \odot C = (A \odot (\text{-}B)) \odot C = (A \odot B) \odot (\text{-}C)$$
$$\text{-}(A \odot (B \odot C)) = ((\text{-}A) \odot (B \odot C)) = A \odot ((\text{-}B) \odot C) = A \odot (B \odot (\text{-}C))$$

(v) *For all* $A, B, C \in \mathbb{R}$,

$$A \odot (B \odot C) = A \odot ((\text{-}B) \odot (\text{-}C)) = (\text{-}A) \odot ((\text{-}B) \odot C) = (\text{-}A) \odot (B \odot (\text{-}C))$$
$$(A \odot B) \odot C = ((\text{-}A) \odot (\text{-}B)) \odot C = ((\text{-}A) \odot B) \odot (\text{-}C) = (A \odot (\text{-}B)) \odot (\text{-}C)$$

PROOF. It follows from Definition 1.119 and Lemma 1.120 that $A \odot B$ is a cut.

We first observe that if one of A, B is equal to O then all expressions in (ii) and (iii) are equal to O. In what follows assume that not both A, B are equal to O.

If $A \in \mathbb{R}^+$ and $B \in \mathbb{R}^-$ or if $A \in \mathbb{R}^-$ and $B \in \mathbb{R}^-$ then the identities in (ii) are already in Observation 1.121.

If $A \in \mathbb{R}^+$ and $B \in \mathbb{R}^+$ or if $A \in \mathbb{R}^-$ and $B \in \mathbb{R}^-$ then we already know (ii) for the pairs A, $\text{-}B$ and $\text{-}A, B$ (from Observation 1.121) and we get

$$\text{-}(A \odot (\text{-}B)) = (\text{-}A) \odot (\text{-}B) = A \odot (\text{--}B)$$
$$\text{-}((\text{-}A) \odot B) = (\text{--}A) \odot B = (\text{-}A) \odot (\text{-}B).$$

Since $B = \text{--}B$, $A = \text{--}A$, we obtain $A \odot B = \text{-}(A \odot (\text{-}B))$ and $A \odot B = \text{-}((\text{-}A) \odot B)$. Applying a negative yields (ii).

For part (iii) the claimed identity follows from Definition 1.119 if A, B are either both in $\mathbb{R}^+$ or both in $\mathbb{R}^-$. If $A \in \mathbb{R}^+$, $B \in \mathbb{R}^-$ or if $A \in \mathbb{R}^-$, $B \in \mathbb{R}^+$ then we use (ii) to get

$$A \odot B = \text{--}(A \odot B) = \text{-}((\text{-}A) \odot B) = \text{-}((\text{--}A) \odot (\text{-}B)) = \text{-}(A \odot (\text{-}B)) = (\text{-}A) \odot (\text{-}B)\,.$$

Repeated applications of parts (i), (ii), (iii) and $\text{-}(\text{-}E) = E$ allow to distribute one or two minus signs in any order and we get (iv), (v). The proofs of these identities follow the same patterns. As an illustration we check the identity $A \odot (B \odot C) = (\text{-}A) \odot (B \odot (\text{-}C))$:

$$\begin{aligned}(\text{-}A) \odot (B \odot (\text{-}C)) &= \text{-}\big(A \odot (B \odot (\text{-}C))\big) = A \odot \big(\text{-}(B \odot (\text{-}C))\big) \\ &= A \odot (B \odot (\text{--}C)) = A \odot (B \odot C).\end{aligned}$$

□

1.123. LEMMA. *The axioms (mul-comm), (mul-assoc), (mul-one), (mul-inv) hold for general cuts in* $\mathbb{R}$.

PROOF. Let $A, B, C \in \mathbb{R}$. We need to check the axioms for the cases that the cuts are not all in $\mathbb{R}^+$. Here we use Definition 1.119 in conjunction with Lemma 1.120 and the identity $\text{-}(\text{-}E) = E$ for $E \in \mathbb{R}$.

(mul-comm) First note that the claim holds by definition if one of A, B is equal to O. We verify

$$\begin{aligned}
A \odot B &= (\text{-}A) \odot (\text{-}B) = (\text{-}B) \odot (\text{-}A) = B \odot A, && \text{if } A \in \mathbb{R}^-, B \in \mathbb{R}^-\,, \\
A \odot B &= \text{-}\big((\text{-}A) \odot B\big) = \text{-}(B \odot (\text{-}A)) = B \odot A, && \text{if } A \in \mathbb{R}^-, B \in \mathbb{R}^+\,, \\
A \odot B &= \text{-}\big(A \odot (\text{-}B)\big) = \text{-}\big((\text{-}B) \odot A\big) = B \odot A, && \text{if } A \in \mathbb{R}^+, B \in \mathbb{R}^-\,.
\end{aligned}$$

(mul-assoc) Associativity follows from the definitions in the case that one of A, B, C is equal to 0. Excluding these cases, each of the sets A, B, C is either contained in $\mathbb{R}^+$ or in $\mathbb{R}^-$ leading us to distinguish 8 different cases. In the case that all sets A, B, C are in $\mathbb{R}^+$ we have already proved that (mul-assoc) holds in Lemma 1.120.

In each of the remaining seven cases we use (mul-assoc) for $\mathbb{R}^+$ in combination with Lemma 1.122 (v). Here are details for one of these cases with the other six cases being analogous: suppose A, B, C are all in $\mathbb{R}^-$. Then

$$\begin{aligned} A \odot (B \odot C) &= (\text{-}(\text{-}A)) \odot ((\text{-}B) \odot (\text{-}C)) = \text{-}\big((\text{-}A) \odot ((\text{-}B) \odot (\text{-}C))\big) \\ &= \text{-}\big((\text{-}A) \odot (\text{-}B)) \odot (\text{-}C)\big) = \text{-}\big((A \odot B) \odot (\text{-}C)\big) = (A \odot B) \odot C \end{aligned}$$

where in the middle identity we have used (mul-assoc) for $\mathbb{R}^+$.

(mul-one) This is a consequence of Definition 1.119: for $A \in \mathbb{R}^-$,

$$\mathbb{1} \odot A = \text{-}(\mathbb{1} \odot (\text{-}A)) = \text{-}(\text{-}A) = A.$$

(mul-inv) Define

$$A^{-1} = \text{-}(\text{-}A)^{-1} \quad \text{if } A \in \mathbb{R}^-.$$

Then for $A \in \mathbb{R}^-$ we get $A \odot A^{-1} = A \odot (\text{-}(\text{-}A)^{-1}) = (\text{-}A) \odot (\text{-}A)^{-1} = \mathbb{1}$ (using (mul-inv) for positive cuts). This proves (mul-inv) for general cuts. □

1.124. Lemma. *The distributivity law $A \odot (B \oplus C) = A \odot B \oplus A \odot C$ holds for general cuts $A, B, C \in \mathbb{R}$.*

Proof. The result is immediate if $A = O$ so let us assume $A \neq O$. We may use the already proved distributivity law for $A, B, C \in \mathbb{R}^+$.

(i) The case $A \in \mathbb{R}^+$. We need to distinguish several subcases.

Let $B \in \mathbb{R}^-$, $B \oplus C \in \mathbb{R}^+$. Since $A \odot (\text{-}B) = \text{-}(A \odot B)$ we obtain

$$A \odot C = A \odot (B \oplus C) \oplus A \odot (\text{-}B) = A \odot (B \oplus C) \oplus (\text{-}(A \odot B)).$$

By Lemma 1.118 and its consequence for cancellations (from Lemma 1.104) we get $A \odot C \oplus A \odot B = A \odot (B \oplus C)$.

Next assume $B \in \mathbb{R}^+$, $B \oplus C \in \mathbb{R}^-$ then $C \in \mathbb{R}^-$. Then

$$\text{-}(A \odot C) = A \odot (\text{-}C) = A \odot B \oplus A \odot (\text{-}(B \oplus C)) = A \odot B \oplus (\text{-}(A \odot (B \oplus C)))$$

and using cancellation as above we get $A \odot B \oplus A \odot C = A \odot (B \oplus C)$.

Now assume $B \in \mathbb{R}^+$ and $B \oplus C \in \mathbb{R}^+$. If $C \in \mathbb{R}^+$ we may apply the distributivity law for $\mathbb{R}^+$ entries, so let us assume that $C \in \mathbb{R}^-$. Then

$$A \odot B = A \odot (B \oplus C \oplus (\text{-}C)) = A \odot (B \oplus C) \oplus A \odot (\text{-}C) = A \odot (B \oplus C) \oplus (\text{-}(A \odot C))$$

and again, using cancellation we get $A \odot B \oplus A \odot C = A \odot (B \oplus C)$.

Last, assume $B \in \mathbb{R}^-$ and $B \oplus C \in \mathbb{R}^-$ which gives $\text{-}B \in \mathbb{R}^+$ and $(\text{-}B) \oplus (\text{-}C) \in \mathbb{R}^+$. In this instance we can apply the previous case to get

$$A \odot (B \oplus C) = \text{-}(A \odot ((\text{-}B) \oplus (\text{-}C))) = \text{-}\big(A \odot (\text{-}B) \oplus A \odot (\text{-}C)\big) = \text{-}(A \odot (\text{-}B)) \oplus (\text{-}(A \odot (\text{-}C))),$$

and by Lemma 1.122 this equals $A \odot B \oplus A \odot C$.

Finally, if $B \oplus C = 0$ then $A \odot (B \oplus C) = O$ and $A \odot B \oplus A \odot (\text{-}B) = A \odot B \oplus (\text{-}(A \odot B)) = O$ so the distributivity law holds in this case as well.

(ii) The case $A \in \mathbb{R}^-$. This follows from the previous case and Lemma 1.122:

$$\begin{aligned} A \odot (B \oplus C) &= \text{-}\big((\text{-}A) \odot (B \oplus C)\big) = \text{-}\big((\text{-}A) \odot B \oplus (\text{-}A) \odot C\big) \\ &= \text{-}\big((\text{-}A) \odot B\big) \oplus \big(\text{-}\big((\text{-}A) \odot C\big)\big) = A \odot B \oplus A \odot C. \end{aligned}$$

□

Finally, we observe that the order axioms (ord-add), (ord-mul) for $\mathbb{R}$ hold:

1.125. Lemma.

(ord-add) *If $A, B, C \in \mathbb{R}$ and $B \subseteq C$, then $A \oplus B \subseteq A \oplus C$.*
(ord-mul) *If $A, B \in \mathbb{R}^+ \cup \{O\}$, then $A \odot B \in \mathbb{R}^+ \cup \{O\}$.*

Both of these follow directly from the definitions of $A \oplus B$, $A \oplus C$, $A \odot B$.

We are now ready to finish the proof of the existence result in Theorem 1.30.

1.126. THEOREM. *The set $\mathbb{R}$ is a Dedekind complete ordered field which contains a subfield that is isomorphic to the rational numbers.*

PROOF. The combination of Lemmas 1.117, 1.118, 1.123, 1.124 and 1.125 proves that $\mathbb{R}$ is a Dedekind ordered field. For any rational number r the set

$$\imath(r) = \{p \in \mathbb{Q} : p < r\}$$

is a Dedekind cut, and $\imath : \mathbb{Q} \to \mathbb{R}$ is a well defined injective map. It satisfies for $p, q \in \mathbb{Q}$

$$\begin{aligned} &\imath(p+q) = \imath(p) \oplus \imath(q) \\ &\imath(pq) = \imath(p) \odot \imath(q) \\ &p \leqslant q \text{ if and only if } \imath(p) \subseteq \imath(q) \end{aligned}$$

which establishes the image $\imath(\mathbb{Q})$ as an ordered field. It is left to the reader to check these relations, using the definitions of cuts and the field operations on $\mathbb{Q}$ and $\mathbb{R}$. $\imath(\mathbb{Q})$ is the set of *rational cuts*. It is an ordered field. By Theorem 1.110 it is equal to the unique ordered subfield $\mathbb{R}_{\mathbb{Q}}$ of $\mathbb{R}$ which is isomorphic to $\mathbb{Q}$ as an ordered field. □

1.6.3. Uniqueness of the real numbers. Let $\mathbb{K}$ be a Dedekind complete ordered field, with the operations $(+, \cdot, \leq)$. Recall that Dedekind completeness means by definition that the least upper bound axiom is satisfied. We use the notation 0, 1 for the neutral elements for addition and multiplication, also using $-X$ for the negative of $X \in \mathbb{K}$. We keep using the notation of the previous section for the field $\mathbb{R}$ of Dedekind cuts, using $\oplus$, $\odot$, $\subseteq$, O, $\mathbb{1}$, - for addition, multiplication, order, zero element, one element, negative.

To complete the proof of Theorem 1.30 we show the existence of an isomorphism

$$F : \mathbb{R} \to \mathbb{K}.$$

The field $\mathbb{K}$ contains a unique subfield of rational numbers $\mathbb{K}_{\mathbb{Q}}$ which is isomorphic to $\mathbb{Q}$. Likewise the field $\mathbb{R}_{\mathbb{Q}}$ of rational Dedekind cuts is the unique ordered subfield of $\mathbb{R}$ that is isomorphic to $\mathbb{Q}$. Let $\phi_1 : \mathbb{Q} \to \mathbb{K}_{\mathbb{Q}}$ and $\phi_2 : \mathbb{Q} \to \mathbb{R}_{\mathbb{Q}}$ be the canonical isomorphisms from Theorem 1.110. Then

$$\varphi : \phi_1 \circ \phi_2^{-1} : \mathbb{R}_{\mathbb{Q}} \to \mathbb{K}_{\mathbb{Q}}$$

is an isomorphism of ordered fields.

1.127. THEOREM. *Let $\mathbb{K}$ be a Dedekind complete ordered field. Then*

(i) *$\mathbb{K}$ is Archimedean.*
(ii) *If $a, b \in \mathbb{K}$ such that $a \precnsim b$ then there exists $q \in \mathbb{K}_{\mathbb{Q}}$ such that $a \precnsim q \precnsim b$.*

PROOF. The implication (i) $\Longrightarrow$ (ii) has already been noted in Theorem 1.113. (i) is proved just as the corresponding result for the real numbers (Theorem 1.31). For the Archimedean property of $\mathbb{K}$, let $\phi : \mathbb{Q} \to \mathbb{K}_{\mathbb{Q}}$ be the canonical isomorphism. Note that $\phi(n)a$ equals the sum $a + \cdots + a$ of n copies of a. We show that for every b there exists $n \in \mathbb{N}$ such that $b \precnsim \phi(n)a$. If not the set $\{\phi(n)a : n \in \mathbb{N}\}$ has a least upper bound. Now use the argument in the proof of Theorem 1.31 to get a contradiction. □

For any cut $X \in \mathbb{R}$ let

$$I_X = \{\varphi(r) \in \mathbb{K} : r \in \mathbb{R}_{\mathbb{Q}}, r \subsetneq X\}$$

I_X is a nonempty set which is bounded above. Indeed if $q \in \mathbb{R}_{\mathbb{Q}}$ with $X \subseteq q$ then $\varphi(r) \preceq \varphi(q)$ for all $\varphi(r) \in I_X$. By the Dedekind completeness of $\mathbb{K}$, I_X has a least upper bound and we may define $F : \mathbb{R} \to \mathbb{K}$ by

$$F(X) = \sup I_X$$

where $\sup I_X$ is the least upper bound of I_X (with respect to the order $\preceq$). This is our candidate for an isomorphism. When restricted to $\mathbb{R}_\mathbb{Q}$ we need it to coincide with the unique isomorphism $\varphi : \mathbb{R}_\mathbb{Q} \to \mathbb{K}_\mathbb{Q}$.

1.128. LEMMA. *If $r \in \mathbb{R}_\mathbb{Q}$ (i.e a rational cut) then $F(r) = \varphi(r)$.*

PROOF. Let $s \in \mathbb{R}_\mathbb{Q}$ with $s \subsetneq r$ we see from Theorem 1.110 that $\varphi(s) \precnsim \varphi(r)$. This means every $s \in I_r$ satisfies $\varphi(s) \precnsim \varphi(r)$ and hence $\varphi(r) \preceq \sup I_r$.

It remains to rule out the possibility that $\sup I_r \precnsim \varphi(r)$. Let us assume this. By Theorem 1.113 there is $q \in \mathbb{K}_\mathbb{Q}$ such that $\sup I_r \precnsim q \precnsim \varphi(r)$. Since $\varphi : \mathbb{R}_\mathbb{Q} \to \mathbb{K}_\mathbb{Q}$ is an isomorphism of ordered fields we get $q = \varphi(r')$ with $r' \in \mathbb{R}_\mathbb{Q}$ and with $r' \subsetneq r$. This implies $\varphi(r') \in I_r$ which contradicts $\sup I_r \precnsim \varphi(r')$.

We conclude that $\varphi(r) = \sup I_r$, i.e. $\varphi(r) = F(r)$. □

The map F preserves the order:

1.129. LEMMA. *If $X, Y \in \mathbb{R}$ such that $X \subseteq Y$ then $F(X) \preceq F(Y)$.*

PROOF. This is a consequence of the definition of $\sup I_X$ and the inequality $\varphi(s) \leqslant \varphi(r)$ for $r, s \in \mathbb{R}_\mathbb{Q}$ with $s \subseteq r$. □

1.130. LEMMA. *$F : \mathbb{R} \to \mathbb{K}$ is a bijective map.*

PROOF. We first show that F is injective. Let $X, Y \in \mathbb{R}$ such that $X \neq Y$. Since X and Y are comparable we may assume $X \subsetneq Y$. Using Theorem 1.113 for $\mathbb{R}$ (which is Theorem 1.32) we find $s, r \in \mathbb{R}_\mathbb{Q}$ such that $X \subsetneq s \subsetneq r \subsetneq Y$. Since $\varphi : \mathbb{R}_\mathbb{Q} \to \mathbb{K}_\mathbb{Q}$ is an isomorphism of ordered fields we have (using Lemma 1.129)

$$F(X) \preceq F(s) = \varphi(s) \precnsim \varphi(r) = F(r) \preceq F(Y)$$

so $F(X) \neq F(Y)$ and we conclude that F is injective.

To show that F is surjective let $b \in \mathbb{K}$. Let

$$J_b = \{r \in \mathbb{R}_\mathbb{Q} : \varphi(r) \precnsim b\}$$

Clearly J_b is a nonempty subset of $\mathbb{R}_\mathbb{Q}$ which is bounded above. Let $X \in \mathbb{R}$ be the least upper bound of J_b. We show that $F(X) = b$.

To see this assume that $F(X) \precnsim b$. Then by Theorem 1.113 there is $r \in \mathbb{R}_\mathbb{Q}$ such that

$$F(X) \precnsim \varphi(r) = F(r) \precnsim b.$$

By the injectivity of F this means $X \subsetneq r$ and $r \in J_b$. But $X = \sup J_b$ and we get a contradiction.

Next assume that $b \precnsim F(X)$. Then by Theorem 1.113 there exists $r \in \mathbb{R}_\mathbb{Q}$ such that

$$b \precnsim \varphi(r) = F(r) \precnsim F(X).$$

This implies that $r \subsetneq X = \sup J_b$ and hence there is $s \in J_b$ with $r \subsetneq s \subsetneq X$. Consequently we get $\varphi(s) \precnsim b$ and $b \precnsim \varphi(s)$ which again gives a contradiction.

Since the possibilities $F(X) \precnsim b$, $b \precnsim F(X)$ both result in contradictions we must have $F(X) = b$. We conclude that F is surjective. □

1.131. LEMMA. *For $X, Y \in \mathbb{R}$ we have*

$$F(X \oplus Y) = F(X) + F(Y), \qquad F(X \odot Y) = F(X) \cdot F(Y).$$

PROOF. We prove the equation $F(X \oplus Y) = F(X) + F(Y)$ by ruling out the possibilities $F(X) + F(Y) \precnsim F(X \oplus Y)$ and $F(X \oplus Y) \precnsim F(X) + F(Y)$, and in doing so we use that the relation $F(r \oplus s) = F(r) + F(s)$ is true for $r, s \in \mathbb{R}_\mathbb{Q}$, by Lemma 1.128.

Assume $F(X) + F(Y) \precnsim F(X \oplus Y)$. By Theorem 1.113 there is $r \in \mathbb{R}_\mathbb{Q}$ such that

$$F(X) + F(Y) \precnsim F(r) \precnsim F(X \oplus Y)$$

By Lemmas 1.129 and 1.130 this implies $r \subsetneq X \oplus Y$. We can write r as the sum of two rational cuts $r = r_1 \oplus r_2$ such that $r_1 \subsetneq X$, $r_2 \subsetneq Y$. Thus

$$F(r) = \varphi(r) = \varphi(r_1) + \varphi(r_2) = F(r_1) + F(r_2) \precsim F(X) + F(Y)$$

which is a contradiction to $F(X) + F(Y) \precsim F(r)$ in the previous display.

The other case is handled similarly: Assume $F(X \oplus Y) \precsim F(X) + F(Y)$. Find $r \in \mathbb{R}_{\mathbb{Q}}$ such that

$$F(X \oplus Y) \precsim F(r) \precsim F(X) + F(Y).$$

Thus $X \oplus Y \subsetneq r$ and therefore we can split $r = r_1 \oplus r_2$ with $r_1, r_2 \in \mathbb{R}_{\mathbb{Q}}$ such that $X \subsetneq r_1$, $Y \subsetneq r_2$. Hence

$$F(X) + F(Y) \precsim F(r_1) + F(r_2) = \varphi(r_1) + \varphi(r_2) = \varphi(r) = F(r)$$

which is a contradiction to $F(r) \precsim F(X) + F(Y)$.

Combining the two cases that both result in a contradiction we deduce that we must have $F(X) + F(Y) = F(X \oplus Y)$.

The multiplication law is proved in a similar fashion. The same type of argument as used for addition yields $F(X \odot Y) = F(X) \cdot F(Y)$ for positive $X, Y \in \mathbb{R}$. In the general case we infer from the addition law

$$0 = F(W \oplus (\text{-}W)) = F(W) + F(\text{-}W)$$

and hence $F(\text{-}W) = -F(W)$ for all $W \in \mathbb{R}$. With this it is straightforward to check the multiplication law for all cases where X, Y are not both positive. □

1.132. Corollary. *All Dedekind complete ordered fields are isomorphic.*

Proof. If $\mathbb{K}$ is any Dedekind complete ordered field, then it is isomorphic to $\mathbb{R}$ by Lemmas 1.129, 1.130, 1.131. □

1.6.4. Equivalent notions of completeness. In this section we examine the relationship of Dedekind completeness in an ordered field and the notion of completeness in a metric space (where completeness means that all Cauchy sequences converge; see Ch. 4).

In an ordered field we may work with a notion of absolute value.

1.133. Definition. Let $\mathbb{K}$ be an ordered field. The *absolute value* of $x \in \mathbb{K}$ is defined by

$$|x| = \begin{cases} x & \text{if } x \succeq 0\,, \\ -x & \text{if } x \precsim 0\,. \end{cases}$$

1.134. Lemma. *Let $\mathbb{K}$ be an ordered field. Then the following properties hold for $x, y \in \mathbb{K}$.*

(i) $|x| = |-x| \succeq 0$, *and* $|x| = 0$ *if and only if* $x = 0$
(ii) $x \preceq |x|$, $-x \preceq |x|$.
(iii) $|xy| = |x||y|$.
(iv) $|x + y| \preceq |x| + |y|$ *(triangle inequality).*
(v) $\big||x| - |y|\big| \preceq |x - y|$.

The proofs of (i)-(iii) are left to the reader.

Proof of (iv) and (v). (iv) follows from $x + y \preceq |x| + |y|$ and $-x - y \preceq |-x| + |-y| = |x| + |y|$. For (v) apply (iv) to get

$$|x| = |x - y + y| \preceq |x - y| + |y|$$

and thus $|x| - |y| \preceq |x - y|$. Switching x and y gives $|y| - |x| \preceq |x - y|$ and combining the two gives (v). □

With this notion we can define Cauchy sequences and limits, by taking the corresponding definitions on the real numbers.

1.135. Definition. Let $\mathbb{K}$ be an ordered field.

(i) A sequence $(x_n)_{n=1}^{\infty}$ with $n \in \mathbb{N}$ is a *Cauchy sequence* if for every positive $\varepsilon \in \mathbb{K}$ there exists $N \in \mathbb{N}$ such that $|x_n - x_m| \preceq \varepsilon$ for all $m, n \geqslant N$.

(ii) (x_n) converges to $x \in \mathbb{K}$ if for every positive $\varepsilon \in \mathbb{K}$ there exists $N \in \mathbb{N}$ such that $|x_n - x| \preceq \varepsilon$ for all $n \geqslant N$.

(iii) $\mathbb{K}$ is *Cauchy complete* if every Cauchy sequence converges to a limit.

1.136. Observation. *Let $(\mathbb{K}, +, \cdot, \preceq)$ be an Archimedean ordered field and $\phi : \mathbb{Q} \mapsto \mathbb{K}_{\mathbb{Q}}$ be the canonical isomorphism.*

(i) *Suppose $|x| \preceq \phi(\frac{1}{n})$ for all $n \in \mathbb{N}$ then $x = 0$.*

(ii) *Let x_n be convergent to x and to $\tilde{x}$. Then $x = \tilde{x}$.*

Note that (i) is an immediate consequence of the archimedean property and (ii) follows by the definition of the limit and the triangle inequality for $|\cdot|$.

1.137. Theorem. *Let $\mathbb{K}$ be an ordered field. Then $\mathbb{K}$ is Dedekind complete if and only if it has the Archimedean property and is Cauchy complete.*

Proof. If $\mathbb{K}$ is Dedekind complete, then it is Archimedean by Theorem 1.127. The fact that Cauchy sequences converge is proved using just notational modifications of the arguments in §1.3.5. We thus focus on the converse. We let as usual $\phi : \mathbb{Q} \to \mathbb{K}_{\mathbb{Q}}$ be the canonical isomorphism.

Let $\mathbb{K}$ be an Archimedean ordered field which is Cauchy complete. Let E be a subset of $\mathbb{K}$ which is bounded above and let M be an upper bound. Let $a \in E$. By the Archimedean property there are, for each n, integers K_n and L_n such that $\phi(\frac{K_n}{n}) \precneqq a$ and $M \precneqq \phi(\frac{L_n}{n})$. For each n we may choose an integer $j(n)$ such that $K_n < j(n) \leqslant L_n$ such that $\phi(\frac{j(n)}{n})$ is an upper bound for E but $\phi(\frac{j(n)-1}{n})$ is not an upper bound. We then have

$$\phi(\tfrac{j(n')-1}{n'}) \preceq \phi(\tfrac{j(n)}{n})$$
$$\phi(\tfrac{j(n)-1}{n}) \preceq \phi(\tfrac{j(n')}{n'})$$

which leads to

$$\phi(-\tfrac{1}{n'}) \leqslant \phi(\tfrac{j(n)}{n}) - \phi(\tfrac{j(n')}{n'}) \leqslant \phi(\tfrac{1}{n})$$

If $n, n' \geqslant N$ this implies

$$|\phi(\tfrac{j(n)}{n}) - \phi(\tfrac{j(n')}{n'})| \leqslant \phi(\tfrac{1}{N})$$

Hence if we set $u_n = \phi(\frac{j(n)}{n})$ then u_n is a Cauchy sequence, which by assumption converges to a unique $u \in \mathbb{K}$. Since u_n is an upper bound for E, we have $x \preceq u_n$ for each $x \in E$ and for each n. It is easy to see $x \preceq u$ for all $x \in E$ (since $u \precneqq x$ would imply $u_n \precneqq x$ for large n). Thus u is an upper bound for E. Now let $u' \precneqq u$. For every n there is $a(n) \in E$ such that $a(n) \succeq \phi(\frac{j(n)}{n}) - \phi(\frac{1}{n}) = u_n - \phi(\frac{1}{n})$; thus $a(n) \succeq u - \phi(\frac{1}{n})$. From the Archimedean property we see that $\lim_{n\to\infty} \phi(\frac{1}{n}) = 0$ which implies $a(n) \succneqq u'$ for large n. Hence u' cannot be an upper bound of E and it follows that u is the desired least upper bound of E. We have thus shown that any Cauchy complete field with the Archimedean property is Dedekind complete. □

The absolute value defines a metric on $\mathbb{Q}$, and one should view $\mathbb{R}$ as the completion of $\mathbb{Q}$ under this metric. We emphasize that we cannot use Theorem 4.158 about completion of metric spaces to construct the real numbers (since the existence of the real numbers was used in the proof of this theorem, and in fact is used in the definition of metric). Nevertheless, given Theorem 1.137 it is natural to construct the real numbers following the same approach, using equivalence classes of Cauchy sequences in $\mathbb{Q}$. We say that two Cauchy sequences (a_n), (b_n) of rational numbers are equivalent if $\lim_{n\to\infty}(a_n - b_n) = 0$. In that case we write $(a_n) \sim (b_n)$. One checks that $\sim$ defines an equivalence relation on the set of Cauchy sequences in $\mathbb{Q}$. We denote the set of equivalence classes by $\mathcal{R}$. The goal is to identify $\mathcal{R}$ with the real

numbers. We use the notation $a = [(a_n)]$ for the equivalence class of the Cauchy sequence $a = (a_n)$. This approach is outlined in the following exercise (for more details see for example [**79**, Ch. 5]).

1.138. Exercise.

(i) Define addition and multiplication on $\mathcal{R}$ by $a + b = [(a_n + b_n)]$, $ab = [(a_n b_n)]$. Show that these operations are well defined as operations on $\mathcal{R}$ and that $\mathcal{R}$ becomes a field.

(ii) An equivalence class $a = [(a_n)]$ of Cauchy sequences in $\mathbb{Q}$ is considered nonnegative if there exists a representative (a_n) with the following property: For every $M \in \mathbb{N}$ there exists $N_M \in \mathbb{N}$ such that $a_n \geqslant -1/M$ for $n \geqslant N_M$. Show that all representatives in the same equivalence class have the same property. Define a partial order on $\mathcal{R}$ by saying that $b \preceq a$ if $[(a_n - b_n)]$ is nonnegative. Prove that this defines a total order on $\mathcal{R}$ and that $\mathcal{R}(+, \cdot, \preceq)$ becomes an ordered field.

(iii) Prove that $\mathcal{R}$ satisfies the Archimedean property.

(iv) Prove that $\mathcal{R}$ is Cauchy complete, and that $\mathcal{R}$ and $\mathbb{R}$ are isomorphic as ordered fields.

CHAPTER 2

Single variable calculus

2.1. Integration

In this section we introduce the Riemann integral for real-valued functions on a closed interval $[a, b]$. The Riemann integral is simple and intuitive, but unfortunately has certain deficiencies; however, it will be sufficient for this text. A better alternative is the Lebesgue integral, which we do not cover in this book.

2.1.1. Upper and lower sums.

2.1. Definition.

(i) A *partition* $P = \{x_0, \dots, x_n\}$ of $[a, b]$ is a finite subset of $[a, b]$ which includes the points a and b and is ordered in the following way:

$$a = x_0 < ... < x_i < x_{i+1} < ... < x_n = b.$$

(ii) If P, P' are partitions of $[a, b]$ with $P \subset P'$, then P' is called a *refinement* of P.

2.2. Definition. Given a partition $P = \{a = x_0 < \cdots < x_n = b\}$ of $[a, b]$ and a bounded function $f : [a, b] \to \mathbb{R}$ define

$$m_i(f) = \inf_{t \in [x_{i-1}, x_i]} f(t), \quad M_i(f) = \sup_{t \in [x_{i-1}, x_i]} f(t).$$

(i) The expression

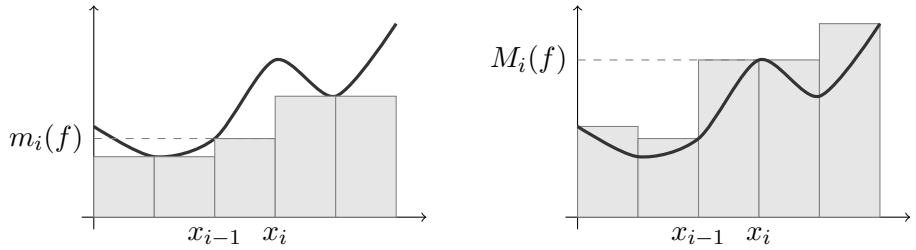
$$L(f, P) = \sum_{i=1}^{n} m_i(f)(x_i - x_{i-1})$$

is called the *lower sum* of f with respect to the partition P.

(ii) The expression

$$U(f, P) = \sum_{i=1}^{n} M_i(f)(x_i - x_{i-1})$$

is called the *upper sum* of f with respect to the partition P.



Lower sum $L(f, P)$ Upper sum $U(f, P)$

Figure 2.1. Lower and upper Riemann sums can each be visualized as the sum of areas of rectangles.

2.3. Observation. *Let P, P' be partitions of $[a,b]$, let $f:[a,b]\to\mathbb{R}$ be bounded, and let P' be a refinement of P. Then*

$$(b-a)\inf_{[a,b]} f \leqslant L(f,P)\leqslant L(f,P')\leqslant U(f,P')\leqslant U(f,P)\leqslant (b-a)\sup_{[a,b]} f.$$

2.4. Corollary. *Let P_1, P_2 be partitions of $[a,b]$. Then $L(f,P_1)\leqslant U(f,P_2)$.*

2.5. Definition. Let $f:[a,b]\to\mathbb{R}$ be a bounded function. The numbers

$$\underline{\mathcal{I}}_a^b(f)=\sup_P L(f,P),\qquad \overline{\mathcal{I}}_a^b(f)=\inf_P U(f,P)$$

are called the *lower and upper Riemann–Darboux integrals of f* on the interval $[a,b]$, respectively. Here the sup and inf are taken over all partitions of $[a,b]$.

2.6. Observation. *Let $f:[a,b]\to\mathbb{R}$ be bounded. Then*

$$(b-a)\inf_{[a,b]} f\leqslant \underline{\mathcal{I}}_a^b(f)\leqslant \overline{\mathcal{I}}_a^b(f)\leqslant (b-a)\sup_{[a,b]} f.$$

2.7. Definition.

(i) A bounded function $f:[a,b]\to\mathbb{R}$ is called *Riemann integrable* if $\underline{\mathcal{I}}_a^b(f)=\overline{\mathcal{I}}_a^b(f)$.

(ii) If f is Riemann integrable, the number $\underline{\mathcal{I}}_a^b(f)=\overline{\mathcal{I}}_a^b(f)$ is called the *Riemann integral* of f, denoted by $\int_{[a,b]} f$ or $\int_a^b f$, or $\int_a^b f(x)\,dx$ or even $\int_a^b f(\text{☞})\,d\text{☞}$.

It is also convenient to define

$$\int_b^a f(x)\,dx=-\int_a^b f(x)\,dx. \tag{2.1}$$

The symbol $\int_a^b f(x)\,dx$ for arbitrary a,b (i.e. not necessarily $a\leqslant b$) is sometimes also referred to as a *directed integral*.

2.8. Lemma. *Let $f:[a,b]\to\mathbb{R}$ be a bounded function. Then f is Riemann integrable if and only if for every $\varepsilon>0$ there is a partition P of $[a,b]$ such that*

$$U(f,P)-L(f,P)<\varepsilon.$$

Proof. Suppose f is Riemann integrable. Then there are partitions P_1, P_2 of $[a,b]$ such that $L(f,P_1)\geqslant \int_a^b f-\frac{\varepsilon}{2}$ and $U(f,P_2)\leqslant \int_a^b f+\frac{\varepsilon}{2}$ and thus

$$U(f,P_2)-L(f,P_1)<\varepsilon.$$

Let P be the refinement $P_1\cup P_2$. Then

$$U(f,P_2)\geqslant U(f,P)\geqslant L(f,P)\geqslant L(f,P_1)$$

and hence $U(f,P)-L(f,P)<\varepsilon$.

Conversely, assume that for every $\varepsilon>0$ there is a partition P_ε of $[a,b]$ such that

$$U(f,P_\varepsilon)-L(f,P_\varepsilon)<\varepsilon.$$

Then $\overline{\mathcal{I}}_a^b(f)-\underline{\mathcal{I}}_a^b(f)\leqslant U(f,P_\varepsilon)-L(f,P_\varepsilon)<\varepsilon$, and since ε was arbitrary we conclude $\overline{\mathcal{I}}_a^b(f)=\underline{\mathcal{I}}_a^b(f)$. Hence f is Riemann integrable. □

2.9. Exercise. A bounded function $s:[a,b]\to\mathbb{R}$ is called a *step function* on $[a,b]$ if there is a partition $P=\{x_0=a<x_1<\cdots<x_N=b\}$ and $a_\nu\in\mathbb{R}$ such that $s(x)=a_\nu$ for $x_{\nu-1}<x<x_\nu$, $\nu=1,\dots,N$. Prove that s is Riemann integrable on $[a,b]$ and that

$$\int_a^b s(x)dx=\sum_{\nu=1}^N a_\nu(x_\nu-x_{\nu-1})\,.$$

Note that we allow s to take any values at the partition nodes x_ν.

2.1.2. Properties of the Riemann integral.

2.10. THEOREM. *Let $f : [a, b] \to \mathbb{R}$ be Riemann integrable and let $a < c < b$. Then f is also Riemann integrable on $[a, c]$, $[c, b]$, respectively, and*

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

PROOF. By passing to a refinement we may always assume that a partition P of $[a, b]$ contains the point c. Then $P = P_1 \cup P_2$ so that P_1 is a partition of $[a, c]$, P_2 is a partition of $[c, b]$. Let f_1 be the restriction of f to $[a, c]$ and f_2 be the restriction of f to $[c, b]$. Then

$$L_P(f) = L_{P_1}(f_1) + L_{P_2}(f_2), \qquad U_P(f) = U_{P_1}(f_1) + U_{P_2}(f_2),$$

and we have

$$\max_{i=1,2} U_{P_i}(f_i) - L_{P_i}(f_i) \leqslant \sum_{i=1,2} U_{P_i}(f_i) - L_{P_i}(f_i) = U_P(f) - L_P(f).$$

From Lemma 2.8 one deduces the Riemann integrability of f_1 and f_2 and the formula $\int_a^b f = \int_a^c f_1 + \int_c^b f_2$ follows from the corresponding formulas for the lower and upper sums. □

2.11. THEOREM. *If $f : [a, b] \to \mathbb{R}$ is continuous, then f is Riemann integrable.*

PROOF. We will use that f is uniformly continuous on $[a, b]$, by Theorem 1.92. Given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|f(x) - f(\tilde{x})| < \tfrac{\varepsilon}{b-a}$$

provided that $|x - \tilde{x}| < \delta$. Let N be such that $(b - a)/N < \delta$ and choose the partition

$$x_j = a + j\tfrac{b-a}{N}, \quad j = 0, \dots, N.$$

Let $I_j = [x_{j-1}, x_j]$ for $j = 1, \dots, N$. Then

$$M_i(f) - m_i(f) = \sup_{I_i} f - \inf_{I_i} f < \tfrac{\varepsilon}{b-a}$$

for $i = 1, \dots, N$ so that

$$\begin{aligned} U(f, P) - L(f, P) &= \sum_{i=1}^N M_i(f)(x_i - x_{i-1}) - \sum_{i=1}^N m_i(f)(x_i - x_{i-1}) \\ &= \sum_{i=1}^N (M_i(f) - m_i(f))(x_i - x_{i-1}) \leqslant \sum_{i=1}^N \tfrac{\varepsilon}{b-a}(x_i - x_{i-1}) = \varepsilon. \end{aligned}$$

By Lemma 2.8, f is Riemann integrable. □

2.12. EXERCISE. For $x \in [0, 1]$ let

$$f(x) = \begin{cases} 1 + x^2 & \text{if } x \text{ is not of the form } 2^{-k} \text{ for some } k > 0 \\ 3 & \text{if } x \text{ is of the form } 2^{-k} \text{ for some } k > 0 \end{cases}$$

Show that f is Riemann integrable on $[0, 1]$ and determine $\int_0^1 f$.

2.13. EXAMPLE. Let $f : [0, 1] \to \{0, 1\}$ be the *Dirichlet function*

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

Note that for all partitions P of $[0, 1]$ we have $U(f, P) = 1$ and $L(f, P) = 0$ so that $\overline{\mathcal{I}}_0^1(f) = 1$ and $\underline{\mathcal{I}}_0^1(f) = 0$ and hence f is not Riemann integrable on $[0, 1]$.

2.14. Exercise. Let $f : [0,1] \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} x^2 & \text{if } x \in \mathbb{Q} \\ x & \text{if } x \in [0,1]\backslash\mathbb{Q}. \end{cases}$$

Determine $\overline{\mathcal{I}}_0^1(f)$ and $\underline{\mathcal{I}}_0^1(f)$.

2.15. Theorem. *Let f and g be Riemann integrable functions on $[a,b]$. Then $f+g$ is Riemann integrable and*

$$\int_a^b (f+g) = \int_a^b f + \int_a^b g.$$

Moreover, for $c \in \mathbb{R}$,

$$\int_a^b c\,f = c\int_a^b f.$$

Proof. Let $P = \{x_0 < \cdots < x_N\}$ be a partition of $[a,b]$. Then

$$m_i(f) + m_i(g) \leqslant m_i(f+g), \qquad M_i(f+g) \leqslant M_i(f) + M_i(g).$$

Consequently,

$$L(f,P) + L(g,P) \leqslant L(f+g,P) \leqslant U(f+g,P) \leqslant U(f,P) + U(g,P).$$

Now let $\varepsilon > 0$. Since f and g are Riemann integrable there exist partitions P_f and P_g such that

$$U(f,P_f) - L(f,P_f) < \tfrac{1}{2}\varepsilon, \quad U(g,P_g) - L(g,P_g) < \tfrac{1}{2}\varepsilon,$$

and these inequalities remain true if P_f, P_g are replaced by a refinement P that contains both P_f and P_g. We then get $U(f,P) + U(g,P) - L(f,P) - L(g,P) < \varepsilon$ and consequently

$$U(f+g,P) - L(f+g,P) < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, $f+g$ is Riemann integrable. Next,

$$L(f,P) + L(g,P) \leqslant L(f+g,P) \leqslant \int_a^b (f+g) \leqslant U(f+g,P) \leqslant U(f,P) + U(g,P)$$

and

$$L(f,P) + L(g,P) \leqslant \int_a^b f + \int_a^b g \leqslant U(f,P) + U(g,P).$$

Thus if we pick a partition P with $U(f,P) - L(f,P) < \varepsilon$ and $U(g,P) - L(g,P) < \varepsilon$ we also obtain $|\int_a^b (f+g) - \int_a^b f - \int_a^b g| < 2\varepsilon$ and since ε was arbitrary this yields $\int_a^b (f+g) = \int_a^b f + \int_a^b g$.

For the second assertion note that

$$U(c\,f,P) = c\,U(f,P) \text{ and } L(c\,f,P) = c\,L(f,P) \text{ if } c \geqslant 0,$$
$$U(c\,f,P) = c\,L(f,P) \text{ and } L(c\,f,P) = c\,U(f,P) \text{ if } c \leqslant 0,$$

which gives $\int_a^b c\,f = c\int_a^b f$. □

2.16. Theorem. *Let f and g be Riemann integrable on $[a,b]$. Then*

(i) $|f|$ is Riemann integrable and

$$\Big|\int_a^b f\Big| \leqslant \int_a^b |f|.$$

(ii) The functions $\min(f,g)$, $\max(f,g)$ are Riemann integrable.

PROOF. Let $P = \{x_0 < \cdots < x_N\}$ be a partition of $[a,b]$. Then one checks

$$M_i(|f|) - m_i(|f|) \leqslant M_i(f) - m_i(f)$$

and therefore

$$U(|f|,P) - L(|f|,P) \leqslant U(f,P) - L(f,P).$$

Hence $|f|$ is Riemann integrable if f is. Also,

$$|U(f,P)| \leqslant U(|f|,P)$$

which, using sufficiently fine partitions leads to $|\int_a^b f| \leqslant \int_a^b |f| + \varepsilon$ for all $\varepsilon > 0$ and thus $|\int_a^b f| \leqslant \int_a^b |f|$.

For part (ii) use

$$\min(f(x),g(x)) = \tfrac{1}{2}\big(f(x) + g(x) - |f(x) - g(x)|\big),$$
$$\max(f(x),g(x)) = \tfrac{1}{2}\big(f(x) + g(x) + |f(x) - g(x)|\big).$$ □

2.17. COROLLARY. *f is Riemann integrable if and only if $f^+ = \max(f,0)$ and $f^- = -\min(f,0)$ are Riemann integrable.*

PROOF. Write $f = f^+ - f^-$ and use part (ii) of Theorem 2.16. □

2.18. THEOREM. *Let f and g be Riemann integrable on $[a,b]$. Then the pointwise product $f \cdot g$ is Riemann integrable on $[a,b]$.*

PROOF. Assume first that f and g are nonnegative and $f \leqslant C$, $g \leqslant C$ on $[a,b]$ for a constant $C \in (0,\infty)$. Then if $P = \{a = x_0 < \cdots < x_N = b\}$ is a partition of $[a,b]$ we get $M_i(f \cdot g) \leqslant M_i(f)M_i(g)$ and $m_i(f \cdot g) \geqslant m_i(f)m_i(g)$. From this,

$$U(fg,P) - L(fg,P) \leqslant \sum_{i=1}^{N} (M_i(f)M_i(g) - m_i(f)m_i(g))(x_i - x_{i-1})$$

and writing $M_i(f)M_i(g) - m_i(f)m_i(g) = (M_i(f) - m_i(f))M_i(g) + (M_i(g) - m_i(g))m_i(f)$ we obtain

$$U(fg,P) - L(fg,P) \leqslant C\sum_{i=1}^{N} (M_i(f) - m_i(f))(x_i - x_{i-1}) + C\sum_{i=1}^{N} (M_i(g) - m_i(g))(x_i - x_{i-1}).$$

Hence

$$U(fg,P) - L(fg,P) \leqslant C(U(f,P) - L(f,P)) + C(U(g,P) - L(g,P))$$

and it follows that $f \cdot g$ is Riemann integrable if f and g are.

To remove the assumption that both f, g are nonnegative, write

$$fg = (f^+ - f^-)(g^+ - g^-) = f^+g^+ - f^-g^+ - f^+g^- + f^-g^-$$

and apply the result for nonnegative functions in conjunction with the linearity of the Riemann integral (Theorem 2.15). □

2.19. THEOREM.

(i) If f, g are Riemann integrable on $[a,b]$ and $f(x) \leqslant g(x)$ for all $x \in [a,b]$, then

$$\int_a^b f \leqslant \int_a^b g.$$

(ii) If f is Riemann integrable on $[a,b]$, then

$$\Big|\int_a^b f\Big| \leqslant (b-a) \sup_{x\in[a,b]} |f(x)|. \tag{2.2}$$

PROOF. Part (i) is a consequence of corresponding estimates for lower and upper sums; we have $L(f,P) \leqslant L(g,P)$ and $U(f,P) \leqslant U(g,P)$. For part (ii) we apply part (i) to the inequality $-\sup_{[a,b]}|f| \leqslant f(x) \leqslant \sup_{[a,b]}|f|$ which holds for all $x \in [a,b]$. □

As a consequence we obtain *mean value theorems for integrals.*

2.20. COROLLARY. *Let $f : [a,b] \to \mathbb{R}$ be continuous on $[a,b]$ and let $w : [a,b] \to \mathbb{R}$ be Riemann integrable on $[a,b]$ with $w(x) \geqslant 0$ for all $x \in [a,b]$. Then there exists $\xi \in (a,b)$ such that*

$$\int_a^b f(x)w(x)\,dx = f(\xi)\int_a^b w(x)\,dx.$$

In particular

$$\frac{1}{b-a}\int_a^b f(x)\,dx = f(\eta)$$

for some $\eta \in [a,b]$.

PROOF. Let $m = \min_{x\in[a,b]} f(x)$, and $M = \max_{x\in[a,b]} f(x)$. From Theorem 2.19 we have

$$m\int_a^b w(x)\,dx \leqslant \int_a^b f(x)w(x)\,dx \leqslant M\int_a^b w(x)\,dx$$

We let $x_1, x_2 \in [a,b]$, $x_1 \neq x_2$ be such that $f(x_1) = m$ and $f(x_2) = M$ and let $I \subset [a,b]$ be the closed interval with endpoints x_1 and x_2. We apply the intermediate value theorem (Theorem 1.88) to the continuous function $x \mapsto f(x)\int_a^b w(t)dt$ on I to see that there exists ξ in the interior of I such that $f(\xi)\int_a^b w(t)dt = \int_a^b f(t)w(t)dt$. The second claim follows if we put $w = 1$. □

2.21. EXERCISE (Riemann sum characterization). Show that a function $f : [a,b] \to \mathbb{R}$ is Riemann integrable if and only if there exists $I \in \mathbb{R}$ such that for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every partition $\{a = x_0 < \cdots < x_n = b\}$ of $[a,b]$ with $\max_i x_i - x_{i-1} < \delta$ and all choices of c_i with $x_{i-1} \leqslant c_i \leqslant x_i$ it holds that

$$\Big|\sum_{i=1}^n f(c_i)(x_i - x_{i-1}) - I\Big| < \varepsilon.$$

Show that in this case, $I = \int_a^b f$.

REMARK. The quantity $\sum_{i=1}^n f(c_i)(x_i - x_{i-1})$ is called a *Riemann sum.* This characterization has the benefit that it allows a Riemann integral to be defined for vector-valued functions (see Exercises 5.124, 5.125).

2.22. EXERCISE. Let $r \geqslant 0$. Compute $\int_0^1 x^r\,dx$ by considering lower and upper sums for the partitions with nodes $x_0 = 0$, $x_j = \alpha^{N-j}$, for $1 \leqslant j \leqslant N$, $\alpha < 1$. *Note:* Do not use the fundamental theorem of calculus.

2.23. EXERCISE. Let f be Riemann integrable on $[a,b]$. Prove that $\lim_{\varepsilon\to 0+}\int_{a+\varepsilon}^b f = \int_a^b f$.

2.24. EXERCISE. Let $f : [a,b] \to \mathbb{R}$ be bounded and $M = \sup_{x\in[a,b]}|f(x)|$. Assume f is Riemann integrable and let $g : [-M,M] \to \mathbb{R}$ be a continuous function. Show that $g \circ f$ is Riemann integrable on $[a,b]$.

The following approximation result lets us approximate Riemann integrable functions by continuous functions using integral means; it will be useful in Chapter 7.

2.25. THEOREM. *Let $f : [a,b] \to \mathbb{R}$ be Riemann integrable, and let $\varepsilon > 0$, and $1 \leqslant q < \infty$. There exists a continuous function $h : [a,b] \to \mathbb{R}$ such that*

(i) $h(a) = h(b) = 0$,

(ii) $\max_{x\in[a,b]}|h(x)| \leqslant \sup_{x\in[a,b]}|f(x)|$, and

(iii) $\int_a^b |f(x)-h(x)|^q\,dx < \varepsilon^q$.

PROOF. We first consider the case $q=1$.

Claim 1. If s is a step function associated to a partition P with partition points $x_0,\dots,x_N$, as in Exercise 2.9, then for every $\varepsilon>0$ there exists a continuous function g such that $\int_a^b |g-s|\,dx<\varepsilon$.

Indeed, let $s(x)=a_\nu$ on the open interval $(x_{\nu-1},x_\nu)$. Let $\delta<\frac14\min_\nu(x_\nu-x_{\nu-1})$ and $M=\max_\nu|a_\nu|$. Let $\varepsilon>0$. Define a trapezoidal function g_ν on $[x_{\nu-1},x_\nu]$ by

$$g_\nu(x)=\begin{cases} a_\nu\delta^{-1}(x-x_{\nu-1}) & \text{if } x_{\nu-1}\leqslant x\leqslant x_{\nu-1}+\delta\\ a_\nu & \text{if } x_{\nu-1}+\delta<x<x_\nu-\delta\\ a_\nu\delta^{-1}(x_\nu-x) & \text{if } x_\nu-\delta\leqslant x\leqslant x_\nu\end{cases}$$

and set $g_\nu(x)=0$ for $x<x_{\nu-1}$ and $x>x_\nu$. Note that $g_\nu\in C([a,b])$ and $g_\nu(x_{\nu-1})=g_\nu(x_\nu)=0$ and thus the sum $g=\sum_{\nu=1}^N g_\nu$ belongs to $C([a,b])$ and satisfies $g(a)=g(b)=0$. Moreover, $s-g=0$ on the union of the intervals $[x_{\nu-1}+\delta,x_\nu-\delta]$. Hence

$$\int_a^b|s-g|\,dx=\sum_{\nu=1}^N\Big(\int_{x_{\nu-1}}^{x_{\nu-1}+\delta}|s-g|\,dx+\int_{x_\nu-\delta}^{x_\nu}|s-g|\,dx\Big)\leqslant\sum_{\nu=1}^N 2\delta|a_\nu|\leqslant 2\delta NM$$

and if we choose $\delta<\frac{\varepsilon}{2NM}$, then $\int_a^b|s-g|\,dx<\varepsilon$. This proves Claim 1.

Now let f be Riemann integrable on $[a,b]$. Let $M=\sup_{[a,b]}|f(x)|$. Choose a partition P such that

$$U(f,P)-L(f,P)<\frac{\varepsilon}{2}.$$

Define a step function

$$s(x)=\inf_{t\in[x_{\nu-1},x_\nu]}f(t)\quad\text{for } x_{\nu-1}<x<x_\nu.$$

Then $s(x)\leqslant f(x)$ and $|s(x)|\leqslant M$ for all x. Moreover, $\int_a^b s(x)\,dx=L(f,P)$. Hence

$$\begin{aligned}\int_a^b|f(x)-s(x)|\,dx=\int_a^b\big(f(x)-s(x)\big)\,dx&=\int_a^b f(x)\,dx-L(f,P)\\&\leqslant U(f,P)-L(f,P)<\frac{\varepsilon}{2}.\end{aligned}$$

By Claim 1, there is a continuous function $h:[a,b]\to\mathbb{R}$ such that $h(a)=h(b)=0$, $\sup_{[a,b]}|h(x)|\leqslant M$ and $\int_a^b|s-h|\,dx<\varepsilon/2$. Then

$$\int_a^b|f-h|\,dx\leqslant\int_a^b|f-s|\,dx+\int_a^b|s-h|\,dx<\varepsilon.$$

This achieves (iii) for $q=1$.

Finally, let $q>1$ and $\varepsilon>0$. For every $\varepsilon_1>0$ we may choose by the above argument a continuous function h so that

$$\int_a^b|f-h|^q\,dx\leqslant(2M)^{q-1}\int_a^b|f-h|\,dx<(2M)^{q-1}\varepsilon_1.$$

Here we have used $|f-h|\leqslant|f|+|h|\leqslant 2M$. Choosing $\varepsilon_1<(2M)^{1-q}\varepsilon^q$ we achieve (iii) for $q>1$. □

2.2. Differentiation

2.2.1. Derivatives. The derivative of a function f at a point x_0 can be interpreted as the rate of change of f at x_0.

2.26. Definition. Let $I \subset \mathbb{R}$ be an interval. A function $f : I \to \mathbb{R}$ is called *differentiable at* $x_0 \in I$ if there is a number $L \in \mathbb{R}$ so that for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every h with $0 < |h| < \delta$ and $x_0 + h \in I$:

$$\left|\frac{f(x_0 + h) - f(x_0)}{h} - L\right| < \varepsilon.$$

The number L is called the *derivative* of f at x_0 and is written $f'(x_0) = L$. Many alternative notations are common, such as $\frac{df}{dx}(x_0)$ or $\frac{df}{dx}|_{x_0}$, and $D_x f(x_0)$. We also write this as

$$f'(x_0) = \lim_{\substack{h \to 0, \\ x_0 + h \in I}} \frac{f(x_0 + h) - f(x_0)}{h} \quad \text{or}$$

$$f'(x_0) = \lim_{\substack{x \to x_0, \\ x \in I}} \frac{f(x) - f(x_0)}{x - x_0}.$$

The function f is called *differentiable (on I)* if it is differentiable at every point in I.

The derivative $f'(x_0)$ can be visualized as the slope of the tangent line to the graph of f at the point x_0 (see Figure 2.2).

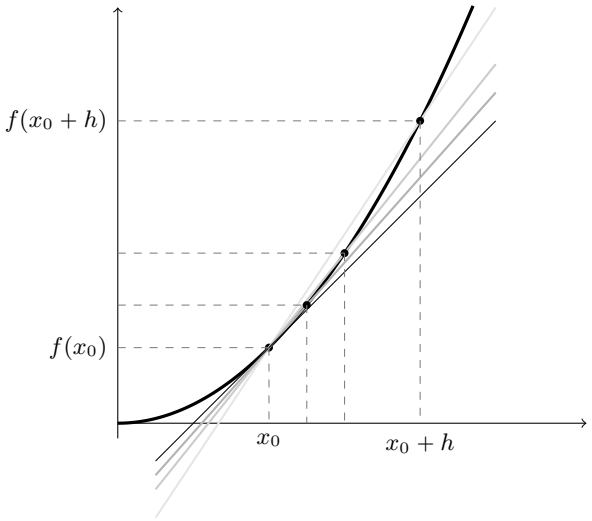


Figure 2.2. The secant slopes $\frac{f(x_0+h)-f(x_0)}{h}$ approach the tangent slope $f'(x_0)$ at x_0 as $h \to 0$.

If x_0 is an interior point, the requirement $x_0 + h \in I$ is automatic as long as $\delta > 0$ is sufficiently small (in particular, we can drop it when writing limits). If x_0 is an endpoint of the interval, then one also refers to the derivative at x_0 as a *one-sided derivative.* We also denote one-sided derivatives using one-sided limits: the *derivative of f at x_0 from the left* is

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

if the limit exists, and the derivative from the right is defined accordingly.

2.27. Example. The function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$, is not differentiable at $x_0 = 0$ since

$$\frac{f(x_0 + h) - f(x_0)}{h} = \frac{|h|}{h}$$

equals 1 if $h > 0$ and -1 if $h < 0$. However, the one-sided derivatives at $x_0 = 0$ exist: the derivative from the left is -1 and the derivative from the right is 1.

Differentiability is equivalent to approximation by a linear function. This is expressed in the following theorem which is a direct consequence of the definition.

2.28. Observation. *Let $f : I \to \mathbb{R}$ be differentiable at $x_0 \in I$. Then*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(x - x_0)$$

as $x \to x_0$ in I.

Vice versa, if $c \in \mathbb{R}$ is such that $f(x) = f(x_0) + c(x - x_0) + o(x - x_0)$ as $x \to x_0$ in I, then f is differentiable at x_0 and $f'(x_0) = c$.

A consequence of this is:

2.29. Observation. *If f is differentiable at x_0, then f is continuous at x_0.*

The converse is not true as Example 2.27 shows.

2.2.2. Rules for differentiation. The properties of limits can be used to derive rules about differentiation for sums and products of functions (the latter is referred to as the *product rule*).

2.30. Theorem. *Let $f, g : (a, b) \to \mathbb{R}$ be differentiable at $x_0 \in (a, b)$. Then $f + g$ and $f \cdot g$ are differentiable at x_0 and*

$$(f + g)'(x_0) = f'(x_0) + g'(x_0),$$
$$(f \cdot g)'(x_0) = f(x_0)g'(x_0) + f'(x_0)g(x_0).$$

Proof. We prove the product rule. Write

$$\frac{f(x_0 + h)g(x_0 + h) - f(x_0)g(x_0)}{h} = \frac{f(x_0 + h) - f(x_0)}{h} g(x_0 + h) + f(x_0) \frac{g(x_0 + h) - g(x_0)}{h}$$

and by the properties of limits and the definition of differentiability the product rule follows. □

2.31. Example. An inductive application of the product rule shows that the derivative of $f(x) = x^n$ is $f'(x) = nx^{n-1}$.

In addition we have the *chain rule*:

2.32. Theorem. *If $g : (a, b) \to (c, d)$ is differentiable at $x_0 \in (a, b)$ and $f : (c, d) \to \mathbb{R}$ is differentiable at $g(x_0) \in (c, d)$, then the composition $f \circ g : (a, b) \to \mathbb{R}$, $(f \circ g)(x) = f(g(x))$ is differentiable at x_0 and*

$$(f \circ g)'(x_0) = f'(g(x_0))g'(x_0).$$

Proof. Let

$$r_1(x) = g(x) - g(x_0) - g'(x_0)(x - x_0),$$
$$r_2(y) = f(y) - f(g(x_0)) - f'(g(x_0))(y - g(x_0)).$$

Then $\lim_{x \to x_0} \frac{r_1(x)}{|x - x_0|} = 0$ since g is differentiable at x_0 and $\lim_{y \to g(x_0)} r_2(y) = 0$ since f is differentiable at $g(x_0)$. Using both formulas we can write, for x near x_0,

$$f(g(x)) = f(g(x_0)) + f'(g(x_0))g'(x_0)(x - x_0) + R(x)$$
$$\text{where } R(x) = f'(g(x_0))r_1(x) + r_2(g(x))$$

and the theorem is proved once we show that $\lim_{x\to x_0}\frac{R(x)}{|x-x_0|}=0$. This is the main idea. In order to accomplish the task we must find, for every given $\varepsilon>0$, a $\delta>0$ such that $|R(x)|\leqslant\varepsilon|x-x_0|$ for $|x-x_0|<\delta$. Define $\varepsilon_1=\frac{\varepsilon}{2(1+|f'(g(x_0))|)}$. Since $\lim_{x\to x_0}\frac{r_1(x)}{|x-x_0|}=0$ there exists a $\delta_1>0$ such that $(x_0-\delta_1,x_0+\delta_1)\subset(a,b)$ and $|r_1(x)|<\varepsilon_1|x-x_0|$ provided that $|x-x_0|<\delta_1$. Let $\varepsilon_2>0$ such that $\varepsilon_2<\frac{\varepsilon}{2(\varepsilon_1+|g'(x_0)|)}$. Since $\lim_{y\to g(x_0)}\frac{r_2(y)}{|y-g(x_0)|}=0$ there is a $\delta_2>0$ such that $(g(x_0)-\delta_2,g(x_0)+\delta_2)\subset(c,d)$ and $|r_2(y)|<\varepsilon_2|y-g(x_0)|$ whenever $|y-g(x_0)|<\delta_2$. Let $\delta_*=\frac{\delta_2}{\varepsilon_1+|g'(x_0)|}$. Then for $|x-x_0|<\delta_*$

$$|g(x)-g(x_0)|\leqslant|g'(x_0)(x-x_0)+r_1(x)|\leqslant(|g'(x_0)|+\varepsilon_1)|x-x_0|<\delta_2$$

and therefore

$$|r_2(g(x))|\leqslant\varepsilon_2|g(x)-g(x_0)|\leqslant\varepsilon_2(\varepsilon_1+|g'(x_0)|)|x-x_0|<\frac{\varepsilon}{2}|x-x_0|.$$

Moreover, if $|x-x_0|<\delta_1$,

$$|f'(g(x_0))r_1(x)|\leqslant|f'(g(x_0))|\varepsilon_1|x-x_0|<\frac{\varepsilon}{2}|x-x_0|.$$

Define $\delta=\min\{\delta_*,\delta_1\}$ then if $|x-x_0|<\delta$ we have both remainder terms estimated by $\frac{\varepsilon}{2}|x-x_0|$. Thus we obtain for $|x-x_0|<\delta$

$$|R(x)|\leqslant|f'(g(x_0))||r_1(x)|+|r_2(g(x))|<\frac{\varepsilon}{2}|x-x_0|+\frac{\varepsilon}{2}|x-x_0|=\varepsilon|x-x_0|.\qquad\square$$

2.2.3. Inverse functions. If a function $f:I\to J$ has an inverse function $f^{-1}:J\to I$, then

$$f^{-1}(f(x))=x$$

holds for all $x\in I$. If in addition f is differentiable at $x_0\in I$ and f^{-1} is differentiable at $f(x_0)$, then by the chain rule

$$(f^{-1})'(f(x_0))f'(x_0)=1.$$

In particular, $f'(x_0)\neq0$ and $(f^{-1})'(f(x_0))\neq0$ and we obtain for $y_0=f(x_0)$ the formula

$$(f^{-1})'(y_0)=\frac{1}{f'(x_0)}.\tag{2.3}$$

Our next theorem yields this same differentiation formula; however, the point is that the differentiability of f^{-1} at $f(x_0)$ is not a hypothesis, but a conclusion.

2.33. Theorem. *Let $f:[a,b]\to\mathbb{R}$ be continuous and strictly monotone. Let $x_0\in[a,b]$ and assume that f is differentiable at x_0, and $f'(x_0)\neq0$. Then f^{-1} is differentiable at $y_0=f(x_0)$ and (2.3) holds.*

Proof. Let $[c,d]$ be the range of f. By Theorem 1.89 f has a continuous inverse

$$f^{-1}:[c,d]\to[a,b].$$

Let $(y_n)_{n=1}^{\infty}$ be an arbitrary sequence in $[c,d]$ such that $\lim_{n\to\infty}y_n=y_0$, and $y_n\neq y_0$. Let $x_n=f^{-1}(y_n)$, then $x_n\neq x_0$. By the continuity of f^{-1} we see that

$$\lim_{n\to\infty}x_n=\lim_{n\to\infty}f^{-1}(y_n)=f^{-1}(y_0)=x_0.$$

Hence (using that $f'(x_0)\neq0$)

$$\frac{f^{-1}(y_n)-f^{-1}(y_0)}{y_n-y_0}=\frac{1}{\frac{f(x_n)-f(x_0)}{x_n-x_0}}\ \to\ \frac{1}{f'(x_0)}\ (\text{as } n\to\infty),$$

and since (y_n) was arbitrary we can use Theorem 1.75 to obtain the differentiability of f^{-1} at y_0 and the claimed formula. $\square$

2.2.4. Local extrema. Derivatives are useful for detecting extreme points of functions. We say that x_0 is a *local minimum point* of a function $f : (a, b) \to \mathbb{R}$ if there is $h > 0$ so that $(x_0 - h, x_0 + h) \subset (a, b)$ and

$$f(x) \geqslant f(x_0)$$

for all $x \in (x_0 - h, x_0 + h)$. Local maximum points are defined accordingly.

2.34. THEOREM. *If $x_0 \in (a, b)$ is a local minimum or maximum point of a differentiable function $f : (a, b) \to \mathbb{R}$, then $f'(x_0) = 0$.*

PROOF. If x_0 is a local minimum, then there exists $\delta > 0$ so that $f(x_0 + h) \geqslant f(x_0)$ for all $|h| < \delta$. Thus for $0 < h < \delta$,

$$\frac{f(x_0 + h) - f(x_0)}{h} \geqslant 0.$$

Letting $h \to 0^+$ we obtain $f'(x_0) \geqslant 0$. Similarly, for $-\delta < h < 0$,

$$\frac{f(x_0 + h) - f(x_0)}{h} \leqslant 0$$

and letting $h \to 0^-$ gives $f'(x_0) \leqslant 0$. This proves $f'(x_0) = 0$. The case of a local maximum follows from the case of a local minimum because a local maximum of f is a local minimum of $-f$. □

2.2.5. The mean value theorem.

2.35. THEOREM (Rolle's theorem). *Let f be continuous in $[a, b]$, differentiable in (a, b) and $f(a) = f(b)$. Then there is $\xi \in (a, b)$ such that $f'(\xi) = 0$.*

PROOF. If f is constant, there is nothing to show. If f is not constant, then by Theorem 1.87 f has a maximum or a minimum point $\xi \in (a, b)$ and by Theorem 2.34 we have $f'(\xi) = 0$. □

As an application, we obtain the *Cauchy mean value theorem.*

2.36. THEOREM. *Let f and g be continuous real-valued functions in $[a, b]$, and differentiable in (a, b). Then there exists $\xi \in (a, b)$ such that*

$$(f(b) - f(a))g'(\xi) = (g(b) - g(a))f'(\xi).$$

PROOF. Let

$$\varphi(x) = f(x)\big(g(b) - g(a)\big) - g(x)\big(f(b) - f(a)\big).$$

Then φ is continuous on $[a, b]$ and differentiable in (a, b) and we have

$$\varphi(a) = f(a)g(b) - g(a)f(b) = \varphi(b).$$

Thus by Rolle's theorem there exists $\xi \in (a, b)$ such that $\varphi'(\xi) = 0$ which gives the asserted formula. □

The standard mean value theorem follows if we let $g(x) = x$ in Theorem 2.36.

2.37. COROLLARY (Mean value theorem). *Let $f : [a, b] \to \mathbb{R}$ be continuous. Suppose f is differentiable on (a, b). Then there exists $\xi \in (a, b)$ such that*

$$\frac{f(b) - f(a)}{b - a} = f'(\xi).$$

2.38. COROLLARY. *If $f : I \to \mathbb{R}$ is differentiable with $f'(x) = 0$ for all $x \in I$, then f is constant on I.*

2.39. COROLLARY. *Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Let $M = \sup\{|f'(x)| : x \in (a, b)\}$. Then for all $x, w \in (a, b)$*

$$|f(x) - f(w)| \leqslant M|x - w|.$$

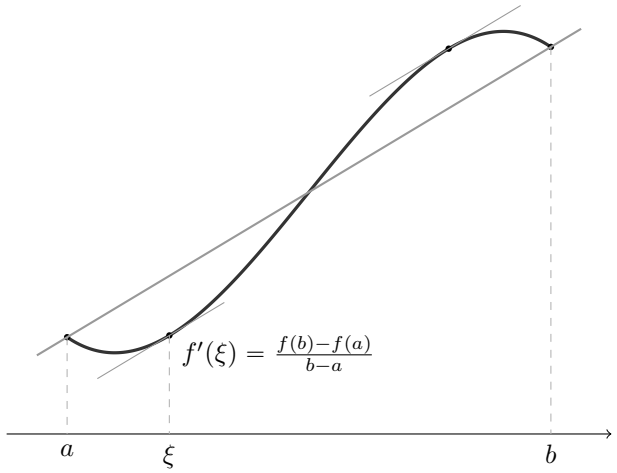


FIGURE 2.3. The mean value theorem: There exists (at least one) $\xi \in (a, b)$ such that the tangent line at ξ is parallel to the secant line through the endpoints.

2.2.6. Arc length of graphs. Given a function $f : [a, b] \to \mathbb{R}$ and a partition

$$P = \{a = x_0 < x_1 \cdots < x_N = b\}$$

we can form a polygonal curve connecting the points $(x_j, f(x_j))$ and compute the *length* as

$$\ell_P(f) = \sum_{j=1}^{N} \Big((x_j - x_{j-1})^2 + (f(x_j) - f(x_{j-1}))^2 \Big)^{1/2}.$$

The idea is to approximate the "length" of the graph

$$\mathcal{G}(f) = \{(x, f(x)) : a \leqslant x \leqslant b\}$$

by the length of such polygons.

2.40. OBSERVATION. *If the partition P' is a refinement of P, then $\ell_{P'}(f) \geqslant \ell_P(f)$.*

This follows from the triangle inequality and leads to the following definition.

2.41. DEFINITION. For $f : [a, b] \to \mathbb{R}$ we say that $\mathcal{G}(f)$ is *rectifiable* if $\sup_P \ell_P(f) < \infty$. In this case we refer to

$$\ell(\mathcal{G}(f)) = \sup_P \ell_P(f)$$

as the *arc length* of $\mathcal{G}(f)$.

The following theorem identifies a class of functions with rectifiable graphs for which the arc length can be computed by an integral. The mean value theorem for derivatives plays an important role in the proof.

2.42. THEOREM. *Let $f : [a, b] \to \mathbb{R}$ be differentiable on $[a, b]$ with continuous derivative. Then the graph $\mathcal{G}(f)$ is rectifiable and its arc length is given by*

$$\ell(\mathcal{G}(f)) = \int_a^b \sqrt{1 + (f'(t))^2}\, dt.$$

PROOF OF THEOREM 2.42. Let $g(t) = \sqrt{1 + (f'(t))^2}$. Then g is continuous on $[a, b]$ and therefore Riemann integrable. Let $P = \{x_0 < \cdots < x_N\}$ be a partition. We may apply the mean-value theorem

for derivatives (Corollary 2.37) for the intervals $[x_{j-1}, x_j]$ to write

$$\ell_P(f) = \sum_{j=1}^{N} \sqrt{1 + \Big(\frac{f(x_j) - f(x_{j-1})}{x_j - x_{j-1}}\Big)^2}(x_j - x_{j-1}) = \sum_{j=1}^{N} g(\xi_j)(x_j - x_{j-1})$$

where $\xi_j \in (x_{j-1}, x_j)$. Thus $\ell_P(f)$ is a Riemann sum for the function g and is bounded above and below by corresponding upper and lower Riemann sums:

$$L(g, P) \leqslant \ell_P(f) \leqslant U(g, P).$$

Refining partitions and using Observation 2.3 we then get

$$\sup_P L(g, P) \leqslant \sup_P \ell_P(f) \leqslant \inf_P U(g, P).$$

Since g is Riemann integrable we see that $\mathcal{G}(f)$ is rectifiable and the asserted formula for the length of $\mathcal{G}(f)$ holds. □

2.3. Fundamental theorem of calculus

There are two fundamental theorems of calculus relating the integral and the derivative. The first one is about differentiating the integral $\int_a^x f(t)dt$ with respect to x, and the second one is about integrating the derivative.

2.43. Theorem (Fundamental theorem of calculus, first part). *Let $f : [a, b] \to \mathbb{R}$ be Riemann integrable and define $F(x) = \int_a^x f$. Let $x_0 \in [a, b]$. If f is continuous at x_0, then F is differentiable at x_0 with*

$$F'(x_0) = f(x_0).$$

If $x_0 = a$, then $F'(a)$ is understood to be the derivative from the right and if $x_0 = b$, then $F'(b)$ is understood to be the derivative from the left.

Proof. Note that $F(a) = 0$. Let $\varepsilon > 0$. Since f is continuous at x_0 there is $\delta > 0$ such that $|f(t) - f(x_0)| < \varepsilon$ for all $t \in [a, b]$ such that $|t - x_0| < \delta$. If $x_0 \in [a, b)$ and $0 < h < \min\{\delta, b - x_0\}$, then (using Theorem 2.19 for the estimation)

$$\begin{aligned}\Big|\frac{F(x_0 + h) - F(x_0)}{h} - f(x_0)\Big| &= \Big|\frac{1}{h}\int_{x_0}^{x_0+h} f(t)\,dt - f(x_0)\Big| = \Big|\frac{1}{h}\int_{x_0}^{x_0+h} \big(f(t) - f(x_0)\big)\,dt\Big| \\ &\leqslant \frac{1}{h}\int_{[x_0, x_0+h]} |f(t) - f(x_0)|\,dt \leqslant \max_{s\in[x_0,x_0+h]} |f(s) - f(x_0)| < \varepsilon,\end{aligned}$$

similarly if $x_0 \in (a, b]$ and $\max\{-\delta, a - x_0\} < h < 0$, then

$$\begin{aligned}\Big|\frac{F(x_0 + h) - F(x_0)}{h} - f(x_0)\Big| &= \Big| - \frac{1}{h}\int_{x_0+h}^{x_0} f(t)dt - f(x_0)\Big| = \Big|\frac{1}{-h}\int_{x_0+h}^{x_0} \big(f(t) - f(x_0)\big)\,dt\Big| \\ &\leqslant \frac{1}{|h|}\int_{[x_0+h, x_0]} |f(t) - f(x_0)|\,dt \leqslant \max_{s\in[x_0+h,x_0]} |f(s) - f(x_0)| < \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, the assertion follows from the definition of derivative. □

2.44. Definition. Let $f : [a, b] \to \mathbb{R}$ be a function. A function $F : [a, b] \to \mathbb{R}$ is called a *primitive* or *antiderivative* of f if it is differentiable and satisfies $F'(x) = f(x)$ for all $x \in [a, b]$.

2.45. Theorem (Fundamental theorem of calculus, second part). *Let F be a primitive of a function $f : [a, b] \to \mathbb{R}$. Then*

$$\underline{\mathcal{I}}_a^b(f) \leqslant F(b) - F(a) \leqslant \overline{\mathcal{I}}_a^b(f).$$

If f is also Riemann integrable, then $\int_a^b f = F(b) - F(a)$.

PROOF. It suffices to prove for every partition $P = \{a = x_0 < x_1 \cdots < x_N = b\}$ of $[a, b]$

$$L(f, P) \leqslant F(b) - F(a) \leqslant U(f, P). \tag{2.4}$$

Since $F' = f$ we get by the mean-value theorem for derivatives

$$F(x_i) - F(x_{i-1}) = f(\xi_i)(x_i - x_{i-1}) \text{ with } \xi_i \in (x_{i-1}, x_i)$$

and after summing over i

$$\sum_{i=1}^{N} f(\xi_i)(x_i - x_{i-1}) = \sum_{i=1}^{N} \big(F(x_i) - F(x_{i-1})\big) = F(b) - F(a).$$

Now we have $m_i(f) = \inf_{[x_{i-1},x_i]} f \leqslant f(\xi_i) \leqslant \sup_{[x_{i-1},x_i]} f = M_i(f)$ and therefore

$$L(f, P) = \sum_{i=1}^{N} m_i(f)(x_i - x_{i-1}) \leqslant \sum_{i=1}^{N} f(\xi_i)(x_i - x_{i-1}) = F(b) - F(a),$$
$$U(f, P) = \sum_{i=1}^{N} M_i(f)(x_i - x_{i-1}) \geqslant \sum_{i=1}^{N} f(\xi_i)(x_i - x_{i-1}) = F(b) - F(a).$$

□

REMARK. Under the stronger assumption that f is continuous, the conclusion in Theorem 2.45 can be derived from the first version of the fundamental theorem of calculus: if $G(x) = \int_a^x f$ we get $G'(x) = f(x)$ and $G(a) = 0$ from Theorem 2.43, thus $G'(x) = F'(x)$ on $[a, b]$ which means that $G(x) = F(x) + c$ for a constant c and evaluating for $x = a$ we get $c = -F(a)$. Hence $G(b) = F(b) - F(a)$ which is the conclusion of Theorem 2.45.

A particularly useful consequence is the *integration by parts* formula.

2.46. THEOREM. *Let f and g be differentiable on $[a, b]$ and assume that f' and g' are Riemann integrable. Then $f'g$ and $g'f$ are Riemann integrable on $[a, b]$ and we have the formula*

$$\int_a^b f'g = f(b)g(b) - f(a)g(a) - \int_a^b fg'.$$

PROOF. By assumption f', g' are bounded and then f, g are (uniformly) continuous and bounded and we see that $f'g$ and fg' are Riemann integrable on $[a, b]$. By the product rule we have $(fg)' = f'g + fg'$ on $[a, b]$ and therefore

$$f(b)g(b) - f(a)g(a) = \int_a^b (fg)' = \int_a^b f'g + \int_a^b fg'.$$

□

Integration by parts is not only a device to compute integrals but also an excellent tool to discover cancellation phenomena when estimating integrals (in these situations the size bound (2.2) can be ineffective). See Exercise 2.140 below. Another consequence of the fundamental theorem of calculus is *integration by substitution.*

2.47. THEOREM. *Let $g : [a, b] \to \mathbb{R}$ be differentiable with continuous derivative g', let I be an interval containing the range of g and let $f : I \to \mathbb{R}$ be continuous. Then*

$$\int_{g(a)}^{g(b)} f(x)\,dx = \int_a^b f(g(t))g'(t)\,dt.$$

This follows by recognizing that both sides are equal to $F(g(b)) - F(g(a))$ where $F(x) = \int_a^x f$ is a primitive of f (by the chain rule, $F \circ g$ is a primitive of $t \mapsto f(g(t))g'(t)$).

2.4. Taylor's theorem

We have learned that the derivative provides an approximation of a function f near a given point a by a linear polynomial. Taylor's theorem provides approximations of a function f near a given point a by higher-order polynomials whose coefficients are determined by the higher-order derivatives of f at a. Higher-order derivatives are derivatives of derivatives and are denoted $f''(x), f'''(x), \dots$ or $f^{(k)}(x)$ for the kth derivative at x.

2.48. Definition (C^k functions). Let $k \in \mathbb{N}_0$ and $I \subset \mathbb{R}$ be an interval. We say that $f : I \to \mathbb{R}$ is *k times continuously differentiable*, or a *$C^k(I)$ function* if the derivatives $f, f', \dots, f^{(k)}$ exist and are continuous on I. We say that f is *smooth* on I, or a *$C^\infty(I)$ function* if all derivatives $f^{(k)}(x)$ exist for all $x \in I$. If I is clear from context, we may also write C^k or C^∞ instead of $C^k(I)$, $C^\infty(I)$.

2.49. Theorem (Taylor). *Let I be an interval and let $f \in C^{n+1}(I)$, $a \in I, a + t \in I$. Then*

$$f(a+t) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} t^k + R_n(a,t),$$

where the remainder term R_n satisfies

$$R_n(a,t) = \frac{t^{n+1}}{n!} \int_0^1 (1-s)^n f^{(n+1)}(a+st)\, ds \tag{2.5}$$

In particular, $R_n(a,t) = O(t^{n+1})$.

Proof. First observe that the formula for the remainder term can be written equivalently as

$$R_n(a,t) = \int_a^{a+t} \frac{(a+t-u)^n}{n!} f^{(n+1)}(u)\, du,$$

by the substitution $u = a + st$, $du = t\, ds$. The theorem is proved by induction on n. For $n = 0$ the claim reads

$$f(a+t) = f(a) + \int_a^{a+t} f'(u)\, du,$$

which follows from the fundamental theorem of calculus. Using integration by parts for $f \in C^{(n+2)}(I)$,

$$\begin{aligned}
&\int_a^{a+t} (a+t-u)^n f^{(n+1)}(u)\, du \\
&\quad = \Big[\frac{-(a+t-u)^{n+1}}{n+1} f^{(n+1)}(u) \Big]_a^{a+t} - \int_a^{a+t} \frac{-(a+t-u)^{n+1}}{n+1} f^{(n+2)}(u)\, du \\
&\quad = \frac{t^{n+1}}{n+1} f^{(n+1)}(a) + \int_a^{a+t} \frac{(a+t-u)^{n+1}}{n+1} f^{(n+2)}(u)\, du,
\end{aligned}$$

which shows

$$R_n(a,t) = \frac{t^{n+1}}{(n+1)!} f^{(n+1)}(a) + R_{n+1}(a,t)$$

and establishes the induction step of the proof of the formula. To be precise, if $(*)_N$ denotes the statement that

$$f(a+t) = \sum_{k=0}^{N} \frac{f^{(k)}(a)}{k!} t^k + R_N(a,t)$$

holds for all $f \in C^{(N+1)}(I)$, then $(*)_N$ implies $(*)_{N+1}$ for all $N = 0, 1, 2, \dots$. □

2.50. Theorem. *Let f, R_n be as in Theorem 2.49. Then there is $\xi \in [a, a+t]$ such that*

$$R_n(a,t) = \frac{t^{n+1}}{(n+1)!} f^{(n+1)}(\xi).$$

PROOF. Let

$$m = \min\{f^{(n+1)}(a+st) : 0 \leqslant s \leqslant 1\},$$
$$M = \max\{f^{(n+1)}(a+st) : 0 \leqslant s \leqslant 1\}.$$

We estimate

$$\int_0^1 (1-s)^n m\, ds \leqslant \int_0^1 (1-s)^n f^{(n+1)}(a+st)\, ds \leqslant \int_0^1 (1-s)^n M\, ds.$$

Thus

$$m \leqslant (n+1)\int_0^1 (1-s)^n f^{(n+1)}(a+st)\, ds \leqslant M.$$

By the intermediate value theorem for continuous real-valued functions there is $\sigma \in (0,1)$ such that

$$f^{(n+1)}(a+\sigma t) = (n+1)\int_0^1 (1-s)^n f^{(n+1)}(a+st)\, ds. \tag{2.6}$$

If we set $\xi = a + \sigma t$ so that ξ is on the open line segment connecting a to $a+t$ we get the claimed statement from (2.5) and (2.6). □

2.51. COROLLARY (Remainder estimate). *With f, R_n as in Theorem 2.49:*

$$|R_n(a,t)| \leqslant \frac{|t|^{n+1}}{(n+1)!} \max\{|f^{(n+1)}(a+st)| : 0 \leqslant s \leqslant 1\}.$$

REMARK. In Theorem 2.50 the assumption of continuity of $f^{(n+1)}$ is needed for our argument but it is not necessary for the stated result; it is enough to assume continuity of $f^{(n)}$ in $I = [a,b]$ and existence of $f^{(n+1)}(t)$ in (a,b). See Exercise 2.131.

2.5. Transcendental functions

There are various ways to define the standard transcendental functions $\log, \exp, \sin, \cos$. The approach we will take is via the fundamental theorem of calculus. In §3.2.4 equivalent definitions using power series are given.

2.5.1. Logarithm and exponential function. Define a function

$$L : (0,\infty) \to (-\infty,\infty), \quad L(t) = \int_1^t \frac{ds}{s}.$$

This is a well-defined continuous and strictly increasing function because $s \mapsto 1/s$ is positive and Riemann integrable on every closed interval $I \subset (0,\infty)$. Also $L(1) = 0$ and L is differentiable with $L'(t) = 1/t$ by the fundamental theorem of calculus.

2.52. OBSERVATION. *For $s, t > 0$,*

$$L(t \cdot s) = L(t) + L(s)\,.$$

PROOF. By a change of variable

$$L(s) + L(t) = \int_1^t \frac{du}{u} + \int_1^s \frac{du}{u} = \int_1^t \frac{du}{u} + \int_t^{ts} \frac{du}{u} = \int_1^{ts} \frac{du}{u} = L(st).$$

□

Also, L is surjective, that is

$$\lim_{t\to\infty} L(t) = \infty, \quad \lim_{t\to 0^+} L(t) = -\infty.$$

Indeed, for $2^N \leqslant t \leqslant 2^{N+1}$,

$$L(t) \geqslant \int_1^{2^N} \frac{ds}{s} = \sum_{k=0}^{N-1} \int_{2^k}^{2^{k+1}} \frac{ds}{s} \geqslant \sum_{k=0}^{N-1} 2^k 2^{-k-1} = \tfrac{1}{2}N.$$

This implies $\lim_{t\to\infty} L(t) = \infty$. Moreover, $L(t) = -L(1/t)$ by a change of variables, so $\lim_{t\to 0^+} L(t) = -\lim_{t\to\infty} L(t) = -\infty$.

2.53. Definition. $L(t)$ is called the *natural logarithm* of t and denoted by $\log(t)$ or $\ln(t)$. For $a > 1$ we also define

$$\log_a(t) = \frac{\log(t)}{\log(a)}.$$

With this notation

$$\log'(t) = \frac{1}{t}, \qquad \log_a'(t) = \frac{1}{t \log a} \quad \text{for } t > 0.$$

Since $L = \log : (0,\infty) \to \mathbb{R}$ is strictly increasing and surjective, it is bijective and has an inverse function.

2.54. Definition. The *exponential function* $\exp : \mathbb{R} \to (0,\infty)$ is defined as the inverse of the function $\log : (0,\infty) \to \mathbb{R}$.

Thus by definition $\exp(\log t) = t$ and $\log(\exp x) = x$ for all $t > 0$, $x \in \mathbb{R}$.

2.55. Theorem.

(i) $\exp(0) = 1$,
(ii) $\exp(x+y) = \exp(x)\exp(y)$ *for all* $x, y \in \mathbb{R}$,
(iii) $\exp(1) = e$ *(Euler's number),*
(iv) $\exp$ *is differentiable on* $\mathbb{R}$ *with* $\exp'(x) = \exp(x)$ *for all* $x \in \mathbb{R}$.

Proof. (i) $\log(1) = 0$ means $\exp(0) = 1$.

(ii) By Observation 2.52

$$\log(\exp(x)\exp(y)) = \log(\exp(x)) + \log(\exp(y)) = x + y.$$

Apply the exponential function on both sides.

(iii) By Theorem 1.55,

$$\log e = \lim_{n\to\infty} n \log\Big(1 + \frac{1}{n}\Big).$$

Since $\log(1) = 0$ and $\log'(1) = 1$,

$$\log(e) = \lim_{n\to\infty} \frac{\log(1+\frac{1}{n}) - \log(1)}{\frac{1}{n}} = \log'(1) = 1.$$

Applying exp on both sides shows $e = \exp(1)$.

(iv) By Theorem 2.33,

$$\exp'(x) = \frac{1}{\log'(\exp(x))} = \frac{1}{\frac{1}{\exp(x)}} = \exp(x).$$

□

2.56. Definition. Define for $a > 0$, $x \in \mathbb{R}$,

$$a^x = \exp(x \log(a)).$$

In particular, $\exp(x) = e^x$.

2.57. Exercise. Let $a > 0, x, y \in \mathbb{R}$. Show each of the following:

(i) $a^{x+y} = a^x a^y$ and $(a^x)^y = a^{xy}$.

(ii) $\frac{d}{dx}a^x = a^x \log a$
(iii) $\frac{d}{dx}x^a = ax^{a-1}$ for $x > 0$.

As a consequence of the derivative formula and Taylor's theorem, we can prove the following result due to Euler.

2.58. THEOREM. *e is irrational.*

PROOF. By Theorem 2.55 exp is a C^∞ function with $\exp^{(k)} = \exp$ and $\exp^{(k)}(0) = 1$ for all $k \in \mathbb{N}_0$. Taylor's theorem (Theorem 2.50) shows that for every $x \in \mathbb{R}$ and $n \in \mathbb{N}$ there is a $\xi = \xi_n(x)$ between 0 and x such that

$$\exp(x) = \sum_{k=0}^{n} \frac{x^k}{k!} + \exp(\xi)\frac{x^{n+1}}{(n+1)!}. \tag{2.7}$$

We argue by contradiction and assume that e is rational. Then $e = p/q$ for some coprime natural numbers p, q. Since for $n \geqslant 2$, $2 < \sum_{k=0}^{n} \frac{1}{k!} < 2 + \sum_{k=2}^{n} 2^{1-k} < 3$ we get $2 < e < 3$. In particular, e is not an integer and $q \geqslant 2$. Using (2.7) with $x = 1$ and $n = q$ yields

$$(q-1)!p = q!e = \sum_{k=0}^{q} \frac{q!}{k!} + \frac{\exp(\xi)}{q+1}$$

with $\xi \in (0,1)$. Hence $0 < \exp(\xi)/(q+1) < 3/(q+1)$ and $3/(q+1) \leqslant 1$. On the other hand,

$$\frac{\exp(\xi)}{q+1} = (q-1)!p - \sum_{k=0}^{q} \frac{q!}{k!}$$

is an integer. This gives a contradiction. □

2.5.2. Trigonometric functions. The upper half of the unit circle can be expressed as the graph of the function $f(t) = \sqrt{1-t^2}$ on $(-1,1)$. When $t \in [0,1)$, let $\alpha(t)$ denote the arc length of the graph of f on $[0,t]$, i.e. the length of the circular arc above $[0,t)$. Since $f'(s) = -s(1-s^2)^{-1/2}$, Theorem 2.42 gives

$$\alpha(t) = \int_0^t \sqrt{1 + (f'(s))^2}\, ds = \int_0^t (1-s^2)^{-1/2}\, ds.$$

This definition extends to $(-1,1)$ and since the integrand is an even function we obtain $\alpha(-t) = -\alpha(t)$. In summary, α is an odd, continuous function on $(-1,1)$, strictly increasing on $[0,1)$. By the fundamental theorem of calculus α is differentiable on $(-1,1)$ with $\alpha'(t) = (1-t^2)^{-1/2}$. We claim that $\lim_{t\to 1-} \alpha(t)$ exists. Since α is monotone increasing, it suffices to show that $\alpha(t)$ remains bounded as t approaches 1. To see this, estimate

$$\alpha(t) \leqslant \int_0^t (1-s)^{-1/2}\, ds = 2 - 2(1-t)^{1/2} < 2.$$

Define $\alpha(1) = \lim_{t\to 1-} \alpha(t)$. The constant π is defined as

$$\pi = 2 \cdot \alpha(1).$$

REMARK. In Theorem 2.66 we will see that π is irrational. The argument above shows $2 < \pi < 4$ which does not come close to the known decimal approximation $\pi = 3.14159\ldots$ The reader may want to derive better numerical estimates for π using the definition directly. For an efficient computation see Example 3.43.

Since α is odd, it is also natural to define $\alpha(-1) = -\pi/2$. This establishes α as a continuous bijection $\alpha : [-1,1] \to [-\pi/2, \pi/2]$. The function $x \mapsto \alpha(x)$ is called *arcsine* and is denoted $\arcsin(x)$. Let

$S : [-\pi/2, \pi/2] \to [-1, 1]$ be the inverse of arcsin. Then S is continuous with $S(0) = 0$, $S(\pi/2) = 1$, $S(-\pi/2) = -1$. For $x \in (-\pi/2, \pi/2)$,

$$S'(x) = \frac{1}{\alpha'(S(x))} = \sqrt{1 - S(x)^2},$$

so by the chain rule

$$S''(x) = -S(x).$$

We now define the familiar *sine* and *cosine* functions by

2.59. Definition. Define

$$\sin x = \begin{cases} S(x), & -\frac{\pi}{2} \leqslant x \leqslant \frac{\pi}{2}, \\ S(\pi - x), & \frac{\pi}{2} \leqslant x \leqslant \frac{3\pi}{2}, \end{cases}$$

and

$$\cos x = \begin{cases} \sqrt{1 - S(x)^2}, & -\frac{\pi}{2} \leqslant x \leqslant \frac{\pi}{2}, \\ -\sqrt{1 - S(\pi - x)^2}, & \frac{\pi}{2} \leqslant x \leqslant \frac{3\pi}{2}. \end{cases}$$

Extend $\sin, \cos$ as 2π-periodic functions so that

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

for all $x \in \mathbb{R}$.

We see that $\sin 0 = 0$, $\sin \pi/2 = 1$, $\sin(-\pi/2) = -1$ from the corresponding properties for S above, and then by Definition 2.59 we get for $k \in \mathbb{Z}$

$$\sin(k\pi) = \cos(k\pi + \tfrac{\pi}{2}) = 0, \qquad \sin(k\pi + \tfrac{\pi}{2}) = \cos(k\pi) = (-1)^k.$$

2.60. Observation. *The functions* sin *and* cos *are differentiable on* $\mathbb{R}$ *and*

$$\sin'(x) = \cos x, \quad \cos'(x) = -\sin x$$

hold for all $x \in \mathbb{R}$.

Proof. For $x \in (-\pi/2, \pi/2)$ the claim holds by the above formulas for the derivative of S. One also checks that with the above definition sin and cos become continuous functions on $\mathbb{R}$ and that the differentiability and formulas for the derivatives hold on all intervals of the form $(k\pi - \pi/2, k\pi + \pi/2)$ for $k \in \mathbb{Z}$. We need to check the differentiability at the endpoint of these intervals, and by the 2π-periodicity it suffices to check them at $\pm\pi/2$. Note that by definition $\sin(\pi/2) = 1$, $\sin(-\pi/2) = -1$, $\cos(\pi/2) = 0$, $\cos(-\pi/2) = 0$. By the mean value theorem we obtain for $h \neq 0$

$$\frac{\sin(\frac{\pi}{2} + h) - \sin(\frac{\pi}{2})}{h} = \cos\xi, \quad \frac{\cos(\frac{\pi}{2} + h) - \cos(\frac{\pi}{2})}{h} = -\sin\tilde{\xi},$$

where $\xi, \tilde{\xi}$ are between $\frac{\pi}{2} + h$ and $\frac{\pi}{2}$. By the continuity of sin and cos we then see that sin and cos are differentiable at $\pi/2$, with $\sin'(\pi/2) = \cos(\pi/2) = 0$ and $\cos'(\pi/2) = -\sin(\pi/2) = -1$. A similar argument works for $-\pi/2$, with $\sin'(-\pi/2) = \cos(-\pi/2) = 0$, $\cos'(-\pi/2) = -\sin(-\pi/2) = 1$. □

2.61. Observation. *The following hold for all* $x, y \in \mathbb{R}$*:*

(i) $|\sin x| \leqslant 1$, *and* $|\cos x| \leqslant 1$, *moreover*

$$\cos^2 x + \sin^2 x = 1.$$

(ii) $\sin(x + y) = \sin x \cos y + \cos x \sin y$, $\cos(x + y) = \cos x \cos y - \sin x \sin y$.

Proof. (i) This holds by definition for $|x| < \frac{\pi}{2}$. We take derivatives and compute $\frac{d}{dx}(\cos^2 x + \sin^2 x) = 0$ using Observation 2.60. Thus $\cos^2 x + \sin^2 x$ is constant on $\mathbb{R}$ and therefore equal to 1 everywhere. This also implies the stated inequalities for $\sin x$ and $\cos x$.

(ii) Consider the function

$$v(t) = \sin t \cos(x + y - t) + \cos t \sin(x + y - t).$$

By Theorem 2.60 $v'(t) = 0$ for all $t \in \mathbb{R}$. Hence $v(t)$ is constant and equal to $v(0) = \sin(x+y)$. Taking $t = x$ we get $\sin(x+y) = \sin x \cos y + \cos x \sin y$. The proof of the formula for $\cos(x+y)$ is analogous. □

2.62. Exercise. Prove that for all $x \in \mathbb{R}$:

(i) $\sin x = 0$ if and only if $x = k\pi$ for some $k \in \mathbb{Z}$,
(ii) $\cos x = 0$ if and only if $x = k\pi + \frac{\pi}{2}$ for some $k \in \mathbb{Z}$,
(iii) $\cos x = \cos(-x)$, $\sin x = -\sin(-x)$,
(iv) $\sin(x \pm \frac{\pi}{2}) = \pm \cos x$, $\cos(x \pm \frac{\pi}{2}) = \mp \sin x$,
(v) $\sin(2x) = 2 \sin x \cos x$, $\cos(2x) = (\cos x)^2 - (\sin x)^2$,
(vi) $\sin \frac{\pi}{4} = \cos \frac{\pi}{4} = \frac{1}{2}\sqrt{2}$, $\sin \frac{\pi}{6} = \cos \frac{\pi}{3} = \frac{1}{2}$, $\cos \frac{\pi}{6} = \sin \frac{\pi}{3} = \frac{1}{2}\sqrt{3}$.

2.63. Exercise. Prove:

(i) $\sin(\frac{\pi}{2} + t) = \sin(\frac{\pi}{2} - t)$.
(ii) $0 < \frac{2}{\pi}x < \sin x < x$ for $0 < x < \frac{\pi}{2}$.

The function

$$\arccos : [-1, 1] \to [0, \pi]$$

is defined as the inverse function of cos restricted to the interval $[0, \pi]$. Since $\cos(x - \pi/2) = \sin x$,

$$\arccos(t) = \tfrac{\pi}{2} - \arcsin(t) \quad \text{for } t \in [-1, 1].$$

2.64. Definition. Define *tangent* and *cotangent* functions by $\tan(x) = \frac{\sin(x)}{\cos(x)}$ when $\cos(x) \neq 0$ and $\cot(x) = \frac{\cos(x)}{\sin(x)}$ when $\sin(x) \neq 0$.

2.65. Exercise. Prove the following:

(i) The functions tan and cot are differentiable in their domains and

$$\tan'(x) = \frac{1}{\cos^2 x} = 1 + \tan^2 x, \qquad \cot'(x) = -\frac{1}{\sin^2 x} = -1 - \cot^2 x.$$

(ii) $\tan(x + \pi) = \tan x$ if $\frac{1}{\pi}(x - \frac{\pi}{2}) \notin \mathbb{Z}$, and $\cot(x + \pi) = \cot x$ if $\frac{x}{\pi} \notin \mathbb{Z}$.

The function tan is strictly increasing on $(-\pi/2, \pi/2)$ with $\lim_{x \to \frac{\pi}{2}-} \tan x = \infty$. Similarly, cot is strictly decreasing on $(0, \pi)$ and $\lim_{x \to 0+} \cot x = \infty$, $\lim_{x \to \pi-} \cot x = -\infty$. Thus

$$\tan : (-\tfrac{\pi}{2}, \tfrac{\pi}{2}) \to \mathbb{R}, \qquad \cot : (0, \pi) \to \mathbb{R}$$

are bijective and their inverses $\tan^{-1} = \arctan$, $\cot^{-1} = \operatorname{arccot}$ are

$$\arctan : \mathbb{R} \to (-\tfrac{\pi}{2}, \tfrac{\pi}{2}), \qquad \operatorname{arccot} : \mathbb{R} \to (0, \pi).$$

arctan is strictly increasing with $\lim_{t \to \infty} \arctan t = \pi/2$ and $\lim_{t \to -\infty} \arctan t = -\pi/2$. From Exercise 2.62 we have $\tan(x - \frac{\pi}{2}) = -\cot x$ for $x \in (0, \pi)$, moreover $\tan(\alpha) = -\tan(-\alpha)$ which gives

$$\operatorname{arccot}(t) = \tfrac{\pi}{2} - \arctan(t), \quad t \in \mathbb{R}.$$

Using Theorem 2.33 and Exercise 2.65 we can compute the derivative of arctan:

$$\arctan'(t) = \frac{1}{\tan'(\arctan t)} = \frac{1}{1 + \tan^2(\arctan t)} = \frac{1}{1 + t^2}.$$

Similarly, $\arcsin'(t) = \frac{1}{\sqrt{1-t^2}}$, $\arccos'(t) = -\frac{1}{\sqrt{1-t^2}}$, and $\operatorname{arccot}'(t) = -\frac{1}{1+t^2}$.

We finish by using properties of sin and cos to prove Lambert's theorem (1760).

2.66. Theorem. *π is irrational.*

PROOF OF THEOREM 2.66. One argues by contradiction, so assume that $\pi = p/q$, with positive integers p, q. On the interval $[0, \pi]$ consider the polynomials

$$f_n(x) = \frac{q^n x^n (\pi - x)^n}{n!}$$

for large n, to be specified. Note that $f_n(x) = f_n(\pi - x)$ and that $f_n(x) > 0$ for $0 < x < \pi$. The idea is to consider the integral

$$J_n = \int_0^\pi f_n(x) \sin x \, dx.$$

We try to evaluate it using a double integration by parts to get

$$\int_0^\pi g(x) \sin x \, dx = g(\pi) + g(0) - \int_0^\pi g''(x) \sin x \, dx$$

for, say, any polynomial g. In the calculation use $\sin 0 = \sin \pi = 0$, $\cos 0 = 1$, $\cos \pi = -1$. We apply the formula with g each of $f_n^{(2l)}$, for $l = 0, \ldots, n$. Taking into account that $f_n^{(2n+2)} = 0$ (since f_n is a polynomial of degree $2n$) this leads to

$$J_n = \sum_{l=0}^{n} (-1)^l \big(f_n^{(2l)}(\pi) + f_n^{(2l)}(0) \big).$$

If $\pi = p/q$ we get $n! f_n(x) = x^n (p - qx)^n = \sum_{j=n}^{2n} b_{j,n} x^j$ where the $b_{j,n}$ are integers. For $k > 2n$ we have $f_n^{(k)}(x) = 0$ for all x and for $k < n$ we have $f_n^{(k)}(0) = 0$. For $n \leqslant k \leqslant 2n$, $f_n^{(k)}(0) = \frac{k!}{n!} b_{k,n}$ is an integer. Thus $f_n^{(k)}(0) \in \mathbb{Z}$ for all $k, n \in \mathbb{N}_0$. Since $f_n(x) = f_n(\pi - x)$ we also get $f_n^{(k)}(\pi) \in \mathbb{Z}$ for all $k, n \in \mathbb{N}_0$. This implies, by the displayed formula, that J_n is an integer, for all $n \in \mathbb{N}_0$.

Now $\sin x > 0$ for $0 < x < \pi$ (see Exercise 2.63) and we get $f_n(x) \sin x > 0$ for $0 < x < \pi$. Hence $J_n > 0$ for all $n \in \mathbb{N}$. We estimate for $x \in [0, \pi]$

$$0 \leqslant n! f_n(x) = x^n (p - qx)^n = (px)^n (1 - \tfrac{x}{\pi})^n \leqslant (px)^n$$

and therefore $f_n(x) \leqslant (px)^n / n!$ for $x \in [0, \pi]$. Consequently

$$J_n \leqslant \int_0^\pi \frac{(px)^n}{n!} \, dx = \frac{p^n \pi^{n+1}}{(n+1)!} \, .$$

We have $\lim_{n\to\infty} \frac{p^n \pi^{n+1}}{(n+1)!} = 0$ (Example 1.45); hence $0 < J_n < 1$ for sufficiently large n. This is the desired contradiction because we have shown above that the assumption $\pi = p/q$ implies that J_n is an integer for all $n \in \mathbb{N}_0$. □

REMARK. This approach to the irrationality of π is due to Niven [**57**].

2.67. EXERCISE. Show that a modification of the above proof also yields that π^2 is irrational.

2.6. Improper integrals

One of the shortcomings of the Riemann integral is that Riemann integrable functions are necessarily bounded and can be integrated only over bounded intervals. To an extent this can be remedied by taking limits of Riemann integrals; these are called *improper integrals*. There are several variants which also lead to considerations of interesting special functions.

For example, the *improper integral* $\int_a^{b-} f$ is defined by

$$\int_a^{b-} f = \lim_{\beta \to b-} \int_a^\beta f,$$

provided f is Riemann integrable on every interval $[a, \beta]$ with $\beta < b$ and the limit exists. Here b is allowed to take the value ∞. Analogously one defines the improper integrals $\int_{a+}^b f$, where a is allowed

to take the value $-\infty$. We write $\int_a^b$ instead of $\int_{a+}^{b-}$, $\int_a^{b+}$, $\int_{a-}^b$ if it is clear from context that the integral is improper.

2.68. EXAMPLE. Let $s \in \mathbb{R}$. The integral $\int_{0+}^1 x^{-s} dx$ exists if and only if $s < 1$. This is because for $\varepsilon > 0$

$$\int_\varepsilon^1 x^{-s}\,dx = \begin{cases} \frac{1}{1-s}(1-\varepsilon^{1-s}) & \text{if } s \neq 1, \\ \log(1/\varepsilon) & \text{if } s = 1. \end{cases}$$

As $\varepsilon \to 0+$, the limit exists if and only if $s < 1$ and in this case

$$\int_{0+}^1 x^{-s}\,dx = \lim_{\varepsilon\to 0+} \int_\varepsilon^1 x^{-s}\,dx = (1-s)^{-1}.$$

Similarly, the integral $\int_1^{\infty-} x^{-s}\,dx$ exists if and only if $s > 1$ and

$$\int_1^{\infty-} x^{-s}\,dx = \lim_{R\to\infty} \int_1^R x^{-s}\,dx = (s-1)^{-1}.$$

2.69. EXERCISE. Show that the improper integral $\int_0^{1-}(1-x^2)^{-\alpha}\,dx$ exists if and only if $\alpha < 1$. For the case $\alpha = -1/2$ the existence has already been shown in §2.5.2 (without using the words improper integral). We defined the number π as $2\int_0^{1-}(1-x^2)^{-1/2}\,dx$.

A useful tool to show that improper integrals exist is absolute convergence:

2.70. OBSERVATION. *If the improper Riemann integral $\int_{a+}^b |f|$ exists, then $\int_{a+}^b f$ also exists. In this case, we also say that $\int_{a+}^b f$* converges absolutely.

PROOF. Let (b_n) be a sequence such that $b_n < b$ for all n and such that $\lim_{n\to\infty} b_n = b$. Let $\mathcal{I}_n = \int_a^{b_n} f$, $\mathcal{J}_n = \int_a^{b_n} |f|$. It suffices to show that the sequence $(\mathcal{I}_n)$ converges. By assumption, $(\mathcal{J}_n)$ converges and is therefore a Cauchy sequence. But

$$|\mathcal{I}_n - \mathcal{I}_m| = \Big|\int_{b_m}^{b_n} f\Big| \leqslant \Big|\int_{b_m}^{b_n} |f|\Big| = |\mathcal{J}_n - \mathcal{J}_m|,$$

so $(\mathcal{I}_n)$ is also a Cauchy sequence and therefore convergent. □

Note that absolute convergence of $\int_a^b f(x)\,dx$ does not allow us to compute the value of $\int_a^b f(x)\,dx$, which may often be difficult.

2.71. EXAMPLE. We show that the integral

$$\int_\pi^{\infty-} \frac{\sin x}{x^s}\,dx$$

exists whenever $s > 0$. To see this we show that for every sequence $R_n \to \infty$, the sequence

$$A_n = \int_\pi^{R_n} \frac{\sin x}{x^s}\,dx$$

is convergent. Then the existence of the improper integral follows by the argument in Theorem 1.75. It suffices to prove that (A_n) is a Cauchy sequence. Using integration by parts,

$$A_m - A_n = \int_{R_n}^{R_m} \frac{\sin x}{x^s}\,dx = R_n^{-s}\cos(R_n) - R_m^{-s}\cos(R_m) - s\int_{R_n}^{R_m} x^{-s-1}\cos(x)\,dx.$$

Since $s > 0$ we get $\lim_{n\to\infty} R_n^{-s}\cos(R_n) = 0$. We estimate the integral, assuming $R_m \geqslant R_n$ by

$$\Big|s\int_{R_n}^{R_m} x^{-s-1}\cos(x)\,dx\Big| \leqslant s\int_{R_n}^{R_m} x^{-s-1}dx = |R_n^{-s} - R_m^{-s}|$$

and a similar consideration applies to the case $R_n \geqslant R_m$. This shows that (A_n) is a Cauchy sequence.

2.72. Exercise.

(i) Prove that $\int_\pi^{\infty-} \frac{\sin x}{x^s}\,dx$ does not exist when $s \leqslant 0$.

(ii) Prove that $\int_\pi^{\infty-} \frac{\sin x}{x^s}\,dx$ converges absolutely if and only if $s > 1$.

2.73. Exercise. Prove, using $\frac{d}{dx}\sin(1/x) = -x^{-2}\cos(1/x)$ that

$$\int_{0+}^{1} x^{-3/2}\sin(1/x)\,dx$$

exists as an improper integral.

One also needs to occasionally consider integrals of the form $\int_{-\infty}^{\infty}$. These are defined via two limits: we say that the improper Riemann integral

$$\int_{a+}^{b-} f$$

exists if there is a $c \in (a,b)$ so that the improper Riemann integrals $\int_{a+}^{c} f$ and $\int_{c}^{b-} f$ exist. In that case define

$$\int_{a+}^{b-} f = \int_{a+}^{c} f + \int_{c}^{b-} f.$$

One checks that the value is independent of the choice of c. Here a is allowed to take the value $-\infty$ and b is allowed to take the value ∞. In practice again one often writes $\int_a^b$ instead of $\int_{a+}^{b-}$. It is important to keep the two limiting processes separate; they generally cannot be replaced by a single limiting process. For example, the improper Riemann integral $\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,dx$ does not exist, but the limit

$$\lim_{R\to\infty}\int_{-R}^{R} \frac{x}{1+x^2}\,dx$$

exists (indeed, since $\frac{x}{1+x^2}$ is an odd function, $\int_{-R}^{R} \frac{x}{1+x^2}\,dx = 0$ for all $R > 0$).

2.74. Exercise. Let $f(x) = \frac{x}{1+x^2}$.

(i) Show that $\lim_{R\to\infty}\int_0^R f = \infty$ and $\lim_{R\to\infty}\int_{-R}^0 f = -\infty$.

(ii) Compute $\lim_{R\to\infty}\int_{-c_1R}^{c_2R} f$ for every $c_1, c_2 > 0$.

(iii) Compute $\lim_{R\to\infty}\int_{-R^\alpha}^{R^\beta} f$ for every $\alpha, \beta > 0$.

2.6.1. The Gaussian integral. We now cover an integral involving the function

$$s \mapsto e^{-s^2}$$

whose antiderivative cannot be evaluated in terms of "elementary" functions (for more about this, see Rosenlicht [**64**]). Since $e^{-x^2} \leqslant e^{-x}$ for $x \geqslant 1$, it is straightforward to see that the improper integral $\int_{-\infty}^{\infty} e^{-x^2}\,dx$ exists. However, it is not obvious that one can actually compute it.

2.75. Theorem (Gaussian integral).

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}\,.$$

Proof. Consider

$$F(x) = \Big(\int_0^x e^{-t^2}dt\Big)^2, \quad G(x) = \int_0^1 \frac{e^{-x^2(t^2+1)}}{t^2+1}dt$$

Differentiating we get

$$F'(x) = 2e^{-x^2}\int_0^x e^{-t^2}dt$$

and

$$G'(x) = \int_0^1 \frac{d}{dx} \frac{e^{-x^2(t^2+1)}}{t^2+1}\, dt = -2e^{-x^2} \int_0^1 e^{-x^2t^2} x\, dt.$$

Note that the last differentiation under the integral can be justified by examining the limit of the difference quotients $h^{-1}(G(x+h) - G(x))$ as $h \to 0$. This is left to the reader who may also want to take a peek at Example 4.78 below. Making the substitution $u = xt$ in the integral for G' we get

$$G'(x) = -2e^{-x^2} \int_0^x e^{-u^2}\, du = -F'(x),$$

so $(F+G)'(x) = 0$ for all $x \in \mathbb{R}$. We thus get

$$F(x) + G(x) = F(0) + G(0) = \int_0^1 \frac{1}{t^2+1}\, dt = \arctan 1 = \frac{\pi}{4}.$$

Now we let $x \to \infty$ and estimate $0 \leqslant G(x) \leqslant e^{-x^2}$ and hence $\lim_{x\to\infty} G(x) = 0$. Hence we obtain

$$\int_0^{\infty-} e^{-t^2}\, dt = \lim_{x\to\infty} \sqrt{F(x)} = \lim_{x\to\infty} \sqrt{F(x)+G(x)} = \sqrt{\frac{\pi}{4}}.$$

Finally,

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \int_0^{\infty-} e^{-x^2} dx + \int_{-\infty+}^{0} e^{-x^2} dx = 2\int_0^{\infty-} e^{-x^2} dx = \sqrt{\pi}.$$

□

REMARK. The *Gaussian error function* is defined for $x \geqslant 0$ by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2}\, ds.$$

The function represents the area of the region $\{(t,y) : 0 \leqslant y \leqslant ce^{-t^2},\ 0 \leqslant t \leqslant x\}$, with $c = 2\pi^{-1/2}$. The constant is chosen so that $\lim_{x\to\infty} \operatorname{erf}(x) = 1$ (by Theorem 2.75).

2.6.2. The Gamma function. The Γ-*function* extends the notion of factorials from the natural numbers to all positive real numbers. It is ubiquitous in analysis and number theory.

2.76. DEFINITION. For $x > 0$,

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$$

This is well-defined as an improper integral for $x > 0$. Indeed we use $0 < t^{x-1}e^{-t} \leqslant t^{x-1}$ and Example 2.68 to see that $\int_{0+}^1 t^{x-1}e^{-t}\, dt$ converges. For $t \geqslant 1$ we can estimate $t^{x-1}e^{-t} \leqslant C(x)e^{-t/2}$ (see Example 1.47) to see that $\int_1^{\infty-} t^{x-1}e^{-t}\, dt$ converges for all $x > 0$.

2.77. THEOREM. *The functional equation*

$$\Gamma(x+1) = x\Gamma(x)$$

holds for all $x > 0$. For $n \in \mathbb{N}_0$,

$$\Gamma(n+1) = n!\,.$$

Moreover, for $p > -1$,

$$\int_0^\infty x^p e^{-x^2} dx = \tfrac{1}{2}\Gamma(\tfrac{p+1}{2})$$

In particular $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, and $\Gamma(\frac{2k+1}{2}) = \frac{(2k-1)!}{(k-1)!}\frac{\sqrt{\pi}}{2^{2k-1}}$ for $k = 1, 2, \ldots$

PROOF. By integration by parts

$$\int_\varepsilon^R t^{x-1}e^{-t}dt = x^{-1}(R^x e^{-R} - \varepsilon^x e^{-\varepsilon}) + x^{-1}\int_\varepsilon^R t^x e^{-t}dt.$$

The functional equation follows by letting $R \to \infty$ and $\varepsilon \to 0$. We get

$$\Gamma(1) = \int_0^\infty e^{-t}dt = 1,$$

so $\Gamma(1) = 1 = 0!$ by the usual convention for 0! (and this integral confirms again that this is a good convention). Using the functional equation repeatedly, we get $\Gamma(n+1) = n!$ for $n = 1, 2, \ldots$.

Using the substitution $t = x^2$, $dt = 2xdx$ we get

$$\int_\varepsilon^R x^p e^{-x^2}dx = \frac{1}{2}\int_{\varepsilon^2}^{R^2} t^{\frac{p-1}{2}}e^{-t}dt$$

and letting $\varepsilon \to 0$ and $R \to \infty$ yields that $\int_0^\infty x^p e^{-x^2}dx = \frac{1}{2}\Gamma(\frac{p+1}{2})$. For $p = 0$ we may use Theorem 2.75 to evaluate $\Gamma(1/2) = \sqrt{\pi}$, and the stated formula for $\Gamma(\frac{2k+1}{2})$ for $k \geqslant 1$ follows from the functional equation and induction. □

2.78. EXERCISE. Show that Γ is a continuous function on $(0, \infty)$.

2.7. Complex numbers

A *complex number* $z \in \mathbb{C}$ is a pair (x_1, x_2) of real numbers. We call $x_1 = \mathrm{Re}\,(z)$ the *real part* of z and $x_2 = \mathrm{Im}\,(z)$ the *imaginary part* of z and write

$$z = x_1 + ix_2,$$

where i is the *imaginary unit*, the complex number with $\mathrm{Re}\,(i) = 0$ and $\mathrm{Im}\,(i) = 1$. Complex numbers are in one-to-one correspondence with vectors (x_1, x_2) in the plane.

The real numbers are embedded in $\mathbb{C}$; they are the complex numbers with zero imaginary part. The complex numbers with zero real part are called *imaginary numbers*. The addition of two complex numbers is analogous to the addition of vectors in $\mathbb{R}^2$:

$$(x_1 + ix_2) + (y_1 + iy_2) = (x_1 + y_1) + i(x_2 + y_2)$$

By definition, the imaginary unit satisfies

$$i^2 = -1.$$

This leads to a general multiplication formula for complex numbers:

$$(x_1 + ix_2) \cdot (y_1 + iy_2) = (x_1y_1 - x_2y_2) + i(x_1y_2 + x_2y_1)$$

(by formally multiplying out the product and using $i^2 = -1$). Commutativity, associativity and distributivity of real numbers carry over to the complex numbers. Moreover, if $z = x_1 + ix_2 \neq 0$, then z has a unique multiplicative inverse denoted

$$z^{-1} = \frac{1}{z} = \frac{x_1}{x_1^2 + x_2^2} + i\frac{-x_2}{x_1^2 + x_2^2}.$$

2.79. EXERCISE. Prove that the complex numbers $\mathbb{C}$ form a field which contains the real numbers as a subfield. Check the field axioms in Definition 1.102.

The *complex conjugate* of $z = x_1 + ix_2$ is defined by

$$\overline{z} = x_1 - ix_2.$$

One observes

$$\mathrm{Re}\,(z) = \frac{z + \overline{z}}{2}, \quad \mathrm{Im}\,(z) = \frac{z - \overline{z}}{2i}.$$

The *absolute value* of z is defined by

$$|z| = \sqrt{x_1^2 + x_2^2}.$$

In particular, we have for the distance between two numbers $z = x_1 + ix_2$ and $w = y_1 + iy_2$

$$|w - z| = \big((x_1 - y_1)^2 + (x_2 - y_2)^2\big)^{1/2},$$

which is the Euclidean distance between (x_1, x_2) and (y_1, y_2). The *triangle inequality* for complex numbers $z, w \in \mathbb{C}$ states that

$$|z + w| \leqslant |z| + |w|.$$

Inductively, we immediately extend this to finite sums of complex numbers $z_1, \ldots, z_n$:

$$\Big|\sum_{k=1}^{n} z_k\Big| \leqslant \sum_{k=1}^{n} |z_k|.$$

We have

$$|z|^2 = z\overline{z}.$$

With this notation the above formula for the inverse of z can be rewritten as $\frac{1}{z} = \frac{\bar{z}}{|z|^2}$ and for general $w, z \in \mathbb{C}$ with $z \neq 0$,

$$\frac{w}{z} = \frac{w\bar{z}}{|z|^2}.$$

A key fact is *Euler's formula*, which states that for all real numbers $x \in \mathbb{R}$

$$e^{ix} = \cos x + i \sin x.$$

We may take this formula as the definition of e^{ix}. In §3.2.4 we will introduce an equivalent definition via power series. It is often advantageous to use *polar coordinates*: for $z \neq 0$,

$$z = r(\cos\varphi + i \sin\varphi) = re^{i\varphi},$$

where $r = |z| \geqslant 0$ and the *angle* $\varphi \in \mathbb{R}$ is uniquely determined up to an additive integer multiple of 2π. That is if $e^{i\varphi_1} = e^{i\varphi_2}$, then $\varphi_2 - \varphi_1 = 2k\pi$ for some $k \in \mathbb{Z}$. The set of complex numbers with $|z| = 1$ is called the *unit circle* and is denoted S^1 (it is a subgroup of $\mathbb{C}$, so sometimes it is also called the circle group).

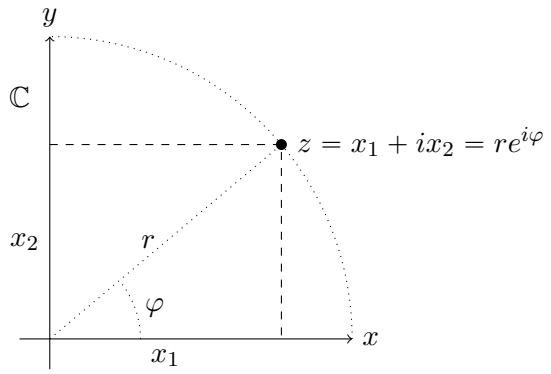


FIGURE 2.4. Polar and Cartesian coordinates in the complex plane.

Complex multiplication is particularly simple in polar coordinates. Using the trigonometric formulas for the cosine and sine of a sum of two angles we get

$$\begin{aligned}\cos(\varphi_1 + \varphi_2) + i\sin(\varphi_1 + \varphi_2) &= (\cos\varphi_1\cos\varphi_2 - \sin\varphi_1\sin\varphi_2) + i(\sin\varphi_1\cos\varphi_2 + \cos\varphi_1\sin\varphi_2)\\ &= (\cos\varphi_1 + i\sin\varphi_1)(\cos\varphi_2 + i\sin\varphi_2).\end{aligned}$$

Using the above notation $\cos\varphi + i\sin\varphi = e^{i\varphi}$ this is succinctly written as

$$(r_1 e^{i\varphi_1}) \cdot (r_2 e^{i\varphi_2}) = r_1 r_2 e^{i(\varphi_1+\varphi_2)}.$$

Geometrically, we see from this formula that multiplication by a complex number is a scaled rotation in the complex plane.

In particular, it is easy to take nth roots of nonzero numbers, i.e. to compute the n solutions $w_1, \dots, w_n$ of the equation $w^n = re^{i\varphi}$. These are given by

$$w_k = r^{1/n} e^{i\frac{\varphi+2\pi k}{n}}, \quad k = 0, \dots, n-1.$$

2.80. Example. Let us prove *Machin's formula*

$$\frac{\pi}{4} = 4\arctan\frac{1}{5} - \arctan\frac{1}{239}.$$

The formula can be used to rather efficiently approximate π; see Example 3.43. It is based on the observation

$$(5+i)^4(239-i) = 2^2 13^4(1+i).$$

The point is that this is a positive real number times $e^{i\pi/4}$ (because $1+i = \sqrt{2}e^{i\pi/4}$). Then by writing $5+i$ and $239-i$ in polar coordinates,

$$5+i = \sqrt{5^2+1}\, e^{i\alpha},\ 239-i = \sqrt{239^2+1}\, e^{-i\beta},$$

where $\alpha = \arctan\frac{1}{5}$, $\beta = \arctan\frac{1}{239}$, one sees that $(5+i)^4(239-i)$ is a positive real number times $e^{i(4\alpha-\beta)}$. Thus $4\alpha - \beta = \pi/4$, which gives the claim.

2.7.1. Complex-valued functions. Definitions for real-valued functions often carry over to complex-valued functions by applying the definitions to the real and imaginary parts separately, unless they involve the ordering of real numbers. In particular, for functions $f : [a,b] \to \mathbb{C}$ we say that f is continuous at a point x if $\operatorname{Re} f$ and $\operatorname{Im} f$ are continuous at x. We define the derivative of a complex-valued function by

$$f'(t) = (\operatorname{Re} f)'(t) + i(\operatorname{Im} f)'(t),$$

if the derivatives on the right hand side exist, and the (Riemann) integral as

$$\int_a^b f(t)dt = \int_a^b \operatorname{Re} f(t)dt + i\int_a^b \operatorname{Im} f(t)dt,$$

if the integrals on the right hand side exist. For a sequence (a_n) of complex numbers we say that it converges to a complex number a if $(\operatorname{Re} a_n)$ converges to $\operatorname{Re} a$ and $(\operatorname{Im} a_n)$ converges to $\operatorname{Im} a$. It is a good exercise to go through the theorems in this chapter and verify that they continue to hold for complex-valued functions. Theorems that involve the ordering of real numbers (such as Theorems 1.87, 1.88,2.34) should be excluded.

2.81. Exercise. Check the useful formula

$$\int_\alpha^\beta e^{i\varphi}\, d\varphi = \frac{1}{i}\big(e^{i\beta} - e^{i\alpha}\big).$$

2.8. Uniform convergence

One often encounters situations where two limits need to be interchanged. This can usually be done under an appropriate *uniform convergence* assumption.

For example, let $(a_{m,n})_{m,n\in\mathbb{N}}$ be a doubly indexed sequence (or a sequence of sequences) and suppose that $\lim_{n\to\infty} a_{m,n}$ exists for all m and $\lim_{m\to\infty} a_{m,n}$ exists for all n. We would like to know whether

$$\lim_{n\to\infty}\Big(\lim_{m\to\infty} a_{m,n}\Big) = \lim_{m\to\infty}\Big(\lim_{n\to\infty} a_{m,n}\Big),$$

and in particular, whether the two outer limits exist. It is clear that this doesn't necessarily hold without further assumptions.

2.82. EXAMPLE. Let $a_{m,n} = 1$ if $m > n$ and $a_{m,n} = -1$ otherwise. Then

$$\lim_{n\to\infty}\Big(\lim_{m\to\infty} a_{m,n}\Big) = 1, \quad \lim_{m\to\infty}\Big(\lim_{n\to\infty} a_{m,n}\Big) = -1.$$

The key concept to exclude such examples is uniform convergence:

2.83. DEFINITION. $(a_{m,n})_{m\in\mathbb{N}}$ *converges uniformly* in $n \in \mathbb{N}$ as $m \to \infty$ with limit (b_n) if for every $\varepsilon > 0$ there exists $M \in \mathbb{N}$ so that for every $n \in \mathbb{N}$ and for every $m \geqslant M$,

$$|a_{m,n} - b_n| < \varepsilon.$$

2.84. THEOREM. *Suppose that the sequences $(a_{m,n})_{m\in\mathbb{N}}$ converge uniformly in $n \in \mathbb{N}$ as $m \to \infty$ and that for every $m \in \mathbb{N}$, the sequence $(a_{m,n})_{n\in\mathbb{N}}$ converges. Then*

$$\lim_{n\to\infty}\Big(\lim_{m\to\infty} a_{m,n}\Big) = \lim_{m\to\infty}\Big(\lim_{n\to\infty} a_{m,n}\Big)$$

and in particular, the limits on both sides exist.

PROOF. Let us write

$$b_n = \lim_{m\to\infty} a_{m,n} \text{ and } c_m = \lim_{n\to\infty} a_{m,n}.$$

These limits exist by assumption. We first show that the sequence $(b_n)_{n\in\mathbb{N}}$ converges. Let $\varepsilon > 0$. By the uniform convergence assumption, there exists $M \in \mathbb{N}$ so that for all $m \geqslant M$ and all $n \in \mathbb{N}$,

$$|a_{m,n} - b_n| < \varepsilon.$$

Next there exists $N \in \mathbb{N}$ so that for all $n \geqslant N$:

$$|a_{M,n} - c_M| < \varepsilon.$$

Thus, for all $n, k \geqslant N$:

$$|b_n - b_k| \leqslant |b_n - a_{M,n}| + |a_{M,n} - c_M| + |c_M - a_{M,k}| + |a_{M,k} - b_k| < 4\varepsilon,$$

so $(b_n)_{n\in\mathbb{N}}$ is a Cauchy sequence and thus converges to a limit L. It remains to show that c_m also converges to L as $m \to \infty$. There exists $N_1 \in \mathbb{N}$ so that for all $n \geqslant N_1$, $|b_n - L| < \varepsilon$. Next, for every $m \in \mathbb{N}$ there exists $N_2(m) \in \mathbb{N}$ so that for all $n \geqslant N_2(m)$,

$$|c_m - a_{m,n}| < \varepsilon.$$

Let $m \geqslant M$ and set $n(m) = \max(N_1, N_2(m))$. Then

$$|c_m - L| \leqslant |c_m - a_{m,n(m)}| + |a_{m,n(m)} - b_{n(m)}| + |b_{n(m)} - L| < 3\varepsilon,$$

so $\lim_{m\to\infty} c_m = L = \lim_{n\to\infty} b_n$. □

2.85. EXERCISE. Suppose that $(a_{m,n})_{m,n\in\mathbb{N}}$ is increasing in n and m, i.e. $a_{m,n} \leqslant a_{m,n'}$ if $n \leqslant n'$ and $a_{m,n} \leqslant a_{m',n}$ if $m \leqslant m'$. Show that

$$\lim_{n\to\infty}\lim_{m\to\infty} a_{m,n} = \lim_{m\to\infty}\lim_{n\to\infty} a_{m,n} = \sup\{a_{m,n} : m, n \in \mathbb{N}\}$$

(where we allow the value ∞).

More generally, let us consider sequences of functions on an arbitrary set X (if X is the set of natural numbers, we recover the case of a doubly indexed sequence).

2.86. DEFINITION. A sequence $(f_n)_{n\in\mathbb{N}}$ of real- or complex-valued functions on a set X is called *uniformly convergent* to a function f if for all $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$ and all $x \in X$,

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

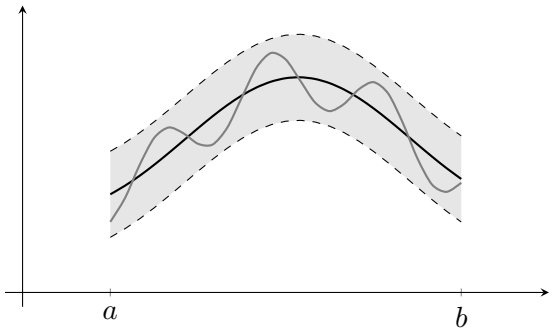


FIGURE 2.5. Uniform convergence on $[a, b]$ means: For every $\varepsilon > 0$, the graph of f_n (gray) for large enough n must be contained in an ε-neighborhood of the graph of f (shaded area).

One should contrast the notion of uniform convergence with the notion of *pointwise convergence* of a sequence of functions f_n. We say that f_n converges *pointwise* to f if for all $x \in X$, the sequence of numbers $f_n(x)$ converges to the number $f(x)$. The difference is that for uniform convergence, N needs to be chosen independently of x (*uniformly* in x), while for pointwise convergence, N is allowed to depend on x.

2.87. EXAMPLE. For every $n \in \mathbb{N}$ define $f_n : [0, 1] \to \mathbb{R}$ by

$$f_n(x) = \begin{cases} 1 - nx & \text{if } 0 \leqslant x \leqslant n^{-1}, \\ 0 & \text{if } n^{-1} < x \leqslant 1. \end{cases}$$

For every $x \in [0, 1]$ the sequence of real or complex numbers $(f_n(x))_{n\in\mathbb{N}}$ converges with

$$\lim_{n\to\infty} f_n(x) = f(x) = \begin{cases} 0 & \text{if } 0 < x \leqslant 1 \\ 1 & \text{if } x = 0 \end{cases}.$$

Thus f_n converges to f pointwise. However, for every $n \in \mathbb{N}$

$$\sup_{x\in[0,1]} |f_n(x) - f(x)| = 1$$

and hence f_n does not converge uniformly.

2.88. OBSERVATION. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions on a set X. Then $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f if and only if*

$$\lim_{n\to\infty} \sup_{x\in X} |f_n(x) - f(x)| = 0.$$

The reader should convince themselves of this fact directly from the definitions.

2.89. THEOREM. *A sequence $(f_n)_{n\in\mathbb{N}}$ of functions on a set X converges uniformly if and only if it is* uniformly Cauchy, *i.e. for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $n, m \geqslant N$ and all $x \in X$,*

$$|f_n(x) - f_m(x)| < \varepsilon.$$

PROOF. '$\Longrightarrow$': Assume that (f_n) converges uniformly to f. Then use

$$|f_n(x) - f_m(x)| = |f_n(x) - f(x) + f(x) - f_m(x)| \leqslant |f_n(x) - f(x)| + |f(x) - f_m(x)|$$

to see that f_n is uniformly Cauchy.

'$\Longleftarrow$': Assume that (f_n) is uniformly Cauchy. Given $\varepsilon > 0$ there is N such that $|f_n(x)-f_m(x)| < \varepsilon/2$ for $n \geqslant N$, $x \in X$. In particular for every $x \in X$ the sequence $(f_n(x))_{n\in\mathbb{N}}$ is Cauchy. Since all Cauchy

sequences in $\mathbb{R}$ converge, this sequence of real numbers has a limit, call it $f(x)$. Letting $m \to \infty$ we see that

$$\lim_{m\to\infty} |f_n(x) - f_m(x)| = |f_n(x) - f(x)|$$

and it follows that for $n \geqslant N$ and all $x \in X$, $|f_n(x) - f(x)| \leqslant \varepsilon/2$. Since ε is arbitrary this means that (f_n) converges to f uniformly. □

A function $f : X \to \mathbb{C}$ is called *bounded* if there exists $C > 0$ such that $|f(x)| \leqslant C$ for all $x \in X$.

2.90. THEOREM. *If $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f and each f_n is bounded, then f is bounded.*

PROOF. By uniform convergence, there is N such that $|f_n(x) - f(x)| < 1$ for all $n \geqslant N$, $x \in X$. Since f_N is bounded there is $M > 0$ such that $|f_N(x)| \leqslant M$ for all $x \in X$. Now by the triangle inequality

$$|f(x)| = |f(x) - f_N(x) + f_N(x)| \leqslant |f(x) - f_N(x)| + |f_N(x)| < 1 + M.$$

□

2.91. THEOREM. *Suppose that f_n is a sequence of functions on an interval $I \subset \mathbb{R}$ and $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f on I. If each f_n is continuous at a, then f is continuous at a. In other words, we may interchange limits:*

$$\lim_{x\to a}\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty}\lim_{x\to a} f_n(x)$$

(the left equals $\lim_{x\to a} f(x)$ and the right equals $\lim_{n\to\infty} f_n(a) = f(a)$).

In §4.1.7 we will generalize this to functions between metric spaces.

PROOF. We have to show that given $\varepsilon > 0$ there is $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ provided that $|x - a| < \delta$ (and $x, a \in I$). Since f_n converges uniformly to f there is an $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \varepsilon/3$ for $n \geqslant N$, and all $x \in I$. Consider the continuous function f_N. There is $\delta > 0$ such that $|f_N(x) - f_N(a)| < \varepsilon/3$ provided that $|x - a| < \delta$. For such x we get

$$\begin{aligned}|f(x) - f(a)| &= |f(x) - f_N(x) + f_N(x) - f_N(a) + f_N(a) - f(a)|\\ &\leqslant |f(x) - f_N(x)| + |f_N(x) - f_N(a)| + |f_N(a) - f(a)| < \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon.\end{aligned}$$

□

We also often want to interchange limits with integrals and derivatives (which are also the result of limiting operations). Uniform convergence again proves itself useful.

2.92. THEOREM. *Let $f_n : [a, b] \to \mathbb{C}$ be Riemann integrable and assume that (f_n) converges uniformly to f on $[a, b]$. Then f is Riemann integrable on $[a, b]$ and*

$$\lim_{n\to\infty} \int_a^b f_n = \int_a^b f.$$

PROOF. We may assume without loss of generality that f is real-valued. We use Lemma 2.8 to show that f is Riemann integrable. Let P be a partition of $[a, b]$. For any partition interval $I_j = [x_{j-1}, x_j] \subset [a, b]$ we have

$$\sup_{x\in I_j} f(x) - \inf_{x\in I_j} f(x) \leqslant \sup_{x\in I_j} f_n(x) - \inf_{x\in I_j} f_n(x) + 2\sup_{x\in I_j} |f_n(x) - f(x)|.$$

For the lower and upper sums this implies

$$U(f, P) - L(f, P) \leqslant U(f_n, P) - L(f_n, P) + 2(b - a)\sup_{x\in[a,b]} |f_n(x) - f(x)|.$$

Let $\varepsilon > 0$. By uniform convergence, there is N such that $|f_n(x)-f(x)| \leqslant \frac{\varepsilon}{3(b-a)}$ for $n \geqslant N$, $x \in [a,b]$. Also since f_N is Riemann integrable there is a partition P such that $U(f_N,P) - L(f_N,P) < \varepsilon/3$. Applying the previous display with $n = N$ gives us

$$U(f,P) - L(f,P) < \varepsilon$$

and since $\varepsilon > 0$ was arbitrary, f is Riemann integrable. By Theorems 2.16, 2.19,

$$\Big|\int_a^b f - \int_a^b f_n\Big| \leqslant \int_a^b |f_n - f| \leqslant (b-a) \sup_{x\in[a,b]} |f_n(x) - f(x)|$$

and uniform convergence implies that $\int_a^b f = \lim_{n\to\infty} \int_a^b f_n$. □

2.93. Exercise. Suppose that $f : [a,b] \to \mathbb{R}$ is *piecewise continuous* in the sense that for all $x_0 \in (a,b]$ the left limit $\lim_{x\to x_0-} f(x)$ exists, for all $x_0 \in [a,b)$ the right limit $\lim_{x\to x_0+} f(x)$ exists and there are only finitely many x_0 at which one of these limits is different from $f(x_0)$. Show that f is Riemann integrable on $[a,b]$.

2.94. Theorem. *Let (f_n) be a sequence of functions differentiable on $[a,b]$ such that f_n' is continuous on $[a,b]$ for every $n \in \mathbb{N}$. Assume there is $x_0 \in [a,b]$ such that $(f_n(x_0))_{n\in\mathbb{N}}$ converges and that (f_n') converges uniformly on $[a,b]$ to a function g. Then $(f_n)_{n\in\mathbb{N}}$ converges uniformly to a function f, and f is differentiable with $f'(x) = g(x)$ for all $x \in [a,b]$.*

Proof. By the fundamental theorem of calculus, we have

$$f_n(x) - f_n(x_0) = \int_{x_0}^x f_n'(t)\,dt.$$

The right hand side converges uniformly to $\int_{x_0}^x g(t)\,dt$, and letting

$$c = \lim_{n\to\infty} f_n(x_0)$$

we see that f_n converges uniformly to $f(x) = c + \int_{x_0}^x g(t)\,dt$. Since g is continuous, the function f is differentiable and $f' = g$. □

2.95. Example. If $f_n \to f$ uniformly and f_n is differentiable on $[a,b]$, then this does *not* imply that f is differentiable. Indeed, we will see in Ch. 7 that all continuous functions on a closed bounded interval can be uniformly approximated by smooth functions (in fact by polynomials). Here we just consider one concrete example. Let the function $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 0 & \text{if } x \leqslant 0, \\ x & \text{if } x > 0. \end{cases}$$

f is not differentiable at 0. We would like to approximate f by differentiable functions f_n. The idea is to replace f on $[0, \frac{1}{n}]$ by a cubic polynomial p_n which satisfies

$$p_n(0) = 0, p_n'(0) = 0, p_n(\tfrac{1}{n}) = \tfrac{1}{n}, p_n'(\tfrac{1}{n}) = 1.$$

A calculation shows that the polynomial

$$p_n(x) = 2nx^2 - n^2x^3$$

has these properties. Then define

$$f_n(x) = \begin{cases} 0 & \text{if } x \leqslant 0, \\ p_n(x) & \text{if } 0 < x \leqslant \frac{1}{n}, \\ x & \text{if } x > \frac{1}{n}. \end{cases}$$

Then f_n is differentiable on $\mathbb{R}$ and f_n' is continuous. Further, $f - f_n$ is nonnegative and has a maximum at $\frac{1}{3n}$. This shows

$$0 \leqslant f(x) - f_n(x) \leqslant \tfrac{4}{27n} \quad \text{for all } x \in \mathbb{R}$$

and thus f_n converges to f uniformly on $\mathbb{R}$.

2.96. Exercise. Give an example of a sequence $(f_n)_{n\in\mathbb{N}}$ of continuous bounded functions on $\mathbb{R}$ that converges pointwise to some function f such that f is unbounded and not continuous.

2.97. Exercise. Let $a \in \mathbb{R}$ and let $f_n : [0,1] \to \mathbb{R}$ be defined by

$$f_n(x) = \frac{n^a x}{(1+n^2x^2)^2}.$$

(i) For which choices of a and x is $(f_n(x))_{n\in\mathbb{N}}$ convergent? In the case of convergence determine the limit. In case of convergence what can you say about continuity and differentiability of the limiting function?
(ii) For which choices of a is (f_n) uniformly convergent on $[0,1]$?
(iii) For which choices of a is $\int_0^1 \lim_{n\to\infty} f_n(x)dx = \lim_{n\to\infty} \int_0^1 f_n(x)dx$?
(iv) For which choices of a and x does $\lim_{n\to\infty} f_n'(x)$ exist?
(v) For which choices of a and x is $(\lim_{n\to\infty} f_n(x))' = \lim_{n\to\infty} f_n'(x)$?

2.98. Exercise. Let $f_n(x) = n^{-1/2}\sin(nx)$.

(i) Show that f_n converges uniformly on $\mathbb{R}$ but for some $x \in \mathbb{R}$ the sequence $(f_n'(x))_{n\in\mathbb{N}}$ does not have a limit.
(ii*) Show that for every $x \in \mathbb{R}$, the sequence $(f_n'(x))_{n\in\mathbb{N}}$ does not have a limit.

2.9. Further exercises

2.99. Exercise. Let $m \in \mathbb{N}$. For $\delta \geqslant 0$ define

$$J(\delta) = \int_{-\delta}^{\delta} (1+x^m)^{-1} dx - 2\delta$$

Determine the asymptotic behavior of $J(\delta)$ as $\delta \to 0$: For which A do we have $J(\delta) = O(\delta^A)$ and for which B do we have $J(\delta) = o(\delta^B)$?

2.100. Exercise. Let $N \in \mathbb{N}$ and $a, b, \varepsilon > 0$. Compute each of the following limits:

(i) $\lim_{x\to\infty} \frac{(\log x)^N}{x}$
(ii) $\lim_{x\to\infty} e^{-x}x^N$
(iii) $\lim_{x\to 0}(\frac{1}{x} - \frac{1}{\sin x})$
(iv) $\lim_{x\to 0} \frac{x-\sin x}{x-\tan x}$
(v) $\lim_{x\to 1+} \frac{\sqrt{x-1}}{\sqrt{x}-1}$
(vi) $\lim_{x\to 0} \frac{a^x-b^x}{\tan x}$

2.101. Exercise. Compute each of the following limits:

(i) $\lim_{x\to 0} \frac{e^x+e^{-x}}{\cos x}$
(ii) $\lim_{x\to 0} x^{-1}\int_0^x \frac{\sin t}{t} dt$
(iii) $\lim_{x\to 0+} x^x$
(iv) $\lim_{x\to\infty} x^{1/x}$
(v) $\lim_{n\to\infty} \frac{e^{\sqrt{n}}}{n^5}$
(vi) $\lim_{\delta\to 0+} \delta^{\varepsilon} e^{\sqrt{\log \frac{1}{\delta}}}$

2.102. Exercise. Use Taylor's theorem to find the limits:

(i) $\lim_{x\to 0} \frac{x-\sin x}{x\sin^2 x}$

(ii) $\lim_{x\to 0} \left(\frac{1}{\sin x} - \frac{1}{x}\right)$

(iii) $\lim_{x\to 0} \frac{\sin(x^2)}{(\sin 3x)^2}$

2.103. Exercise. Determine for which A, B the quantities $\frac{x-\sin x}{x\sin^2 x}$, $\frac{1}{\sin x} - \frac{1}{x}$, $\frac{\sin(x^2)}{(\sin 3x)^2}$ are $O(|x|^A)$ and $o(|x|^B)$ as $x \to 0$.

2.104. Exercise. Let f be a continuous and strictly increasing function on $[a, b]$ and let $f(a) = c$, $f(b) = d$. Prove that

$$\int_c^d f^{-1} + \int_a^b f = bd - ac.$$

2.105. Exercise. Let $k \in \mathbb{N}$. Prove that

$$\lim_{n\to\infty} 2^{-n}\binom{n}{k} = 0.$$

2.106. Exercise. Let $(a_k)_{k\in\mathbb{N}}$ be a convergent sequence of complex numbers. The *arithmetic mean* of $a_1, \ldots, a_n$ is given by

$$A_n = \frac{a_1 + \cdots + a_n}{n}.$$

Prove: If $\lim_{k\to\infty} a_k = L$, then (A_n) converges to L.

2.107. Exercise. Suppose $\lim_{n\to\infty} a_n = A$. Prove that

$$\lim_{n\to\infty} 2^{-n} \sum_{k=0}^{n} \binom{n}{k} a_k = A.$$

2.108. Exercise. Let $(b_k)_{k=1}^{\infty}$ be a sequence of positive numbers. The *geometric mean* of $b_1, \ldots, b_n$ is given by

$$G_n = \sqrt[n]{b_1 \cdots b_n}.$$

(i) Prove: If $\lim_{k\to\infty} b_k = b$ and $b > 0$, then (G_n) converges to b.

(ii) Let $(a_k)_{k=1}^{\infty}$ be a sequence of positive numbers such that $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = b > 0$. Then $(\sqrt[n]{a_n})$ converges to b as $n \to \infty$.

2.109. Exercise. Determine the value of

$$\lim_{n\to\infty} \frac{n}{\sqrt[n]{n!}}.$$

2.110. Exercise. Let $a > 0$ and $d > 0$ and consider the *arithmetic progression*

$$a_n = a + (n-1)d, \quad n \in \mathbb{N}.$$

Denote by A_n and G_n the associated arithmetic and geometric means (as defined in Exercises 2.106 and 2.108). Show that $\lim_{n\to\infty} A_n/G_n$ exists and determine its value.

2.111. Exercise. For a function $f : [a, b] \to \mathbb{R}$ let

$$A_n(f) = \frac{1}{n} \sum_{k=1}^{n} f(a + k\tfrac{b-a}{n}).$$

(i) Prove: If f is continuous on $[a, b]$ then

$$\lim_{n\to\infty} A_n(f) = \frac{1}{b-a} \int_a^b f(x)\, dx$$

The limit is called the *arithmetic mean* of the function f over the interval $[a, b]$.

(ii) Show that the limit relation in (i) remains true if f is merely Riemann integrable over $[a,b]$

2.112. Exercise. Let $f:[a,b]\to\mathbb{R}$ be Riemann integrable such that $\inf_{x\in[a,b]} f(x)>0$. Let

$$c_{k,n}(f)=f(a+k\tfrac{b-a}{n}),\quad k=1,\dots,n.$$

Let $H_n(f)$ and $G_n(f)$ be the harmonic and geometric means of $\{c_{k,n}(f):k=1,\dots,n\}$. Prove:

$$\lim_{n\to\infty} H_n(f)=\frac{b-a}{\int_a^b \frac{1}{f(x)}\,dx},$$

$$\lim_{n\to\infty} G_n(f)=\exp\Big(\frac{1}{b-a}\int_a^b \log(f(x))\,dx\Big).$$

The limits are called the *harmonic and geometric means* of the function f over the interval $[a,b]$.

2.113. Exercise (Jensen's inequality). A function $f:(a,b)\to\mathbb{R}$ is called *convex* if for all $x,y\in(a,b)$ and all $\theta\in[0,1]$

$$f((1-\theta)x+\theta y)\leqslant(1-\theta)f(x)+\theta f(y).$$

(See Figure 2.6.) Show that f is convex if and only if for all collections of points $x_1,\dots,x_n\in(a,b)$ and $\theta_i\in[0,1]$ with $\sum_{i=1}^n\theta_i=1$,

$$f\Big(\sum_{i=1}^n\theta_i x_i\Big)\leqslant\sum_{i=1}^n\theta_i f(x_i).$$

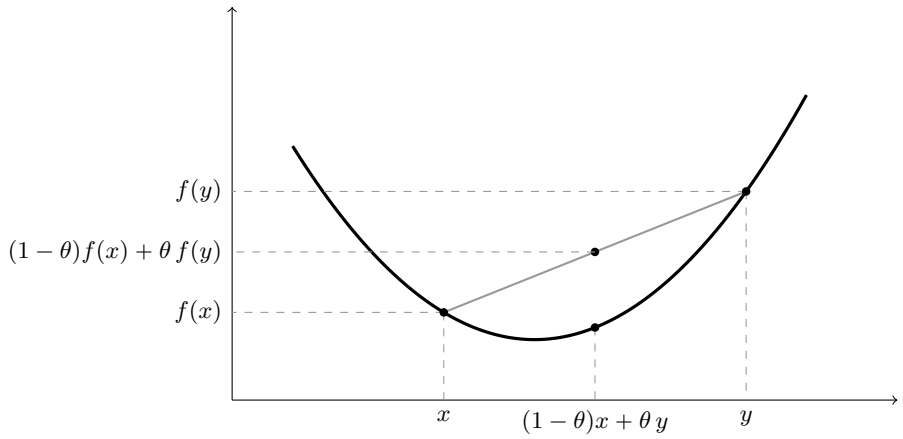


Figure 2.6. Convexity interpreted geometrically: the line connecting any two points on the graph lies on or above the graph.

2.114. Exercise. If f is differentiable on (a,b) and f' is increasing on (a,b) then f is convex on (a,b). In particular, if f is differentiable twice and $f''(t)\geqslant 0$ on (a,b) then f is convex on (a,b).

2.115. Exercise.

(i) Let $a_1,\dots,a_n$ be positive numbers and let A_n, G_n and H_n be the arithmetic, geometric and harmonic means of the a_k, respectively. Show that

$$H_n\leqslant G_n\leqslant A_n.$$

(ii) Let $f:[a,b]\to\mathbb{R}$ be Riemann integrable and $\inf_{x\in[a,b]} f(x)>0$. Prove that

$$\frac{b-a}{\int_a^b\frac{1}{f(x)}\,dx}\leqslant\exp\Big(\frac{1}{b-a}\int_a^b\log(f(x))\,dx\Big)\leqslant\frac{1}{b-a}\int_a^b f(x)\,dx.$$

2.116. Exercise. (i) Evaluate

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n}.$$

(ii) Let $a_{k,n} = \binom{n-1}{k-1}$ and let A_n be the arithmetic mean of $a_{1,n}, \ldots, a_{n,n}$.

Determine $\lim_{n\to\infty} \sqrt[n]{A_n}$.

(iii) Let G_n be the geometric mean of $a_{1,n}, \ldots, a_{n,n}$ and prove that $\lim_{n\to\infty} \sqrt[n]{G_n} = \sqrt{e}$.

2.117. Exercise. Let $b_k > 0$, $k = 1, \ldots, n$. Prove that

(i) The function

$$r \mapsto \Big(\frac{1}{n}\sum_{k=1}^{n} b_k^r\Big)^{1/r}$$

is increasing on $(-\infty, \infty)$.

(ii) Prove

$$\lim_{r\to\infty} \Big(\frac{1}{n}\sum_{k=1}^{n} b_k^r\Big)^{1/r} = \max\{b_k : k = 1, \ldots n\}$$

and

$$\lim_{r\to 0+} \Big(\frac{1}{n}\sum_{k=1}^{n} b_k^r\Big)^{1/r} = \exp\Big(\frac{1}{n}\sum_{k=1}^{n} \log b_k\Big).$$

2.118. Exercise. Show that

$$x\sqrt{\log(\tfrac{2}{x+y})} + y\sqrt{\log(\tfrac{2}{x+y})} \geqslant x\sqrt{\log(\tfrac{1}{x})} + y\sqrt{\log(\tfrac{1}{y})}$$

for all $x, y \in (0, 1)$.

2.119. Exercise (Leibniz rule). Let $I = (a, b)$, let f, g be n times differentiable functions on I. Then the nth derivative of the product is given by

$$(f \cdot g)^{(n)}(x) = \sum_{k=0}^{n} \binom{n}{k} f^{(k)}(x) g^{(n-k)}(x).$$

2.120. Exercise. Let g be a C^n function on (a, b) and let f be a C^n function defined on the range of g. Prove that $f \circ g$ is n-times differentiable.

2.121. Exercise (Faà di Bruno's formula). Prove the precise formula for the nth derivative of $f \circ g$:

$$\frac{d^n}{(dx)^n}[f(g(x))] = \sum \frac{n!}{m_1!m_2!\ldots m_n!} f^{(m_1+\cdots+m_n)}(g(x)) \cdot \prod_{j=1}^{n} \Big(\frac{g^{(j)}(x)}{j!}\Big)^{m_j},$$

where the sum is extended over all n-tuples of nonnegative integers $(m_1, \ldots, m_n)$ satisfying the constraint that $\sum_{j=1}^{n} j m_j = n$.

2.122. Exercise. Find the derivatives of the functions $f(x) = x^x$, $g(x) = x^{x^x}$ (defined for $x > 0$).

2.123. Exercise. Let $f(x) = x^2 \sin(\frac{1}{x})$, $f(0) = 0$. Show that f is differentiable everywhere but f' is not continuous at 0.

2.124. Exercise. Let $a \geqslant 0$ and $b > 0$. Let $f : [-1, 1] \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} |x|^a \sin(\frac{1}{|x|^b}) & \text{if } 0 < |x| \leqslant 1, \\ 0 & \text{if } x = 0. \end{cases}$$

(i) For which a, b is f continuous in $(-1, 1)$?
(ii) For which a, b does $f'(0)$ exist?
(iii) For which a, b is f' a bounded function?
(iv) For which a, b is f' a continuous function?

(v) For which a, b is the graph rectifiable?

2.125. Exercise (Darboux's theorem). Let f be differentiable on an open interval I and let $[a,b] \subset I$. Then for every η between $f'(a)$ and $f'(b)$ there exists a $\xi \in (a,b)$ such that $f'(\xi) = \eta$. Note: It is not assumed that f' is continuous and hence the intermediate value theorem does not apply.

2.126. Exercise. Let f be differentiable in $[a,b]$, with f' continuous in $[a,b]$ and $f(x) \neq 0$ on $[a,b]$. Prove that

$$\int_a^b \frac{f'(x)}{f(x)} dx = \log\Big(\frac{f(b)}{f(a)}\Big).$$

2.127. Exercise. Let f and g be differentiable at a, $g'(a) \neq 0$ and $f(a) = g(a) = 0$. Prove that

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}.$$

2.128. Exercise (l'Hôpital's rule). Let f, g be differentiable functions on (a,b) such that $g'(x) \neq 0$ on (a,b), $\lim_{x\to a+} f(x) = 0$ and $\lim_{x\to a+} g(x) = 0$. Show that

$$\text{If } \lim_{x\to a+} \tfrac{f'(x)}{g'(x)} = L, \text{ then } \lim_{x\to a+} \tfrac{f(x)}{g(x)} = L.$$

(Here L is either finite or $\pm\infty$.)

2.129. Exercise (Schwarz second derivative). Let $a \in \mathbb{R}$ and $\varepsilon > 0$. For a real-valued function f on $(a-\varepsilon, a+\varepsilon)$ we call the limit

$$\lim_{h\to 0} \frac{f(a+h) + f(a-h) - 2f(a)}{h^2}$$

the *Schwarz second derivative* of f at a (if it exists).

Let f be differentiable and assume that $f''(a)$ exists. Prove that the Schwarz second derivative exists and is equal to $f''(a)$.

2.130. Exercise. Let $a \in \mathbb{R}$ and

$$f(x) = \begin{cases} (x-a)^2 & \text{if } x \geqslant a, \\ -(x-a)^2 & \text{if } x < a. \end{cases}$$

Determine whether the Schwarz second derivative of f at a exists and whether $f''(a)$ exists.

2.131. Exercise. The following is a stronger version of Taylor's theorem in Theorem 2.50 ($f^{(n+1)}$ is not assumed to be continuous).

Let I be an open interval, $n \in \mathbb{N}_0$ and let $f \in C^n(I)$. Let $f^{(n)}$ be differentiable on I and let $a \in I$ and $a + t \in I$. Then there exists a ξ between a and $a + t$ such that

$$f(a+t) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} t^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} t^{n+1}.$$

Note that for $n = 0$ this is the mean value theorem.

Hints: Let $P_a(t) = \sum_{k=0}^n \frac{f^{(k)}(a)}{k!} t^k$. Define, for $t \neq 0$, $M(t) = t^{-n-1}(f(a+t) - P_a(t))$ and

$$G(s) = f(a+s) - P_a(s) - M(t)s^{n+1}.$$

Then $G^{(n+1)}(s) = f^{(n+1)}(a+s) - M(t)(n+1)!$ and the proof is done if one shows that $G^{(n+1)}$ has a zero between 0 and t.

To show this verify that $G(t) = 0$ and $G(0) = G'(0) = \cdots = G^{(n)}(0) = 0$. Note that by Rolle's theorem there exists a ξ_1 between 0 and t such that $G'(\xi_1) = 0$. Applying Rolle's theorem again there is a ξ_2 between 0 and ξ_1 such that $G''(\xi_2) = 0$. Iterating, we come to the conclusion that there exists a ξ_{n+1} between 0 and ξ_n (which is between 0 and t) for which $G^{(n+1)}(\xi_{n+1}) = 0$.

2.132. Exercise. Let $f : [0,\infty) \to \mathbb{R}$ be such that f is Riemann integrable on $[0,b]$ for all $b > 0$ and $\lim_{x\to\infty} f(x) = L$. Show that

$$\lim_{b\to\infty} \frac{1}{b} \int_0^b f(t)\,dt = L.$$

2.133. Exercise. Let f be of class C^2 on (b,∞) and let $M_0 = \sup_{x>b} |f(x)|$, $M_1 = \sup_{x>b} |f'(x)|$, and $M_2 = \sup_{x>b} |f''(x)|$.

(i) Prove *Landau's inequality*: $M_1^2 \leqslant 4M_0M_2$.

(ii) If $M_2 < \infty$ and $\lim_{x\to\infty} f(x) = 0$, then $\lim_{x\to\infty} f'(x) = 0$.

2.134. Exercise. Let f be Riemann integrable on $[a,b]$. Show there is a $\xi \in [a,b]$ such that $\int_a^\xi f(x)\,dx = \int_\xi^b f(x)\,dx$.

2.135. Exercise. Prove the *second mean value theorem for integrals:* Let f be a monotone differentiable function on $[a,b]$, and assume that f' is continuous. Let g be continuous on $[a,b]$. Then there exists a $\xi \in (a,b)$ such that

$$\int_a^b f(x)g(x)\,dx = f(a)\int_a^\xi g(x)\,dx + f(b)\int_\xi^b g(x)\,dx.$$

2.136. Exercise. Let f be a differentiable function on $I = (x_0 - r, x_0 + r)$ and let g be continuous on I. Show that if $f(x_0) = B$ and $f'(x) = g(x)f(x)$ on I, then $f(x) = Be^{G(x)}$ for $x \in I$, with $G(x) = \int_{x_0}^x g(t)dt$.

2.137. Exercise. Let f be a continuous function on $\mathbb{R}$ which satisfies $f(x+y) = f(x)f(y)$ for all $x, y \in \mathbb{R}$, and $f(0) = B$. Suppose that $f'(0)$ exists. Prove that there exists an $a > 0$ such that $f(x) = Ba^x$ for all $x \in \mathbb{R}$.

2.138. Exercise. Let $f : \mathbb{R} \to \mathbb{R}$ be twice differentiable everywhere. Assume that

$$f''(x) + f(x) = 0$$

for all $x \in \mathbb{R}$, and $f(0) = a$, $f'(0) = b$. Prove that $f(x) = a\cos x + b\sin x$.

2.139. Exercise. Let $a, B \in \mathbb{R}$. Prove that there is a unique function f which is differentiable on $(0,\infty)$ such that $f(1) = B$ and $xf'(x) = af(x)$ for all $x > 0$. Determine f.

2.140. Exercise. Integration by parts can be used to detect cancellation effects in integrals: For a large parameter $\lambda \geqslant 1$ and for $0 \leqslant a < b$ consider the integrals

$$\mathcal{I}_\lambda(a,b) = \int_a^b \cos(\lambda x^2)\,dx.$$

(i) Show that for all $b > a > 0$,

$$|\mathcal{I}_\lambda(a,b)| < 2(\lambda a)^{-1}.$$

(ii) Show that for all $b > a \geqslant 0$,

$$|\mathcal{I}_\lambda(a,b)| < \sqrt{8}\lambda^{-1/2}.$$

2.141. Exercise. Let I be an open interval, let χ be a C^∞ function such that $\chi(x) = 0$ in the complement of a closed bounded subinterval $[a,b]$. Let ϕ be a C^∞ function satisfying $\phi'(x) \neq 0$ for all $x \in I$. Let

$$\mathcal{J}(\lambda) = \int_I \chi(x)e^{i\lambda\phi(x)}\,dx.$$

Prove that, for every $N \in \mathbb{N}$, $\mathcal{J}(\lambda) = O(\lambda^{-N})$, as $\lambda \to \infty$.

Suggestion: For this, prove using integration by parts

$$\mathcal{J}(\lambda) = \Big(\frac{i}{\lambda}\Big)^N \int_I e^{i\lambda\phi(x)}\mathcal{L}^N[\chi](x)\,dx,$$

where

$$\mathcal{L}[\chi](x) = \frac{d}{dx}\Big(\frac{\chi(x)}{\phi'(x)}\Big) = \frac{\chi'(x)}{\phi'(x)} - \frac{\chi(x)\phi''(x)}{(\phi'(x))^2}$$

and $\mathcal{L}^N[\chi] = \mathcal{L}[\mathcal{L}^{N-1}[\chi]]$. Argue by induction that $\mathcal{L}^N[\chi]$ is a C^∞-function.

2.142. Exercise. In the situation of Exercise 2.141 let $\mathcal{L}[\chi](x) = \frac{d}{dx}\Big(\frac{\chi(x)}{\phi'(x)}\Big)$. In some instances (in particular when ϕ' is small) it may be useful to have a more explicit description of $\mathcal{L}^N[\chi]$.

Prove that

$$\mathcal{L}^N[\chi](x) = \frac{\chi^{(N)}(x)}{(\phi'(x))^N} + \sum_{\substack{(j,\kappa_1,\dots,\kappa_M):\\ j+\sum_{l=1}^M \kappa_l = N}} c_{j,\kappa_1,\dots,\kappa_M} \frac{\chi^{(j)}(x)}{(\phi'(x))^j} \prod_{l=1}^M \frac{\phi^{(\kappa_l+1)}(x)}{(\phi'(x))^{\kappa_l+1}}$$

for suitable constants $c_{j,\kappa_1,\dots,\kappa_M}$ where $(j,\kappa_1,\dots,\kappa_M)$ are $M+1$ tuples of integers satisfying $j \geqslant 0$, $\kappa_l \geqslant 1$ and subject to the condition $j + \sum_{l=1}^M \kappa_l = N$.

2.143. Exercise. Let $f : [0,1] \to \mathbb{R}$ be continuous. Show that

$$\lim_{a\to 0+} a \int_a^1 \frac{f(x)}{x^2}\,dx$$

exists and evaluate it.

2.144. Exercise. Determine for which real α, β, γ each of the following improper integrals exists:

(i) $\int_{0+}^1 x^{-\alpha}\,dx$

(ii) $\int_{0+}^{1/2} x^{-1}(\log\frac{1}{x})^{-\alpha}\,dx$

(iii) $\int_{0+}^{1/2} x^{-1}(\log\frac{1}{x})^{-\beta}\,dx$

(iv) $\int_{0+}^{1/100} x^{-1}(\log\frac{1}{x})^{-1}(\log\log\frac{1}{x})^{-\gamma}\,dx$

2.145. Exercise. Determine for which real α, β, γ each of the following improper integrals exists:

(i) $\int_1^\infty x^{-\alpha}\,dx$

(ii) $\int_2^\infty x^{-1}(\log x)^{-\beta}\,dx$

(iii) $\int_{100}^\infty x^{-1}(\log x)^{-1}(\log\log x)^{-\gamma}\,dx$

2.146. Exercise. Prove that the *Fresnel integrals*

$$\int_0^\infty \sin(x^2)\,dx \quad \text{and} \quad \int_0^\infty \cos(x^2)\,dx$$

exist as improper integrals.

2.147. Exercise. Which of the following improper integrals exist?

(i) $\int_0^\infty \frac{\cos x}{\sqrt{x}}dx$

(ii) $\int_0^\infty \frac{|\cos x|}{\sqrt{x}}\,dx$

(iii) $\int_0^\infty x\sin(x)\,dx$

(iv) $\int_0^\infty x\sin(x^2)\,dx$

(v) $\int_0^\infty x\sin(x^3)\,dx$

(vi) $\int_0^\infty \sin(x)\,dx$

(vii) $\int_0^1 x^{-5/2}\sin(\frac{1}{x})\,dx$

(viii) $\int_0^\infty e^x\sin(e^{2x})\,dx$

2.148. Exercise. Prove

$$\int_0^\infty \frac{\sin x}{1+x}dx + \int_0^\infty \frac{\cos x}{(1+x)^2}\,dx = 1.$$

2.149. Exercise. Does there exist a continuous function f on $[0,\infty)$ so that $f(x) \geqslant 0$ for all $x \geqslant 0$, $\limsup_{x\to\infty} e^{-e^x} f(x) = \infty$ and $\int_0^\infty f(x)\,dx$ exists as an improper Riemann integral? Either disprove or give an example.

2.150. Exercise. Suppose that for every $a, b \in (0,\infty)$ the function $x \mapsto \frac{f(x)}{x}$ is Riemann integrable on $[a,b]$ and assume that $\lim_{x\to 0+} f(x) = A$, $\lim_{x\to\infty} f(x) = B$.

(i) Prove that for $r > 0$, $s > 0$

$$\int_0^\infty \frac{f(rx) - f(sx)}{x}\,dx = (A - B)\log(s/r)$$

(ii) Evaluate

$$\int_0^\infty \frac{e^{-rx} - e^{-sx}}{x}\,dx.$$

2.151. Exercise. Prove for $r, s > 0$

$$\int_0^\infty \frac{\cos(rx) - \cos(sx)}{x}\,dx = \log(s/r).$$

CHAPTER 3

Sums and series

3.1. Convergence of series

Let $(a_n)_{n=0}^{\infty}$ denote a sequence of complex numbers. We wish to study infinite sums (series), written as $\sum_{n=0}^{\infty} a_n$. The first step is to give a precise meaning to these symbols.

3.1. Definition. We say that the *series*

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

is *convergent* with sum $S \in \mathbb{C}$ if the sequence $(s_n)_{n=0}^{\infty}$ of *partial sums*

$$s_n = \sum_{k=0}^{n} a_k = a_0 + a_1 + \cdots + a_n$$

converges to S as $n \to \infty$. In this case we write

$$\sum_{k=0}^{\infty} a_k = S.$$

If $(s_n)_{n=0}^{\infty}$ does not converge, then we say that $\sum_n a_n$ *diverges*.

Changing a finite number of terms changes the value of an infinite series but it does not change whether a series converges or not. In particular, for issues of convergence it is not relevant where the sum starts. The notation $\sum_n a_n$ is sometimes used when the limits of the sum are implied from the context or not relevant.

3.2. Example. If $|z| < 1$, then the *geometric series* $\sum_{n=0}^{\infty} z^n$ is convergent with

$$\sum_{n=0}^{\infty} z^n = (1-z)^{-1}. \tag{3.1}$$

This follows from the formula

$$(1-z)\sum_{k=0}^{n} z^k = 1 - z^{n+1} \tag{3.2}$$

and $\lim_{n\to\infty} z^{n+1} = 0$ for $|z| < 1$.

3.3. Theorem (Cauchy criterion). *The series $\sum_{k=0}^{\infty} a_k$ is convergent if and only if for every $\varepsilon > 0$ there exists an $N = N(\varepsilon) \in \mathbb{N}$ such that for all $n > m \geqslant N$,*

$$\Big| \sum_{k=m+1}^{n} a_k \Big| < \varepsilon.$$

Proof. Let $s_n = \sum_{k=0}^{n} a_k$. Then for $n > m$,

$$s_n - s_m = \sum_{k=m+1}^{n} a_k.$$

Thus, (s_n) is a Cauchy sequence. Cauchy sequences of complex numbers are convergent (see §1.3). □

The 'if' part in the Cauchy criterion, with $m = n - 1$, implies the following necessary condition.

3.4. Corollary. *If $\sum_{n=0}^{\infty} a_n$ converges, then $\lim_{n\to\infty} a_n = 0$.*

The converse of this statement fails (see e.g. Example 3.11 below). We also get an analogous statement for the tails of the series:

3.5. Corollary. *Suppose $\sum_{n=0}^{\infty} a_n$ converges and let $r_n = \sum_{k=n+1}^{\infty} a_k$. Then*

$$\lim_{n\to\infty} r_n = 0.$$

Proof. If $s_n = \sum_{k=0}^{n} a_k$ and $s = \lim_{n\to\infty} s_n$ we have $r_n = s - s_n$ and r_n converges to $s - s = 0$. □

From basic results about convergence of sequences, applied to the partial sums of a series we also get a rule about linear combinations of series (prove it!).

3.6. Observation. *If $\sum_{n=0}^{\infty} a_n = A$, $\sum_{n=0}^{\infty} b_n = B$ and if $\lambda, \mu \in \mathbb{C}$, then*

$$\sum_{n=0}^{\infty} (\lambda a_n + \mu b_n) = \lambda A + \mu B.$$

We will also sometimes consider *doubly infinite series* of the form

$$\sum_{n=-\infty}^{\infty} a_n$$

for a sequence of complex numbers $(a_n)_{n\in\mathbb{Z}}$. Such a series is considered convergent if each of the series $\sum_{n=0}^{\infty} a_{-n}$ and $\sum_{n=1}^{\infty} a_n$ converges (and its value is in this case the sum of the values of these two series).

3.1.1. Absolute convergence and comparison. A series $\sum_{n=0}^{\infty} a_n$ is said to *converge absolutely* if the series $\sum_{n=0}^{\infty} |a_n|$ converges. It is important to notice here that the summands of the series $\sum_{n=0}^{\infty} |a_n|$ are real and non-negative. If this series converges we also write

$$\sum_{n=0}^{\infty} |a_n| < \infty.$$

The divergence of a non-negative series is equivalent to

$$\lim_{n\to\infty} \sum_{k=0}^{n} |a_k| = \infty$$

(which we also write as $\sum_{k=0}^{\infty} |a_k| = \infty$).

3.7. Theorem. *Let $(a_n)_{n=0}^{\infty}$ be a sequence of complex numbers. Suppose that $\sum_{n=0}^{\infty} a_n$ converges absolutely. Then $\sum_{n=0}^{\infty} a_n$ converges. Moreover, the* triangle inequality *holds:*

$$\Big|\sum_{n=0}^{\infty} a_n\Big| \leqslant \sum_{n=0}^{\infty} |a_n| \tag{3.3}$$

Proof. By the triangle inequality for complex numbers,

$$\Big|\sum_{m<k\leqslant n} a_k\Big| \leqslant \sum_{m<k\leqslant n} |a_k|.$$

Now the convergence of $\sum_{n=0}^{\infty} a_n$ follows from Theorem 3.3. The inequality (3.3) also follows from the triangle inequality for finite sums,

$$\Big|\sum_{k=0}^{n} a_k\Big| \leqslant \sum_{k=0}^{n} |a_k|.$$

Taking $n \to \infty$ on both sides and using continuity of the absolute value concludes the proof. □

3.8. Corollary (Comparison test). *Let $(b_n)_{n=0}^{\infty}$ be a sequence of nonnegative numbers such that $\sum_n b_n$ converges. Let $(a_n)_{n=0}^{\infty}$ be a sequence of complex numbers, and assume that $|a_n| \leqslant b_n$ for all except possibly finitely many n. Then $\sum_{n=0}^{\infty} a_n$ converges absolutely.*

Proof. Let $N \in \mathbb{N}$ be such that $|a_n| \leqslant b_n$ for all $n > N$. For those n,

$$\sum_{k=m+1}^{n} |a_k| \leqslant \sum_{k=m+1}^{n} b_k$$

and the assertion follows from the Cauchy criterion (Theorem 3.3). □

The most familiar convergence tests are based on a comparison with a geometric series, i.e. with $b_n = c^n$ for $0 < c < 1$.

3.9. Observation (Root test). *For $a_n \in \mathbb{C}$ the following hold:*

(i) If $\limsup_{n\to\infty} |a_n|^{1/n} < 1$, then $\sum_{n=0}^{\infty} a_n$ converges absolutely.
(ii) If $\liminf_{n\to\infty} |a_n|^{1/n} > 1$, then $\sum_{n=0}^{\infty} a_n$ diverges.

Proof.

(i) By assumption $\limsup_{n\to\infty} |a_n|^{1/n} = 1 - \varepsilon$ with $\varepsilon > 0$. Then there exists N such that $|a_n|^{1/n} \leqslant 1 - \varepsilon/2$ for $n \geqslant N$. For those n, $|a_n| \leqslant (1 - \varepsilon/2)^n$ and the result follows by applying the comparison test (Corollary 3.8, with $b_n = (1 - \varepsilon/2)^n$).

(ii) By assumption $\liminf_{n\to\infty} |a_n|^{1/n} = 1 + \varepsilon$ with $\varepsilon > 0$. Then there exists N such that $|a_n|^{1/n} \geqslant 1 + \varepsilon/2$ for $n \geqslant N$. Hence $|a_n| \geqslant (1 + \varepsilon/2)^n$ for such n. We have $\lim_{n\to\infty} |a_n| = \infty$ in violation of the necessary condition for convergence in Corollary 3.4.

□

3.10. Observation (Ratio test). *Given a sequence $(a_n)_{n=0}^{\infty}$ of complex numbers with $a_n \neq 0$, the following hold:*

(i) If $\limsup_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} < 1$, then $\sum_{n=0}^{\infty} a_n$ converges absolutely.
(ii) If $a_n > 0$ for all n and there is $N > 0$ such that $\frac{a_{n+1}}{a_n} \geqslant 1$ for $n \geqslant N$, then $\sum_{n=0}^{\infty} a_n$ diverges.

Proof.

(i) If $\limsup_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} = 1 - \varepsilon$ with $\varepsilon > 0$ there is N such that $a_n \neq 0$ and $\frac{|a_{n+1}|}{|a_n|} \leqslant 1 - \varepsilon/2$ for $n \geqslant N$. Then for $n \geqslant N + 1$ we have

$$|a_n| = |a_N| \prod_{k=N}^{n-1} \frac{|a_{k+1}|}{|a_k|} \leqslant |a_N|(1 - \varepsilon/2)^{n-N-1}$$

and the comparison test involving the geometric series shows absolute convergence.

(ii) If $\frac{a_{n+1}}{a_n} \geqslant 1$ for $n \geqslant N$, then $a_k \geqslant a_N$ for $k > N$ and thus (a_n) does not converge to 0. Hence $\sum_k a_k$ does not converge.

□

Remark. The ratio test is often easier to apply than the root test, but the root test is somewhat stronger, see Exercise 3.126 below. In many cases a direct comparison can be more effective than the application of convergence tests.

In the following example we consider convergence of the series $\sum_{n=1}^{\infty} n^{-p}$ for $p \in \mathbb{R}$. This cannot be settled by direct termwise comparison with a geometric series (because $\lim_{n\to\infty} n^{-p/n} = 1$).

3.11. Observation. *Let $p \in \mathbb{R}$. The series $\sum_{n=1}^{\infty} n^{-p}$ converges if and only if $p > 1$.*

PROOF. When $p \leqslant 0$, we have divergence by Corollary 3.4. Therefore, it suffices to consider the case $p > 0$. The idea is to decompose the partial sums into subsums where all the terms are comparable. Specifically, this is the case for the sums with $2^k \leqslant n < 2^{k+1}$. We have for $k = 0, 1, 2, \ldots$ the estimate

$$2^{-p}2^{k(1-p)} \leqslant \sum_{n=2^k}^{2^{k+1}-1} n^{-p} < 2^{k(1-p)}.$$

Here the lower bound was obtained by replacing n^{-p} with $(2^{k+1})^{-p}$ and the upper bound by replacing n^{-p} with $(2^k)^{-p}$. Consider the case $p > 1$. The partial sums $s_N = \sum_{n=1}^{N} n^{-p}$ are increasing and dominated by

$$\sum_{k=0}^{\infty} 2^{(1-p)k} = (1 - 2^{1-p})^{-1},$$

since $2^{1-p} < 1$ (see Example 3.2). Thus the series converges for $p > 1$. For $p = 1$ we get for every $M \in \mathbb{N}$,

$$\sum_{n=1}^{2^M} n^{-p} \geqslant \tfrac{1}{2}M \tag{3.4}$$

and this implies divergence of $\sum_{n=1}^{\infty} n^{-p}$ for $0 < p \leqslant 1$ (and we can specify larger lower bounds of the form $c_p 2^{M(1-p)}$ for $0 < p < 1$). □

3.12. EXERCISE. Prove Observation 3.11 by a comparison with integrals.

REMARK. For $p > 1$ the value of the series $\sum_{k=1}^{\infty} k^{-p}$ is usually denoted by $\zeta(p)$. One can extend the function $p \to \zeta(p)$ to complex choices of p, and the extension is referred to as the *Riemann ζ-function*. It plays a crucial role in many different areas of mathematics, chiefly in number theory, concerning problems on the distribution of prime numbers.

3.13. EXERCISE. Prove that $\sum_{n=2}^{\infty} n^{-1}(\log n)^{-p}$ converges if and only if $p > 1$. Do this first by the method outlined in Example 3.11, and then again by the method in Exercise 3.12.

The divergent series $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$ is also known as *harmonic series*. The proof of Fact 3.11 with $p = 1$ shows that the partial sums of the harmonic series, $\sum_{k=1}^{n} \frac{1}{k}$, are comparable to $\log n$ for large n. The following exercise gives more precise information.

3.14. EXERCISE. Show that the limit

$$\gamma = \lim_{n\to\infty} \Big(\sum_{k=1}^{n} \frac{1}{k} - \log n \Big) \tag{3.5}$$

exists. Its value γ is called the *Euler–Mascheroni constant.*

REMARK. At the time of this writing it is unknown whether γ is an irrational number.

If the sequence (a_k) is monotone, then we can strengthen the necessary condition for convergence in Corollary 3.4.

3.15. EXERCISE. Assume that the sequence $(a_k)_{k=0}^{\infty}$ is monotone.

(i) Prove that if $\sum_{k=0}^{\infty} a_k$ converges, then $\lim_{k\to\infty} ka_k = 0$.
(ii) Give an example of a monotone sequence with $\lim_{k\to\infty} ka_k = 0$ so that $\sum_{k=1}^{\infty} a_k$ diverges.

The following exercise generalizes the method from the proof of Fact 3.11.

3.16. EXERCISE (Cauchy condensation test). Assume that $(a_n)_{n=1}^{\infty}$ is a decreasing sequence of positive numbers. Show that $\sum_{n=1}^{\infty} a_n$ converges if and only if $\sum_{k=0}^{\infty} 2^k a_{2^k}$ converges.

The comparison test raises the question whether there is a universal convergent series that could be used as a comparison for all other series with nonnegative terms. The following exercise shows that there can be no such universal test.

3.17. Exercise. Let $(a_n)_{n=0}^{\infty}$, $(b_n)_{n=0}^{\infty}$ be sequences with nonnegative terms such that $\sum_{n=0}^{\infty} a_n$ converges and $\sum_{n=0}^{\infty} b_n = \infty$. Prove the following statements:

(i) There is a sequence $(M_n)_{n=0}^{\infty}$ of positive numbers such that $\lim_{n\to\infty} M_n = \infty$ and $\sum_{n=0}^{\infty} a_n M_n$ converges.

(ii) There is a sequence $(\varepsilon_n)_{n=0}^{\infty}$ of positive numbers such that $\lim_{n\to\infty} \varepsilon_n = 0$ and $\sum_{n=0}^{\infty} b_n \varepsilon_n$ diverges.

For more sophisticated comparison tests with a systematic treatment, see Knopp [**37, 38**] We now discuss a test for series $\sum_{k=0}^{\infty} a_k$ with nonzero terms for which $\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = 1$ (as in examples above where the ratio test is inconclusive).

3.18. Theorem (Raabe's test). *Let $(a_n)_{n=0}^{\infty}$ be a sequence with positive terms.*

(i) Suppose

$$\limsup_{n\to\infty} n\left(\frac{a_{n+1}}{a_n} - 1\right) < -1.$$

Then $\sum_{n=0}^{\infty} a_n$ is convergent.

(ii) Suppose that there is $N \geqslant 1$ such that $n(\frac{a_{n+1}}{a_n} - 1) \geqslant -1$ for all $n \geqslant N$. Then $\sum_{n=0}^{\infty} a_n$ is divergent.

Proof.

(i) The condition $\limsup_{n\to\infty} n(\frac{a_{n+1}}{a_n} - 1) < -1$ implies that there exist $b > 0$ and $N \in \mathbb{N}$ such that $n(\frac{a_{n+1}}{a_n} - 1) < -1 - b$ for $n \geqslant N$. This means

$$na_{n+1} < (n-1)a_n - ba_n \text{ for } n \geqslant N.$$

In particular na_{n+1} is decreasing for $n \geqslant N+1$ (it is also bounded below by zero). Therefore, the limit $\lim_{n\to\infty} na_{n+1} = \beta$ exists with $\beta \geqslant 0$. If we set

$$c_k = (k-1)a_k - ka_{k+1} \geqslant 0$$

we can calculate the telescoping sum

$$\sum_{k=N+1}^{n} c_k = Na_{N+1} - na_{n+1}$$

which converges to $Na_{N+1} - \beta$ as $n \to \infty$. Hence the series $\sum_{k=N+1}^{\infty} c_k$ converges. Since we have by construction that

$$0 \leqslant a_k \leqslant b^{-1}c_k$$

for $k \geqslant N+1$, the series $\sum_{n=0}^{\infty} a_n$ converges by the comparison test.

(ii) If $n(\frac{a_{n+1}}{a_n} - 1) \geqslant -1$ for all $n \geqslant N$ we get $na_{n+1} - (n-1)a_n \geqslant 0$ for $n \geqslant N$, so na_{n+1} is increasing for $n \geqslant N$. Therefore $a_{n+1} \geqslant n^{-1}Na_{N+1}$, and Fact 3.11 for $p = 1$ and the comparison test give the divergence of $\sum_{n=0}^{\infty} a_n$.

□

For relatives of Raabe's test, see exercises in §3.6.

3.1.2. Rearrangements and unconditional convergence. We will consider the question of when infinite sums depend on the order in which their summands occur.

3.19. Definition. Let $\omega : \mathbb{N}_0 \to \mathbb{N}_0$ be a bijection. Given a convergent series $\sum_{n=0}^{\infty} a_n$ we call $\sum_{n=0}^{\infty} a_{\omega(n)}$ its *rearrangement* by ω. We say that a convergent series $\sum_{n=0}^{\infty} a_n$ converges *unconditionally* if all its rearrangements converge, with the same sum.

3.20. Theorem. *Let (a_n) be a sequence of complex numbers such that $\sum_{n=0}^{\infty} a_n$ is absolutely convergent. Then all rearrangements of $\sum_{n=0}^{\infty} a_n$ converge absolutely and $\sum_{n=0}^{\infty} a_n$ converges unconditionally.*

Proof. Consider a bijection $\omega : \mathbb{N}_0 \to \mathbb{N}_0$. Let $\varepsilon > 0$. Since $\sum_{n=0}^{\infty} |a_n|$ converges there exists $N = N(\varepsilon)$ such that $\sum_{n=N+1}^{\infty} |a_n| \leqslant \varepsilon$. Let $M = M(\varepsilon)$ be large enough that

$$\{0, \dots, N\} \subset \{\omega(0), \dots, \omega(M)\}. \tag{3.6}$$

Then for all $L \geqslant M$,

$$\sum_{k=M+1}^{L} |a_{\omega(k)}| \leqslant \sum_{k=N+1}^{\infty} |a_k| < \varepsilon.$$

Hence the partial sums of the series $\sum_{n=0}^{\infty} |a_{\omega(n)}|$ form a Cauchy sequence which implies that $\sum_{n=0}^{\infty} a_{\omega(n)}$ converges absolutely. It remains to prove unconditional convergence. Let $S = \sum_{n=0}^{\infty} a_n$. Then for $n > M$,

$$\begin{aligned}\Big| \sum_{k=0}^{n} a_{\omega(k)} - S \Big| &= \Big| \sum_{k=0}^{n} a_{\omega(k)} - \sum_{\nu=0}^{N} a_\nu - \sum_{\nu=N+1}^{\infty} a_\nu \Big| \\ &\leqslant \Big| \sum_{k=0}^{n} a_{\omega(k)} - \sum_{\nu=0}^{N} a_\nu \Big| + \sum_{\nu=N+1}^{\infty} |a_\nu| \leqslant 2 \sum_{\nu=N+1}^{\infty} |a_\nu| \leqslant 2\varepsilon.\end{aligned}$$

Here we used (3.6) and $n > M$ so that the terms a_ν with $\nu \leqslant N$ cancel. Since ε was arbitrary this shows that $\sum_{k=0}^{\infty} a_{\omega(k)} = S$. □

The assumption of absolute convergence was crucial in the above proof. If it is dropped and replaced by mere convergence the conclusion in Theorem 3.20 no longer holds. We first consider an example.

3.21. Example. Consider the convergent alternating series

$$S = \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} = 1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + \tfrac{1}{5} - \tfrac{1}{6} + \tfrac{1}{7} + \cdots$$

(in Example 3.56 we shall see using Taylor's formula that $S = \log 2$). Let $a_n = \frac{(-1)^n}{n+1}$. Consider the rearrangement of this series given by

$$\sum_{n=0}^{\infty} b_n = 1 + \tfrac{1}{3} - \tfrac{1}{2} + \tfrac{1}{5} + \tfrac{1}{7} - \tfrac{1}{4} + \tfrac{1}{9} + \tfrac{1}{11} - \tfrac{1}{6} + \cdots$$

where two positive terms are followed by one negative term. We shall show that this rearranged series converges to $\frac{3}{2}S$ thus a different value from $\sum_{n=0}^{\infty} a_n$. We have $b_n = a_{\omega(n)}$ where $\omega : \mathbb{N}_0 \to \mathbb{N}_0$ is defined by

$$\omega(3L) = 4L, \qquad \omega(3L+1) = 4L+2, \qquad \omega(3L+2) = 2L+1,$$

for $L = 0, 1, 2, \ldots$. Let S_n denote the partial sums of this rearrangement. For $M \geqslant 1$ we examine S_{3M-1} and then S_{3M} and S_{3M+1}. We calculate

$$\begin{aligned}\sum_{n=0}^{3M-1} b_n &= \sum_{L=0}^{M-1} (b_{3L} + b_{3L+1} + b_{3L+2}) = \sum_{L=0}^{M-1} \left(\tfrac{1}{4L+1} + \tfrac{1}{4L+3} - \tfrac{1}{2L+2}\right) \\ &= \Big(\sum_{L=0}^{M-1} \tfrac{1}{4L+1} - \tfrac{1}{4L+2} + \tfrac{1}{4L+3} - \tfrac{1}{4L+4}\Big) + \Big(\sum_{L=0}^{M-1} \tfrac{1}{4L+2} - \tfrac{1}{2L+2} + \tfrac{1}{4L+4}\Big) \\ &= \sum_{n=1}^{4M} \tfrac{(-1)^{n+1}}{n} + \sum_{L=0}^{M-1} \tfrac{1}{2}\big(\tfrac{1}{2L+1} - \tfrac{1}{2L+2}\big) = \sum_{n=0}^{4M-1} \tfrac{(-1)^n}{n+1} + \tfrac{1}{2} \sum_{n=0}^{2M-1} \tfrac{(-1)^n}{n+1}\end{aligned}$$

and taking the limit as $M \to \infty$ we see that $\lim_{M\to\infty} S_{3M-1} = \frac{3}{2}S$. Since $\lim_{n\to\infty} b_n = 0$ we also get $\lim_{M\to\infty}(S_{3M-1} - S_{3M}) = 0$ and $\lim_{M\to\infty}(S_{3M-1} - S_{3M+1}) = 0$, hence (S_N) converges with limit $\sum_{n=0}^{\infty} a_{\omega(n)} = \frac{3}{2}S$.

3.22. Exercise. Let $\sum_{k=0}^{\infty} b_k$ be the rearrangement of $\sum_{k=0}^{\infty} \frac{(-1)^k}{k+1}$ where N positive terms are followed by M negative terms, *i.e.* $\sum_{k=0}^{\infty} b_k$ is given by

$$\begin{aligned}&1 + \tfrac{1}{3} + \cdots + \tfrac{1}{2N-1} - \tfrac{1}{2} - \tfrac{1}{4} - \cdots - \tfrac{1}{2M} \\ &+ \tfrac{1}{2N+1} + \cdots + \tfrac{1}{4N-1} - \tfrac{1}{2M+2} - \cdots - \tfrac{1}{4M} \\ &+ \tfrac{1}{4N+1} + \ldots + \tfrac{1}{6N-1} - \tfrac{1}{4M+2} - \cdots - \tfrac{1}{6M} - \ldots\end{aligned}$$

Show that $\sum_{k=0}^{\infty} b_k = \log(2) + \frac{1}{2}\log(N/M)$.

3.23. Exercise. Let $a_k = (-1)^k \frac{1}{k+1}$. Given the alternating series $\sum_{k=0}^{\infty} a_k$ construct a rearrangement with $\lim_{N\to\infty} \sum_{k=1}^{N} a_{\omega(k)} = \infty$.

Example 3.21 reflects the general case of a convergent series with real terms which is not absolutely convergent.

3.24. Theorem. *A convergent series $\sum_{n=0}^{\infty} a_n$ with $a_n \in \mathbb{R}$ is unconditionally convergent if and only if it is absolutely convergent.*

The 'if' part is Theorem 3.20. The 'only if' part is an immediate consequence of (for example) the following Theorem 3.25 which specifies for a convergent but not absolutely convergent series that we can achieve any value or make the rearrangement divergent (see also Exercise 3.27 for other erratic behavior). In view of Theorem 3.24 we also say that a convergent series which is not absolutely convergent, converges *conditionally*.

3.25. Theorem (Riemann). *Assume that $\sum_{n=0}^{\infty} a_n$ is a convergent series of real numbers which does not converge absolutely. Let $-\infty \leqslant L_1 \leqslant L_2 \leqslant \infty$. Then there exists a rearrangement $\sum_{n=0}^{\infty} a_{\omega(n)}$ such that*

$$\liminf_{n\to\infty} \sum_{k=0}^{n} a_{\omega(k)} = L_1, \qquad \limsup_{n\to\infty} \sum_{k=0}^{n} a_{\omega(k)} = L_2.$$

Sketch of proof. Before starting a construction it is helpful to verify the following observations about general series. Set

$$a_k^+ = \begin{cases} a_k & \text{if } a_k \geqslant 0 \\ 0 & \text{if } a_k < 0 \end{cases}, \qquad a_k^- = \begin{cases} -a_k & \text{if } a_k \leqslant 0 \\ 0 & \text{if } a_k > 0 \end{cases}.$$

Then $a_k = a_k^+ - a_k^-$ and $|a_k| = a_k^+ + a_k^-$. Also $a_k^+ = \frac{1}{2}(|a_k| + a_k)$ and $a_k^- = \frac{1}{2}(|a_k| - a_k)$. From this one verifies that a series $\sum_{k=0}^{\infty} a_k$ converges absolutely if and only if both non-negative series $\sum_{k=0}^{\infty} a_k^+$ and $\sum_{k=0}^{\infty} a_k^-$ converge.

In the case of Theorem 3.25, $\sum_{k=0}^{\infty} a_k$ converges, but does not converge absolutely. Thus we must have that both $\sum_{k=0}^{\infty} a_k^+ = \infty$ and $\sum_{k=0}^{\infty} a_k^- = \infty$. This means $\sum_{k=P}^{\infty} a_k^+ = \infty$ and $\sum_{k=P'}^{\infty} a_k^- = \infty$ for all P, P'. We also have $\lim_{n\to\infty} a_n^{\pm} = 0$. These observations can now be used to construct a rearrangement with the required properties.

If $-\infty < L_1 \leqslant L_2 < \infty$ the idea is to start summing positive terms until the partial sums exceed the number L_2 for the first time. Then we continue subtracting negative terms until the partial sums go below L_1. Then we continue adding positive terms until the sums again exceed L_2 and then adding negative terms until they go again below L_1. One then continues this process. It is important to use $\lim_{n\to\infty} a_n = 0$ to verify the claim about the lim inf and lim sup of the partial sums.

It is instructive to fill in all the details of this argument, and modify the arguments when $L_2 = \infty$ or $L_1 = -\infty$.

3.26. Exercise. Extend Theorem 3.24 to series with complex terms.

3.27. Exercise. Let $\sum_{k=0}^{\infty} a_k$ be convergent but not absolutely convergent. Show that there exists a rearrangement $\sum_{k=0}^{\infty} a_{\omega(k)}$ with partial sums $S_n = \sum_{k=0}^{n} a_{\omega(k)}$ such that for every $L \in \mathbb{R}$ there is a subsequence of $(S_n)_{n=1}^{\infty}$ which converges to L.

3.1.3. Multiplication of series. Let us for the moment consider the product of two finite sums of complex numbers:

$$\Big(\sum_{k=0}^{N} a_k\Big)\Big(\sum_{l=0}^{N} b_l\Big) = \sum_{0\leqslant k,l\leqslant N} a_k b_l.$$

Since this is a finite sum we may order the terms $a_k b_l$ in any way we like. For infinite sums we have already seen that the order of summation may influence whether or not a series converges and even its value. Thus it is not entirely clear how to define the product of two infinite series. It will turn out to be particularly convenient to order the sum of products $a_k b_l$ according to the value of $k + l$ (which ranges from 0 to $2N$):

$$\sum_{0\leqslant k,l\leqslant N} a_k b_l = a_0 b_0 + (a_0 b_1 + a_1 b_0) + (a_0 b_2 + a_1 b_1 + a_2 b_0) + \cdots + a_N b_N$$

If $a_k = 0$ for $k > N$ and $b_l = 0$ for $l > N$, then this sum equals $\sum_{n=0}^{2N} c_n$, where

$$c_n = \sum_{k=0}^{n} a_{n-k} b_k = \sum_{k=0}^{n} a_k b_{n-k}. \tag{3.7}$$

3.28. Definition. Given two convergent infinite series $\sum_{k=0}^{\infty} a_k$, $\sum_{l=0}^{\infty} b_l$, their *Cauchy product* is defined as the series $\sum_{n=0}^{\infty} c_n$ where c_n is given by (3.7).

Unfortunately, it is possible that the Cauchy product of two convergent series does not converge, as the following example shows.

3.29. Example. Let

$$a_k = (-1)^k (k+1)^{-1/4}, \quad b_l = (-1)^l (l+1)^{-1/4}.$$

Then the series $\sum_{k=0}^{\infty} a_k$ and $\sum_{l=0}^{\infty} b_l$ converge by the alternating series test (Exercise 3.47), but their Cauchy product $\sum_{n=0}^{\infty} c_n$ does not converge, because the coefficients c_n are unbounded. Indeed,

$$c_n = \sum_{k=0}^{n} a_k b_{n-k} = (-1)^n \sum_{k=0}^{n} (k+1)^{-1/4} (n-k+1)^{-1/4}$$

and therefore

$$|c_n| = \sum_{k=0}^{n} (k+1)^{-1/4} (n-k+1)^{-1/4} \geqslant \sum_{k=0}^{n} (n+1)^{-1/4} (n+1)^{-1/4} = (n+1)^{1/2}.$$

The situation can be rectified if both series converge absolutely.

3.30. THEOREM. *Suppose that the series $\sum_{k=0}^{\infty} a_k$, $\sum_{l=0}^{\infty} b_l$ converge absolutely. Then their Cauchy product converges absolutely and its value is given by*

$$\sum_{n=0}^{\infty} c_n = \Big(\sum_{k=0}^{\infty} a_k\Big)\Big(\sum_{l=0}^{\infty} b_l\Big). \tag{3.8}$$

PROOF. Define

$$M = \max\Big(\sum_{k=0}^{\infty} |a_k|, \sum_{l=0}^{\infty} |b_l|\Big).$$

By the triangle inequality,

$$\sum_{n=0}^{N} |c_n| = \sum_{n=0}^{N} \Big|\sum_{k=0}^{n} a_{n-k}b_k\Big| \leqslant \sum_{n=0}^{N}\sum_{k=0}^{n} |a_k||b_{n-k}| \leqslant \sum_{k=0}^{N} |a_k| \sum_{l=0}^{\infty} |b_l| \leqslant M^2.$$

Since $\sum_{n=0}^{N} |c_n|$ is increasing in N and bounded above we see that $\sum_{n=0}^{\infty} |c_n| \leqslant M^2 < \infty$, so the Cauchy product converges absolutely. To determine the value of $\sum_{n=0}^{\infty} c_n$ write

$$\sum_{n=0}^{2N} c_n = \Big(\sum_{k=0}^{N} a_k\Big)\Big(\sum_{k=0}^{N} b_k\Big) + R_N,$$

with

$$R_N = \sum_{k=N+1}^{2N} a_k \sum_{\ell=0}^{2N-k} b_\ell + \sum_{\ell=N+1}^{2N} b_\ell \sum_{k=0}^{2N-\ell} a_k.$$

We estimate the remainder term

$$|R_N| \leqslant \Big(\sum_{k=N+1}^{\infty} |a_k|\Big)M + \Big(\sum_{\ell=N+1}^{\infty} |b_\ell|\Big)M.$$

In view of the assumptions of absolute convergence of the series $\sum_k a_k$, $\sum_k b_k$ the sequences $\sum_{k=N+1}^{\infty} |a_k|$ and $\sum_{\ell=N+1}^{\infty} |b_\ell|$ converge to 0 and thus $\lim_{N\to\infty} R_N = 0$. □

The assumptions in the previous theorem can be weakened at the expense of a more sophisticated argument:

3.31. THEOREM (Mertens). *Suppose that the series $A = \sum_{n=0}^{\infty} a_n$ and $B = \sum_{n=0}^{\infty} b_n$ both converge and that at least one of them converges absolutely. Then their Cauchy product converges and $\sum_{n=0}^{\infty} c_n = AB$.*

PROOF. We may assume that $\sum_{n=0}^{\infty} b_n$ converges absolutely.

Let $A_n = \sum_{k=0}^{n} a_k$, $B_n = \sum_{k=0}^{n} b_k$ and $S_n = \sum_{k=0}^{n} c_k$. We write

$$\begin{aligned} S_n &= a_0b_0 + (a_0b_1 + a_1b_0) + \cdots + (a_0b_n + a_1b_{n-1} + \cdots + a_nb_0) \\ &= A_nb_0 + A_{n-1}b_1 + \cdots + A_0b_n \\ &= (A_n - A)b_0 + (A_{n-1} - A)b_1 + \cdots + (A_0 - A)b_n + AB_n \end{aligned}$$

Since $\lim_{n\to\infty} AB_n = AB$ it suffices to show that

$$\lim_{n\to\infty} \sum_{k=0}^{n} (A_k - A)b_{n-k} = 0. \tag{3.9}$$

To prove this let $\varepsilon > 0$. Let $V = \sum_{n=0}^{\infty} |b_n|$ and $U = |A| + \sup_{k\geqslant 0} |A_k|$. Since

$$\lim_{k\to\infty} A_k - A = 0$$

we can find M_0 such that $|A_k - A| < \varepsilon/(2V)$ for all $k \geqslant M_0$. Since $\sum_{k=0}^{\infty} |b_k|$ converges we can find M_1 such that

$$\sum_{k=M_1}^{\infty} |b_k| < \frac{\varepsilon}{2U}.$$

Now let $n > M_0 + M_1$. We split

$$\sum_{k=0}^{n} (A_k - A)b_{n-k} = \sum_{k=0}^{M_0-1} (A_k - A)b_{n-k} + \sum_{k=M_0}^{n} (A_k - A)b_{n-k}.$$

We can estimate the second term by

$$\Big| \sum_{k=M_0}^{n} (A_k - A)b_{n-k} \Big| \leqslant \sum_{k=M_0}^{n} \frac{\varepsilon}{2V} |b_{n-k}| \leqslant \frac{\varepsilon}{2V} V = \frac{\varepsilon}{2}.$$

In the first term we have $n - k \geqslant M_0 + M_1 - (M_0 - 1) \geqslant M_1$ and therefore

$$\Big| \sum_{k=0}^{M_0-1} (A_k - A)b_{n-k} \Big| \leqslant U \sum_{k=0}^{M_0-1} |b_{n-k}| \leqslant U \sum_{l=M_1}^{\infty} |b_l| < U \frac{\varepsilon}{2U} = \frac{\varepsilon}{2}.$$

Combining the two estimates we get $|\sum_{k=0}^{n} (A_k - A)b_{n-k}| < \varepsilon$ for $n > M_0 + M_1$. This implies (3.9) which finishes the proof. □

REMARK. In Example 3.29 we saw that convergence of the Cauchy product $\sum_n c_n$ may fail. One may wonder whether it is possible for $\sum_{n=0}^{\infty} c_n$ to converge to a value distinct from $(\sum_{n=0}^{\infty} a_n)(\sum_{n=0}^{\infty} b_n)$. This is not the case as we will see in Exercise 3.75 below.

3.1.4. Doubly indexed series. Let $(a_{m,n})_{(m,n)\in\mathbb{N}_0^2}$ be a double indexed sequence of complex numbers. We investigate the precise meaning of *doubly indexed series*

$$\sum_{m,n} a_{m,n}.$$

Such a sum is not problematic if m and n live in finite index sets E and F, respectively. In particular, it is irrelevant in which order the sum is computed:

$$\sum_{m\in E} \sum_{n\in F} a_{m,n} = \sum_{n\in F} \sum_{m\in E} a_{m,n}.$$

If we drop the finiteness assumption, convergence questions arise and the order of summation may matter. Assuming that for each $n \in \mathbb{N}_0$ the series $\sum_{m=0}^{\infty} a_{m,n}$ converges, the series

$$\sum_{n=0}^{\infty} \Big(\sum_{m=0}^{\infty} a_{m,n} \Big) \tag{3.10}$$

can be interpreted as usual: it converges if the sequence of partial sums $\sum_{n=0}^{N} \Big(\sum_{m=0}^{\infty} a_{m,n} \Big)$ converges as $N \to \infty$. In the same way, the double series

$$\sum_{m=0}^{\infty} \Big(\sum_{n=0}^{\infty} a_{m,n} \Big) \tag{3.11}$$

converges if for every $m \in \mathbb{N}_0$ the series $\sum_{n=0}^{\infty} a_{m,n}$ converges and the sequence of partial sums $\sum_{m=0}^{M} \Big(\sum_{n=0}^{\infty} a_{m,n} \Big)$ converges as $M \to \infty$. If the series (3.10) and (3.11) both converge, we would like to know when their values coincide. In other words, when can we interchange the order of summation? A first observation is that this is always the case if $a_{m,n} \geqslant 0$ for all m, n since all partial sums are then monotone.

3.32. THEOREM. *Let $a_{m,n} \geqslant 0$ for all m, n. Then*

$$\sum_{n=0}^{\infty} \Big(\sum_{m=0}^{\infty} a_{m,n} \Big) = \sum_{m=0}^{\infty} \Big(\sum_{n=0}^{\infty} a_{m,n} \Big)$$

(with the usual convention that the value may equal ∞).

PROOF. Since $a_{m,n} \geqslant 0$, all involved partial sum sequences are monotone increasing, which implies that both sides of the equation in question are equal to

$$\sup \Big\{ \sum_{n=0}^{N} \sum_{m=0}^{M} a_{m,n} \,:\, N, M \in \mathbb{N} \Big\}. \qquad \square$$

In particular, for complex numbers $(a_{m,n})$ we always have

$$\sum_{n=0}^{\infty} \Big(\sum_{m=0}^{\infty} |a_{m,n}| \Big) = \sum_{m=0}^{\infty} \Big(\sum_{n=0}^{\infty} |a_{m,n}| \Big),$$

including in the case when one of the two sides diverges (i.e. equals ∞). Thus it makes sense to *define* the non-negative doubly indexed series

$$\sum_{n,m \geqslant 0} |a_{m,n}|$$

as the value in the previous display (which may be ∞). Interchange of summation may fail when summing terms with sign changes as the next example shows.

3.33. EXAMPLE. Define $a_{m,n} = 1$ if $m = n$, $a_{m,n} = -1$ if $m = n + 1$, and $a_{m,n} = 0$ in all other cases. Then for every $n \in \mathbb{N}_0$, we have $\sum_{m=0}^{\infty} a_{m,n} = 0$ and thus

$$\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} a_{m,n} = 0.$$

On the other hand, for $m \geqslant 1$ we have $\sum_{n=0}^{\infty} a_{m,n} = 0$ but for $m = 0$, $\sum_{n=0}^{\infty} a_{0,n} = a_{0,0} = 1$. Hence

$$\sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{m,n} = 1.$$

The problem disappears when we add an appropriate absolute convergence assumption.

3.34. THEOREM. *Let $(a_{m,n})_{(m,n) \in \mathbb{N}_0^2}$ be a doubly indexed sequence of complex numbers and assume that*

$$\sum_{n,m \geqslant 0} |a_{m,n}| < \infty.$$

Then

$$\sum_{n=0}^{\infty} \Big(\sum_{m=0}^{\infty} a_{m,n} \Big) = \sum_{m=0}^{\infty} \Big(\sum_{n=0}^{\infty} a_{m,n} \Big) \tag{3.12}$$

and the series on both sides converge.

When the assumption of the theorem holds, we define the doubly indexed series $\sum_{n,m \geqslant 0} a_{m,n}$ as the value in (3.12).

PROOF. We apply Theorem 2.84 on interchanging limits of sequences to the doubly indexed sequence of partial sums $(\sum_{n=0}^{k} \sum_{m=0}^{l} a_{m,n})_{k,l \in \mathbb{N}}$. Indeed, the assumption implies that for all $k \in \mathbb{N}$, the sequence

$(\sum_{m=0}^{l}\sum_{n=0}^{k} a_{m,n})_{l=0}^{\infty}$ converges uniformly in k as $l \to \infty$ to the value

$$b_k = \sum_{m=0}^{\infty}\sum_{n=0}^{k} a_{m,n}$$

(this series converges absolutely for each k). To see this we use the triangle inequality:

$$\Big|\sum_{m=0}^{l}\sum_{n=0}^{k} a_{m,n} - b_k\Big| = \Big|\sum_{m>l}\sum_{n=0}^{k} a_{m,n}\Big| \leqslant \sum_{m>l}\sum_{n=0}^{\infty} |a_{m,n}|.$$

The right-hand side does not depend on k and converges to 0 as $l \to \infty$ since $\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}|a_{m,n}|$ converges. The same is true, by the same reasoning, with the roles of m, n interchanged, which verifies the hypothesis of Theorem 2.84. □

Remark. The proof shows that (3.12) also holds under the weaker, but somewhat more technical assumption that the sequences $(\sum_{n=0}^{k}\sum_{m=0}^{l} a_{m,n})_{l=0}^{\infty}$ converge uniformly in $k \in \mathbb{N}$ as $l \to \infty$ and that $\sum_{n=0}^{\infty} a_{m,n}$ converges for every $m \in \mathbb{N}$.

Remark. This theorem is sometimes referred to as Fubini's theorem (or Tonelli–Fubini theorem) in reference to a more general theorem in integration theory.

3.1.5. Uniform convergence. It is common to encounter series of the form $\sum_{n=0}^{\infty} f_n(x)$ with f_n functions depending on a parameter x. We are then often interested in examining the continuous dependence of the sum on the parameter, and in integrating and differentiating with respect to this parameter. To validate such operations the concept of *uniform convergence* (as in §2.8) is extremely useful. The idea is to simply apply results on uniform convergence for sequences to the partial sums of the series.

3.35. Definition. Given a sequence $(f_n)_{n=0}^{\infty}$ of complex-valued functions on a set X we say that $\sum_{n=0}^{\infty} f_n(x)$ *converges uniformly* on X if the sequence of partial sums $s_N(x)$ with

$$s_N(x) = \sum_{n=0}^{N} f_n(x)$$

converges uniformly on X.

3.36. Theorem (Weierstrass test). *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of complex-valued functions on a set X such that*

$$|f_n(x)| \leqslant C_n \text{ for all } x \in X,$$

where (C_n) is a sequence of nonnegative numbers. Assume that $\sum_{n=1}^{\infty} C_n$ converges. Then the series $\sum_{n=1}^{\infty} f_n(x)$ converges uniformly on X.

Proof. Let $s_m(x) = \sum_{n=1}^{m} f_n(x)$. For $\ell < m$ we observe the estimate,

$$|s_m(x) - s_\ell(x)| = \Big|\sum_{n=\ell+1}^{m} f_n(x)\Big| \leqslant \sum_{n=\ell+1}^{m} |f_n(x)| \leqslant \sum_{n=\ell+1}^{m} C_n.$$

Since $\sum_{n=1}^{\infty} C_n$ converges, $\sum_{n=\ell+1}^{m} C_n$ converges to zero as $\ell \to \infty$. Therefore the sequence of partial sums $(s_N(x))$ is uniformly Cauchy on X and converges by Theorem 2.89. □

We now rewrite various results on uniform convergence for sequences, so that they apply to uniform convergence of series.

3.37. THEOREM. *Let $f_n : [a,b] \to \mathbb{C}$ and assume that $F(x) = \sum_{n=0}^{\infty} f_n(x)$ converges uniformly on $[a,b]$. If f_n is continuous at $x_0 \in [a,b]$ for every n, then F is continuous at x_0. That is,*

$$\lim_{x\to x_0} \sum_{n=0}^{\infty} f_n(x) = \sum_{n=0}^{\infty} f_n(x_0).$$

3.38. THEOREM. *Let $f_n : [a,b] \to \mathbb{C}$ be Riemann integrable and assume that $F(x) = \sum_{n=0}^{\infty} f_n(x)$ converges uniformly on $[a,b]$. Then F is Riemann integrable on $[a,b]$ and, for $a \leqslant x \leqslant b$,*

$$\int_a^x F(t)\,dt = \sum_{n=0}^{\infty} \int_a^x f_n(t)\,dt.$$

3.39. THEOREM. *Let $f_n : [a,b] \to \mathbb{C}$ be differentiable in $[a,b]$ such that f_n' is continuous in $[a,b]$ for every $n \in \mathbb{N}$. Suppose that there is $x_0 \in [a,b]$ such that the series $\sum_{n=1}^{\infty} f_n(x_0)$ converges and that the formally differentiated series $\sum_{n=0}^{\infty} f_n'(x)$ converges uniformly on $[a,b]$.*

Then $F(x) = \sum_{n=0}^{\infty} f_n(x)$ converges uniformly and is differentiable on $[a,b]$ and for all $x \in [a,b]$,

$$F'(x) = \sum_{n=0}^{\infty} f_n'(x).$$

These are direct consequences of the corresponding theorems for sequences (Theorems 2.91, 2.92 and 2.94) applied to the sequence of partial sums $s_N(x) = \sum_{n=0}^{N} f_n(x)$.

3.40. EXAMPLE (Weierstrass' function). Let $A > 2$ be an integer and consider the series

$$F(x) = \sum_{n=0}^{\infty} 2^{-n} \cos(A^n x)$$

Since $|\cos(A^n x)| \leqslant 1$, the Weierstrass test shows that the series is uniformly convergent in x. Therefore F is a continuous function. This example illustrates that uniform convergence does not by itself give any conclusions about derivatives and whether the derivatives of the partial sums,

$$\sum_{n=0}^{N} 2^{-n} A^n \sin(A^n x)$$

converge.

A related, but different question is whether the continuous function F is differentiable. Suppose that A is large (imagine $A > 100$ for now) and divisible by 4. As we will see it is not too difficult to show that F is then nowhere differentiable. The realization that such functions exist marked an important milestone in the development of modern analysis.

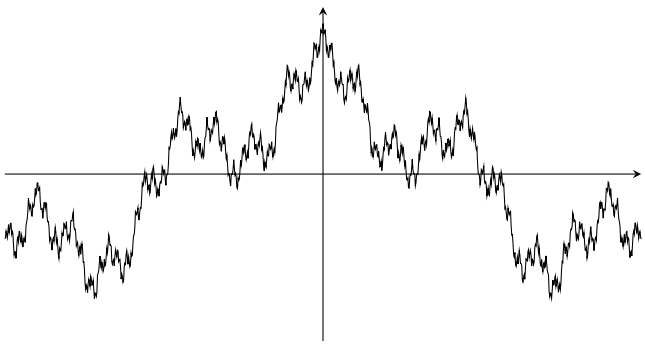


FIGURE 3.1. Weierstrass' function

The claim is that for every $x_0 \in \mathbb{R}$,

$$\limsup_{x \to x_0} \Big| \frac{F(x) - F(x_0)}{x - x_0} \Big| = \infty.$$

To show this we should control the signs of $\cos(A^n x_0)$ and $\cos(A^n x)$ in order to pinpoint enough destructive interference. Let $N \in \mathbb{N}$ be a (large) integer. Let us assume without loss of generality that

$$\cos(A^N x_0) \geqslant 0$$

(the same argument works for the case $\leqslant 0$, with appropriate changes of signs). Then there is $k \in \mathbb{Z}$ so that

$$|A^N x_0 - 2\pi k| \leqslant \tfrac{1}{2}\pi.$$

We define

$$x = x_0 + A^{-N}\delta$$

where

$$\delta = \begin{cases} \pi & \text{if } -\frac{3}{8}\pi \leqslant A^N x_0 - 2\pi k \leqslant \frac{3}{8}\pi, \\ \pi/2 & \text{if } \frac{3}{8}\pi \leqslant A^N x_0 - 2\pi k \leqslant \frac{1}{2}\pi, \\ -\pi/2 & \text{if } -\frac{1}{2}\pi \leqslant A^N x_0 - 2\pi k \leqslant -\frac{3}{8}\pi. \end{cases}$$

In the first case,

$$\cos(A^N x_0) - \cos(A^N x) = 2\cos(A^N x_0) \geqslant 2\cos(\tfrac{3}{8}\pi).$$

In the second case,

$$\tfrac{7}{8}\pi \leqslant A^N x - 2\pi k \leqslant \pi,$$

and since $\cos(x)$ is non-positive and decreasing on the interval $[\frac{7}{8}\pi, \pi]$,

$$\cos(A^N x_0) - \cos(A^N x) \geqslant -\cos(A^N x_0 + \pi/2) \geqslant -\cos(\tfrac{7}{8}\pi) = \sin(\tfrac{3}{8}\pi)$$

Similarly, in the third case also

$$\cos(A^N x_0) - \cos(A^N x) \geqslant \sin(\tfrac{3}{8}\pi).$$

Therefore, in each case

$$\cos(A^N x_0) - \cos(A^N x) \geqslant c_1$$

where $c_1 = \min(\sin(\frac{3}{8}\pi), 2\cos(\frac{3}{8}\pi)) > 0$. For $m \geqslant 1$ we have

$$\cos(A^{N+m} x) = \cos(A^{N+m} x_0 + A^m \delta) = \cos(A^{N+m} x_0)$$

since δ is a multiple of $\pi/2$ and A is a multiple of 4, so for $m \geqslant 1$,

$$\cos(A^{N+m} x_0) - \cos(A^{N+m} x) = 0.$$

Moreover, for every $n \in \mathbb{N}$ we may use the mean value theorem to estimate

$$|\cos(A^n x_0) - \cos(A^n x)| \leqslant A^n |x - x_0| \leqslant \pi A^{n-N}.$$

Thus,

$$\begin{aligned} |F(x_0) - F(x)| &\geqslant 2^{-N}(\cos(A^N x_0) - \cos(A^N x)) - \sum_{n=0}^{N-1} 2^{-n} |\cos(A^n x_0) - \cos(A^n x)| \\ &\geqslant 2^{-N} c_1 - \pi A^{-N} \sum_{n=0}^{N-1} (A/2)^n \end{aligned}$$

and thus with $|x - x_0| \geqslant \frac{1}{2}\pi A^{-N}$,

$$\Big| \frac{F(x) - F(x_0)}{x - x_0} \Big| \geqslant 2\pi^{-1} c_1 (A/2)^N - 2\frac{(A/2)^N - 1}{A/2 - 1} \geqslant (A/2)^N (2\pi^{-1} c_1 - \tfrac{2}{A/2 - 1})$$

If A is large ($A = 12$ works), then the right hand side tends to ∞ as $N \to \infty$. Since N was arbitrary the claim is proved. It is instructive to modify this argument to work for additional values of A.

3.41. Exercise. Let $\beta : \mathbb{R} \to \mathbb{R}$ denote the distance of x to the nearest integer, i.e. β is the continuous 1-periodic function which is defined on $[0,1]$ by

$$\beta(x) = \begin{cases} x & \text{if } 0 \leqslant x \leqslant 1/2, \\ 1-x & \text{if } 1/2 \leqslant x \leqslant 1. \end{cases}$$

The *van der Waerden function* is defined by

$$f(x) = \sum_{n=1}^{\infty} 2^{-n} \beta(2^n x)$$

Show that f is continuous but nowhere differentiable.

3.42. Example. Consider the series

$$\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{2k+1} = x - \frac{x^3}{3} + \frac{x^5}{5} - \dots.$$

It diverges for $|x| > 1$. We now argue that the series converges uniformly on $[-1,1]$ and identify the limit as

$$\sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{2k+1} = \arctan x \quad \text{for} \quad x \in [-1,1]. \tag{3.13}$$

Since $\arctan(1) = \pi/4$, this implies in particular that

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \dots, \tag{3.14}$$

a relation known as *Leibniz' formula.* Consider the partial sums

$$s_n(x) = \sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{2k+1}.$$

As there is substantial cancellation between consecutive positive and negative terms we examine the differences

$$d_\ell(x) = s_{2\ell+1}(x) - s_{2\ell-1}(x) = \frac{x^{4\ell+1}}{4\ell+1} - \frac{x^{4\ell+3}}{4\ell+3}.$$

Using calculus one verifies that on $[0,1]$ the maximum of d_ℓ occurs at $x = 1$. Moreover, $d_\ell(x) = -d_\ell(-x)$, so

$$\max_{x \in [-1,1]} |d_\ell(x)| = \frac{2}{(4\ell+1)(4\ell+3)}.$$

This is enough to deduce that the sequence $(s_{2\ell+1}(x))_\ell$ is uniformly Cauchy on $[-1,1]$ (why?), and in view of the estimate

$$|s_{2\ell+1}(x) - s_{2\ell}(x)| \leqslant \frac{1}{4\ell+3}$$

we get that also $(s_n(x))_n$ is uniformly Cauchy on $[-1,1]$. Since the s_n are continuous on $[-1,1]$ the uniform limit

$$s(x) = \lim_{n \to \infty} s_n(x)$$

is a well-defined continuous function (see Theorem 2.91). Let us now compute the value $s(x)$ for a given $x \in (-1,1)$ (for now excluding the endpoints). The idea is to write

$$\frac{x^{2k+1}}{2k+1} = \int_0^x t^{2k}\, dt$$

and interchange the limit in n and the (Riemann) integral using uniform convergence (see Theorem 2.92). This is justified since the geometric series

$$\sum_{k=0}^{\infty}(-1)^k t^{2k}$$

converges uniformly on every interval $[-a,a]$ with $0<a<1$. For $|x|\leqslant a<1$ we then have

$$\begin{aligned} s(x) &= \lim_{n\to\infty} s_n(x) = \lim_{n\to\infty}\int_0^x \sum_{k=0}^{n}(-1)^k t^{2k}dt \\ &= \int_0^x \lim_{n\to\infty}\sum_{k=0}^{n}(-1)^k t^{2k}dt = \int_0^x (1+t^2)^{-1}dt = \arctan x. \end{aligned}$$

Since $a<1$ was arbitrary we get $s(x)=\arctan x$ for all $x\in(-1,1)$. But s and arctan are continuous functions on $[-1,1]$ and thus we see that the relation $s(x)=\arctan x$ is valid for all $x\in[-1,1]$.

3.43. Example. The series (3.14) representing $\pi/4$ converges, but does so extremely slowly and is therefore inefficient for computing the digits of π. Recall Machin's formula

$$\frac{\pi}{4} = 4\arctan\frac{1}{5} - \arctan\frac{1}{239}$$

established in Example 2.80. In combination with the formula (3.13) this can be used to efficiently approximate π. One gets

$$\pi = \frac{16}{5} - \frac{16}{3\cdot 5^3} + \frac{16}{5\cdot 5^5} - \frac{16}{7\cdot 5^7} + \frac{16}{9\cdot 5^9} - \frac{16}{11\cdot 5^{11}} + \frac{16}{13\cdot 5^{13}} - \frac{16}{15\cdot 5^{15}} - \frac{4}{239} + \frac{4}{3\cdot 239^3} + \tau,$$

where $|\tau|<10^{-12}$. For efficient computation one uses a trick called *Horner's scheme* to rewrite the last formula as

$$\begin{aligned} \pi =& \frac{16}{5}\Big(1-\frac{4}{100}\Big(\frac{1}{3}-\frac{4}{100}\Big(\frac{1}{5}-\frac{4}{100}\Big(\frac{1}{7}-\frac{4}{100}\Big(\frac{1}{9}-\frac{4}{100}\Big(\frac{1}{11}-\frac{4}{100}\Big(\frac{1}{13}-\frac{4}{100}\frac{1}{15}\Big)\Big)\Big)\Big)\Big)\Big)\Big) \\ &- \frac{4}{239}\Big(1-\frac{1}{3\cdot 239^2}\Big) + \tau \end{aligned}$$

which can be used to show that $\pi = 3.1415926535...$ Remarkably, already in 1706 Machin used his formula to manually compute 100 digits in the decimal expansion of π.

3.1.6. Summation by parts. A useful method to investigate convergence of sums with cancellation is *summation by parts,* a formula which resembles the familiar integration by parts formula: If f, g are continuous on $[a,b]$, g in addition differentiable with continuous derivative, then

$$\int_a^b f(x)g(x)dx = F(b)g(b) - F(a)g(a) - \int_a^b F(x)g'(x)dx, \tag{3.15}$$

where F is such that $F'=f$. Integration by parts can be an effective tool to estimate integrals with cancellation, see e.g. Exercise 2.140. *Summation by parts*, also known as *Abel summation,* can likewise detect cancellation phenomena in sums. Here we consider functions on the integers and replace derivatives in the integration by parts formula with differences.

3.44. Theorem (Summation by parts). *Let $f:\mathbb{Z}\to\mathbb{C}$, $g:\mathbb{Z}\to\mathbb{C}$. Let $L,a,b\in\mathbb{Z}$ with $L\leqslant a\leqslant b$. Define*

$$F(n) = \begin{cases} \sum_{k=L}^{n} f(k) & \text{for } n\geqslant L \\ 0 & \text{for } n<L \end{cases} \qquad \text{and} \qquad \triangle g(n) = g(n+1)-g(n)\,.$$

Then

$$\sum_{n=a}^{b} f(n)g(n) = F(b)g(b) - F(a-1)g(a) - \sum_{n=a}^{b-1} F(n)\triangle g(n).$$

PROOF. We use $F(n) - F(n-1) = f(n)$ for $n \geqslant L$ and since $a \geqslant L$ we see that

$$\begin{aligned}\sum_{n=a}^{b} f(n)g(n) &= \sum_{n=a}^{b} (F(n) - F(n-1))g(n) \\ &= \sum_{n=a}^{b} F(n)g(n) - \sum_{n=a}^{b} F(n-1)g(n) = \sum_{n=a}^{b} F(n)g(n) - \sum_{n=a-1}^{b-1} F(n)g(n+1) \\ &= \Big(\sum_{n=a}^{b-1} F(n)g(n)\Big) + F(b)g(b) - \Big(\sum_{n=a}^{b-1} F(n)g(n+1)\Big) - F(a-1)g(a)\end{aligned}$$

which equals the right hand side of the asserted formula. □

The summation by parts formula plays a key role in the following exercises.

3.45. EXERCISE. Assume that $\sum_{k=0}^{\infty} b_k$ and $\sum_{k=0}^{\infty} |a_k - a_{k+1}|$ are convergent. Prove that $\sum_{n=0}^{\infty} a_n b_n$ is convergent.

3.46. EXERCISE. Let $a_k \geqslant 0$ and assume that a_k is a monotone bounded sequence. Assume further the partial sums $\sum_{k=1}^{n} b_k$ are bounded. Prove:

(i) The sequence $S_n = \sum_{k=1}^{n} a_k b_k$ is bounded.
(ii) If in addition $\lim_{n\to\infty} a_n = 0$, then $\sum_{k=1}^{\infty} a_k b_k$ converges.

3.47. EXERCISE. Let $a_n \geqslant 0$ with $a_{n+1} \leqslant a_n$ for all $n \geqslant 0$.

(i) Show that for every $x \in \mathbb{R}$ there is a constant $C_x \in (0, \infty)$ such that

$$\Big|\sum_{k=1}^{n} a_k \sin(kx)\Big| \leqslant C_x \text{ for all } n \geqslant 1.$$

(ii) If in addition $\lim_{n\to\infty} a_n = 0$, show that $\sum_{k=1}^{\infty} a_k \sin(kx)$ converges for all $x \in \mathbb{R}$.
(iii) Prove the *alternating series test* (aka *Leibniz test*): If (a_k) is a monotone sequence converging to 0, then $\sum_{k=1}^{\infty} (-1)^k a_k$ converges.

3.48. EXERCISE. Let $a, \theta \in \mathbb{R}$. Consider the series

$$\sum_{n=1}^{\infty} n^a e^{in\theta}.$$

(i) Show that it converges absolutely if $a < -1$ and diverges if $a \geqslant 0$.
(ii) Let $-1 \leqslant a < 0$. Show that the series converges if and only if θ is *not* of the form $2\pi m$ with $m \in \mathbb{Z}$.

3.2. Power series

A *power series* is a series of the form

$$\sum_{n=0}^{\infty} c_n x^n$$

where the coefficients c_n may be arbitrary complex numbers. Power series naturally arise when taking a limit in Taylor's formula (see Theorem 2.49). We view a power series as a function of the variable x, which may be real or complex. When x denotes a complex variable, it is customary to use the letter z instead.

3.2.1. Radius of convergence. The set of $z \in \mathbb{C}$ for which the power series

$$\sum_{n=0}^{\infty} c_n z^n$$

converges has a simple structure. It is characterized by a unique number $R \in [0, \infty]$ called the *radius of convergence*, which we may define by

$$R = \Big(\limsup_{n\to\infty} |c_n|^{1/n} \Big)^{-1}, \tag{3.16}$$

with the convention that $R = \infty$ if $\limsup_{n\to\infty} |c_n|^{1/n} = 0$. The identity (3.16) is known as the *Cauchy–Hadamard formula*.

3.49. Theorem. *Suppose the power series $\sum_{n=0}^{\infty} c_n z^n$ has radius of convergence $R \in [0, \infty]$. Then*

(i) If $r \in (0, R)$, then $\sum_{n=0}^{\infty} c_n z^n$ converges uniformly on $\overline{B}(0, r) = \{z \in \mathbb{C} : |z| \leqslant r\}$. In particular, it converges for all $z \in \mathbb{C}$ with $|z| < R$.

(ii) $\sum_{n=0}^{\infty} c_n z^n$ diverges for all $z \in \mathbb{C}$ with $|z| > R$.

Proof.

(i) The series $\sum_{n=0}^{\infty} |c_n| r^n$ converges by the root test since $\limsup |c_n|^{1/n} r < 1$. By Theorem 3.36, $\sum_{n=0}^{\infty} c_n z^n$ converges uniformly on $\overline{B}(0, r)$.

(ii) If $R = \infty$, there is nothing to show. If $R < \infty$, we may choose a subsequence $(c_{\kappa(n)})$ such that $\lim_{n\to\infty} |c_{\kappa(n)}|^{1/\kappa(n)} = R^{-1}$. Then for $|z| > R$ the sequence $|c_{\kappa(n)} z^{\kappa(n)}|$ is unbounded as $n \to \infty$ and hence the series does not converge. □

As an immediate consequence of this theorem (and Theorem 2.91) we obtain continuity of power series.

3.50. Corollary. *Suppose the power series $\sum_{n=0}^{\infty} c_n x^n$ has radius of convergence $R \in [0, \infty]$. Then the function $x \mapsto f(x) = \sum_{n=0}^{\infty} c_n x^n$ is continuous on $(-R, R)$.*

Within the radius of convergence, power series can be differentiated and integrated *termwise*.

3.51. Theorem. *Suppose that the power series $f(x) = \sum_{n=0}^{\infty} c_n x^n$ has radius of convergence R. Then*

(i) f is differentiable on $(-R, R)$ and

$$f'(x) = \sum_{n=1}^{\infty} c_n n x^{n-1} = \sum_{n=0}^{\infty} c_{n+1}(n+1) x^n.$$

(ii) f is Riemann integrable on $[0, x]$ for every $x \in (-R, R)$ and

$$\int_0^x f(t)\, dt = \sum_{n=0}^{\infty} c_n \frac{x^{n+1}}{n+1} = \sum_{n=1}^{\infty} \frac{c_{n-1}}{n} x^n.$$

Proof of (i). Observe that $\lim_{n\to\infty} (n+1)^{1/n} = 1$ and therefore

$$\limsup_{n\to\infty} |c_{n+1}(n+1)|^{1/n} = \lim_{n\to\infty} (n+1)^{1/n} \limsup_{n\to\infty} |c_{n+1}|^{1/n} = \limsup_{n\to\infty} |c_{n+1}|^{1/(n+1)}$$

which shows that the formally differentiated power series has the same radius of convergence as the original power series. Therefore it is uniformly convergent on $(-r, r)$, for any $r < R$ and the claim follows from Theorem 3.39. □

The analogous argument for part (ii) is left to the reader. Termwise differentiation can be applied iteratively. We then see that all derivatives exist in the open interval $(-R, R)$. In other words, the power series defines a C^∞ (smooth) function on $(-R, R)$.

3.52. Corollary. *Suppose the series $f(x) = \sum_{n=0}^{\infty} c_n x^n$ has radius of convergence $R > 0$. Then f is a C^∞ function on $(-R, R)$ and the derivatives of f are given by the power series*

$$f^{(k)}(x) = \sum_{n=0}^{\infty} c_{n+k}(k+n)(k+n-1)\cdots(n+1)x^n.$$

In particular,

$$c_k = \frac{f^{(k)}(0)}{k!} \quad \text{for} \quad k = 0, 1, 2, \ldots.$$

The proof is by induction on k and using Theorem 3.51 in the induction step. An important consequence of Corollary 3.52 is that power series are uniquely determined by their coefficients:

3.53. Corollary. *If $\sum_{n=0}^{\infty} a_n x^n$, $\sum_{n=0}^{\infty} b_n x^n$ converge on $(-r, r)$ for some $r > 0$ and are equal on this interval, then $a_n = b_n$ for all $n \in \mathbb{N}_0$.*

This follows because Corollary 3.52 gives $a_n = f^{(n)}(0)/n! = b_n$ for all $n \in \mathbb{N}$.

3.2.2. Taylor series. Corollary 3.53 says that if f is given by a power series with radius of convergence $R > 0$, then for $|x| < R$,

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

The series on the right is referred to as a *Taylor series* at the origin in reference to Taylor's theorem, Theorem 2.49. Given a function f that is C^∞ on an interval $(-r, r)$, it is natural to ask when its Taylor series at the origin

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

converges and in case of convergence, whether its value equals $f(x)$. As hinted by the remainder term in Taylor's theorem, this will depend on the magnitude of $|f^{(n)}(x)|$ as n grows when x is near 0.

3.54. Proposition. *Suppose that f is C^∞ on an interval $(-R, R)$ and that there exist constants $A, C \in (0, \infty)$ so that*

$$|f^{(n)}(x)| \leqslant CA^n n!$$

for all $|x| < R$. Then f is represented by its Taylor series near the origin. That is, there exists $r > 0$ so that for all $|x| < r$,

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

(and in particular the radius of convergence of the series on the right is $\geqslant r$). Moreover, one may set $r = \min(R, A^{-1})$.

Proof. From the assumption and the Cauchy–Hadamard formula, the radius of convergence of the Taylor series is $\geqslant A^{-1}$. From Taylor's theorem (in the form of Theorem 2.50) we obtain for $|x| < r = \min(R, A^{-1})$,

$$\Big| f(x) - \sum_{n=0}^{N} \frac{f^{(n)}(0)}{n!} x^n \Big| = \frac{|x|^{N+1}}{(N+1)!} |f^{(N+1)}(\xi)|$$

for some ξ with $|\xi| \leqslant |x|$. Using the assumption, the previous is $\leqslant C(|x|A)^{N+1}$ and this converges to zero as $N \to \infty$ since $|x|A < rA \leqslant 1$. □

3.55. Example. The exponential function exp as defined in §2.5 is represented by its Taylor series on the entire real line. Indeed from (2.7), $\exp(x) = \sum_{k=0}^{n} \frac{x^k}{k!} + \exp(\xi)\frac{x^{n+1}}{(n+1)!}$ with ξ between 0 and x. Thus

$$\Big|\exp(x) - \sum_{k=0}^{n} \frac{x^k}{k!}\Big| \leqslant \max\{1, \exp x\}\frac{x^{n+1}}{(n+1)!}. \tag{3.17}$$

Since $\frac{x^{n+1}}{(n+1)!}$ converges to zero as $n \to \infty$, we get

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} \quad \text{for all } x \in \mathbb{R}.$$

3.56. Example. The series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$$

converges by the alternating series test. We claim that it has value $\log 2$. Let $f(x) = \log(1+x)$ for $x > -1$ and compute inductively that for every $k \in \mathbb{N}$

$$f^{(k)}(x) = \frac{(-1)^{k-1}(k-1)!}{(1+x)^k}.$$

Hence

$$\sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}x^k = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}x^k.$$

From Theorem 2.50 we get the remainder estimate

$$\sup_{0\leqslant x\leqslant 1} |R_n(0,x)| \leqslant \max_{0\leqslant x\leqslant 1} \frac{1}{(n+1)!} n!\, x^{n+1} = \frac{1}{n+1}.$$

Taking the limit $n \to \infty$,

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}x^k = \log(1+x), \quad 0 \leqslant x \leqslant 1$$

and setting $x = 1$ we obtain $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = \log 2$ as claimed.

3.57. Example. Not every smooth function is represented by its Taylor series. Consider

$$f(x) = \begin{cases} e^{-1/x} & \text{if } x > 0, \\ 0 & \text{if } x \leqslant 0. \end{cases} \tag{3.18}$$

This function is smooth at every point $x \in \mathbb{R}$. If $x \neq 0$, this follows by repeated applications of the chain rule. Let us consider the behavior at 0. Observe that f is continuous at 0, since

$$\lim_{x\to 0+} e^{-1/x} = \lim_{t\to\infty} e^{-t} = 0.$$

For the derivative at $x = 0$ we use the definition of differentiability and observe that for $x > 0$

$$\frac{f(x) - f(0)}{x - 0} = x^{-1}e^{-1/x}.$$

Now $\lim_{x\to 0+} x^{-1}e^{-1/x} = \lim_{t\to\infty} te^{-t} = 0$ (one can see this by using the Taylor expansion for the exponential function, which implies $e^t \geqslant t^2/2$ for $t > 0$). For $x < 0$ we have $\frac{f(x)-f(0)}{x-0} = 0$. Thus f is differentiable at 0 with $f'(0) = 0$.

We argue similarly for the higher derivatives and for $x > 0$ observe that $f^{(n)}(x)$ is given by $R(x)e^{-1/x}$ where R is a rational function. Using $\lim_{x\to 0+} x^{-d}e^{-1/x} = 0$ for all $d > 0$ (which follows from $e^t \geqslant t^{d+1}/(d+1)!$ for $t > 0$) we see that all derivatives of f at 0 exist and

$$f^{(n)}(0) = 0 \text{ for all } n = 0, 1, 2, \dots.$$

In particular, the Taylor series at the origin for f is identically equal to zero and thus does not represent f in any neighborhood of the origin.

3.58. Exercise. Let f be as in Example 3.57. Prove that

$$f^{(n)}(x) = \begin{cases} x^{-2n} P_{n-1}(x) e^{-1/x} & \text{if } x > 0 \\ 0 & \text{if } x \leqslant 0 \end{cases}$$

where P_{n-1} is a polynomial of degree $n-1$. Find a recursion formula for P_n.

It is possible for the Taylor series of a smooth function to have radius of convergence equal to zero. One might ask which power series are the Taylor series of some smooth function. Borel's theorem gives a satisfying answer to these questions.

3.59. Theorem (Borel). *Let $(b_n)_{n\geqslant 0}$ be real numbers. There exists a function $f \in C^\infty(\mathbb{R})$ such that*

$$f^{(n)}(0) = b_n \text{ for all } n = 0, 1, 2, \ldots.$$

We first construct *smooth bump functions,* an indispensable tool for many constructions and techniques in analysis.

3.60. Proposition. *Given $x_0 \in \mathbb{R}$ and $0 < r < R$ there is a function $\chi \in C^\infty(\mathbb{R})$ such that*

$$\begin{cases} 0 \leqslant \chi(x) \leqslant 1 & \text{for all } x \in \mathbb{R}, \\ \chi(x) = 1 & \text{for } x \in [x_0 - r, x_0 + r], \\ \chi(x) = 0 & \text{for } x \notin [x_0 - R, x_0 + R]. \end{cases}$$

Proof of Proposition 3.60. We may assume without loss of generality that $x_0 = 0$ (by means of the translated function $\chi(\cdot - x_0)$). Define

$$g(x) = f(1 - x^2)$$

with f given by (3.18). Then g is a C^∞ function on $\mathbb{R}$ with the property that $g(x) > 0$ for $x \in (-1, 1)$ and $g(x) = 0$ for $x \notin (-1, 1)$. Let $M = \int_{-1}^{1} g(t)dt \in (0, \infty)$. For $\varepsilon > 0$ small enough ($\varepsilon < (R - r)/2$ will suffice) let

$$h_\varepsilon(t) = M^{-1} \varepsilon^{-1} g(\varepsilon^{-1} t).$$

By construction, h_ε is a C^∞ function that vanishes outside the interval $(-\varepsilon, \varepsilon)$, satisfies $h_\varepsilon(t) \geqslant 0$ for all $t \in \mathbb{R}$ and

$$\int_{-\varepsilon}^{\varepsilon} h_\varepsilon(t)\, dt = 1.$$

Finally we define

$$\chi(x) = \int_{-r-\varepsilon}^{r+\varepsilon} h_\varepsilon(x - t)\, dt = \int_{x-r-\varepsilon}^{x+r+\varepsilon} h_\varepsilon(t)\, dt.$$

By the fundamental theorem of calculus, $\chi'(x) = h_\varepsilon(x + r + \varepsilon) - h_\varepsilon(x - r - \varepsilon)$ and in particular χ is C^∞. For all $x \in \mathbb{R}$,

$$0 \leqslant \chi(x) \leqslant \int_{-\varepsilon}^{\varepsilon} h_\varepsilon(t)\, dt = 1.$$

For $|x| \leqslant r$ the interval $[x - r - \varepsilon, x + r + \varepsilon]$ contains $[-\varepsilon, \varepsilon]$ and therefore $\chi(x) = 1$ for $|x| \leqslant r$. For $x \geqslant R$ we have $x - r - \varepsilon > \varepsilon$ and for $x \leqslant -R$ we have $x + r + \varepsilon < -\varepsilon$, so the intervals $[x - r - \varepsilon, x + r + \varepsilon]$ and $[-\varepsilon, \varepsilon]$ are disjoint in those cases (here we used that $\varepsilon < (R - r)/2$). This implies $\chi(x) = 0$ for $|x| \geqslant R$. □

We now sketch the construction of a function f satisfying the requirements of Theorem 3.59. Let $\chi \in C^\infty(\mathbb{R})$ be a smooth bump function so that $\chi(x) = 0$ for $|x| \geqslant 1$ and $\chi(x) = 1$ for $|x| \leqslant 1/2$. For $x \in \mathbb{R}$ define

$$f(x) = \sum_{n=0}^{\infty} b_n \frac{x^n}{n!} \chi(\varepsilon_n^{-1} x),$$

where $\varepsilon_n \in (0,1)$ are suitably chosen small numbers converging to zero as $n \to \infty$. In particular, note that the right hand side is not a power series in any neighborhood of the origin. In fact, the sum has only finitely many non-zero terms for every fixed value of $x \in \mathbb{R}$. Since χ is constant near zero, it is clear from the product rule that $f^{(n)}(0) = b_n$ for all n as long as all the differentiated series converge uniformly so that we are allowed to differentiate termwise. The numbers ε_n must be chosen to guarantee this. Observe that each of the derivatives of the function χ is bounded on $\mathbb{R}$. Setting

$$A_n = \max_{j=0,\ldots,n-1} \sup_{t \in \mathbb{R}} |\chi^{(j)}(t)|$$

a permissible choice for ε_n is

$$\varepsilon_n = \frac{2^{-2n}}{1 + |b_n| A_n}.$$

One may then check using the Leibniz rule (Exercise 2.119) that for $n > k$ the nth term of the k times differentiated series is bounded by 2^{-n}, for all $x \in \mathbb{R}$.

3.2.3. Analytic functions. So far we have considered power series centered at the origin. It is also common to consider power series centered at a given point $x_0 \in \mathbb{R}$:

$$f(x) = \sum_{n=0}^{\infty} c_n (x - x_0)^n \tag{3.19}$$

Changing variables $y = x - x_0$ we may apply all of our results on power series centered at the origin. In particular, we have convergence for all $x \in (x_0 - R, x_0 + R)$, where $R \in [0, \infty]$ denotes the radius of convergence and f is a smooth function on that interval with

$$\frac{f^{(n)}(x_0)}{n!} = c_n, \quad n = 0, 1, 2, \ldots$$

Functions given by power series are also called *analytic functions.*

3.61. Definition. A function f defined on an open interval around $x_0 \in \mathbb{R}$ is called *(real) analytic at* x_0 if there exists $r > 0$ such that f is represented by a convergent power series centered at x_0, i.e. (3.19) holds for all $x \in (x_0 - r, x_0 + r)$.

3.62. Exercise. Let $w \in \mathbb{C}$. Show that the function $f(x) = \frac{1}{x-w}$ is analytic at every point $x_0 \in \mathbb{R}$ as long as $x_0 \neq w$. Determine the power series expansion at x_0 and observe that its radius of convergence equals $|w - x_0|$.

If f is analytic at x_0, then it is smooth on an interval around x_0 and the power series (3.19) coincides with the Taylor series of f at x_0. A function $f : I \to \mathbb{C}$ for an open interval I is called *real analytic* (on I) if f is analytic at every $x_0 \in I$.

3.63. Theorem. *Let $f(x) = \sum_{n=0}^{\infty} c_n (x - x_0)^n$ be a power series converging for $|x - x_0| < R$. Then f is real analytic on $(x_0 - R, x_0 + R)$.*

Specifically, if $x_1 \in (x_0 - R, x_0 + R)$, then f can be expanded in a power series about the point x_1 which converges at least for those x with $|x - x_1| < R - |x_0 - x_1|$ and for those x we have

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_1)}{n!} (x - x_1)^n.$$

PROOF. Expanding $(x-x_0)^n = (x-x_1+x_1-x_0)^n = \sum_{k=0}^n \binom{n}{k}(x-x_1)^k(x_1-x_0)^{n-k}$ we would like to write

$$f(x) = \sum_{n=0}^{\infty}\sum_{k=0}^{n} c_n\binom{n}{k}(x-x_1)^k(x_1-x_0)^{n-k} = \sum_{k=0}^{\infty} a_k(x-x_1)^k,$$

$$\text{where } a_k = \sum_{n=0}^{\infty} c_n\binom{n}{k}(x_1-x_0)^{n-k}.$$

In the last step we interchanged the sums over k and n. This needs to be justified. By Theorem 3.34 it is justified if

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n} |c_n|\binom{n}{k}|x-x_1|^k|x_1-x_0|^{n-k} < \infty.$$

Assuming $|x-x_1| \leqslant R - |x_0-x_1| - \varepsilon$ for some $\varepsilon > 0$ we have

$$|c_n|\sum_{k=0}^{n}\binom{n}{k}|x-x_1|^k|x_1-x_0|^{n-k} = |c_n|(|x-x_1|+|x_1-x_0|)^n \leqslant |c_n|(R-\varepsilon)^n$$

so the convergence holds by definition of radius of convergence. By Theorem 3.34 we get the power series expansion

$$f(x) = \sum_{k=0}^{\infty} a_k(x-x_1)^k$$

convergent at least for $|x-x_1| \leqslant R - |x_0-x_1| - \varepsilon$. By Corollary 3.52 we have $k!a_k = f^{(k)}(x_1)$, which yields the asserted Taylor series expansion about the point x_1. □

3.64. EXAMPLE. The power series $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ has radius of convergence $R = \infty$, so exp is a real analytic function on $\mathbb{R}$ and for every $x, x_1 \in \mathbb{R}$:

$$\exp(x) = \sum_{n=0}^{\infty} \frac{e^{x_1}}{n!}(x-x_1)^n.$$

3.65. EXAMPLE. It is natural to wonder whether functions real analytic on $\mathbb{R}$ can always be represented by power series with infinite radius of convergence. This is not the case. Define for $x \in \mathbb{R}$:

$$f(x) = \frac{1}{1+x^2}.$$

For $|x| < 1$ we have the geometric series expansion

$$f(x) = \sum_{n=0}^{\infty}(-1)^n x^{2n}$$

which has radius of convergence equal to 1. Nevertheless, f is analytic at every point $x_0 \in \mathbb{R}$. Indeed, we can see this from writing

$$f(x) = \frac{1}{2i}\Big(\frac{1}{x-i} - \frac{1}{x+i}\Big)$$

and applying the result from Exercise 3.62. The power series expansion at x_0 has radius of convergence equal to the distance of x_0 to the set of *singularities* $\{\pm i\}$ in the complex plane, i.e. $R = \min\{|x_0 - i|, |x_0 + i|\}$.

3.2.4. Complex analyticity and complex derivatives. As we have already seen, it is natural to consider power series as functions of a complex variable, i.e. given $z_0 \in \mathbb{C}$ and $c_n \in \mathbb{C}$ we may consider the function

$$f(z) = \sum_{n=0}^{\infty} c_n (z - z_0)^n \tag{3.20}$$

depending on the complex variable $z \in \mathbb{C}$. The value $f(z)$ is well-defined as long as z is contained in the open disk

$$B(z_0, R) = \{z \in \mathbb{C} : |z - z_0| < R\},$$

where R is the radius of convergence. We then say that f is *complex analytic* on $B(z_0, R)$. If $R = \infty$, then f is also called an *entire function.*

It is natural to ask whether our results on termwise differentiation and Taylor series carry over to the case of functions of a complex variable. In particular, in what sense can the formally differentiated series

$$\sum_{n=1}^{\infty} n c_n (z - z_0)^{n-1}$$

be interpreted as a derivative of f at z? We may extend the familiar definition of derivative to the case of complex variables.

3.66. Definition. Let $f : B(z_0, R) \to \mathbb{C}$ be a complex-valued function defined on the open disk $B(z_0, R) \subset \mathbb{C}$. f is *complex differentiable at* z_0 if

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists. In that case, the limit is denoted by $f'(z_0)$ and called the *complex derivative* of f at z_0.

Many familiar properties from real variable differentiation carry over without much difficulty to the complex case. For example, one verifies that the function $f(z) = z^n$ is complex differentiable with $f'(z) = nz^{n-1}$. Similarly a function which is complex analytic on $B(z_0, R)$ and given by (3.20) is complex differentiable, and

$$f'(z) = \sum_{n=1}^{\infty} n c_n (z - z_0)^{n-1}$$

holds for all $z \in B(z_0, R)$. One may also formulate a complex-variable version of Proposition 3.54. In particular, if f is given by (3.20), then

$$c_n = \frac{f^{(n)}(z_0)}{n!}$$

holds for all $n \geqslant 0$ (as long as $R > 0$).

The reader is invited to prove these and further properties by carefully reviewing the real-variable arguments. When a real analytic function is given by a power series with radius of convergence equal to ∞, then we may extend it to an entire function.The most important examples of this for the purpose of this book are the exponential function and the trigonometric functions. The familiar formula for the exponential function

$$\exp(z) = \sum_{n=0}^{\infty} \frac{1}{n!} z^n$$

serves as its definition when z is not a real number.

3.67. Observation. *For all $w, z \in \mathbb{C}$ we have the functional equation*

$$\exp(w + z) = \exp(w) \exp(z).$$

It is instructive to prove this by using the Cauchy product for the multiplication of the series for $\exp(w)$ and $\exp(z)$. In view of the functional equation and the definition $e = \sum_{k=0}^{\infty} \frac{1}{k!}$ it makes sense to use the notation e^z for $\exp(z)$ for all $z \in \mathbb{C}$.

Recall that we defined the trigonometric functions sin and cos for real arguments in §2.5. From Proposition 3.54 one can observe that the formulas

$$\cos(z) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} z^{2n} \quad \text{and} \quad \sin(z) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} z^{2n+1}$$

hold for all real z. The formulas may serve as definitions for $\sin(z)$ and $\cos(z)$ for all $z \in \mathbb{C}$. The reader should verify the following facts that are already familiar in the real-variable setting:

3.68. Observation (Euler's identity). *For all $z \in \mathbb{C}$,*

$$e^{iz} = \cos(z) + i \sin(z),$$

and

$$\cos(z) = \frac{e^{iz} + e^{-iz}}{2}, \quad \sin(z) = \frac{e^{iz} - e^{-iz}}{2i}.$$

3.69. Observation. *For all $z \in \mathbb{C}$,*

$$(\cos z)^2 + (\sin z)^2 = 1.$$

(For $z \in \mathbb{R}$ this is known as the Pythagorean theorem.)

Remark. As innocuous as it may seem at first glance, the definition of complex differentiability leads to fundamental differences between real- and complex-differentiable functions. Specifically, it may surprise the reader that if f is complex differentiable in an open disk around z_0, then it is automatically complex analytic there (and in particular, infinitely often complex differentiable) – as we have learned, this is very much not the case for real-variable functions. The proof of this fact is the entry point to a rich area of mathematics called *complex analysis,* which we do not discuss in this book. For a good introduction to this topic, see e.g. Stein–Shakarchi [**75**] or Rudin [**66**].

3.2.5. Abel summability. We introduce a *summability method* which attaches a value to some divergent series, and also allows one to compute certain convergent sums.

3.70. Definition. Let $(c_n)_{n=0}^{\infty}$ be a sequence of complex numbers. Then the series $\sum_{k=0}^{\infty} c_k$ is called *Abel summable* to S if the series $A(r) = \sum_{k=0}^{\infty} c_k r^k$ converges for every r with $0 \leqslant r < 1$ and if

$$\lim_{r \to 1-} A(r) = S\,.$$

3.71. Theorem (Abel). *If $\sum_{k=0}^{\infty} c_k = S$ converges, then $\sum_{k=0}^{\infty} c_k$ is Abel summable to S.*

3.72. Corollary. *Suppose $f(z) = \sum_{n=0}^{\infty} c_n z^n$ has radius of convergence $R > 0$. Assume that $|w| = R$ and that $\sum_{n=0}^{\infty} c_n w^n$ converges. Then*

$$\lim_{r \to 1-} f(rw) = \sum_{n=0}^{\infty} c_n w^n.$$

This follows by applying Abel's theorem to the convergent series $\sum_{n=0}^{\infty} c_n w^n$.

The corollary expresses a continuity property of $f(z)$ when restricted to the ray $\{tw : t > 0\} \subset \mathbb{C}$. Abel summability means that the limit of $f(z)$ as z tends to the boundary point w on this ray exists and under the assumption of the corollary we can identify this limit as $\sum_{n=0}^{\infty} c_n w^n$. The converse of Abel's theorem is not true:

3.73. Exercise.

(i) Consider the series $1 - 2 + 3 - 4 + 5 - 6 \pm \ldots = \sum_{k=0}^{\infty} (-1)^k (k+1)$ which is not convergent. Show that it is Abel summable to some S and determine S.

(ii) What can you say about Abel summability of $\sum_{k=0}^{\infty}(-1)^k(k+1)z^k$ for $|z|=1$?

PROOF OF THEOREM 3.71. The assumption of convergence of $\sum_{n=0}^{\infty} c_n$ implies that the series $\sum_{n=0}^{\infty} c_n r^n$ converges for $0 \leqslant r < 1$, by Theorem 3.49. The proof is based on summation by parts (Theorem 3.44). To apply it we set $s_n = \sum_{k=0}^{n} c_k$, $s_{-1} = 0$ and obtain

$$\sum_{n=0}^{N} c_n r^n = \sum_{n=0}^{N} (s_n - s_{n-1}) r^n = s_N r^N + (1-r) \sum_{n=0}^{N-1} s_n r^n.$$

Let $0 < r < 1$. We let $N \to \infty$ to get

$$\sum_{n=0}^{\infty} c_n r^n = (1-r) \sum_{n=0}^{\infty} s_n r^n.$$

By assumption, $s_n \to S$. Let $\varepsilon > 0$ and choose $N \in \mathbb{N}$ such that

$$|s_n - S| < \varepsilon$$

for all $n > N$. Then,

$$\Big| \sum_{n=0}^{\infty} c_n r^n - S \Big| = \Big| (1-r) \sum_{n=0}^{\infty} (s_n - S) r^n \Big|,$$

because $(1-r)\sum_{n=0}^{\infty} r^n = 1$. Now we use the triangle inequality and split the sum at $n = N$, so we get

$$\begin{aligned} \Big| \sum_{n=0}^{\infty} c_n r^n - S \Big| &\leqslant (1-r) \sum_{n=0}^{N} |s_n - S| r^n + (1-r) \sum_{n=N+1}^{\infty} \overbrace{|s_n - S|}^{\leqslant \varepsilon} r^n \\ &\leqslant (1-r) \sum_{n=0}^{N} |s_n - S| r^n + \varepsilon. \end{aligned}$$

By making r sufficiently close to 1 we can achieve that

$$(1-r) \sum_{n=0}^{N} |s_n - S| r^n \leqslant \varepsilon.$$

Since ε was arbitrary this concludes the proof. □

Abel's theorem provides a convenient tool to evaluate convergent series.

3.74. EXAMPLE. Leibniz' formula (Example 3.42) can be proved via Abel's theorem. For $x \in (-1,1)$ consider the power series

$$\arctan x = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1}$$

by integrating the geometric series expansion for $(1+t^2)^{-1} = \sum_{n=0}^{\infty}(-1)^n t^{2n}$. For $x = 1$ the series $\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}$ converges by the alternating series test. Thus, Abel's theorem implies

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} = \lim_{x \to 1-} \arctan(x) = \arctan(1) = \frac{\pi}{4}.$$

Unlike in our first approach in Example 3.42 we did not have to establish uniform convergence of the series on $[-1,1]$.

3.75. EXERCISE. The following complements the results on the Cauchy product in §3.1.3. Suppose that $(a_n)_{n=0}^{\infty}$, $(b_n)_{n=0}^{\infty}$ are given and $c_n = \sum_{k=0}^{n} a_{n-k} b_k$. If $\sum_n a_n$, $\sum_n b_n$ and $\sum_n c_n$ converge, then

$$\sum_{n=0}^{\infty} c_n = \Big(\sum_{n=0}^{\infty} a_n \Big) \Big(\sum_{n=0}^{\infty} b_n \Big).$$

3.3. Infinite products

Much like an infinite sum, an infinite product arises from successively multiplying infinitely many numbers:

$$\prod_{n=1}^{\infty} p_n = p_1 \cdot p_2 \cdot p_3 \cdots$$

As for sums, we may give a rigorous meaning to this by considering it a limit of a sequence:

3.76. Definition. Let (p_k) be a sequence of non-zero complex numbers. If the sequence $(\prod_{k=1}^{n} p_k)_{n=1}^{\infty}$ of *partial products* converges to a non-zero complex number, then we say that the *infinite product* converges and denote its limit by

$$\prod_{k=1}^{\infty} p_k = \lim_{n\to\infty} \prod_{k=1}^{n} p_k.$$

Otherwise, we say that the infinite product *diverges*. If the sequence of partial products converges to zero, then we say that the infinite product *diverges to zero* and write $\prod_{k=1}^{\infty} p_k = 0$.

If zero were allowed as a limit of a convergent product, many of the following theorem statements would need to be revised to account for ultimately uninteresting special cases. Throughout the remainder of this section we will assume that (p_k) denote non-zero complex numbers. At least when the p_k are positive, one may apply a logarithm:

$$\log\Big(\prod_{k=1}^{n} p_k\Big) = \sum_{k=1}^{n} \log(p_k).$$

With this in mind it will not surprise the reader that we will be able to reduce much of the study of infinite products to that of series. In particular:

3.77. Observation. *Let $p_k > 0$ for all $k \in \mathbb{N}$. The product $\prod_{k=1}^{\infty} p_k$ converges if and only if the series $\sum_{k=1}^{\infty} \log(p_k)$ converges.*

The Cauchy criterion we have already seen for sequences and series also has a formulation for infinite products:

3.78. Theorem (Cauchy criterion). *Let (p_n) be a sequence of nonzero numbers. The infinite product $\prod_{k=1}^{\infty} p_k$ is convergent if and only if for every $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that for all $n > m \geqslant N$*

$$\big|1 - \prod_{k=m+1}^{n} p_k\big| < \varepsilon.$$

Proof. Set $P_n = \prod_{k=1}^{n} p_k$. Then

$$1 - \prod_{k=m+1}^{n} p_k = \frac{P_m - P_n}{P_m}.$$

'$\Longrightarrow$': By assumption (P_n) is convergent to a nonzero limit. Hence there exists $a > 0$ such that $|P_k| \geqslant a$ for all $k \geqslant 1$. Since (P_n) is a Cauchy sequence there exists for given $\varepsilon > 0$ a positive integer N such that $|P_n - P_m| < a\varepsilon$ for $n \geqslant N$. Thus for $n > m \geqslant N$:

$$\big|1 - \prod_{k=m+1}^{n} p_k\big| = \frac{|P_n - P_m|}{|P_m|} < \frac{a\varepsilon}{a} = \varepsilon.$$

'$\Longleftarrow$': We first apply the condition with $\varepsilon = 1/2$, and $n = m+1$, and see that there is N_0 such that $|p_{m+1} \cdots p_n - 1| < \frac{1}{2}$ for all $n \geqslant N_0$, i.e. $\frac{1}{2} < \prod_{k=m+1}^{n} p_k < \frac{3}{2}$ for $n > m \geqslant N_0$. In particular $p_k \neq 0$ for $k \geqslant N_0$. Now, for $n \geqslant N_0$ set $Q_n = P_n/P_{N_0} = \prod_{k=N_0+1}^{n} p_k$. We have $|Q_n| > 1/2$. In order to show that (Q_n) is convergent it suffices to prove that it is a Cauchy sequence.

For every $\varepsilon > 0$ we find $N \geqslant N_0$ such that $|(\prod_{k=m+1}^{n} p_k) - 1| < \frac{\varepsilon}{2}$ for $n > m \geqslant N$ and for those m, n

$$|Q_n - Q_m| \leqslant 2\frac{|Q_n - Q_m|}{|Q_m|} = 2|p_{m+1} \cdots p_n - 1| < \varepsilon.$$

Hence (Q_n) is a Cauchy sequence and thus convergent. And this implies that $P_n = P_{N_0} Q_n$ converges, too. □

Applying the Cauchy criterion with $n = m + 1$ gives a necessary condition for convergence:

3.79. Corollary. *If $\prod_{n=1}^{\infty} p_n$ is convergent, then $\lim_{n\to\infty} p_n = 1$*

In view of this we will write infinite products from now on as

$$\prod_{n=1}^{\infty} p_n = \prod_{n=1}^{\infty} (1 + a_n),$$

so that the necessary condition reads

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

in analogy with Corollary 3.4.

3.80. Theorem. *Let $a_n > 0$ for all n. Then*

$$\prod_{n=1}^{\infty} (1 + a_n) \quad \textit{converges if and only if} \quad \sum_{n=1}^{\infty} a_n \quad \textit{converges.}$$

Proof. We will derive this from Observation 3.77. By Taylor's theorem,

$$\log(1 + t) = t - (1 + \xi)^{-2} \frac{t^2}{2} \tag{3.21}$$

with $t > 0$, $\xi \in [0, t]$. This implies

$$t - \frac{t^2}{2} \leqslant \log(1 + t) \leqslant t \quad \text{for } t \geqslant 0.$$

Since $a_n \to 0$ in either case by Corollaries 3.4 and 3.79, there exists $N \in \mathbb{N}$ so that for all $n \geqslant N$, $0 \leqslant a_n \leqslant \frac{1}{2}$, which then implies

$$\tfrac{3}{4} a_n \leqslant \log(1 + a_n) \leqslant a_n.$$ □

3.81. Definition. We say that the product $\prod_{n=1}^{\infty}(1 + a_n)$ *converges absolutely* if $\prod_{n=1}^{\infty}(1 + |a_n|)$ converges.

The following is a consequence of Theorem 3.80.

3.82. Corollary. *The product $\prod_{n=1}^{\infty}(1 + a_n)$ converges absolutely if and only if the series $\sum_{k=1}^{\infty} a_k$ converges absolutely.*

In the absence of absolute convergence, convergence of $\sum_{n=1}^{\infty} a_n$ generally does not imply convergence of $\prod_{n=1}^{\infty}(1 + a_n)$.

3.83. Example. Let $a_n = (-1)^n n^{-1/2}$. Then $\sum_{n=2}^{\infty} a_n$ is convergent by the alternating series test (but not absolutely convergent). The product $\prod_{n=2}^{\infty}(1 + a_n)$ is not convergent. This can be shown by direct computation or as a consequence of the following theorem:

3.84. Theorem. *Let (a_n) be a sequence of real numbers such that $a_n \neq -1$. Suppose that $\sum_{n=1}^{\infty} a_n$ converges. Then*

$$\prod_{n=1}^{\infty} (1 + a_n) \quad \textit{converges if and only if} \quad \sum_{n=1}^{\infty} a_n^2 \quad \textit{converges.}$$

PROOF. Since $\sum_{n=1}^{\infty} a_n$ converges, there is N such that $|a_n| < 1/2$ for $n \geqslant N$. We argue as before using Observation 3.77 to see that $\prod_{n=N}^{\infty}(1+a_n)$ converges if and only if $\sum_{n=N}^{\infty}\log(1+a_n)$ converges. By Taylor expansion,

$$\log(1+a_n) = a_n - (1+\xi_n)^{-2}\frac{a_n^2}{2}, \quad |\xi_n| < 1/2.$$

Since $\sum_{n=N}^{\infty} a_n$ converges we see that $\sum_{n=N}^{\infty}\log(1+a_n)$ converges if and only if $\sum_{n=N}^{\infty} a_n^2$ converges, and the assertion follows. □

3.85. EXERCISE. Prove that the infinite product

$$\prod_{n=2}^{\infty}(1+(-1)^n n^{-\beta})$$

(i) converges absolutely if $\beta > 1$,
(ii) does not converge absolutely when $\frac{1}{2} < \beta \leqslant 1$,
(iii) diverges to 0 when $0 < \beta \leqslant \frac{1}{2}$.

3.86. EXAMPLE. The *Wallis product* is an interesting infinite product representation of the number $\pi/2$ as a limit of rational numbers:

$$\frac{\pi}{2} = \prod_{k=1}^{\infty}\frac{4k^2}{(2k-1)(2k+1)}. \tag{3.22}$$

The proof is based on the computation of the integrals

$$c_m = \int_0^{\pi/2}(\sin x)^m\,dx. \tag{3.23a}$$

Integration by parts gives for $m \geqslant 2$

$$c_m = \int_0^{\pi/2}(\sin x)^{m-1}\sin x\,dx = \int_0^{\pi/2}(m-1)(\sin x)^{m-2}\cos^2 x\,dx.$$

Since $\cos^2 x = 1-\sin^2 x$ this leads to $c_m = (m-1)(c_{m-2}-c_m)$ hence to $c_m = \frac{m-1}{m}c_{m-2}$ for $m \geqslant 2$. Thus we can compute the c_m, separately for even and odd m and get

$$c_0 = \pi/2, \quad c_1 = 1 \tag{3.23b}$$

and for $n \geqslant 1$

$$c_{2n} = \frac{\pi}{2}\cdot\prod_{k=1}^{n}\frac{2k-1}{2k}, \qquad c_{2n+1} = \prod_{k=1}^{n}\frac{2k}{2k+1}. \tag{3.23c}$$

Then

$$\frac{c_{2n+1}}{c_{2n}} = \frac{2}{\pi}\prod_{k=1}^{n}\frac{(2k)(2k)}{(2k-1)(2k+1)}.$$

Use the inequalities $(\sin x)^{2n+1} \leqslant (\sin x)^{2n} \leqslant (\sin x)^{2n-1}$ on $[0,\pi/2]$ which imply

$$1 \geqslant \frac{c_{2n+1}}{c_{2n}} \geqslant \frac{c_{2n+1}}{c_{2n-1}} = \frac{2n}{2n+1}.$$

Hence $\lim_{n\to\infty}\frac{c_{2n+1}}{c_{2n}} = 1$ and the Wallis product formula follows.

3.87. EXAMPLE (Sums of reciprocals of primes). If $(p_j)_{j=1}^{\infty}$ is the sequence of all prime numbers arranged in increasing order, then

$$\sum_{j=1}^{\infty}\frac{1}{p_j} = \infty, \qquad \prod_{j=1}^{\infty}\Big(1-\frac{1}{p_j}\Big)^{-1} = \infty.$$

This will follow once we show that for all $n \in \mathbb{N}$

$$\sum_{\substack{p \leqslant n \\ p \text{ prime}}} \frac{1}{p} > \log\log(n+1) - \log 2.$$

Assume $n \geqslant 2$. Every $k \in \mathbb{N}$ can be uniquely factored as

$$k = p_1(k) \cdots p_m(k) \cdot R(k)^2$$

where $R(k)$ is an integer and $p_1(k)$, ..., $p_m(k)$ are distinct prime numbers $\leqslant k$. If one multiplies out the product

$$\prod_{p \leqslant n, p \text{ prime}} (1 + \frac{1}{p}),$$

then for each $2 \leqslant k \leqslant n$ the expression $\frac{1}{p_1(k)\cdots p_m(k)}$ occurs as a summand. By the uniqueness of prime factorization one gets the inequality

$$\Big(\sum_{j=1}^{n} \frac{1}{j^2}\Big) \prod_{\substack{p \leqslant n \\ p \text{ prime}}} \Big(1 + \frac{1}{p}\Big) \geqslant \sum_{k=1}^{n} \frac{1}{k}.$$

We have $\sum_{k=1}^{n} \frac{1}{k} \geqslant \int_1^{n+1} t^{-1} dt \geqslant \log(n+1)$ and $\sum_{j=1}^{n} j^{-2} \leqslant 1 + \int_1^n t^{-2} dt \leqslant 2$. We also use the inequality $1 + x < e^x$ for $x > 0$ (from the power series expansion of the exponential function) and apply it to $x = 1/p$ to get

$$2 \prod_{\substack{p \leqslant n \\ p \text{ prime}}} e^{1/p} > \log(n+1).$$

Taking the logarithm we get for $n \geqslant 2$

$$\log 2 + \sum_{\substack{p \leqslant n \\ p \text{ prime}}} \frac{1}{p} > \log\log(n+1)$$

and thus the first assertion. This implies $\sum_{j=1}^{\infty} p_j^{-1} = \infty$. Since $(1 - p_j^{-1})^{-1} - 1 = O(p_j^{-2})$ the divergence of $\prod_{j=1}^{\infty} \left(1 - {p_j}^{-1}\right)^{-1}$ follows.

REMARK. This proof is from Niven [**58**], based on arguments of Euler.

3.88. EXERCISE.

(i) Prove by induction that

$$\sin(x) = 2^n \sin(2^{-n}x) \prod_{k=1}^{n} \cos(2^{-k}x).$$

(ii) Prove that

$$\sin(x) = x \prod_{k=1}^{\infty} \cos(2^{-k}x).$$

3.89. EXERCISE. The binomial coefficient $\binom{\alpha}{n}$ is usually defined for natural numbers α and n, but the definition extends to arbitrary real numbers α: define $\binom{\alpha}{0} = 1$ and

$$\binom{\alpha}{n} = \frac{1}{n!} \prod_{k=1}^{n} (\alpha - k + 1) = (-1)^n \prod_{k=1}^{n} \left(1 - \frac{\alpha + 1}{k}\right).$$

Prove:
(i) If $\alpha > -1$ then $\lim_{n\to\infty} \binom{\alpha}{n} = 0$.
(ii) If $\alpha \leqslant -1$ then the sequence $\binom{\alpha}{n}$ (with $n = 1, 2, \dots$) is divergent.

3.90. Exercise. Consider for $\alpha \in \mathbb{R}$ the binomial coefficients

$$\binom{n+\alpha}{n} = \prod_{k=1}^{n} (1 + \frac{\alpha}{k}).$$

We are interested in their behavior for large n.

(i) Show that $\prod_{k=1}^{\infty}(1 + \frac{\alpha}{k})$ diverges to 0 when $\alpha < 0$ and diverges to ∞ when $\alpha > 0$.

(ii) Show that for all $\alpha \in \mathbb{R}$ the quantities n^α and $\binom{n+\alpha}{n}$ are asymptotically proportional as $n \to \infty$, in the sense that

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\binom{n+\alpha}{n}$$

exists and is positive.

3.4. Nested radicals

In this section we explore an interesting variation on the theme of infinite series and investigate infinite expressions such as

$$\sqrt{1+\sqrt{2+\sqrt{3+\sqrt{4+\cdots}}}}. \tag{3.24}$$

We first give a precise meaning to such an expression. In analogy to the notion of partial sums of an infinite series it is natural to consider this expression as the limit of the sequence of finite nested radicals

$$\sqrt{1}, \sqrt{1+\sqrt{2}}, \sqrt{1+\sqrt{2+\sqrt{3}}}, \ldots$$

Nested radicals occur naturally in solutions of polynomial equations and in trigonometry. Let us give a general definition.

3.91. Definition. Let $(a_n)_{n=1}^{\infty}$ be a sequence of real numbers and $p \in \mathbb{R}$. We say that the *nested radical*

$$S = (a_1 + (a_2 + (a_3 + \cdots)^p)^p)^p \tag{3.25}$$

converges if the sequence of *partial nested radicals* $(s_n)_{n=1}^{\infty}$ given by

$$a_1^p, (a_1 + a_2^p)^p, (a_1 + (a_2 + a_3^p)^p)^p, \ldots$$

converges. More formally, for each n let the finite sequence $(s_{n,k})_{k=0}^{n}$ be defined recursively by letting $s_{n,0} = a_n$ and[1]

$$s_{n,k} = s_{n,k-1}^p + a_{n-k} \text{ for } k = 1, 2, \ldots, n$$

With the convention $a_0 = 0$ we then define $s_n = s_{n,n}$ and the value of the nested radical (3.25) is defined as

$$S = \lim_{n\to\infty} s_n$$

if the limit exists. Other terms used for nested radicals are *continued radicals*, *nested/continued pth powers*, *nested/continued* $\frac{1}{p}$*th roots*, etc. We prefer the term nested radical if $p \in (0, 1)$ and nested pth power otherwise.

[1] We only consider a nested radical to be defined if $s_{n,k-1}^p$ is a well-defined real number for all n, k; in particular this is the case if $a_n > 0$ for all n.

For $p = 1$ we recover the familiar definition of an infinite series, while for $p = -1$ we obtain a *(regular) continued fraction*

$$\cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}.$$

There exists an abundant literature on continued fractions. They are particularly useful in rational approximation. Our focus here is on the case $p \in (0, 1)$. A popular choice is $p = 1/2$ and the resulting nested radicals are also known as *nested or continued square-root.*

3.92. Example. The nested radical

$$S = \sqrt{1 + \sqrt{1 + \sqrt{1 + \cdots}}}$$

converges to the value $(\sqrt{5}+1)/2$. To see this notice that the sequence of partially nested radicals (s_n) satisfies $s_1 = 1$ and for $n = 1, 2, \ldots,$

$$s_{n+1} = \sqrt{1 + s_n}.$$

We see from the definition that $s_{n+1} > s_n > 0$ for all n. To check that the sequence converges, we write for $n \geqslant 2$,

$$s_{n+1} - s_n = \sqrt{1 + s_n} - \sqrt{1 + s_{n-1}} \leqslant \tfrac{1}{2}(s_n - s_{n-1}),$$

where we have used the mean value theorem for the function $f(x) = \sqrt{1 + x}$. Thus for all $m > n \geqslant 2$,

$$s_m - s_n \leqslant 2^{-(n-2)}(s_2 - s_1),$$

so (s_n) is a Cauchy sequence and therefore converges to a limit $S > 0$ satisfying the equation

$$S = \sqrt{1 + S}.$$

Solving the quadratic equation gives $S = (1 + \sqrt{5})/2$.

3.93. Exercise. Determine all $a > 0$ such that

$$\sqrt{a + \sqrt{a + \sqrt{a + \cdots}}} = a + \cfrac{1}{a + \cfrac{1}{a + \cdots}}$$

(in particular, show that the expressions on both sides are convergent).

3.94. Exercise. For $a > 0$ and $p \in \mathbb{R}$ consider the nested pth powers

$$(a + (a + (a + \cdots)^p)^p)^p.$$

Show the following:

(i) If $p < 1$, then the nested pth power converges.

(ii) If $p > 1$, show that the nested pth power converges if and only if $a \leqslant (p-1)p^{-\frac{p}{p-1}}$.

Let us from now on fix $p \in (0, 1)$, $r = 1/p$ and write

$$\sqrt[r]{a_1 + \sqrt[r]{a_2 + \sqrt[r]{a_3 + \cdots}}} = (a_1 + (a_2 + (a_3 + \cdots)^p)^p)^p.$$

How do we determine convergence of a nested radical if (a_n) is a given non-constant sequence of non-negative reals? As we shall see, there is a remarkably simple necessary and sufficient condition. We first record two important properties that are direct analogues of the corresponding properties for infinite series.

3.95. Observation (Comparison test). *If $(a_n)_{n=1}^{\infty}, (A_n)_{n=1}^{\infty}$ are sequences of reals satisfying*

$$0 \leqslant a_n \leqslant A_n \text{ for all } n = 1, 2, \ldots,$$

and $(s_n), (S_n)$ *denote the respective partial nested radicals, then* $s_n \leqslant S_n$ *for all* $n = 1, 2, \ldots$ *In particular, if* (S_n) *converges, then* (s_n) *converges and*

$$\sqrt[r]{a_1 + \sqrt[r]{a_2 + \sqrt[r]{a_3 + \cdots}}} \leqslant \sqrt[r]{A_1 + \sqrt[r]{A_2 + \sqrt[r]{A_3 + \cdots}}}.$$

PROOF. The definition of the partial nested radical sequence directly implies $s_{n,k} \leqslant S_{n,k}$ and thus $s_n \leqslant S_n$. The convergence claim follows, because (s_n) is an increasing sequence. □

3.96. OBSERVATION. *If* $a_n \geqslant 0$ *and* $c > 0$*, then the nested radical associated with* (a_n) *converges if and only if the nested radical associated with* $(c^{r^n} a_n)$ *converges and we have*

$$c \cdot \sqrt[r]{a_1 + \sqrt[r]{a_2 + \sqrt[r]{a_3 + \cdots}}} = \sqrt[r]{c^r a_1 + \sqrt[r]{c^{r^2} a_2 + \sqrt[r]{c^{r^3} a_3 + \cdots}}}.$$

The reader should convince themselves of this fact using the definitions directly.

3.97. THEOREM (Herschfeld). *Let* $(a_n)_{n=1}^{\infty}$ *be a non-negative sequence and* $p \in (0, 1)$*. Then the nested radical*

$$\sqrt[r]{a_1 + \sqrt[r]{a_2 + \sqrt[r]{a_3 + \cdots}}}$$

converges if and only if

$$\sup_{n=1,\ldots,\infty} a_n^{r^{-n}} < \infty.$$

PROOF. First suppose that the sequence s_n of partial nested radicals converges. Using that $a_n \geqslant 0$,

$$s_n \geqslant a_n^{r^{-n}}.$$

Indeed, for $1 \leqslant k \leqslant n$,

$$s_{n,k} = \sqrt[r]{s_{n,k-1}} + a_{n-k} \geqslant \sqrt[r]{s_{n,k-1}},$$

so $s_n = s_{n,n} \geqslant s_{n,0}^{r^{-n}} = a_n^{r^{-n}}$. Since convergent sequences are bounded we obtain that $(a_n^{r^{-n}})$ must be bounded, as claimed. Next, suppose that $\sup_n a_n^{r^{-n}} < \infty$. Then there exists a constant $C \in (0, \infty)$ so that

$$a_n \leqslant C^{r^n} \text{ for all } n = 1, 2, \ldots \tag{3.26}$$

By Observation 3.96 the nested radical

$$\sqrt[r]{C^r + \sqrt[r]{C^{r^2} + \sqrt[r]{C^{r^3} + \cdots}}}$$

associated with the sequence (C^{r^n}) converges with value

$$C \sqrt[r]{1 + \sqrt[r]{1 + \sqrt[r]{1 + \cdots}}},$$

because the nested radical in this display converges (by the argument from Example 3.92). By (3.26) and the comparison test, Observation 3.95, we thus have

$$\sqrt[r]{a_1 + \sqrt[r]{a_2 + \sqrt[r]{a_3 + \cdots}}} \leqslant C \sqrt[r]{1 + \sqrt[r]{1 + \sqrt[r]{1 + \cdots}}} < \infty.$$

□

3.98. EXAMPLE. The nested radical (3.24) with $a_n = n$, $r = 2$ converges because

$$\lim_{n \to \infty} n^{2^{-n}} = 1.$$

3.99. Corollary (Pólya's criterion). *Let $r \in (1,\infty)$ and let $(a_n)_{n=1}^{\infty}$ be a non-negative sequence and let*

$$L = \limsup_{n\to\infty} \frac{\log\log a_n}{n},$$

where we agree by convention that if $a_n \leqslant 1$ along an infinite subsequence, then $L = -\infty$. Then the nested radical $\sqrt[r]{a_1 + \sqrt[r]{a_2 + \sqrt[r]{a_3 + \cdots}}}$

(i) converges if $L < \log r$, and
(ii) diverges if $L > \log r$.

Proof.

(i) If $L < \log r$, there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$,

$$\log\log a_n < n \log r.$$

Therefore $\log a_n < e^{n\log(r)} = r^n$, so

$$a_n^{r^{-n}} < e,$$

which by Theorem 3.97 implies convergence of the nested radical.

(ii) If $L > \log r$, then there exist $\varepsilon > 0$ and infinitely many $n \in \mathbb{N}$ so that

$$\log\log a_n \geqslant n(\log r + \varepsilon).$$

Thus, $\log a_n \geqslant r^n e^{n\varepsilon}$ and

$$a_n^{r^{-n}} \geqslant e^{e^{n\varepsilon}},$$

which is unbounded as $n \to \infty$ and thus gives divergence by Theorem 3.97.

□

3.100. Exercise. Pólya's criterion leaves open the case $L = \log r$. Show that in this case, the nested radical $\sqrt[r]{a_1 + \sqrt[r]{a_2 + \sqrt[r]{a_3 + \cdots}}}$ converges if and only if

$$\limsup_{n\to\infty} n \cdot \Big(\frac{\log\log a_n}{n} - \log r\Big) < \infty.$$

3.101. Exercise. Extend the notion of nested radical to infinite expressions of the form

$$\sqrt[r]{a_1 + b_1\sqrt[r]{a_2 + b_2\sqrt[r]{a_3 + b_3\sqrt[r]{\cdots}}}}$$

and show that for $a_n \geqslant 0, b_n > 0$ this nested radical converges if and only if

$$\sup_n a_n^{r^{-n}} \prod_{k=1}^{n-1} b_k^{r^{-k}} < \infty.$$

3.102. Exercise. Define the nested radical sequence

$$s_n = \underbrace{\sqrt{2 + \sqrt{2 + \cdots\sqrt{2}}}}_{n \text{ square roots}}$$

(i) Show by induction on n that $s_n = 2\cos(2^{-n-1}\pi) < 2$ and conclude that

$$2 = \sqrt{2 + \sqrt{2 + \sqrt{2 + \cdots}}}.$$

(ii) Prove

$$\pi = \lim_{n\to\infty} 2^{n+1}\sqrt{2 - s_n}$$

3.103. Exercise. Prove *Viète's formula*

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2} \cdot \frac{\sqrt{2+\sqrt{2}}}{2} \cdot \frac{\sqrt{2+\sqrt{2+\sqrt{2}}}}{2} \cdot \dots.$$

3.104. Exercise. Show that the sequence

$$\sqrt{1}, \sqrt{1+2\sqrt{1}}, \sqrt{1+2\sqrt{1+3\sqrt{1}}}, \dots$$

converges to 3. (This example is due to Ramanujan.)

3.105. Exercise. Let a and b be positive real numbers, and let $n > 1$ be an integer. Show that the polynomial equation $x^n - bx - a = 0$ has a unique solution among positive real numbers x and that its value is given by

$$\sqrt[n]{a + b\sqrt[n]{a + b\sqrt[n]{a + b\sqrt[n]{\cdots}}}}.$$

Remark. Theorem 3.97 and the above proof of Corollary 3.99 are from Herschfeld [**31**]. Pólya's criterion and other exercises on nested radicals can be found in Pólya–Szegö [**59**].

3.5. The Euler–Maclaurin formula

Let a, b be integers and f a function defined on the real numbers. How can we compare the value of the sum

$$\sum_{a<n\leqslant b} f(n)$$

to the value of the integral

$$\int_a^b f(t)\,dt\ ?$$

It turns out there is a natural way to do this using integration by parts that leads to a powerful tool, called the Euler–Maclaurin formula. This formula can be used to derive a plethora of interesting formulas and approximations. We will discuss a few of them: Faulhaber's formula for sums of powers, precise error terms for the trapezoid rule of numerical integration, approximations of the Euler–Mascheroni constant, and an asymptotic for $n!$ for large n (Stirling's formula).

To start we define a sequence of polynomials, called the *Bernoulli polynomials* by the following recursion:

$$\begin{aligned} &b_0(x) = 1 \\ &b_n'(x) = n b_{n-1}(x), \quad \int_0^1 b_n(x)\,dx = 0, \quad n \geqslant 1. \end{aligned} \tag{3.27}$$

Inductively, this uniquely determines $b_n(x)$ as a polynomial of degree n. The constants $B_n = b_n(0)$ are the *Bernoulli numbers*. The first few Bernoulli polynomials are given by

$$\begin{aligned} &b_0(x) = 1, \quad b_1(x) = x - \tfrac{1}{2}, \quad b_2(x) = x^2 - x + \tfrac{1}{6}, \quad b_3(x) = x^3 - \tfrac{3}{2}x^2 + \tfrac{1}{2}x, \\ &b_4(x) = x^4 - 2x^3 + x^2 - \tfrac{1}{30}, \quad b_5(x) = x^5 - \tfrac{5}{2}x^4 + \tfrac{5}{3}x^3 - \tfrac{1}{6}x, \\ &b_6(x) = x^6 - 3x^5 + \tfrac{5}{2}x^4 - \tfrac{1}{2}x^2 + \tfrac{1}{42}. \end{aligned}$$

The corresponding values for the Bernoulli numbers $B_n = b_n(0)$ are

$$B_0 = 1,\ B_1 = -\tfrac{1}{2},\ B_2 = \tfrac{1}{6},\ B_3 = 0,\ B_4 = -\tfrac{1}{30},\ B_5 = 0,\ B_6 = \tfrac{1}{42}.$$

Note that the condition $\int_0^1 b_k(x)dx = 0$ for $k \geqslant 1$ yields $b_n(0) = b_n(1)$ for all $n \geqslant 2$.

Finally, define the *Bernoulli functions* ψ_k as 1-periodic functions on $\mathbb{R}$ that coincide on $[0,1)$ with b_k. In other words $\psi_k(x) = b_k(x - \lfloor x \rfloor)$ for $x \in \mathbb{R}$. In particular we get

$$\psi_1(x) = x - \lfloor x \rfloor - \tfrac{1}{2},$$

a sawtooth function with jumps at the integers, but for $k \geqslant 2$ the functions ψ_k are 1-periodic continuous functions on $\mathbb{R}$. Moreover, the derivatives up to order $k-2$ are continuous on $\mathbb{R}$.

The proof of the following formula uses integration by parts. The reader may notice the formal similarity to the proof of Taylor's theorem 2.49.

3.106. Theorem (Euler–Maclaurin formula). *Let $k \in \mathbb{N}_0$. Let $a < b$ be integers and let $f \in C^{k+1}([a,b])$. Then*

$$\sum_{a<n\leqslant b} f(n) = \int_a^b f(t)\,dt + \sum_{j=0}^{k} \frac{(-1)^{j+1}B_{j+1}}{(j+1)!}\big(f^{(j)}(b) - f^{(j)}(a)\big) + \frac{(-1)^k}{(k+1)!}\int_a^b \psi_{k+1}(t) f^{(k+1)}(t)\,dt.$$

Proof. Denote the statement of the theorem by $\mathcal{A}(k)$; we prove it by induction.

The case $k = 0$. Let $m \in \mathbb{Z}$ such that $a \leqslant m \leqslant b-1$. We use that for $m \leqslant t \leqslant m+1$ the Bernoulli function ψ_1 is given by $\psi_1(t) = (t - m - \frac{1}{2})$. Integration by parts yields

$$\int_m^{m+1} (t - m - \tfrac{1}{2}) f'(t)\,dt = \frac{1}{2}(f(m+1) + f(m)) - \int_m^{m+1} f(t)\,dt$$

and summing m from a to $b-1$ yields

$$\frac{1}{2}\sum_{m=a}^{b-1}(f(m+1)+f(m)) = \sum_{m=a+1}^{b-1} f(m) + \frac{1}{2}(f(b)+f(a)) = \sum_{m=a+1}^{b} f(m) + \frac{1}{2}(f(a)-f(b))$$

and thus

$$\int_a^b \psi_1(t) f'(t) dt = \sum_{n=a+1}^{b} f(n) - \frac{1}{2}(f(b)-f(a)) - \int_a^b f(t)\,dt.$$

Since $B_1 = -1/2$ this yields statement $\mathcal{A}(0)$.

The implication $\mathcal{A}(k-1) \implies \mathcal{A}(k)$ for $k \geqslant 1$. We can now assume $f \in C^{k+1}([a,b])$. Recall that for $k \geqslant 1$ the function ψ_{k+1} is continuous in $[a,b]$. We integrate by parts on the intervals $[m, m+1]$, for $a \leqslant m \leqslant b-1$, and get

$$\begin{aligned}
&\int_m^{m+1} \frac{(-1)^k}{(k+1)!}\psi_{k+1}(t) f^{(k+1)}(t)\,dt \\
&= \frac{(-1)^k}{(k+1)!}\Big(\psi_{k+1}(m+1)f^{(k)}(m+1) - \psi_{k+1}(m) f^{(k)}(m) - \int_m^{m+1} \psi'_{k+1}(t) f^{(k)}(t)\,dt\Big) \\
&= \frac{(-1)^k}{(k+1)!} B_{k+1}\big(f^{(k)}(m+1) - f^{(k)}(m)\big) - \frac{(-1)^k}{k!}\int_m^{m+1} \psi_k(t) f^{(k)}(t)\,dt
\end{aligned}$$

where in the last line we have used $\psi_{k+1}(m+1) = \psi_{k+1}(m) = B_{k+1}$ for $k \geqslant 1$ and $\psi'_{k+1} = (k+1)\psi_k$. Consequently, after summing in $m = a, \dots, b-1$,

$$\frac{(-1)^{k-1}}{k!}\int_a^b \psi_k(t) f^{(k)}(t)\,dt = \frac{(-1)^{k+1}}{(k+1)!} B_{k+1}\big(f^{(k)}(b) - f^{(k)}(a)\big) + \frac{(-1)^k}{(k+1)!}\int_a^b \psi_{k+1}(t) f^{(k+1)}(t)\,dt.$$

Using this together with the inductive hypothesis $\mathcal{A}(k-1)$ yields $\mathcal{A}(k)$. □

3.107. Proposition.

(i) The series $\sum_{k=0}^{\infty} b_k(x)\frac{t^k}{k!}$ *converges absolutely for* $x \in [0,1]$ *and* $|t| < 1$.

(ii) For $x \in [0,1]$ *and* $t \neq 0$

$$\sum_{k=0}^{\infty} b_k(x)\frac{t^k}{k!} = \frac{te^{tx}}{e^t - 1}.$$

In particular,

$$\sum_{k=0}^{\infty} B_k\frac{t^k}{k!} = \frac{t}{e^t - 1}\,.$$

(iii) For $\ell = 1, 2, \ldots,$

$$B_{2\ell+1} = 0.$$

Proof.

(i) We prove $|b_k(x)| \leqslant k!$ by induction. This holds for $k = 0$. Since

$$\int_0^1 b_{k+1}(x)ds = 0$$

there is $x_0 \in [0,1]$ with $b_{k+1}(x_0) = 0$ and therefore we get

$$b_{k+1}(x) = (k+1)\int_{x_0}^{x} b_k(t)dt,$$

which implies $|b_{k+1}(x)| \leqslant (k+1)!$ by inductive hypothesis.

(ii) Define

$$G(t,x) = \sum_{k=0}^{\infty} b_k(x)\frac{t^k}{k!}.$$

From the recursion formula this can be differentiated in x and we get for fixed t

$$\frac{d}{dx}G(t,x) = \sum_{k=1}^{\infty} kb_{k-1}(x)\frac{t^k}{k!} = t \cdot G(t,x),$$

which implies that

$$G(t,x) = A(t)e^{tx}$$

for some number $A(t)$. By the relation $\int_0^1 b_k(x)dx = 0$ for $k \geqslant 1$ and $\int_0^1 b_0(x)dx = 1$ we get

$$\int_0^1 G(t,x)\,dx = 1$$

for all t. This implies

$$A(t) = \frac{t}{e^t - 1}.$$

(iii) From

$$A(t) - A(-t) = -t,$$

we conclude $A(t) + t/2 = A(t) - B_1 t$ is an even function. With (ii) in mind this shows $B_{2\ell+1} = 0$, for $\ell = 1, 2, \ldots$. □

3.108. Exercise. Let B_n be the Bernoulli numbers and $b_n(x)$ be the Bernoulli polynomials.

(i) Show that $b_n(x) = \sum_{k=0}^{n} B_k\binom{n}{k}x^{n-k}$ for $n > 1$.

(ii) Show that the Bernoulli numbers can be defined by the recursion $B_0 = 1$, and $\sum_{k=0}^{n-1}\binom{n}{k}B_k = 0$, $n > 1$.

3.109. Exercise. Show the following identities for Bernoulli numbers and polynomials:

(i) $b_n(x+1) - b_n(x) = nx^{n-1}$,

(ii) $b_n(1-x) = (-1)^n b_n(x)$,
(iii) $b_n(\frac{1}{2}) = (2^{1-n}-1)B_n$.

REMARK. As ℓ increases, $B_{2\ell}$ grows rapidly. In Exercise 7.85 we will prove the formula

$$B_{2\ell} = (-1)^{\ell-1}\frac{2(2\ell)!}{(2\pi)^{2\ell}}\zeta(2\ell), \tag{3.28}$$

where

$$\zeta(s) = \sum_{\nu=1}^{\infty} \nu^{-s}$$

for $s > 1$. In particular, this provides an asymptotic formula for $B_{2\ell}$ as $\ell \to \infty$ (see Exercise 3.117).

3.5.1. Summing powers of integers. The *Bernoulli formulas* (also known as the *Faulhaber formulas*) generalize the familiar formulas

$$\sum_{k=1}^{n} k = \frac{n^2}{2} + \frac{n}{2}, \qquad \sum_{k=1}^{n} k^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$$

to higher powers.

3.110. THEOREM. *For all $p \in \mathbb{N}$,*

$$\sum_{k=1}^{n} k^p = \frac{n^{p+1}}{(p+1)} + \frac{n^p}{2} + \sum_{\substack{1 \leqslant j \leqslant p \\ j \text{odd}}} B_{j+1}\binom{p+1}{j+1}\frac{n^{p-j}}{p+1}$$

PROOF. Apply Theorem 3.106 for $f(x) = x^p$ (with $p \in \mathbb{N}_0$, setting $k = p$, $a = 0$ and $b = n$, and $\int_0^n x^p = \frac{n^{p+1}}{p+1}$). This yields

$$\sum_{k=1}^{n} k^p = \frac{n^{p+1}}{p+1} - B_1 n^p + \sum_{j=1}^{p} \frac{(-1)^{j+1}}{(j+1)!} B_{j+1} n^{p-j} p(p-1)\cdots(p-j+1).$$

Now observe $\frac{p(p-1)\cdots(p-j+1)}{(j+1)!} = \frac{p!}{(j+1)!(p-j)!} = \frac{1}{p+1}\binom{p+1}{j+1}$. □

3.5.2. The trapezoid rule. The purpose of the trapezoid rule is to approximate the value of $\int_a^b f(t)\,dt$. We now take a general interval $[a,b]$, with $a, b \in \mathbb{R}$ (not necessarily integers) and split into n subintervals I_k of equal length $\frac{b-a}{n}$, i.e. $I_k = [x_{k-1}, x_k]$ where $x_k = a + k\frac{b-a}{n}$. The idea is to approximate the integral over I_k by the area of the trapezoid with two vertical sides and corners at $(x_{k-1}, f(x_{k-1}))$ and $(x_k, f(x_k))$. Thus the trapezoid sum

$$\begin{aligned} \mathfrak{T}_n(f;a,b) &= \sum_{k=1}^{n} \frac{1}{2}\big(f(a+(k-1)\tfrac{b-a}{n}) + f(a+k\tfrac{b-a}{n})\big)\frac{b-a}{n} \\ &= \frac{b-a}{n}\Big(\frac{f(a)}{2} + \sum_{k=1}^{n-1} f(a+k\tfrac{b-a}{n}) + \frac{f(b)}{2}\Big) \end{aligned} \tag{3.29}$$

should be a good approximation for the integral provided that f is sufficiently smooth.

3.111. THEOREM (Trapezoid rule).

(i) Let $f \in C^2([a,b])$. Then

$$\mathfrak{T}_n(f;a,b) - \int_a^b f(x)\,dx = \frac{(b-a)^3}{12n^2} f''(\xi)$$

for some $\xi \in [a,b]$.

(ii) Set $h = (b-a)/n$. For $M = 0, 1, 2, \dots$ and $f \in C^{2M+2}([a,b])$,

$$\mathfrak{T}_n(f;a,b) - \int_a^b f(x)\,dx = \sum_{1\leqslant 2\ell+1\leqslant 2M} B_{2\ell}\frac{h^{2\ell}}{(2\ell)!}\big(f^{(2\ell-1)}(b) - f^{(2\ell-1)}(a)\big) + E_{2M+1}(a,b),$$

where

$$E_{2M+1}(a,b) = -\frac{h^{2M+2}}{(2M+2)!}\int_a^b \big(\psi_{2M+2}(\tfrac{x-a}{h}) - B_{2M+2}\big) f^{(2M+2)}(x)\,dx.$$

Proof. (ii) Set $g(s) = f(a+sh)$, for $0 \leqslant s \leqslant n$. Compute

$$\mathfrak{T}_n(f;a,b) = \frac{h}{2}\big(f(a) - f(b)\big) + h\sum_{k=1}^{n} f(a+kh) = \frac{h}{2}\big(g(0) - g(n)\big) + h\sum_{k=1}^{n} g(k)$$
$$= \frac{h}{2}\big(g(0) - g(n)\big) + h\int_0^n g(s)ds - B_1 h(g(n) - g(0))$$
$$+ h\sum_{1\leqslant j\leqslant 2M} (-1)^{j+1}\frac{B_{j+1}}{(j+1)!}\big(g^{(j)}(n) - g^{(j)}(0)\big) + \frac{h}{(2M+1)!}\int_0^n \psi_{2M+1}(t)g^{(2M+1)}(t)dt,$$

by the Euler–Maclaurin theorem. Since $B_1 = -1/2$ the terms with $g(n) - g(0)$ drop out. The terms with even $j \geqslant 2$ also drop out since $B_{j+1} = 0$ for $j = 2, 4, 6, \dots$ (see Proposition 3.107). Moreover we use the recursion formula for the Bernoulli polynomials and functions, i.e. $(2M+2)\psi_{2M+1} = \psi'_{2M+2}$, and integration by parts to see that

$$\frac{1}{(2M+1)!}\int_0^n \psi_{2M+1}(t)g^{(2M+1)}(t)\,dt = \frac{-1}{(2M+2)!}\int_0^n \big(\psi_{2M+2}(t) - B_{2M+2}\big)g^{(2M+2)}(t)\,dt.$$

Using this above and expressing everything in terms of f yields

$$\mathfrak{T}_n(f;a,b) = \int_a^b f(x)dx + \sum_{1\leqslant 2\ell+1\leqslant 2M} B_{2\ell}\frac{h^{2\ell}}{(2\ell)!}\big(f^{(2\ell-1)}(b) - f^{(2\ell-1)}(a)\big)$$
$$- \frac{h^{2M+2}}{(2M+2)!}\int_a^b \big(\psi_{2M+2}(\tfrac{x-a}{h}) - \psi_{2M+2}(0)\big) f^{(2M+2)}(x)\,dx$$

and we have proven part (ii).

(i) We use the special case $M = 0$ with $f \in C^2$ to get

$$\mathfrak{T}_n(f;a,b) - \int_a^b f(x)dx = -\frac{h^2}{2}\int_a^b \big(\psi_2(\tfrac{x-a}{h}) - B_2\big) f''(x)dx.$$

Let $x_{k-1} = a + (k-1)h$ and use on $[x_{k-1}, x_k]$ the change of variable $x = x_{k-1} + sh$. We compute (using the 1 periodicity of ψ_2, and $\psi_2(s) - B_2 = s^2 - s$ on $[0,1]$)

$$\int_{x_{k-1}}^{x_k} \big(\psi_2(\tfrac{x-a}{h}) - B_2\big) f''(x)dx = h\int_0^1 (\psi_2(k-1+s) - B_2) f''(x_{k-1} + sh)ds$$
$$= -h\int_0^1 (s - s^2) f''(x_{k-1} + sh)ds.$$

Using this in the preceding formula and summing in k we get

$$\mathfrak{T}_n(f;a,b) - \int_a^b f(x)dx = \frac{nh^3}{2}\int_0^1 (s-s^2)\frac{1}{n}\sum_{k=1}^{n} f''(x_{k-1} + sh)ds$$

Note that $s - s^2$ is nonnegative on $[0,1]$, $\int_0^1 (s-s^2)ds = \frac{1}{6}$ and hence

$$\frac{1}{6}\min_{[a,b]} f'' \leqslant \int_0^1 (s-s^2)\frac{1}{n}\sum_{k=1}^{n} f''(x_{k-1}+sh)ds \leqslant \frac{1}{6}\max_{[a,b]} f''$$

Hence by the intermediate value theorem there is a $\xi \in [a,b]$ such that the integral in the display is equal to $\frac{1}{6}f''(\xi)$. Consequently

$$\mathfrak{T}_n(f;a,b) - \int_a^b f(x)dx = \frac{nh^3}{12}f''(\xi) = \frac{(b-a)^3}{12n^2}f''(\xi).$$

□

The following exercise gives an example (a special case of *Richardson extrapolation*) where part (i) of the theorem can be used to generate other formulas for numerical integration that allow for better error terms as $n \to \infty$.

3.112. Exercise. Suppose $f \in C^4([a,b])$. Show that there is a constant A (independent of a, b, n, f) such that

$$\Big|\frac{4\mathfrak{T}_{2n}(f;a,b) - \mathfrak{T}_n(f;a,b)}{3} - \int_a^b f(x)dx\Big| \leqslant A\frac{(b-a)^5}{n^4}\max_{[a,b]}|f^{(4)}|$$

Give a concrete (not necessarily best possible) value for A.

3.5.3. Revisiting the Euler–Mascheroni limit. As a further illustration of how the Euler–Maclaurin formula works we apply it to obtain refinements of the limiting relation $\gamma = \lim_{n\to\infty}(\sum_{k=1}^{n}\frac{1}{k} - \log n)$ from Exercise 3.14.

3.113. Theorem. *For $n \in \mathbb{N}$ and $k \geqslant 2$ we have*

$$\sum_{\ell=1}^{n}\frac{1}{\ell} = \log n + \gamma - \sum_{j=1}^{k}\frac{B_j}{j}\frac{1}{n^j} + \int_n^\infty \psi_k(t)t^{-k-1}dt$$

and

$$\gamma = \frac{1}{2} + \sum_{j=2}^{k}\frac{B_j}{j} - \int_1^\infty \psi_k(t)t^{-k-1}dt$$

3.114. Corollary.

(i)

$$\Big|\gamma - \Big(\frac{1}{2} + \frac{1}{12} - \frac{1}{120}\Big)\Big| \leqslant \frac{1}{120}.$$

(ii) For $n \in \mathbb{N}$ there is $\theta_n \in [0,1]$ such that

$$\sum_{\ell=1}^{n}\frac{1}{\ell} = \log n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{\theta_n}{60n^4}.$$

Proof. Use Theorem 3.113 with $k = 4$ and since $B_1 = -1/2$, $B_2 = 1/6$, $B_3 = 0$ and $B_4 = -1/30$ we get

$$\sum_{\ell=1}^{n}\frac{1}{\ell} = \log n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} + \int_n^\infty t^{-5}\psi_4(t)dt$$

Now ψ_4 is the 1-periodization of $b_4(x) = x^4 - 2x^3 + x^2 - \frac{1}{30}$ and it is straightforward to check that $|b_4(x)| \leqslant \frac{1}{30}$ on $[0,1]$. Hence $|\int_n^\infty t^{-5}\psi_4(t)dt| \leqslant \frac{1}{120n^4}$. We get

$$0 \leqslant \frac{1}{120n^4} + \int_n^\infty t^{-5}\psi_4(t)dt \leqslant \frac{1}{60n^4}$$

which establishes the last error term. In the same way we obtain the approximation of γ from the second statement in Theorem 3.113. Moreover

$$\gamma = \frac{1}{2} + \frac{1}{12} - \frac{1}{120} - \int_1^\infty t^{-5}\psi_4(t)dt$$

and $\int_1^\infty t^{-5}|\psi_4(t)|dt \leqslant \frac{1}{120}$. □

Proof of Theorem 3.113. We apply the Euler–Maclaurin formula with $f(t) = t^{-1}$, $a = 1$, $b = n$. Since $f^{(j)}(t) = (-1)^j j! t^{-j-1}$ and $B_1 = -1/2$ we get for $k \geqslant 1$

$$\begin{aligned} 1 + \sum_{2\leqslant \ell \leqslant n} \frac{1}{\ell} &= 1 + \log n - \sum_{j=0}^{k} \frac{B_{j+1}}{j+1}\Big(\frac{1}{n^{j+1}} - 1\Big) - \int_1^n \frac{\psi_{k+1}(t)}{t^{k+2}}dt \\ &= \log n + \frac{1}{2} + \sum_{j=2}^{k+1} \frac{B_j}{j} - \sum_{j=1}^{k+1} \frac{B_j}{j}\frac{1}{n^j} - \int_1^n \frac{\psi_{k+1}(t)}{t^{k+2}}dt. \end{aligned}$$

Letting $n \to \infty$ yields

$$\gamma = \frac{1}{2} + \sum_{j=2}^{k+1} \frac{B_j}{j} - \int_1^\infty \frac{\psi_{k+1}(t)}{t^{k+2}}dt$$

Putting the two formulas together we get

$$\sum_{\ell=1}^{n} \frac{1}{\ell} = \log n + \gamma - \sum_{j=1}^{k+1} \frac{B_j}{j}\frac{1}{n^j} + \int_n^\infty \frac{\psi_{k+1}(t)}{t^{k+2}}dt$$

and the asserted formulas in Theorem 3.113 are established by replacing k with $k-1$. □

3.5.4. Factorials and shifted variants. Finally, we discuss a useful asymptotic formula for $n!$ as $n \to \infty$ and some of its relatives. For $\alpha \in \mathbb{R}$ we may consider the *shifted factorial*

$$\mathfrak{p}_\alpha(n) = \prod_{k=0}^{n-1}(\alpha + k) = \alpha(\alpha+1)\cdots(\alpha+n-1)$$

also known as *ascending factorial* and sometimes as *Pochhammer function*. Observe that for $\alpha = 1$ we recover $n! = 1 \cdot 2 \cdots n$, the standard factorial. For $n = 0$ we set $\mathfrak{p}_\alpha(0) = 1$.

Remark.

(i) When $\alpha = N \in \mathbb{N}$, then $\mathfrak{p}_N(n)$ gives the number of injective functions from a set of cardinality n to a set of cardinality $n + N - 1$. This is also known as the number of n-permutations in a set of cardinality $n + N - 1$.

(ii) The shifted factorials $\mathfrak{p}_\alpha(n)$ are sometimes denoted by $(\alpha)^n$ or $(\alpha)_n$ (e.g. Andrews–Askey–Roy [**2**]), also sometimes known as the *Pochhammer symbols.* Shifted factorials can also be expressed in terms of the Γ-function (Definition 2.76) as $\mathfrak{p}_\alpha(n) = \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)}$, provided $\alpha \notin \{0, -1, -2, \dots\}$.

(iii) There is also the notion of a *descending factorial*, defined by the quantity $\prod_{k=0}^{n-1}(\alpha - k)$. This quantity is equal to $n!\binom{\alpha}{n}$, with the generalized binomial coefficients $\binom{\alpha}{n}$ defined in Exercise 3.89.

3.115. Theorem (Asymptotics for the shifted factorials). *Let $\alpha \in \mathbb{R}\backslash\{0, -1, -2, \dots\}$. There exists a constant $\mathcal{C}(\alpha) \neq 0$, such that for $n \geqslant 1$*

$$\mathfrak{p}_\alpha(n) = \mathcal{C}(\alpha) n^{n+\alpha-\frac{1}{2}} e^{-n}(1 + O(\tfrac{1}{n})).$$

Moreover $\mathcal{C}(\alpha) > 0$ if $\alpha > 0$, and $(-1)^m \mathcal{C}(\alpha) > 0$ if $\alpha \in (-m, -m+1)$, $m \in \mathbb{N}$.

PROOF. We shall first consider the case $\alpha > 0$. In this case we have $\log(\mathfrak{p}_\alpha(n)) = \sum_{k=0}^{n-1} \log(\alpha+k)$ and therefore we may apply the Euler-Maclaurin formula with $f(t) = \log(\alpha+t)$, for $t \geqslant 0$. This yields (ignoring the first term in the sum)

$$\begin{aligned}\sum_{k=1}^{n-1} \log(\alpha+k) &= \int_0^{n-1} \log(\alpha+t)dt - B_1(\log(\alpha+n-1) - \log\alpha) + \int_0^{n-1} \frac{\psi_1(t)}{\alpha+t}dt \\ &= (\alpha+n-1)\log(\alpha+n-1) - (\alpha+n-1) - \alpha\log\alpha + \alpha \\ &\qquad + \frac{1}{2}\big(\log(\alpha+n-1) - \log\alpha\big) + \int_0^{n-1} \frac{\psi_1(t)}{\alpha+t}dt.\end{aligned}$$

Recall that $\int_k^{k+1} \psi_1(t)dt = 0$ and therefore

$$\begin{aligned}\Big|\int_k^{k+1} \psi_1(t)\frac{dt}{\alpha+t}\Big| &= \Big|\int_k^{k+1} \psi_1(t)\Big(\frac{1}{\alpha+t} - \frac{1}{\alpha+k}\Big)dt\Big| \\ &\leqslant \int_k^{k+1} \frac{|k-t|}{(\alpha+t)(\alpha+k)}dt \leqslant \frac{1}{2(\alpha+k)^2}\end{aligned}$$

(use $|\psi_1(t)| \leqslant 1/2$). Hence

$$\Big|\int_n^{\infty} \psi_1(t)\frac{dt}{\alpha+t}\Big| \leqslant \sum_{k=n}^{\infty} \frac{1}{2(\alpha+k)^2} \leqslant cn^{-1}$$

and thus $\int_0^{n-1} \frac{\psi_1(t)}{\alpha+t}dt = A_\alpha + O(1/n)$ for some $A_\alpha \in \mathbb{R}$. We are taking exponentials and obtain

$$\mathfrak{p}_\alpha(n) = \alpha\exp\big(\sum_{k=1}^{n-1} \log(\alpha+k)\big) = \alpha(\alpha+n-1)^{\alpha+n-1+\frac{1}{2}} e^{1-n}\alpha^{-\alpha-\frac{1}{2}} e^{A_\alpha+O(1/n)}.$$

Check (provide details!) that

$$\frac{(\alpha+n-1)^{\alpha+n-\frac{1}{2}}}{n^{n+\alpha-\frac{1}{2}}} = \big(1+\frac{\alpha-1}{n}\big)^n\big(1+\frac{\alpha-1}{n}\big)^{\alpha-1/2} = e^{\alpha-1}(1+O(\tfrac{1}{n}))$$

and then

$$\begin{aligned}\mathfrak{p}_\alpha(n) &= n^{n+\alpha-\frac{1}{2}} e^{\alpha-1}(1+O(\tfrac{1}{n}))e^{1-n}\alpha^{\frac{1}{2}-\alpha} e^{A_\alpha+O(\frac{1}{n})} \\ &= \mathcal{C}(\alpha)n^{n+\alpha-\frac{1}{2}}e^{-n}(1+O(\tfrac{1}{n}))\end{aligned}$$

where $\mathcal{C}(\alpha) = e^\alpha\alpha^{\frac{1}{2}-\alpha}e^{A_\alpha} > 0$. Hence we obtain the stated asymptotics for $\alpha > 0$.

Finally assume $\alpha < 0$, with $\alpha \in (-m, 1-m)$, $m \in \mathbb{N}$. Set $\beta = \alpha + m$ so that $\beta \in (0,1)$. We observe $\mathfrak{p}_\alpha(n) = u_m(\beta)\mathfrak{p}_\beta(n-m)$ with $u_m(\beta) = \prod_{k=1}^m(\beta-k) \neq 0$; note that $u_m(\beta) > 0$ if m is even and $u_m(\beta) < 0$ if m is odd. Also for $n \to \infty$

$$\begin{aligned}&\mathfrak{p}_\beta(n-m) = \mathcal{C}(\beta)(n-m)^{n-m+\beta-\frac{1}{2}}e^{-n+m}(1+O(\tfrac{1}{n})) \\ &= \mathcal{C}(\beta)e^{-n}n^{n+\alpha-\frac{1}{2}}(1+O(\tfrac{1}{n}))e^m(1-\tfrac{m}{n})^{n+\alpha-\frac{1}{2}} = \mathcal{C}(\beta)e^{-n}n^{n+\alpha-\frac{1}{2}}(1+O(\tfrac{1}{n})).\end{aligned}$$

The asserted asymptotics follows with

$$\mathcal{C}(\alpha) = u_m(\beta)\mathcal{C}(\beta) \neq 0,$$

and we have $(-1)^m\mathcal{C}(\alpha) > 0$ if $\alpha \in (-m, 1-m)$. □

For the special case $\alpha = 1$ we get an asymptotic for $n!$. In this case we compute the constant $\mathcal{C}(1)$ using the Wallis product in Example 3.86.

3.116. THEOREM (Stirling's formula). *For $n \geqslant 1$,*

$$n! = n^n e^{-n}\sqrt{2\pi n}(1 + O(\tfrac{1}{n}))$$

PROOF. By Theorem 3.115 for $\alpha = 1$ there is $\mathcal{C} > 0$ such that

$$n! = \mathcal{C} n^{n+\frac{1}{2}} e^{-n}\big(1 + O(\tfrac{1}{n})\big).$$

It remains to compute $\mathcal{C}$. By Example 3.86 we have

$$\frac{\pi}{2} = \lim_{n\to\infty} W_n \text{ where } W_n = \prod_{k=1}^{n} \frac{4k^2}{(2k-1)(2k+1)} = \frac{2^{4n}(n!)^4}{((2n)!)^2(2n+1)}\,.$$

Using the above asymptotics for the two factorial terms we get

$$\frac{\pi}{2} = \lim_{n\to\infty} \frac{2^{4n}\mathcal{C}^4 n^{4n+2} e^{-4n}(1 + O(\frac{1}{n}))}{(\mathcal{C}(2n)^{2n+\frac{1}{2}}e^{-2n})^2(1 + O(\frac{1}{n}))(2n+1)} = \mathcal{C}^2 \lim_{n\to\infty} \frac{n}{2(2n+1)} = \frac{\mathcal{C}^2}{4}$$

which yields $\mathcal{C} = \sqrt{2\pi}$. □

3.117. EXERCISE. Using the formula (3.28) derive the asymptotics

$$|B_{2\ell}| \sim 4\sqrt{\pi\ell}\big(\frac{\ell}{\pi e}\big)^{2\ell} \qquad \text{as } \ell \to \infty,$$

in the sense that the quotient of the left and right hand sides converges to 1 as $\ell \to \infty$.

3.6. Further exercises

3.118. EXERCISE. Investigate convergence or divergence for each of the following series (here a, b, c are real parameters and convergence may depend on their values).

$$\text{(i)} \sum_{n=10}^{\infty} \frac{1}{n^a (\log n)^b (\log\log\log n)^c} \qquad \text{(ii)} \sum_{n=3}^{\infty} (\log n)^{a\frac{\log n}{\log\log n}} \qquad \text{(iii)} \sum_{n=1}^{\infty} \Big(e^{1/n} - \tfrac{n+1}{n}\Big)$$

$$\text{(iv)} \sum_{k=1}^{\infty} \binom{3k}{k} a^k \qquad \text{(v)} \sum_{k=1}^{\infty} (-1)^k \log(\tfrac{k+1}{k}) \qquad \text{(vi)} \sum_{n=1}^{\infty} \frac{(n!)^2}{2^{n^2}}$$

3.119. EXERCISE. Investigate convergence or divergence for each of the following series (a is a real parameter and convergence may depend on its value):

$$\text{(i)} \sum_{n=1}^{\infty} \cos(\pi n)\sin(\pi n^{-1}) \qquad \text{(ii)} \sum_{n=2}^{\infty} \left(\left(1 + \frac{1}{n}\right)^n - e\right)^2 \qquad \text{(iii)} \sum_{n=1}^{\infty} \frac{1}{n(n^{1/n})^{100}}$$

$$\text{(iv)} \sum_{n=1}^{\infty} \frac{n^n}{n!} a^n \qquad \text{(v)} \sum_{n=2}^{\infty} 2^{-(\log(n))^a} \qquad \text{(vi)} \sum_{n=1}^{\infty} \Big(\sum_{k=0}^{10n} (-1)^k \frac{n^k}{k!}\Big)$$

3.120. EXERCISE. Evaluate each of the following series:

(i) $\sum_{k=1}^{\infty} \frac{1}{k(k+1)}$

(ii) $\sum_{k=1}^{\infty} \frac{1}{k(k+2)}$

(iii) $\sum_{k=1}^{\infty} \frac{1}{k(k+1)(k+2)}$

3.121. EXERCISE. Let $(a_k)_{k=1}^{\infty}$ be a sequence such that $\lim_{k\to\infty} k a_k = 1$. Is $\sum_{n=1}^{\infty}(-1)^n a_n$ necessarily convergent? (Proof or counterexample).

3.122. Exercise.

(i) Determine the radii of convergence of

$$\sum_{k=1}^{\infty}(2^{1/k}-1)z^k \quad \text{and} \quad \sum_{k=1}^{\infty}(k^{1/k}-1)z^k.$$

(ii) For each $z\in\mathbb{C}$ determine whether $\sum_{k=1}^{\infty}(2^{1/k}-1)z^k$ converges.

(iii) For each $z\in\mathbb{C}$ determine whether $\sum_{k=1}^{\infty}(k^{1/k}-1)z^k$ converges.

3.123. Exercise. Let $(a_n)_{n=1}^{\infty}$ be a sequence of decimal digits, i.e. $a_n\in\{0,1,\dots,9\}$ for all $n\geqslant 1$.

(i) Prove that $\sum_{n=1}^{\infty}a_n10^{-n}$ converges to a value in $[0,1]$.

(ii) Say that (a_n) has a 9-tail if there exists an $N\geqslant 1$ such that $a_n=9$ for all $n\geqslant N$. Show that for every $x\in[0,1)$ there exists a unique sequence of decimal digits (a_n) with no 9-tail such that

$$x=\sum_{n=1}^{\infty}a_n10^{-n}.$$

(iii) Let $b\in\mathbb{N}$, $b\geqslant 2$. Generalize the above to expansions in base b.

3.124. Exercise. For $n\geqslant 0$ let $a_n=(\log(2+n))^{-1}$.

(i) For which complex numbers z does the series $\sum_{n=0}^{\infty}a_n\,z^n$ converge?

(ii) For which complex numbers z does the series $\sum_{n=0}^{\infty}a_n\,z^n$ converge absolutely?

(iii) On which closed and bounded sets of the complex plane does the series $\sum_{n=0}^{\infty}a_n\,z^n$ converge uniformly?

3.125. Exercise. For which $x\in\mathbb{R}$ do the following series converge? On which sets do these series converge uniformly?

$$\text{(i)}\ \sum_{n=1}^{\infty}n^2x^n \qquad \text{(ii)}\ \sum_{n=1}^{\infty}(3^{1/n}-1)^nx^n \qquad \text{(iii)}\ \sum_{n=1}^{\infty}\tan(n^{-2})e^{nx}$$

$$\text{(iv)}\ \sum_{n=1}^{\infty}\frac{x^n}{n^n} \quad \text{(v)}\ \sum_{n=1}^{\infty}\frac{\sin(nx)}{n^2} \quad \text{(vi)}\ \sum_{n=1}^{\infty}2^{-n}\tan(\lfloor x\rfloor+1/n)$$

3.126. Exercise. This exercise is relevant in assessing the strength of the root and ratio tests for series. Assume that $b_n>0$ for all $n\in\mathbb{N}_0$.

(i) Prove that

$$\liminf_{n\to\infty}\frac{b_{n+1}}{b_n}\leqslant\liminf_{n\to\infty}b_n^{1/n}\leqslant\limsup_{n\to\infty}b_n^{1/n}\leqslant\limsup_{n\to\infty}\frac{b_{n+1}}{b_n}.$$

(ii) Find a sequence with positive terms where the four expressions are all different.

3.127. Exercise. Let $(a_n)_{n\geqslant 2}$ be a sequence of positive numbers.

(i) Show that if $\liminf_{k\to\infty}\frac{\log(1/a_k)}{\log k}>1$, then $\sum_n a_n$ converges.

(ii) Show that if $\frac{\log(1/a_k)}{\log k}\leqslant 1$ for all sufficiently large k, then $\sum_n a_n$ diverges.

3.128. Exercise. Assume $a_n>0$ for all $n\in\mathbb{N}$ and $\lim_{n\to\infty}a_n=\infty$. What can you say about the convergence of the series

$$\sum_{n=1}^{\infty}a_n^{-n},\ \sum_{n=1}^{\infty}a_n^{-\sqrt{n}},\ \sum_{n=10}^{\infty}a_n^{-\log n},\ \sum_{n=100}^{\infty}a_n^{-\log\log n},\ \sum_{n=100}^{\infty}a_n^{-\log\log\log n}\ ?$$

In each case give a proof or counterexample.

3.129. Exercise. Let (a_n) be a decreasing sequence with $\lim_{n\to\infty}a_n=0$. Prove that there exists a decreasing sequence b_n of positive numbers such that $\lim_{n\to\infty}b_n=0$, $\sum_{n=0}^{\infty}b_n=\infty$, and $\sum_{n=0}^{\infty}a_nb_n$ converges.

3.130. Exercise. Let $(a_n)_{n\geqslant 0}$ be a sequence of positive numbers. Let (b_n) be any sequence of positive numbers and define

$$\kappa_n = b_n - b_{n+1}\frac{a_{n+1}}{a_n}.$$

Prove *Kummer's test:*
(i) If $\liminf_{n\to\infty}\kappa_n > 0$, then $\sum_{n=0}^{\infty} a_n$ converges.
(ii) If $\kappa_n \leqslant 0$ for all sufficiently large n and if $\sum_{n=0}^{\infty}\frac{1}{b_n}$ diverges, then $\sum_{n=0}^{\infty} a_n$ diverges.

3.131. Exercise. Show how to derive Raabe's test (Theorem 3.18) from Kummer's test in Exercise 3.130.

3.132. Exercise. Derive *Gauss' test* : Let $a_n > 0$ for all $n \geqslant 0$, and let $b \in \mathbb{R}$. Suppose that

$$\lim_{k\to\infty} k(\log k)\Big(\frac{a_{k+1}}{a_k} - 1 - \frac{b}{k}\Big) = 0$$

Then $\sum_{n=0}^{\infty} a_n$ converges if and only if $b < -1$.

3.133. Exercise. Show that

$$\sum_{n=1}^{\infty}(-1)^n\binom{\alpha}{n}$$

converges for $\alpha > 0$ and diverges for $\alpha < 0$.

3.134. Exercise. Let $(a_n)_{n=1}^{\infty}$ be a sequence of numbers with $a_n > -1$. Prove or disprove the following statement: If $\sum_{k=1}^{\infty}\log(1+a_k)$ converges then $\sum_{k=1}^{\infty} a_k$ converges.

3.135. Exercise. Assume that $a_k > 0$ and $\sum_{k=0}^{\infty} a_k = \infty$. Assume that (b_k) is a bounded sequence. Show that one can choose a strictly increasing sequence of natural numbers $\kappa(n)$ such that

$$\sum_{n=0}^{\infty} a_{\kappa(n)} = \infty$$

and the sequence $\big(b_{\kappa(n)}\big)$ has a limit.

3.136. Exercise. Consider on $[0,1]$ the functions $f_n(x) = \frac{n^a x}{(1+n^2x^2)^2}$.
(i) For which choices of a and x is $\sum_{n=1}^{\infty} f_n(x)$ convergent?
(ii) For which choices of a is $\sum_{n=1}^{\infty} f_n(x)$ uniformly convergent on $[0,1]$?
(iii) For which choices of a is $\sum_{n=1}^{\infty} f_n(x)$ bounded on $[0,1]$?
(iv) For which choices of a and x is $\sum_{n=1}^{\infty} f_n(x)$ continuous at x?

3.137. Exercise. Let

$$s_N(x) = \sum_{n=1}^{N}(-1)^n\frac{x^{3n}}{n^{2/3}}.$$

Prove that there is a constant C so that for all $N \geqslant 1$,

$$\sup_{x\in[0,1]}|s_N(x) - s_\infty(x)| \leqslant CN^{-2/3}.$$

(In particular, s_N converges uniformly.)

3.138. Exercise. Consider $\sum_{k=1}^{\infty}(-1)^k(x^2+k)^{-1}$ and $\sum_{k=1}^{\infty}(-1)^k k^{-2}(x^2+k)$.
(i) For which x do these series converge?
(ii) For which x do these series converge absolutely?
(iii) On which intervals do these series converge uniformly?

3.139. Exercise. Does the series

$$\sum_{n=1}^{\infty} e^{-x/n}\frac{(-1)^n}{n}$$

converge uniformly on $[0,\infty)$? Prove or disprove.

3.140. Exercise. Let λ_1, λ_2,..., λ_n, ... be real numbers. Argue that

$$f(x) = \sum_{n=1}^{\infty} \frac{e^{i\lambda_n x}}{n^2}$$

defines a continuous bounded function on $\mathbb{R}$ and then show that the limit

$$\lim_{T\to\infty} \frac{1}{T} \int_{-T}^{T} f(x)dx$$

exists.

3.141* Exercise. Consider the series

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

Does it converge or not?

3.142. Exercise. Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series and assume that $R > 0$ is its radius of convergence. For $|z| < R$ let $f(z) = \sum_{n=0}^{\infty} a_n z^n$. Prove:

(i) f satisfies $f(x) = f(-x)$ for all $x \in (-R, R)$ if and only if $a_{2k+1} = 0$ for all $k \geqslant 0$.
(ii) f satisfies $f(z) = f(-z)$ for all $|z| < R$ if and only if $a_{2k+1} = 0$ for all $k \geqslant 0$.
(iii) f satisfies $f(x) = -f(-x)$ for all $x \in (-R, R)$ if and only if $a_{2k} = 0$ for all $k \geqslant 0$.
(iv) f satisfies $f(z) = -f(-z)$ for all $|z| < R$ if and only if $a_{2k} = 0$ for all $k \geqslant 0$.

3.143. Exercise. Compute all the derivatives $(\frac{d}{dx})^k \arctan x$ when evaluated at 0.

3.144. Exercise. For $|x| < 1$ compute

$$\text{(i)} \sum_{k=1}^{\infty} k^2 x^{2k+1}, \qquad \text{(ii)} \sum_{k=1}^{\infty} k^{-1} x^{3k+1}, \qquad \text{(iii)} \sum_{k=1}^{n} k^2 (x^k - x^{k-1}) \ .$$

3.145. Exercise. Suppose $f(z) = \sum_{k=0}^{\infty} a_k z^k$ has radius of convergence $R > 0$. For $|z| < R$ find formulas for

$$F_n(z) = \sum_{k=0}^{\infty} k^n a_k z^k, \quad n = 1, 2, \ldots$$

3.146. Exercise. Determine the value of the series

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

3.147. Exercise. For every positive integer k, determine the value of the series

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

3.148. Exercise. For a positive real number x define

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{n(n+1)+x}.$$

(i) Show that $f : (0,\infty) \to (0,\infty)$ is a well-defined and continuous function.
(ii) Prove that there exists a unique $x_0 \in (0,\infty)$ such that $f(x_0) = 2\pi$.
(iii) Determine the value of x_0.

3.149. Exercise. Consider the function

$$\zeta(s) = \sum_{k=1}^{\infty} k^{-s}.$$

Show that the series converges uniformly for $s \in [1+\varepsilon, \infty)$ if $\varepsilon > 0$. Prove that ζ is differentiable for $s > 1$ and show that

$$\zeta'(s) = \sum_{k=1}^{\infty} \frac{\log k}{k^s}$$

for $s > 1$. Show that ζ is a C^∞ function on $(1, \infty)$ and obtain formulas for the higher derivatives.

3.150. Exercise. Show that for all $a \notin 0$ and $|x| < 1$,

$$(1+x)^a = \sum_{k=0}^{\infty} \binom{a}{k} x^k.$$

3.151. Exercise. Use the definition of arcsin as in §2.5 to derive the Taylor series

$$\arcsin(x) = \sum_{k=0}^{\infty} \frac{1}{2^{2k}(2k+1)} \binom{2k}{k} x^{2k+1}.$$

Show that equality holds for all $x \in [-1, 1]$.

3.152. Exercise. Let $(a_k)_{k\in\mathbb{N}}$ be a sequence of complex numbers with $\lim_{k\to\infty} a_k = L$. Let $q > -1$. Prove that the limit

$$\lim_{N\to\infty} \frac{1}{N^{1+q}} \sum_{k=1}^{N} a_k k^q$$

exists and determine its value.

3.153. Exercise (Cesàro sums). Given a sequence a_k, form the partial sums $s_n = \sum_{k=0}^{n} a_k$ and let

$$\sigma_N = \frac{s_0 + \cdots + s_{N-1}}{N}.$$

σ_N is called the Nth *Cesàro mean* of the sequence s_k or the Nth *Cesàro sum* of the series $\sum_{k=0}^{\infty} a_k$. If σ_N converges to a limit S we say that the series $\sum_{k=0}^{\infty} a_k$ is *Cesàro summable to* S.

(i) Show that if $\sum_{k=1}^{\infty} a_k$ is convergent with sum S, then $\sum_{k=0}^{\infty} a_k$ is Cesàro summable to S.
(ii) Prove that the sum $\sum_{k=0}^{\infty} (-1)^k$ does not converge but is Cesàro summable to some limit S and determine S. Show that it is Abel summable to S.
(iii) Consider the series $1 - 2 + 3 - 4 + \cdots$ which is Abel summable by Exercise 3.73. Show that this series is not Cesàro summable to any number.

3.154. Exercise (Cesàro vs. Abel sums). Abel's theorem 3.71 tells us that if $\sum_{k=1}^{\infty} a_k$ is summable to S then the series is also Abel summable to S. This exercise strengthens this implication to

$$\sum_{k=0}^{\infty} a_k \text{ summable} \implies \sum_{k=0}^{\infty} a_k \text{ Cesàro summable} \implies \sum_{k=0}^{\infty} a_k \text{ Abel summable}, \quad (*)$$

where the first implication is the result of part (ii) of Exercise 3.153. We are concerned with the second implication:

(i) Prove first that if $\sum_k a_k$ is Cesàro summable, then for $r < 1$,

$$\sum_{k=0}^{\infty} a_k r^k = (1-r)^2 \sum_{n=1}^{\infty} n\sigma_n r^n.$$

Here $\sigma_n = \frac{1}{n}(s_0 + \cdots + s_{n-1})$ with $s_k = \sum_{n=0}^{k} a_n$.

(ii) Split the right hand side of the previous formula as

$$(1-r)^2\sum_{n=1}^{\infty} n(\sigma_n - S)r^n + (1-r)^2\sum_{n=1}^{\infty} nSr^n$$

and finish the proof of the second implication in $(*)$.

3.155. Exercise (Tauberian theorem). Tauber's theorem says that under an additional condition on the a_k some of the implications in Exercise 3.154 can be reversed. Prove:

(i) If $\sum_{k=1}^{\infty} a_k$ is Cesàro summable to S and if $\lim_{k\to\infty} ka_k = 0$, then $\sum_{k=1}^{\infty} a_k = S$.

Remark. Note that part (i) is immediate from (ii) and Exercise 3.154. But it is still instructive to prove part (i) directly.

(ii) If $\sum_{k=1}^{\infty} a_k$ is Abel summable to S and if $\lim_{k\to\infty} ka_k = 0$, then $\sum_{k=1}^{\infty} a_k = S$.

3.156. Exercise (Hardy). Suppose $\sum_{k=1}^{\infty} a_k$ is Cesáro summable and $\sup_k k|a_k| < \infty$. Then $\sum_{k=0}^{\infty} a_k$ converges.

Remark. Littlewood showed the following: If $\sum_{k=0}^{\infty} a_k$ is Abel summable to S and if ka_k is merely a bounded sequence then $\sum_{k=1}^{\infty} a_k = S$. In view of Exercise 3.154 this is stronger than Hardy's result (and the proof is harder, see Zygmund [**85**, (1-41)]). Littlewood also showed the optimality of the condition $\sup_k |ka_k| < \infty$.

3.157. Exercise. Show that the series $f(z) = \sum_{m=0}^{\infty} z^{2^m}$ converges for $|z| < 1$. Let $\Sigma \subset S^1$ be the set of points w in the unit circle S^1 at which $\lim_{r\to 1-} f(rw)$ does not exist. Show that every w satisfying $w^{2^\ell} = 1$ for some nonnegative integer ℓ belongs to Σ (i.e. for those w the series $\sum_{n=0}^{\infty} w^{2^n}$ is not Abel summable). Show that Σ is dense in the unit circle (i.e. every ball centered at a point in S^1 contains points in Σ).

3.158. Exercise. Suppose $f(x) = \sum_{n=0}^{\infty} a_n x^n$ has radius of convergence $R = 1$ and assume that $\sum_{n=0}^{\infty} a_n = \infty$. Prove that $\lim_{x\to 1-} f(x) = \infty$.

3.159. Exercise. Suppose f is a nonnegative C^∞ function on $[0,\rho]$, and all its derivatives are nonnegative on $[0,\rho]$. Prove the Taylor series

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

converges to f on $[0,\rho)$.

3.160. Exercise. The sequence of *Fibonacci numbers* is defined by

$$a_0 = a_1 = 1, \qquad a_n = a_{n-1} + a_{n-2} \text{ for } n \geqslant 2.$$

(i) Show that for sufficiently small z the power series $\sum_{n=0}^{\infty} a_n z^n$ converges and that

$$(1 - z - z^2)\sum_{n=0}^{\infty} a_n z^n = 1.$$

(ii) Use a partial fraction decomposition $\frac{1}{1-z-z^2} = \frac{A_1}{z-z_1} + \frac{A_2}{z-z_2}$ to derive the explicit formula

$$a_n = \frac{1}{\sqrt{5}}(\varphi^{n+1} + (-1)^n\varphi^{-n-1}),$$

where $\varphi = \frac{\sqrt{5}+1}{2}$.

(iii) Show that $\sum_{n=0}^{\infty} a_n z^n$ has radius of convergence $R = \varphi^{-1}$. For each z with $|z| = \varphi^{-1}$ determine whether $\sum_{n=0}^{\infty} a_n z^n$ is convergent, and whether it is Abel summable.

(iv) For

$$f(x) = (1 - x - x^2)^{-1} = \sum_{n=0}^{\infty} a_n x^n$$

determine the value of all derivatives when evaluated at 0.

3.161. Exercise. Consider the series

$$\sum_{k=0}^{\infty} (-1)^k \frac{x^{k+1}}{k+1}.$$

(i) Use the argument in Example 3.42 to prove that for every $\beta \in (-1, 1)$ the series converges uniformly in $[\beta, 1]$.

(ii) Prove that the series does not converge uniformly on $(-1, 1]$.

3.162. Exercise. Let a be a decreasing C^1-function in $[0, \infty)$ such that $\lim_{t\to\infty} a(t) = 0$.

(i) Show that $\lim_{N\to\infty} \int_0^N a(t) \sin(tx) dt$ exists for all $x > 0$.

(ii) For $\epsilon > 0$ show that $\lim_{N\to\infty} \int_0^N a(t) \sin(tx) dt$ converges uniformly for $x \in [\epsilon, \infty)$.

(iii) Show that uniform convergence may fail in $(0, \infty)$, for a suitable choice of a.

3.163. Exercise. Let

$$s_n(x) = \sum_{k=1}^{n} \sin(kx).$$

Show that there exists a constant C, independent of N, x, such that

$$\sum_{n=1}^{N} \frac{|s_n(x)|}{n^2} < C, \quad 0 < x < \pi, \quad N = 1, 2, 3, \ldots.$$

3.164. Exercise. Suppose $\sum_{k=0}^{\infty} a_k$ converges.

(i) Let $R > 1$, and

$$b_n(R) = R^{-n} \sum_{k=0}^{n} R^k a_k.$$

Prove that $\sum_{n=0}^{\infty} b_n(R)$ converges and express this sum in terms of the value of $\sum_{k=0}^{\infty} a_k$.

(ii) Let $c_n = \frac{1}{n^2+n} \sum_{k=1}^{n} k a_k$. Show that $\sum_{n=1}^{\infty} c_n$ converges and $\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} a_n$.

3.165. Exercise. Verify the relation $\frac{z}{e^z - 1} + \frac{z}{2} = \frac{z}{2} \coth(\frac{z}{2})$ where $\coth z = \frac{\cosh z}{\sinh z}$. Show that for sufficiently small z,

(i) $\frac{z}{2} \coth(\frac{z}{2}) = \sum_{\ell=0}^{\infty} \frac{B_{2\ell} z^{2\ell}}{(2\ell)!}$.

(ii) $z \cot z = \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{B_{2\ell} 2^{2\ell} z^{2\ell}}{(2\ell)!}$.

(iii) Given Exercise 3.117, what are the radii of convergence in (i) and (ii)?

3.166. Exercise. Examine convergence or divergence of the series

$$\text{(i)} \sum_{n=0}^{\infty} (-1)^n \frac{n^n}{e^n n!}, \qquad \text{(ii)} \sum_{n=0}^{\infty} (-1)^n \frac{e^n n!}{n^n},$$

and, more generally for $|z| = 1$,

$$\text{(iii)} \sum_{n=0}^{\infty} \frac{n^n}{e^n n!} z^n, \qquad \text{(iv)} \sum_{n=0}^{\infty} \frac{e^n n!}{n^n} z^n.$$

3.167. Exercise. Let $\alpha, \beta, \gamma \in \mathbb{R}\backslash\{0, -1, -2, \dots\}$. Let $\mathfrak{p}_\alpha(n) = \prod_{k=0}^{n-1}(\alpha + k)$ and

$$\mathfrak{h}_{\alpha,\beta,\gamma}(n) = \frac{1}{n!}\frac{\mathfrak{p}_\alpha(n)\mathfrak{p}_\beta(n)}{\mathfrak{p}_\gamma(n)}$$

(i) Use Raabe's test to show that the *hypergeometric series*

$$\sum_{n=1}^{\infty} \mathfrak{h}_{\alpha,\beta,\gamma}(n)$$

converges if $\alpha + \beta < \gamma$ and diverges if $\alpha + \beta > \gamma$.
(ii) Use Gauss' test in Exercise 3.132 to show that it diverges for $\alpha + \beta = \gamma$.

3.168. Exercise. Let $\alpha, \beta, \gamma, \mathfrak{h}_{\alpha,\beta,\gamma}$ be as in Exercise 3.167.

(i) Show using the asymptotics for factorials and shifted factorials in Theorem 3.115 that

$$\mathfrak{h}_{\alpha,\beta,\gamma}(n) = c(\alpha, \beta, \gamma)n^{\alpha+\beta-\gamma-1}(1 + O(\tfrac{1}{n}))$$

as $n \to \infty$, where $c(\alpha, \beta, \gamma)$ is a non-zero constant.

(ii) The hypergeometric power series with parameters (α, β, γ) is defined by

$$\sum_{n=0}^{\infty} \mathfrak{h}_{\alpha,\beta,\gamma}(n)z^n.$$

For all $z \in \mathbb{C}$ and admissible parameters α, β, γ determine whether the series converges.

In case of convergence the hypergeometric power series with parameters α, β, γ is also called *hypergeometric function* and denoted by ${}_2F_1(\alpha, \beta; \gamma; z)$. This is a special case of a more general hypergeometric series:

$${}_pF_q(\alpha_1, \dots, \alpha_p; \gamma_1, \dots, \gamma_q; z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}\frac{\prod_{i=1}^{p}\mathfrak{p}_{\alpha_i}(n)}{\prod_{j=1}^{q}\mathfrak{p}_{\gamma_j}(n)}.$$

(iii) Investigate what can be said about convergence of the series defining ${}_pF_q$.

Remark. For more on hypergeometric functions and their generalizations, see e.g. Andrews–Askey–Roy [**2**].

3.169. Exercise. For $x \in \mathbb{R}$ and $y > 0$ let

$$f(x, y) = \sum_{n=1}^{\infty} \frac{x}{x^2 + yn^2}.$$

(i) Show that for each $y > 0$, the limit

$$g(y) = \lim_{x\to+\infty} f(x, y)$$

exists. Evaluate the limit function g.

(ii) Determine if $f(x, y)$ converges to $g(y)$ uniformly in $y \in (0, \infty)$ as $x \to \infty$.

3.170. Exercise. Suppose that for $n \in \mathbb{N}_0$ the complex-valued functions f_n and the real-valued functions g_n are defined on a set I and suppose that

(i) the partial sums $\sum_{k=0}^{n} f_k(x)$ are uniformly bounded on I,
(ii) $g_n(x) \to 0$ uniformly on I,
(iii) $g_k(x) \geqslant g_{k+1}(x)$ for all $k \in \mathbb{N}_0$ and all $x \in I$.

Prove that $\sum_{n=0}^{\infty} f_n(x)g_n(x)$ converges uniformly on I.

3.171. Exercise. Let $p > 0$. Discuss convergence and uniform convergence (on suitable intervals) of the series $\sum_{k=1}^{\infty} k^{-p}\sin(kx)$ and $\sum_{k=1}^{\infty} k^{-p}\cos(kx)$.

3.172. Exercise. Suppose $(a_n)_{n=0}^{\infty}$ are complex numbers and $(b_n)_{n=0}^{\infty}$ real numbers such that

(i) the series $\sum_{n=0}^{\infty} a_n$ converges,

(ii) $\sup_{k\geqslant 0} |b_k| < \infty$,

(iii) (b_k) is a monotone sequence.

Prove that $\sum_{n=0}^{\infty} a_n b_n$ converges.

3.173. Exercise. Suppose that for $n \in \mathbb{N}_0$ the complex-valued functions f_n and the real-valued functions g_n are defined on a set I and suppose that

(i) the series $\sum_{k=0}^{\infty} f_k(x)$ converges uniformly on I,

(ii) g_k converges uniformly on I to a bounded function,

(iii) $g_k(x) \geqslant g_{k+1}(x)$ for all $k \in \mathbb{N}_0$ and all $x \in I$.

Prove that $\sum_{n=0}^{\infty} f_n(x)g_n(x)$ converges uniformly on I.

CHAPTER 4

Metric spaces

The notion of a metric space serves as a convenient abstract setting that underlies all topics discussed in this text. A metric space can be thought of as a collection of distinct objects that come with a *distance* between them. This provides a structure that makes it meaningful to speak of notions such as *convergence* and *continuity*. It will allow us to use the same terminology for potentially very different kinds of objects.

4.1. Definition (Metric space). A set X equipped with a map $d : X \times X \to [0, \infty)$ is called a *metric space* if X is not the empty set and for all $x, y, z \in X$,

(i) $d(x, y) = d(y, x)$,
(ii) $d(x, z) \leqslant d(x, y) + d(y, z)$,
(iii) $d(x, y) = 0$ if and only if $x = y$.

d is called a *metric*. Property (ii) is called the *triangle inequality*,

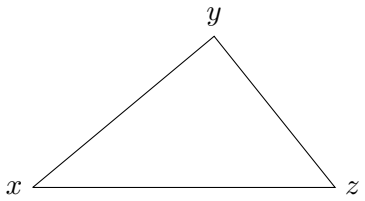


Figure 4.1. The length of any side of a triangle is no greater than the sum of the lengths of two of its sides.

If multiple metric spaces are relevant at the same time, then we may also write d_X for the metric d on the metric space X. One also refers to the pair (X, d) (consisting of the set and the associated metric) as a metric space.

4.2. Examples. Some fundamental examples of metric spaces that will be important in this book are:

- The real numbers $\mathbb{R}$ with $d(x, y) = |x - y|$.
- Closed and open intervals of real numbers (with the same metric),
- The complex numbers $\mathbb{C}$ with $d(z, w) = |z - w|$,
- The n-dimensional Euclidean space $\mathbb{R}^n$ consisting of vectors $x = (x_1, \ldots, x_n)$ with the *Euclidean metric*
$$d(x, y) = \Big(\sum_{i=1}^{n} |x_i - y_i|^2 \Big)^{1/2}.$$
- The space $C([a, b])$ of continuous functions $[a, b] \to \mathbb{C}$ with
$$d(f, g) = \sup_{x \in [a,b]} |f(x) - g(x)| \quad (\textit{supremum metric}).$$
- The space ℓ^∞ of bounded sequences $(a_n)_{n \in \mathbb{N}}$ of complex numbers with
$$d(a, b) = \sup_{n \in \mathbb{N}} |a_n - b_n| .$$

- The space c_0 of sequences $(a_n)_{n\in\mathbb{N}}$ of complex numbers with $\lim_{n\to\infty} a_n = 0$, with the same metric as for ℓ^∞.

4.3. Exercise. Verify that each of the preceding examples is really a metric space.

4.4. Exercise (Product metric). Let (X_i, d_i) be metric spaces, $i = 1, \dots, n$. On $X_1 \times \cdots \times X_n$ we define a *product metric* d, setting for $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$

$$d(x, y) = \max_{i=1,\dots,n} d_i(x_i, y_i).$$

Verify the metric space properties.

4.1. Topology

4.1.1. Open and closed sets. In the following, let X be a metric space with metric d. For every $x_0 \in X$ and $r > 0$ define the *open ball*

$$B(x_0, r) = \{x \in X \,:\, d(x, x_0) < r\},$$

and the *closed ball*

$$\overline{B}(x_0, r) = \{x \in X \,:\, d(x, x_0) \leqslant r\}.$$

For a set $E \subset X$ we define the *diameter* of E to be

$$\operatorname{diam}(E) = \sup\{d(x, x') : x, x' \in E\}.$$

A ball of radius r is of diameter $\leqslant 2r$.

4.5. Example. If $X = \mathbb{R}$ (always with the usual metric), then the open balls are open intervals and the closed balls are closed intervals.

Should multiple metric spaces be involved, we use subscripts on the metric and balls to indicate which metric space we mean, *i.e.* $B_X(x_0, r)$ is a ball in the metric space X.

4.6. Definition (Open and closed sets). Let X be a metric space and $U \subset X$.

(i) A point $x \in U$ is called *interior* in U if there exists $r > 0$ such that $B(x, r) \subset U$.
(ii) A set $U \subset X$ is called *open* if every $x \in U$ is an interior point.
(iii) A set $A \subset X$ is called *closed* if its complement $A^\complement = X \backslash A$ is open.

Clarification of notation: $A \subset B$ means for us that A is a subset of B, not necessarily a proper subset. That is, we also allow $A = B$. We will write $A \subsetneq B$ to refer to proper subsets.

4.7. Observations. *Let (X, d) be a metric space.*

(i) Unions of open sets in X are open.
(ii) Intersections of closed sets in X are closed.
(iii) The sets X and $\varnothing$ are both open and closed.

4.8. Definition. The family $(U)_{U\subset X \text{ open}}$ of open sets is also called the *topology of X*.

Note from the definition that the topology of a metric space X is determined by the open balls $(B(x, r))_{x\in X, r>0}$. The notion of open sets can be generalized and leads to the concept of *topological spaces*, which we will not need in this text.

4.9. Definition (Closure). If $A \subset X$ is an arbitrary set, then the *closure of A*, denoted by $\overline{A}$, is the intersection of all closed sets containing A.

Note that X is a closed set containing A, and therefore $A \subset \overline{A}$. Since the intersection of closed sets is closed, the closure $\overline{A}$ is indeed closed. It is the 'smallest' closed set containing A in the sense that if A' is a closed set with $A \subset A'$, then $\overline{A} \subset A'$. Therefore we have

4.10. Observation. *A set A is closed if and only if $A = \overline{A}$.*

4.11. Exercise. Verify that open balls are open and closed balls are closed.

Note that $\overline{B(x_0,r)}$, the closure of the open ball $B(x_0,r)$, often coincides with the closed ball $\overline{B}(x_0,r)$. While this is the case in most of the metric spaces encountered in this course, it is generally only true that

$$B(x_0,r) \subset \overline{B(x_0,r)} \subset \overline{B}(x_0,r).$$

4.12. Example. Let X be a non-empty set. For $x,y \in X$ let

$$d(x,y) = \begin{cases} 0, & \text{if } x = y, \\ 1, & \text{if } x \neq y. \end{cases}$$

This defines a metric on X (called the *trivial metric*). The topology on X is not very interesting: every set is open (hence also every set is closed). If X consists of at least two points then for every $x \in X$,

$$B(x,1) = \overline{B(x,1)} = \{x\} \subsetneq \overline{B}(x,1) = X.$$

4.13. Definition (Accumulation point). Let $A \subset X$. A point $x \in X$ is called an *accumulation point* of A if for every $r > 0$, there exists $y \in B(x,r) \cap A$ with $y \neq x$.

4.14. Lemma. *Let X be a metric space and $A \subset X$. Then $\overline{A}$ is equal to the union of A and the set of accumulation points of A.*

Proof. Let $\mathrm{Acc}(A)$ be the set of accumulation points of A. We know from the definition of closure that $A \subset \overline{A}$ and show that $\mathrm{Acc}(A) \subset \overline{A}$. To see this it suffices to show for an arbitrary closed set C containing A that every accumulation point x belongs to C. If x were in $X \backslash C$ (an open set not intersecting A), then there would be an $\varepsilon > 0$ and a ball $B(x,\varepsilon)$ such that $B(x,\varepsilon) \subset X \backslash C$ and hence $B(x,\varepsilon) \cap A \subset B(x,\varepsilon) \cap C = \varnothing$, in contradiction to x being an accumulation point. Hence we have found that $A \cup \mathrm{Acc}(A) \subset \overline{A}$.

To show the reverse inclusion $\overline{A} \subset A \cup \mathrm{Acc}(A)$ let $x \in \overline{A} \backslash A$; we have to show that x is an accumulation point. Again argue by contradiction and suppose x is not an accumulation point. Then there exists a ball $B(x,\varepsilon)$ containing no points in A other than x itself. Hence $C = (X \backslash B(x,\varepsilon)) \cap \overline{A}$ is a closed set containing A, with $C \subsetneq \overline{A}$, in contradiction to the definition of closure of A. □

4.1.2. Relative topology. Let X be a metric space, with metric d, and let $A \subset X$ be a non-empty subset. Then A can be made a metric space by restricting the metric: we define a metric $d_A : A \times A \to [0,\infty)$ by setting

$$d_A(x,y) = d(x,y) \quad \text{for all } x,y \in A.$$

In other words, d_A is the restriction of d to the set $A \times A \subset X \times X$, also denoted by $d_A = d|_{A\times A}$. As a metric space, A comes with its own open sets: unpacking the definition, a set $U \subset A$ is open in A if and only if for every $x \in U$ there exists $r > 0$ such that

$$B_A(x,r) = \{y \in A : d(x,y) < r\} \subset U.$$

Observe that the open balls in A are not necessarily open balls in X. As a consequence, a set $U \subset A$ that is open in A is not necessarily open in X. However, the open sets in A can be characterized by the open sets in X.

4.15. Lemma. *Let $A \subset X$. A set $U \subset A$ is open in A if and only if there exists an open set $V \subset X$ such that $U = V \cap A$.*

Proof. Suppose that $U = V \cap A$ with V open. We have to show that U is open in A. Let $x \in U \subset V$ then there is a ball $B(x,r_x) = \{y \in X : d(x,y) < r_x\}$ contained in V. Then $B_A(x,r_x) \subset U$, so x is an interior point of U (with respect to the metric on A).

Vice versa, let U be open in A. Then for every $x \in U$ there is $r_x > 0$ such that $x \in B_A(x,r_x) \subset U$, and thus $U = \bigcup_{x\in U} B_A(x,r_x)$. Define $V = \bigcup_{x\in U} B(x,r_x)$. Then V is open in X and $V \cap A = U$. □

4.16. Example. Let $X = \mathbb{R}$, $A = [0,1]$. Then $U = [0, \frac{1}{2}) \subset A \subset X$ is open in A, but not open in $\mathbb{R}$. However, there exists $V \subset \mathbb{R}$ open such that $U = V \cap A$: for example, $V = (-1, \frac{1}{2})$.

4.1.3. Convergence.

4.17. Definition (Convergence). Let X be a metric space, $(x_n)_{n\in\mathbb{N}} \subset X$ a sequence and $x \in X$. We say that $(x_n)_{n\in\mathbb{N}}$ *converges* to x if for every $\varepsilon > 0$ there exists $N = N(\varepsilon) \in \mathbb{N}$ such that for all $n \geqslant N$ it holds that $d(x_n, x) < \varepsilon$.

If $(x_n)_{n\in\mathbb{N}}$ converges to x we also call x the *limit* of the sequence and write

$$x = \lim_{n\to\infty} x_n \quad \text{or} \quad x_n \to x \text{ in } X.$$

4.18. Exercise. Show that a convergent sequence in a metric space only has one limit.

4.19. Definition (Cauchy sequence). Let X be a metric space. A sequence $(x_n)_{n\in\mathbb{N}}$ in X is called a *Cauchy sequence* if for every $\varepsilon > 0$ there exists $N = N(\varepsilon) \in \mathbb{N}$ such that for all $n, m \geqslant N$,

$$d(x_n, x_m) < \varepsilon.$$

4.20. Lemma. *Every convergent sequence is a Cauchy sequence.*

Proof. Let $\varepsilon > 0$. Since $x_n \to x$ there is N so that for all $k \geqslant N$ we have $d(x_k, x) < \varepsilon/2$. If $m \geqslant N$, $n \geqslant N$ we get by the triangle inequality

$$d(x_n, x_m) \leqslant d(x_m, x) + d(x, x_n) < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

and since ε was arbitrary the result is proved. □

4.21. Definition (Completeness). A metric space X is called *complete* if every Cauchy sequence $(x_n)_{n\in\mathbb{N}} \subset X$ converges.

4.22. Example. The metric space of rational numbers, $\mathbb{Q}$ (with the usual metric) is not complete: the sequence of rational numbers

$$(10^{-n}\lfloor 10^n\sqrt{2}\rfloor)_{n\in\mathbb{N}} = (1.4, 1.41, 1.414, \dots)$$

is a Cauchy sequence, but it does not converge in $\mathbb{Q}$. This is because it converges as a sequence of real numbers to the irrational number $\sqrt{2} \notin \mathbb{Q}$. Here we have used the notation $\lfloor x \rfloor$ for the largest integer $\leqslant x$.

The real numbers form an example of a complete metric space (see §1.3).

4.23. Lemma. *If X is complete and $A \subset X$ is closed, then A is a complete metric space.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in A. Since $d_A = d_X|_{A\times A}$, (x_n) is a Cauchy sequence in X, it has a limit $x \in X$. By Lemma 4.14 $x \in \overline{A}$, but by assumption $\overline{A} = A$. Hence x_n converges to x in A. □

Note that we cannot drop the assumption that X is complete; for example, every metric space is a closed subset of itself, but not every metric space is complete.

The following lemma is often useful in proving completeness of a metric space.

4.24. Lemma. *Let (x_n) be a Cauchy sequence in a metric space and assume that (x_n) has a convergent subsequence with limit x. Then $\lim_{n\to\infty} x_n = x$.*

Proof. By assumption there is a strictly increasing sequence of integers $(\kappa(n))$ such that $\lim_{n\to\infty} x_{\kappa(n)} = x$. Let $\varepsilon > 0$. Then there is $N_1 \in \mathbb{N}$ such that $d(x_{\kappa(n)}, x) < \varepsilon/2$ for $n \geqslant N_1$.

Since (x_n) is Cauchy we find N_2 such that $d(x_\ell, x_m) < \varepsilon/2$ for $m, \ell \geqslant N_2$. Since $\kappa(n) \geqslant n$ this implies in particular that $d(x_n, x_{\kappa(n)}) < \varepsilon/2$ for $n \geqslant N_2$.

Now if $n \geqslant \max\{N_1, N_2\}$ we have

$$d(x_n, x) \leqslant d(x_n, x_{\kappa(n)}) + d(x_{\kappa(n)}, x) \leqslant \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary we obtain $\lim_{n\to\infty} x_n = x$. □

4.25. Lemma. *Let X be a metric space and (x_n) be a sequence in X. Then x_n converges to x if and only if for every subsequence (y_n) of (x_n) there is a subsequence (z_n) of (y_n) such that (z_n) converges to x.*

Proof. If (x_n) converges to x, then all subsequences converge to x. Now assume that (x_n) does not converge to x. Then there is an open set O containing x such that for an infinite number of indices x_n does not belong to O, i.e. there is a subsequence $(y_n) = (x_{\kappa(n)})$ with $\kappa : \mathbb{N} \to \mathbb{N}$ strictly increasing such that $x_{\kappa(n)} \notin O$. Thus no subsequence of $(x_{\kappa(n)})$ can converge to x. □

4.1.4. Continuity. The familiar definition of continuous functions of a real variable can be formulated for general maps between metric spaces.

4.26. Definition (Continuity). Let X, Y be metric spaces.

(i) A map $f : X \to Y$ is called *continuous at $x \in X$* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $d_X(x, y) < \delta$, then $d_Y(f(x), f(y)) < \varepsilon$.
(ii) f is called *continuous* if it is continuous at every $x \in X$. We also write $f \in C(X, Y)$.

4.27. Lemma. *Let $f : X \to Y$ and $x \in X$. The following are equivalent.*

(i) f is continuous at x.
(ii) For every sequence $(x_n)_{n\in\mathbb{N}} \subset X$ convergent to x, the sequence $(f(x_n))_{n\in\mathbb{N}}$ converges to $f(x)$.

Proof. *(i) $\Longrightarrow$ (ii).* Suppose $x_n \to x$ in X. Let $\varepsilon > 0$. Since by assumption f is continuous at x we find $\delta_\varepsilon > 0$ such that $d_Y(f(\tilde{x}), f(x)) < \varepsilon$ whenever $d_X(\tilde{x}, x) < \delta_\varepsilon$. Since $\lim_{n\to\infty} x_n = x$, we find N_ε so that $d_X(x_n, x) < \delta_\varepsilon$ for $n > N_\varepsilon$ and thus $d_Y(f(x_n), f(x)) < \varepsilon$ for $n > N_\varepsilon$. Since $\varepsilon > 0$ was arbitrary this shows that $f(x_n) \to f(x)$ in X.

(ii) $\Longrightarrow$ (i). We argue by contradiction and assume that f is not continuous at x. Then there exists an $\varepsilon > 0$ so that there is no $\delta > 0$ with the property that $d_Y(f(\tilde{x}), f(x)) < \varepsilon$ for all $\tilde{x} \in B(x, \delta)$. Taking δ to be a reciprocal of a positive integer, we see that for every $n \in \mathbb{N}$ there exists an x_n with $d_X(x_n, x) < 1/n$ but $d_Y(f(x_n), f(x)) \geqslant \varepsilon$. Then $x_n \to x$ in X but $(f(x_n))_{n\in\mathbb{N}}$ does not converge to $f(x)$. This is a contradiction. □

4.28. Definition (Inverse images). Let $f : X \to Y$ be a function and $A \subset Y$. The inverse image of A under f is defined as

$$f^{-1}(A) = \{x \in X : f(x) \in A\}.$$

4.29. Exercise. Definition 4.28 makes sense for all functions from X to Y. The notation does *not* imply that f is invertible. Find examples to illustrate this. Convince yourself that *if* f is invertible then $f^{-1}(A)$ is the image of A under the inverse map.

Prove that for $A \subset X$ we have $A \subset f^{-1}(f(A))$ and for $B \subset Y$ we have $f(f^{-1}(B)) \subset B$, but equality need not hold in either case.

4.30. Lemma. *Let $f : X \to Y$ be a map between metric spaces X and Y. The following are equivalent:*

(i) f is continuous.
(ii) $f^{-1}(U) \subset X$ is open for every open set $U \subset Y$.

Proof. *(i) $\Longrightarrow$ (ii).* Let $U \subset Y$ be open, and let $x \in f^{-1}(U)$. Since U is open there is an open ball $B(f(x), \varepsilon)$ contained in U. Since f is continuous at x there is an open ball $B(x, \delta)$ in X such that $d_Y(f(\tilde{x}), f(x)) < \varepsilon$ for every $\tilde{x} \in B(x, \delta)$. In particular $f(\tilde{x}) \in U$ for every $\tilde{x} \in B(x, \delta)$. Hence $B(x, \delta) \subset f^{-1}(U)$ so that x is an interior point of $f^{-1}(U)$. Since x was chosen arbitrary in $f^{-1}(U)$ we conclude that $f^{-1}(U)$ is open.

(ii) $\Longrightarrow$ (i). Let $\varepsilon > 0$. Let $x \in X$. By assumption $f^{-1}(B(f(x), \varepsilon))$ is open. The point x belongs to this set and thus is an interior point. Hence there exists a $\delta > 0$ (depending on x and ε) so that

$$B(x, \delta) \subset f^{-1}(B(f(x), \varepsilon)).$$

This means that $d_Y(f(\tilde{x}), f(x)) < \varepsilon$ provided that $d_X(\tilde{x}, x) < \delta$. Since $\varepsilon > 0$ was arbitrary we conclude that f is continuous at x. Since $x \in X$ was arbitrary f is continuous on X. □

As a consequence of Lemma 4.27 and Lemma 4.30, continuity of f can also be characterized by saying that f *commutes with limits*. That is,

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

provided that $(x_n)_{n\in\mathbb{N}}$ is a convergent sequence in X. This is also called *sequential continuity*. We consider as an example the continuity properties of the metric. We define a metric on $X \times X$ by

$$d_{X\times X}((x, \tilde{x}), (x', \tilde{x}')) = d(x, x') + d(\tilde{x}, \tilde{x}').$$

4.31. Example. For $x, y, x', y' \in X$

$$|d(x, y) - d(x', y')| \leqslant d(x, x') + d(y, y') \tag{4.1}$$

and thus $d : X \times X \to \mathbb{R}$ is continuous. (4.1) holds since by the triangle inequality $d(x, y) \leqslant d(x, x') + d(x', y') + d(y', y)$ and $d(x', y') \leqslant d(x', x) + d(x, y) + d(y, y')$.

4.32. Example. Continuity depends on the domain of definition of f. For instance, consider the function $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) = 1$ if $x \in \mathbb{Q}$ and $f(x) = 0$ otherwise. Then f is not continuous (anywhere), but the restricted function $f|_{\mathbb{Q}} : \mathbb{Q} \to \mathbb{R}$ is continuous (it is a constant function on $\mathbb{Q}$).

4.33. Definition (Isometries). Let (X, d_X), (Y, d_Y) be metric spaces. A function $f : X \to Y$ is called an *isometry* of metric spaces if for all $x, x' \in X$

$$d_Y(f(x), f(x')) = d_X(x, x').$$

4.1.5. Connected sets.

4.34. Definition. A metric space X is called *connected* if the only subsets that are both open and closed are the whole space X and the empty set. A subset of a metric space is *connected* if it is connected as a metric space.

4.35. Observation. *A metric space X is connected if and only if there do not exist non-empty, disjoint open subsets U, V such that $X = U \cup V$.*

This follows because if $X = U \cup V$ with U, V disjoint and open, then $U = V^{\complement}$ and $V = U^{\complement}$, so U and V are also closed.

4.36. Theorem. *A subset of $\mathbb{R}$ is connected if and only if it is an interval.*

Proof. First let E be a connected subset of $\mathbb{R}$. Let $A = \inf E$ and $B = \sup E$. In order to show that E is an interval we need to show that $(A, B) \subset E$. Suppose not. Then there exists x with $A < x < B$ so that $x \notin E$. Let $U = E \cap (-\infty, x)$ and $V = E \cap (x, \infty)$. Then U and V are open subsets of E, also disjoint and nonempty and thus E is not connected. This is a contradiction.

Now we show that an interval I is connected. Suppose not, then there are non-empty disjoint open subsets U, V of I such that $U \cup V = I$. Define a function f on $I = U \cup V$ by $f(x) = 1$ if $x \in U$ and $f(x) = 0$ if $x \in V$. Since $U \cap V = \varnothing$, f is well defined on $U \cup V = I$. Since U and V are open, f is continuous on I. f takes the values 0 and 1. By the intermediate value theorem (Theorem 1.88) all numbers between 0 and 1 are in the range of f; this gives a contradiction. Hence I is connected. □

4.37. Exercise. Show that every open subset of $\mathbb{R}$ is the union of an at most countable collection of open intervals.

The following theorem can be seen as a generalization of the intermediate value theorem.

4.38. THEOREM. *Let $f : X \to Y$ be a continuous map between metric spaces X and Y. If X is connected, then $f(X) \subset Y$ is connected.*

PROOF. Suppose $f(X) = U \cup V$ with U, V disjoint subsets of $f(X)$, both open in the relative topology of $f(X)$. Then $X = f^{-1}(U) \cup f^{-1}(V)$. Since U, V are open in $f(X)$, there exist $\tilde{U}, \tilde{V}$ open in Y so that $U = \tilde{U} \cap f(X)$ and $V = \tilde{V} \cap f(X)$. Thus $f^{-1}(U) = f^{-1}(\tilde{U})$ and $f^{-1}(V) = f^{-1}(\tilde{V})$ are open sets in X, since f is continuous. They are also disjoint since U and V are disjoint. Since X is connected, one of $f^{-1}(U)$, $f^{-1}(V)$ must be empty, so one of U, V must be empty. □

4.39. EXERCISE.

(i) Let A, B be connected subsets of X such that $A \cap B \neq \varnothing$. Then $A \cup B$ is connected.

(ii) The same conclusion holds if we replace the intersection assumption in (i) by the weaker assumption $\overline{A} \cap B \neq \varnothing$ (however, it is not possible to weaken this further to $\overline{A} \cap \overline{B} \neq \varnothing$).

4.40. DEFINITION. Let X be a metric space. Two points x, y in X are *connected in* X if there exists a connected subset of X that contains them both.

4.41. EXERCISE. Prove that the relation of being connected in X is an equivalence relation on X.

The equivalence classes of this relation are called the *connected components* of X.

4.42. DEFINITION. Let X be a metric space.

(i) Let $x, y \in X$. A *path* from x to y is a continuous function $\gamma : [0,1] \to X$ such that $\gamma(0) = x$ and $\gamma(1) = y$.

(ii) X is *path connected* if for every pair $x, y \in X$ there exists a path from x to y.

4.43. EXERCISE. We say that x and y are path equivalent in X if there exists a path from x to y. Show that this indeed defines an equivalence relation.

4.44. THEOREM. *A path connected metric space X is connected.*

PROOF. Argue by contradiction, and assume that X is not connected. Then X is the union of two disjoint non-empty open sets U, V. Let $x \in U$ and $y \in V$ and let γ be a path from x to y. Let $\Gamma = \{\gamma(t) : t \in [0,1]\}$. Then Γ is connected by Theorem 4.38. The sets $\Gamma \cap U$, $\Gamma \cap V$ are nonempty disjoint sets in Γ, open in the relative topology, and their union is Γ. This is a contradiction. □

4.45. EXAMPLE. Open balls in $\mathbb{R}^n$ are path connected (and therefore connected). Setting $\|x\| = (\sum_{j=1}^n |x_j|^2)^{1/2}$ let $B = B(z,r) = \{x : \|x - z\| < r\}$. Let $x, y \in B$ and define

$$\gamma(t) = x + t(y - x), \quad 0 \leqslant t \leqslant 1.$$

γ is continuous, and a path from x to y in $\mathbb{R}^n$. Since $\|x - z\| < r$, $\|y - z\| < r$ we also get for $0 \leqslant t \leqslant 1$

$$\|\gamma(t) - z\| = \|(1-t)x + ty - z\| = \|(1-t)(x-z) + t(y-z)\| \leqslant (1-t)\|x - z\| + t\|y - z\| < r.$$

Hence γ is a path from x to y within B which proves the statement.

4.46. EXAMPLE. Closed balls in $\mathbb{R}^n$ are path connected (and therefore connected). The argument in Example 4.45 applies, with a modification.

4.47. EXERCISE. Let X be a connected metric space, or a path connected vector space. In both cases answer the following questions with a proof or counterexample.

(i) Is it necessarily true that an open ball in X is connected?

(ii) Is it necessarily true that an open ball in X is path connected?

4.48. EXERCISE. The converse of Theorem 4.44 does not hold: Consider the following subsets of $\mathbb{R}^2$: $A = \{(0,0)\}$, $B = \{(x, \sin(1/x)) : 0 < x \leqslant 1\}$.

Prove that A and B are path connected, $A \cup B$ is connected, but $A \cup B$ is not path connected.

4.1.6. Oscillation as a quantification of discontinuity. In this section we let (X, d_X), (Y, d_Y) be metric spaces and $f : X \to Y$ be a function. For a set S consider

$$\operatorname{diam}(f(S)) = \sup_{x,x' \in S} d_Y(f(x), f(x'))$$

for S being a ball $B(x, \delta)$. Notice that $\operatorname{diam}(f(B(x, \delta)))$ decreases when δ decreases.

4.49. Definition. For a point $x \in X$ we define the *oscillation of f at x* by

$$\operatorname{osc}_f(x) = \inf_{\delta > 0} \operatorname{diam}(f(B(x, \delta)))$$

The number $\operatorname{osc}_f(x)$ can be used to quantify discontinuities:

4.50. Observation. *A function $f : X \to Y$ is continuous at $a \in X$ if and only if* $\operatorname{osc}_f(a) = 0$.

This is a consequence of the definition of continuity.

4.51. Lemma. *Given a function $f : X \to Y$ and $\gamma > 0$, the set $\{x : \operatorname{osc}_f(x) \geqslant \gamma\}$ is closed.*

Proof. For $\gamma > 0$ we show that $\Omega_\gamma = \{x \in X : \operatorname{osc}_f(x) < \gamma\}$ is open. Either Ω_γ is empty (hence open) or there exists $x_0 \in \Omega_\gamma$. Then there exists $\delta > 0$ such that $\operatorname{diam}(f(B(x_0, \delta))) < \gamma$. We are done if we can show that for x with $d(x, x_0) < \delta/2$ we have $\operatorname{diam}(f(B(x, \delta/2))) < \gamma$. So let $x \in B(x_0, \delta/2)$. Then by the triangle inequality, $B(x, \delta/2) \subset B(x_0, \delta)$. Therefore,

$$\operatorname{osc}_f(x) \leqslant \operatorname{diam}(f(B(x, \delta/2))) \leqslant \operatorname{diam}(f(B(x_0, \delta))) < \gamma$$

hence $x \in \Omega_\gamma$. □

4.52. Exercise. Define $f : [-10, 10] \to \mathbb{R}$ by $f(x) = -4x$ for $x \leqslant 0$, $f(x) = \sin(\pi/x)$ for $0 < x < 3/2$, $f(x) = \cos(\pi/x)$ for $x \geqslant 3/2$. Determine $\operatorname{osc}_f(x)$ for all $x \in [-10, 10]$.

4.53. Exercise. Recall *Thomae's function* from Exercise 1.84: $f : [0, 1] \to \mathbb{R}$, defined by

$$f(x) = \begin{cases} 0 & \text{if } x \in [0,1] \backslash \mathbb{Q}, \\ \frac{1}{n} & \text{if } x = \frac{m}{n} \text{ with } \gcd(m, n) = 1. \end{cases}$$

Find $\operatorname{osc}_f(x)$ for all $x \in [0, 1]$.

4.1.7. Uniform convergence. We return to the concept of uniform convergence which was covered in §2.8 for sequences of functions with values in $\mathbb{R}$ or $\mathbb{C}$. Now consider sequences of functions that take values in a metric space Y.

4.54. Definition. Given $f_n : X \to Y$ we say that the sequence $(f_n)_{n \in \mathbb{N}}$ is *uniformly convergent* to a function f if for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$ and all $x \in X$,

$$d_Y(f_n(x), f(x)) < \varepsilon.$$

One can rephrase uniform convergence by taking a supremum over x. Indeed, (f_n) converges uniformly to f if and only if

$$\lim_{n \to \infty} \sup_{x \in X} d_Y(f_n(x), f(x)) = 0.$$

We will now state a few theorems about uniform convergence whose proofs are all modifications of the proofs in the special case of §2.8, and will be left to the reader.

4.55. Theorem. *Let X be a set, and (Y, d) be a complete metric space. Let $f_n : X \to Y$. Then the sequence $(f_n)_{n \in \mathbb{N}}$ converges uniformly if and only if it is* uniformly Cauchy, *i.e. for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $n, m \geqslant N$ and all $x \in X$, $d(f_n(x), f_m(x)) < \varepsilon$.*

4.56. Exercise. Show by example that the assumption of completeness of Y in Theorem 4.55 cannot be dropped.

4.57. Definition. Let Y be a metric space. A subset $A \subset Y$ is *bounded* if the *diameter* of A,

$$\operatorname{diam}(A) = \sup\{d(y, y') \,:\, y, y' \in A\}$$

is finite. Let X be a set and $f : X \to Y$ be a function. f is *bounded* or *has bounded range* if the image $f(X) = \{f(x) : x \in X\}$ has finite diameter.

4.58. Theorem. *If $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f and each f_n has bounded range, then f has bounded range.*

We shall now assume that X is a metric space, with metric d_X.

4.59. Theorem. *Let (X, d_X), (Y, d_Y) be metric spaces and $a \in X$. Let $f_n : X \to Y$ and let $(f_n)_{n\in\mathbb{N}}$ converge uniformly to f. Then, if each f_n is continuous at a, f is continuous at a.*

4.60. Exercise. Let (X, d_X) and (Y, d_Y) be metric spaces. Prove each of the following:

(i) The space $C_b(X, Y)$ of continuous functions $f : X \to Y$ with bounded range is a metric space with the *supremum metric*

$$d_\infty(f, g) = \sup_{x\in X} d_Y(f(x), g(x)).$$

(ii) If (Y, d_Y) is complete, then $(C_b(X, Y), d_\infty)$ is also complete.

(iii) Let $(f_n)_{n\in\mathbb{N}} \subset C_b(X, Y)$ be a sequence. Then $(f_n)_{n\in\mathbb{N}}$ converges in $C_b(X, Y)$ (with respect to d_∞) if and only if it converges uniformly to some $f \in C_b(X, Y)$.

4.2. The contraction principle

The contraction principle is a powerful tool for solving equations.

4.61. Definition. A map $\varphi : X \to X$ is called a *contraction* (of X) if there exists a constant $c \in (0, 1)$ such that

$$d(\varphi(x), \varphi(y)) \leqslant c \cdot d(x, y) \tag{4.2}$$

holds for all $x, y \in X$.

In other words, a map $\varphi : X \to X$ is a contraction if it uniformly diminishes distances between points in X. In particular, observe that contractions are continuous.

4.62. Exercise. Let $f : (a, b) \to (a, b)$ be differentiable and suppose that $c \in (0, 1)$ is such that $|f'(x)| \leqslant c$ for all $x \in (a, b)$. Show that f is a contraction of (a, b).

4.63. Definition. A point $x \in X$ such that $\varphi(x) = x$ is called a *fixed point* of φ.

The following important *Banach fixed point theorem* (also known as *contraction mapping theorem*) tells us that contractions on complete metric spaces have unique fixed points which can be computed by an iterative procedure.

4.64. Theorem. *Let X be a complete metric space and $\varphi : X \to X$ a contraction. Then the following hold:*

(i) There exists a unique $x_* \in X$ such that $\varphi(x_*) = x_*$.

(ii) Let $x_0 \in X$. Define a sequence $(x_n)_{n\geqslant 0}$ recursively by

$$x_{n+1} = \varphi(x_n), \quad n \geqslant 0. \tag{4.3}$$

Then for all $n \geqslant 1$,

$$d(x_n, x_{n+1}) \leqslant c^n d(x_0, x_1) \tag{4.4a}$$

and

$$d(x_n, x_*) \leqslant \frac{c^n}{1 - c} d(x_0, x_1), \tag{4.4b}$$

PROOF. *Uniqueness of the fixed point:* Suppose $x_0, x_1 \in X$ are fixed points of φ. Then

$$0 \leqslant d(x_0, x_1) = d(\varphi(x_0), \varphi(x_1)) \leqslant c \cdot d(x_0, x_1),$$

which implies $d(x_0, x_1) = 0$, since $c \in (0, 1)$. Thus $x_0 = x_1$.

Existence of the fixed point: Pick $x_0 \in X$ arbitrarily and use the iteration (4.3). We claim that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Indeed, as $x_{n+1} = \varphi(x_n)$, we see by induction that

$$d(x_{n+1}, x_n) \leqslant cd(x_n, x_{n-1}) \leqslant c^2 d(x_{n-1}, x_{n-2}) \leqslant \cdots \leqslant c^n d(x_1, x_0).$$

Thus, for $n < m$ we can use the triangle inequality to obtain

$$\begin{aligned} d(x_m, x_n) &\leqslant d(x_m, x_{m-1}) + d(x_{m-1}, x_{m-2}) + \cdots + d(x_{n+1}, x_n) \\ &= \sum_{i=n}^{m-1} d(x_{i+1}, x_i) \leqslant \sum_{i=n}^{m-1} c^i d(x_1, x_0) \leqslant d(x_1, x_0) \sum_{i=n}^{\infty} c^i = c^n \frac{d(x_1, x_0)}{1-c}. \end{aligned} \tag{4.5}$$

Since $\lim_{n\to\infty} c^n = 0$ this shows that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. By completeness of X, it must converge to a limit which we call $x_* \in X$. By continuity of φ,

$$\varphi(x_*) = \varphi(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} \varphi(x_n) = \lim_{n\to\infty} x_{n+1} = x_*.$$

This shows the existence of the fixed point.

To prove the a priori estimate (4.4b) just let $m \to \infty$ in (4.5). □

REMARKS.

1. The Banach fixed point theorem does not just show the existence and uniqueness of the fixed point x_*, but also gives an efficient algorithm to compute it. The iteration (4.3) is called the *method of successive approximation* and the a priori estimate (4.4b) gives information on how quickly the algorithm converges. The speed of convergence depends only on the parameter $c \in (0, 1)$ and the quality of the initial guess $x_0 \in X$.

2. The conclusion does not necessarily hold if we drop the assumption that X is complete: the map $f : (0, 1) \to (0, 1)$ defined by $f(x) = x/2$ is a contraction (in which metric space?) but has no fixed point.

3. If we drop the uniformity in the distance diminishing assumption (4.2) and replace it by the weaker condition

$$|\varphi(x) - \varphi(y)| < |x - y| \tag{4.6}$$

for all x, y with $x \neq y$, then φ may not have a fixed point in X. Consider $\varphi(x) = x + e^{-x}$ on the complete metric space $X = [0, \infty)$. One verifies that $\varphi'(x) = 1 - e^{-x} \in (0, 1)$ for $x \geqslant 0$ thus (4.6) is satisfied if $x \neq y$. But clearly $\varphi(x) - x > 0$ for $x \geqslant 0$ so φ does not have a fixed point.

4.65. EXERCISE. Show that the system of equations

$$\begin{aligned} x_1 + \tfrac{1}{10}\cos(\sin(2x_2 + x_1)) &= 6 \\ x_2 + \tfrac{1}{12}e^{-x_1^2} + \tfrac{1}{10}\cos(x_1 + x_2) &= 7 \end{aligned}$$

has a unique solution $(x_1, x_2) \in \mathbb{R}^2$.

4.66. EXERCISE. Show that there exists a unique point $(x, y) \in \mathbb{R}^2$ such that

$$\cos(\sin(x)) = y \quad \text{and} \quad \sin(\cos(y)) = x.$$

4.67. EXERCISE.

(i) Show there is exactly one $u \in C([-1, 1])$ that satisfies the integral equation

$$u(x) \;=\; x \int_0^x t^2 \cos\big(u(t)\big)\, dt, \qquad x \in [-1, 1].$$

(ii) Show that u is differentiable. Is u' differentiable?

4.68. Exercise. Let $u : \mathbb{R} \to \mathbb{R}$ be a C^1-function, such that $|u'(t)| \leqslant a < \frac{1}{2}$ for all $t \in \mathbb{R}$. Define a C^1-function $g : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$g(x, y) = (x + u(2y), y + u(x)).$$

Show that the range of g is all of $\mathbb{R}^2$.

4.3. Compactness

Compactness is a fundamental topological property of metric spaces and plays an important role in analysis.

4.69. Definition. Let X be a metric space.

(i) An *open cover* is a collection $(G_i)_{i\in I}$ of open sets $G_i \subset X$ with $X \subset \bigcup_{i\in I} G_i$. Here I is an arbitrary index set.
(ii) X is called *compact* if every open cover of X contains a finite sub-cover. This means for any open cover $(G_i)_{i\in I}$ there exists $m \in \mathbb{N}$ and $i_1, \ldots, i_m \in I$ such that $X \subset \bigcup_{j=1}^m G_{i_j}$.
(iii) The defining property in (ii) is also called the *Heine–Borel property.*
(iv) A subset $A \subset X$ is called *compact* if $(A, d|_{A\times A})$ is a compact metric space.
(v) A subset $A \subset X$ is called *relatively compact* if the closure $\overline{A} \subset X$ is compact.

It seems hard to verify compactness of a given set directly from the definition, so the Heine–Borel property appears unmotivated at first glance. However, it turns out to be extremely useful for proving interesting theorems. We will see this right away, in §4.3.1. Efficient tools for the verification of the compactness property, with various equivalent characterizations, will be subsequently developed. The main result will be Theorem 4.92 and its proof in §4.3.2.

4.70. Examples.

- If X has only finitely many elements, then it is compact.
- Borel showed that bounded and closed intervals $[a, b]$ satisfy the Heine-Borel property (this is known as *Borel's lemma*). More generally, we will see from Corollary 4.93 below that the compact subsets of Euclidean space $\mathbb{R}^n$ are exactly the closed and bounded sets. For example, the unit spheres in $\mathbb{R}^n$,

$$\{x \in \mathbb{R}^n \,:\, \sum_{i=1}^n |x_i - a_i|^2 = 1\}$$

  are compact. However, this is not easy to see directly from the above definition.
- For general metric spaces X, the "spheres"

$$S(x_0, r) = \{x \in X \,:\, d(x, x_0) = r\}$$

  are *not* necessarily compact.

4.71. Example. Let $A \subset \ell^\infty$ be the "unit vectors", i.e the functions $e_k : \mathbb{N} \to \mathbb{R}$ with $e_k(k) = 1$ and $e_k(n) = 0$ for $n \neq k$. Then

$$d(e_k, e_l) = \sup_n |e_k(n) - e_l(n)| = 1 \quad \text{for } k \neq l.$$

Thus A is a closed bounded subset of ℓ^∞ and every singleton set $\{e_k\}$ is open and closed in the relative topology on A, so the singleton sets form an open cover of A that does not have a finite subcover. Therefore, A is not compact.

As a warm-up in dealing with the definition of compactness, let us prove the following.

4.72. Proposition. *A closed subset of a compact metric space is compact.*

PROOF. Let $\mathcal{G} = (G_i)_{i\in I}$ be an open cover of a closed subset $A \subset X$. That is, $G_i \subset A$ is open with respect to the relative topology on A. Then $G_i = U_i \cap A$ for some open $U_i \subset X$ (see Lemma 4.15). Since A is closed, $X\backslash A$ is open. Thus,

$$\{U_i \,:\, i \in I\} \cup \{X\backslash A\}$$

is an open cover of X, which by the compactness assumption has a finite subcover

$$\{U_{i_k} \,:\, k = 1, \dots, M\} \cup \{X\backslash A\}.$$

Now observe that $\{G_{i_k} \,:\, k = 1, \dots, M\} \subset \mathcal{G}$ is a finite subcover of A. □

4.73. EXERCISE. A collection $\{F_\alpha : \alpha \in A\}$ of closed sets has the *finite intersection property* if for every finite subset A_o of A the intersection $\bigcap_{\alpha\in A_o} F_\alpha$ is not empty. Prove that the following are equivalent:

(i) A metric space X, with metric d, is compact.

(ii) For every collection $\{F_\alpha\}_{\alpha\in A}$ of closed sets with the finite intersection property it follows that

$$\bigcap_{\alpha\in A} F_\alpha \neq \varnothing.$$

4.74. EXERCISE. Let X be a compact metric space. Prove that there exists a countable, dense set $E \subset X$ (recall that $E \subset X$ is called *dense* if $\overline{E} = X$).

4.75. EXERCISE. Construct a compact subset of $\mathbb{R}$ whose accumulation points form an infinite countable set.

4.3.1. Compactness and continuity.

4.76. DEFINITION. Let X and Y be metric spaces. A function $f : X \to Y$ is called *uniformly continuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d_Y(f(x), f(x')) < \varepsilon$ holds for all x, x' with $d_X(x, x') < \delta$.

Heine (1870) proved that continuous functions on closed bounded intervals satisfy the stronger property of uniform continuity. Heine's theorem holds on compact metric spaces:

4.77. THEOREM. *Let X, Y be metric spaces, assume that X is compact, and let $f : X \to Y$ be continuous. Then f is uniformly continuous.*

PROOF. Let $\varepsilon > 0$. We need to demonstrate the existence of a number $\delta > 0$ such that for all $x, y \in X$ we have that $d_X(x, y) \leqslant \delta$ implies $d_Y(f(x), f(y)) < \varepsilon$. By continuity, for every $x \in X$ there exists a number $\delta_x > 0$ such that for all $y \in X$, $d_X(x, y) \leqslant \delta_x$ implies $d_Y(f(x), f(y)) < \varepsilon/2$. Let

$$B_x = B(x, \delta_x/2) = \{y \in X \,:\, d_X(x, y) < \delta_x/2\}.$$

Then $(B_x)_{x\in X}$ is an open cover of X. By compactness, there exists a finite subcover by $B_{x_1}, \dots, B_{x_m}$. Now we set

$$\delta = \tfrac{1}{2} \min(\delta_{x_1}, \dots, \delta_{x_m}).$$

We claim that this δ does the job. Indeed, let $x, y \in X$ satisfy $d_X(x, y) \leqslant \delta$. There exists $i \in \{1, \dots, m\}$ such that $x \in B_{x_i}$. Then

$$d_X(x_i, y) \leqslant d_X(x_i, x) + d_X(x, y) \leqslant \tfrac{1}{2}\delta_{x_i} + \delta \leqslant \delta_{x_i}.$$

Thus, by definition of δ_{x_i},

$$d_Y(f(x), f(y)) \leqslant d_Y(f(x), f(x_i)) + d_Y(f(x_i), f(y)) < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$ □

4.78. EXAMPLE. As an illustration of how uniform continuity is often used, we present a formula about differentiation of integrals depending on a parameter. Let f be a continuous real-valued function on the compact rectangle $R = [a, b] \times [c, d]$ and consider the integral

$$\mathcal{I}(y) = \int_a^b f(x, y)\, dx.$$

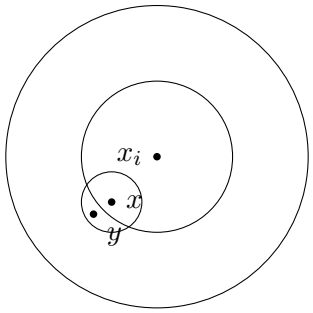


FIGURE 4.2. The balls B_{x_i}, $B(x_i, \delta_{x_i})$, $B(x, \delta)$.

Assume that for every $x \in [a, b]$ the function $y \mapsto f(x, y)$ is differentiable and denote the derivative by $\frac{\partial f}{\partial y}(x, y)$; moreover assume that $(x, y) \mapsto \frac{\partial f}{\partial y}(x, y)$ is continuous on R. We show that $\mathcal{I}$ is differentiable on $[c, d]$ with

$$\mathcal{I}'(y) = \int_a^b \frac{\partial f}{\partial y}(x, y)\, dx.$$

To see this we write (for $y, y + h \in [c, d]$)

$$\frac{\mathcal{I}(y + h) - \mathcal{I}(y)}{h} - \int_a^b \frac{\partial f}{\partial y}(x, y)\, dx = \int_a^b \frac{1}{h}\Big(f(x, y + h) - f(x, y) - h\frac{\partial f}{\partial y}(x, y)\Big)\, dx$$

By the mean value inequality in Corollary 2.39

$$\Big|\frac{1}{h}\Big(f(x, y + h) - f(x, y) - h\frac{\partial f}{\partial y}(x, y)\Big)\Big| \leqslant \max\Big\{\Big|\frac{\partial f}{\partial y}(x, \eta) - \frac{\partial f}{\partial y}(x, y)\Big| : |\eta - y| \leqslant |h|,\ (x, \eta), (x, y) \in R\Big\}.$$

By Theorem 4.77, $\frac{\partial f}{\partial y}$ is uniformly continuous on R. Therefore there exists for every $\varepsilon > 0$ a $\delta > 0$ so that $|\frac{\partial f}{\partial y}(x, y + h) - \frac{\partial f}{\partial y}(x, y)| < \frac{\varepsilon}{b-a}$ provided that $(x, y + h), (x, y) \in R$ and $|h| \leqslant \delta$. Hence we get

$$\Big|\frac{\mathcal{I}(y + h) - \mathcal{I}(y)}{h} - \int_a^b \frac{\partial f}{\partial y}(x, y)\, dx\Big| < \varepsilon$$

and, since $\varepsilon > 0$ was arbitrary, the differentiation result holds.

4.79. THEOREM. *Let X be a compact metric space, Y a metric space, and $f : X \to Y$ a continuous map. Then $f(X) = \{f(x) : x \in X\}$ is compact.*

PROOF. Let $(V_i)_{i\in I}$ be an open cover of $f(X)$. Since f is continuous, the sets $U_i = f^{-1}(V_i) \subset X$ are open. We have $f(X) \subset \bigcup_{i\in I} V_i$. So,

$$X \subset f^{-1}(f(X)) \subset \bigcup_{i\in I} f^{-1}(V_i) = \bigcup_{i\in I} U_i.$$

$(U_i)_{i\in I}$ is an open cover of X. By compactness there exists a finite subcover $\{U_{i_1}, \dots, U_{i_m}\}$, that is, $X \subset \bigcup_{k=1}^m U_{i_k}$. Consequently,

$$f(X) \subset \bigcup_{k=1}^m f(U_{i_k}) \subset \bigcup_{k=1}^m V_{i_k}.$$

Thus $\{V_{i_1}, \dots, V_{i_m}\}$ is a finite open cover of $f(X)$. □

4.80. Definition. Let (X, d_X), (Y, d_Y) be metric spaces and let $f : X \to Y$ be *bijective*. If $f : X \to Y$ and $f^{-1} : Y \to X$ are continuous then f is called a *homeomorphism* from X to Y. The spaces X and Y are called *homeomorphic* or *homeomorphically equivalent* if there exists a homeomorphism from X to Y.

The reader will check that this defines an equivalence relation between metric spaces.

4.81. Theorem. *Let X, Y be metric spaces, let X be compact, and let $f : X \to Y$ be a bijective continuous map. Then Y is compact and f is a homeomorphism from X to Y.*

Proof. By Theorem 4.79 $Y = f(X)$ is compact. Moreover if $U \subset X$ is open then $X\backslash U$ is closed and therefore compact and thus $f(X\backslash U)$ is compact. Hence $f(U) = Y\backslash f(X\backslash U)$ is open in Y. But $f(U) = (f^{-1})^{-1}(U)$ (the inverse image of U under the map f^{-1}). This shows (see Lemma 4.30) that $f^{-1} : Y \to X$ is continuous and thus f is a homeomorphism. □

In the following we apply the Heine–Borel property to monotone sequences of continuous real-valued functions.

4.82. Theorem (Dini). *Let X be a compact metric space and let $f_n : X \to \mathbb{R}$, and $f : X \to \mathbb{R}$ be such that $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in X$. Also assume that all f_n and f are continuous on X and that $f_n(x) \leqslant f_{n+1}(x)$ for all $n \in \mathbb{N}$ and $x \in X$. Then f_n converges uniformly to f on X.*

Proof. Notice that for the limiting function f we have $f(x) - f_n(x) \geqslant 0$ for all $x \in X$. Let $\varepsilon > 0$. Let $U_n = \{x \in X : f(x) - f_n(x) < \varepsilon\}$. Since $f_n(x)$ is increasing and pointwise convergent to f we see that $U_n \subset U_{n+1}$ and $\cup_{n=1}^{\infty} U_n = X$. Since X is compact and the U_n are increasing there exists an N such that $U_N = X$, then also $U_n = X$ for all $n \geqslant N$ and so we have $0 \leqslant f(x) - f_n(x) < \varepsilon$ for all $x \in X$ and all $n \geqslant N$. Since ε was arbitrary we conclude that f_n converges uniformly to f. □

4.83. Exercise. Is the conclusion in Dini's theorem still true if we remove the assumption that X is compact? Give a proof or a counterexample.

4.84. Example. Define a sequence of polynomials recursively by

$$P_0(x) = 0, \qquad P_{n+1}(x) = P_n(x) + \tfrac{1}{2}\big(x^2 - P_n(x)^2\big).$$

Then (P_n) is a sequence of polynomials, increasing for every $x \in [-1, 1]$ such that

$$\lim_{n\to\infty} P_n(x) = |x|$$

with uniform convergence on $[-1, 1]$. Moreover $P_n(0) = 0$ for all $n \in \mathbb{N}$ which is immediate from the recursion relation and induction.

We examine $|x| - P_n(x)$ and note

$$|x| - P_{n+1}(x) = |x| - P_n(x) - \tfrac{1}{2}\big(|x| - P_n(x)\big)\big(|x| + P_n(x)\big)$$

which gives

$$|x| - P_{n+1}(x) = \big(|x| - P_n(x)\big)\big(1 - \tfrac{1}{2}(|x| + P_n(x))\big).$$

By induction on n this identity implies, for all $n \in \mathbb{N}_0$,

$$0 \leqslant P_n(x) \leqslant P_{n+1}(x) \leqslant |x|$$

for all $x \in [-1, 1]$. The increasing sequence $(P_n(x))_{n\in\mathbb{N}}$ is bounded above and thus converges to a limit $f(x)$ which satisfies

$$f(x) = f(x) + \frac{1}{2}(x^2 - f(x)^2),$$

by taking the limit in the recursion relation. Hence $|x|^2 = f(x)^2$ and since $f(x) \geqslant 0$ this implies $f(x) = |x|$. Dini's theorem shows that P_n converges to f uniformly on $[-1, 1]$. For more on this example, see Exercise 4.173.

Our next exercise shows that in Dini's Theorem 4.82 the assumption of continuity of the limiting function cannot be dropped.

4.85. Exercise. Define the sequence $f_n : [0,\infty) \to \mathbb{R}$, $n \in \mathbb{N}$ recursively by

$$f_1(x) = x^{1/2}, \qquad f_{n+1}(x) = \sqrt{x + f_n(x)},\ n \in \mathbb{N}.$$

(i) Prove that $f_n(0) = 0$, and for $x > 0$,

$$f_n(0) < f_n(x) < f_{n+1}(x) < 1 + x.$$

Show that for every $x \in [0,\infty)$ the sequence $f_n(x)$ converges and determine the limit.

(ii) Show that the limit function $f : [0,\infty) \to \mathbb{R}$ is not continuous at $x = 0$.

(iii) Let $0 < a < b < \infty$. Show that f_n converges uniformly on $[a, b]$.

(iv) Show that f_n does not converge uniformly on $[0, b]$ for any $b > 0$.

4.3.2. Sequential compactness and total boundedness. The reader may recall the Bolzano–Weierstrass theorem for the real line which states that every bounded sequence of real numbers has a convergent subsequence (see Theorem 1.60).

4.86. Definition. A metric space X is *sequentially compact* if every sequence in X has a convergent subsequence. This is also called the *Bolzano–Weierstrass property.*

4.87. Definition. A metric space X is *totally bounded* if for every $\varepsilon > 0$ there exist finitely many balls of radius ε that cover X.

Similarly, we define these terms for subsets $A \subset X$ by considering $(A, d|_{A\times A})$ as its own metric space. In applications the following lemma is useful (it takes away the requirement that in order to check total boundedness of a subset A the covering balls need to be centered in A).

4.88. Lemma. *Let (X, d) be a metric space and let $A \subset X$, with the metric inherited from X. A is totally bounded if and only if for every $\varepsilon > 0$ there is a finite collection $\{X_1, \dots, X_N\}$ of subsets of X, with $\operatorname{diam}(X_i) < \varepsilon$, such that $A \subset \bigcup_{j=1}^N X_j$.*

Proof. If A is totally bounded then we can cover A with finitely many balls centered in A of radius $\varepsilon/2$ (which are of diameter ε), so the stated condition holds.

Vice versa assume the stated condition holds for all $\varepsilon > 0$. Fix $\varepsilon > 0$ and find sets $X_1, \dots, X_N$, each of diameter $< \varepsilon$, such that $A \subset \bigcup_{j=1}^N X_j$. We may assume that each X_j contains a point a_j in A (otherwise discard those X_j which don't contain points in A). For $x \in X_j$ we then have $d(x, a_j) < \varepsilon$ and thus $X_j \subset B_X(a_j, \varepsilon)$, the ball in X centered at a_j with radius ε. Now let $B_A(a_j, \varepsilon) = \{a \in A : d(a, a_j) < \varepsilon\}$. Then A is contained in $A \cap \bigcup_{j=1}^N B_X(a_j, \varepsilon) = \bigcup_{j=1}^N B_A(a_j, \varepsilon)$. Since $\varepsilon > 0$ was arbitrary, A is totally bounded. □

4.89. Lemma. *Let X be a metric space, $A \subset X$. Then A is totally bounded if and only if $\overline{A}$ is totally bounded.*

Proof. A covering of $\overline{A}$ with ε-balls is a covering of A. On the other hand, if $(B(x_i, \frac{\varepsilon}{2}))_{i=1,\dots,N}$ is a covering of A with $\frac{\varepsilon}{2}$-balls then the expanded ε-balls $B(x_i, \varepsilon)$ cover $\overline{A}$. □

Recall from Definition 4.57 that a metric space is called bounded if it has finite diameter, i.e. if there exist $x_0 \in X$ and $r > 0$ such that X is contained in the single fixed ball $B(x_0, r)$. Total boundedness is a significant strengthening of boundedness. Indeed, by the triangle inequality, every totally bounded metric space is bounded, but the converse is generally false (cf. Examples 4.71, 4.95). However, in the special case of subsets of our finite-dimensional Euclidean space $\mathbb{R}^n$ (or $\mathbb{C}^n$) the two notions are equivalent:

4.90. Lemma. *Consider $X = \mathbb{R}^n$ or $\mathbb{C}^n$ with the metric $d(x, y) = (\sum_{j=1}^n |x_j - y_j|^2)^{1/2}$. Then $A \subset X$ is totally bounded if and only if A is bounded.*

PROOF. It suffices to prove this for $X = \mathbb{R}^n$, since the corresponding results for $\mathbb{R}^{2n}$ also give the result for $\mathbb{C}^n$. For the nontrivial implication assume that A is bounded. Then there is $R > 1$ such that $|x| < R$ for all $x \in A$. Let $\varepsilon > 0$. We will cover the set A with finitely many ε-balls centered in A. Let $\delta = (2n)^{-1/2}\varepsilon$ and consider the family of cubes

$$Q_k = [k_1\delta, (k_1+1)\delta] \times \cdots \times [k_n\delta, (k_n+1)\delta], \quad k \in \mathbb{Z},$$

these are cubes with sidelength δ, and sides parallel to the coordinate axes. We have $\operatorname{diam}(Q_k) = \sqrt{n}\delta < \varepsilon$. Notice that every point in $\mathbb{R}^n$ lies in a Q_k and A is contained in the union of those Q_k which satisfy $-R - \delta \leqslant k_i\delta \leqslant R$ for $i = 1, \ldots, n$. These cubes form a finite family of sets of diameter $< \varepsilon$ which cover A and since ε was arbitrary it follows from Lemma 4.88 that A is totally bounded. □

4.91. EXERCISE. Prove the same result when the Euclidean metric is replaced with $d_1(x, y) = \sum_{i=1}^n |x_i - y_i|$ or $d_\infty(x, y) = \max_{i=1,\ldots,n} |x_i - y_i|$.

The main theorem on *characterizations of compactness* in general metric spaces is as follows.

4.92. THEOREM. *Let X be a metric space. The following are equivalent:*

(i) X is compact.
(ii) X is sequentially compact.
(iii) X is totally bounded and complete.

The proof of the theorem is given in §4.3.4 below. For $X = \mathbb{R}^n$, we get

4.93. COROLLARY.

(i) (Heine–Borel theorem) A subset $A \subset \mathbb{R}^n$ is compact if and only if it is bounded and closed.
(ii) (Bolzano–Weierstrass theorem) A subset $A \subset \mathbb{R}^n$ is sequentially compact if and only if it is bounded and closed.

PROOF OF COROLLARY 4.93. A subset $A \subset \mathbb{R}^n$ is closed if and only if A is complete as a metric space (this is because $\mathbb{R}^n$ is complete). Moreover from Lemma 4.90 we know that $A \subset \mathbb{R}^n$ is bounded if and only if it is totally bounded. Therefore, both claims follow from Theorem 4.92. □

We discuss a few examples.

4.94. EXAMPLE. Let $\ell^\infty \equiv \ell^\infty(\mathbb{N})$ be the space of functions $\mathbb{N} \to \mathbb{R}$ which are bounded (also known as the space of bounded sequences). The metric is given by

$$d(a, b) = \sup_{n \in \mathbb{N}} |a(n) - b(n)|.$$

Note that $\ell^\infty = C_b(\mathbb{N})$. By Example 4.71 the closed unit ball centered at the origin,

$$\overline{B}(0, 1) = \{f \in \ell^\infty \,:\, |f(n)| \leqslant 1 \,\forall n \in \mathbb{N}\}$$

is not totally bounded. Indeed check that for every $\varepsilon < 1$ no cover of $\overline{B}(0, 1)$ with ε-balls can be finite.

4.95. EXAMPLE. Now, with ℓ^∞ as above, consider the subset

$$\mathcal{F} = \{f \in \ell^\infty : 0 \leqslant f(n) \leqslant n^{-1} \text{ for all } n \in \mathbb{N}.\}$$

We show that $\mathcal{F}$ is totally bounded. To see this we fix $0 < \varepsilon < 1$ and construct a finite cover with $\varepsilon/2$-balls. Observe that $|f(n)| \leqslant \varepsilon/2$ for all $n > 2/\varepsilon$ and all $f \in \mathcal{F}$. Now let $\mathcal{G}_\varepsilon$ be the subset of $\mathcal{F}$ consisting of sequences $g \in \mathcal{F}$ with the following properties:

(i) $g(n) = 0$ for $n > 2\varepsilon^{-1}$.
(ii) For $1 \leqslant k \leqslant N_\varepsilon$, $g(k)$ takes only values $m\frac{\varepsilon}{2}$ where $m \in \mathbb{N}_0$ and $0 \leqslant m\varepsilon/2 \leqslant 1$.

Here $N_\varepsilon = 1 + 2\varepsilon^{-1}$. For $g \in \mathcal{G}_\varepsilon$, $1 \leqslant k \leqslant N_\varepsilon$, the number of values that $g(k)$ can take does not exceed N_ε. Thus the cardinality of $\mathcal{G}_\varepsilon$ is bounded by $N_\varepsilon^{N_\varepsilon}$, certainly finite. Also note that for every $f \in \mathcal{F}$ we can find a $g \in \mathcal{G}_\varepsilon$ such that $\sup_{n\in\mathbb{N}} |f(n) - g(n)| \leqslant \varepsilon/2$. Hence the open balls in $\mathcal{F}$ of radius ε centered at some $g \in \mathcal{G}_\varepsilon$ form a finite ε-cover of $\mathcal{F}$. Since $\varepsilon > 0$ was arbitrary, $\mathcal{F}$ is totally bounded.

$\mathcal{F}$ is a closed subset of the closed unit ball in ℓ^∞, and as a closed subset of the complete space ℓ^∞, $\mathcal{F}$ is also complete. We have thus shown that $\mathcal{F}$ is compact.

4.96. Exercise. We are given a sequence $\gamma = (\gamma_n)_{n=1}^\infty$ of nonnegative numbers such that $\lim_{n\to\infty} \gamma_n = 0$. Prove that $\mathcal{F}(\gamma) = \{f \in \ell^\infty(\mathbb{N}) : |f(n)| \leqslant \gamma_n \text{ for all } n \in \mathbb{N}\}$ is totally bounded. Also show that $\mathcal{F}(\gamma)$ is compact.

4.97. Exercise. Let $\mathcal{F}$ be a bounded subset of c_0 (the space of bounded sequences converging to zero). Prove that $\mathcal{F}$ is totally bounded if and only if

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} |f(n)| = 0.$$

4.3.3. Cantor diagonalization. The proof of Theorem 4.92 will reference the Cantor diagonalization process.

4.98. Definition. Let $(x_n)_{n\in\mathbb{N}}$ be a given sequence, set $x_n^{(0)} = x_n$ and let, for all $k \in \mathbb{N}$, $(x_n^{(k)})_{n\in\mathbb{N}}$ be a subsequence of $(x_n^{(k-1)})_{n\in\mathbb{N}}$, i.e $x_n^{(k)} = x_{\kappa_k(n)}^{(k-1)}$, with a strictly increasing function $\kappa_k : \mathbb{N} \to \mathbb{N}$. Let $\kappa(n) = \kappa_1 \circ \cdots \circ \kappa_n(n)$.

The sequence $(x_{\kappa(n)})_{n\in\mathbb{N}}$ is the *Cantor diagonalization sequence* with respect to the subsequences $\{x^{(k)} : k \in \mathbb{N}\}$.

In the following display, each row represents a subsequence of $(x_n) = (x_n^{(0)})$. Each row is a subsequence of the previous row, and the boxed diagonal entries form the Cantor diagonalization sequence.

$$\begin{array}{cccccc}
\boxed{x_1^{(1)}} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} & x_5^{(1)} & \dots \\
x_1^{(2)} & \boxed{x_2^{(2)}} & x_3^{(2)} & x_4^{(2)} & x_5^{(2)} & \dots \\
x_1^{(3)} & x_2^{(3)} & \boxed{x_3^{(3)}} & x_4^{(3)} & x_5^{(3)} & \dots \\
x_1^{(4)} & x_2^{(4)} & x_3^{(4)} & \boxed{x_4^{(4)}} & x_5^{(4)} & \dots \\
x_1^{(5)} & x_2^{(5)} & x_3^{(5)} & x_4^{(5)} & \boxed{x_5^{(5)}} & \dots \\
x_1^{(6)} & x_2^{(6)} & x_3^{(6)} & x_4^{(6)} & x_5^{(6)} & \boxed{\ddots}
\end{array}$$

4.99. Exercise. Given subsequences $x^{(k)} = (x_n^{(k)})_{n\in\mathbb{N}}$, $k = 1, 2, 3, \dots$, as in Definition 4.98:

(i) Show that the corresponding Cantor diagonalization sequence $(x_{\kappa(n)})_{n\in\mathbb{N}}$ satisfies $x_{\kappa(n)} = x_n^{(n)}$ for all $n \in \mathbb{N}$.

(ii) Prove that for every $k \in \mathbb{N}$, the sequence $(x_{\kappa(n+k-1)})_{n\in\mathbb{N}}$ is a subsequence of $x^{(l)}$ for $l \leqslant k$ and of $(x_{\kappa(n)})$.

4.3.4. Equivalence of the characterizations of compactness. We now give the proof of the main theorem on compactness in metric spaces:

Proof of Theorem 4.92. We establish it by showing in turn the implications (1) $\Longrightarrow$ (2), (2) $\Longrightarrow$ (3), (3) $\Longrightarrow$ (2), (2) $\Longrightarrow$ (1).

X compact $\Rightarrow$ X sequentially compact: Suppose that X is compact, but not sequentially compact. Then there exists a sequence $(x_n)_{n\in\mathbb{N}} \subset X$ without a convergent subsequence. Let $A = \{x_n : n \in \mathbb{N}\} \subset X$. Note that A must be an infinite set (otherwise $(x_n)_{n\in\mathbb{N}}$ has a constant subsequence). Since A has no accumulation points, there exists, for every $n \in \mathbb{N}$, an open ball B_n such that $B_n \cap A = \{x_n\}$. Also, A is a closed set, so $X\backslash A$ is open. Thus, $\{B_n : n \in \mathbb{N}\} \cup \{X\backslash A\}$ is an open cover of X. By compactness

of X, it has a finite subcover, which leads to a cover of A by finitely many of the B_n. This implies that A is finite, a contradiction.

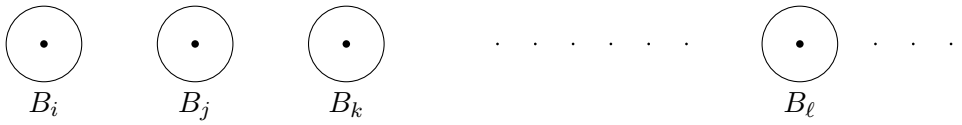


FIGURE 4.3.

X sequentially compact $\Rightarrow$ X totally bounded and complete: Suppose X is sequentially compact. Then it is complete, because every Cauchy sequence that has a convergent subsequence must converge (Lemma 4.24). Suppose that X is not totally bounded. Then there exists $\varepsilon > 0$ such that X cannot be covered by finitely many ε-balls.

Claim: There exists a sequence $p_1, p_2, \dots$ in X such that $d(p_i, p_j) \geqslant \varepsilon$ for all $i \neq j$.

To prove the claim, pick p_1 arbitrarily and then proceed inductively: say that we have constructed $p_1, \dots, p_n$ already. Then there exists p_{n+1} such that $d(p_i, p_{n+1}) \geqslant \varepsilon$ for all $i = 1, \dots, n$ since otherwise we would have $\bigcup_{i=1}^{n} B(p_i, \varepsilon) \supset X$. We have verified the claim.

It remains to observe that the sequence $(p_n)_{n\in\mathbb{N}}$ has no convergent subsequence (no subsequence can be Cauchy). This is a contradiction to the assumption that X is sequentially compact. We conclude that X must be totally bounded.

X totally bounded and complete $\Rightarrow$ X sequentially compact: Assume that X is totally bounded and complete. Let $(x_n)_{n\in\mathbb{N}} \subset X$ be a sequence. We will construct a convergent subsequence. Cover X with finitely many 1-balls. For at least one of them we must have $x_n \in B_0$ for infinitely many indices n. Hence there is a subsequence $(x_n^{(0)})_{n=1}^{\infty} \subset B_0$. That is, we may write $x_n^{(0)} = x_{\kappa_0(n)}$ for a strictly increasing function $\kappa_0 : \mathbb{N} \to \mathbb{N}$.

Next, cover X by finitely many $\frac{1}{2}$-balls. There is at least one of those balls, labeled B_1, that contains $x_n^{(0)}$ for infinitely many indices n. Thus $(x_n^{(0)})$ has a subsequence $(x_n^{(1)})_{n=1}^{\infty}$ contained in B_1. Inductively, we obtain sequences $x^{(k)} = (x_n^{(k)})_{n=1}^{\infty}$ such that $x^{(k)}$ is a subsequence of $x^{(k-1)}$ and all $x_n^{(k)}$, $n \in \mathbb{N}$ are contained in a ball B_k of radius 2^{-k}. Now let $a_n = x_n^{(n)}$ be the Cantor diagonalization subsequence which is a subsequence of (x_n).

We now show that $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. To prove this, let $\varepsilon > 0$ and N be large enough such that $2^{-N+1} < \varepsilon$. Then for $m > n \geqslant N$ we have

$$d(a_m, a_n) \leqslant 2 \cdot 2^{-n} \leqslant 2^{-N+1} < \varepsilon,$$

because $a_n, a_m \in B_n$ and B_n is a ball of radius 2^{-n}. Since ε was arbitrary, $(a_n)_{n\in\mathbb{N}}$ is indeed a Cauchy sequence. By our assumption that X is complete $(a_n)_{n\in\mathbb{N}}$ converges.

X sequentially compact $\Rightarrow$ X compact: Assume that X is sequentially compact. Let $(G_i)_{i\in I}$ be an open cover of X. We first establish the following

Claim: There exists $\varepsilon > 0$ such that every ball of radius ε is contained in one of the G_i.

To prove this, we argue by contradiction, so suppose that no such ε exists. Then for every $n \in \mathbb{N}$ there is a ball B_n of radius $\frac{1}{n}$ that is not contained in any of the G_i. Let p_n be the center of B_n. By sequential compactness, the sequence $(p_n)_{n\in\mathbb{N}}$ has a convergent subsequence $(p_{\kappa(n)})_{n\in\mathbb{N}}$ (here $\kappa : \mathbb{N} \to \mathbb{N}$ strictly increasing) with some limit $p \in X$. Let $i_0 \in I$ be such that $p \in G_{i_0}$. Since G_{i_0} is open there exists $\delta > 0$ such that $B(p, \delta) \subset G_{i_0}$. Let n be large enough such that $d(p_{\kappa(n)}, p) < \delta/2$ and $\frac{1}{\kappa(n)} < \delta/2$. Then $B_{\kappa(n)} \subset B(p, \delta)$ because if $x \in B_{\kappa(n)}$, then

$$d(p, x) \leqslant d(p, p_{\kappa(n)}) + d(p_{\kappa(n)}, x) < \delta/2 + \delta/2 = \delta.$$

Thus, $B_{\kappa(n)} \subset B(p,\delta) \subset G_{i_0}$.

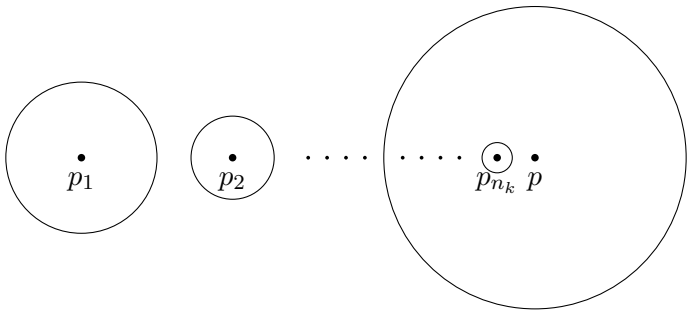


FIGURE 4.4.

This is a contradiction, because we assumed that the B_n are not contained in any of the G_i, and the Claim is proved.

Now let $\varepsilon > 0$ be such that every ε-ball is contained in one of the G_i. We have already proven earlier that X is totally bounded if it is sequentially compact. Thus there exist $p_1, \ldots, p_M$ such that the balls $B(p_j, \varepsilon)$ cover X. But each $B(p_j, \varepsilon)$ is contained in a G_i, say in G_{i_j}, so we have found a finite subcover:

$$X \subset \bigcup_{j=1}^{M} B(p_j, \varepsilon) \subset \bigcup_{j=1}^{M} G_{i_j}. \qquad \square$$

4.100. COROLLARY. *Compact subsets of metric spaces are bounded and closed.*

4.101. COROLLARY. *Let X be a complete metric space and $A \subset X$. Then A is totally bounded if and only if it is relatively compact.*

PROOF. If $\overline{A}$ is compact then it is totally bounded, by Theorem 4.92, and so is A. If A is totally bounded, then $\overline{A}$ is both totally bounded, by Lemma 4.89. Moreover $\overline{A}$ is complete, as a closed subset of a complete space. Hence $\overline{A}$ is compact, by Theorem 4.92. $\square$

4.3.5. Applications of compactness. In this section we discuss the existence of maxima and minima and then as an application prove the algebraic completeness of the complex numbers.

4.102. THEOREM. *Let X be a compact metric space and $f : X \to \mathbb{R}$ a continuous function. Then there exists $x_0, x_1 \in X$ such that*

$$f(x_0) = \sup_{x \in X} f(x), \qquad f(x_1) = \inf_{x \in X} f(x),$$

in other words, f has a maximum at x_0 and a minimum at x_1.

PROOF. By Theorem 4.79, $f(X) \subset \mathbb{R}$ is compact. By the Heine–Borel Theorem (Corollary 4.93) it is therefore closed and bounded. By completeness of the real numbers, $f(X)$ has a finite supremum. Since $f(X)$ is closed we have $\sup f(X) \in f(X)$, and there exists $x_0 \in X$ such that $f(x_0) = \sup f(X) = \sup_{x \in X} f(x)$.

Applying this result to $-f$ and noting that $\sup(-f)(X) = -\inf f(X)$ yields the second statement. $\square$

4.103. COROLLARY. *Let X be a compact metric space. Then every continuous function on X is bounded: $C(X) = C_b(X)$.*

For a converse of this statement, see Exercise 4.183 below.

PROOF. Let $f \in C(X)$. Then $|f| : X \to [0,\infty)$ is also continuous. By Theorem 4.102 there exists $x_0 \in X$ such that $|f(x_0)| = \sup_{x\in X} |f(x)|$. Set $C = |f(x_0)|$. Then $|f(x)| \leqslant C$ for all $x \in X$, so f is bounded. □

As an application of Theorem 4.102 we include a proof of the *algebraic completeness* of the complex numbers, also known as the *fundamental theorem of algebra*. We follow Rudin's [**65**] elegant presentation.

4.104. THEOREM (Fundamental theorem of algebra). *Let $n \geqslant 1$, $a_0, a_1, \ldots, a_n$ be complex numbers, with $a_n \neq 0$ and define for $z \in \mathbb{C}$*

$$P(z) = \sum_{k=0}^{n} a_k z^k.$$

Then there exists $z \in \mathbb{C}$ with $P(z) = 0$.

PROOF. Let $L = \inf_{z\in\mathbb{C}} |P(z)|$. We consider $P(z)$ on the circles of radius R centered at the origin. By the triangle inequality we have the lower bound

$$\begin{aligned} |P(z)| &\geqslant |a_n||z|^n - |a_{n-1}||z|^{n-1} - \cdots - |a_0| \\ &\geqslant |a_n|R^n\Big(1 - \frac{|a_{n-1}|}{|a_n|}R^{-1} - \cdots - \frac{|a_0|}{|a_n|}R^{-n}\Big), \qquad |z| = R \end{aligned}$$

and the quantity on the right side tends to ∞ as $R \to \infty$. It follows that there exists an R_0 such that $|P(z)| > L$ for all z with $|z| > R_0$. The continuous function $z \mapsto |P(z)|$ has a minimum on the compact disk $\overline{B(0,R_0)}$, at a point z_0, and we have $|P(z_0)| = L$. If $L = 0$ we are done.

We assume $L > 0$ (so that $P(z_0) \neq 0$) and derive a contradiction. Let

$$p(z) = \frac{P(z_0 + z)}{P(z_0)}.$$

Then p is a nonconstant complex polynomial of degree $\leqslant n$ with $p(0) = 1$, and we have

$$|p(z)| \geqslant 1, \quad \text{for all } z \in \mathbb{C}.$$

p can be written as

$$p(z) = 1 + \beta_k z^k + \cdots + \beta_n z^n$$

for some $k \in \{1, \ldots, n\}$ with $\beta_k \neq 0$. The key is now to examine $p(z)$ on a ray on which the term $\beta_k z^k$ is real and negative. Observe that for small $|z|$ this term dominates the terms with higher powers of z. Write $\beta_k = \rho_k e^{i\theta_k}$ in polar coordinates and then consider $p(z)$ for $z = re^{i(\pi-\theta_k)/k}$ for small $r > 0$. We then have $\beta_k z^k = e^{i\pi}\rho_k r^k = -\rho_k r^k$. Hence for $\rho_k r^k < 1$

$$|p(re^{i(\pi-\theta_k)/k})| \leqslant 1 - \rho_k r^k + |\beta_{k+1}|r^{k+1} + \cdots + |\beta_n|r^n.$$

We may find $r_0 > 0$ such that $|\beta_{k+1}|r^{k+1} + \cdots + |\beta_n|r^n \leqslant \frac{1}{2}\rho_k r^k$ for $r \in [0, r_0]$, and obtain

$$|p(re^{i(\pi-\theta_k)/k})| \leqslant 1 - \frac{1}{2}\rho_k r^k < 1, \quad 0 < r \leqslant r_0.$$

Since $\inf_{z\in\mathbb{C}} |p(z)| \geqslant 1$ we get a contradiction. □

Multiple applications of Theorem 4.104 lead to

4.105. COROLLARY. *Let P be a nonconstant complex polynomial of degree n, i.e. $P(z) = \sum_{k=0}^{n} a_k z^k$ with $a_n \neq 0$ and $n \geqslant 1$. Then there are $z_1, \ldots, z_n \in \mathbb{C}$ so that*

$$P(z) = a_n \prod_{j=1}^{n} (z - z_j)$$

for all $z \in \mathbb{C}$.

PROOF. We argue by induction. For $n = 1$ the statement is true. Let $n > 1$. By Theorem 4.104 there is $z_1 \in \mathbb{C}$ such that $P(z_1) = 0$. By using (an elementary version of) Theorem 3.63 we expand $P(z)$ in powers of $(z - z_1)$ and because of $P(z_1) = 0$ we get $P(z) = \sum_{k=1}^{n} \frac{P^{(k)}(z_1)}{k!}(z - z_1)^k = (z - z_1)Q(z)$ where $Q(z) = \sum_{j=0}^{n-1} \beta_j z^j$ and the highest coefficient satisfies $\beta_{n-1} = \frac{P^{(n)}(z_1)}{n!} = a_n$. By the induction hypothesis there are $n-1$ complex numbers which we label $z_2, \ldots, z_n$ such that $Q(z) = a_n \prod_{j=2}^{n}(z - z_j)$. Hence $P(z) = a_n \prod_{j=1}^{n}(z - z_j)$. □

4.3.6. Equicontinuity and the Arzelà-Ascoli theorem. Throughout this section let (K, d) be a *compact metric space.* By Corollary 4.103, continuous functions on K are bounded. Thus, $C(K) = C_b(K)$ is a complete metric space with the supremum metric

$$d_\infty(f, g) = \sup_{x \in K} |f(x) - g(x)|$$

and convergence with respect to d_∞ is uniform convergence. In this section we consider the problem of characterizing compact subsets of $C(K)$.

4.106. EXAMPLE. Let $\mathcal{F} = \{f_n : n \in \mathbb{N}\} \subset C([0, 1])$, where

$$f_n(x) = x^n, \quad x \in [0, 1].$$

$\mathcal{F}$ is not compact, because no subsequence of $(f_n)_{n\in\mathbb{N}}$ converges. This is because the pointwise limit

$$f(x) = \begin{cases} 0, & x \in [0, 1), \\ 1, & x = 1. \end{cases}$$

is not continuous, i.e. not in $C([0, 1])$.

The key concept that characterizes compactness in $C(K)$ is equicontinuity.

4.107. DEFINITION (Equicontinuity). A family $\mathcal{F} \subset C(K)$ is called *equicontinuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ for all $f \in \mathcal{F}$, $x, y \in K$ with $d(x, y) < \delta$.

4.108. DEFINITION. A family $\mathcal{F} \subset C(K)$ is called *uniformly bounded* if there exists $C \geqslant 0$ such that $|f(x)| \leqslant C$ for all $x \in K$ and $f \in \mathcal{F}$.

4.109. DEFINITION. $\mathcal{F} \subset C(K)$ is called *pointwise bounded* if for every $x \in K$ there exists $C(x) \geqslant 0$ such that $|f(x)| \leqslant C(x)$ for all $f \in \mathcal{F}$.

Note that $\mathcal{F} \subset C(K)$ is uniformly bounded if and only if it is bounded as a metric space (see Definition 4.57). We have

$$\mathcal{F} \text{ uniformly bounded } \Rightarrow \mathcal{F} \text{ pointwise bounded.}$$

The converse is false in general.

4.110. OBSERVATION. *If $(f_n)_{n\in\mathbb{N}} \subset C(K)$ is uniformly convergent (on K), then $\{f_n : n \in \mathbb{N}\}$ is equicontinuous.*

PROOF. Let $\varepsilon > 0$. By uniform convergence there exists $N \in \mathbb{N}$ such that

$$\sup_{x \in K} |f_n(x) - f_N(x)| \leqslant \varepsilon/3$$

for $n \geqslant N$. By uniform continuity (using Theorem 4.77) there exists $\delta > 0$ such that

$$|f_k(x) - f_k(y)| \leqslant \varepsilon/3$$

for all $x, y \in K$ with $d(x, y) < \delta$ and all $k = 1, \ldots, N$. Thus, for $n \geqslant N$ and $x, y \in K$ with $d(x, y) < \delta$ we have

$$|f_n(x) - f_n(y)| \leqslant |f_n(x) - f_N(x)| + |f_N(x) - f_N(y)| + |f_N(y) - f_n(y)| \leqslant 3 \cdot \varepsilon/3 = \varepsilon. \qquad \square$$

4.111. OBSERVATION. *If $\mathcal{F} \subset C(K)$ is pointwise bounded and equicontinuous, then it is uniformly bounded.*

PROOF. Choose $\delta > 0$ such that

$$|f(x) - f(y)| \leqslant 1$$

for all $x, y \in K$ with $d(x, y) < \delta$ and all $f \in \mathcal{F}$. Since K is totally bounded (by Theorem 4.92) there exist $p_1, \ldots, p_m \in K$ such that the balls $B(p_j, \delta)$ cover K. By pointwise boundedness, for every $x \in K$ there exists $C(x)$ such that $|f(x)| \leqslant C(x)$ for all $f \in \mathcal{F}$. Set

$$C = \max\{C(p_1), \ldots, C(p_m)\}.$$

Then for $f \in \mathcal{F}$ and $x \in K$,

$$|f(x)| \leqslant |f(p_j)| + |f(x) - f(p_j)| \leqslant C + 1,$$

where j is chosen such that $x \in B(p_j, \delta)$. □

The next theorem tells us that a stronger conclusion is true, namely that $\mathcal{F}$ is totally bounded. This is our main theorem in this section; it provides a characterization of the totally bounded subsets of $C(K)$.

4.112. THEOREM (Arzelà–Ascoli). *A subset $\mathcal{F}$ of $C(K)$ is totally bounded if and only if it is pointwise bounded and equicontinuous.*

PROOF OF NECESSITY. We show that if $\mathcal{F}$ is a totally bounded subset of $C(K)$ then $\mathcal{F}$ is pointwise bounded and equicontinuous.

By the definition of $\mathcal{F}$ being totally bounded, there are functions $f_1, \ldots, f_N$ in $\mathcal{F}$ so that for every $f \in \mathcal{F}$ there is an index $i \in \{1, \ldots, N\}$ with $\sup_{x \in K} |f_i(x) - f(x)| < \varepsilon/4$. Note that for every $x \in K$,

$$|f(x)| \leqslant |f_i(x)| + |f(x) - f_i(x)| \leqslant \max_{i=1,\ldots,N} \sup_{x \in K} |f_i(x)| + \varepsilon/4$$

so that $\mathcal{F}$ is pointwise bounded.

Now we show the equicontinuity of the family $\mathcal{F}$. By Theorem 4.77 each f_i is uniformly continuous. Thus for each i there exists a $\delta_i > 0$ such that $|f_i(x) - f_i(x')| < \varepsilon/2$ whenever $d(x, x') < \delta_i$. Let $\delta = \min\{\delta_1, \ldots, \delta_N\}$. Then $\delta > 0$ and we have $|f_i(x) - f_i(x')| < \varepsilon/2$ for every i whenever $d(x, x') < \delta$.

Now pick any $f \in \mathcal{F}$, and let i be so that $\sup_{x \in K} |f_i(x) - f(x)| < \varepsilon/4$, and let x, x' be so that $d(x, x') < \delta$. Then

$$\begin{aligned} |f(x) - f(x')| &\leqslant |f(x) - f_i(x)| + |f_i(x) - f_i(x')| + |f_i(x') - f(x')| \\ &\leqslant \varepsilon/4 + |f_i(x) - f_i(x')| + \varepsilon/4 < \varepsilon\,. \end{aligned}$$

□

PROOF OF SUFFICIENCY. We show that if $\mathcal{F} \subset C(K)$ is equicontinuous and pointwise bounded then $\mathcal{F}$ is totally bounded.

Fix a small $\varepsilon > 0$, and assume that $\varepsilon < 1$. By Lemma 4.88 it suffices to find a finite collection $\mathcal{G}$ of functions in $\mathcal{B}(K)$ (the space of functions bounded on K) such that for every $f \in \mathcal{F}$ there exists a $g \in \mathcal{G}$ with $\sup_{x \in K} |f(x) - g(x)| < \varepsilon$.

By the equicontinuity of $\mathcal{F}$ there is $\delta > 0$ such that $|f(x) - f(x')| < \varepsilon/2$ whenever $d(x, x') < \delta$ and $f \in \mathcal{F}$. Again we use the compactness of K and cover K with finitely many balls $B(x_i, \delta)$, $i = 1, \ldots, L_\varepsilon$. There is M_i so that $|f(x_i)| \leqslant M_i$ for all $f \in \mathcal{F}$. Let $M = 1 + \max_{i=1,\ldots,L_\varepsilon} M_i$.

We now construct a finite partition of K into pairwise disjoint sets A_i, $i = 1, \ldots, L_\varepsilon$ such that $A_i \subset B(x_i, \delta)$. We set $A_1 = B(x_1, \delta)$, and, for $2 \leqslant i \leqslant L_\varepsilon$, $A_i = B(x_i, \delta) \setminus \bigcup_{\nu=1}^{i-1} B(x_\nu, \delta)$. These sets are pairwise disjoint (some possibly empty) and $K = \bigcup_{i=1}^{L_\varepsilon} A_i$. We define $\mathcal{G}$ as the collection of all functions which are constant on each non-empty A_i and which take only values of the form $n\varepsilon/2$ where $n \in \mathbb{N}$ satisfies $-2\varepsilon^{-1}M \leqslant n \leqslant 2\varepsilon^{-1}M$. Observe that $\mathcal{G}$ is finite. More precisely, since every $g \in \mathcal{G}$ is constant on each A_i and the number of possible values on each A_i does not exceed $1 + 4\varepsilon^{-1}M$, the cardinality of $\mathcal{G}$ is bounded by $(1 + 4M\varepsilon^{-1})^{L_\varepsilon}$.

Let $f \in \mathcal{F}$. For each $i \in \{1, \dots, L_\varepsilon\}$,

$$|f(x) - f(x_i)| < \varepsilon/2 \text{ for all } x \in A_i,$$

since A_i is a subset of $B(x_i, \delta)$; the statement is vacuous if A_i is empty. Now $|f(x_i)| \leqslant M$ and therefore there exists an integer n_i with the property that $-2\varepsilon^{-1}M \leqslant n_i \leqslant 2\varepsilon^{-1}M$ and $|f(x_i) - n_i\varepsilon/2| < \varepsilon/2$. Then for $i = 1, \dots, L_\varepsilon$

$$|f(x) - n_i\varepsilon/2| \leqslant |f(x) - f(x_i)| + |f(x_i) - n_i\varepsilon/2| < \varepsilon/2 + \varepsilon/2 = \varepsilon, \text{ for every } x \in A_i.$$

Thus if g is defined by setting $g(x) = n_i\varepsilon/2$ for $x \in A_i$ we have $g \in \mathcal{G}$ and

$$\sup_{x \in K} |f(x) - g(x)| < \varepsilon. \qquad \square$$

As a corollary of the theorem of Arzelà–Ascoli we obtain a characterization of compactness.

4.113. Corollary. *A closed subset $\mathcal{F}$ of $C(K)$ is compact if and only if it is pointwise bounded and equicontinuous.*

Proof. $\mathcal{F}$ is complete, as a closed subspace of the complete space $C(K)$. Thus by the characterization of compactness ($\mathcal{F}$ compact $\iff$ $\mathcal{F}$ totally bounded and complete) the corollary follows from the theorem. $\square$

4.114. Corollary. *An equicontinuous and bounded sequence $\{f_n\}$ of functions in $C(K)$ has a uniformly convergent subsequence.*

Proof. The closure of $\mathcal{F} = \{f_n : n \in \mathbb{N}\}$ is bounded, complete, and equicontinuous, thus compact. By one part of Theorem 4.92 it is also sequentially compact, therefore f_n has a convergent subsequence. $\square$

We now discuss a special case of the Arzelà–Ascoli theorem.

4.115. Corollary. *Let $\mathcal{F}$ be a bounded subset of $C(K)$. Assume that every $f \in \mathcal{F}$ is continuously differentiable and the set of derivatives $\mathcal{F}' = \{f' : f \in \mathcal{F}\}$ is a bounded subset of $C([a, b])$. Then $\mathcal{F}$ is totally bounded.*

Proof of Corollary 4.115. Using the mean value theorem we see that for every $x, y \in [a, b]$ there exists a $\xi \in [a, b]$ such that

$$f(x) - f(y) = f'(\xi)(x - y).$$

Since $\mathcal{F}'$ is bounded there exists $C > 0$ such that $|f'(\tau)| \leqslant C$ for all $f \in \mathcal{F}$, $\tau \in [a, b]$. Thus,

$$|f(x) - f(y)| \leqslant C|x - y|$$

for all $x, y \in [a, b]$ and all $f \in \mathcal{F}$. This yields equicontinuity. Indeed, given $\varepsilon > 0$ we get $|f(x) - f(y)| < \varepsilon$ provided that $|x - y| \leqslant C^{-1}\varepsilon$ (note that $\delta_\varepsilon = C^{-1}\varepsilon$ does not depend on $x, y, f \in \mathcal{F}$). Therefore the claim follows from Theorem 4.112. $\square$

4.116. Example. Let $\mathcal{F} = \{x \mapsto \sum_{n=0}^{\infty} c_n x^n : |c_n| \leqslant 1\} \subset C([-\frac{1}{2}, \frac{1}{2}])$. The set $\mathcal{F}$ is bounded, because

$$\Big| \sum_{n=0}^{\infty} c_n x^n \Big| \leqslant \sum_{n=0}^{\infty} 2^{-n} = 2$$

for all sequences $(c_n)_{n \in \mathbb{N}}$ with $|c_n| \leqslant 1$ and for all $x \in [-1/2, 1/2]$. Moreover,

$$\mathcal{F}' = \Big\{ \sum_{n=1}^{\infty} n c_n x^{n-1} : |c_n| \leqslant 1 \Big\}$$

is also bounded; indeed $\sup_{g \in \mathcal{F}'} \sup_{x \in [-1/2, 1/2]} |g(x)| \leqslant \sum_{n=1}^{\infty} n 2^{-n} < \infty$. Thus, $\mathcal{F} \subset C([-\frac{1}{2}, \frac{1}{2}])$ is relatively compact.

4.117. Example. Let $\mathcal{F} = \{x \mapsto \sum_{n=0}^{\infty} c_n x^n : \sum_{n=0}^{\infty} |c_n| \leqslant 1\} \subset C([-1, 1])$. This set is not relatively compact, because it contains the set in Example 4.106.

4.118. Example. The set

$$\mathcal{F} = \Big\{ x \mapsto \sin(\pi n x) \,:\, n \in \mathbb{Z} \Big\} \subset C([0,1])$$

is bounded, but not relatively compact. To see this, argue by contradiction and assume that $\mathcal{F}$ is relatively compact. Then by Arzelà–Ascoli it is equicontinuous, so there exists $\delta > 0$ such that for all $n \in \mathbb{N}$ and for all $x, y \in [0,1]$ with $|x-y| < \delta$ we have $|\sin(\pi n x) - \sin(\pi n y)| < 1/2$. Set $x = 0$ and $y = 1/(2n)$ for $n > \delta^{-1}/2$. Then $|\sin(\pi n x) - \sin(\pi n y)| = 1$. This is a contradiction.

4.119. Example. Condition (i) from Corollary 4.115 is necessary, because relatively compact sets are bounded. Condition (ii) however is certainly not necessary. Consider $f_n(x) = n^{-1/2}\sin(nx)$ as functions on $[0,1]$ and let $\mathcal{F} = \{f_n : n \in \mathbb{N}\} \subset C([0,1])$. The set $\mathcal{F}$ is bounded, but $\mathcal{F}'$ is unbounded, since $f_n'(0) = n^{1/2}$. But the sequence $(f_n)_{n\in\mathbb{N}}$ is uniformly convergent, so by Observation 4.110, $\mathcal{F}$ is equicontinuous and hence relatively compact.

4.120. Exercise. Let f_n be as in Example 4.119. Prove that there is a constant $C > 0$ such that for all $x, y \in \mathbb{R}$ and all $n \in \mathbb{N}$

$$|f_n(x) - f_n(y)| \leqslant C|x-y|^{1/2}.$$

4.3.7. A characterization of the Riemann integral. In this section we discuss a characterization of Riemann integrable functions by their set of continuity points. Compactness plays a major role here (indeed the Heine-Borel property was an important tool in works by Borel and Lebesgue towards what is now known as the Lebesgue theory of measure and integral). The characterization of the class of Riemann integrable functions involves the notion of a (Lebesgue) null set, which we now discuss.

For a compact interval $I = [c,d]$ we call $d-c$ the length of I, also denoted by $\ell(I)$.

4.121. Definition. A set $E \subset [a,b]$ is called a *Lebesgue null set* if for every $\varepsilon > 0$ there is a sequence $(I_n)_{n\in\mathbb{N}}$ of intervals such that $E \subset \bigcup_{n\in\mathbb{N}} I_n$ and $\sum_{n=1}^{\infty} \ell(I_n) < \varepsilon$.

4.122. Lemma. *Countable unions of Lebesgue null sets are Lebesgue null sets.*

Proof. Let $\varepsilon > 0$. Let E_k, $k \in \mathbb{N}$ be Lebesgue null sets. For each k find a countable family of intervals $\{I_{k,n}\}_{n=1}^{\infty}$ such that $\sum_{n=1}^{\infty} \ell(I_{k,n}) < \varepsilon 2^{-k-1}$ and $E_k \subset \cup_{n=1}^{\infty} I_{k,n}$.

The family of intervals $\{I_{k,n}\}_{(k,n)\in\mathbb{N}^2}$ is countable (and thus the double sum can be arranged in a series) and we have

$$\sum_{k=1}^{\infty}\sum_{n=1}^{\infty} \ell(I_{k,n}) < \sum_{k=1}^{\infty} \varepsilon 2^{-k-1} = \varepsilon/2.$$

□

4.123. Exercise. Which theorems about series with nonnegative terms have been used in this proof?

4.124. Definition. A set $E \subset [a,b]$ has *content zero* if for every $\varepsilon > 0$ there is a finite set of intervals $I_1, \ldots, I_N$ such that $E \subset \bigcup_{n=1}^{N} I_n$ and $\sum_{n=1}^{N} \ell(I_n) < \varepsilon$.

Note that any set of content zero is a Lebesgue null set, but the converse is not true (see Exercise 4.128 below).

4.125. Lemma. *Let $\{I_\nu\}_{\nu=1}^{N}$ be a finite collection of intervals such that $[a,b] \subset \bigcup_{\nu=1}^{N} I_\nu$. Then $\sum_{\nu=1}^{N} \ell(I_\nu) \geqslant b-a$. In particular $[a,b]$ does not have content zero.*

Proof. Let $J_\nu = \overline{I}_\nu \cap [a,b]$. Arrange the finite set formed by all the endpoints of those intervals in increasing order, written as $a \leqslant x_1 \leqslant \cdots < x_M = b$. Then every interval $[x_{i-1}, x_i]$ is contained in at least one J_ν.

Define inductively sets of indices A_ν. For $\nu = 1$ set

$$A_1 = \big\{ i \in \{1, \ldots, M\} : [x_{i-1}, x_i] \subset J_1 \big\}.$$

For any $\nu > 1$ we are either in the situation that $A_1 \cup \cdots \cup A_{\nu-1}$ contains all $i \in \{1, \ldots, M\}$ (then we stop the construction) or, if not, then we form

$$A_\nu = \big\{ i \in \{1, \ldots, M\} : [x_{i-1}, x_i] \subset J_\nu \text{ and } [x_{i-1}, x_i] \subsetneq J_l \text{ for } l \leqslant \nu - 1 \big\}.$$

The construction stops after K steps, where $K \leqslant N$. Note that each index i is in exactly one family A_ν and also for each ν we have $\sum_{i \in A_\nu} (x_i - x_{i-1}) \leqslant \ell(J_\nu)$. Consequently

$$b - a = \sum_{i=1}^{M} (x_i - x_{i-1}) = \sum_{\nu=1}^{K} \sum_{i \in A_\nu} (x_i - x_{i-1}) \leqslant \sum_{\nu=1}^{K} \ell(J_\nu) \leqslant \sum_{\nu=1}^{N} \ell(I_\nu). \qquad \square$$

The Heine–Borel property is crucial for the following lemma.

4.126. Lemma. *Let E be a compact Lebesgue null set. Then E has content zero.*

Proof. Let $\varepsilon > 0$. Since E is a null set there is a countable family $\{I_\nu\}_{\nu \in \mathbb{N}}$ of closed intervals such that $\sum_{\nu=1}^{\infty} \ell(I_\nu) < \varepsilon/2$. Write $I_\nu = [a_\nu, b_\nu]$ and form the slightly larger open intervals $\widetilde{I}_\nu = (a_\nu - \varepsilon 2^{-\nu-2}, b_\nu + \varepsilon 2^{-\nu-2})$ so that $\ell(\widetilde{I}_\nu) = \ell(I_\nu) + \varepsilon 2^{-\nu-1}$ and thus

$$\sum_{\nu=1} \ell(\widetilde{I}_\nu) \leqslant \sum_{\nu=1}^{\infty} \ell(I_\nu) + \sum_{\nu=1}^{\infty} \varepsilon 2^{-\nu-1} < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Since E is compact we may choose finitely many $\widetilde{I}_{\nu_1}, \ldots, \widetilde{I}_{\nu_M}$ such that $E \subset \cup_{l=1}^{M} \widetilde{I}_{\nu_l}$ and $\sum_{l=1}^{M} \ell(\widetilde{I}_{\nu_l}) \leqslant \sum_{\nu=1}^{\infty} \ell(\widetilde{I}_\nu) < \varepsilon$. Hence E has content zero. $\square$

4.127. Corollary. *Let $a < b$. Then $[a, b]$ is not a Lebesgue null set.*

Proof. This is an immediate consequence of Lemma 4.125 together with Lemma 4.126. $\square$

4.128. Exercise. Let E be the set of rational numbers in $[a, b]$. Show that E is a Lebesgue null set but E is not of content zero.

We now formulate the main theorem characterizing Riemann integrable functions.

4.129. Theorem (Lebesgue). *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Then f is Riemann integrable if and only if the set of discontinuities of f,*

$$D_f = \{x \in [a, b] : \ f \text{ is not continuous at } x\},$$

is a null set.

We recall the notion of oscillation of $f : [a, b] \to \mathbb{R}$ at a point x, $\operatorname{osc}_f(x)$ (see Definition 4.49). If $x \in [a, b]$ and $\delta > 0$ we form the expressions

$$M_{f,\delta}(x) = \sup\{f(y) : |x - y| < \delta, \ y \in [a, b]\}$$
$$m_{f,\delta}(x) = \inf\{f(y) : |x - y| < \delta, \ y \in [a, b]\}$$

and observe that

$$\operatorname{osc}_f(x) = \lim_{\delta \to 0+} \big(M_{f,\delta}(x) - m_{f,\delta}(x) \big).$$

The following lemma linking oscillation to lower and upper sums is a very helpful tool in the proof of Theorem 4.129.

4.130. Lemma. *Let $f : [a, b] \to \mathbb{R}$ be a bounded function and assume that $\operatorname{osc}_f(x) < \gamma$ for all $x \in [a, b]$. Then there is a partition P of $[a, b]$ such that $U(f, P) - L(f, P) < \gamma(b - a)$.*

Proof. By definition of $\operatorname{osc}_f(x)$ we can find a $\delta_x > 0$ such that

$$M_{f,2\delta_x}(x) - m_{f,2\delta_x}(x) < \gamma.$$

Since $[a, b]$ is compact we find $x_1, \ldots, x_N$ such that $[a, b]$ is contained in the union of the intervals $(x_i - \delta_{x_i}, x_i + \delta_{x_i})$. Consider the finite set consisting of a and b, the x_i, the points $x_i - \delta_{x_i}$ and the

points $x_i + \delta_{x_i}$, where $i = 1, \dots, N$. Discard those points which do not lie in $[a, b]$. The resulting set P is a partition of $[a, b]$ with nodes $a = t_0 < \cdots < t_M = b$ such that, if t_{i-1}, t_i are consecutive nodes in this partition, then

$$\sup\{f(t) \,:\, t \in [t_{i-1}, t_i]\} - \inf\{f(t) \,:\, t \in [t_{i-1}, t_i]\} < \gamma.$$

Hence

$$U(f, P) - L(f, P) < \gamma \sum_{i=1}^{M} (t_i - t_{i-1}) = \gamma(b - a). \qquad \square$$

Proof of Theorem 4.129.

Part 1: Set of discontinuities is a null set $\implies$ f is Riemann integrable. By Lemma 2.8 it suffices to construct, for given $\varepsilon > 0$, a partition $\mathcal{P}$ such that

$$U(f, \mathcal{P}) - L(f, \mathcal{P}) < \varepsilon. \tag{4.7}$$

The function f is bounded and thus there is $C > 0$ such that $|f(x)| \leqslant C$ for $x \in [a, b]$.

Now let $\varepsilon_1 \ll \varepsilon$ depending on ε; we will see later that

$$\varepsilon_1 = \frac{\varepsilon}{2C + b - a}$$

is an appropriate choice. Consider the set

$$D(\varepsilon_1) = \big\{x \in [a, b] : \operatorname{osc}_f(x) \geqslant \varepsilon_1\big\}.$$

$D(\varepsilon_1)$ is a Lebesgue null set since $D(\varepsilon_1) \subset D_f$ and D_f is a Lebesgue null set. Also, $D(\varepsilon_1)$ is a closed subset of $[a, b]$, and thus compact. Hence $D(\varepsilon_1)$ has *content zero* (by Lemma 4.126). Hence there is a *finite* collection $\{I_\nu\}_{\nu=1}^N$ of closed intervals such that $\sum_{\nu=1}^N \ell(I_\nu) < \varepsilon_1$ and $D(\varepsilon_1) \subset \bigcup_{\nu=1}^N (I_\nu)^\circ$ (where $(I_\nu)^\circ$ denotes the interior of I_ν).

We may choose a partition $P = \{a = x_0 < \cdots < x_N = b\}$ such that each index i belongs to (at least) one of the following sets:

$$\mathcal{J}_1 = \{i : [x_{i-1}, x_i] \subset I_\nu \text{ for some } \nu \text{ in } [1, N]\},$$
$$\mathcal{J}_2 = \{i : [x_{i-1}, x_i] \cap D(\varepsilon_1) = \varnothing\}.$$

Regarding the intervals $[x_{i-1}, x_i]$ with $i \in \mathcal{J}_1$ we have

$$\sum_{i \in \mathcal{J}_1} (x_i - x_{i-1}) \leqslant \sum_{\nu=1}^{N} \sum_{\substack{i: \\ [x_{i-1}, x_i] \subset I_\nu}} (x_i - x_{i-1}) \leqslant \sum_{\nu=1}^{N} \ell(I_\nu) < \varepsilon_1. \tag{4.8}$$

Observe that for all $i \in \mathcal{J}_2$ we have $\operatorname{osc}_f(x) < \varepsilon_1$ for all $x \in [x_{i-1}, x_i]$. Thus by Lemma 4.130, we find a partition P_i of $[x_{i-1}, x_i]$, labeled $\{x_{i,0} = x_{i-1}, \dots, x_{i,N_i} = x_i\}$ such that with

$$U^i(f, P_i) = \sum_{j=1}^{N_i} (x_{i,j} - x_{i,j-1}) \sup_{[x_{i,0}, x_{i,N_i}]} f(x),$$

$$L^i(f, P_i) = \sum_{j=1}^{N_i} (x_{i,j} - x_{i,j-1}) \inf_{[x_{i,0}, x_{i,N_i}]} f(x)$$

we have

$$U^i(f, P_i) - L^i(f, P_i) < \varepsilon_1 (x_i - x_{i-1}). \tag{4.9}$$

Now we shall use the partition $\mathcal{P}$ defined by

$$\mathcal{P} = \{x_i : i \in \mathcal{J}_1\} \cup \bigcup_{i \in \mathcal{J}_2} \{x_{i,0}, \dots, x_{i,N_i}\}.$$

We split $U(f,\mathcal{P})-L(f,\mathcal{P})$ and then estimate

$$\begin{aligned}U(f,\mathcal{P})-L(f,\mathcal{P}) &= \sum_{i\in\mathcal{J}_1}\Big(\sup_{[x_{i-1},x_i]} f-\inf_{[x_{i-1},x_i]} f\Big)(x_i-x_{i-1})+\sum_{i\in\mathcal{J}_2}(U^i(f,P_i)-L^i(f,P_i))\\ &\leqslant 2C\sum_{i\in\mathcal{J}_1}(x_i-x_{i-1})+\sum_{i\in\mathcal{J}_2}(U^i(f,P_i)-L^i(f,P_i)).\end{aligned}$$

By (4.8) and (4.9) we get

$$U(f,\mathcal{P})-L(f,\mathcal{P})<2C\varepsilon_1+\varepsilon_1\sum_{i\in\mathcal{J}_2}(x_i-x_{i-1})\leqslant 2C\varepsilon_1+(b-a)\varepsilon_1.$$

In view of our choice $\varepsilon_1=\varepsilon/(2C+b-a)$ we have proved the desired inequality (4.7).

Part 2: f is Riemann integrable $\implies$ Set of discontinuities is a null set.

For each $n\in\mathbb{N}$ we define $D^n=\{x\in[a,b]:\operatorname{osc}_f(x)\geqslant 1/n\}$. Observe that $D_f=\bigcup_{n=1}^{\infty}D^n$, by Lemma 4.50. Thus by Lemma 4.122 it suffices to show that each D^n is a Lebesgue null set.

Fix $n\in\mathbb{N}$. Since f is Riemann integrable there exists, by Lemma 2.8 a partition $P=\{x_0<\cdots<x_N\}$ of $[a,b]$ such that

$$U(f,P)-L(f,P)<\varepsilon/n.$$

Let $\mathcal{J}$ be the set of indices i for which $I_i=[x_{i-1},x_i]$ contains a point in D^n, and then $D^n\subset\bigcup_{i\in\mathcal{J}}I_i$. Clearly we have

$$M_i(f)-m_i(f)\geqslant\inf_{x\in[x_{i-1},x_i]}\operatorname{osc}_f(x)\geqslant\frac{1}{n}\text{ if }i\in\mathcal{J}.$$

Hence

$$\begin{aligned}\sum_{i\in\mathcal{J}}\ell(I_i)&=\sum_{i\in\mathcal{J}}\frac{(M_i(f)-m_i(f))(x_i-x_{i-1})}{M_i(f)-m_i(f)}\\ &\leqslant\frac{1}{1/n}\sum_{i=1}^{N}(M_i(f)-m_i(f))(x_i-x_{i-1})<n\cdot\frac{\varepsilon}{n}=\varepsilon.\end{aligned}$$

Since $\varepsilon>0$ was arbitrary, we have proved that each D^n is a Lebesgue null set, and thus D_f is a Lebesgue null set. □

Theorem 4.129 gives a very quick proof of the following result (which can also be proved directly from the Riemann integrability criterion using upper and lower sums, see Theorems 2.15, 2.16 and 2.18).

4.131. Theorem.

(i) The set $\mathcal{R}([a,b])$ of real-valued Riemann integrable functions on $[a,b]$ forms an algebra, that is $\mathcal{R}([a,b])$ is a vector space with respect to the usual addition and scalar multiplication, and if f and g are in $\mathcal{R}([a,b])$ then the pointwise product $fg:x\mapsto f(x)g(x)$ is in $\mathcal{R}([a,b])$.

(ii) $\mathcal{R}([a,b])$ is a lattice, i.e. for $f,g\in\mathcal{R}([a,b])$ the functions $x\mapsto\max\{f(x),g(x)\}$, $x\mapsto\min\{f(x),g(x)\}$ belong to $\mathcal{R}([a,b])$.

This is immediate from Theorem 4.129 and the fact that the discontinuity sets of $f+g$, cf, fg, $\max\{f,g\}$, $\min\{f,g\}$ are contained in the union of the discontinuity sets of f and g.

4.132. Exercise. Define $f(0)=0$, and for $0<x\leqslant 1$,

$$f(x)=\begin{cases}x^2 & \text{if } x \text{ is irrational}\\ x^2-q^{-1/4} & \text{if } x \text{ is rational and } x=p/q \text{ in lowest terms.}\end{cases}$$

(a) Determine the set of discontinuities of f on $[0,1]$.

(b) Determine whether f is Riemann integrable on $[0,1]$ and if so determine the value of the integral.

4.133. Exercise. Let $f : [a, b] \to \mathbb{R}$ be an increasing function. Show that the set of discontinuities of f is countable, hence a Lebesgue null set (and thus f is Riemann integrable).

4.134. Exercise. Let $f : [a, b] \to \mathbb{R}$ be a nonnegative Riemann integrable function such that $\int_a^b f = 0$. Show that $\{x : f(x) \neq 0\}$ is a Lebesgue null set.

4.3.8. Covering numbers and Minkowski dimension. We further quantify the concept of total boundedness.

4.135. Definition. Let E be a totally bounded subset of a metric space X, i.e. for every $\delta > 0$ it is contained in a finite collection of δ-balls.

For $\delta > 0$ let $\mathcal{N}(E, \delta)$ be the minimal number of δ-balls needed to cover E (the centers of these balls are not required to belong to E). This number is called the *δ-covering number of E*; note that it depends not only on E but also on the underlying metric space X and the given metric d. The function $\delta \mapsto \log \mathcal{N}(E, \delta)$ is called the *metric entropy function* of E.

The metric entropy function serves as a quantification of total boundedness, and hence, for compact sets, as a *quantitative measure of compactness.* The definition of $\mathcal{N}(E, \delta)$ is extended to sets that are not totally bounded if we allow the value ∞. If E is not totally bounded then there exists a δ_0 such that $\mathcal{N}(E, \delta) = \infty$ for $\delta < \delta_0$.

4.136. Definition. Let E be totally bounded. The number

$$\overline{\dim}_{\mathrm{M}} E = \limsup_{\delta \to 0+} \frac{\log \mathcal{N}(E, \delta)}{\log(\frac{1}{\delta})}$$

is called the *upper Minkowski dimension* or *upper box counting dimension* of E. The expression

$$\underline{\dim}_{\mathrm{M}} E = \liminf_{\delta \to 0+} \frac{\log \mathcal{N}(E, \delta)}{\log(\frac{1}{\delta})}$$

is called *lower Minkowski dimension* or *lower box counting dimension of E*. If $\underline{\dim}_{\mathrm{M}} E = \overline{\dim}_{\mathrm{M}} E = \alpha$ we say that E has *Minkowski dimension* or *box counting dimension* α and write $\dim_{\mathrm{M}} E = \alpha$.

4.137. Example. Let $k \leqslant n$ and let E denote a k-dimensional box in $\mathbb{R}^n$:

$$E = [0, 1]^k \times \{0\}^{n-k} = \{x \in \mathbb{R}^n \,:\, x_j \in [0, 1] \text{ for } 1 \leqslant j \leqslant k,\ x_j = 0 \text{ for } k < j \leqslant n\}.$$

Then there exist constants $c, c' > 0$ such that

$$c' \delta^{-k} \leqslant \mathcal{N}(E, \delta) \leqslant c \delta^{-k} \tag{4.10}$$

for all $\delta \in (0, 1)$. Hence E has Minkowski dimension k.

This notion of dimension coincides in many (but not all) cases with our intuition about dimension. For instance, the set E from Example 4.137 which we referred to as a "k-dimensional box" has box counting dimension k. Note that there is no reason why the box counting dimension of some given set A should always be an integer. In fact, there are lots of compact sets with a non-integer box counting dimension. We refer to such sets as *fractals* (because they have fractional dimension); see Example 4.142 below.

4.138. Exercise. Let $E \subset \mathbb{R}^n$ be a compact set. Show that there exists a constant $c \in (0, \infty)$ such that

$$\mathcal{N}(E, \delta) \leqslant c \cdot \delta^{-n}$$

holds for all $\delta > 0$. Hence $\overline{\dim}_{\mathrm{M}} E \leqslant n$

4.139. Exercise.

(i) Show that if we replace the natural log in the above definitions by another $\log_b$ with base $b > 1$ then the definitions of the dimensions do not change.

(ii) Let $\alpha > 0$. Suppose that for every $\varepsilon > 0$ there is a $\delta(\varepsilon) > 0$ and a positive constant $C_\varepsilon \geqslant 1$ such that $C_\varepsilon^{-1}\delta^{-\alpha+\varepsilon} \leqslant \mathcal{N}(E,\delta) \leqslant C_\varepsilon\delta^{-\alpha-\varepsilon}$ for $0 < \delta < \delta(\varepsilon)$. Show that E has Minkowski dimension α.

(iii) Let $E \subset X$ be totally bounded and let $\overline{E}$ be the closure (which is also totally bounded by Lemma 4.89). Show:

$$\mathcal{N}(E,\delta) \leqslant \mathcal{N}(\overline{E},\delta) \leqslant \mathcal{N}(E,\delta') \text{ if } 0 < \delta' < \delta.$$

(iv) Define $N^{\text{cent}}(E,\delta)$ to be the minimal number of δ-balls with center in E needed to cover E. Show that

$$\mathcal{N}(E,\delta) \leqslant N^{\text{cent}}(E,\delta) \leqslant \mathcal{N}(E,\delta/2).$$

(v) Let $B_1,\dots,B_M$ be balls of radius δ in X, so that each ball has nonempty intersection with the set E. For each $i = 1,\dots,M$ denote by B_i^* the ball with same center as B_i and radius 3δ. Assume that the balls $B_1^*,\dots,B_M^*$ are disjoint. Prove that $M \leqslant \mathcal{N}(E,\delta)$.

Remark. This can be an effective tool to prove lower bounds for the covering numbers.

4.140. Exercise. Consider the following metrics in $\mathbb{R}^n$.

- $d_1(x,y) = \sum_{i=1}^n |x_i - y_i|$,
- $d_2(x,y) = \left(\sum_{i=1}^n |x_i - y_i|^2\right)^{1/2}$,
- $d_\infty(x,y) = \max_{i=1,\dots,n} |x_i - y_i|$.

(i) Let $E \subset \mathbb{R}^n$ and let $\mathcal{N}_1(E,\delta), \mathcal{N}_2(E,\delta), \mathcal{N}_\infty(E,\delta)$ be the metric entropy numbers of E associated with the metrics d_1, d_2, d_∞, respectively. Show that

$$\mathcal{N}_\infty(E,\delta) \leqslant \mathcal{N}_2(E,\delta) \leqslant \mathcal{N}_1(E,\delta) \leqslant \mathcal{N}_2(E,\delta/\sqrt{n}) \leqslant \mathcal{N}_\infty(E,\delta/n).$$

(ii) Let $Q = [0,1]^n$ be the unit cube in $\mathbb{R}^n$. Show that Q has Minkowski dimension n (with respect to any of the metrics d_1, d_2, d_∞).

(iii) Let f be a differentiable function on $[0,1]$ with bounded derivative. Let E be the graph of f over $[0,1]$, i.e. the set of all $x = (x_1,x_2) \in \mathbb{R}^2$ for which $0 \leqslant x_1 \leqslant 1$ and $x_2 = f(x_1)$. What is the Minkowski dimension of E?

(iv) Let E be the set of all $x = (x_1,x_2) \in \mathbb{R}^2$ for which $0 \leqslant x_1 \leqslant 1$ and $x_2 = \sqrt{x_1}$. What is the Minkowski dimension of E?

4.141. Exercise. Let $\beta > 0$. Consider the subset E of $\mathbb{R}$ consisting of the numbers $n^{-\beta}$, for $n = 1,2,\dots$. Show that E has a Minkowski dimension, $\dim_M E_\beta = \frac{1}{\beta+1}$.

4.142. Example. The *Cantor middle third set* $\mathfrak{C}$ is given as the subset of $[0,1]$ consisting of numbers of the form

$$\sum_{k=1}^{\infty} a_k 3^{-k} \text{ where } a_k \in \{0,2\}.$$

It can be written as

$$\mathfrak{C} = [0,1] \setminus \bigcup_{\ell=0}^{\infty} \bigcup_{k=0}^{3^\ell-1} \left(\tfrac{3k+1}{3^{\ell+1}}, \tfrac{3k+2}{3^{\ell+1}}\right). \tag{4.11}$$

$\mathfrak{C}$ is a compact subset of $[0,1]$, with the property that for each N there are 2^N disjoint closed intervals of length 3^{-N} which cover $\mathfrak{C}$.

4.143. Exercise. Show that $\mathfrak{C}$ has Minkowski dimension $\frac{\log 2}{\log 3}$.

4.144. Exercise. Let A be the space of functions $f : \mathbb{N} \to \mathbb{R}$ (aka sequences) so that $|f(n)| \leqslant 2^{-n}$ for all $n \in \mathbb{N}$. It is a subset of the space of bounded sequences with norm $\|f\|_\infty = \sup_{n\in\mathbb{N}} |f(n)|$ and associated metric d_∞. Show that for $\delta < 1/2$ the covering numbers $\mathcal{N}(A,\delta)$ satisfy the bounds

$$\mathcal{N}(A,\delta) \leqslant \left(\frac{1}{\delta}\right)^{C + \frac{1}{2}\log_2 \frac{1}{\delta}}$$

where C is independent of δ.

Also provide a lower bound which shows that A does not have finite lower Minkowski dimension.

4.145. Exercise. For $\gamma = (\gamma_n)_{n\in\mathbb{N}}$, with $\gamma_n \geqslant 0$, let $\mathcal{A}(\gamma)$ be the space of functions $f : \mathbb{N} \to \mathbb{R}$ with $|f(n)| \leqslant \gamma_n$. Find satisfactory estimates for the metric entropy numbers $N(\mathcal{A}(\gamma), \delta)$ with $\delta = 2^{-\ell}$ when (i) $\gamma_n = \frac{1}{n}$, (ii) $\gamma_n = \frac{1}{\log(n+1)}$, (iii) $\gamma_n = 10^{-n}$, (iv) $\gamma_n = 2^{-2^n}$, (v) $\gamma_n = \log\log(e+n)$.

4.146. Exercise.

(i) Let $\mathcal{E}_1 \subset C([0,1])$ be the set of all continuous functions with $\sup_{[0,1]} |f(x)| \leqslant 1$ satisfying

$$|f(x) - f(x')| \leqslant |x - x'| \text{ for all } x, x' \in [0,1].$$

Use the standard sup-norm on $C([0,1])$. Then $\mathcal{E}_1$ is a totally bounded subset of $C([0,1])$. Show that $N(\mathcal{E}_1, \delta) \leqslant 4\delta^{-1}3^{(2/\delta)^n}$.

(ii) Let $\mathcal{E}_s$ be the set of all functions with $\sup_{[0,1]} |f(x)| \leqslant 1$ satisfying

$$|f(x) - f(x')| \leqslant |x - x'|^s \text{ for all } x, x' \in [0,1].$$

What can you say about $N(\mathcal{E}_s, \delta)$?

Minkowski dimension is related to the notion of Lebesgue null sets in the following way.

4.147. Lemma. *Let $A \subset \mathbb{R}$ be a compact set such that $\underline{\dim}_{\mathrm{M}}(A) < 1$. Then A is a Lebesgue null set.*

Proof. Let $\nu = \frac{1}{2}(1 - \underline{\dim}_{\mathrm{M}}(A)) > 0$. By assumption, there exists a sequence $(\delta_m)_{m=1}^{\infty}$ such that $\delta_m \to 0$ and

$$\frac{\log \mathcal{N}(A, \delta_m)}{\log(\delta_m^{-1})} \leqslant \underline{\dim}_{\mathrm{M}}(A) + \nu = 1 - \nu.$$

Hence, for each m there exist open intervals $(B(x_{m,j}, \delta_m))_{j=1,\dots,N_m}$, with $N_m \leqslant \delta_m^{-(1-\nu)}$ such that $A \subset \bigcup_{j=1}^{N_m} B(x_{m,j}, \delta_m)$. Then

$$\sum_{j=1}^{N_m} \delta_m = N_m \delta_m \leqslant \delta_m^{\nu} \longrightarrow 0 \text{ as } m \to \infty. \qquad \square$$

4.148. Exercise. Lemma 4.147 does not have a converse: Show that

$$A = \left\{ \tfrac{1}{\log n} : n \geqslant 3 \right\} \cup \{0\}$$

is a compact Lebesgue null set with $\dim_{\mathrm{M}}(A) = 1$.

4.149. Exercise. Construct a set $E \subset \mathbb{R}$ where $0 < \underline{\dim}_{\mathrm{M}} E < \overline{\dim}_{\mathrm{M}} E < 1$.

4.4. Densely defined functions

Let X, Y be metric spaces, with metrics d_X and d_Y, respectively. Let $D \subset X$ be a *dense* subset of X (i.e. D has the property that every nonempty open set in X contains a point in D). Let

$$F : D \to Y$$

be a continuous function. We would like to extend F to a continuous function defined on all of X. A sufficient condition for the existence will be uniform continuity (see Definition 4.76), and it is important to assume the completeness of Y.

4.150. Theorem. *Let D be dense in X, let Y be complete and let $F : D \to Y$ be a uniformly continuous function. Then there is a unique continuous function $\widetilde{F} : X \to Y$ so that $F(x) = \widetilde{F}(x)$ for $x \in D$; moreover $\widetilde{F}$ is uniformly continuous.*

We first show that any extension is *unique*; here we just need to assume continuity.

4.151. Lemma. *Let $f : X \to Y$, $g : X \to Y$ be continuous. Then $A = \{x \in X : f(x) = g(x)\}$ is closed in X.*

Proof. We need to show that $X\backslash A$ is open. If $b \in X\backslash A$, then $d_Y(f(b), g(b)) =: c > 0$. By the continuity of f and g, there is a neighborhood U of b so that $d_Y(f(x), f(b)) < c/3$ and $d_Y(g(x), g(b)) < c/3$ for all $x \in U$. From the triangle inequality it follows that for $x \in U$

$$d_Y(f(x), g(x)) \geqslant d_Y(f(b), g(b)) - d_Y(f(x), f(b)) - d_Y(g(x), g(b)) \geqslant c - c/3 - c/3 > 0.$$

This shows that b is an interior point of $X\backslash A$. □

4.152. Corollary. *Let D be dense in X and let f and g be continuous functions on D, with values in Y. If $f(x) = g(x)$ on D, then $f(x) = g(x)$ for all $x \in X$.*

Proof. This is immediate from Lemma 4.151, since the set $A = \{x \in X : f(x) = g(x)\}$ is closed and contains D. Hence $A = X$. □

We now prove the *existence* of the extension.

4.153. Lemma. *Let $F : D \to Y$ be continuous, D dense in X. A necessary and sufficient condition for the existence of a continuous function $\widetilde{F} : X \to Y$ which coincides with F on D is that the limits $\lim_{\substack{x \to b \\ x \in D}} F(x)$ exist in Y.*

Proof. The condition about the limits is obviously necessary for the existence of a continuous extension of F. For the sufficiency of this condition we define

$$\widetilde{F}(x) = \lim_{\substack{x \to b \\ x \in D}} F(x).$$

One needs to check that $\widetilde{F}$ is continuous on X. By the existence of the limit there is, for every $\varepsilon > 0$ a $\delta_\varepsilon > 0$ such that $d_Y(F(a), \widetilde{F}(x)) < \varepsilon/2$ for every $a \in D$ with $a \neq x$ and $d_X(a, x) < \delta_\varepsilon$. For an arbitrary x' with $d_X(x, x') < \delta_\varepsilon$ the value $\widetilde{F}(x')$ is given as the limit of $F(a)$ for $a \to x'$, for $a \in D$, $d_X(a, x) < \delta_\varepsilon$. Hence $\widetilde{F}(x')$ is contained in the closure of $F(D \cap B_{\delta_\varepsilon}(x))$, and thus $d_Y(\widetilde{F}(x), \widetilde{F}(x')) \leqslant \varepsilon/2 < \varepsilon$. □

Proof of Theorem 4.150. Uniqueness has been shown in Corollary 4.152. For the existence of a continuous extension we need to check, by Lemma 4.153, that $\lim_{\substack{x \to b \\ x \in D}} F(x)$ exists for all $b \in X$. Let's fix such a b, and let $\{x_n\}$ be a sequence in D with $\lim_{n\to\infty} x_n = b$.

By the uniform continuity of F (on D) there exists for every $\varepsilon > 0$ a $\delta_\varepsilon > 0$ so that $d_Y(F(x), F(x')) < \varepsilon$ whenever $d_X(x, x') < \delta_\varepsilon$ and $x, x' \in D$. Since $x_n \to b$ there is N_ε so that $d_X(x_n, x_m) < \delta_\varepsilon$ for all $m, n > N_\varepsilon$, hence also $d_Y(F(x_n), F(x_m)) < \varepsilon$. This shows that $(F(x_n))_{n=1}^\infty$ is a Cauchy sequence in Y and by the assumption of completeness of Y it has a limit $\widetilde{F}(b)$.

Claim 1: $F(a) \to \widetilde{F}(b)$ as $a \to b$, $a \in D$.

This follows if we show that for arbitrary $\varepsilon > 0$ and $a \in D$ with $d_X(a, b) < \delta_\varepsilon/2$ we have $d_Y(F(a), \widetilde{F}(b)) \leqslant \varepsilon$. Since $x_n \to b$ there is an $N_1 \in \mathbb{N}$ such that $d_X(b, x_n) < \delta_\varepsilon/2$ for all $n \geqslant N_1$. Hence by the triangle inequality $d_X(a, x_n) < \delta_\varepsilon$ for $n \geqslant N_1$, and thus $d_Y(F(a), F(x_n)) < \varepsilon$ for $n \geqslant N_1$. The function $z \mapsto d_Y(F(a), z)$ is continuous on Y and it follows (letting $n \to \infty$) that

$$d_Y(F(a), \widetilde{F}(b)) \leqslant \varepsilon \text{ for } a \in D \text{ and } d_X(a, b) < \delta_\varepsilon.$$

This shows Claim 1.

Claim 2: $\widetilde{F}$ is uniformly continuous.

Let $x, x' \in X$ with $d_X(x, x') < \delta_\varepsilon/3$. Let $\{x_n\}$, $\{x'_n\}$ be sequences in D with $x_n \to x$ and $x'_n \to x'$. Then there is $N_2 \in \mathbb{N}$ such that $d_X(x_n, x) < \delta_\varepsilon/3$ and $d(x'_n, x') < \delta_\varepsilon/3$ for $n > N_2$. Now for $m > n \geqslant N_2$.

$$d_X(x_m, x'_n) \leqslant d_X(x_m, x) + d_X(x, x') + d_X(x', x'_n) < \delta_\varepsilon$$

and thus

$$d_Y(F(x_m), F(x'_n)) < \varepsilon.$$

We now let $m \to \infty$ and $n \to \infty$ and conclude that $d_Y(\widetilde{F}(x), \widetilde{F}(x')) \leqslant \varepsilon$. This proves Claim 2 and finishes the proof of the theorem. □

REMARK. An especially useful version of Theorem 4.150 will be given in the chapter on bounded linear operators (see Theorem 5.22).

4.154. EXAMPLE. The assumption of uniform continuity is needed in Theorem 4.150, as the following example shows. Take $X = \mathbb{R}$, $Y = \mathbb{R}$ and $D = \mathbb{R}\backslash\{0\}$. Define $f(x) = \cos 1/x$, then $f : D \to \mathbb{R}$ is continuous, but not uniformly continuous on D. f does not have a continuous continuation to $\mathbb{R}$: Take $a_n = (2\pi n + \pi/2)^{-1} \in D$, $b_n = (2\pi n)^{-1} \in D$, then $a_n \to 0$, $b_n \to 0$, as $n \to \infty$ and we have $\lim_{n\to\infty} f(a_n) = 0$, $\lim_{n\to\infty} f(b_n) = 1$. For a similar earlier example see also Exercise 1.91.

4.155. EXERCISE. Let D be dense in X, let Y be complete and let $F : (D, d_X) \to (Y, d_Y)$ be an isometry. Prove: There is a unique isometry $\widetilde{F} : (X, d_X) \to (Y, d_Y)$ such that $F(x) = \widetilde{F}(x)$ for $x \in D$ (argue as in the proof of Theorem 4.150).

4.5. Completion

4.156. DEFINITION. A *completion* of a metric space (X, d) is a pair consisting of a complete metric space $(\widetilde{X}, \tilde{d})$ and an isometry $\imath : X \to \widetilde{X}$ such that the image $\imath(X)$ is dense in $\widetilde{X}$.

4.157. EXAMPLE. The real numbers $\mathbb{R}$ can be viewed as a completion of the rational numbers $\mathbb{Q}$ (with the natural choices of Euclidean distance and isometry).

The existence of the real numbers is used in the proof of the following theorem.

4.158. THEOREM. *Every metric space has a completion.*

PROOF. We begin by noticing that given Cauchy sequences $\mathfrak{x} = (x_n)_{n\in\mathbb{N}}$ and $\mathfrak{y} = (y_n)_{n\in\mathbb{N}}$ in X, the sequence $(d(x_n, y_n))_{n\in\mathbb{N}}$ is a Cauchy sequence of real numbers. This is a consequence of (4.1). By the completeness of $\mathbb{R}$ it is convergent in $\mathbb{R}$.

We consider the set of all Cauchy sequences on X, labeled Cauchy(X). We define an equivalence relation on Cauchy(X) by

$$\mathfrak{x} \sim \mathfrak{y} \text{ if } \lim_{n\to\infty} d(x_n, y_n) = 0.$$

Consider for $\mathfrak{x}, \mathfrak{y}$ in Cauchy(X) the functional $(\mathfrak{x}, \mathfrak{y}) \mapsto \lim_{n\to\infty} d(x_n, y_n)$. This functional depends only on the equivalence classes of $\mathfrak{x}$ and $\mathfrak{y}$: Namely if $\mathfrak{x} \sim \mathfrak{x}'$ and $\mathfrak{y} \sim \mathfrak{y}'$ it follows from $|d(x_n, y_n) - d(x'_n, y'_n)| \leqslant d(x_n, x'_n) + d(y_n, y'_n)$ that $\lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(x'_n, y'_n)$. We conclude that the function

$$\tilde{d} : \widetilde{X} \times \widetilde{X} \to \mathbb{R}$$
$$\tilde{d}([\mathfrak{x}], [\mathfrak{y}]) = \lim_{n\to\infty} d(x_n, y_n)$$

is well-defined. We check that $\tilde{d}$ is a metric on $\widetilde{X}$. Note that $\tilde{d}$ is symmetric. Assuming $\tilde{d}([\mathfrak{x}], [\mathfrak{y}]) = 0$ we have $\lim_{n\to\infty} d(x_n, y_n) = 0$, that is $\mathfrak{x} \sim \mathfrak{y}$ and thus $[\mathfrak{x}] = [\mathfrak{y}]$. In order to check the triangle inequality for $\tilde{d}$ let $[\mathfrak{x}]$, $[\mathfrak{y}]$, $[\mathfrak{z}]$ be in $\widetilde{X}$. Using the triangle inequality in X we get for Cauchy-sequences $\mathfrak{x}$, $\mathfrak{y}$, $\mathfrak{z}$,

$$d(x_n, z_n) \leqslant d(x_n, y_n) + d(y_n, z_n).$$

As noted above, the sequences $d(x_n, z_n)$, $d(x_n, y_n)$ and $d(y_n, z_n)$ are convergent as sequences of real numbers and we get

$$\lim_{n\to\infty} d(x_n, z_n) \leqslant \lim_{n\to\infty} d(x_n, y_n) + \lim_{n\to\infty} d(y_n, z_n).$$

Hence

$$\tilde{d}([\mathfrak{x}], [\mathfrak{z}]) \leqslant \tilde{d}([\mathfrak{x}], [\mathfrak{y}]) + \tilde{d}([\mathfrak{y}], [\mathfrak{z}]).$$

We now define an isometry $\imath : X \to \widetilde{X}$ by assigning to $x \in X$ the equivalence class of the constant sequence $(x, x, x, \dots)$. Note that $\tilde{d}(\imath(x), \imath(y)) = d(x, y)$. We show that the image $\imath(X)$ is dense in $\widetilde{X}$. Let $[\mathfrak{x}] \in \widetilde{X}$ where $\mathfrak{x} = (x_n)_{n\in\mathbb{N}} \in \text{Cauchy}(X)$, and let $\varepsilon > 0$. Since $(x_n)_{n\in\mathbb{N}}$ is Cauchy there is $N > 0$ such that $d(x_m, x_n) < \varepsilon/2$ for $m, n \geqslant N$. Then

$$\tilde{d}(\imath(x_N), [\mathfrak{x}]) = \lim_{n\to\infty} d(x_N, x_n) \leqslant \varepsilon/2 < \varepsilon$$

and $\imath(x_N)$ is in the open $\tilde{d}$-ball of radius ε centered at $[\mathfrak{x}]$. Since $\varepsilon > 0$ was arbitrary this shows that $\overline{\imath(X)} = \widetilde{X}$, i.e. $\imath(X)$ is dense in $\widetilde{X}$.

It remains to show that $(\widetilde{X}, \tilde{d})$ is complete. Let $(\tilde{\mathfrak{x}}^{(n)})_{n\in\mathbb{N}}$ be a Cauchy sequence in $\widetilde{X}$ (each $\tilde{\mathfrak{x}}^{(n)}$ is the equivalence class of a Cauchy sequence in X). By the density of $\imath(X)$ in $\widetilde{X}$ there is a $w_n \in X$ such that $\tilde{d}(\imath(w_n), \tilde{\mathfrak{x}}^{(n)}) \leqslant 2^{-n}$ and $(\imath(w_n))_{n\in\mathbb{N}}$ is a Cauchy sequence in $\widetilde{X}$. Notice that $(\tilde{\mathfrak{x}}^{(n)})_{n\in\mathbb{N}}$ converges if and only if $(\imath(w_n))_{n\in\mathbb{N}}$ converges (and then both have the same limit). It remains to show that $(\imath(w_n))_{n\in\mathbb{N}}$ converges. Observe that $\mathfrak{w} = (w_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. We claim that $\imath(w_n)$ converges to $[\mathfrak{w}]$. Since $\mathfrak{w} \in \text{Cauchy}(X)$, for every $\varepsilon > 0$ there is an $N_\varepsilon > 0$ such that $d(w_k, w_l) < \varepsilon/2$ for $k, l > N_\varepsilon$. Now for $m > N_\varepsilon$

$$\tilde{d}(\imath(w_m), [\mathfrak{w}]) = \lim_{n\to\infty} d(w_m, w_n) \leqslant \varepsilon/2 < \varepsilon.$$

Since ε was arbitrary we conclude that $(\imath(w_n))_{n\in\mathbb{N}}$ and therefore $(\tilde{\mathfrak{x}}^{(n)})_{n\in\mathbb{N}}$ converge in $\widetilde{X}$ to $[\mathfrak{w}]$, and we have shown the completeness of $\widetilde{X}$. □

4.159. Exercise (Universal mapping property). Let (X, d_X) be a metric space, let (Y, d_Y) be a complete metric space and let $\imath : X \to Y$ be an isometry. Prove that $\imath(X)$ is dense in Y if and only if for every complete metric space (Z, d_Z), and every isometry $f : (X, d_X) \to (Z, d_Z)$ there is a unique isometry $g : Y \to Z$ such that $f = g \circ \imath$.

4.160. Exercise (Uniqueness of completion up to isometry). The completion is unique up to isometry in the following sense. Prove that if $(\widetilde{X}_1, \tilde{d}_1, \imath_1)$, $(\widetilde{X}_2, \tilde{d}_2, \imath_2)$, are two completions of (X, d) then there is a unique isometry $f : (\widetilde{X}_1, \tilde{d}_1) \to (\widetilde{X}_2, \tilde{d}_2)$ such that $f \circ \imath_1 = \imath_2$.

Hint: Use Exercise 4.159.

4.6. Further exercises

4.161. Exercise. Let X be a metric space with metric d.

(i) Define $\rho : X \times X \to \mathbb{R}$ by

$$\rho(x, y) = \frac{d(x, y)}{1 + d(x, y)}.$$

Prove that ρ is a metric on X.

(ii) Show that a subset U of X is open with respect to the metric d if and only if it is open with respect to the metric ρ.

4.162. Exercise. Let X, Y be metric spaces, let $f_n : X \to Y$ be uniformly continuous on X, for all $n \in \mathbb{N}$. Show that if f_n converges uniformly to f, then f is uniformly continuous.

4.163. Exercise. Let (X, d_X), (Y, d_Y) be metric spaces. Let $f : X \to Y$ be an isometry. Use the metric d_Y on the range $f(X)$. Prove that f is uniformly continuous and injective, and that the inverse $f^{-1} : f(X) \to X$ is also an isometry.

4.164. Exercise. Let $K \in C([a, b]^2)$ and $f \in C([a, b])$. Define

$$G(x) = \int_a^b K(x, t) f(t) dt.$$

Show that $G \in C([a, b])$.

4.165. Exercise. The *torus group* $\mathbb{T}$ is defined as $\mathbb{R}/\mathbb{Z}$, the set of equivalence classes over $\mathbb{R}$ under the equivalence relation $x \sim y$ if and only if $x - y \in \mathbb{Z}$.

(i) Show that $d\big([x],[y]\big) = \min_{n\in\mathbb{Z}} \big| x - y + n\big|$ is a well-defined metric on $\mathbb{T}$.
(ii) Show that the metric space $\mathbb{T}$ is compact.
(iii) Show that the map $\varphi : \mathbb{T} \to S^1$, $\varphi([x]) = e^{2\pi i x}$ is a well-defined homeomorphism.

4.166. Exercise. Let $K \in C([0,1]^2)$ such that $|K(x,t)| \leqslant \frac{3}{4}$ for all $(x,t) \in [0,1]^2$ and let $h \in C([0,1])$. Show that there is a unique $f \in C([0,1])$ such that

$$f(x) = \int_0^1 K(x,t)f(t)dt + h(x) \quad \text{for all } x \in [0,1].$$

4.167. Exercise. Let $A, B > 0$, $h \in C([0,A])$ and $K \in C([0,A]^2)$ such that $|K(x,t)| \leqslant B$ for all $(x,t) \in [0,A]^2$. Consider the *Volterra integral equation*

$$f(x) = \int_0^x K(x,t)f(t)dt + h(x) \quad \text{for all } x \in [0,A] \tag{4.12}$$

(i) Show that there exists a unique $f \in C([0,A])$ such that (4.12) holds.
Hint: Use the metric $d_M(f,g) = \sup_{x\in[0,A]} |f(x) - g(x)|e^{-Mx}$ for a suitable choice of $M > 0$.

(ii) If h is continuous on $[0,\infty)$ and K is continuous on $[0,\infty)^2$ is it still true that there exists a unique $f \in C([0,\infty))$ such that the integral equation (4.12) holds for all $x \in [0,\infty)$?

4.168. Exercise. Let K and h be as in Exercise 4.167 and define functions $f_n \in C([0,A])$ by $f_0 = h$, and $f_{n+1} = V(f_n)$, $n = 0,1,2,\ldots$, where $V(f)(x)$ denotes the right-hand side in (4.12). This exercise examines the speed of convergence of this sequence without explicit reference to the contraction principle.

(i) Prove that

$$|f_{n+1}(x) - f_n(x)| \leqslant \frac{(Bx)^{n+1}}{(n+1)!}, \quad n = 0,1,2,\ldots$$

(ii) Deduce that f_n converges uniformly to $f \in C([0,A])$ satisfying $V[f] = f$ on $[0,A]$ and prove that for $n \geqslant 0$

$$\sup_{0\leqslant x\leqslant A} |f_n(x) - f(x)| \leqslant \frac{e^{BA}(BA)^{n+1}}{(n+1)!}.$$

For large n, how does this compare with (4.4b)?

4.169. Exercise. Let X be a complete metric space, let $K_R = \{x \in X : d(x,x_0) \leqslant R\}$ and $c \in (0,1)$. Let $f : K_R \to X$ be a map satisfying $d(f(x),f(x')) \leqslant cd(x,x')$ for $x,x' \in K_R$ and assume that $d(x_0,f(x_0)) \leqslant R(1-c)$. Show that there is a unique fixed point in K_R.

4.170. Exercise (Regula falsi). Let $I = [x_0 - R, x_0 + R]$, $g \in C^2(I)$ such that $|g'(x)| \geqslant m > 0$, $|g''(x)| \leqslant M$ on I, and assume that $g(x_0 + R)$, $g(x_0 - R)$ have opposite signs. Let $\alpha \in I$ such that $\alpha \neq x_0$ and define

$$F(x) = \begin{cases} \alpha - g(\alpha)\frac{x-\alpha}{g(x)-g(\alpha)}, & \text{if } x \neq \alpha, \\ \alpha - \frac{g(\alpha)}{g'(\alpha)}, & \text{if } x = \alpha. \end{cases}$$

(i) Show that $F \in C^1(I)$, and $\max_{x\in I} |F'(x)| \leqslant \frac{|g(\alpha)|M}{2m^2}$.

(ii) Assume that $c = \frac{|g(\alpha)|M}{2m^2} < 1$ and $|F(x_0) - x_0| \leqslant R(1-c)$. Show that the *Regula falsi* iteration starting at x_0, with

$$x_n = F(x_{n-1}), \quad n \geqslant 1,$$

satisfies $x_n \neq \alpha$ for all $n \in \mathbb{N}$ and converges to the unique zero of g.

4.171. Exercise (Newton's method). Let I be an interval and let f be differentiable twice on I, with f'' being continuous. Suppose

$$|f'(x)| \geqslant m_1 > 0, \qquad |f''(x)| \leqslant M_2 \text{ for all } x \in I.$$

Suppose that $\beta \in I$ is such that $f(\beta) = 0$. For $x \in I$ let

$$\varphi(x) = x - \frac{f(x)}{f'(x)}.$$

Note $\varphi(x) = x$ for $x \in I$ if and only if $f(x) = 0$. The point $(\varphi(x), 0)$ is where the tangent line to the graph at $(x, f(x))$ intersects the x-axis. The fixed point iteration

$$x_{n+1} = \varphi(x_n)$$

can be used to efficiently compute β if the initial guess x_0 is already close to β.

(i) Prove for $x \in I$ the bound

$$|\varphi(x) - \beta| \leqslant \tfrac{M_2}{2m_1}|x - \beta|^2.$$

(ii) Suppose r is sufficiently small so that $[\beta - r, \beta + r] \subset I$ and $r \leqslant \frac{2m_1}{M_2}$. Then $|\varphi(x) - \beta| \leqslant r$ whenever $|x - \beta| \leqslant r$.

(iii) Show that if $r < (\sqrt{3} - 1)\frac{m_1}{M_2}$, then φ is a contraction on $[\beta - r, \beta + r]$.

(iv) Prove: If $|x_0 - \beta| \leqslant r \leqslant \frac{2m_1}{M_2}$, then for $n \in \mathbb{N}$,

$$|x_n - \beta| \leqslant \tfrac{2m_1}{M_2}\big(\tfrac{M_2 r}{2m_1}\big)^{2^n}.$$

(In particular, for small r the speed of convergence is much higher than predicted by the contraction principle.)

4.172. Exercise (Heron's method). Let $a > 0$ and consider the algorithm

$$x_{n+1} = \frac{1}{2}\Big(x_n + \frac{a}{x_n}\Big)$$

with initial guess $x_0 > \sqrt{a}$.

(i) Show that $\sqrt{a} < x_{n+1} < x_n$ for all $n \in \mathbb{N}$ and that $\lim_{n\to\infty} x_n = \sqrt{a}$.

(ii) If $x_0 > \sqrt{a}$ show that for $n \geqslant 1$,

$$0 < x_n - \sqrt{a} \leqslant 2\sqrt{a}\Big(\frac{x_0 - \sqrt{a}}{2\sqrt{a}}\Big)^{2^n}.$$

4.173. Exercise. Consider the polynomial sequence P_n in Example 4.84, with $P_0(x) = 0$, $P_{n+1}(x) = P_n(x) + \frac{1}{2}(x^2 - P_n(x)^2)$ for $n \geqslant 0$. Analyzing this example further, one can avoid the use of Dini's theorem to prove uniform convergence. Show by induction that $|x| - P_n(x) \leqslant |x|(1 - \frac{|x|}{2})^n$. Compute or derive a useful upper bound for $\max_{[0,1]} |x|(1 - \frac{|x|}{2})^n$ to deduce that $\max_{x\in[-1,1]} \big||x| - P_n(x)\big| \to 0$ as $n \to \infty$.

4.174. Exercise. Let (X, d) be a metric space and $A \subset X$ a subset.

(i) Show that A is totally bounded if and only if $\overline{A}$ is totally bounded.

(ii) Assume that X is complete. Show that A is totally bounded if and only if A is relatively compact. Which direction is still always true if X is not complete?

4.175. Exercise. Let ℓ^1 denote the space of all sequences $(f(n))_{n\in\mathbb{N}}$ of complex numbers such that $\sum_{n=1}^{\infty} |f(n)| < \infty$, equipped with the metric $d(f, g) = \sum_{n=1}^{\infty} |f(n) - g(n)|$.

(i) Prove that $\overline{B}(0, 1) = \{f \in \ell^1 : \sum_{n=1}^{\infty} |f(n)| \leqslant 1\}$ is bounded and closed, but not compact.

(ii) Let $b \in \ell^1$ with $b(n) \geqslant 0$ for all $n \in \mathbb{N}$. Show that

$$B = \{f \in \ell^1 : |f(n)| \leqslant b(n) \text{ for all } n \in \mathbb{N}\}$$

is compact.

4.176. Exercise. Let $\mathcal{F}$ be a bounded subset of $\ell^1(\mathbb{N})$. Show that $\mathcal{F}$ is totally bounded if and only if

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \sum_{k=n}^{\infty} |f(k)| = 0.$$

4.177. Exercise. For each of the following subsets of $C([0,1])$ prove or disprove compactness:

(i) $A_1 = \{f \in C([0,1]) : \max_{x\in[0,1]} |f(x)| \leqslant 1\}$.
(ii) $A_2 = A_1 \cap \{p : p \text{ polynomial of degree } \leqslant d\}$ (where $d \in \mathbb{N}$ is given).
(iii) $A_3 = A_1 \cap \{f : f \text{ is a power series with infinite radius of convergence}\}$.

4.178. Exercise. Let $\omega : [0,\infty) \to [0,\infty)$ be such that

$$\lim_{t\to 0+} \omega(t) = \omega(0) = 0.$$

Let $\mathcal{F} \subset C([a,b])$ be a bounded set such that for all $x, y \in [a,b]$, $f \in \mathcal{F}$,

$$|f(x) - f(y)| \leqslant \omega(|x-y|).$$

Show that $\mathcal{F} \subset C([a,b])$ is relatively compact.

4.179. Exercise. Let $C^k([a,b])$ denote the space of k-times continuously differentiable functions on $[a,b]$ endowed with the metric

$$d(f,g) = \sum_{j=0}^{k} \sup_{x\in[a,b]} |f^{(j)}(x) - g^{(j)}(x)|.$$

Let $0 \leqslant \ell < k$ be integers and consider the canonical embedding map

$$\iota : C^k([a,b]) \to C^\ell([a,b]) \text{ with } \iota(f) = f.$$

Prove that if $B \subset C^k([a,b])$ is bounded, then the image $\iota(B) \subset C^\ell([a,b])$ is relatively compact.

4.180. Exercise. Consider the sequence of functions $f_n : [0,1] \to \mathbb{R}$ defined by $f_n(x) = 2nx^2 - n^2x^3$ for $0 \leqslant x \leqslant 1/n$ and $f_n(x) = x$ for $1/n \leqslant x \leqslant 1$. Show that $f_n \in C^1([0,1])$ and that the derivatives f_n' do not form an equicontinuous family of functions. Does f_n have a subsequence that converges in $C^1([0,1])$? Here the norm on $C^1([0,1])$ is as in Exercise 4.179.

4.181. Exercise. Let $C_0(\mathbb{R})$ be the space of continuous functions $f : \mathbb{R} \to \mathbb{C}$ satisfying $\lim_{|x|\to\infty} f(x) = 0$. Set $d_\infty(f,g) = \sup_{x\in\mathbb{R}} |f(x) - g(x)|$. Prove:

(i) Every $f \in C_0(\mathbb{R})$ is uniformly continuous on $\mathbb{R}$.
(ii) $(C_0(\mathbb{R}), d_\infty)$ is a complete metric space.

4.182. Exercise. Let $(C_0(\mathbb{R}), d_\infty)$ be as in Exercise 4.181. Prove: $\mathcal{F} \subset C_0(\mathbb{R})$ is totally bounded if and only if

- $\sup_{f\in\mathcal{F}} \sup_{x\in\mathbb{R}} |f(x)| < \infty$,
- $\lim_{R\to\infty} \sup_{f\in\mathcal{F}} \sup_{|x|>R} |f(x)| = 0$, and
- $\mathcal{F}$ is equicontinuous on $\mathbb{R}$.

4.183. Exercise. Let X be a metric space. Assume that for every continuous function $f : X \to \mathbb{C}$ there exists a constant $C_f > 0$ such that $|f(x)| \leqslant C_f$ for all $x \in X$. Show that X is compact.

4.184. Exercise. Let (X, d_1), (Y, d_2) be metric spaces and choose

$$d((x,y),(x',y')) = \max\{d_1(x,x'), d_2(y,y')\}$$

as a metric in $X \times Y$. Prove:

(i) If $f : X \to Y$ is continuous on X, then the graph

$$\text{graph}(f) = \{(x, f(x)) : x \in X\}$$

is closed as a subset of $X \times Y$.

(ii) If Y is compact and if the graph of $f : X \to Y$ is closed in $X \times Y$, then f is continuous on X.

(iii) Give an example to show that the conclusion in (ii) may fail without the compactness assumption.

4.185. Exercise. Suppose (X, d) is a metric space with a countable dense subset, i.e. a set $A = \{x_1, x_2, \dots\} \subset X$ with $\overline{A} = X$. Let ℓ^∞ denote the metric space of bounded sequences $a = (a_n)_{n\in\mathbb{N}}$ of real numbers with metric $d_\infty(a, b) = \sup_{n\in\mathbb{N}} |a_n - b_n|$. Show that there exists an isometry $\iota : X \to \ell^\infty$ with $d_\infty(\iota(x), \iota(x')) = d(x, x')$ for every $x, x' \in X$ (i.e. (X, d) can be isometrically embedded into (ℓ^∞, d_∞)).

4.186. Exercise. The following is a localized variant of Minkowski's dimension focused on the behavior of the set on small intervals. The *upper Assouad dimension* $\overline{\dim}_{\mathrm{A}} E$ of a totally bounded set E is defined by the infimum over all $a > 0$ for which there exists $c > 0$ such that for all balls B and all positive $\delta < \operatorname{diam}(B)$

$$N(E \cap B, \delta) \leqslant c \left(\frac{\delta}{\operatorname{diam}(B)}\right)^{-a}.$$

(i) Prove that $0 \leqslant \overline{\dim}_{\mathrm{M}} E \leqslant \overline{\dim}_{\mathrm{A}} E$.

(ii) If $E = \mathfrak{C}$ is the Cantor middle third set as in Example 4.142, then $\overline{\dim}_{\mathrm{A}} \mathfrak{C} = \dim_{\mathrm{M}} \mathfrak{C} = \frac{\log 2}{\log 3}$.

(iii) Let $\beta > 0$. Consider the subset E of $\mathbb{R}$ consisting of the numbers $n^{-\beta}$, as in Exercise 4.141. Show that $\overline{\dim}_{\mathrm{M}} E < \overline{\dim}_{\mathrm{A}} E = 1$.

4.187. Exercise. The *upper Assouad spectrum* of dimensions interpolates between the upper Minkowski and the upper Assouad dimension. For a totally bounded set E and $\theta \in [0, 1)$ define $\overline{\dim}_{\mathrm{A},\theta} E$ to be the infimum over all $a > 0$ for which there exists $c > 0$ such that for all balls B and all positive δ with $\operatorname{diam}(B) \geqslant \delta^\theta$:

$$N(E \cap B, \delta) \leqslant c \left(\frac{\delta}{\operatorname{diam}(B)}\right)^{-a}.$$

(i) Prove that the function $\theta \mapsto \overline{\dim}_{\mathrm{A},\theta} E$ is continuous and increasing on $[0, 1)$.

(ii) Prove that $\overline{\dim}_{\mathrm{M}} E = \overline{\dim}_{\mathrm{A},0} E$.

(iii) Define $\overline{\dim}_{\mathrm{qA}} E = \lim_{\theta\to 1-} \overline{\dim}_{\mathrm{A},\theta} E$, the *upper quasi-Assouad dimension of* E. Prove that the limit exists. Show that $\overline{\dim}_{\mathrm{qA}} E \leqslant \overline{\dim}_{\mathrm{A}} E$.

(iv) Let E be totally bounded and $\overline{\dim}_{\mathrm{q}A} E < \infty$. Prove that

$$\overline{\dim}_{\mathrm{M}} E \leqslant \overline{\dim}_{\mathrm{A},\theta} E \leqslant \min\left(\frac{\overline{\dim}_{\mathrm{M}} E}{1-\theta}, \overline{\dim}_{\mathrm{qA}} E\right)$$

4.188. Exercise. Compute the upper Assouad spectrum and the upper Assouad dimension for each of the following sets $E \subset \mathbb{R}$:

(i) $E = \mathfrak{C}$ (the middle third Cantor set as in Example 4.142)

(ii) $E = \{n^{-\beta} : n \in \mathbb{N}\}$, where $\beta > 0$

(iii) $E = \{2^{-\sqrt{n}} : n \in \mathbb{N}\}$

Remark. For more on Assouad dimension and spectra, see the monograph by Fraser [**23**].

CHAPTER 5

Linear operators

5.1. Norms

Let $\mathbb{K}$ denote either $\mathbb{R}$ or $\mathbb{C}$ and let X be a vector space over $\mathbb{K}$.

5.1. Definition. A map $\|\cdot\| : X \to [0,\infty)$ is called a *seminorm* if for all $x, y \in X$ and $\lambda \in \mathbb{K}$,

$$\|\lambda x\| = |\lambda| \cdot \|x\|, \quad \|x+y\| \leqslant \|x\| + \|y\|.$$

A seminorm $\|\cdot\|$ is called a *norm* if

$$\|x\| = 0 \iff x = 0.$$

A $\mathbb{K}$-vector space equipped with a norm is called a *normed vector space.* On every normed vector space we have a natural metric space structure defined by

$$d(x,y) = \|x-y\|.$$

A normed vector space which is also complete as a metric space is called a *Banach space.*

5.2. Exercise. Verify that the following are examples of Banach spaces.

- $\mathbb{K}^n$ with $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$
- $\mathbb{K}^n$ with $\|x\|_1 = \sum_{i=1}^n |x_i|$
- $\mathbb{K}^n$ with the Euclidean norm $\|x\|_2 = (\sum_{i=1}^n |x_i|^2)^{1/2}$
- $C(K)$, where K is a compact metric space with the *supremum norm* $\|f\|_\infty = \sup_{x\in K} |f(x)|$

Remark. The Euclidean norm on $\mathbb{R}^n$ is an example of a norm that can be defined by $\|x\|_2 = \sqrt{\langle x, x\rangle}$ where $\langle x, y\rangle = \sum_{j=1}^n x_j y_j$ is the usual scalar (real) scalar product on $\mathbb{R}^n$. Similarly on $\mathbb{C}^n$ we have $\|x\|_2 = \sqrt{\langle x, x\rangle}$ where $\langle x, y\rangle = \sum_{j=1}^n x_j \overline{y_j}$ is the complex scalar product. For a more systematic treatment of scalar products see §5.6.

5.3. Example. The set of bounded sequences $(a_n)_{n\in\mathbb{N}}$ of complex numbers equipped with the *ℓ^∞-norm,*

$$\|a\|_\infty = \sup_{n\in\mathbb{N}} |a_n|$$

is a Banach space. As a metric space, ℓ^∞ coincides with $C_b(\mathbb{N})$. Occasionally we will work with the subspace of convergent sequences $\mathfrak{c}$, or with the subspace $\mathfrak{c}_0$ of sequences with $\lim_{n\to\infty} a_n = 0$, both with the inherited ℓ^∞ norm. These spaces are closed subspaces of ℓ^∞ and as such, Banach spaces.

5.4. Exercise.
Define $\ell^1 = \{(a_n)_{n\in\mathbb{N}} \subset \mathbb{C} : \sum_{n=1}^\infty |a_n| < \infty\}$. We equip ℓ^1 with the norm defined by

$$\|a\|_1 = \sum_{n=1}^\infty |a_n|.$$

Prove that ℓ^1 is a Banach space.

5.5. Exercise. Prove that the space of continuous functions on $[0,1]$ equipped with the integral-norm (aka L^1-norm) $\|f\|_1 = \int_0^1 |f(x)|dx$ is a normed vector space, but not a Banach space.

5.6. Exercise. Let $0 < s \leqslant 1$.

We say that a function $f : [a,b] \to \mathbb{C}$ belongs to the space $C^{0,s}([a,b])$ of *s-Hölder continuous* functions (or *Lipschitz continuous* if $s = 1$) if

$$|f(x) - f(x')| \leqslant A|x - x'|^s$$

for all $x, x' \in [a,b]$ and some nonnegative constant A. For $f \in C^{0,s}([a,b])$ we set

$$\|f\|_{C^{0,s}} = \sup_{x\in[a,b]} |f(x)| + \sup_{\substack{x,x'\in[a,b]\\ x\neq x'}} \frac{|f(x) - f(x')|}{|x - x'|^s}.$$

Show that this defines a Banach space. Show that the closed unit ball $\{f : \|f\|_{C^{0,s}} \leqslant 1\}$ is a compact subset of $C([a,b])$.

Remark. See Exercise 5.107 for the higher-order analogs $C^{k,s}([a,b])$.

Completeness can be characterized via absolutely convergent series.

5.7. Definition.

(a) We say that $\sum_{n=0}^{\infty} x_n$ converges absolutely if $\sum_{n=0}^{\infty} \|x_n\| < \infty$.
(b) We say that a series $\sum_{n=0}^{\infty} x_n$ converges if the sequence of partial sums $\sum_{k=0}^{n} x_k$ converges in X. That is, there is $x \in X$ such that $\lim_{n\to\infty} \|x - \sum_{k=0}^{n} x_k\| = 0$.

5.8. Theorem. *Let X be a normed space. Then the following statements are equivalent.*

(i) X is complete.
(ii) Every absolutely convergent series in X converges.

Proof. The implication (i) $\Longrightarrow$ (ii) is proved exactly as in the case $X = \mathbb{C}$ (see Theorem 3.7, replace absolute values by norms).

For the implication (ii) $\Longrightarrow$ (i) let $(x_n)_{n=0}^{\infty}$ be a Cauchy sequence. In order to show that (x_n) converges it suffices to show that it has a convergent subsequence (Lemma 4.24). Using the Cauchy sequence condition with $\varepsilon_n = 2^{-n}$ we construct a subsequence $x_{\kappa(n)}$ (with $\kappa(n)$ an increasing sequence of positive integers) such that

$$\|x_{\kappa(n+1)} - x_{\kappa(n)}\| \leqslant 2^{-n}.$$

Then $\sum_{n=2}^{\infty}(x_{\kappa(n)} - x_{\kappa(n-1)})$ converges absolutely, and thus, by assumption, converges to some $y \in X$. This is equivalent to

$$\lim_{N\to\infty}(x_{\kappa(N)} - x_{\kappa(1)}) = \lim_{N\to\infty} \sum_{\ell=2}^{N} (x_{\kappa(\ell)} - x_{\kappa(\ell-1)}) = y.$$

Hence the subsequence $x_{\kappa(N)}$ converges, with limit $y + x_{\kappa(1)}$. □

We can also consider rearrangements of series, defined exactly as in Definition 3.19. We say that a convergent series $\sum_{n=0}^{\infty} x_n$ converges *unconditionally* if all rearrangements converge, with the same sum.

5.9. Exercise. Let X be a Banach space.

(i) Prove that every absolutely convergent series in X converges unconditionally. Follow the proof of Theorem 3.20.

(ii) Show that the converse is no longer true in this generality, i.e. an unconditionally convergent series does not have to be absolutely convergent (and thus the analogue of Theorem 3.24 does not hold in the present generality). As an example consider the space ℓ^∞. Let $e^{(m)}$ be the sequence $(e_n^{(m)})$ with $e_n^{(m)} = 0$ if $n \neq m$ and $e_m^{(m)} = 1$. Let $(c_m)_{m=0}^{\infty}$ a sequence of real numbers with $\lim_{m\to\infty} c_m = 0$, but with $\sum_{m=1}^{\infty} |c_m| = \infty$. Then prove that $\sum_{m=1}^{\infty} c_m e^{(m)}$ is unconditionally convergent in ℓ^∞, but not absolutely convergent.

5.2. Bounded linear operators

5.2.1. Continuity. Let X, Y be normed vector spaces. Recall that a map $T : X \to Y$ is called *linear* if

$$T(\lambda x + \mu y) = \lambda T(x) + \mu T(y)$$

for every $x, y \in X, \lambda, \mu \in \mathbb{K}$. For ease of notation we will adopt the convention that whenever T is a linear map we may write Tx instead of $T(x)$ (unless brackets are necessary because of operator precedence).

5.10. Definition. A linear map $T : X \to Y$ is called *bounded* if there exists $C \geqslant 0$ such that

$$\|Tx\|_Y \leqslant C\|x\|_X$$

for all $x \in X$.

Note that there are two norms involved here: the norm on X and the norm on Y. We may omit the subscript when there is no risk of confusion. Linear maps between normed vector spaces are also referred to as *linear operators.* There is a simple characterization for the continuity of a linear operator:

5.11. Theorem. *Let $T : X \to Y$ be a linear map. The following are equivalent:*

(i) T is bounded.
(ii) T is continuous.
(iii) T is continuous at 0.
(iv) $\sup_{\|x\|_X=1} \|Tx\|_Y < \infty$.

Proof.

(i) $\Rightarrow$ (ii): By assumption and linearity, for $x, y \in X$,

$$\|Tx - Ty\|_Y = \|T(x - y)\|_Y \leqslant C\|x - y\|_X.$$

This implies continuity.

(ii) $\Rightarrow$ (iii): There is nothing to prove.

(iii) $\Rightarrow$ (iv): By continuity at 0 there exists $\delta > 0$ such that for $x \in X$ with $\|x\|_X \leqslant \delta$ we have $\|Tx\|_Y \leqslant 1$. Let $x \in X$ with $\|x\|_X = 1$. Then $\|\delta x\|_X = \delta$, so

$$\|T(\delta x)\|_Y \leqslant 1$$

By linearity of T, $\|Tx\|_Y \leqslant \delta^{-1}$. Thus, $\sup_{\|x\|_X=1} \|Tx\|_Y \leqslant \delta^{-1} < \infty$.

(iv) $\Rightarrow$ (i): Let $C = \sup_{\|x\|_X=1} \|Tx\|_Y < \infty$. Let $x \in X$ with $x \neq 0$. Then $\left\|\frac{x}{\|x\|_X}\right\|_X = 1$, and therefore

$$\left\|T\left(\tfrac{x}{\|x\|_X}\right)\right\|_Y \leqslant C.$$

By linearity of T and the homogeneity of the norm in Y this implies for $x \neq 0$

$$\|Tx\|_Y = \left\|\|x\|_X T\left(\tfrac{x}{\|x\|_X}\right)\right\|_Y = \left\|T\left(\tfrac{x}{\|x\|_X}\right)\right\|_Y \|x\|_X \leqslant C\|x\|_X. \qquad \square$$

5.2.2. The space of bounded linear operators. Let X, Y be normed spaces, with scalar field $\mathbb{K}$ ($= \mathbb{R}$ or $\mathbb{C}$). For bounded linear maps $S, T : X \to Y$ and scalars $\alpha, \beta \in \mathbb{K}$, we can consider the linear combination $\alpha S + \beta T$ whose value at $x \in X$ is $\alpha S(x) + \beta T(x)$. This is a bounded linear transformation. Namely if $\|Sx\|_Y \leqslant C_S\|x\|_X$ and $\|Tx\|_Y \leqslant C_T\|x\|_X$ for all $x \in X$, then $\|\alpha S(x) + \beta T(x)\|_Y \leqslant (|\alpha|C_S + |\beta|C_T)\|x\|_X$ for all $x \in X$. Thus the set of bounded linear maps from X to Y is a vector space over $\mathbb{K}$.

5.12. Definition. By $L(X, Y)$ we denote the space of bounded linear maps $T : X \to Y$. For every $T \in L(X, Y)$ we define its *operator norm* by

$$\|T\|_{\mathrm{op}} = \sup_{x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X}.$$

We also denote $\|T\|_{\mathrm{op}}$ by $\|T\|_{X \to Y}$. In the case where $X = Y$ we also write $L(X)$ for $L(X, X)$.

One should think of $\|T\|_{\mathrm{op}}$ as the best (i.e. smallest) constant $C > 0$ for which

$$\|Tx\|_Y \leqslant C\|x\|_X$$

holds for all $x \in X$. We have by definition that

$$\|Tx\|_Y \leqslant \|T\|_{\mathrm{op}}\|x\|_X.$$

Observe that by linearity of T and homogeneity of the norms in X and Y,

$$\|T\|_{\mathrm{op}} = \sup_{\|x\|_X=1} \|Tx\|_Y = \sup_{\|x\|_X\leqslant 1} \|Tx\|_Y.$$

This also gives $\|cS\|_{\mathrm{op}} = |c|\|S\|_{\mathrm{op}}$ and $\|S+T\|_{\mathrm{op}} \leqslant \|S\|_{\mathrm{op}} + \|T\|_{\mathrm{op}}$; also, $\|T\|_{\mathrm{op}} = 0$ if and only if $Tx = 0$ for all $x \in X$. To summarize:

5.13. Observation. *The operator norm $\|\cdot\|_{\mathrm{op}}$ is a norm on the vector space $L(X,Y)$ of bounded linear operators.*

We shall assume that we use $\|\cdot\|_{\mathrm{op}}$ as a norm on $L(X,Y)$, unless explicitly stated otherwise.

5.14. Example. Let $\mathcal{P}$ be the space of real polynomials, with norm $\|p\|_\infty = \sup\{|p(x)| : x \in [0,1]\}$. Define $D : \mathcal{P} \to \mathcal{P}$ by $Dp = p'$, the differentiation operator. D is a linear operator, but unbounded. Verify that for $f_n(x) = x^n$ we have $\|f_n\|_\infty = 1$, and $\|Df_n\|_\infty = n$.

5.15. Example. Let $A \in \mathbb{R}^{m\times n}$ be a real $m \times n$ matrix. We view A as a linear map $\mathbb{R}^n \to \mathbb{R}^m$: for $x \in \mathbb{R}^n$, $A(x) = A \cdot x \in \mathbb{R}^m$. Let us equip $\mathbb{R}^n$ and $\mathbb{R}^m$ with the corresponding $\|\cdot\|_\infty$ norms. Consider the operator norm $\|A\|_{\infty\to\infty} = \sup_{\|x\|_\infty=1} \|Ax\|_\infty$ with respect to these normed spaces:

$$\|Ax\|_\infty = \max_{i=1,\dots,m} \Big|\sum_{j=1}^n A_{ij}x_j\Big| \leqslant \Big(\max_{i=1,\dots,m} \sum_{j=1}^n |A_{ij}|\Big)\|x\|_\infty.$$

This implies $\|A\|_{\infty\to\infty} \leqslant \max_{i=1,\dots,m}\sum_{j=1}^n |A_{ij}|$. On the other hand, for given $i = 1,\dots,m$ we choose $x \in \mathbb{R}^n$ with $x_j = |A_{ij}|/A_{ij}$ if $A_{ij} \neq 0$ and $x_j = 0$ if $A_{ij} = 0$. Then $\|x\|_\infty \leqslant 1$ and

$$\|A\|_{\infty\to\infty} \geqslant \|Ax\|_\infty = \sum_{j=1}^n |A_{ij}|.$$

Since i was arbitrary, we get $\|A\|_{\infty\to\infty} \geqslant \max_{i=1,\dots,m}\sum_{j=1}^n |A_{ij}|$. Altogether we proved

$$\|A\|_{\infty\to\infty} = \max_{i=1,\dots,m} \sum_{j=1}^n |A_{ij}|.$$

5.16. Exercise. Let $A \in \mathbb{R}^{m\times n}$. For $x \in \mathbb{R}^n$ we define $\|x\|_1 = \sum_{i=1}^n |x_i|$. Show that the value of $\|A\|_{1\to 1} := \sup_{\|x\|_1=1} \|Ax\|_1$ is

$$\|A\|_{1\to 1} = \max_{j=1,\dots,n} \sum_{i=1}^m |A_{ij}|.$$

5.17. Exercise. Let $A \in \mathbb{R}^{m\times n}$. Let $\|A\|_{2\to 2} = \sup_{\|x\|_2=1} \|Ax\|_2$.

(i) Show that

$$\|A\|_{2\to 2} \leqslant \Big(\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2\Big)^{1/2}.$$

Note that if $m = n$ and A is a diagonal matrix, then this bound can be substantially improved for $n \geqslant 2$ (more about this for more general matrices in Example 5.18 and Exercise 5.102).

(ii) Show *Schur's test*

$$\|A\|_{2\to 2} \leqslant \Big(\max_{j=1,\dots,n} \sum_{i=1}^m |A_{ij}|\Big)^{1/2} \Big(\max_{i=1,\dots,m} \sum_{j=1}^n |A_{ij}|\Big)^{1/2}.$$

That is, $\|A\|_{2\to 2} \leqslant \|A\|_{1\to 1}^{1/2}\|A\|_{\infty\to\infty}^{1/2}$ by Example 5.15 and Exercise 5.16.

5.18. Example. For $x \in \mathbb{R}^n$, let $\|x\|_2 = (\sum_{j=1}^n |x_j|^2)^{1/2}$, the Euclidean norm. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix (i.e. $A = A^\intercal$) and let $\|A\|_{2\to 2} = \sup_{\|x\|_2=1} \|Ax\|_2$. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of A. Then

$$\|A\|_{2\to 2} = \max_{i=1,\dots,n} |\lambda_i|.$$

To see this, we first note that if v_k is an eigenvector with eigenvalue λ_k then $\|Av_k\| = |\lambda_k|\|v_k\|$ which shows $\|A\|_{2\to 2} \geqslant \max_{k=1,\dots,n} |\lambda_k|$, for $k = 1, \dots, n$.

To prove the upper bound we first consider a diagonal matrix D with diagonal entries $D_{ii} = \lambda_i$. We can then estimate

$$\|Dx\| = \Big(\sum_{i=1}^n |\lambda_i x_i|^2\Big)^{\frac{1}{2}} \leqslant \max_{k=1,\dots,n} |\lambda_k| \Big(\sum_{i=1}^n |x_i|^2\Big)^{\frac{1}{2}}$$

which shows $\|D\| \leqslant \max_k |\lambda_k|$.

For a general symmetric matrix A we use the spectral theorem according to which $A = U^\intercal DU$ where $Ue_i = v_i$, $Av_i = \lambda_i v_i$, D is a diagonal matrix with $D_{ii} = \lambda_i$, and $U^{-1} = U^\intercal$; the latter implies that U and $U^\intercal$ preserve the Euclidean norm, i.e. $\|Ux\| = \|x\|$, and $\|U^\intercal y\| = \|y\|$ for all x, y. Hence

$$\|Ax\| = \|U^\intercal DUx\| = \|DUx\| \leqslant \max_k |\lambda_k| \|Ux\| \leqslant \max_k |\lambda_k| \|x\|$$

which shows $\|A\| \leqslant \max_k |\lambda_k|$. For the calculation of the norm $\|A\|_{2\to 2}$ for not necessarily symmetric matrices see Exercise 5.102 below.

When the target space Y is complete the space $L(X, Y)$ can also be proved to be complete.

5.19. Theorem. *Let X be a normed vector space and Y a Banach space. Then $L(X, Y)$ is a Banach space (with the operator norm).*

Proof. Let $(T_n)_{n\in\mathbb{N}} \subset L(X, Y)$ be a Cauchy sequence. Then for every $x \in X$, $(T_n x)_{n\in\mathbb{N}} \subset Y$ is Cauchy and by completeness of Y it therefore converges to some limit which we call Tx. This defines a linear operator $T : X \to Y$. We claim that T is bounded. Since $(T_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, it is a bounded sequence. Thus there exists $M > 0$ such that $\|T_n\|_{\mathrm{op}} \leqslant M$ for all $n \in \mathbb{N}$. We have for $x \in X$,

$$\|Tx\|_Y \leqslant \|Tx - T_n x\|_Y + \|T_n x\|_Y \leqslant \|Tx - T_n x\|_Y + M\|x\|_X.$$

Letting $n \to \infty$ we get $\|Tx\|_Y \leqslant M\|x\|_X$. So T is bounded with $\|T\|_{\mathrm{op}} \leqslant M$. It remains to show that $T_n \to T$ in $L(X, Y)$. That is, for all $\varepsilon > 0$ we need to find $N \in \mathbb{N}$ such that

$$\|T_n x - Tx\|_Y \leqslant \varepsilon\|x\|_X$$

for all $n \geqslant N$ and $x \in X$. Since $(T_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, there exists $N \in \mathbb{N}$ such that

$$\|T_n x - T_m x\|_Y \leqslant \tfrac{\varepsilon}{2}\|x\|_X$$

for all $n, m \geqslant N$ and $x \in X$. Fix $x \in X$. Then there exists $m_x \geqslant N$ such that

$$\|T_{m_x} x - Tx\|_Y \leqslant \tfrac{\varepsilon}{2}\|x\|_X.$$

Then if $n \geqslant N$ and $x \in X$,

$$\|T_n x - Tx\|_Y \leqslant \|T_n x - T_{m_x} x\|_Y + \|T_{m_x} x - Tx\|_Y \leqslant \varepsilon\|x\|_X.$$

□

The case $Y = \mathbb{K}$ ($\mathbb{R}$ or $\mathbb{C}$) is of particular interest.

5.20. Definition. Let X be a normed vector space. Elements of $L(X, \mathbb{K})$ are called *bounded linear functionals*. $L(X, \mathbb{K})$ is called the *dual space* of X and denoted X'.

5.21. Corollary. *Dual spaces of normed vector spaces are Banach spaces.*

PROOF. This follows from Theorem 5.19 because $\mathbb{K}$ (which is $\mathbb{R}$ or $\mathbb{C}$) is complete. □

5.2.3. Densely defined operators. In practice a bounded linear operator is often a priori defined only on a dense subspace of X. However, under the assumption of completeness of Y, Theorem 4.150 allows us to extend it to the full space.

5.22. THEOREM. *Let X be a normed space, Y be a Banach space and let V be a dense linear subspace of X. Let $T : V \to Y$ be a linear operator such that*

$$\|Tx\|_Y \leqslant C\|x\|_X \text{ for all } x \in V.$$

Then there is a unique linear operator $\widetilde{T} : X \to Y$ such that $\widetilde{T}x = Tx$ for all $x \in V$ and $\|\widetilde{T}\|_{X\to Y} = \|T\|_{V\to Y}$.

PROOF. Linear bounded operators are uniformly continuous. Thus the existence of a unique extension follows from Theorem 4.150. By the continuity of addition and scalar multiplication it is immediate that the extension is also a linear operator. By the definition of operator norms we have $\|\widetilde{T}\|_{X\to Y} \geqslant \|T\|_{V\to Y}$. By the continuity of the norms we get $\|\widetilde{T}x\|_Y \leqslant \|T\|_{V\to Y}\|x\|_X$ for all $x \in X$ and thus also $\|\widetilde{T}\|_{X\to Y} \leqslant \|T\|_{V\to Y}$. □

REMARK. To avoid notational overloading it is common practice in applications of Theorem 5.22 to keep using the notation T for the extension to X (in place of $\tilde{T}$).

5.3. Finite dimensional spaces

5.3.1. Equivalence of norms.

5.23. DEFINITION. Two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on a vector space X are called *equivalent* if there exist constants $c, C > 0$ such that

$$c\|x\|_a \leqslant \|x\|_b \leqslant C\|x\|_a$$

for all $x \in X$.

5.24. EXERCISE. Let X be a vector space. For two norms write $\|\cdot\|_a \sim \|\cdot\|_b$ if they are equivalent in the sense of Definition 5.23. Show that $\sim$ is an equivalence relation on the set of norms on X.

5.25. EXERCISE. Prove that equivalent norms generate the same topologies: if $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent, then a set $U \subset X$ is open with respect to $\|\cdot\|_a$ if and only if it is open with respect to $\|\cdot\|_b$.

Two given norms on a vector space are generally not equivalent. For example, the supremum norm and the L^2-norm on $C([0,1])$ are not equivalent (as a consequence of Exercise 7.156). However, for finite-dimensional vector spaces we have

5.26. THEOREM. *Let X be a finite-dimensional $\mathbb{K}$-vector space. Then all norms on X are equivalent.*

PROOF. Let $\{b_1, \dots, b_n\}$ be a basis. Then for every $x \in X$ we can write $x = \sum_{i=1}^n x_i b_i$ with uniquely determined coefficients $x_i \in \mathbb{K}$. Moreover the coefficients, as functions on X, are given by $x_i = \xi_i(x)$ where $\xi_i : X \to \mathbb{K}$ are linear functionals. Then

$$\|x\|_* = \max_i |\xi_i(x)|$$

defines a norm on X. Let $\|\cdot\|$ be any norm on X. Since equivalence of norms is an equivalence relation, it suffices to show that $\|\cdot\|_*$ and $\|\cdot\|$ are equivalent. We have

$$\|x\| \leqslant \sum_{i=1}^n |\xi_i(x)|\|b_i\| \leqslant (\max_{i=1,\dots,n} |\xi_i(x)|) \sum_{i=1}^n \|b_i\| = C\|x\|_*,$$

where $C = \sum_{i=1}^n \|b_i\| \in (0, \infty)$. Now define

$$S = \{x \in X \,:\, \|x\|_* = 1\}.$$

We claim that this is a compact set with respect to $\|\cdot\|_*$. Indeed, define the canonical isomorphism $\phi : \mathbb{K}^n \to X$, $(x_1, \dots, x_n) \mapsto \sum_{i=1}^n x_i b_i$, then $x_j = \xi_j(\sum_{i=1}^n x_i b_i)$. Equip $\mathbb{K}^n$ with the Euclidean metric, say, and observe that $\phi : \mathbb{K}^n \to X$ is a continuous map with $S = \phi(K)$, where $K = \{x \in \mathbb{K}^n : \max_i |x_i| = 1\}$. K is compact by the Heine-Borel Theorem (see Corollary 4.93). Thus S is compact by Theorem 4.79. Next note that the function $x \mapsto \|x\|$ is continuous with respect to the $\|\cdot\|_*$ norm. Indeed by the triangle inequality

$$|\|x\| - \|y\|| \leqslant \|x - y\| \leqslant C\|x - y\|_*.$$

Thus by Theorem 4.102, $x \mapsto \|x\|$ attains its infimum on the compact set S and therefore there exists $c > 0$ such that we have the lower bound

$$\|y\| \geqslant c$$

for all $y \in S$. For $x \in X, x \neq 0$ we have $\frac{x}{\|x\|_*} \in S$. Hence using the lower bound with $y = \frac{x}{\|x\|_*}$, and homogeneity of norms, we get

$$\|x\| \geqslant c\|x\|_*.$$

Thus we proved that $\|\cdot\|$ and $\|\cdot\|_*$ are equivalent norms. □

5.27. Corollary.

(i) If X is finite-dimensional then every linear map $T : X \to Y$ is bounded.

(ii) If X and Y are finite-dimensional normed spaces of the same dimension then every bijective linear map $\imath : X \to Y$ is an isomorphism of normed spaces, i.e. both $\imath$ and $\imath^{-1}$ are bounded operators.

Proof. Let $\{x_1, \dots, x_n\} \subset X$ be a basis. Then for $x = \sum_{i=1}^n c_i x_i$ with $c_i \in \mathbb{K}$,

$$\|Tx\|_Y \leqslant \sum_{i=1}^n |c_i|\|Tx_i\|_Y \leqslant C \max_{i=1,\dots,n} |c_i|,$$

where $C = \sum_{i=1}^n \|Tx_i\|_Y$. By equivalence of norms we may assume that $\max_i |c_i|$ is the norm on X. This implies (i), and (ii) follows immediately. □

The corollary implies that we can transfer statements about $\mathbb{R}^n$ or $\mathbb{C}^n$ to finite dimensional normed spaces. For example, we get from Corollary 4.93 a generalization of the Heine–Borel theorem:

5.28. Corollary. *Let X be a finite dimensional normed space. Then the following hold.*

(i) X is a Banach space.

(ii) A subset $A \subset X$ is compact if and only if A is bounded and closed.

5.29. Example. Let X be the set of sequences of complex numbers $(a_n)_{n\in\mathbb{N}}$ such that the condition $\sup_{n\in\mathbb{N}} n|a_n| < \infty$ holds and let Y be the space of bounded complex sequences. Then $X \subset Y$. Equip both spaces with the norm $\|a\| = \sup_{n\in\mathbb{N}} |a_n|$. The map $T : X \to Y$, $(Ta)_{n\in\mathbb{N}} = na_n$ is not bounded: let $e_n^{(k)} = 1$ if $k = n$ and $e_n^{(k)} = 0$ if $k \neq n$. Then $e^{(k)} \in X$ and $Te^{(k)} = ke^{(k)}$ and $\|e^{(k)}\| = 1$. So

$$\|Te^{(k)}\| = k$$

for every $k \in \mathbb{N}$ and therefore $\sup_{\|x\|=1} \|Tx\| = \infty$.

5.30. Exercise. Let $C^1([a, b])$ be the space of (real or complex-valued) continuously differentiable functions. Show that the following two norms are not equivalent on $C^1([a, b])$:

$$\|f\|_{C^0} = \sup_{x\in[a,b]} |f(x)|,$$

$$\|f\|_{C^1} = \sup_{x\in[a,b]} |f(x)| + \sup_{x\in[a,b]} |f'(x)|.$$

5.31. Exercise. Let X be the set of continuously differentiable functions on $[0, 1]$ and let $Y = C([0, 1])$. We consider X and Y as normed vector spaces with the norm $\|f\| = \sup_{x\in[0,1]} |f(x)|$. Define a linear map $T : X \to Y$ by $Tf = f'$. Show that T is not bounded.

5.32. Theorem. *Finite dimensional subspaces of normed spaces are closed.*

Proof. Let $V \subset X$ be a finite dimensional subspace of a normed space. Let $v_1, \dots, v_N$ be a basis of V. For $v = \sum_{k=1}^N \alpha_k v_k \in V$ we define $\|v\|_* = \max\{|\alpha_k| : k = 1, \dots, N\}$. This norm on V is equivalent to $\|\cdot\|_X$ on V. Now let $x^{(n)}$ be a sequence in V which converges to x. We write $x^{(n)} = \sum_{k=1}^N \alpha_k^{(n)} v_k$. Then $(\alpha_k^{(n)})_{n=1}^\infty$ is a Cauchy sequence in $\mathbb{K}$, and thus convergent. Let $\alpha_k = \lim_{n\to\infty} \alpha_k^{(n)}$. Then $x^{(n)}$ converges to $\sum_{k=1}^N \alpha_k v_k \in V$. Hence V is closed. □

5.3.2. Locally compact normed spaces.

5.33. Definition. A normed space is called *locally compact* if the closed unit ball $\{x \in X : \|x\| \leqslant 1\}$ is compact.

In the proof of Theorem 5.32 the crucial property of finite dimensional spaces used is that they are locally compact. The following theorem shows that this property actually characterizes the finite dimensional vector spaces.

5.34. Theorem (F. Riesz). *A normed space X is locally compact if and only if it is finite dimensional.*

Proof. If X is finite dimensional, then X is locally compact, by Corollary 5.28.

Assume that X is locally compact. Let $\overline{B} = \overline{B}(0,1)$. Since $\overline{B}$ is compact we may cover it with a finite collection of open balls of radius $1/2$, centered at $u_1, \dots, u_N \in X$. Define $V = \operatorname{span}\{u_1, \dots, u_N\}$, so that $\dim V \leqslant N$. By Theorem 5.32, V is closed.

We show that $V = X$. Suppose not. Then there is $x \in X \backslash V$ so that $\inf_{y\in V} \|x - y\| = d > 0$. Pick $y \in V$ with $\|x - y\| < 2d$. Let $e = \frac{x-y}{\|x-y\|} \in \overline{B}$. Since the balls $B(u_i, \frac{1}{2})$ cover $\overline{B}$ there is an $i \in \{1, \dots, N\}$ such that $\|e - u_i\| \leqslant \frac{1}{2}$. Then we write

$$x = y + \|x - y\|e = y + \|x - y\|u_i + \|x - y\|(e - u_i)$$

and since $y + \|x - y\|u_i \in V$ and $\|x - y\| < 2d$ we have

$$d \leqslant \|x - (y + \|x - y\|u_i)\| = \|e - u_i\|\|x - y\| \leqslant \tfrac{1}{2}\|x - y\| < d$$

which is the desired contradiction. □

5.4. The uniform boundedness principle

The following theorem is one of the cornerstones of functional analysis.

5.35. Theorem (Uniform boundedness principle). *Let X be a Banach space and Y a normed vector space. Let $\mathcal{G} \subset L(X, Y)$ be a family of bounded linear operators. Then*

$$\sup_{T\in\mathcal{G}} \|Tx\|_Y < \infty \text{ for all } x \in X \quad \Longleftrightarrow \quad \sup_{T\in\mathcal{G}} \|T\|_{\mathrm{op}} < \infty.$$

In other words, a family of bounded linear operators is uniformly bounded if and only if it is pointwise bounded.

This theorem is also known as the *Banach–Steinhaus theorem*. Banach and Steinhaus published a non-constructive proof, with a stronger conclusion, based on the Baire category theorem (see Theorem 8.27), which is the proof given in most textbooks. However the original proofs were constructive, and based on the *gliding hump method*. See the papers by Hahn [**30**] and by Banach [**5**]. Hahn assumed that the operators T_n are linear functionals (i.e. $Y = \mathbb{R}$ or $Y = \mathbb{C}$); this was later generalized. We present the gliding hump method here, following Sokal [**70**].

5.4.1. The gliding hump method. We begin with an auxiliary lemma (which does not assume completeness).

5.36. Lemma. *Let X and Y be normed spaces and let $T : X \to Y$ be a bounded linear operator. Then for all $x_0 \in X$ and all $r > 0$*

$$\|T\|_{X\to Y} \leqslant r^{-1} \sup_{w:\|w-x_0\|\leqslant r} \|Tw\|_Y .$$

Proof. By homogeneity we may assume without loss of generality that $r = 1$. Then it suffices to show that for all $h \in X$ with $\|h\| \leqslant 1$:

$$\|Th\| \leqslant \sup_{w:\|w-x_0\|\leqslant 1} \|Tw\|.$$

This follows by the triangle inequality, because

$$\|Th\| = \|\tfrac{1}{2}T(x_0+h) - \tfrac{1}{2}T(x_0-h)\| \leqslant \tfrac{1}{2}\|T(x_0+h)\| + \tfrac{1}{2}\|T(x_0-h)\| \leqslant \sup_{w:\|w-x_0\|\leqslant 1} \|Tw\|.$$

□

Proof of Theorem 5.35. We argue by contradiction and assume that

$$\sup_{T\in\mathcal{G}} \|T\|_{\mathrm{op}} = \infty.$$

The goal now is to construct an example $x \in X$ for which $\sup_{T\in\mathcal{G}} \|Tx\|_Y = \infty$, which achieves a contradiction. Since $\sup_{T\in\mathcal{G}} \|T\|_{X\to Y} = \infty$ we find a sequence of bounded operators $T_n \in \mathcal{G}$, $n = 1, 2, \ldots$, such that

$$\|T_n\|_{X\to Y} \geqslant 10^n.$$

Pick x_0 arbitrarily and let $\varepsilon \in (0, \frac{1}{4}]$. By the definition of supremum we can choose a sequence x_n such that

$$\|x_n - x_{n-1}\| \leqslant 5^{-n}$$
$$(1-\varepsilon) \sup_{w:\|w-x_{n-1}\|_X\leqslant 5^{-n}} \|T_n w\| \leqslant \|T_n x_n\|.$$

We now apply Lemma 5.36 to the operator T_n, with $r = 5^{-n}$ to get

$$5^n \sup_{w:\|w-x_{n-1}\|_X\leqslant 5^{-n}} \|T_n w\| \geqslant \|T_n\|_{X\to Y}.$$

Hence

$$\|T_n x_n\| \geqslant (1-\varepsilon)5^{-n}\|T_n\|_{X\to Y}.$$

This is the *gliding hump*.

Now the condition $\|x_{n+1} - x_n\| \leqslant 5^{-n-1}$ implies that (x_n) is a Cauchy sequence. Indeed, we get for $m > n$

$$\|x_m - x_n\| \leqslant \sum_{k=1}^{m-n} \|x_{n+k} - x_{n+k-1}\| \leqslant \sum_{k=1}^{\infty} 5^{-n-k} = \tfrac{1}{4}5^{-n}.$$

Since by assumption X is complete we see that the Cauchy sequence converges to a limit x. Letting $m \to \infty$ in the previous display gives

$$\|x - x_n\| \leqslant \tfrac{1}{4}5^{-n}.$$

Finally,

$$\begin{aligned}\|T_n x\| &\geqslant \|T_n x_n\| - \|T_n(x - x_n)\| \geqslant \|T_n x_n\| - \|T_n\|_{X\to Y}\|x - x_n\| \\ &\geqslant (1-\varepsilon)5^{-n}\|T_n\|_{X\to Y} - \|T_n\|_{X\to Y}\tfrac{1}{4}5^{-n} = (1-\varepsilon-\tfrac{1}{4})5^{-n}\|T_n\|_{X\to Y}.\end{aligned}$$

Since we chose $\varepsilon \leqslant 1/4$ and constructed the $T_n \in \mathcal{G}$ such that $\|T_n\|_{X\to Y} \geqslant 10^n$ the preceding inequality gives $\|T_n x\|_Y \geqslant \frac{1}{2}5^{-n}10^n = 2^{n-1}$ so that

$$\sup_{T\in\mathcal{G}} \|Tx\| \geqslant \sup_n \|T_n x\| = \infty.$$

We get a contradiction to the pointwise boundedness of $\mathcal{G}$. □

5.37. Example. If X is not complete, then the conclusion of the theorem may fail. For instance, let X be the space of all sequences $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ such that at most finitely many of the x_n are non-zero. Equip X with the norm $\|x\|_\infty = \sup_{n\in\mathbb{N}} |x_n|$. Define $\lambda_n : X \to \mathbb{R}$ by $\lambda_n(x) = nx_n$. λ_n is a bounded linear map because

$$|\lambda_n(x)| = |nx_n| \leqslant n\|x\|_\infty.$$

For every $x \in X$ there exists $N_x \in \mathbb{N}$ such that $x_n = 0$ for all $n > N_x$. This implies that

$$\sup_{n\in\mathbb{N}} |\lambda_n(x)| = \max\{|\lambda_n(x)| \,:\, n = 1, \dots, N_x\} < \infty.$$

But $\|\lambda_n\|_{\text{op}} \geqslant n$ because $|\lambda_n(e_n)| = n$ (where e_n denotes the sequence such that $e_n(m) = 0$ for every $m \neq n$ and $e_n(n) = 1$). Thus,

$$\sup_{n\in\mathbb{N}} \|\lambda_n\|_{\text{op}} = \infty.$$

As a first application of the uniform boundedness principle we prove that the pointwise limit of a sequence of bounded linear operators on a Banach space must be a bounded linear operator.

5.38. Corollary. *Let X be a Banach space and Y a normed vector space. Suppose $(T_n)_{n\in\mathbb{N}} \subset L(X, Y)$ is such that $(T_n x)_{n\in\mathbb{N}}$ converges to some Tx for every $x \in X$. Then $T \in L(X, Y)$.*

Proof. Linearity of T follows from linearity of limits. It remains to show that T is bounded. Let $x \in X$. Since $(T_n x)_{n\in\mathbb{N}}$ converges, we have $\sup_n \|T_n x\|_Y < \infty$ (convergent sequences are bounded). By the uniform boundedness principle, there exists $C \in (0, \infty)$ such that $\|T_n\|_{\text{op}} \leqslant C$ for every n. Now, for every $x \in X$,

$$\|Tx\|_Y = \lim_{n\to\infty} \|T_n x\|_Y \leqslant C\|x\|_X.$$

Hence $T \in L(X, Y)$ with $\|T\| \leqslant C$. □

Remark. Note that in the context of Corollary 5.38 it does not follow that $T_n \to T$ in $L(X, Y)$. For instance, let $T_n : \ell^1 \to \ell^1$ and $T_n(x) = x_n e_n$. Then $T_n(x) \to 0$ as $n \to \infty$ for every $x \in \ell^1$, but $\|T_n\|_{\text{op}} = 1$ for every $n \in \mathbb{N}$, so T_n does not converge to 0 in $L(X, Y)$.

5.4.2. An application to limiting methods. We are given an infinite matrix $A = (a_{nk})_{n,k=1,2,\dots}$ with the property that $\sum_{k=1}^\infty a_{nk}x_k$ converges whenever $\lim_{k\to\infty} x_k$ exists (this is certainly true if $\sum_{k=1}^\infty |a_{nk}| < \infty$ for all $n \in \mathbb{N}$).

By assumption we may attach to every convergent sequence x_n a new sequence $y = \{y_n\}$ with $y = Ax$, i.e. $y_n = \sum_{k=1}^\infty a_{nk}x_k$ for $n \in \mathbb{N}$.

5.39. Definition. The matrix A defines a *limiting method* if the sequence $y = Ax$ has limit L whenever the sequence x has limit L. In other words $\lim_{n\to\infty}(Ax)_n = \lim_{n\to\infty} x_n$ whenever the latter limit exists.

We are aiming to characterize those infinite matrices A for which A defines a limiting method. The following definition is crucial.

5.40. Definition. The infinite matrix A is called *regular* if the following conditions are satisfied.

(i) $\lim_{n\to\infty} a_{nk} = 0$, for every $k \in \mathbb{N}$.
(ii) $\lim_{n\to\infty} \sum_{k=1}^\infty a_{nk} = 1$
(iii) $\sup_{n\in\mathbb{N}} \sum_{k=1}^\infty |a_{nk}| < \infty$.

Remark. The uniform condition (iii) is stronger than the finiteness condition $\sum_{k=1}^\infty |a_{nk}| < \infty$ for all $n \in \mathbb{N}$, mentioned above.

5.41. Example. The Cesàro means

$$y_n = \frac{x_1 + \cdots + x_n}{n}$$

define a limiting method; here the regular matrix A is defined by $a_{nk} = 1/n$ if $1 \leqslant k \leqslant n$ and $a_{nk} = 0$ for $k > n$.

The main theorem is:

5.42. Theorem (Toeplitz). *An infinite matrix $A = (a_{nk})_{n,k\geqslant 1}$ defines a limiting method if and only if it is regular.*

Proof. The interesting part of the proof is the *necessity*. Assume that A defines a limiting method. Condition (i) is seen to be necessary by testing against the sequence δ^k, with $\delta^k_k = 1$ and $\delta^k_n = 0$ if $n \neq k$. Condition (ii) is seen to be necessary by testing against the sequence $(1, 1, 1, ...)$.

To prove the necessity of (iii) we first prove that $\sum_{k=1}^{\infty} |a_{nk}| < \infty$ for all $n \in \mathbb{N}$. Consider the normed space $\mathfrak{c}$ of convergent sequences with the ℓ^∞ norm, defined in Example 5.3. $\mathfrak{c}$ is a closed subspace of the complete space ℓ^∞ and thus a Banach space. For each $n \in \mathbb{N}$ and each $M \in \mathbb{N}$ define a linear functional $\lambda_{n,M}$ on $\mathfrak{c}$ by

$$\lambda_{n,M}(x) = \sum_{k=1}^{M} a_{nk} x_k.$$

Note that $|\lambda_{n,M}(x)| \leqslant \sum_{j=1}^{M} |a_{nj}| \|x\|_\infty$. The constant $\sum_{j=1}^{M} |a_{nj}|$ is sharp as one can test the inequality of the sequence with $x_k = \operatorname{sign}(a_{nk})$ for $k \leqslant M$ and $x_k = 0$ for $k > M$. Thus

$$\|\lambda_{n,M}\| = \sum_{j=1}^{M} |a_{nj}|.$$

Now fix $x \in \mathfrak{c}$. Then by assumption on A the sum $\sum_{k=1}^{\infty} a_{nk} x_k$ converges for every n. Thus for fixed n we have

$$\sup_M |\lambda_{n,M}(x)| \leqslant C(n, x) < \infty.$$

By the uniform boundedness principle we also get $\sup_M \|\lambda_{n,M}\| \leqslant C(n) < \infty$, i.e. $\sum_k |a_{nk}| < \infty$ for every $n \in \mathbb{N}$.

This means that for every $n \in \mathbb{N}$

$$\lambda_n(x) = \sum_{k=1}^{\infty} a_{nk} x_k$$

defines a bounded linear functional on $\mathfrak{c}$ (indeed on ℓ^∞); moreover by testing as above we see that

$$\|\lambda_n\| = \sup_M \sum_{k=1}^{M} |a_{nk}| = \sum_{k=1}^{\infty} |a_{nk}|.$$

Now if A is a limiting method then the $\lim_{n\to\infty} \lambda_n(x)$ exists for every $x \in \mathfrak{c}$ and this limit is equal to $\lim_{n\to\infty} x_n$. In particular we have

$$\sup_{n\in\mathbb{N}} |\lambda_n(x)| \leqslant C(x) < \infty$$

for every $x \in \mathfrak{c}$. By the uniform boundedness principle we have $\sup_n \|\lambda_n\| < \infty$ and hence $\sup_n \sum_{k=1}^{\infty} |a_{nk}| < \infty$ which is the necessary condition (iii).

Proof of sufficiency. This is a standard argument in calculus, similar to the one for the example of Cesàro means.

We assume that the limiting method is regular and we have to show that for any $x \in \mathfrak{c}$ with $\lim_{n\to\infty} x_n = L$ we also have $\lim_{n\to\infty} \sum_{k=1}^{\infty} a_{nk} x_k = L$.

Thus let $\varepsilon > 0$ and we are going to verify that $|\sum_{k=1}^{\infty} a_{nk}x_k - L| < \varepsilon$ for large n.

Let $M = \sup_n \sum_{k=1}^{\infty} |a_{nk}|$. We split

$$\sum_{k=1}^{\infty} a_{nk}x_k - L = \sum_{k=1}^{\infty} a_{nk}(x_k - L) + L(-1 + \sum_{k=1}^{\infty} a_{nk}).$$

By condition (ii) there is $N_1(\varepsilon)$ so that $|L(-1 + \sum_{k=1}^{\infty} a_{nk})| < \varepsilon/3$ for $n \geqslant N_1(\varepsilon)$.

Since $\lim_{n\to\infty} x_n = L$, there is $N_2(\varepsilon)$ so that $|x_k - L| < \varepsilon/(3M)$ for $k \geqslant N_2(\varepsilon)$. We then have *for all* $n \in \mathbb{N}$

$$\Big|\sum_{k\geqslant N_2(\varepsilon)} a_{nk}(x_k - L)\Big| \leqslant \sum_{k\geqslant N_2(\varepsilon)} |a_{nk}||(x_k - L)| \leqslant \frac{\varepsilon}{3M}\sum_{k=1}^{\infty} |a_{nk}| \leqslant \frac{\varepsilon}{3}.$$

We still have to take care of the terms $\sum_{1\leqslant k\leqslant N_2(\varepsilon)} a_{nk}(x_k - L)$. But for each k we have $a_{nk} \to 0$ as $n \to \infty$ by condition (i). Thus also $\lim_{n\to\infty} \sum_{1\leqslant k\leqslant N_2(\varepsilon)} a_{nk}(x_k - L) = 0$ and there is $N_3(\varepsilon)$ so that $|\sum_{1\leqslant k\leqslant N_2(\varepsilon)} a_{nk}(x_k - L)| < \varepsilon/3$ for $n \geqslant N_3(\varepsilon)$. We can combine the three $\varepsilon/3$-estimates provided that $n \geqslant \max_{i=1,2,3} N_i(\varepsilon)$ and get $|\sum_{k=1}^{\infty} a_{nk}x_k - L| < \varepsilon$ for this range of n. □

5.5. ℓ^p spaces

5.43. Definition. Let $0 < p < \infty$. Define ℓ^p as the set of all sequences $(x_n)_{n=1,2,\dots} \subset \mathbb{C}$ such that $\sum_{n=1}^{\infty} |x_n|^p < \infty$. The *$\ell^p$-(quasi)-norm* is defined as

$$\|x\|_p = \Big(\sum_{n=1}^{\infty} |x_n|^p\Big)^{1/p}.$$

ℓ^∞ is the set of bounded sequences with the norm $\|x\|_\infty = \sup_{n=1,2,\dots} |x_n|$. If $1 \leqslant p \leqslant \infty$ then the number $p' \in [1,\infty]$ such that $\frac{1}{p} + \frac{1}{p'} = 1$ is called the *Hölder dual exponent* of p.

One sees that ℓ^p is a vector space over $\mathbb{C}$. We shall restrict our attention to the case $1 \leqslant p \leqslant \infty$ since the expressions $\|x\|_p$ are not in fact norms when $p < 1$ (see Exercise 5.113). It is immediate from the definition of $\|x\|_p$ that we have the homogeneity identity

$$\|cx\|_p = |c|\|x\|_p$$

for all scalars $c \in \mathbb{C}$ and all $x \in \ell^p$; moreover $\|x\|_p = 0$ if and only if $x = 0$. We will now show that $\|\cdot\|_p$ really is a norm for $p \geqslant 1$. In the proof of the triangle inequality we will use another key inequality.

5.44. Theorem (Hölder's inequality). *Let $p \in [1,\infty]$ and $x \in \ell^p, y \in \ell^{p'}$. Then $\sum_{n=1}^{\infty} x_n y_n$ converges absolutely, and*

$$\sum_{n=1}^{\infty} |x_n y_n| \leqslant \|x\|_p \|y\|_{p'}$$

We get equality if and only if there exist constants α and β with $\alpha\beta \neq 0$ and $\alpha|x_n|^p = \beta|y_n|^{p'}$ for all $n \in \mathbb{N}$.

Remark. The special case $p = 2$ is known as the Cauchy–Schwarz inequality (which holds with equality if and only if x and y are linearly dependent). For a generalization, see §5.6.

In order to prove Hölder's inequality we need a generalization of the arithmetic versus geometric mean inequality for two nonnegative numbers a, b:

$$\sqrt{ab} \leqslant \frac{a+b}{2}.$$

Here we compare the geometric mean of a, b (i.e. the sidelength of the square with the same area as the rectangle with sides a and b) with the arithmetic mean (the number half way between a and b). The generalization is the following:

5.45. LEMMA. *Let $a, b \geqslant 0$.*

(i) *Let $0 < \vartheta < 1$. Then*

$$a^{1-\vartheta} b^{\vartheta} \leqslant (1-\vartheta)a + \vartheta b$$

and the inequality is strict if and only if $a \neq b$.

(ii) *(Young's inequality) Let $p \in (1, \infty)$. Then*

$$ab \leqslant \frac{a^p}{p} + \frac{b^{p'}}{p'}$$

and the inequality is strict if and only if $a^p \neq b^{p'}$.

REMARK. Applying log on both sides, one sees that (i) expresses the concavity of the logarithm. The expression $a^{1-\vartheta} b^{\vartheta}$ is sometimes referred to as a *generalized geometric mean* of a and b.

PROOF. The inequality holds if one of a, b is 0. Also check that if the inequality is true for some a, b then it is also true for ta, tb where $t > 0$.

Assume now $0 < b \leqslant a$ and let $s = b/a$. Then the stated inequality is equivalent to $s^{\vartheta} \leqslant (1-\vartheta) + \vartheta s$ for $0 \leqslant s \leqslant 1$. Set $f(s) = 1 - \vartheta + \vartheta s - s^{\vartheta}$. Then $f(1) = 0$ and $f'(s) < 0$ for $0 < s < 1$, thus $f(s) \geqslant 0$ for $0 \leqslant s \leqslant 1$ which implies the desired inequality. The same argument shows $f(s) > 0$ for $0 \leqslant s < 1$, which implies strict inequality in (i) when $b < a$.

The case $0 < a \leqslant b$ is shown in the same way (in fact follows from the previous case by interchanging a, b and replacing ϑ by $1 - \vartheta$). This proves part (i).

For part (ii) set $x = a^p$, $y = b^{p'}$, $\vartheta = 1 - 1/p$ and observe that the inequality is then equivalent to $x^{1-\vartheta} y^{\vartheta} \leqslant (1-\vartheta)x + \vartheta y$ which holds by part (i). Again we have strict inequality for $x \neq y$, i.e. $a^p \neq b^{p'}$. □

PROOF OF HÖLDER'S INEQUALITY. Observe that the inequality is true if either x or y is 0. Check that if the inequality is true for some choice of x and y then it is also true for sx, ty with $s > 0$, $t > 0$. When $p = 1$ or $p = \infty$ there is nothing to show. So we assume $p \in (1, \infty)$. By Young's inequality, $|x_n y_n| \leqslant |x_n|^p/p + |y_n|^{p'}/p'$ and thus

$$\sum_{n=1}^{\infty} |x_n y_n| \leqslant \frac{1}{p} \sum_{n=1}^{\infty} |x_n|^p + \frac{1}{p'} \sum_{n=1}^{\infty} |y_n|^{p'}.$$

Observe that this yields the asserted inequality when $\|x\|_p = 1$ and $\|y\|_{p'} = 1$; moreover in this special case we have strict inequality if and only if $|x_n|^p = |y_n|^{p'}$ for all $n \in \mathbb{N}$.

Now consider the general case with $\|x\|_p \neq 0$, and $\|y\|_{p'} \neq 0$. Then $\| \frac{x}{\|x\|_p} \|_p = 1$, $\| \frac{y}{\|y\|_{p'}} \|_{p'} = 1$, and since the assertion holds for $x/\|x\|_p$ and $y/\|y\|_{p'}$ we can then deduce the desired inequality for general x and y. Moreover we get equality if and only if $|x_n|^p \|y\|_{p'}^{p'} = |y_n|^{p'} \|x\|_p^p$. □

We now prove the triangle inequality for the ℓ^p norm $\| \cdot \|_p$.

5.46. THEOREM (Minkowski's inequality). *Let $p \in [1, \infty]$. For $x, y \in \ell^p$,*

$$\|x + y\|_p \leqslant \|x\|_p + \|y\|_p.$$

PROOF. If $p = 1$ or $p = \infty$ this follows from the triangle inequality for real or complex numbers. If $\|x + y\|_p = 0$, the inequality is immediate, so we can assume $\|x + y\|_p > 0$. Now we write

$$\|x + y\|_p^p \leqslant \sum_{n=1}^{\infty} |x_n| |x_n + y_n|^{p-1} + \sum_{n=1}^{\infty} |y_n| |x_n + y_n|^{p-1}.$$

Using Hölder's inequality we estimate

$$\sum_{n=1}^{\infty} |x_n||x_n+y_n|^{p-1} \leqslant \Big(\sum_{n=1}^{\infty} |x_n|^p\Big)^{1/p}\Big(\sum_{n=1}^{\infty} |x_n+y_n|^{(p-1)p'}\Big)^{1/p'} = \|x\|_p\|x+y\|_p^{p-1};$$

here we have used $p'(p-1) = \frac{p}{p-1}(p-1) = p$. Similarly we estimate the second term. We then obtain

$$\|x+y\|_p^p \leqslant \big(\|x\|_p + \|y\|_p\big)\|x+y\|_p^{p-1}.$$

and dividing by $\|x+y\|_p^{p-1}$ gives the claim. □

5.47. Corollary. *The class ℓ^p is a normed vector space, with norm $\|\cdot\|_p$.*

We now give a useful result relating the various ℓ^p spaces; they are increasing in p:

5.48. Theorem. *Let $0 < p_1 \leqslant p_2 \leqslant \infty$. Then $\ell^{p_1} \subset \ell^{p_2}$ and*

$$\|x\|_{p_2} \leqslant \|x\|_{p_1} \tag{5.1}$$

for all $x \in \ell^{p_1}$.

Proof. We first assume that $\|x\|_{p_1} = 1$. This implies that $|x_j| \leqslant 1$ for all $j \in \mathbb{N}$, and thus $\|x\|_\infty \leqslant 1$. Moreover if $p_1 \leqslant p_2 < \infty$ we also get $|x_j|^{p_2} \leqslant |x_j|^{p_1}$ for all $j \in \mathbb{N}$. Summing in j leads to $\|x\|_{p_2} \leqslant 1$.

Now for general $x \in \ell^{p_1}$ with $x \neq 0$ consider $y = x/\|x\|_{p_1}$ which satisfies $\|y\|_{p_1} = 1$. We have just shown that this implies $\|y\|_{p_2} \leqslant 1$. This means $\|x/\|x\|_{p_1}\|_{p_2} \leqslant 1$ and by homogeneity the inequality $\|x\|_{p_2} \leqslant \|x\|_{p_1}$ follows. □

5.49. Exercise. Show the strict inclusion $\ell^p \subsetneq \ell^q$ if $0 \leqslant p < q \leqslant \infty$.

5.50. Exercise. Let $1 < p \leqslant \infty$. Modify the example in Exercise 5.9 to construct a series in ℓ^p which converges unconditionally but not absolutely.

Remark. Using a deeper result of Dvoretzky and Rogers (1950) one can construct in every infinite-dimensional Banach space a series which converges unconditionally but not absolutely. We refer to the monograph of Albiac and Kalton [**1**] for more information.

We now study the bounded linear functionals on the normed space ℓ^p, for $1 \leqslant p < \infty$.

5.51. Theorem. *Let $1 \leqslant p < \infty$. For each $\Lambda \in (\ell^p)'$ there exists $y \in \ell^{p'}$ such that*

$$\Lambda(x) = \sum_{k=1}^{\infty} x_k y_k$$

for all $x \in \ell^p$ and $\|\Lambda\|_{(\ell^p)'} = \|y\|_{\ell^{p'}}$. That is, the dual space $(\ell^p)'$ is isometrically isomorphic to $\ell^{p'}$.

Proof. We write $\|v\|_{op} = \sup_{\|x\|_p \leqslant 1} |v(x)|$ for the norm on $(\ell^p)'$.

The case $p = 1$. If $x \in \ell^1$ then for every $y \in \ell^\infty$ there is a bounded linear functional $\psi^y : \ell^1 \to \mathbb{C}$, defined by

$$\psi^y(x) = \sum_{k=1}^{\infty} x_k y_k.$$

Indeed, the sum converges absolutely and we have

$$|\psi^y(x)| \leqslant \sum_{k=1}^{\infty} |x_k y_k| \leqslant \|x\|_1\|y\|_\infty.$$

The map $y \to \psi^y$ is linear and injective, and the above shows $\|\psi^y\|_{op} \leqslant \|y\|_\infty$. By e_k we denote the sequence which is 1 at position k and 0 everywhere else, and testing ψ^y on the unit vectors e_k we see

that $\|\psi^y\|_{op} \geqslant \|y\|_\infty$. Hence the map

$$\psi : \begin{array}{l} \ell^\infty \to (\ell^1)' \\ y \mapsto \psi^y \end{array}$$

is an isometry. It remains to show that it is surjective. Let $v \in (\ell^1)'$ and set $y_k = v(e_k)$, $k \in \mathbb{N}$. Clearly $|v(e_k)| \leqslant \|v\|_{op}$ for all k so that $\|y\|_\infty \leqslant \|v\|_{op}$. Now if $x \in \ell^1$ we see that $\lim_{N\to\infty} \sum_{k=1}^N x_k e_k = x$ with convergence in the ℓ^1 norm. That means

$$v(x) = \lim_{N\to\infty} v\big(\sum_{k=1}^N x_k e_k\big) = \lim_{N\to\infty} \sum_{k=1}^N x_k y_k = \sum_{k=1}^\infty x_k y_k$$

and we see that $v = \psi^y$ for the ℓ^∞ sequence $y = (v(e_k))_{k\in\mathbb{N}}$. Moreover $\|v\|_{op} \leqslant \|y\|_\infty$ so that we get $\|y\|_\infty = \|\psi^y\|_{op} = \|v\|_{op}$. We have shown that ψ is an isometry from ℓ^∞ onto $(\ell^1)'$.

The case $1 < p < \infty$. First by Hölder's inequality, every $y \in \ell^{p'}$ defines a bounded linear functional ψ^y on ℓ^p by setting $\psi^y(x) = \sum_{n=1}^\infty x_n y_n$, and we have $|\psi^y(x)| \leqslant \|x\|_p \|y\|_{p'}$, hence $\|\psi^y\|_{op} \leqslant \|y\|_{p'}$. Clearly $\psi : y \mapsto \psi^y$ defines a linear injective and continuous map from $\ell^{p'}$ to $(\ell^p)'$. We now show that this map is surjective, and that it is an isometry.

Let $v \in (\ell^p)'$, and, as before define $y_k = v(e_k)$ and as in the case $p = 1$ we find that y is a sequence in ℓ^∞. We now show that it belongs to $\ell^{p'}$. Define

$$u_k = \begin{cases} \frac{|y_k|^{p'}}{y_k} & \text{if } y_k \neq 0, \\ 0 & \text{otherwise;} \end{cases}$$

moreover for each n we define the sequence $u^{(n)} \in \ell^p$ by $u_k^{(n)} = u_k$ if $k \leqslant n$ and $u_k^{(n)} = 0$ if $k > n$. For the ℓ^p norm of $u^{(n)}$ we get

$$\|u^{(n)}\|_p = \Big(\sum_{k=1}^n |y_k|^{(p'-1)p}\Big)^{1/p} = \Big(\sum_{k=1}^n |y_k|^{p'}\Big)^{1/p}.$$

here we have used that $p(p'-1) = p(\frac{p}{p-1} - 1) = \frac{p}{p-1} = p'$.

Now we have on the one hand

$$v(u^{(n)}) = \sum_{k=1}^n u_k y_k = \sum_{k=1}^n |y_k|^{p'},$$

and on the other hand

$$|v(u^{(n)})| \leqslant \|v\|_{\mathrm{op}} \|u^{(n)}\|_p = \|v\|_{\mathrm{op}} \Big(\sum_{k=1}^n |y_k|^{p'}\Big)^{1/p}.$$

Combining these two we get

$$\Big(\sum_{k=1}^n |y_k|^{p'}\Big)^{\frac{1}{p'}} \leqslant \|v\|_{\mathrm{op}},$$

for all $n \in \mathbb{N}$ and letting $n \to \infty$ we see that $y \in \ell^{p'}$ with $\|y\|_{\ell^{p'}} \leqslant \|v\|_{op}$.

Then also for all $x \in \ell^p$ we get $v(x) = \sum_{k=1}^\infty x_k y_k$ so that $v = \psi^y$. Above we have shown that $\|y\|_{p'} \leqslant \|\psi^y\|_{op} = \|v\|_{op}$ and also $\|\psi^y\|_{op} \leqslant \|y\|_{p'}$. Hence $\|v\|_{op} = \|\psi^y\|_{op} = \|y\|_{p'}$ and we have proved that $\psi : y \to \psi^y$ is an isometric isomorphism from $\ell^{p'}$ onto $(\ell^p)'$. □

5.52. Exercise. Let $\mathfrak{c}_0$ be the subspace of ℓ^∞ consisting of all sequences $x = (x_k)_{k\in\mathbb{N}}$ for which $\lim_{k\to\infty} x_k = 0$. Then the dual space $(\mathfrak{c}_0)'$ is isometrically isomorphic to ℓ^1.

Remark. Theorem 5.51 does not have an analog for $p = \infty$. Of course the linear functional $x \mapsto \sum_{k=1}^\infty x_k y_k$, with $y \in \ell^1$ belongs to $(\ell^\infty)'$ and has operator norm $\|y\|_{\ell^1}$. However, not every bounded linear functional on ℓ^∞ is of this form; see Example 5.90.

5.53. Corollary. *ℓ^p is a Banach space for all $p \in [1, \infty]$.*

Proof. The case $1 < p \leqslant \infty$ follows from Theorem 5.51 and Corollary 5.21. ℓ^1 is also a Banach space which can be deduced either from Exercise 5.4 or Exercise 5.52. □

5.54. Exercise. Give a direct proof of completeness of ℓ^p which does not use duality.

5.55. Exercise. Let f and g be Riemann integrable on $[a, b]$, and let $1 < p < \infty$, $\frac{1}{p} + \frac{1}{p'} = 1$. Show that $|fg|$, $|f|^p$, $|g|^{p'}$ are Riemann integrable and prove that

$$\int_a^b |f(x)g(x)|\,dx \leqslant \Big(\int_a^b |f(x)|^p\,dx\Big)^{1/p}\Big(\int_a^b |g(x)|^{p'}\,dx\Big)^{1/p'}$$

5.56. Exercise. Let f, g be Riemann integrable, $1 \leqslant p < \infty$. Show that $|f+g|^p$ is Riemann integrable and prove that

$$\Big(\int_a^b |f(x)+g(x)|^p\,dx\Big)^{1/p} \leqslant \Big(\int_a^b |f(x)|^p\,dx\Big)^{1/p} + \Big(\int_a^b |g(x)|^p\,dx\Big)^{1/p}.$$

5.57. Exercise. Show that

$$\|f\|_p = \Big(\int_a^b |f(x)|^p dx\Big)^{1/p}$$

is a norm on the space $C([a, b])$.

5.6. Inner product spaces

5.6.1. Inner products. The space ℓ^2 is unique among the ℓ^p spaces in that it carries an inner product structure.

5.58. Definition. Let V be a vector space with scalar field $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. An *inner product* on V is a map

$$V \times V \to \mathbb{K}$$
$$(f, g) \mapsto \langle f, g\rangle$$

satisfying the following properties, for $f, g, h \in V$ and $\lambda \in \mathbb{K}$:

- Linearity in the first entry: $\langle f + g, h\rangle = \langle f, h\rangle + \langle g, h\rangle$, and $\langle \lambda f, g\rangle = \lambda\langle f, g\rangle$.
- (Anti-)Symmetry: $\langle f, g\rangle = \overline{\langle g, f\rangle}$,
- Nonnegativity: $\langle f, f\rangle \geqslant 0$,

An inner product is called a *scalar product* if it satisfies in addition

- Positive definiteness: $\langle f, f\rangle > 0$ if f is not zero.

A vector space V with an inner product is called an *inner product space.* A vector space with a scalar product is called a *pre-Hilbert space.*

Remarks.

(i) In a real pre-Hilbert space, the anti-symmetry property becomes an actual symmetry property $\langle f, g\rangle = \langle g, f\rangle$ (since $a = \overline{a}$ for any $a \in \mathbb{R}$). Furthermore, using the linearity and the (anti-)symmetry property we also get

$$\langle h, \lambda f + \mu g\rangle = \overline{\lambda}\langle h, f\rangle + \overline{\mu}\langle h, g\rangle.$$

This property together with the linearity property in the first entry is sometimes called the *sesquilinear property* of the inner product. Note that on real inner product spaces we have the linearity property both with respect to the first entry (fixing the second entry) and with respect to the second entry (fixing the first entry).

(ii) A pre-Hilbert space is sometimes aptly called a *Euclidean space* since it carries many properties of the Euclidean geometry of $\mathbb{R}^n$ or $\mathbb{C}^n$. However since one often associates finite-dimensionality with

the term "Euclidean" we shall use this term only for finite-dimensional spaces with a scalar product. Pre-Hilbert spaces are named after David Hilbert (1862-1943); the terminology will be motivated by the use of *Hilbert spaces* in §5.6.2 below.

(iii) Note that different authors may use slightly different terminology. In the broader literature, inner product spaces are often defined to include the property of positive definiteness, whereas in this text we will find it convenient in later chapters to only require non-negativity.

5.59. Examples. Each of the following is a pre-Hilbert space.

- $\mathbb{R}^n$ with $\langle x, y\rangle = \sum_{j=1}^n x_j y_j$.
- $\mathbb{C}^n$ with $\langle w, z\rangle = \sum_{j=1}^n w_j \overline{z_j}$.
- $\mathbb{R}^{m\times n}$ with $\langle A, B\rangle = \operatorname{tr}(AB^\intercal) = \sum_{i=1}^m \sum_{j=1}^n a_{ij} b_{ij}$
- $\mathbb{C}^{m\times n}$ with $\langle A, B\rangle = \operatorname{tr}(A\overline{B}^\intercal) = \sum_{i=1}^m \sum_{j=1}^n a_{ij} \overline{b_{ij}}$
- ℓ^2 with $\langle f, g\rangle = \sum_{j=1}^\infty f(j)\overline{g(j)}$.

Verify the defining properties of an inner product for the above examples. Regarding the inner products on $\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$ and its formulation using the trace see Exercise 6.114.

5.60. Example (Square-integrable functions). Our main examples are classes of square-integrable functions. The vector space of complex-valued Riemann integrable functions on $[a, b]$ is an inner product space with the inner product

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}\, dx. \tag{5.2}$$

The right hand side is well-defined since the product of Riemann integrable functions is Riemann integrable (Theorem 2.18). The inner product is not a scalar product as it does not have the positive-definiteness property. For example, if f is nonzero only at finitely many points then $\langle f, f\rangle = 0$. The norm associated with this inner product is called L^2 *norm*:

$$\|f\|_2 = \Big(\int_a^b |f(x)|^2\, dx\Big)^{1/2}.$$

5.61. Example. Consider the subspace of continuous complex-valued functions on $[a, b]$ with the same inner product as in (5.2) for Riemann integrable functions. The inner product then becomes a scalar product (making the space of continuous functions on $[a, b]$ a pre-Hilbert space). To check the positive definiteness assume that $f(x_0) \neq 0$ for some $x_0 \in [a, b]$. Then by continuity of f there exists a $\delta > 0$ such that $|f(x)| > |f(x_0)|/2$ for all x in a subinterval $I \subset [a, b]$ of length δ, with $x_0 \in I$, and we get using Theorem 2.19

$$\langle f, f\rangle = \int_a^b |f(t)|^2\, dt \geqslant \int_I |f(t)|^2\, dt \geqslant \int_I \frac{|f(x_0)|^2}{4}\, dt = \frac{\delta |f(x_0)|^2}{4} > 0.$$

Hence $\langle f, f\rangle = 0$ only holds if $f = 0$.

5.62. Definition. Let I be an interval. A function $f : I \to \mathbb{C}$ is called *piecewise continuous* if f has only finitely many points of discontinuity and if all such discontinuities are jump discontinuities.

5.63. Examples.

(i) Piecewise continuous functions form a subspace of the inner product space of Riemann integrable functions. The inner product (5.2) on this space does not have the positive-definiteness property.

(ii) The subspace of right continuous functions (that is, for $a \leqslant x < b$ we have $\lim_{\delta\to 0+} f(x + \delta) = f(x)$) does have the positive-definiteness property and thus is a pre-Hilbert space.

The following generalizes the special case of Hölder's inequality for the exponent $p = 2$ (Theorem 5.44).

5.64. Theorem (Cauchy–Schwarz inequality). *An inner product on V satisfies for all $f, g \in V$ the inequality*

$$|\langle f, g\rangle| \leqslant \sqrt{\langle f, f\rangle}\sqrt{\langle g, g\rangle}.$$

Proof. We first prove the inequality

$$\operatorname{Re}\langle f, g\rangle \leqslant \sqrt{\langle f, f\rangle}\sqrt{\langle g, g\rangle} \tag{5.3}$$

which already gives the Cauchy–Schwarz inequality if $\mathbb{K} = \mathbb{R}$. We consider for $t \in \mathbb{R}$ the inner product of $f + tg$ with itself (which is nonnegative). Thus using the properties of the inner product

$$\begin{aligned} 0 \leqslant \langle f + tg, f + tg\rangle &= \langle f, f\rangle + t\langle f, g\rangle + t\langle g, f\rangle + t^2\langle g, g\rangle \\ &= \langle f, f\rangle + 2t \operatorname{Re}\langle f, g\rangle + t^2\langle g, g\rangle = p(t). \end{aligned}$$

If $\langle g, g\rangle = 0$, then the resulting inequality $0 \leqslant \langle f, f\rangle + 2t \operatorname{Re}\langle f, g\rangle$ for all t implies that $\operatorname{Re}\langle f, g\rangle = 0$ and (5.3) holds in this case. Now let us assume that $\langle g, g\rangle \neq 0$, i.e. $\langle g, g\rangle > 0$. It is elementary to verify that the nonnegative quadratic polynomial $p(t)$ in the last display has a minimum at $t_0 = -\frac{\operatorname{Re}\langle f, g\rangle}{\langle g, g\rangle}$. The inequality $p(t_0) \geqslant 0$ yields

$$0 \leqslant \langle f, f\rangle + 2t_0 \operatorname{Re}\langle f, g\rangle + t_0^2\langle g, g\rangle = \langle f, f\rangle - \frac{(\operatorname{Re}\langle f, g\rangle)^2}{\langle g, g\rangle},$$

that is $(\operatorname{Re}\langle f, g\rangle)^2 \leqslant \langle f, f\rangle\langle g, g\rangle$ and (5.3) follows.

Now let us assume that $\mathbb{K} = \mathbb{C}$. Consider $\langle f, g\rangle \in \mathbb{C}$. We can find a complex number a such that $|a| = 1$ and $a\langle f, g\rangle = \langle af, g\rangle$ is real and nonnegative. Then using (5.3),

$$|\langle f, g\rangle| = a\langle f, g\rangle = \langle af, g\rangle \leqslant \sqrt{\langle af, af\rangle}\sqrt{\langle g, g\rangle} = \sqrt{\langle f, f\rangle}\sqrt{\langle g, g\rangle}$$

which proves the Cauchy–Schwarz inequality in the complex category as well. □

5.65. Exercise. Show that in a pre-Hilbert space we have equality in $|\langle f, g\rangle|^2 = \langle f, f\rangle\langle g, g\rangle$ if and only if f and g are linearly dependent.

5.66. Theorem (Triangle inequality). *Let V be an inner product space. Set $\|f\| = \sqrt{\langle f, f\rangle}$. Then for $f, g \in V$*

$$\|f + g\| \leqslant \|f\| + \|g\|.$$

Proof. We may assume $\|f + g\| \neq 0$ because otherwise there is nothing to prove. Then

$$\|f + g\|^2 = \langle f + g, f + g\rangle = \langle f, f + g\rangle + \langle g, f + g\rangle \leqslant \|f\|\|f + g\| + \|g\|\|f + g\|$$

and the assertion follows after dividing by $\|f + g\|$. □

5.67. Corollary. *Let V be an inner product space. Then the expression $\|f\| = \sqrt{\langle f, f\rangle}$ defines a seminorm on V. In a pre-Hilbert space this seminorm becomes a norm and $d(f, g) = \|f - g\|$ defines a metric.*

5.68. Proposition. *The operation of the inner product $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$ is continuous.*

Proof. This follows with the usual trick for product type expressions. We write

$$|\langle a, b\rangle - \langle c, d\rangle| = |\langle a - c, b\rangle + \langle c, b - d\rangle| \leqslant |\langle a - c, b\rangle| + |\langle c, b - d\rangle| \leqslant \|b\|\|a - c\| + \|c\|\|b - d\|$$

where in the last step we have used the Cauchy-Schwarz inequality. □

5.69. Definition. Two vectors f, g in an inner product space V are called *orthogonal* if $\langle f, g\rangle = 0$.

Note that by the (anti-)symmetry of the scalar product, f, g are orthogonal if and only if g, f are orthogonal.

5.70. Proposition (The Pythagorean Theorem). *Let f, g be orthogonal vectors in the inner product space V. Then*

$$\|f+g\|^2 = \|f\|^2 + \|g\|^2.$$

Proof. We compute

$$\begin{aligned}\|f+g\|^2 = \langle f+g, f+g\rangle &= \langle f,f\rangle + \langle f,g\rangle + \langle g,f\rangle + \langle g,g\rangle \\ &= \langle f,f\rangle + \langle g,g\rangle = \|f\|^2 + \|g\|^2 .\end{aligned}$$

□

5.71. Proposition. *Let $\langle\cdot,\cdot\rangle$ be an inner product on the vector space V. Then:*

(i) For all f, g, the identity

$$\|f+g\|^2 + \|f-g\|^2 = 2\|f\|^2 + 2\|g\|^2$$

holds.

(ii) If V is a real vector space then for all $f, g \in V$

$$\langle f,g\rangle = \tfrac{1}{4}\big(\|f+g\|^2 - \|f-g\|^2\big).$$

(iii) If V is a complex vector space then for all $f, g \in V$

$$\langle f,g\rangle = \tfrac{1}{4}\big(\|f+g\|^2 - \|f-g\|^2 + i\|f+ig\|^2 - i\|f-ig\|^2\big).$$

The identity in (i) is called the *parallelogram identity.* It reflects a result from Euclidean geometry relating the lengths of the diagonals in a parallelogram to the lengths of the sides. The formulas in (ii), (iii) are also known as *polarization identities.*

Proof. For (i) we use the additivity in the first and second entry to compute that

$$\begin{aligned}\langle f+g, f+g\rangle &= \langle f,f\rangle + \langle g,g\rangle + 2\operatorname{Re}\langle f,g\rangle,\\ \langle f-g, f-g\rangle &= \langle f,f\rangle + \langle g,g\rangle - 2\operatorname{Re}\langle f,g\rangle.\end{aligned}$$

Hence $\langle f+g, f+g\rangle + \langle f-g, f-g\rangle = 2\langle f,f\rangle + 2\langle g,g\rangle$ which is the asserted parallelogram identity.

We use the additivity properties of the inner product to compute

$$\langle f+g, f+g\rangle - \langle f-g, f-g\rangle = 4\operatorname{Re}\langle f,g\rangle$$

which implies (ii). If $z = a+ib \in \mathbb{C}$ then $\operatorname{Im}(z) = \operatorname{Re}(-iz)$ and the last display with ig in place of g yields

$$\langle f+ig, f+ig\rangle - \langle f-ig, f-ig\rangle = 4\operatorname{Re}\langle f, ig\rangle = 4\operatorname{Im}\langle f,g\rangle.$$

Now combine the formulas for the real and imaginary parts of $\langle f,g\rangle$ and (iii) follows. □

5.72. Exercise. For each of the following, prove that no inner product $\langle\cdot,\cdot\rangle$ exists with $\|x\| = \sqrt{\langle x,x\rangle}$ for all $x \in X$:

(i) $X = \mathbb{R}^n$, $n \geqslant 2$, with $\|x\| = \max_{i=1,\dots,n} |x_i|$.

(ii) $X = \mathbb{C}^n$, $n \geqslant 2$, with $\|x\| = \sum_{i=1}^n |x_i|$.

(iii) $X = C([0,1])$ with $\|f\| = \sup_{t\in[0,1]} |f(t)|$.

5.73. Exercise. Let $(V, \|\cdot\|)$ be a normed space. Show that there is a continuous scalar product on V with $\|f\| = \langle f,f\rangle^{1/2}$ if and only if the parallelogram identity

$$\|f+g\|^2 + \|f-g\|^2 = 2\|f\|^2 + 2\|g\|^2$$

holds.

5.6.2. Hilbert spaces and the Riesz representation theorem. We now investigate the role of completeness for an inner product space.

5.74. Definition. A pre-Hilbert space which is complete as a normed space is called a *Hilbert space*.

5.75. Example. The space $\ell^2(\mathbb{N})$ of square summable sequences $(f(n))_{n\in\mathbb{N}}$ (with the inner product $\langle f,g\rangle = \sum_{n=1}^{\infty} f(n)\overline{g(n)}$) is a Hilbert space. The completeness property follows from Corollary 5.53.

5.76. Exercise. Consider the pre-Hilbert space $V = C([0,1])$ (with the scalar product $\langle f,g\rangle = \int_0^1 f\bar{g}dx$ as in Examples 5.60, 5.61). It is not a Hilbert space: Define

$$f_n(x) = \begin{cases} x^{-1/2} & \text{if } 2^{-n} \leqslant x \leqslant 1 \\ 2^{n/2} & \text{if } 0 \leqslant x < 2^{-n} \end{cases}$$

and show that f_n is a Cauchy sequence (with respect to the norm $\|\cdot\|_2 = \langle\cdot,\cdot\rangle^{1/2}$) which does not converge to a limit in V.

The Cauchy-Schwarz inequality can be viewed as a statement about continuity and operator norm of the linear functional $\lambda_g \in H'$ defined by $\lambda_g(f) = \langle f,g\rangle$. The following lemma states some simple facts about the map which assigns to g the linear form λ_g.

5.77. Lemma. *Let H be a pre-Hilbert space and define*

$$\jmath : \begin{cases} H \to H' \\ g \mapsto \lambda_g \text{ where } \lambda_g(f) = \langle f,g\rangle \text{ for all } f \in H. \end{cases} \tag{5.4}$$

Then

(i) *$\jmath$ is an antilinear map, i.e. for $f,g \in H$, $\alpha \in \mathbb{K}$ we have*

$$\jmath(f+g) = \jmath(f) + \jmath(g)$$
$$\jmath(\alpha f) = \overline{\alpha}\jmath(f).$$

(ii) *$\jmath$ is an isometry, i.e. $\|g\|_H = \|\lambda_g\|_{H'}$; in particular $\jmath$ is injective.*

Proof. (i) follows from $\langle f, g+h\rangle = \langle f,g\rangle + \langle f,h\rangle$ for all $f \in H$ and from $\langle f,\alpha g\rangle = \overline{\alpha}\langle f,g\rangle$ for all $f \in H$.

By the Cauchy-Schwarz inequality, $|\lambda_g(f)| = |\langle f,g\rangle| \leqslant \|f\|_H\|g\|_H$ and because of $\jmath(g) = \lambda_g$ we get $\|\jmath(g)\|_{H'} \leqslant \|g\|_H$. For the opposite inequality we get, for $g \neq 0$ that $\jmath(g)(\frac{g}{\|g\|}) = \|g\|$ so $\|\jmath(g)\|_{H'} \geqslant \|g\|_H$ and (ii) is proved. □

If in addition the pre-Hilbert space is complete, i.e. H is a Hilbert space then we show that the map $\jmath$ in (5.4) is also surjective; this fact is known as the *Riesz representation theorem* for Hilbert spaces.

5.78. Theorem. *Let H be a Hilbert space. For every $\lambda \in H'$ there exists a unique $g \in H$ such that $\lambda = \jmath(g)$, i.e. $\lambda(f) = \langle f,g\rangle$ for all $f \in H$.*

Proof. The uniqueness is already proved in Lemma 5.77.

For the existence part consider a normalized $\lambda \in H'$ with $\|\lambda\|_{H'} = 1$. The first step is to show that there is a $g \in H$ such that $\|g\| = 1$ and $\lambda(g) = 1$. The assumption $\|\lambda\|_{H'} = 1$ means that there is a sequence $g_n \in H$ such that $\|g_n\| = 1$ and $\lim_{n\to\infty}|\lambda(g_n)| = 1$. We have $\lambda(g_n) = e^{i\theta_n}c_n$ where $\theta_n \in \mathbb{R}$, $c_n \in \mathbb{R}$ with $0 < c_n \leqslant 1$. Let $\tilde{g}_n = e^{-i\theta_n}g_n$ then $\lambda(\tilde{g}_n)$ is real and belongs to $(0,1]$, moreover $\lim_{n\to\infty}\lambda(\tilde{g}_n) = 1$.

We show that $\tilde{g}_n$ is a Cauchy sequence. Let $\varepsilon > 0$ then there is N_ε such that $\lambda(\tilde{g}_n) \geqslant 1-\varepsilon$ for $n > N_\varepsilon$. This also implies that $\lambda(\tilde{g}_n + \tilde{g}_m) \geqslant 2-2\varepsilon$ for $n,m > N_\varepsilon$. Since $\|\lambda\|_{H'} = 1$ we get $\|\tilde{g}_n + \tilde{g}_m\| \geqslant 2-2\varepsilon$ for $n,m \geqslant N_\varepsilon$ and since $\|\tilde{g}_n\| = 1$ we obtain by the parallelogram identity

$$\|\tilde{g}_n - \tilde{g}_m\|^2 = 2\|\tilde{g}_n\|^2 + 2\|\tilde{g}_m\|^2 - \|\tilde{g}_n + \tilde{g}_m\|^2 = 4 - \|\tilde{g}_n + \tilde{g}_m\|^2 \leqslant 4 - (2-2\varepsilon)^2 = 8\varepsilon - 4\varepsilon^2,$$

provided that $n, m > N_\varepsilon$. This clearly shows that $\tilde{g}_n$ is a Cauchy sequence. By the assumed completeness of H its limit exists; we set $g = \lim_{n\to\infty} \tilde{g}_n$. Note that Re $\lambda(g) = 1$ and $\|g\| = 1$.

We now show that $\lambda(f) = \langle f, g\rangle$ for all $f \in H$ (which means that $\jmath(g) = \lambda$). We first consider the real part of $\lambda(f)$ (which defines a linear functional with $\mathbb{R}$ as a scalar field).

Note that Re $\lambda(g + tf) \leqslant |\lambda(g + tf)| \leqslant \|g + tf\|$ and since Re $\lambda(g) = 1 = \|g\|$ we get

$$\text{Re } \lambda(f) = \frac{\text{Re } \lambda(g + tf) - \text{Re } \lambda(g)}{t} \leqslant \frac{\|g + tf\| - \|g\|}{t}.$$

Similarly (also for $t > 0$)

$$\text{Re } \lambda(f) = -\frac{\text{Re } \lambda(g - tf) - \text{Re } \lambda(g)}{t} \geqslant -\frac{\|g - tf\| - \|g\|}{t}.$$

The two inequalities imply

$$-\frac{\|g - tf\|^2 - \|g\|^2}{t(\|g - tf\| + \|g\|)} \leqslant \text{Re } \lambda(f) \leqslant \frac{\|g + tf\|^2 - \|g\|^2}{t(\|g + tf\| + \|g\|)},$$

and since $\|g \pm tf\|^2 = \|g\|^2 \pm 2t \,\text{Re}\langle g, f\rangle + t^2\|f\|^2$ we obtain

$$\frac{2\,\text{Re}\langle g, f\rangle - t\|f\|^2}{\|g - tf\| + \|g\|} \leqslant \text{Re } \lambda(f) \leqslant \frac{2\,\text{Re}\langle g, f\rangle + t\|f\|^2}{\|g + tf\| + \|g\|}.$$

We recall that $\|g\| = 1$ and let $t \to 0+$ to obtain $\text{Re}\,(\lambda(f)) = \text{Re}\langle g, f\rangle = \text{Re}\langle f, g\rangle$.

For any complex number z we have $-\,\text{Re}\,(iz) = \text{Im}\,(z)$ and therefore

$$\text{Im}\,(\lambda(f)) = -\,\text{Re}\,(i\lambda(f)) = -\,\text{Re}\,(\lambda(if)) = -\,\text{Re}\langle if, g\rangle = -\,\text{Re}\,(i\langle f, g\rangle) = \text{Im}\langle f, g\rangle$$

and thus we have verified $\lambda(f) = \langle f, g\rangle$ (under the assumption that $\|\lambda\|_{H'} = 1$).

Now consider arbitrary $\lambda \in H'$. If $\lambda = 0$ then $\lambda = \jmath(0)$. If $\lambda \neq 0$ we apply the above to $\lambda/\|\lambda\|$ and obtain $g_1 \in H$ with $\|g_1\| = 1$ such that $\lambda(f)/\|\lambda\| = \langle f, g_1\rangle$ for all $f \in H$. This gives $\lambda = \jmath(g)$ with $g = \|\lambda\|g_1$ and the proof is complete. □

REMARK. The completeness of H is a crucial assumption in Theorem 5.78. The sequence space ℓ^2 is an example of a Hilbert space. The space $C([a, b])$ with the inner product $\int f\overline{g}dx$ is not a Hilbert space and the conclusion of the Riesz representation theorem does not hold in this case; simply consider the functionals $f \mapsto \int fg$ where g is bounded but has jump discontinuities. The inner product space $C([a, b])$ is embedded in the complete Hilbert space $L^2([a, b])$ of square integrable functions in the sense of Lebesgue; this definition requires the Lebesgue theory of integration which is not considered in this text.

The use of the parallelogram identity was crucial in the proof of Theorem 5.78. For a similar application see Exercise 5.121.

The following theorem is a straightforward and very useful application of the Riesz representation theorem.

5.79. THEOREM. *Let H_1, H_2 be Hilbert spaces and let $T \in L(H_1, H_2)$. Then there exists a unique operator $T^* \in L(H_2, H_1)$ such that $\langle Tf, g\rangle_{H_2} = \langle f, T^*g\rangle$ for all $f \in H_1$ and $g \in H_2$. Moreover,*

$$\|T\|_{H_1\to H_2} = \|T^*\|_{H_2\to H_1} = \|T^*T\|^{1/2}_{H_1\to H_1} = \|TT^*\|^{1/2}_{H_2\to H_2}$$

Given $T \in L(H_1, H_2)$ we refer to the unique T^* in Theorem 5.79 as the *adjoint operator*.

PROOF OF THEOREM 5.79. The uniqueness assertion follows from the fact that if a vector in H_1 is orthogonal to all other vectors then it is 0.

Let $g \in H_2$ and consider the linear functional $\Lambda_g : H_1 \to \mathbb{C}$, given by $\Lambda_g(f) = \langle Tf, g\rangle$. It is bounded since

$$|\Lambda_g(f)| \leqslant \|Tf\|_{H_2}\|g\|_{H_2} \leqslant \|T\|_{H_1\to H_2}\|f\|_{H_1}\|g\|_{H_2}$$

and we see that $\|\Lambda_g\|_{H_1'} \leqslant \|T\|_{H_1\to H_2}\|g\|_{H_2}$. By the Riesz representation theorem there exists a unique vector in H_1, labeled T^*g, such that $\Lambda_g(f) = \langle f, T^*g\rangle$ for all $f \in H_1$. We show that T^* is linear. This follows from

$$\begin{aligned}\langle f, T^*(\alpha g + \beta h)\rangle &= \langle Tf, \alpha g + \beta h\rangle = \overline{\alpha}\langle Tf, g\rangle + \overline{\beta}\langle Tf, h\rangle\\ &= \overline{\alpha}\langle f, T^*g\rangle + \overline{\beta}\langle f, T^*h\rangle = \langle f, \alpha T^*g + \beta T^*h\rangle.\end{aligned}$$

Indeed, this shows that $T^*(\alpha g + \beta h) - (\alpha T^*g + \beta T^*h)$ is orthogonal to all $f \in H_1$, hence equal to 0 and the linearity of T^* is proved.

Since $\|\Lambda_g\|_{H_1'} = \|T^*g\|_{H_1}$ we also have $\|T^*g\|_{H_1} \leqslant \|T\|_{H_1\to H_2}\|g\|_{H_2}$ (from the above bound for $\|\Lambda_g\|_{H_1'}$). This implies $T^* \in L(H_2, H_1)$, and also $\|T^*\|_{H_2\to H_1} \leqslant \|T\|_{H_1\to H_2}$. Vice versa,

$$\|Tf\|_{H_2}^2 = \langle Tf, Tf\rangle = \langle f, T^*Tf\rangle \leqslant \|f\|_{H_1}\|T^*Tf\|_{H_1} \leqslant \|f\|_{H_1}\|T^*\|_{H_2\to H_1}\|Tf\|_{H_2}$$

which implies $\|Tf\|_{H_2} \leqslant \|T^*\|_{H_2\to H_1}\|f\|_{H_1}$ and thus $\|T\|_{H_1\to H_2} \leqslant \|T^*\|_{H_2\to H_1}$. Combining the two estimates we get $\|T\|_{H_1\to H_2} = \|T^*\|_{H_2\to H_1}$.

From the last display we also get

$$\|Tf\|_{H_2}^2 \leqslant \|T^*T\|_{H_1\to H_1}\|f\|_{H_1}^2$$

which implies $\|T\|_{H_1\to H_2} \leqslant \|T^*T\|_{H_1\to H_1}^{1/2}$. On the other hand

$$\|T^*Tf\|_{H_1} \leqslant \|T^*\|_{H_2\to H_1}\|T\|\|f\|_{H_1} = \|T\|^2\|f\|_{H_1}$$

which yields $\|T^*T\|_{H_1\to H_1}^{1/2} \leqslant \|T\|$, so we have proved $\|T\| = \|T^*T\|^{1/2}$. Applying the same to T^* and using $\|T\| = \|T^*\|$ gives $\|T\| = \|TT^*\|^{1/2}$. □

5.7. Hahn–Banach theorem

From linear algebra we know that the dual space of a finite dimensional space is again a finite dimensional space of equal dimension. Moreover, we have previously discussed the duals of some interesting Banach spaces spaces (e.g. Theorem 5.78 and Theorem 5.51). However, for an arbitrary normed space X, it is not yet clear that there are any nontrivial linear functionals. In this section we prove a theorem of Hahn and Banach which states that there are nontrivial linear functionals in X, indeed the space is quite rich as we will see that for any vector $x \neq 0$ there is a linear functional with $\lambda(x) \neq 0$. The main idea is to have a mechanism on extending linear functionals on a subspace to a larger space while preserving some prescribed quantitative control. This quantitative control is formulated via a norm (such as for "bounded" linear functionals), or, more generally, via a sublinear functional.

5.7.1. Sublinear functionals.

5.80. Definition. Let X be a vector space. A functional $p : X \to \mathbb{R}$ is *sublinear* if it satisfies

- $p(x + y) \leqslant p(x) + p(y)$ for all $x, y \in X$ (subadditivity).
- $p(tx) = tp(x)$ for all $t \geqslant 0$ (homogeneity).

5.81. Examples.

(i) Any norm or seminorm on X is a sublinear functional.
(ii) Every real linear functional $X \to \mathbb{R}$ is sublinear.
(iii) Let $p : \mathbb{R} \to \mathbb{R}$ be defined by $p(0) = 0$, $p(x) = \beta x$ when $x > 0$, $p(x) = \alpha x$ when $x < 0$ where $\alpha, \beta \in \mathbb{R}$ with $\alpha \leqslant \beta$. Then p is sublinear, and all sublinear functionals $\mathbb{R} \to \mathbb{R}$ are of this form (provide a graph).
(iv) If $\{p_\alpha\}_{\alpha\in A}$ is a family of sublinear functionals with $p(x) = \sup_\alpha p_\alpha(x) < \infty$ for all $x \in X$, then p defines a sublinear functional on X.

We now introduce the Minkowski functional of a convex body which is often used for quantitative control of linear functionals.

5.82. Definition. Let X be a vector space.

(i) For $x, y \in X$ and $0 \leqslant t \leqslant 1$ we call $tx + (1-t)y$ a *convex combination* of x and y. The set $\{tx + (1-t)y : 0 \leqslant t \leqslant 1\}$ of convex combinations of x and y is the *line segment* joining x and y.

(ii) A subset E of a vector space X is *convex* if for every two points $x, y \in E$ the line segment joining x and y is contained in E.

(iii) For a convex set E we define $\mathrm{Int}(E)$ as the set of points x such that for every $y \in X$ there is an $\varepsilon = \varepsilon(x,y) > 0$ with $x + ty \in E$ whenever $|t| < \varepsilon$. A convex set $\Omega \subset X$ is called a *convex body* if $0 \in \mathrm{Int}(\Omega)$.

(iv) Let Ω be a convex body in X. Then the functional p_Ω given for $x \in X$ by

$$p_\Omega(x) = \inf\left\{t > 0 : \tfrac{x}{t} \in \Omega\right\}$$

is called the *Minkowski functional* associated to the convex body Ω.

5.83. Exercise. Let $R > 0$. Verify that the open ball $\Omega_R = \{x : \|x\| < R\}$ centered at the origin of a normed space and its closure are convex bodies, and that $p_{\Omega_R}(x) = p_{\overline{\Omega}_R}(x) = \|x\|/R$ for all $x \in X$.

5.84. Proposition. *The Minkowski functional of a convex body Ω is finite and sublinear.*

Proof. Since $0 \in \mathrm{Int}(\Omega)$, for any fixed $x \in X$, there is an $R > 0$ such that $x/R \in \Omega$. Hence $p_\Omega(x) < \infty$. For $\lambda > 0$

$$\inf\{R > 0 : \tfrac{\lambda x}{R} \in \Omega\} = \lambda \inf\{R' > 0 : \tfrac{x}{R'} \in \Omega\}$$

which implies the homogeneity property $p_\Omega(\lambda x) = \lambda p_\Omega(x)$. It remains to check subadditivity. Let $x_1, x_2 \in X$ and $\varepsilon > 0$. Choose $R_1, R_2 > 0$ such that $x_i/R_i \in \Omega$ for $i = 1, 2$ and $R_i \leqslant p_\Omega(x_i) + \varepsilon$. Since Ω is convex, the convex combination

$$\frac{x_1 + x_2}{R_1 + R_2} = \frac{R_1}{R_1 + R_2}\frac{x_1}{R_1} + \frac{R_2}{R_1 + R_2}\frac{x_2}{R_2}$$

is also in Ω. Therefore,

$$p_\Omega(x_1 + x_2) \leqslant R_1 + R_2 \leqslant p_\Omega(x_1) + p_\Omega(x_2) + 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, subadditivity follows. □

5.85. Exercise. A convex body Ω is called *symmetric* if it has the property that $-x \in \Omega$ for every $x \in \Omega$. Sketch some examples of symmetric and nonsymmetric convex bodies in $\mathbb{R}^2$. Prove that a convex body in X is symmetric if and only if its Minkowski functional is a seminorm.

5.7.2. Hahn–Banach extension theorem. The Hahn–Banach theorem allows us to extend linear functionals defined on a subspace of a real vector space to the whole space, with suitable quantitative control.

5.86. Theorem. *Let X be a real vector space, and let $p : X \to \mathbb{R}$ be a sublinear functional. Let V be a subspace of X, $\lambda : V \to \mathbb{R}$ be a linear functional such that $\lambda(v) \leqslant p(v)$ for all $v \in V$. Then:*

(i) There is a linear functional $\Lambda : X \to \mathbb{R}$ such that $\Lambda(v) = \lambda(v)$ for all $v \in V$ and $-p(-x) \leqslant \Lambda(x) \leqslant p(x)$ for all $x \in X$.

(ii) Let $\mathcal{E}(\lambda)$ denote the set of all real linear functionals $\Lambda : X \to \mathbb{R}$ such that $\Lambda|_V = \lambda$ and $-p(-x) \leqslant \Lambda(x) \leqslant p(x)$ for all $x \in X$. For $x_1 \in X \backslash V$ let $\mathcal{E}(\lambda, x_1) = \big\{\Lambda(x_1) : \Lambda \in \mathcal{E}(\lambda)\big\}$. Then

$$\mathcal{E}(\lambda, x_1) = \left[\alpha_{\min}(V, x_1), \alpha_{\max}(V, x_1)\right],$$

with

$$\alpha_{\min}(V, x_1) = \sup_{v \in V}\{\lambda(v) - p(v - x_1)\}, \qquad \alpha_{\max}(V, x_1) = \inf_{v \in V}\{p(v + x_1) - \lambda(v)\}. \tag{5.5}$$

REMARK. Part (i) is usually stated as the *Hahn–Banach extension theorem.* The quantification in part (ii) can be quite useful in some applications (see Theorem 5.99 below).

We shall first state a basic lemma which allows us to extend a linear functional on a subspace V of a real vector space, and $x_1 \notin V$, to the larger vector space $V + \mathbb{R}x_1$. While this is easy to accomplish the main point here is to preserve the quantitative control by the sublinear functional.

5.87. LEMMA. *Let X be a real vector space, and $p : X \to \mathbb{R}$ a sublinear functional. Let V be a subspace of X, $\lambda : V \to \mathbb{R}$ be a linear functional such that $\lambda(v) \leqslant p(v)$ for all $v \in V$, and let $x_1 \notin V$. Let $\alpha_{\min}(V, x_1)$, $\alpha_{\max}(V, x_1)$ be as in* (5.5)*. Then*

$$\alpha_{\min}(V, x_1) \leqslant \alpha_{\max}(V, x_1). \tag{5.6}$$

If $V_1 = V + \mathbb{R}x_1$ and if $\lambda_1 : V_1 \to \mathbb{R}$ is a linear functional such that $\lambda_1|_V = \lambda$ and $\lambda_1(x) \leqslant p(x)$ for all $x \in V_1$ then we must have

$$\alpha_{\min}(V, x_1) \leqslant \lambda_1(x_1) \leqslant \alpha_{\max}(V, x_1).$$

Vice versa, for every α with $\alpha_{\min}(V, x_1) \leqslant \alpha \leqslant \alpha_{\max}(V, x_1)$ the linear functional $\lambda_1 : V_1 \to \mathbb{R}$ defined by $\lambda_1(v + tx_1) = \lambda(v) + t\alpha$ satisfies $\lambda_1|_V = \lambda$ and $\lambda_1(x) \leqslant p(x)$ for all $x \in V_1$.

PROOF. For $v, w \in V$

$$\lambda(v) + \lambda(w) = \lambda(v + w) \leqslant p(v + w) \leqslant p(v - x_1) + p(x_1 + w)$$

where in the last inequality we have used the subadditivity of p on X. The above implies

$$\lambda(v) - p(v - x_1) \leqslant p(w + x_1) - \lambda(w)$$

for all $v, w \in V$. Taking the supremum on the left hand side, we see that it is finite, and with the definition in (5.5), $\alpha_{\min}(V, x_1) \leqslant p(w + x_1) - \lambda(w)$ for all $w \in V$. This means that the infimum on the right hand side in the latter inequality is also finite and we obtain

$$\alpha_{\min}(V, x_1) \leqslant \alpha_{\max}(V, x_1).$$

Now if λ_1 is an extension of λ to V_1 with $\lambda_1 \leqslant p$ on V_1 we have

$$\lambda_1(x_1) = \lambda_1(w + x_1) - \lambda_1(w) = \lambda_1(w + x_1) - \lambda(w) \leqslant p(w + x_1) - \lambda(w)$$

and taking the infimum over $w \in V$ we see that $\lambda_1(x_1) \leqslant \alpha_{\max}(V, x_1)$. Similarly,

$$\lambda_1(x_1) = \lambda_1(v) - \lambda_1(v - x_1) \geqslant \lambda(v) - p(v - x_1)$$

for all $v \in V$ and taking the supremum over v, tells us $\lambda_1(x_1) \geqslant \alpha_{\min}(V, x_1)$.

Vice versa, if $\alpha_{\min}(V, x_1) \leqslant \alpha \leqslant \alpha_{\max}(V, x_1)$ we *define* $\lambda_1(v + tx_1) = \lambda(v) + t\alpha$. Then for $t > 0$, using the homogeneity of p and the definition of $\alpha_{\max}$,

$$\begin{aligned} \lambda_1(v + tx_1) &= \lambda(v) + t\alpha \leqslant \lambda(v) + t\alpha_{\max} \leqslant \lambda(v) + t\big(p(\tfrac{v}{t} + x_1) - \lambda(\tfrac{v}{t})\big) \\ &= tp(\tfrac{v}{t} + x_1) = p(v + tx_1) \end{aligned}$$

and, similarly (now using the definition of $\alpha_{\min}$),

$$\begin{aligned} \lambda_1(v - tx_1) &= \lambda(v) - t\alpha \leqslant \lambda(v) - t\alpha_{\min} \leqslant \lambda(v) - t\big(\lambda(\tfrac{v}{t}) - p(\tfrac{v}{t} - x_1)\big) \\ &= tp(\tfrac{v}{t} - x_1) = p(v - tx_1). \end{aligned}$$

Hence $\lambda_1(x) \leqslant p(x)$ for all $x \in V_1$. □

PROOF OF THEOREM 5.86. First let Λ be any real linear functional with $\Lambda|_V = \lambda$ and $\Lambda(x) \leqslant p(x)$ for all $x \in X$. Then

$$\Lambda(x_1) \in [\alpha_{\min}(V, x_1), \alpha_{\max}(V, x_1)]$$

by Lemma 5.87 applied to $\Lambda|_{V+\mathbb{R}x_1}$. Hence $\mathcal{E}(\lambda, x_1) \subset [\alpha_{\min}(V, x_1), \alpha_{\max}(V, x_1)]$. Conversely, for every proper subspace V and every real linear functional $\lambda : V \to \mathbb{R}$ such that $\lambda(v) \leqslant p(v)$ for $v \in V$, for every

$x_1 \in X\backslash V$, and for every $\alpha \in [\alpha_{\min}(V,x_1), \alpha_{\max}(V,x_1)]$ we have to show the existence of an extension Λ, satisfying $\Lambda(x_1) = \alpha$ and

$$-p(-x) \leqslant \Lambda(x) \leqslant p(x) \quad \text{for all } x \in X.$$

Fix such V, λ, x_1, α. Let $V_1 = V + \mathbb{R}x_1$, and let $\lambda_1 : V_1 \to \mathbb{R}$ be the linear functional defined by $\lambda_1(v + tx_1) = \lambda(v) + t\alpha$. According to Lemma 5.87 it satisfies $\lambda_1(x) \leqslant p(x)$ for all $x \in V_1$. Applying this to $-x$ we also get $-p(-x) \leqslant \lambda_1(x)$ for all $x \in V_1$. Also $\lambda_1(x_1) = \alpha$.

For the construction of the desired linear functional on X we aim to use Zorn's lemma (Theorem 1.8). Consider the set $\mathfrak{S}$ of pairs $(\mathcal{V}, \Lambda)$ such that $\mathcal{V}$ is a linear subspace of X with $V_1 \subset \mathcal{V}$, $\Lambda : \mathcal{V} \to \mathbb{R}$ is a linear functional such that $\Lambda(v) = \lambda_1(v)$ for all $v \in V_1$. Note that $(V_1, \lambda_1) \in \mathfrak{S}$ and that $\Lambda(x_1) = \alpha$ for all $(\mathcal{V}, \Lambda) \in \mathfrak{S}$. On $\mathfrak{S}$ we define a partial order by setting

$$(\mathcal{V}, \Lambda) \preceq (\mathcal{V}', \Lambda')$$

if $\mathcal{V}$ is a subspace of $\mathcal{V}'$, $\Lambda'(v) = \Lambda(v)$ for all $v \in \mathcal{V}$, and

$$-p(-v') \leqslant \Lambda'(v') \leqslant p(v') \quad \text{for all } v' \in \mathcal{V}'.$$

We show that any chain in $\mathfrak{S}$ has an upper bound. If $\mathfrak{S}$ is empty then (V_1, λ_1) is an upper bound. If $\mathfrak{C} = \{(\mathcal{V}_\beta, \Lambda_\beta) : \beta \in \mathcal{I}\}$ is a nonempty totally ordered collection then $\bigcup_\beta \mathcal{V}_\beta$ is a linear subspace of X and we can define a linear functional Λ_{top} on $\bigcup_\beta \mathcal{V}_\beta$ by setting $\Lambda_{\text{top}}(v) = \Lambda_\beta(v)$ for $v \in \mathcal{V}_\beta$. Indeed one checks from the definition of $\preceq$ that the total ordering makes Λ_{top} well defined and that $(\bigcup_\beta \mathcal{V}_\beta, \Lambda_{\text{top}})$ belongs to $\mathfrak{S}$ and is an upper bound for $\mathfrak{C}$. Zorn's lemma tells us that $\mathfrak{S}$ has a maximal element $(\mathcal{V}_{\max}, \Lambda_{\max})$. It suffices to show that $\mathcal{V}_{\max} = X$; then $\Lambda = \Lambda_{\max}$ has the required properties.

We argue by contradiction and assume $\mathcal{V}_{\max} \neq X$. Then there exists $x_* \in X\backslash\mathcal{V}_{\max}$, and we can apply Lemma 5.87 to extend $\Lambda_{\max}$ to $\Lambda_{\text{ext}} : \mathcal{V}_{\max} + \mathbb{R}x_* \to \mathbb{R}$ so that $\Lambda_{\text{ext}}|_{\mathcal{V}_{\max}} = \Lambda_{\max}$ and $-p(-x) \leqslant \Lambda_{\text{ext}}(x) \leqslant p(x)$ for all $x \in \mathcal{V}_{\max} + \mathbb{R}x_*$. Thus $(\mathcal{V}_{\max} + \mathbb{R}x_*, \Lambda_{\text{ext}}) \in \mathfrak{S}$ and $(\mathcal{V}_{\max}, \Lambda_{\max}) \precneqq (\mathcal{V}_{\max} + \mathbb{R}x_*, \Lambda_{\text{ext}})$, in contradiction to the maximality of $(\mathcal{V}_{\max}, \Lambda_{\max})$. □

We now give a version of the Hahn–Banach extension theorem for complex vector spaces where control by sublinear functionals is replaced with control by seminorms.

We need a preliminary observation on $\mathbb{C}$-linear functionals and $\mathbb{R}$-linear functionals.

5.88. Observation. *Let X be a complex vector space. If λ is a $\mathbb{C}$-linear functional, then $u = \operatorname{Re}(\lambda)$ is $\mathbb{R}$-linear and $\lambda(x) = u(x) - iu(ix)$.*

Vice versa, if u is an $\mathbb{R}$-linear functional on X then $\lambda(x) = u(x) - iu(ix)$ defines a $\mathbb{C}$-linear functional, with $u = \operatorname{Re}(\lambda)$.

Proof. If λ is $\mathbb{C}$-linear then clearly u is $\mathbb{R}$-linear. Moreover let $v = \operatorname{Im}(\lambda)$, so that $\lambda(x) = u(x) + iv(x)$, then $i\lambda(x) = -v(x) + iu(x)$ and by the $\mathbb{C}$-linearity $i\lambda(x) = \lambda(ix) = u(ix) + iv(ix)$, hence $-v(x) = u(ix)$. Thus $\lambda(x) = u(x) - iu(ix)$.

Vice versa, if u is $\mathbb{R}$-linear and $\lambda(x) = u(x) - iu(ix)$, then the $\mathbb{C}$-valued functional λ is $\mathbb{R}$-linear. The $\mathbb{C}$-linearity follows from the $\mathbb{R}$-linearity and the calculation

$$\lambda(ix) = u(ix) - iu(-x) = u(ix) + iu(x) = i\big(u(x) - iu(ix)\big) = i\lambda(x). \qquad \square$$

5.89. Theorem. *Let X be a vector space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, and let p be a seminorm on X. Let V be a subspace of X and $\lambda : V \to \mathbb{K}$ a linear functional satisfying $|\lambda(v)| \leqslant p(v)$ for all $v \in V$. Then there exists a linear functional $\Lambda : X \to \mathbb{K}$ such that $\Lambda|_V = \lambda$ and $|\Lambda(x)| \leqslant p(x)$ for all $x \in X$.*

Proof. For $\mathbb{K} = \mathbb{R}$ this is a special case of Theorem 5.86 since $p(x) = p(-x)$ for seminorms. Let us assume $\mathbb{K} = \mathbb{C}$ and let $u = \operatorname{Re}(\lambda)$. By Theorem 5.86 we can extend u to an $\mathbb{R}$-linear functional on X satisfying $U|_V = u$ and $|U(x)| \leqslant p(x)$ for all $x \in X$. Then $\Lambda(x) = U(x) - iU(ix)$ defines a $\mathbb{C}$-linear

functional on X with $\Lambda|_V = \lambda$. For $x \in X$ choose an angle α such that $e^{i\alpha}\Lambda(x) = |\Lambda(x)|$ and we get by $\mathbb{C}$-linearity of Λ

$$|\Lambda(x)| = \Lambda(e^{i\alpha}x) = U(e^{i\alpha}x) \leqslant p(e^{i\alpha}x) = p(x)$$

since p is a seminorm on X. □

5.90. Example. Let $\mathfrak{c} \subset \ell^\infty$ be the space of all convergent sequences $(x_n)_{n=1}^\infty$. For $x \in \mathfrak{c}$ let $\lambda_0(x) = \lim_{n\to\infty} x_n$ which is a bounded linear functional on $\mathfrak{c}$ and satisfies $|\lambda_0(x)| \leqslant \|x\|_\infty$ for all $x \in \mathfrak{c}$. By Theorem 5.89 λ_0 extends to a linear functional λ on ℓ^∞ such that $\lambda_0 = \lambda|_{\mathfrak{c}}$ and such that $|\lambda(x)| \leqslant \|x\|_\infty$ for all $x \in \ell^\infty$. Argue that λ is not of the form $x \mapsto \sum_{k=1}^\infty x_k y_k$ with $y \in \ell^1$.

5.91. Corollary. *Let X be a normed space. For every $v \in X$ there is a bounded linear functional with $\Lambda(v) = \|v\|_X$ and $\|\Lambda\|_{X'} \leqslant 1$. Moreover, for every $x \in X$,*

$$\|x\| = \sup\{|\lambda(x)| : \lambda \in X', \|\lambda\|_{X^*} \leqslant 1\}.$$

Proof. Clearly the statement is true when $v = 0$ and Λ is the zero functional.

Let $v \neq 0$. Take V to be the one-dimensional space $\operatorname{span}(\{v\})$ and define $\lambda(cv) = c\|v\|$. Then λ is a linear functional on V with $\lambda(v) = \|v\|$ and $|\lambda(w)| \leqslant \|w\|_X$ for all $w \in V$. By Theorem 5.89 with p being the norm in X we may extend λ to a linear functional Λ on X satisfying $\Lambda(v) = \|v\|$ and $|\Lambda(x)| \leqslant \|x\|_X$ for all $x \in X$. By definition of the operator norm $\|\Lambda\|_{X'} = 1$. □

5.92. Corollary. *Let X be a normed vector space over $\mathbb{R}$ or $\mathbb{C}$ and let $x, \tilde{x} \in X$. Then*

$$\lambda(x) = \lambda(\tilde{x}) \text{ for all } \lambda \in X' \quad \text{if and only if} \quad x = \tilde{x}.$$

Denote by $V^\perp$ the space of linear functionals $\lambda \in X'$ such that $\lambda|_V = 0$, i.e. $\lambda(v) = 0$ for all $v \in V$. $V^\perp$ is called the *annihilator* of V. We can characterize best approximations from V via a statement about $V^\perp$.

5.93. Theorem. *Let V be a closed subspace of the normed space X, and let $x \in X\backslash V$. Then a vector $v_0 \in V$ is a best approximation to x from V if and only if there exists a bounded linear functional $\Lambda \in V^\perp$ such that $\|\Lambda\|_{X'} = 1$ and $\Lambda(x) = \|x - v_0\|$. Moreover,*

$$\operatorname{dist}(x, V) = \sup\big\{|\lambda(x)| : \lambda \in V^\perp, \|\lambda\|_{X'} = 1\big\}.$$

We note that by taking $V = \{0\}$ we can recover Corollary 5.91 from Theorem 5.93.

Proof of Theorem 5.93. If $x \notin V$ then $\operatorname{dist}(x, V) > 0$ since V is closed. If v_0 is not a best approximation then $\|x - v_0\| > \operatorname{dist}(x, V)$. For $\Lambda \in V^\perp$ with $\|\Lambda\|_{X'} = 1$ we have for all $v \in V$

$$|\Lambda(x)| = |\Lambda(x - v)| \leqslant \|x - v\|$$

so $|\Lambda(x)| \leqslant \operatorname{dist}(x, V) < \|x - v_0\|$. We have found that if $\Lambda \in V^\perp$ then $\Lambda(x) = \|x - v_0\|$ can only hold if v_0 is a best approximation to x from V.

Vice versa let v_0 be a best approximation to x from V, so $\|x - v_0\| = \operatorname{dist}(x, V) > 0$. Since $V \cap \mathbb{K}x = \{0\}$, we can define a bounded linear functional $\lambda \in (V \oplus \mathbb{K}x)'$ by $\lambda(cx + v) = c\|x - v_0\|_X$ for all $v \in V$. Note that $\lambda(v) = 0$ for $v \in V$. Now $\lambda(x - v_0) = \lambda(x) = \|x - v_0\|$ and so $\|\lambda\| \geqslant 1$. Also

$$|\lambda(cx + v)| = |c||\lambda(x)| = |c|\|x - v_0\| \leqslant |c|\|x + v/c\| = \|cx + v\|$$

so $\|\lambda\| \leqslant 1$, altogether $\|\lambda\| = 1$. By Theorem 5.89 we can extend λ to a bounded linear functional $\Lambda \in X'$ with $\|\Lambda\|_{X'} = 1$. Then $\Lambda \in V^\perp$ and $\Lambda(x) = \lambda(x) = \|x - v_0\|$. □

5.7.3. Banach limit functionals. To illustrate the Hahn–Banach theorem we study a family of examples of linear functionals on $\ell^\infty(\mathbb{N})$ which generalize the notion of limit of a sequence to all real ℓ^∞ sequences. In what follows let $\mathrm{sh} : \ell^\infty \to \ell^\infty$ be the *shift operator* which assigns to the sequence x the shifted sequence with

$$(\mathrm{sh}(x))_n = x_{n+1}, \quad n = 1, 2, \dots.$$

Naturally, any generalized limit functional should be linear, should assign to convergent sequences the classical limit and should assign the same value to x and $\mathrm{sh}(x)$. This motivates the following definition.

5.94. Definition. A linear functional Λ on ℓ^∞ is called *shift-invariant* if $\Lambda(\mathrm{sh}(x)) = \Lambda(x)$ for all $x \in \ell^\infty$. A linear functional $\Lambda : \ell^\infty \to \mathbb{R}$ is called a *Banach limit functional* (or shortly a *Banach limit*) if

(i) $\liminf_{n\to\infty} x_n \leqslant \Lambda(x) \leqslant \limsup_{n\to\infty} x_n$ for all $x \in \ell^\infty$ and
(ii) Λ is shift invariant.

The set of Banach limit functionals is denoted by BL.

Let $\mathfrak{c}$ be the space of convergent sequences. By condition (i) in the definition any $\Lambda \in \mathrm{BL}$ satisfies $\lim_{n\to\infty} x_n = \Lambda(x)$ for all $x \in \mathfrak{c}$. Also observe that $|\Lambda(x)| \leqslant \|x\|_\infty$ and thus any $\Lambda \in \mathrm{BL}$ belongs to $(\ell^\infty)'$ with $\|\Lambda\|_{(\ell^\infty)'} = 1$.

5.95. Lemma. *A shift invariant linear functional $\Lambda : \ell^\infty \to \mathbb{R}$ is a Banach limit functional if and only if*

(a) $\Lambda(\mathbb{1}) = 1$ *where* $\mathbb{1}_n = 1$ *for all* $n \in \mathbb{N}$.
(b) $\Lambda(x) \geqslant 0$ *if* $x_n \geqslant 0$ *for all* $n \in \mathbb{N}$.

Proof. If Λ is a Banach limit and $x_n \geqslant 0$ then $\liminf_{n\to\infty} x_n \geqslant 0$ so $\Lambda(x) \geqslant 0$, so (b) follows from (i) in the definition. (a) is clearly a special case of (i).

Vice versa, assume that (a), (b) hold. Let $\gamma = \limsup_{n\to\infty} x_n$. Given $\varepsilon > 0$ there is $N = N(\varepsilon)$ such that $x_n \leqslant \gamma + \varepsilon$ for $n \geqslant N$. If sh^N denotes the N-fold application of the shift operator we get $(\mathrm{sh}^N x)_n \leqslant \gamma + \varepsilon$ for all $n \geqslant 1$, and therefore, if $v_n = \gamma + \varepsilon - (\mathrm{sh}^N x)$ we have $v_n \geqslant 0$ for all n. By (b), $\Lambda(v_n) \geqslant 0$. By linearity of Λ and (a), $\Lambda(\mathrm{sh}^N x) \leqslant \gamma + \varepsilon$, and by shift invariance, $\Lambda(x) \leqslant \gamma + \varepsilon$. Since ε was arbitrary we obtain $\Lambda(x) \leqslant \limsup_{n\to\infty} x_n$. Applying this to $-x$ we get $-\Lambda(x) = \Lambda(-x) \leqslant \limsup_{n\to\infty}(-x_n) = -\liminf_{n\to\infty} x_n$, hence $\Lambda(x) \geqslant \liminf_{n\to\infty} x_n$. This proves (i) in Definition 5.94. □

We now use the Hahn–Banach theorem to show that the set BL is not empty.

5.96. Theorem. *Banach limit functionals exist.*

Proof. Define $\lambda : \mathfrak{c} \to \mathbb{R}$ by $\lambda(x) = \lim_{n\to\infty} x_n$. To get a linear functional that extends λ to ℓ^∞ one may apply the Hahn–Banach theorem with domination by the sublinear functional $p_1(x) = \sup_n |x_n|$, but this would not give us the control in condition (ii) of the definition of Banach limit. To achieve this, we can work with domination by $p_2(x) = \limsup_{n\to\infty} x_n$. Note that $-p_2(-x) = \liminf_{n\to\infty} x_n$. p_2 is a sublinear functional (*cf.* Exercise 1.100). However control by p_2 is still not enough to prove the shift invariance, i.e. property (ii) in Definition 5.94. Instead we work with a refinement p_3, that for $x \in \ell^\infty$ is defined by

$$p_3(x) = \limsup_{n\in\mathbb{N}} \frac{1}{n} \sum_{j=1}^{n} x_j.$$

Notice that p_3 is a sublinear functional on ℓ^∞. By the Hahn–Banach theorem we get an extension Λ of λ with $-p_3(-x) \leqslant \Lambda(x) \leqslant p_3(x)$. We have $-p_3(-x) = \liminf_{n\in\mathbb{N}} \frac{1}{n} \sum_{j=1}^{n} x_j$. The two properties (a), (b) in Lemma 5.95 are thus satisfied and it remains to check that Λ is shift invariant. For $x \in \ell^\infty$ we consider

$$\sigma_n = \frac{1}{n} \sum_{k=1}^{n} \big(\mathrm{sh}(x) - x\big)_k = \frac{1}{n}(x_{n+1} - x_1).$$

Observe $|\sigma_n| \leqslant 2\|x\|_\infty/n$ and thus $\sigma = (\sigma_n)_{n=1}^\infty \in \mathfrak{c}$ with $\lim_{n\to\infty} \sigma_n = 0$. This implies $p_3(\mathrm{sh}(x) - x) = 0$ and $p_3(x - \mathrm{sh}(x)) = 0$ and consequently $\Lambda(\mathrm{sh}(x) - x) = 0$. By linearity $\Lambda(\mathrm{sh}(x)) = \Lambda(x)$ and we have shown the existence of a Banach limit functional. □

5.97. Definition.

(i) For every bounded sequence of real numbers $x = (x_n)_{n=1}^\infty$ we define the *Banach limit set* of x by $\mathrm{BL}(x) = \{\Lambda(x) : \Lambda \in \mathrm{BL}\}$.

(ii) A bounded sequence $x = (x_n)$ of real numbers is *almost convergent* (in the sense of Lorentz) if $\mathrm{BL}(x) = \{\alpha\}$ for some $\alpha \in \mathbb{R}$. In this case we also say that (x_n) is almost convergent to α.

A natural question is to characterize all $x \in \ell^\infty$ which are almost convergent. This is addressed by a theorem of Lorentz. For its proof it will be crucial to work with a different controlling sublinear functional than in the proof of Theorem 5.96 (indeed that argument misses some Banach limit functionals).

5.98. Lemma. *Let $x = (x_n) \in \ell^\infty$ and*

$$\rho^*(x,n) = \sup_{k\geqslant 0} \frac{1}{n} \sum_{j=1}^n x_{j+k}, \qquad \rho_*(x,n) = \inf_{k\geqslant 0} \frac{1}{n} \sum_{j=1}^n x_{j+k}\,. \tag{5.7a}$$

Then:

(i) For every $x \in \ell^\infty$ the limits

$$\rho^*(x) = \lim_{n\to\infty} \rho^*(x,n), \qquad \rho_*(x) = \lim_{n\to\infty} \rho_*(x,n), \tag{5.7b}$$

exist and we have $\rho_*(x) = -\rho^*(-x)$.

(ii) ρ^* is a sublinear functional. For all $\Lambda \in \mathrm{BL}$, and all $x \in \ell^\infty$,

$$\rho_*(x) \leqslant \Lambda(x) \leqslant \rho^*(x).$$

Proof. The sequence $\rho^*(x,n)$ is not monotone decreasing, but $\rho^*(x,\ell m) \leqslant \rho^*(x,m)$. Indeed

$$\rho^*(x,\ell m) = \sup_{k\geqslant 0} \frac{1}{\ell m} \sum_{j=1}^{\ell m} x_{j+k} = \sup_{k\geqslant 0} \frac{1}{\ell} \sum_{\nu=1}^{\ell} \frac{1}{m} \sum_{r=1}^{m} x_{(\nu-1)m+r+k} \leqslant \sup_{k'\geqslant 0} \frac{1}{m} \sum_{r=1}^{m} x_{r+k'} = \rho^*(x,m).$$

For any $\ell \in \mathbb{N}$, $m \in \mathbb{N}$, $r \in \{1,\dots,m\}$

$$\begin{aligned}(\ell m + r)\rho^*(\ell m + r) &= \sup_{k\geqslant 0} \sum_{j=1}^{\ell m+r} x_{j+k} \leqslant \sup_{k\geqslant 0} \sum_{j=1}^{\ell m} x_{j+k} + \sup_{k\geqslant 0} \sum_{j=\ell m+1}^{\ell m+r} x_{j+k} \\ &\leqslant \ell m \rho^*(x,\ell m) + \sup_{k'\geqslant 0} \sum_{i=1}^{r} x_{i+k'} \leqslant \ell m\, \rho^*(x,m) + r\rho^*(x,r).\end{aligned}$$

We divide by $\ell m + r$ and consider the behavior for large ℓ. We get for any m, $1 \leqslant r \leqslant m$,

$$\limsup_{\ell\to\infty} \rho^*(\ell m + r) \leqslant \rho^*(x,m).$$

This implies $\limsup_{n\to\infty} \rho^*(x,n) \leqslant \rho^*(x,m)$ for all m, and therefore

$$\limsup_{n\to\infty} \rho^*(x,n) \leqslant \liminf_{m\to\infty} \rho^*(x,m)$$

which gives the existence of the limit defining $\rho^*(x)$. Since $\rho_*(x,n) = -\rho^*(-x,n)$ we get $\lim_{n\to\infty} \rho_*(x,n) = -\rho^*(-x)$ and part (i) is proved.

Now let $\Lambda \in \mathrm{BL}$. Note that $\inf w_k \leqslant \Lambda(w) \leqslant \sup w_k$ for any $w \in \ell^\infty$ (apply Λ to the nonnegative sequences $(\sup_\nu w_\nu - w_k)_{k=1}^\infty$ and $(w_k - \inf_\nu w_\nu)_{k=1}^\infty$). Consider for fixed n,

$$w_k^{(n)} = \frac{1}{n}\sum_{j=1}^{n} x_{j+k}.$$

Then by linearity and shift invariance of Λ we have $\Lambda(x) = \Lambda(w^{(n)})$ for all n, and this implies

$$\Lambda(x) \leqslant \sup_k w_k^{(n)} = \rho^*(x,n)$$

for all $n \in \mathbb{N}$. Letting $n \to \infty$ we obtain $\Lambda(x) \leqslant \rho^*(x)$. Replacing x by $-x$ we also get

$$\Lambda(x) \geqslant -\rho^*(-x) = \rho_*(x).$$

□

5.99. Theorem (Lorentz). *Let $y = (y_n)$ be a real bounded sequence, and let $\rho_*(y)$, $\rho^*(y)$ be as in* (5.7b). *Then*

(i) $\mathrm{BL}(y) = [\rho_*(y), \rho^*(y)]$.

(ii) A bounded sequence (y_n) is almost convergent to α if and only if

$$\lim_{n\to\infty} \sup_k \Big|\frac{1}{n}\sum_{j=1}^{n} y_{j+k} - \alpha\Big| = 0\,.$$

Proof. We use the Hahn–Banach theorem with domination by the sublinear functional ρ^*. As before consider $\lambda : \mathfrak{c} \to \mathbb{R}$, with $\lambda(y) = \lim_{n\to\infty} y_n$ for $y \in \mathfrak{c}$. Then $\rho_*(y) = \lambda(y) = \rho^*(y)$ for $y \in \mathfrak{c}$, and the assertion on $\mathrm{BL}(y)$ is true for $y \in \mathfrak{c}$.

Fix $y \in \ell^\infty \backslash \mathfrak{c}$. We define $\alpha_{\min}, \alpha_{\max}$ as in (5.5) and claim that

$$\begin{aligned} \rho^*(y) &= \alpha_{\max}(\mathfrak{c}, y) = \inf\{\rho^*(v+y) - \lambda(v) : v \in \mathfrak{c}\} \\ \rho_*(y) &= \alpha_{\min}(\mathfrak{c}, y) = \sup\{\lambda(v) - \rho^*(v-y) : v \in \mathfrak{c}\} \end{aligned} \tag{5.8}$$

To see this we use $\lambda(v) = \lim_{n\to\infty} v_n$ for $v \in \mathfrak{c}$ which also implies that for every $\varepsilon > 0$ there is N such that $|\frac{1}{n}\sum_{j=1}^{n} v_{j+k} - \lambda(v)| < \varepsilon$ for all $n \geqslant N$ (and all $k \geqslant 0$). Thus for $y \in \ell^\infty$, $v \in \mathfrak{c}$ we get $\rho^*(y+v) = \rho^*(y) + \lambda(v)$. Similarly also $-\rho^*(v-y) = -\lambda(v) - \rho^*(-y) = -\lambda(v) + \rho_*(y)$ and (5.8) holds.

By Theorem 5.86, for every

$$\alpha \in [\alpha_{\min}(\mathfrak{c}, y), \alpha_{\max}(\mathfrak{c}, y)] = [\rho_*(y), \rho^*(y)]$$

there exists a linear functional $\Lambda \in (\ell^\infty)'$ such that $\Lambda|_{\mathfrak{c}} = \lambda$, $\rho_*(x) \leqslant \Lambda(x) \leqslant \rho^*(x)$ for all $x \in \ell^\infty$ and $\Lambda(y) = \alpha$. To verify that $\Lambda \in \mathrm{BL}$ we argue as in the proof of Theorem 5.96. We have $\Lambda(\mathbb{1}) = 1$ and if $x_n \geqslant 0$ for all $n \in \mathbb{N}$ we also get $\rho_*(x) \geqslant 0$, so (a), (b) in Lemma 5.95 are satisfied. Note that $\rho^*(\mathrm{sh}(x) - x, n) = n^{-1}\sup_k(x_{n+k+1} - x_{k+1})$, so $|\rho^*(\mathrm{sh}(x) - x, n)| \leqslant \frac{2}{n}\|x\|_\infty$ and thus $\rho^*(\mathrm{sh}(x) - x) = 0 = \rho_*(x - \mathrm{sh}(x))$. This implies $\Lambda(\mathrm{sh}(x) - x) = 0$ and thus Λ is shift invariant, hence a Banach limit functional. We have also shown $\mathrm{BL}(y) \supset [\rho_*(y), \rho^*(y)]$. By Lemma 5.98, the opposite inclusion $\mathrm{BL}(y) \subset [\rho_*(y), \rho^*(y)]$ holds and thus part (i) is proved.

Regarding part (ii), the condition that $\lim_{n\to\infty} \sup_{k\geqslant 0} |\frac{1}{n}\sum_{j=1}^{n} y_{j+k} - \alpha| = 0$ is equivalent to the condition that for every $\varepsilon > 0$ there exists an N_ε such that for all $n \geqslant N_\varepsilon$

$$\alpha - \varepsilon \leqslant \frac{1}{n}\sum_{j=1}^{n} y_{j+k} \leqslant \alpha + \varepsilon \text{ for all } k \geqslant 0.$$

Equivalently, for every $\varepsilon > 0$ there is N_ε such that $\alpha - \varepsilon \leqslant \rho_*(y,n) \leqslant \rho^*(y,n) \leqslant \alpha + \varepsilon$ for $n \geqslant N_\varepsilon$. By part (i) this is equivalent to $\rho^*(y) = \alpha = \rho_*(y)$, i.e. to (y_n) being almost convergent to α. □

Remark. Here we followed Sucheston [**78**].

5.100. Exercise. (i) Show that the sequence $(\cos(n\theta))_{n=1}^{\infty}$ is almost convergent to some $\alpha(\theta)$ and determine $\alpha(\theta)$. Do the same for $(\sin n\theta)_{n=1}^{\infty}$.

(ii) Let $(x_n)_{n=1}^{\infty}$ be periodic with period $q \in \mathbb{N}$, i.e. $x_{(m-1)q+k} = x_k$ for all $m \in \mathbb{N}$ and $k = 1, \dots, q$. Show that (x_n) is almost convergent to $\frac{1}{q}\sum_{k=1}^{q} x_k$.

(iii) Prove that if (x_n) is almost convergent to α then $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k = \alpha$.

(iv) For given $b \leqslant \alpha \leqslant c$ construct a sequence (x_n) such that $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k = \alpha$ and $\mathrm{BL}(x) = [b, c]$.

(v) Show by example that not every Banach limit functional satisfies the quantitative control by p_3 in the proof of Theorem 5.96.

5.8. Further exercises

5.101. Exercise. Let $C(\mathbb{R})$ be the set of continuous functions on $\mathbb{R}$. Let $\omega(t) = \frac{t}{1+t}$ for $t \geqslant 0$. Define

$$d(f,g) = \sum_{k=0}^{\infty} 2^{-k}\, \omega\big(\sup_{x\in[-k,k]} |f(x) - g(x)|\big).$$

(i) Show that d is a well-defined metric.
(ii) Show that $C(\mathbb{R})$ is complete with this metric.
(iii) Show that there exists no norm $\|\cdot\|$ on $C(\mathbb{R})$ such that $d(f,g) = \|f - g\|$.

5.102. Exercise. Let $A \in \mathbb{R}^{m\times n}$. Define for $x \in \mathbb{R}^n$, $\|x\|_2 = (\sum_{j=1}^{n} |x_j|^2)^{1/2}$ (Euclidean norm) and likewise for $y \in \mathbb{R}^m$, $\|y\|_2 = (\sum_{i=1}^{m} |y_i|^2)^{1/2}$. Let

$$\|A\|_{2\to 2} = \sup_{\|x\|_2=1} \|Ax\|_2.$$

Observe that $A^{\intercal}A$ is a symmetric $m \times m$ matrix and has only non-negative eigenvalues. Denote the largest eigenvalue of $A^{\intercal}A$ by ρ. Prove that $\|A\|_{2\to 2} = \sqrt{\rho}$.

5.103. Exercise. Let $A \in \mathbb{R}^{n\times n}$ and define the *Hilbert–Schmidt norm*

$$\|A\|_{\mathrm{HS}} = \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} |A_{ij}|^2\Big)^{1/2}.$$

This is a norm on $\mathbb{R}^{n\times n}$. Prove the following properties for all $A, B \in \mathbb{R}^{n\times n}$:
(i) $\|AB\|_{\mathrm{HS}} \leqslant \|A\|_{\mathrm{HS}}\|B\|_{\mathrm{HS}}$
(ii) $\|A\|_{\mathrm{HS}} = \sqrt{\mathrm{trace}(A^{\intercal}A)}$
(iii) $\|UA\|_{\mathrm{HS}} = \|AU\|_{\mathrm{HS}} = \|A\|_{\mathrm{HS}}$ for all orthogonal $U \in \mathbb{R}^{n\times n}$
(iv) $\|A\|_{2\to 2} \leqslant \|A\|_{\mathrm{HS}} \leqslant \sqrt{n}\|A\|_{2\to 2}$.

Remark. For $A \in \mathbb{R}^{m\times n}$ one defines $\|A\|_{\mathrm{HS}} = \Big(\sum_{i=1}^{m}\sum_{j=1}^{n} |A_{ij}|^2\Big)^{1/2}$. Find suitable extensions of the above statements.

5.104. Exercise. If X is finite-dimensional, then X' is isomorphic to X.

5.105. Exercise. Let V be the vector space of continuous functions on $[0,1]$ equipped with the norm $\|f\| = \int_0^1 |f(t)|dt$. Define a linear map $T : V \to V$ by

$$Tf(x) = \int_0^x f(t)dt.$$

Show that T is well-defined and bounded and determine the value of $\|T\|_{\mathrm{op}}$.

5.106. Exercise. For $k \in \mathbb{N}$ let $C^k([a,b])$ be the space of k-times differentiable functions, with $f^{(k)}$ continuous, as defined in Exercise 4.179. Prove that

$$\|f\|_{C^k} = \sum_{j=0}^{k} \sup_{x\in[a,b]} |f^{(j)}(x)|$$

$$\|f\|_{C^k_{mod}} = \sup_{x\in[a,b]} |f(x)| + \sup_{x\in[a,b]} |f^{(k)}(x)|$$

are equivalent norms on $C^k([a,b])$. Prove that $C^k([a,b])$ is a Banach space.

5.107. Exercise. For a nonnegative integer k and $0 < s \leqslant 1$, the space $C^{k,s}([a,b])$ consists of functions $f \in C^k([a,b])$ such that the kth derivative $f^{(k)}$ belongs to $C^{0,s}([a,b])$ (see Exercise 5.6). Show that this defines a Banach space with the norm

$$\|f\|_{C^{k,s}} = \|f\|_{C^k} + \|f^{(k)}\|_{C^{0,s}}.$$

5.108. Exercise. Is there a constant C such that for all continuous functions f on $[0,2]$ the inequality

$$\int_0^2 |f(t)|dt \leqslant C \max_{0\leqslant x\leqslant 2} |f(x)|$$

holds? Is there a constant C such that for all continuous functions f on $[0,2]$ the reverse inequality

$$\max_{0\leqslant x\leqslant 2} |f(x)| \leqslant C \int_0^2 |f(t)|dt$$

holds? The expressions on both sides of these inequalities define norms on $C([0,2])$. Are these equivalent norms?

5.109. Exercise. Prove or disprove: There is a constant C_n such that for all polynomials P of degree $\leqslant n$ we have

$$\max_{0\leqslant x\leqslant 2} |P(x)| \leqslant C_n \int_0^2 |P(t)|dt\,.$$

What about the analogous question concerning the inequality

$$\max_{0\leqslant x\leqslant 200} |P(x)| \leqslant C_n \int_0^{10^{-10}} |P(t)|dt\,?$$

5.110. Exercise. Consider the space ℓ^1 of absolutely summable sequences of complex numbers. Let $p, q \in [1,\infty]$ with $p \neq q$. Then $\|\cdot\|_p$ and $\|\cdot\|_q$ are norms on ℓ^1 (recall that $\|a\|_p = (\sum_{n=1}^{\infty} |a_n|^p)^{1/p}$ for $p \in [1,\infty)$ and $\|a\|_\infty = \sup_{n\in\mathbb{N}} |a_n|$). Show that $\|\cdot\|_p$ and $\|\cdot\|_q$ are not equivalent.

5.111. Exercise. Consider $\mathcal{F} = \{f_N : N \in \mathbb{N}\} \subset C([0,1])$ with

$$f_N(x) = \sum_{n=0}^{N} b^{-sn} \sin(b^n x),$$

where $0 < s < 1$ and $b > 1$ are fixed.
(a) Show that $\mathcal{F}$ is relatively compact in $C([0,1])$.
(b) Show that $\mathcal{F}'$ is not a bounded subset of $C([0,1])$.
(c) Show that there exists $c > 0$ such that for all $x, y \in \mathbb{R}$ and $N \in \mathbb{N}$,

$$|f_N(x) - f_N(y)| \leqslant c|x-y|^s.$$

That is, the functions f_N form a bounded subset in the Banach space $C^{0,s}([0,1])$ defined in Exercise 5.6.

5.112. Exercise. Let $p < \infty$. The purpose of this exercise is to prove a theorem of Fréchet that characterizes compactness in ℓ^p. Let $\mathcal{F} \subset \ell^p$.

(i) Assume that $\mathcal{F}$ is bounded and *equisummable* in the following sense: for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$\sum_{n=N}^{\infty} |a_n|^p < \varepsilon \text{ for all } a \in \mathcal{F}.$$

Then show that $\mathcal{F}$ is totally bounded.

(ii) Conversely, assume that $\mathcal{F}$ is totally bounded. Then show that it is equisummable in the above sense.

5.113. Exercise. For $x, y \in \mathbb{R}^n$ define

$$\rho_p(x, y) = \sum_{i=1}^{n} |x_i - y_i|^p.$$

(i) Let $0 < p \leqslant 1$. Prove that ρ_p is a metric on $\mathbb{R}^n$.

(ii) Let $1 < p < \infty$. Prove that ρ_p is not a metric on $\mathbb{R}^n$.

(iii) Let $0 < p < 1$, $n \geqslant 2$. Prove that $\rho_p^{1/p}$ is not a metric on $\mathbb{R}^n$.

(iv) Let $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. Prove that if $0 < p < 1$, $n \geqslant 2$, neither $\|x\|_p$ nor $\|x\|_p^p$ define a norm on $\mathbb{R}^n$.

5.114. Exercise. Let $x \in \mathbb{R}^n$. Define $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $0 < p < \infty$ and $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$.
(i) Show that $\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$.
(ii) Show that $\lim_{p\to 0} \|x\|_p$ exists and determine its value (here we also allow ∞ as a limit).

5.115. Exercise (L^p norms). Let $p \in [1, \infty)$.

(i) Show that the L^p *norm*

$$\|f\|_p = \Big(\int_0^1 |f(t)|^p dt \Big)^{1/p}$$

defines a norm on the space of continuous functions on $[0, 1]$. You may choose to do part (iv) below first.

(ii) Let $\alpha < 1/p$ and define

$$f_n(t) = \begin{cases} t^{-\alpha} & \text{for } n^{-1} \leqslant t \leqslant 1, \\ n^{\alpha} & \text{for } 0 \leqslant t < n^{-1}. \end{cases}$$

Show that $\{f_n\}_{n=0}^{\infty}$ is a Cauchy sequence with respect to the norm $\|\cdot\|_p$. Show that it is not a Cauchy sequence with respect to the usual sup-norm on $C([0, 1])$.

(iii) Show that the space of continuous functions on $[0, 1]$ with norm $\|\cdot\|_p$ is not complete.

(iv) Show that if f is Riemann integrable then $|f|^p$ is also Riemann integrable. Show that the L^p norm defines a seminorm on the space of Riemann integrable functions on $[0, 1]$, but not a norm.

(v) Show for f, g Riemann integrable on $[0, 1]$ and $p \in (1, \infty)$, $\frac{1}{p} + \frac{1}{p'} = 1$,

$$\Big| \int_0^1 f(t)g(t)dt \Big| \leqslant \|f\|_p \|g\|_{p'}.$$

(vi) Show for f Riemann integrable on $[0, 1]$,

$$\|f\|_{p_1} \leqslant \|f\|_{p_2} \text{ if } p_1 \leqslant p_2.$$

5.116. Exercise. Let $1 \leqslant p \leqslant \infty$, $V = \mathbb{R}^n$, $W = \mathbb{R}^m$ and use on V the norm $\|\cdot\|_1$ and on W the norm $\|\cdot\|_p$. Let $T : V \to W$ be defined by $T(x) = Ax$ where A is an $m \times n$ matrix. Show that the operator

norm of T is given by

$$\|T\|_{1\to p} = \max_{j=1,\dots n} \Big(\sum_{i=1}^{m} |a_{ij}|^p \Big)^{1/p}.$$

5.117. Exercise. Let $1 \leqslant p \leqslant \infty$, $\frac{1}{p} + \frac{1}{p'} = 1$, $V = \mathbb{R}^n$, $W = \mathbb{R}^m$ and use on V the norm $\|\cdot\|_p$ and on W the norm $\|\cdot\|_\infty$. Let $T : V \to W$ be defined by $T(x) = Ax$ where A is an $m \times n$ matrix. Show that the operator norm of T is given by

$$\|T\|_{p\to\infty} = \max_{i=1,\dots m} \Big(\sum_{j=1}^{n} |a_{ij}|^{p'} \Big)^{1/p'}.$$

5.118. Exercise. Let V be a normed space and $(\widetilde{V}, \widetilde{d})$ its completion of (V, d) as a metric space. Show that there is a vector space structure on $\widetilde{V}$, and a norm on $\widetilde{V}$ whose restriction to V is $\|\cdot\|_V$, such that $\widetilde{V}$ is a Banach space.

5.119. Exercise. Let $(V, \langle\cdot,\cdot\rangle)$ be a pre-Hilbert space. Let $\widetilde{V}$ be the completion of V as a normed space. Show that there is a scalar product on $\widetilde{V}$ whose restriction to V is the original scalar product, so that $\widetilde{V}$ becomes a Hilbert space. Show that there is an isometry $\imath : V \to \widetilde{V}$ such that $\overline{\imath(V)} = \widetilde{V}$.

We introduce a class of bounded linear operators on Hilbert spaces which generalizes the notion of a self-adjoint operator on a finite-dimensional Euclidean space which is given by a *Hermitian matrix*, i.e. an $n \times n$ matrix A satisfying $A = \overline{A}^{\mathsf{T}}$.

5.120. Exercise. Let H be a Hilbert space. An operator $T \in L(H)$ is called *selfadjoint* if $T = T^*$.

Consider $H = \ell^2(\mathbb{N})$. Assume that for $j \in \mathbb{N}$, $k \in \mathbb{N}$ the complex numbers $a_{j,k}$ satisfy $\sup_{j\in\mathbb{N}} \sum_{k=1}^{\infty} |a_{j,k}| < \infty$, $\sup_{k\in\mathbb{N}} \sum_{j=1}^{\infty} |a_{j,k}| < \infty$. For $f = (f(j))_{j\in\mathbb{N}} \in \ell^2(\mathbb{N})$ set

$$Tf(j) = \sum_{k=1}^{\infty} a_{j,k} f(k).$$

Show $Tf \in \ell^2(\mathbb{N})$ and that $T : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ is a bounded operator on $\ell^2(\mathbb{N})$. Determine T^* and prove that T is self-adjoint if and only if $a_{j,k} = \overline{a_{k,j}}$ for all $(j, k) \in \mathbb{N}^2$.

5.121. Exercise. Let C be a nonempty, closed, and convex subset of a Hilbert space H. Then C contains a unique element of smallest norm.

5.122. Exercise. Let X be a normed vector space, and X'' the dual space of the dual space. We define a natural map $X \to X''$, the *evaluation map*:

$$\mathrm{ev} : X \longrightarrow X''$$

For $x \in X$ let $\mathrm{ev}(x)$ be the linear map $X' \to \mathbb{C}$ defined by sending $v \in X'$ to $v(x) \in \mathbb{C}$. That is, $\mathrm{ev}(x)(v) = v(x)$.

(i) Verify that $\mathrm{ev}(x)$ is a bounded linear functional on X'. Determine $\|\mathrm{ev}(x)\|_{X''}$.

(ii) Prove that ev is a bounded linear map. Determine the operator norm $\|\mathrm{ev}\|_{X\to X''}$.

(iii) Prove that ev is injective.

5.123. Exercise. Let $\mathrm{ev} : X \to X''$ be as in Exercise 5.122. We call X *reflexive* if ev is surjective.

(i) Show that finite dimensional spaces are reflexive.

(ii) Show that ℓ^p is reflexive for $1 < p < \infty$ but $\mathfrak{c}_0$ and ℓ^1 are not reflexive.

5.124. Exercise. Let Y be a normed space. A function $f : [a, b] \to Y$ is bounded if $\sup_x \|f(x)\|_Y < \infty$. For any partition $P = \{x_0 < \cdots < x_N\}$ of $[a, b]$ we define the P-variation of f by

$$\Omega(f, P) = \sum_{i=1}^{N} (x_i - x_{i-1}) \sup \left\{ \|f(\xi) - f(\tilde{\xi})\|_Y : x_{i-1} \leqslant \xi, \tilde{\xi} \leqslant x_i \right\}.$$

A *tagged partition* is a partition $P = \{x_0 < \cdots < x_N\}$ together with nodes $\xi = (\xi_1, \ldots, \xi_N)$ so that $x_{i-1} \leqslant \xi_i \leqslant x_i$ for $i = 1, \ldots, N$. The width of a partition is $\Delta(P) = \max_i(x_i - x_{i-1})$. For every tagged partition (P, ξ) define the Riemann sum

$$S_{P,\xi}(f) = \sum_{i=1}^{N}(x_i - x_{i-1})f(\xi_i).$$

We say that a bounded function $f : [a, b] \to Y$ is *Riemann integrable* if for every $\varepsilon > 0$ there exists a partition P such that $\Omega(f, P) < \varepsilon$. We let $\mathcal{R}([a, b]; Y)$ be the class of all Riemann integrable functions $f : [a, b] \to Y$.Prove:

(i) If P' is a refinement of P, then $\Omega(f, P') \leqslant \Omega(f, P)$.

(ii) If $Y = \mathbb{R}$ (with the usual norm $|\cdot|$), then $\Omega(f, P) = U(f, P) - L(f, P)$.

(iii) If (P, ξ) and (P', ξ') are tagged partitions, and if P' is a refinement of P, then

$$\|S_{P,\xi}(f) - S_{P',\xi'}(f)\|_Y \leqslant \Omega(f, P).$$

(iv) If (P, ξ) and (P', ξ') are tagged partitions, then

$$\|S_{P,\xi}(f) - S_{P',\xi'}(f)\|_Y \leqslant \Omega(f, P) + \Omega(f, P')$$

(v) $\mathcal{R}([a, b]; Y)$ is a vector space.

(vi) Let $f : [a, b] \to Y$ be continuous. Prove that $f \in \mathcal{R}([a, b]; Y)$.

(vii) Let $f : [a, b] \to Y$ be bounded and continuous on the complement of a Lebesgue null set (see Definition 4.121). Prove that $f \in \mathcal{R}([a, b]; Y)$.

5.125. Exercise. To define for a Riemann integrable function the integral we need completeness of Y. Let Y be a Banach space, and $f \in \mathcal{R}([a, b]; Y)$.

(i) Prove: There is a unique $v \in Y$ such that

$$\|S_{P,\xi}(f) - v\|_Y \leqslant \Omega(f, P)$$

for all partitions P of $[a, b]$. v is called the *Riemann integral* of f and denoted by $\int_a^b f$.

(ii) Prove that for each $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\Big\|S_{P,\xi}(f) - \int_a^b f\Big\|_Y < \varepsilon$$

for all tagged partitions (P, ξ) with $\Delta(P) < \delta$.

5.126. Exercise. Let Y, Y_1, Y_2 be Banach spaces.

(i) If $f \in \mathcal{R}([a, b]; Y)$, then $\|f\| : x \mapsto \|f(x)\|_Y$ belongs to $\mathcal{R}([a, b], \mathbb{R})$ and

$$\Big\| \int_a^b f\Big\|_Y \leqslant \int_a^b \|f(x)\|_Y dx.$$

(ii) If $f \in \mathcal{R}([a, b]; Y_1)$ and $T : Y_1 \to Y_2$ is a bounded linear operator, then $Tf \in \mathcal{R}([a, b]; Y_2)$ and

$$\int_a^b Tf = T\Big(\int_a^b f\Big).$$

(iii) Provide the formulas for Riemann integrable $\mathbb{R}^n$-valued or $\mathbb{C}^n$-valued functions.

CHAPTER 6

Differential calculus

In this chapter we shall develop the differential calculus of functions depending on several variables, or more generally the differential calculus of functions on normed vector spaces, with values in normed vector spaces. Our emphasis in this text is on *real analysis*, and therefore we shall restrict ourselves to vector spaces with the scalar field $\mathbb{R}$. The field of complex numbers, or vector spaces of complex-valued functions can also be considered as vector spaces over $\mathbb{R}$.

6.1. Derivatives

Recall that a real function f on an open interval I is called differentiable at $a \in I$ if $\lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$ exists. In other words, f is differentiable at a if there exists a number $T \in \mathbb{R}$ such that

$$\lim_{h\to 0} \frac{|f(a+h) - f(a) - Th|}{|h|} = 0.$$

This expresses that at the point a the function f is well approximated by the graph of a line which is given by a constant, $f(a)$, plus a linear term, $Th = T(a+h) - Ta$. In that case we denote that real number T by $f'(a)$. A real number can also be understood as a linear map $\mathbb{R} \to \mathbb{R}$, i.e. we associate to $T \in \mathbb{R}$ the map $a \mapsto Ta$, given by multiplication with T. Interpreting the derivative at a given point a as a linear map, we can formulate the definition of derivative in the general setting of normed vector spaces.

6.1.1. The Fréchet derivative. In what follows we are working with functions between normed vector spaces, with the scalar field $\mathbb{R}$. Recall that given two normed spaces X and Y we denote by $L(X,Y)$ the space of bounded linear transformations from X to Y.

6.1. Definition. Let X, Y be normed vector spaces over $\mathbb{R}$. Let $U \subset X$ be open. A map $f : U \to Y$ is called *Fréchet differentiable* (or just *differentiable*) at $a \in U$ if there exists $T \in L(X,Y)$ such that

$$\lim_{h\to 0} \frac{\|f(a+h) - f(a) - Th\|_Y}{\|h\|_X} = 0. \tag{6.1}$$

In that case we call T the *(Fréchet) derivative* of f at a and write $T = Df|_a$. f is called *(Fréchet) differentiable* in U if it is differentiable at every point $x \in U$.

Before moving on, we need to verify that $Df|_a$ is well defined, that is, T in (6.1) is uniquely determined by f and a. Suppose $T, \widetilde{T} \in L(X,Y)$ both satisfy (6.1). Then, by the triangle inequality,

$$\|Th - \widetilde{T}h\|_Y \leqslant \|f(a+h) - f(a) - Th\|_Y + \|f(a+h) - f(a) - \widetilde{T}h\|_Y .$$

Thus, by (6.1), $\|Th - \widetilde{T}h\|_Y / \|h\|_X \to 0$ as $h \to 0$. Pick any $u \neq 0$ and choose $h = tu$, letting $t \to 0$. Since by linearity and homogeneity of norms

$$\frac{\|T(tu) - \widetilde{T}(tu)\|_Y}{\|tu\|_X} = \frac{\|Tu - \widetilde{T}u\|}{\|u\|_X}$$

is independent of t we see that $Tu - \widetilde{T}u = 0$. Since $u \neq 0$ was arbitrary we get $T = \widetilde{T}$.

We can use little-o notation to restate the definition of derivatives in an equivalent way:

6.2. Observation. *f is Fréchet differentiable at a, with derivative $Df|_a$, if and only if*

$$f(a+h) = f(a) + Df|_a h + o(\|h\|) \quad (\text{as } h \to 0).$$

Remarks (Terminology).

(i) As in the one-dimensional setting, the Fréchet derivative map $Df|_a$ provides a linear approximation to $f(a+h)-f(a)$ when $\|h\|$ is small. One also refers to the linear operator $Df|_a$ as the *linearization* of the map f at a.

(ii) In the case $Y = \mathbb{R}$, $Df|_a$ is a linear form and also called the *differential* of f at a, sometimes denoted by df_a.

6.3. Example. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x_1, x_2) = x_1 \cos(x_2)$. We claim that f is Fréchet differentiable at every $a = (a_1, a_2) \in \mathbb{R}^2$. Indeed, let $h = (h_1, h_2) \in \mathbb{R}^2 \backslash \{0\}$. Then

$$f(a+h) = (a_1 + h_1)\cos(a_2 + h_2) = a_1 \cos(a_2 + h_2) + h_1 \cos(a_2 + h_2)$$

From Taylor's theorem,

$$\cos(t+\varepsilon) = \cos(t) - \sin(t)\varepsilon + O(\varepsilon^2) \quad \text{as } \varepsilon \to 0.$$

Thus,

$$f(a+h) = a_1\cos(a_2) - a_1\sin(a_2)h_2 + O(\|h\|^2) + h_1\cos(a_2) - h_1\sin(a_2)h_2 + O(\|h\|^3)$$

and hence

$$f(a+h) - f(a) = \cos(a_2)h_1 - a_1\sin(a_2)h_2 + O(\|h\|^2)\,.$$

This implies

$$f(a+h) = f(a) + Th + o(\|h\|),$$

where we have set $Th = \cos(a_2)h_1 - a_1\sin(a_2)h_2$. This is a linear map $\mathbb{R}^2 \to \mathbb{R}$ and bounded by the finite dimensionality of $\mathbb{R}^2$. So we have proven that f is differentiable at a, and $Df|_a \in L(\mathbb{R}^2, \mathbb{R})$ is given by

$$Df|_a h = \cos(a_2)h_1 - a_1\sin(a_2)h_2.$$

6.4. Example. Let $\mathbb{R}^{n\times n}$ denote the space of real $n \times n$ matrices equipped with the matrix norm $\|A\|_{\text{op}} = \sup_{\|x\|=1} \|Ax\|$ (with $x \mapsto \|x\|$ any given norm on $\mathbb{R}^n$). Define

$$f : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{n\times n}, \qquad f(A) = A^2.$$

We show that f is Fréchet differentiable and compute $Df|_A$. Expand

$$(A+H)^2 = A^2 + AH + HA + H^2$$

and $\|H^2\|_{\text{op}} \leqslant \|H\|_{\text{op}}^2 = o(\|H\|)$. The term $T[H] = AH + HA$ is linear and T is a bounded operator since $\|T[H]\| \leqslant 2\|A\|\|H\|$. Hence

$$Df|_A[H] = AH + HA.$$

Observe that when $n = 1$ the 1×1 matrices A and H commute and so we recognize the derivative operator as $H \mapsto 2AH$, acting on numbers H, thus recovering the familiar formula of differentiation of the function $x \mapsto x^2$ when evaluated at $x = A$. However for $n \geqslant 2$ the matrices A and H do not always commute (for most A), so no such simplification is possible.

6.5. Example. The linearization of a bounded linear map should be the linear map itself. More generally, let y_0 be any vector in Y, $T \in L(X, Y)$, and let $f : X \to Y$ be defined by

$$f(x) = y_0 + Tx.$$

Then for every $a \in X$ we have $f(a+h) - f(a) = Th$ so the $o(\|h\|)$ term in the definition of derivative is actually zero. This shows that f is differentiable everywhere and $Df|_a = T$. We also recognize that in this case the linear transformation $Df|_a$ is "constant" in the sense that it is independent of a.

6.6. Example. Let $\mathcal{J} : C([0,1]) \to C([0,1])$ where $\mathcal{J}(g)$ is the $C([0,1])$-function defined by

$$\mathcal{J}(g)(x) = \int_0^x g(t)^2 dt.$$

We show that $\mathcal{J}$ is Fréchet differentiable at every $g \in C([0,1])$. Compute

$$\mathcal{J}(g+h)(x) - \mathcal{J}(g)(x) = \int_0^x (g(t)+h(t))^2 dt - \int_0^x g(t)^2 dt = 2\int_0^x g(t)h(t)dt + \int_0^x h(t)^2 dt$$

Set $Th(x) = 2\int_0^x g(t)h(t)dt$. This is a bounded linear map:

$$\|Th\|_\infty \leqslant 2\int_0^1 |g(t)h(t)|dt \leqslant C_g \|h\|_\infty,$$

where $C_g = 2\int_0^1 |g(t)|dt$. We have

$$\mathcal{J}(g+h)(x) - \mathcal{J}(g)(x) - Th(x) = \int_0^x h(t)^2 dt.$$

Thus

$$\begin{aligned}\|\mathcal{J}(g+h) - \mathcal{J}(g) - Th\|_\infty &\leqslant \sup_{x\in[0,1]} \Big| \int_0^x h(t)^2 dt \Big| \\ &\leqslant \int_0^1 |h(t)|^2 dt \leqslant \sup_{x\in[0,1]} |h(x)|^2 = \|h\|_\infty^2\end{aligned}$$

This implies

$$\frac{1}{\|h\|_\infty}\|\mathcal{J}(g+h) - \mathcal{J}(g) - Th\|_\infty \leqslant \|h\|_\infty \to 0$$

as $h \to 0$. Thus, $\mathcal{J}$ is Fréchet differentiable at g and $D\mathcal{J}|_g$ is the linear operator from $C([0,1])$ to $C([0,1])$ which maps the function h to the function $D\mathcal{J}|_g[h]$, defined by

$$D\mathcal{J}|_g[h](x) = 2\int_0^x g(t)h(t)dt.$$

6.7. Example. Consider the nonlinear functional $\mathcal{I} : C([0,1]) \to \mathbb{R}$ given by

$$\mathcal{I}(g) = \int_0^1 g(t)^2 dt.$$

Then $\mathcal{I}$ is Fréchet differentiable at every $g \in C([0,1])$. By similar calculations as in Example 6.6 we see that

$$\mathcal{I}(g+h) - \mathcal{I}(g) - \Lambda[h] = \int_0^1 h(t)^2 dt$$

where $\Lambda : C([0,1]) \to \mathbb{R}$ is the linear functional that maps h to the number $2\int_0^1 g(t)h(t)\,dt$. We conclude that $\mathcal{I}$ is Fréchet differentiable at g and that $D\mathcal{I}|_g : C([0,1]) \to \mathbb{R}$ is the bounded linear functional on $C([0,1])$ given by

$$D\mathcal{I}|_g[h] = \Lambda[h] = 2\int_0^1 g(t)h(t)\,dt.$$

Next, we discuss some familiar properties of derivatives. It follows directly from the definition that $Df|_a$ is linear in f. That is, if $f : U \to Y$, $g : U \to Y$ are differentiable at $a \in U$ and $\lambda, \mu \in \mathbb{R}$, then the function $\lambda f + \mu g : U \to Y$ defined by $(\lambda f + \mu g)(x) = \lambda f(x) + \mu g(x)$ is differentiable at a and

$$D(\lambda f + \mu g)|_a = \lambda\, Df|_a + \mu\, Dg|_a.$$

More substantial is a generalization of the chain rule regarding differentiation of a composition of two differentiable functions, $f \circ g(x) = f(g(x))$.

6.8. THEOREM (Chain rule). *Let X_1, X_2, X_3 be normed vector spaces and $U_1 \subset X_1, U_2 \subset X_2$ open. Let $a \in U_1$ and $g : U_1 \to X_2$, $f : U_2 \to X_3$ such that g is Fréchet differentiable at a, $g(U_1) \subset U_2$ and f is Fréchet differentiable at $g(a)$. Then the function $f \circ g : U_1 \to X_3$ defined by $(f \circ g)(x) = f(g(x))$ is Fréchet differentiable at a and*

$$D(f \circ g)|_a h = Df|_{g(a)} Dg|_a h \tag{6.2}$$

for all $h \in X_1$.

Note that $D(f \circ g)|_a \in L(X_1, X_3)$, $Dg|_a \in L(X_1, X_2)$, $Df|_{g(a)} \in L(X_2, X_3)$ and the product on the right hand side of (6.2) refers to the composition of the two linear transformations.

PROOF OF THEOREM 6.8. Let $a, a + h \in U_1$. We write

$$k(h) = g(a + h) - g(a)$$

and

$$\begin{aligned} &f(g(a + h)) - f(g(a)) - Df|_{g(a)} Dg|_a h \\ &= f(g(a) + k(h)) - f(g(a)) - Df|_{g(a)} k(h) + Df|_{g(a)}(g(a + h) - g(a) - Dg|_a h). \end{aligned}$$

We use the triangle inequality and that $Df|_{g(a)}$ is a bounded linear map to estimate

$$\begin{aligned} &\big\| f(g(a + h)) - f(g(a)) - Df|_{g(a)} Dg|_a h \big\|_{X_3} \\ &\leqslant \big\| f(g(a) + k(h)) - f(g(a)) - Df|_{g(a)}[k(h)] \big\|_{X_3} + \|Df|_{g(a)}\|_{\mathrm{op}} \|g(a + h) - g(a) - Dg|_a h\|_{X_2} . \end{aligned}$$

Now the second term is $o(\|h\|_{X_1})$ since $\|g(a + h) - g(a) - Dg|_a h\|_{X_2} = o(\|h\|_{X_1})$. In order to prove that the first term is $o(\|h\|_{X_1})$ we have to prove that given $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that

$$\big\| f(g(a) + k(h)) - f(g(a)) - Df|_{g(a)}[k(h)] \big\|_{X_3} \leqslant \varepsilon \|h\|_{X_1} \text{ for } \|h\|_{X_1} < \delta. \qquad (\#)$$

By the differentiability of g at a we have $k(h) = Dg|_a h + o(\|h\|_{X_1})$. Setting $C = 1 + \|Dg|_a\|_{\mathrm{op}}$ we find δ_1 such that

$$\|k(h)\|_{X_2} \leqslant C\|h\|_{X_1} \text{ for } \|h\|_{X_1} \leqslant \delta_1.$$

By the differentiability of f at $g(a)$ there exists a $\delta_2 > 0$ such that

$$\|f(g(a) + k) - f(g(a)) - Df|_{g(a)} k\|_{X_3} \leqslant \frac{\varepsilon}{C} \|k\|_{X_2} \quad \text{for } \|k\|_{X_2} \leqslant \delta_2.$$

Now let $\delta = \min\{\delta_1, \delta_2/C\}$. Then $\|k(h)\|_{X_2} \leqslant C\delta \leqslant \delta_2$ for $\|h\|_{X_1} \leqslant \delta$. Therefore, for $\|h\|_{X_1} \leqslant \delta$,

$$\big\| f(g(a) + k(h)) - f(g(a)) - Df|_{g(a)}[k(h)] \big\|_{X_3} \leqslant \frac{\varepsilon}{C} \|k(h)\|_{X_2} \leqslant \varepsilon \|h\|_{X_1}$$

and (#) is verified. □

6.9. COROLLARY. *Let X_1, X_2, X_3, X_4 be normed spaces and let $U \subset X_2$ be open, $a \in U$. Let $f : U \to X_3$ be differentiable at a.*

(i) If $S \in L(X_1, X_2)$, $S(b) = a$ then $f \circ S$ is differentiable at b and $D(f \circ S)|_b[h] = Df|_a[Sh]$ for all $h \in X_1$.

(ii) If $T \in L(X_3, X_4)$ then $T \circ f$ is differentiable at a and $D(T \circ f)|_a[h] = TDf|_a[h]$ for all $h \in X_3$.

To prove this apply the chain rule and Example 6.5.

6.10. EXERCISE (Product rule). Let X be a normed vector space, $U \subset X$ open and assume that $f, g : U \to \mathbb{R}$ are differentiable at $a \in U$. We consider the pointwise product $fg : x \mapsto f(x)g(x)$. Show that fg is also differentiable at a and

$$D(fg)\,|_a = f(a)\, Dg\,|_a + g(a)\, Df\,|_a .$$

Formulate and prove other variants of the product rule (you might take a peek at Example 6.174).

6.1.2. Directional derivatives. A directional derivative (for a direction vector v) measures the rate of change of a function in the direction v.

6.11. Definition. Let X, Y be normed vector spaces, $U \subset X$ open, $f : U \to Y$. Let $a \in U$, $v \in X$ with $v \neq 0$. If the function

$$t \mapsto f(a + tv)$$

(defined for small $t \in \mathbb{R}$) is differentiable at $t = 0$, then we denote its derivative at $t = 0$ by $D_v f|_a$ and call it the *directional derivative* of f at a in direction v. In other words,

$$D_v f|_a = \lim_{t \to 0} \frac{f(a + tv) - f(a)}{t}$$

exists in Y. The function f is called *Gâteaux differentiable* at a if $D_v f|_a$ exists for all $v \in X$.

6.12. Theorem. *Let X, Y be normed vector spaces, $U \subset X$ open and $f : U \to Y$ Fréchet differentiable at $a \in U$. Then f is Gâteaux differentiable at a and for every $v \in X$ the directional derivative $D_v f|_a$ is given by*

$$D_v f|_a = Df|_a[v]. \tag{6.3}$$

Proof. By Observation 6.2, $f(a + tv) - f(a) - Df|_a[tv] = o(t)$ as $t \to 0$. Therefore,

$$\frac{f(a + tv) - f(a)}{t} = Df|_a[v] + o(1) \quad \text{as } t \to 0.$$

This means that the limit $\lim_{t \to 0}(f(a + tv) - f(a))/t$ exists and equals $Df|_a[v]$. □

6.13. Example. Revisit Example 6.3 and consider $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = x_1 \cos(x_2)$. The directional derivative $D_v f$ at $a = (a_1, a_2)$, with $v = (v_1, v_2)$, is given by

$$D_v f|_a = Df|_a[v] = \cos(a_2)v_1 - a_1 \sin(a_2)v_2.$$

6.14. Example. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x) = \frac{x_1^3}{x_1^2 + x_2^2}$ if $x \neq 0$ and $f(0) = 0$. Note that f is continuous. We show that all directional derivatives $D_v f|_0$ for $v \neq 0$ exist: for $v = (v_1, v_2)$,

$$f(tv) - f(0) = t \frac{v_1^3}{v_1^2 + v_2^2},$$

so $D_v f|_0 = \frac{v_1^3}{v_1^2 + v_2^2}$. But f is not Fréchet differentiable at 0, because, by linearity of the Fréchet derivative, $D_v f|_0$ would have to be $v_1 D_{e_1} f|_0 + v_2 D_{e_2} f|_0 = v_1$, which is false when $v_2 \neq 0$.

6.1.3. Partial derivatives. We now study functions of n variables, i.e. defined on an open $U \subset \mathbb{R}^n$. The target space is allowed to be any normed space Y, but special emphasis will be put on the case where $Y = \mathbb{R}^m$.

6.15. Definition. Let e_k denote the kth unit vector in $\mathbb{R}^n$. Consider a function $f : U \to Y$, $a = \sum_{j=1}^n a_j e_j \in U$. For $k = 1, \ldots, n$ the *partial derivative* $\partial_k f(a)$ is the directional derivative of f in the direction e_k,

$$\partial_k f(a) = D_{e_k} f|_a = \lim_{t \to 0} \frac{f(a + te_k) - f(a)}{t}.$$

One also uses the notations $\partial_{x_k} f(a)$ or $\frac{\partial f}{\partial x_k}(a)$ for $\partial_k f(a)$. Observe that $\partial_k f(a)$ is the derivative of the one-variable function $t \mapsto f(a_1, \ldots, a_{k-1}, t, a_{k+1}, \ldots, a_n)$ at a_k.

If f is Fréchet differentiable at a, then we can compute its Fréchet derivative in terms of the partial derivatives using Theorem 6.12. From (6.3) we get

$$Df|_a h = \sum_{j=1}^n h_j Df|_a e_j = \sum_{j=1}^n h_j \partial_j f(a) \in Y. \tag{6.4}$$

The case $Y = \mathbb{R}^m$ is of particular interest. We now assume that f is defined in some open subset U of $\mathbb{R}^n$ and has values in $\mathbb{R}^m$. We let $\{e_1, \dots, e_n\}$ be the standard basis of unit vectors in $\mathbb{R}^n$ and usually use the same notation $\{e_1, \dots, e_m\}$ for the standard basis in $\mathbb{R}^m$. We shall use the Euclidean norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, i.e. $\|x\| = (\sum_{j=1}^n |x_j|^2)^{1/2}$ for $x \in \mathbb{R}^n$ and $\|y\| = (\sum_{i=1}^m |y_i|^2)^{1/2}$ for $y \in \mathbb{R}^m$. However, in view of the equivalence of norms in finite dimensional spaces (see §5.3) the choice of norms on $\mathbb{R}^n$, $\mathbb{R}^m$ will usually make not much difference.

For $\mathbb{R}^m$-valued functions we may write $f(x) = \sum_{i=1}^m f_i(x)e_i$ and call the f_i the *coordinate functions* of f. In matrix notation

$$f : \mathbb{R}^n \supset U \to \mathbb{R}^m, \qquad \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_n) \end{pmatrix}.$$

In this case the partial derivative $\partial_k f(a)$ is the vector of the partial derivatives $\partial_k f_i(a)$ of the coordinate functions,

$$\partial_k f(a) = \sum_{i=1}^m \partial_k f_i(a) e_i = \begin{pmatrix} \partial_k f_1(a) \\ \vdots \\ \partial_k f_m(a) \end{pmatrix}.$$

For $\mathbb{R}^m$-valued functions of n real variables, the Fréchet derivative is also called the *total derivative* and Fréchet differentiability is also referred to as *total differentiability.* If f is totally differentiable at a point $a = (a_1, \dots, a_n) \in U$, then we can compute its total (or Fréchet) derivative in terms of the partial derivatives by using again (6.3):

$$Df|_a h = \sum_{j=1}^n h_j Df|_a e_j = \sum_{j=1}^n h_j \begin{pmatrix} \partial_j f_1(a) \\ \vdots \\ \partial_j f_m(a) \end{pmatrix}. \tag{6.5}$$

By definition, $Df|_a$ is a linear map $\mathbb{R}^n \to \mathbb{R}^m$. It is therefore given by multiplication with a real $m \times n$ matrix. We will denote this matrix also by $(Df|_a)$ and call it the *Jacobian matrix of f at a.* One often uses the notations $Df|_a$ or $(Df|_a)$ interchangeably if it is clear from the context whether we mean the linear transformation, or the matrix associated to it. From (6.3) we conclude that the jth column vector of this matrix is given by the $\partial_j f(a) \in \mathbb{R}^m$. Hence

$$Df|_a = (\partial_j f_i(a))_{i,j} = \begin{pmatrix} \partial_1 f_1(a) & \cdots & \partial_n f_1(a) \\ \vdots & \ddots & \vdots \\ \partial_1 f_m(a) & \cdots & \partial_n f_m(a) \end{pmatrix} \in \mathbb{R}^{m \times n}. \tag{6.6}$$

6.16. Definition. If $m = 1$, the *gradient* of f at a is defined as

$$\nabla f(a) = (Df|_a)^\intercal = \sum_{j=1}^n \partial_j f(a) e_j = \begin{pmatrix} \partial_1 f(a) \\ \vdots \\ \partial_n f(a) \end{pmatrix} \in \mathbb{R}^n.$$

One also uses the notation $\operatorname{grad} f(a)$ for $\nabla f(a)$.

Summary of matrix notation:

- Column vectors with m entries, i.e. $m \times 1$-matrices are identified with vectors in $\mathbb{R}^m$.
- Row vectors with n entries, i.e. $1 \times n$ matrices, are identified with linear functionals acting on vectors in $\mathbb{R}^n$.
- The $m \times 1$ columns of the Jacobian matrix $(Df)_a$ are the partial derivatives $\partial_k f(a)$.
- The $1 \times n$ rows of $(Df|_a)$ are the transposes of the gradients of the coordinate functions at a.

- The rows of $(Df|_a)$ are identified with the differentials of the coordinate functions at a. If $\langle\cdot,\cdot\rangle$ denotes the standard Euclidean scalar product on $\mathbb{R}^n$ the ith row is identified with the linear form $h \mapsto (\nabla f_i(a))^\intercal h = \langle \nabla f_i(a), h\rangle$.

6.17. Exercise. Let $F : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by $F(x) = (x_1 x_2 \sin(x_3), x_2^2 - e^{x_1})$. Use a combination of product and chain rules to show that F is totally differentiable everywhere. The Jacobian matrix is given by

$$(DF|_x) = \begin{pmatrix} x_2\sin(x_3) & x_1\sin(x_3) & x_1x_2\cos(x_3) \\ -e^{x_1} & 2x_2 & 0 \end{pmatrix}.$$

6.18. Exercise. Let $f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ be functions which are differentiable at $0 \in \mathbb{R}$. Define a function $F : \mathbb{R}^2 \to \mathbb{R}$ as

$$F(x_1, x_2) = \begin{cases} f(x_1) & \text{if } x_2 = 0 \\ g(x_2) & \text{if } x_1 = 0 \\ \dots & \text{if } x_1 \neq 0,\ x_2 \neq 0 \end{cases}$$

(i) Show that the partial derivatives $\partial_1 F(0,0)$ and $\partial_2 F(0,0)$ exist and compute them.

(ii) Demonstrate that the definition of F in $\mathbb{R}^2\backslash(\mathbb{R}\times\{0\}\cup\{0\}\times\mathbb{R})$ has no impact on the partial derivatives at $(0,0)$. The function F may even be unbounded near $(0,0)$. Conclude that one can compute the Jacobian matrix $\left(\frac{\partial F}{\partial x_1} \quad \frac{\partial F}{\partial x_2}\right)|_{0,0}$ but its existence does not imply Fréchet differentiability at $(0,0)$.

6.19. Exercise. Let $F : \mathbb{R}^2 \to \mathbb{R}$ be defined by $F(x) = \frac{x_1x_2}{x_1^2+x_2^2}$ if $x \neq 0$ and $F(0) = 0$.
(i) Show that the partial derivatives $\partial_1 F(x)$, $\partial_2 F(x)$ exist for every $x \in \mathbb{R}^2$.
(ii) Show that F is not continuous at $(0,0)$.
(iii) Determine at which points F is totally differentiable.

The formulas on directional and partial derivatives were derived *assuming* that F is Fréchet differentiable. A natural question is: When does the existence of directional derivatives near a (or just the existence of partial derivatives near a) imply the Fréchet differentiability of f at a? Exercises 6.18, 6.19 tell us that additional assumptions are needed. One usually imposes some continuity assumption on the directional or partial derivatives. This will be addressed in the next section (see Theorems 6.30, 6.32 below).

6.1.4. Mean value theorems. We begin with a version of the mean value theorem for *real-valued* differentiable functions on a normed space X. Recall (*cf.* Definition 5.82) that a set $A \subset X$ is called convex if for every two points $x, y \in A$ the line segment connecting x and y is contained in A.

6.20. Theorem (Mean value theorem). *Let X be a normed space and let $U \subset X$ be open and convex. Let $f : U \to \mathbb{R}$ be differentiable in U. Then for every $x, y \in U$, there exists a $\tau \in (0,1)$ such that*

$$f(x) - f(y) = Df|_{\tau x+(1-\tau)y}(x - y).$$

Proof. Let $g(t) = f(tx+(1-t)y)$. By the chain rule $g'(t) = Df|_{tx+(1-t)y}(x-y)$. By the standard mean value theorem (Corollary 2.37), $g(1) - g(0) = g'(\tau)$ for some $\tau \in (0,1)$ and we get the assertion. □

6.21. Example. It is crucial in Theorem 6.20 that f takes values in $\mathbb{R}$ (and not in $\mathbb{C}$ or a higher dimensional vector space). Consider $f : [0,1] \to \mathbb{R}^2$ given by

$$f(t) = (f_1(t), f_2(t)) = (t^2, t^3).$$

We have $f'(t) = (2t, 3t^2)$. Then $\tau_1 = 1/2$ is the unique value of t such that $1 = f_1(1) - f_1(0) = f_1'(\tau_1)(1-0)$ and $\tau_2 = 1/\sqrt{3}$ is the unique value of t such that $1 = f_2(1) - f_2(0) = f_2'(\tau_2)(1-0)$. Thus there does not exist any $\xi \in (0,1)$ with $f(1) - f(0) = f'(\xi)(1-0)$. This example also shows that the theorem does not in general hold for complex-valued functions (consider $t \mapsto t^2 + it^3$).

We discuss an integral version of the mean value theorem which is closely related to the fundamental theorem of calculus. Since we integrate derivatives we need to make an additional continuity assumption.

6.22. Definition. Let X be a normed space and $U \subset X$ be an open subset. Let Y be a normed space. A map $f : U \to Y$, is called *continuously differentiable* (on U) if it is differentiable on U and the map

$$Df : \begin{cases} U \to L(X,Y), \\ x \mapsto Df|_x \end{cases}$$

is continuous. We denote the collection of continuously differentiable maps by $C^1(U,Y)$. If $Y = \mathbb{R}$ we also write $C^1(U,\mathbb{R}) = C^1(U)$.

Remark. For $f : U \to \mathbb{R}$, continuity of the map $U \to L(\mathbb{R}^n, \mathbb{R})$, $x \mapsto Df|_x$ is equivalent to continuity of the gradient map $U \to \mathbb{R}^n$, $x \mapsto \nabla f(x)$.

6.23. Theorem (Mean value theorem, integral version). *Let X be a normed space and let $U \subset X$ be open and convex. Let $f \in C^1(U,\mathbb{R})$. Then for every $x, y \in U$,*

$$f(x) - f(y) = \int_0^1 Df|_{tx+(1-t)y}(x-y)dt.$$

Proof. Let $g(t) = f(tx + (1-t)y)$. By the fundamental theorem of calculus

$$f(x) - f(y) = g(1) - g(0) = \int_0^1 g'(s)ds = \int_0^1 Df|_{tx+(1-t)y}(x-y)dt.$$

□

We will now aim for versions of the above results for *vector-valued functions*. The integral version in Theorem 6.23 has an analogue for functions with values in $\mathbb{R}^m$, and more generally for functions with values in a Banach space Y, see Exercise 6.120. This involves a Riemann integral for Banach-space valued functions (see Exercises 5.124, 5.125, 5.126). On the other hand, from Example 6.21 we know that the original mean value theorem in the form of Theorem 6.20 has *no direct analogue* for vector-valued functions.

In many situations one wants to derive estimates and may not need the full mean value theorem; indeed one is often happy with just an efficient inequality, such as Corollary 2.39 in the scalar case. We now obtain extensions of such inequalities to the vector-valued case.

We first consider functions of one real variable with values in a normed space Y. In the following theorem we do not assume any continuity on the derivatives (this mirrors the assumptions in the scalar version of the mean value Theorem 6.20 above). Note that the space $L(\mathbb{R}, Y)$ of bounded linear operators from $\mathbb{R}$ to Y is identified with Y.

6.24. Theorem. *Let $f : (a,b) \to Y$ be differentiable on (a,b). Then for all $t_1, t_2 \in (a,b)$*

$$\|f(t_1) - f(t_2)\|_Y \leqslant |t_1 - t_2| \sup\{\|Df|_{t_1+s(t_2-t_1)}\|_Y : 0 \leqslant s \leqslant 1\}.$$

Proof. Fix $t_1 \neq t_2$, let $M_\circ = \sup\{\|Df|_{t_1+s(t_2-t_1)}\|_Y : 0 \leqslant s \leqslant 1\}$ and let $M > M_\circ$. Let

$$E = \{s \in [0,1] : \|f(t_1 + \sigma(t_2 - t_1)) - f(t_1)\|_Y \leqslant M\sigma|t_1 - t_2| \text{ for } 0 \leqslant \sigma \leqslant s\}.$$

Observe that E is not empty since $0 \in E$. We wish to show that E is open and closed in $[0,1]$; this implies $E = [0,1]$, by the connectedness of $[0,1]$ (see §4.1.5).

Since f is continuous we see that E is closed. To see that E is open in $[0,1]$ we assume $s_0 \in E$. By definition of E, $s \in E$ for $0 \leqslant s \leqslant s_0$. Thus we may assume $s_0 < 1$, and it suffices to verify that there is a $\delta > 0$ such that $\|f(t_1 + s(t_2 - t_1)) - f(t_1)\|_Y \leqslant Ms|t_2 - t_1|$ for $s_0 \leqslant s < s_0 + \delta$. By assumption f is differentiable at $t_1 + s_0(t_2 - t_1)$ which implies that

$$\|f(t_1 + s(t_2 - t_1)) - f(t_1 + s_0(t_2 - t_1))\|_Y \leqslant M_\circ(s - s_0)|t_2 - t_1| + r(s - s_0)|t_2 - t_1|$$

with $r(s-s_0) = o(|s-s_0|)$ as $s \to s_0$. Since $M > M_\circ$ there exists $\delta > 0$ such that

$$\left|M_\circ(s-s_0)(t_2-t_1) + r(s-s_0)(t_2-t_1)\right| \leqslant M(s-s_0)|t_2-t_1|$$

for $s_0 \leqslant s < s_0 + \delta$ and we get for such s

$$\begin{aligned}
&\|f(t_1 + s(t_2-t_1)) - f(t_1)\|_Y \\
&\leqslant \|f(t_1 + s(t_2-t_1)) - f(t_1 + s_0(t_2-t_1))\|_Y + \|f(t_1 + s_0(t_2-t_1)) - f(t_1)\|_Y \\
&\leqslant M(s-s_0)|t_2-t_1| + Ms_0|t_2-t_1| = Ms|t_2-t_1|\,.
\end{aligned}$$

This establishes that E is open in $[0,1]$. We conclude $E = [0,1]$ and thus $\|f(t_2) - f(t_1)\| \leqslant M|t_2-t_1|$. Since $M > M_\circ$ was arbitrary, the inequality with $M_\circ$ in place of M holds. □

6.25. Theorem. *Let U be a convex open set in X and let $f : U \to Y$ be differentiable on U. Let $x, w \in U$ and $M_{x,w} = \sup\{\|Df|_{x+t(w-x)}\|_{X\to Y} : 0 \leqslant t \leqslant 1\}$. Then*

(i)

$$\|f(x) - f(w)\|_Y \leqslant M_{x,w}\|x-w\|_X\,.$$

(ii) If in addition $t \mapsto Df(x+t(w-x))$ is continuous in $[0,1]$, then there exists $\tau \in [0,1]$ such that

$$\|f(x) - f(w)\|_Y \leqslant \|Df|_{x+\tau(w-x)}\|_{X\to Y}\|x-w\|_X\,.$$

Proof. To get (i), we consider the function $g : [0,1] \to Y$ with

$$g(t) = f(tx + (1-t)w).$$

It is differentiable, with $Dg|_t = Df|_{tx+(1-t)w}(x-w)$. Now apply Theorem 6.24 to g.

For part (ii) consider the function $t \mapsto \|Df|_{tx+(1-t)w}\|_{X\to Y}$ which under the assumption in (ii) is continuous and has a maximum in $[0,1]$. This yields the asserted result. □

6.26. Corollary. *Let U be a convex open set in X and let $f : U \to Y$ be differentiable on U. If $Df|_x = 0$ for all $x \in U$, then f is constant.*

6.27. Exercise. Show that the conclusion of the corollary also holds under the weaker assumption that U is open and connected (rather than convex).

6.28. Exercise. There are alternative proofs of Theorems 6.24, 6.25 which rely on the standard mean value theorem for $\mathbb{R}$-valued functions and Corollary 5.92 (a consequence of the Hahn-Banach theorem). For a Y-valued function $x \mapsto f(x)$ and $\lambda \in Y^*$ consider $x \to \lambda(f(x))$. Work out the proofs.

6.29. Exercise. Let X be a Banach space. Let $U \subset X$ be open and convex, and let $\overline{U}$ its closure. Let $F : \overline{U} \to \overline{U}$ be a continuous map, such that F is differentiable on U. Show: If $\sup_{x\in U}\|DF|_x\|_{\mathrm{op}} < 1$ then $F : \overline{U} \to \overline{U}$ has a unique fixed point in $\overline{U}$.

We will now give an application of the mean value inequality to a result which elucidates the relation between Fréchet and Gâteaux differentiability. Recall from Theorem 6.12 that Fréchet differentiability at a point a implies Gâteaux differentiability at a, and we know that the converse does not hold. We give additional conditions under which such a converse can be established.

6.30. Theorem. *Let X, Y be normed spaces, $a \in X$, and U a neighborhood of a. Let $f : U \to Y$ be Gâteaux differentiable in U. Then the following holds:*

(i) If for all $u \in X$ the directional derivatives $D_u f$ are continuous at a, then the map $u \to D_u f|_a$ is linear.

(ii) If $\sup_{\|u\|_X\leqslant 1}\|D_u f|_a\|_Y < \infty$ and $\lim_{x\to a}\sup_{\|u\|_X\leqslant 1}\|D_u f|_x - D_u f|_a\| = 0$ then f is Fréchet differentiable at a, with $D_u f|_a = Df|_a u$.

PROOF. The formula $D_{cu}f|_x = cD_uf|_x$ for $c \in \mathbb{R}$, $x \in U$ follows directly from the definition of directional derivative. We need to show $D_{u+v}f|_a = D_uf|_a + D_vf|_a$. For small t we write

$$\frac{f(a+t(u+v))-f(a)}{t} - D_vf|_a - D_uf|_a = \\ \frac{f(a+tu+tv)-f(a+tu)}{t} - D_vf|_a + \frac{f(a+tu)-f(a)}{t} - D_uf|_a$$

We use the existence of D_vf near a in order to apply Theorem 6.24 to

$$s \mapsto g(s) = \frac{f(a+tu+stv)-f(a+tu)}{t} - sD_vf|_a$$

which is defined in a neighborhood of $[0,1]$ and Fréchet differentiable there. This yields

$$\Big\| \frac{f(a+tu+tv)-f(a+tu)}{t} - D_vf|_a \Big\|_Y = \|g(1)-g(0)\|_Y \\ = \sup_{0\leqslant\sigma\leqslant 1} \|g'(\sigma)\|_Y = \sup_{0\leqslant\sigma\leqslant 1} \big\|D_vf|_{a+tu+\sigma tv} - D_vf|_a\big\|_Y.$$

Similarly,

$$\Big\| \frac{f(a+tu)-f(a)}{t} - D_uf|_a \Big\|_Y \leqslant \sup_{0\leqslant\sigma\leqslant 1} \big\|D_uf|_{a+\sigma tu} - D_uf|_a\big\|_Y.$$

Choosing t small enough and using the assumed continuity of D_uf and D_vf at a we can make the right hand sides of the last two displays arbitrarily small. Letting $t \to 0$ we conclude that $D_{u+v}f|_a = D_uf|_a + D_vf|_a$.

Turning to (ii) the assumption implies that for $\|u\| \leqslant 1$ the function $x \to D_uf|_x$ is continuous at a, and since $D_{su}f|_x = sD_uf|_x$ we get this for all u. Thus by (i) $u \mapsto D_uf|_a$ is a linear function of u. The assumption $\sup_{\|u\|_X\leqslant 1} \|D_uf|_a\|_Y < \infty$ then implies that $u \mapsto D_uf|_a$ belongs to $L(X,Y)$. Now fix h small and let $u = h/\|h\|_X$. Again by Theorem 6.24,

$$\|f(a+h)-f(a)-D_hf|_a\| = \sup_{0\leqslant s\leqslant 1} \|D_hf|_{a+sh} - D_hf|_a\|_X \\ = \|h\|_X \sup_{0\leqslant s\leqslant 1} \|D_uf|_{a+sh} - D_uf|_a\|_X$$

and since $\lim_{h\to 0} \sup_{\|u\|=1} \sup_{0\leqslant s\leqslant 1} \|D_uf|_{a+sh} - D_uf|_a\|_X = 0$ by the uniform continuity assumption in (ii) we obtain $\|f(a+h)-f(a)-D_hf|_a\| = o(\|h\|_X)$. This means that f is Fréchet differentiable at a. □

6.31. EXAMPLE. The uniformity in u in part (ii) of Theorem 6.30 is needed (unless X is finite dimensional): Consider an arbitrary unbounded linear map $T : X \to Y$ (see Example 5.14). It is Gâteaux differentiable everywhere, and $D_hT(a) = Th$ for all $h \in X$. Notice that all directional derivatives are continuous (since constant) but T is not Fréchet differentiable anywhere.

Here is a useful analog of Theorem 6.30 (ii), for functions of n variables, which involves the partial derivatives.

6.32. THEOREM. *Let Y be a normed space, $U \subset \mathbb{R}^n$ be an open set with $a \in U$. Assume that the partial derivatives $\partial_j f(x)$ exist for all $x \in U$ and that they are continuous at a. Then f is Fréchet differentiable at a.*

PROOF. For small h we write $h = \sum_{j=1}^n h_je_j$ and define $v_k = v_k(h) = \sum_{j=1}^k h_je_j$ for $1 \leqslant k \leqslant n$. Moreover set $v_0 = 0$. Let $\delta_0 > 0$ be such that the cube $\{a+h : \max_i |h_i| \leqslant 2\delta_0\}$ is contained in U. Then,

$$f(a+h)-f(a) = \sum_{j=1}^n \big(f(a+v_j)-f(a+v_{j-1})\big).$$

By the mean value inequality (Theorem 6.24) applied to

$$t \mapsto g_j(t) = f(a + v_{j-1} + th_je_j) - f(a + v_{j-1}) - th_j\partial_j f(a)$$

we obtain (estimating $g_j(1) - g_j(0) = g_j(1)$ for $j = 1, \dots, n$)

$$\big\|f(a + v_{j-1} + h_je_j) - f(a + v_{j-1}) - h_j\partial_j f(a)\big\|_Y \leqslant \sup_{0\leqslant\tau\leqslant 1} \|\partial_j f(a + v_{j-1} + \tau h_je_j) - \partial_j f(a)\|_Y |h_j|$$

By the continuity of $\partial_j f$ at a, there exists for every $\varepsilon > 0$ a positive $\delta < \delta_0$ such that

$$\sup_{0\leqslant\tau\leqslant 1} \|\partial_j f(a + v_{j-1} + \tau h_je_j) - \partial_j f(a)\|_Y < \varepsilon/n$$

provided that $\|h\| \leqslant \delta$. Then if $\|h\| \leqslant \delta$ (and therefore also $\|v_j\| \leqslant \delta, \|v_{j-1} + t_jh_je_j\| \leqslant \delta$) we get

$$\begin{aligned}\Big\|f(a + h) - f(a) - \sum_{j=1}^n h_j\partial_j f(a)\Big\|_Y &= \Big\|\sum_{j=1}^n \big(f(a + v_j) - f(a + v_{j-1}) - h_j\partial_j f(a)\big)\Big\|_Y \\ &\leqslant \sum_{j=1}^n \big\|f(a + v_j) - f(a + v_{j-1}) - h_j\partial_j f(a)\big\|_Y \leqslant \varepsilon\|h\|.\end{aligned}$$

Therefore, $\|f(a + h) - f(a) - Df|_a h\| = o(\|h\|)$, where $Df|_a h = \sum_{j=1}^n h_j\partial_j f(a)$ and clearly the linear map $h \mapsto Df|_a h$ is bounded (see Corollary 5.27). We conclude that f is Fréchet differentiable at a. □

6.33. Corollary. *Let $U \subset \mathbb{R}^n$ be open. Let $f : U \to \mathbb{R}^m$. Then $f \in C^1(U, \mathbb{R}^m)$ if and only if for every $j \in \{1, \dots, n\}$, $i \in \{1, \dots, m\}$, the partial derivatives $\partial_j f_i$ exist in U and are continuous there.*

Proof. Let $f \in C^1(U, \mathbb{R}^m)$. If π_i denotes the projection $y \to y_i$ then $\pi_i \circ \partial_j f$ are continuous, i.e. $\partial_j f_i$ are continuous.

Vice versa if the $\partial_j f_i$ are continuous for $i = 1, \dots, m$ so is $\partial_j f$. And if this holds for all $1 \leqslant j \leqslant n$ and all $x \in U$ then by Theorem 6.32 the function is Fréchet differentiable in U. By our assumptions $x \mapsto Df|_x$ is continuous in U, so $f \in C^1(U, \mathbb{R}^m)$. □

6.2. Inverse function theorem

An important result in the differential calculus of one variable is the local invertibility of differentiable functions and the computation of the derivatives of the inverse function, see §2.2.3. Invertibility for functions of one variable is strongly related to the concept of strict monotonicity. This concept is of no use when we consider maps from $\mathbb{R}^n$ to $\mathbb{R}^n$ (or more generally maps between Banach spaces) and it is much harder to construct inverses of functions, i.e. solve in x an equation of the form $f(x) = y$. An important tool here is the contraction principle which we will use to find (local) inverses of maps between open sets in Banach spaces.

6.2.1. Inverses of linear operators. We will first study the more basic concept of inverting linear operators on a *Banach space* X.

6.34. Definition. Let X, Y be normed spaces. A linear operator $T : X \to Y$ is an *isomorphism of normed spaces* if $T \in L(X, Y)$, T is bijective, and $T^{-1} \in L(Y, X)$.

The following lemma deals with inveribility of small perturbations of the identity map $\mathrm{id} : X \to X$.

6.35. Lemma. *Let X be a Banach space. Let $H \in L(X)$ such that $\|H\|_{\mathrm{op}} < 1$. Then $\mathrm{id} + H$ is an isomorphism with inverse given by*

$$(\mathrm{id} + H)^{-1} = \sum_{n=0}^{\infty} (-1)^n H^n = \mathrm{id} - H + H^2 - H^3 + \cdots .$$

PROOF. The partial sums of the series $\sum_{n=0}^{\infty}(-1)^n H^n$ form a Cauchy sequence provided that $\|H\|_{\mathrm{op}} < 1$, and by the assumed completeness of X it is convergent. We get for $\|H\|_{\mathrm{op}} < 1$

$$(\mathrm{id} + H)\sum_{k=0}^{n}(-1)^k H^k = \Big(\sum_{k=0}^{n}(-1)^k H^k\Big)(\mathrm{id} + H) = \mathrm{id} + (-1)^n H^{n+1}.$$

Thus $(\mathrm{id} + H)^{-1} = \sum_{k=0}^{\infty}(-1)^k H^k$. □

We say that normed spaces X and Y are *isomorphic* if there exists an isomorphism of normed spaces from X to Y. This defines an equivalence relation between normed spaces. From linear algebra we know that finite dimensional vector spaces are isomorphic if and only if they have the same (nonzero) dimension (in the sense of linear algebra).

6.36. THEOREM. *Let X, Y be Banach spaces and let $T \in L(X,Y)$ be an isomorphism.*

(i) Assume $H \in L(X,Y)$ and

$$\|H\|_{X\to Y} < \|T^{-1}\|_{Y\to X}^{-1}.$$

Then $T + H : X \to Y$ is an isomorphism, such that

$$\|(T+H)^{-1}\|_{Y\to X} \leqslant \frac{\|T^{-1}\|_{Y\to X}}{1 - \|T^{-1}\|_{Y\to X}\|H\|_{X\to Y}}.$$

Moreover,

$$\|(T+H)^{-1} - T^{-1} + T^{-1}HT^{-1}\|_{Y\to X} \leqslant \frac{\|H\|_{X\to Y}^2 \|T^{-1}\|_{Y\to X}^3}{1 - \|T^{-1}\|_{Y\to X}\|H\|_{X\to Y}}.$$

(ii) The map $\mathrm{Inv} : S \to S^{-1}$ *defined for* $\|S - T\|_{X\to Y} < \|T^{-1}\|_{Y\to X}^{-1}$ *is differentiable at T and the Fréchet derivative* $D(\mathrm{Inv})_T : L(X,Y) \to L(Y,X)$ *is given by*

$$D(\mathrm{Inv})_T[H] = -T^{-1}HT^{-1}.$$

PROOF. We write

$$T + H = T(\mathrm{id}_X + T^{-1}H).$$

If $\|T^{-1}H\|_{X\to X} < 1$, then Lemma 6.35 implies that $T + H$ is invertible, with

$$(T+H)^{-1} = (\mathrm{id}_X + T^{-1}H)^{-1}T^{-1} = \Big(\sum_{k=0}^{\infty}(-1)^k (T^{-1}H)^k\Big)T^{-1}.$$

Since $\|T^{-1}H\|_{X\to X} \leqslant \|T^{-1}\|_{Y\to X}\|H\|_{X\to Y} < 1$, we get for the operator norm of $(T+H)^{-1}$

$$\|(T+H)^{-1}\|_{Y\to X} \leqslant \|T^{-1}\|_{Y\to X}\sum_{k=0}^{\infty}\|T^{-1}H\|_{X\to X}^k \leqslant \frac{\|T^{-1}\|_{Y\to X}}{1 - \|T^{-1}\|_{Y\to X}\|H\|_{X\to Y}}.$$

The assertion in (ii) follows from the definition of Fréchet derivative and the inequality in (i). Using again the series representation of $(\mathrm{id} + T^{-1}H)^{-1}$ we get

$$(T+H)^{-1} - T^{-1} + T^{-1}HT^{-1} = \sum_{k=2}^{\infty}(-1)^k \big(T^{-1}H\big)^k T^{-1}$$

and then,

$$\begin{aligned}\big\|(T+H)^{-1} - T^{-1} + T^{-1}HT^{-1}\big\|_{Y\to X} &= \Big\|\sum_{k=2}^{\infty}(-1)^k (T^{-1}H)^k T^{-1}\Big\|_{Y\to X}\\ &\leqslant \sum_{k=2}^{\infty}\|T^{-1}\|_{Y\to X}^{k+1}\|H\|_{X\to Y}^k = \frac{\|H\|_{X\to Y}^2\|T^{-1}\|_{Y\to X}^3}{1 - \|T^{-1}\|_{Y\to X}\|H\|_{X\to Y}}.\end{aligned}$$

This shows that Inv is of class C^1 and $D\mathrm{Inv}|_T(H) = D_H\mathrm{Inv}|_T = -T^{-1}HT^{-1}$. □

As a consequence of part (i) of Theorem 6.36 we obtain:

6.37. Corollary. *The set of isomorphisms $X \to Y$ is an open subset of $L(X,Y)$.*

6.2.2. Local invertibility. We shall now consider locally invertible maps defined on an open set Ω in a Banach space X.

6.38. Definition. Let X, Y be normed spaces, let $\Omega \subset X$. We say that a function $f : \Omega \to Y$ is *locally invertible* at $a \in \Omega$ if there exist open sets $U \subset \Omega \subset X$, $V \subset Y$ such that $a \in U$, $f(a) \in V$ and a function $g : V \to U$ such that $g(f(x)) = x$ for all $x \in U$ and $f(g(y)) = y$ for all $y \in V$. In this case we call g a *local inverse* of f (at a) and denote it by $f|_U^{-1}$.

6.39. Theorem (Inverse function theorem). *Let X, Y be Banach spaces, $\Omega \subset X$ be open and let $f : \Omega \to Y$ be differentiable on Ω. Let $a \in \Omega$ and assume that $Df|_a$ is invertible and that $x \mapsto Df|_x$ is continuous at a. Then*

(i) f is locally invertible at a in some open neighborhood $U \subset \Omega$ of a with $(f|_U)^{-1}$ differentiable on $V = f(U)$, and we have for all $x \in U$

$$D(f|_U^{-1})|_{f(x)} = (Df|_x)^{-1}.$$

(ii) If $f \in C^1(\Omega, Y)$, then $(f|_U)^{-1} \in C^1(V, U)$.

Proof. We are aiming to apply the contraction principle and first rewrite the equation $f(x) = y$ as a fixed point equation for a map $\varphi_y : \Omega \to X$.

For fixed $y \in Y$, consider

$$\varphi_y(x) = x + Df|_a^{-1}(y - f(x)) \quad (x \in \Omega)$$

Then $f(x) = y$ if and only if x is a fixed point of φ_y. Calculate

$$D\varphi_y|_x = \mathrm{id} - Df|_a^{-1} Df|_x = Df|_a^{-1}(Df|_a - Df|_x).$$

Let $\lambda = \|Df|_a^{-1}\|_{Y \to X}$. By continuity of Df at a, there exists an open ball $U \subset \Omega$ centered at a such that

$$\|Df|_a - Df|_x\|_{X \to Y} \leqslant \frac{1}{2\lambda} \quad \text{for } x \in U.$$

Then for $x, x' \in U$ we use the mean-value inequality in Theorem 6.25 to get

$$\begin{aligned} \|\varphi_y(x) - \varphi_y(x')\|_X &\leqslant \sup_{\xi \in U} \|D\varphi_y|_\xi\|_{X \to X} \|x - x'\|_X \\ &\leqslant \|Df|_a^{-1}\|_{Y \to X} \sup_{\xi \in U} \|Df|_a - Df|_\xi\|_{X \to Y} \|x - x'\|_X \leqslant \tfrac{1}{2}\|x - x'\|_X. \end{aligned}$$

Note that this does not show that φ_y is a contraction, because $\varphi_y(U)$ may not be contained in U. However, it does show that φ_y has at most one fixed point (this is the same argument used to show uniqueness in the Banach fixed point theorem). Now we can deduce that for every $y \in Y$ we have $f(x) = y$ for at most one $x \in U$. Let $V = f(U)$. Then $f|_U : U \to V$ is a bijection and has an inverse $g : V \to U$.

Claim. V is open.

Proof of the claim. Let $y_0 \in V$. We need to show that there exists an open ball around y_0 that is contained in V. Since $V = f(U)$ there exists $x_0 \in U$ such that $f(x_0) = y_0$. Let $r > 0$ be small enough so that $\overline{B_r(x_0)} \subset U$ (possible because U is open). Let $\varepsilon > 0$ and $y \in B_\varepsilon(y_0)$. We will demonstrate that if $\varepsilon > 0$ is small enough (specifically $\varepsilon \leqslant r/2\lambda$), then φ_y maps $\overline{B_r(x_0)}$ into itself. First note

$$\|\varphi_y(x_0) - x_0\|_X = \|Df|_a^{-1}(y - y_0)\|_X \leqslant \lambda\varepsilon.$$

Hence, with $\varepsilon \leqslant \frac{r}{2\lambda}$, we get for $x \in \overline{B_r(x_0)}$ that

$$\begin{aligned}\|\varphi_y(x) - x_0\|_X &\leqslant \|\varphi_y(x) - \varphi_y(x_0)\|_X + \|\varphi_y(x_0) - x_0\|_X \\ &\leqslant \tfrac{1}{2}\|x - x_0\| + \tfrac{r}{2} \leqslant \tfrac{r}{2} + \tfrac{r}{2} = r.\end{aligned}$$

Thus $\varphi_y(x) \in \overline{B_r(x_0)}$. This proves $\varphi_y(\overline{B_r(x_0)}) \subset \overline{B_r(x_0)}$, so φ_y is a contraction of $\overline{B_r(x_0)}$. By the Banach fixed point theorem, φ_y must have a unique fixed point $x \in \overline{B_r(x_0)}$. So by definition of φ_y we have $f(x) = y$, so $y \in f(U) = V$. Therefore we have shown that $B_\varepsilon(y_0) \subset V$, so V is open.

Conclusion. It remains to show that g is differentiable on V and $Dg|_y = (Df|_{g(y)})^{-1}$ for $y \in V$, or equivalently $Dg|_{f(x)} = Df|_x^{-1}$ for $x \in U$.

By Theorem 6.36 $Df|_a + H$ is invertible if $\|H\|_{X\to Y} < \|Df|_a^{-1}\|_{Y\to X}^{-1} = \lambda^{-1}$. By the definition of U above we have $\|Df|_x - Df|_a\|_{X\to Y} \leqslant (2\lambda)^{-1}$, so indeed, for all $x \in U$, $Df|_x$ is invertible with $(Df|_x)^{-1} \in L(Y, X)$.

We need to show that g is differentiable at y and that the Fréchet derivative is equal to $(Df|_x)^{-1}$. Let k be such that $y + k \in V$. Then there exists a unique h such that $y + k = f(x + h)$. We have

$$\|h\|_X \leqslant \|h - Df|_a^{-1}k\|_X + \|Df|_a^{-1}k\|_X$$

and

$$h - Df|_a^{-1}k = h + Df|_a^{-1}(f(x) - f(x + h)) = \varphi_y(x + h) - \varphi_y(x),$$

so $\|h - Df|_a^{-1}k\|_X \leqslant \frac{1}{2}\|h\|_X$. Therefore, $\|h\|_X \leqslant 2\lambda\|k\|_Y \to 0$ as $\|k\|_Y \to 0$. Now we compute

$$\begin{aligned}&g(y + k) - g(y) - Df|_x^{-1}k = x + h - x - Df|_x^{-1}k \\ &= h - Df|_x^{-1}(f(x + h) - f(x)) = -Df|_x^{-1}(f(x + h) - f(x) - Df|_x h)\end{aligned}$$

and so

$$\begin{aligned}&\frac{1}{\|k\|_Y}\big\|g(y + k) - g(y) - Df|_x^{-1}k\big\|_X \leqslant \|Df|_x^{-1}\|_{Y\to X}\frac{\big\|f(x + h) - f(x) - Df|_x h\big\|_Y}{\|h\|_X}\frac{\|h\|_X}{\|k\|_Y} \\ &\leqslant \|Df|_x^{-1}\|_{Y\to X}\frac{\big\|f(x + h) - f(x) - Df|_x h\big\|_Y}{\|h\|_X}2\lambda \longrightarrow 0 \quad \text{as } k \to 0.\end{aligned}$$

Therefore g is differentiable at y with $Dg|_y = Df|_x^{-1}$. This finishes the proof of part (i) of Theorem 6.39.

In part (ii) we assume that f is of class C^1, and the formula $Dg|_y = (Df|_{g(y)})^{-1}$ together with the continuity of the map Inv in Theorem 6.36 shows that Dg is continuous. □

If f is locally invertible at every point, it is not necessarily (globally) invertible (that is, bijective):

6.40. Example. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $f(x) = (e^{x_2}\sin(x_1), e^{x_2}\cos(x_1))$. Then the matrix of Df_x is given by

$$Df|_x = \begin{pmatrix} e^{x_2}\cos(x_1) & e^{x_2}\sin(x_1) \\ -e^{x_2}\sin(x_1) & e^{x_2}\cos(x_1) \end{pmatrix}.$$

Thus $\det Df|_x = e^{2x_2}(\cos^2(x_1) + \sin^2(x_1)) = e^{2x_2} \neq 0$, so Df_x is invertible. By Theorem 6.39, f is locally invertible at every point $x \in \mathbb{R}^2$. f is not bijective: it is not injective because, for instance, $f(0, 0) = f(2\pi, 0)$.

6.41. Exercise. We are given the concrete case of a nonlinear equation

$$\sin(\alpha(t) - t) + \frac{\alpha(t)^2}{10}\int_0^1 e^{\alpha(s)}ds = \beta(t).$$

Given $\alpha \in C([0, 1])$ the left hand side defines a continuous function of t which we call $\mathcal{K}[\alpha]$. Check that for $\alpha_0(t) = t$ we have $\mathcal{K}[\alpha_0] = \beta_0$ where $\beta_0(t) = \frac{e-1}{10}t^2$.

Prove using the inverse function theorem: If β is close to β_0 (with respect to the norm on $C([0,1])$) then the equation $\mathcal{K}[\alpha] = \beta$ has a solution in $C([0,1])$, with some α near α_0.

We have seen that the inverse function theorem is very useful in solving equations. In some situations the solution of an equation is not unique (even locally) and it is convenient to have a version of the inverse function theorem which covers such situations. Here we assume that the assumption of bijectivity of $Df|_{x_0}$ is replaced by the weaker assumption of surjectivity. We prove such a theorem under the assumption that the target space Y is finite dimensional (a nice application will be given in Theorem 6.65 below).

6.42. THEOREM. *Let X, Y be normed spaces over $\mathbb{R}$, Y finite dimensional. Let $x_0 \in X$, let Ω be a neighborhood of x_0, $f : \Omega \to Y$ be of class C^1, and assume that $Df|_{x_0}$ is surjective. Then there is a neighborhood V of $f(x_0)$, and a C^1 function $g : V \to \Omega$ such that $f(g(y)) = y$ for all $y \in V$.*

PROOF. Let $v_1, \dots, v_n$ be a basis of Y. Since $Df|_{x_0}$ is surjective, there are linearly independent vectors $u_1, \dots, u_n$ in X such that $Df|_{x_0} u_j = v_j$. Let Ω_0 be an open neighborhood of $0 \in \mathbb{R}^n$ such that $x_0 + \sum_{j=1}^n t_j u_j \in \Omega$ for all $t \in \Omega_0$. Define F on Ω_0, with values in Y, by

$$F(t_1, \dots, t_n) = f(x_0 + \sum_{j=1}^n t_j u_j).$$

Then F is differentiable for $t = (t_1, \dots, t_n)$ near the origin and for $h \in \mathbb{R}^n$ we have

$$DF|_t[h] = \sum_{j=1}^n h_j Df|_{x_0 + \sum_{j=1}^n t_j u_j}[u_j].$$

In particular, at the origin $(t_1, \dots, t_n) = 0$,

$$DF|_0[h] = \sum_{j=1}^n h_j Df|_{x_0} u_j = \sum_{j=1}^n h_j v_j.$$

Hence $DF|_0 : \mathbb{R}^n \to Y$ is a linear isomorphism and $t \mapsto DF|_t$ is continuous near 0. By the inverse function theorem there is a neighborhood $U_0 \subset \Omega_0$ of $0 \in \mathbb{R}^n$, a neighborhood V of $f(x_0)$ and an inverse function $G : V \to U_0$ of class C^1 such that $F(G(y)) = y$ for all $y \in V$. Let G_j be the coordinate functions of G, and define

$$g : V \to \Omega, \qquad g(y) = x_0 + \sum_{j=1}^n G_j(y) u_j.$$

Then $f(g(y)) = F(G(y)) = y$ for $y \in V$. □

REMARK. For a corresponding result for injective $Df|_{x_0}$, and some versions without the finite dimensionality assumptions, see Exercises 6.128, 6.129 below.

6.2.3. Further quantification and Newton iterations. In practice, one often needs more quantitative statements about the neighborhoods U and V in the inverse function theorem. We are now given Banach spaces X, Y, $a \in X$ and a Y-valued function f which is defined on an open neighborhood of $a \in X$. Moreover, we assume that the closed ball $\overline{B}(a, r)$ is contained in the domain of f. The objective is to solve the equation $f(x) = y$ for y close to $f(a)$, with a unique x in a neighborhood of a. The crucial assumption in the inverse function theorem is that $Df|_a$ is invertible and that the inverse is in $L(Y, X)$. To obtain an effective quantitative condition, we add a regularity hypothesis on $Df|_x$ for x close to a. We say that on $\overline{B(a,r)}$ the function Df satisfies a *K-Lipschitz condition* at a if

$$\|Df|_{a+h} - Df|_a\|_{X \to Y} \leqslant K\|h\|_X, \quad \|h\|_X \leqslant r. \tag{6.7}$$

In Theorem 6.44 we shall then make quantifications on the neighborhoods U, V in the inverse function theorem, depending on the Lipschitz constant K and the operator norm of $Df|_a^{-1}$. Before discussing this, we mention a consequence of the Lipschitz condition that will be frequently applied.

6.43. Lemma. *Let X, Y be Banach spaces, let $a, a+h \in X$ and let f be a Y-valued C^1-function defined in a neighborhood of the line segment ℓ connecting a and $a+h$. Assume Df satisfies the K-Lipschitz condition* (6.7)*, for all $x \in \ell$. Then*

$$\big\|f(a+h) - f(a) - Df|_a h\big\|_Y \leqslant K\frac{\|h\|_X^2}{2}.$$

Proof. Let λ be any real bounded linear functional in Y^*, and $\|\lambda\|_{Y*} \leqslant 1$. Since $\lambda \circ f \in C^1$ near ℓ we use the fundamental theorem of calculus to write

$$\lambda(f(a+h) - f(a)) = \int_0^1 \frac{d}{ds}\big(\lambda(f(a+sh) - f(a))\big)\, ds = \int_0^1 \lambda(Df|_{a+sh} h)\, ds$$

and then

$$\begin{aligned}
\big|\lambda(f(a+h) - f(a) - Df|_a h)\big| &= \Big|\int_0^1 \lambda(Df|_{a+sh}h - Df|_a h)\, ds\Big| \\
&\leqslant \int_0^1 \big\|Df|_{a+sh}h - Df|_a h\big\|_Y\, ds \leqslant \int_0^1 \big\|Df|_{a+sh} - Df|_a\big\|_{X\to Y}\|h\|_X\, ds \\
&\leqslant \|h\|_X \int_0^1 K\|sh\|_X\, ds = K\|h\|_X^2 \int_0^1 s\, ds = \frac{1}{2}K\|h\|_X^2.
\end{aligned}$$

We conclude using Corollary 5.91:

$$\|f(a+h) - f(a) - Df|_a h\|_Y = \sup_{\|\lambda\|_{Y^*}\leqslant 1} \big|\lambda(f(a+h) - f(a) - Df|_a h)\big| \leqslant \frac{1}{2}K\|h\|_X^2. \qquad \square$$

Lemma 6.43 is not only useful to quantify the neighborhoods in the inverse function theorem; it can also be used to prove effective theorems about the numerical methods used to compute the solutions. We consider two methods (x_n) of iteration, where in both we may choose the initial point $x_0 = a$.

- The *Newton-Kantorovich iteration* given by

$$x_{n+1} = \Phi_y(x_n), \quad \text{where} \quad \Phi_y(x) = x - Df|_x^{-1}(f(x) - y). \tag{6.8}$$

- The *modified Newton iteration* given by

$$x_{n+1} = \varphi_y(x_n), \quad \text{where} \quad \varphi_y(x) = x - Df|_a^{-1}(f(x) - y). \tag{6.9}$$

The map φ_y already featured in the proof of the inverse function theorem.

In both cases suitable assumptions have to be formulated to guarantee that the iterations are well defined for all n. The modified Newton iteration is computationally more robust, since at every step we use the inverse of the same matrix, namely $Df|_a^{-1}$. The original Newton iteration for the computation of zeros of functions on the real line is (6.8) for a real valued function on an interval; this was already covered in Exercise 4.171. The Newton iteration converges extremely fast if one begins the iteration with a good guess (compare the estimates (6.10a), (6.11) below for the modified Newton iteration and the Newton-Kantorovich iteration, respectively). However, in the higher-dimensional (or even infinite dimensional) generalization of Kantorovich it is computationally problematic to recompute the inverse $Df|_{x_n}^{-1}$ at every step of the iteration. We first state the more robust result where the modified Newton iteration is used.

6.44. Theorem. *Let X, Y be Banach spaces, $\Omega \subset X$ be open. Let $f : \Omega \to Y$ be of class C^1 and $a \in \Omega$, and $r > 0$ such that $\{x : \|x - a\|_X \leqslant r\} \subset \Omega$. Assume that $Df|_a$ is invertible with $M = \|Df|_a^{-1}\|_{\mathrm{op}}$, and*

that there exists $K > 0$ such that Df satisfies the K-Lipschitz condition (6.7) at a, for all $x \in \overline{B(a,r)}$. Let $\beta < 1$ and

$$\delta \leqslant \min\Big\{r, \frac{\beta}{KM}\Big\}, \qquad \|y - f(a)\|_Y \leqslant \frac{(2-\beta)\delta}{2M}.$$

Then there is a unique $x_* \in \{x \in X : \|x-a\|_X \leqslant \delta\}$ such that $y = f(x_*)$. If $\|x_0 - a\|_X \leqslant \delta$ the modified Newton iteration

$$x_{n+1} = x_n - Df|_a^{-1}\big(f(x_n) - y\big)$$

satisfies $\|x_n - a\|_X \leqslant \delta$ for $n \geqslant 1$ and converges to x_*. Moreover, for all $n \geqslant 1$,

$$\|x_{n+1} - x_n\|_X \leqslant (\delta + \|x_0 - a\|_X)\beta^n \tag{6.10a}$$

and

$$\|x_n - x_*\|_X \leqslant \tfrac{\delta + \|x_0 - a\|_X}{1-\beta}\beta^n. \tag{6.10b}$$

Proof of Theorem 6.44. We let $\varphi_y(x) = x - Df|_a^{-1}(f(x) - y)$ which satisfies

$$\varphi_y(x) - a = Df|_a^{-1}\big(Df|_a(x-a) - f(x) + f(a) - f(a) + y\big).$$

Thus we get for $\|x-a\|_X \leqslant \delta$, using Lemma 6.43,

$$\begin{aligned}\|\varphi_y(x) - a\|_X &\leqslant \|Df|_a^{-1}\|_{\mathrm{op}}\big(\|f(x) - f(a) - Df|_a(x-a)\|_Y + \|y - f(a)\|_Y\big)\\ &\leqslant M\Big(K\frac{\|x-a\|^2}{2} + \frac{(2-\beta)\delta}{2M}\Big) \leqslant M\big(K\frac{\delta^2}{2} + \frac{(2-\beta)\delta}{2M}\big) = \frac{\delta}{2}\big(MK\delta + 2 - \beta\big) \leqslant \delta\end{aligned}$$

since $MK\delta \leqslant \beta$ by assumption. Hence φ_y maps $\overline{B}(a,\delta)$ into itself. Moreover, φ_y is a contraction with contraction constant β since for $x, x' \in \overline{B}(a,\delta)$

$$\begin{aligned}\|\varphi_y(x) - \varphi_y(x')\|_X &= \|x - Df|_a^{-1}(f(x)) - x' + Df|_a^{-1}(f(x'))\|_X\\ &= \big\|Df|_a^{-1}\big(f(x') - f(x) - Df|_a(x'-x)\big)\big\|_Y\\ &\leqslant M \sup_{\|w-a\|\leqslant\delta} \|Df|_w - Df|_a\|_{\mathrm{op}}\|x - x'\|_X\\ &\leqslant MK\delta\|x - x'\|_X \leqslant \beta\|x - x'\|_X\,.\end{aligned}$$

Here we used the mean value inequality for the function $x \mapsto f(x) - Df|_a x$.

Since $\beta < 1$ we may use the Banach fixed point theorem and find that φ_y has a unique fixed point x^* in $\overline{B}(a,\delta)$. Note $\varphi_y(x_*) = x_*$ is equivalent to $f(x_*) = y$. The sequence (x_n) converges to x_* and by (4.4) we obtain $\|x_{n+1} - x_n\|_X \leqslant \beta^n\|x_1 - x_0\|_X$ and

$$\|x_n - x_*\|_X \leqslant \frac{\beta^n}{1-\beta}\|x_0 - x_1\|_X.$$

Since $\|x_0 - x_1\|_X \leqslant \delta + \|x_0 - a\|_X$ we are done. □

We now turn to the Newton-Kantorovich iteration. Here we will slightly strengthen our assumptions, as we need a Lipschitz condition on Df for all points in a neighborhood of a and impose more restrictive assumptions on δ and $\|y - f(a)\|_Y$.

6.45. Theorem. *Let X, Y be Banach spaces, $\Omega \subset X$ be open. Let $f : \Omega \to Y$ be of class C^1 and $a \in \Omega$, and $r > 0$ such that $\{x : \|x-a\| \leqslant r\} \subset \Omega$. Assume that $Df|_a$ is invertible and let $M = \|Df|_a^{-1}\|_{\mathrm{op}}$. Assume that there exists $K > 0$ such that Df satisfies the Lipschitz condition at a*

$$\|Df|_x - Df|_{x'}\|_{\mathrm{op}} \leqslant K\|x - x'\|_X$$

for x, x' in $\overline{B}(a,r)$. Let $\delta > 0$ and y satisfy

$$\delta < \min\big\{r, \frac{2}{3MK}\big\}, \qquad \|y - f(a)\|_Y \leqslant \frac{\delta}{2M}(2 - 3MK\delta).$$

Define the Newton–Kantorovich iteration by

$$x_0 = a, \qquad x_n = \Phi_y(x_{n-1}),\ n \geqslant 1,\ \textit{with } \Phi_y(x) = x - Df|_x^{-1}(f(x) - y).$$

Then $x_n \in \overline{B}(a,\delta)$ for all $n \in \mathbb{N}$, and (x_n) converges to the unique $x_* \in \overline{B}(a,\delta)$ with $f(x_*) = y$. With $\gamma = 3MK\delta/2$, i.e. $\frac{\gamma}{3-2\gamma} < \gamma < 1$ the sequence (x_n) satisfies

$$\|x_{n+1} - x_n\|_X \leqslant \delta\Big(\frac{\gamma}{3-2\gamma}\Big)^{2^n-1}. \tag{6.11}$$

Proof. By Theorem 6.36, $Df|_x$ is invertible if $\|Df|_x - Df|_a\|_{\mathrm{op}} < M^{-1}$, which by the Lipschitz condition holds if $K\|x-a\|_X < M^{-1}$. Theorem 6.36 also gives

$$\big\|Df|_x^{-1}\big\|_{\mathrm{op}} \leqslant \frac{M}{1 - MK\|x-a\|_X} \leqslant \frac{M}{1-MK\delta}\,.$$

We first show that Φ_y maps $\overline{B} = \overline{B(a,\delta)}$ to itself. Here we need the assumed conditions $MK\delta < 2/3$ and $\|y - f(a)\|_Y \leqslant \frac{2-3MK\delta}{2M}$ (indeed, the latter condition is manufactured exactly to achieve $\Phi_y : \overline{B} \to \overline{B}$). We compute

$$\begin{aligned}\Phi_y(x) - a &= x - a - Df|_x^{-1}(f(x) - y)\\ &= Df|_x^{-1}\big(Df|_x(x-a) - f(x) + y\big)\\ &= Df|_x^{-1}\big(f(a) - f(x) - Df|_x(a-x)\big) + Df|_x^{-1}\big(y - f(a)\big)\,.\end{aligned}$$

Thus, using Lemma 6.43,

$$\begin{aligned}\|\Phi_y(x) - a\|_X &\leqslant \frac{M}{1-MK\delta}\Big(\frac{K\|x-a\|_X^2}{2} + \|y - f(a)\|_Y\Big)\\ &\leqslant \frac{M}{1-MK\delta}\Big(\frac{K\delta^2}{2} + \frac{\delta}{2M}(2 - 3MK\delta)\Big) = \delta.\end{aligned}$$

To estimate $x_{n+1} - x_n$ in the iteration we observe

$$\Phi_y(\Phi_y(x)) - \Phi_y(x) = -Df|_{\Phi_y(x)}^{-1}(f(\Phi_y(x)) - y)$$

and since $f(x) + Df|_x(\Phi_y(x) - x) = y$, by definition of $\Phi_y(x)$, we obtain

$$\Phi_y(\Phi_y(x)) - \Phi_y(x) = -Df|_{\Phi_y(x)}^{-1}\big(f(\Phi_y(x)) - f(x) - Df|_x(\Phi_y(x) - x)\big)\,.$$

By Lemma 6.43,

$$\|\Phi_y(\Phi_y(x)) - \Phi_y(x)\|_X \leqslant \frac{M}{1-MK\delta}\,\frac{K\|\Phi_y(x) - x\|_X^2}{2}\,.$$

We apply this successively to $x_1, x_2, \ldots$ and get

$$\|x_{n+1} - x_n\|_X \leqslant \frac{MK}{2(1-MK\delta)}\|x_n - x_{n-1}\|_X^2$$

for $n = 0, 1, 2, \ldots$. Now an iteration leads to

$$\|x_{n+1} - x_n\|_X \leqslant \delta\Big(\frac{MK\delta}{2(1-MK\delta)}\Big)^{2^n-1}. \tag{6.12}$$

We check this by induction. For $n = 0$ we have $x_0 = a$ and therefore $\|x_1 - x_0\|_X \leqslant \delta$ which is (6.12) in this case. For the induction step we assume the formula for $\|x_{n+1} - x_n\|_X$ and obtain for $n \geqslant 0$

$$\begin{aligned}\|x_{n+2} - x_{n+1}\|_X &\leqslant \frac{MK}{2(1-MK\delta)}\|x_{n+1} - x_n\|_X^2\\ &\leqslant \frac{MK}{2(1-MK\delta)}\delta^2\Big(\frac{MK\delta}{2(1-MK\delta)}\Big)^{2^{n+1}-2} = \delta\Big(\frac{MK\delta}{2(1-MK\delta)}\Big)^{2^{n+1}-1}.\end{aligned}$$

This proves (6.12). Setting $\gamma = 3MK\delta/2$, which satisfies $\gamma < 1$ by assumption, we can rewrite this estimate as (6.11). The sequence (x_n) converges to a point x_* satisfying $f(x_*) = y$ and the uniqueness of such a solution follows already from Theorem 6.44, even on the larger open ball $\{x : \|x - a\|_X < (MK)^{-1}\}$. □

REMARK. As an illustration consider in the assumptions the case $\delta = (15MK)^{-1}$ which gives us $\beta = \frac{1}{15}$ in Theorem 6.44 and $\gamma = \frac{1}{10}$, i.e. $\frac{\gamma}{3-2\gamma} = \frac{1}{28}$ in Theorem 6.45. The allowable size of $\|y - f(a)\|_Y$ is about twice as much in the modified Newton iteration compared to the assumption in the Newton-Kantorovich iteration. For the speed of convergence compare the bound with $\|x_{n+1} - x_n\|_X = O((\frac{1}{15})^n)$ in the modified Newton iteration with the corresponding $O((\frac{1}{28})^{2^n-1})$ bound in the Newton-Kantorovich iteration. In the first term we roughly add a fixed number of digits of accuracy (one or two) with each iteration. In the second we more than double the number of accurate digits with each iteration, which appears to be a spectacular improvement. Of course, in practice, the computation of $Df|_{x_n}^{-1}$ at every step poses many additional problems.

6.2.4. Implicit function theorem. Let $\Omega \subset \mathbb{R}^n \times \mathbb{R}^m$ be an open set and $f : \Omega \to \mathbb{R}^m$ a C^1 map. Consider the zero set of f given by

$$\mathcal{Z} = \{(x, y) \in \Omega \,:\, f(x, y) = 0\}.$$

It is natural to ask when $\mathcal{Z}$ is locally the graph of a function. The implicit function theorem gives a satisfactory answer. More precisely, given a point $(x_0, y_0) \in \mathcal{Z}$ we ask whether there exists an open neighborhood of (x_0, y_0) so that $\mathcal{Z}$ intersected with this neighborhood is given as the graph of a C^1 function, in the sense that there exists g so that $f(x, g(x)) = 0$ for x close to x_0. In other words, we are aiming to solve the system of equations given by $f(x, y) = 0$ for y, when x is given (this is reasonable as the number of equations and the number of dependent variables should be the same, here m). Notice that it is often difficult or impossible to algebraically determine a solution $g(x)$.

Let us assume that $(x, y) \to f(x, y)$ is an $\mathbb{R}^m$-valued function of class C^1 and that we already know the existence of a C^1 function $x \mapsto g(x)$ so that $f(x, g(x)) = 0$ in some open set $U \subset \mathbb{R}^n$. We can then differentiate this equation and express the derivative $Dg|_x$ as a function of x and $g(x)$. For this we split the Jacobian matrix for f as

$$(Df)|_{(x,y)} = \begin{pmatrix} D_x f & D_y f \end{pmatrix}\Big|_{(x,y)}$$

where $(Df)|_{(x,y)}$ is the $m \times (n + m)$ matrix representing the total derivative of f, $D_x f|_{(x,y)}$ is the $m \times n$ matrix representing for fixed y the derivative of $x \mapsto f(x, y)$, and $D_y f|_{(x,y)}$ is the $m \times m$ matrix representing for fixed x the derivative of $y \mapsto f(x, y)$. Denote by I_n the $n \times n$ identity matrix, and by $O_{m\times n}$ the $m \times n$ zero matrix. Differentiate the function $x \mapsto f(x, g(x))$ using the chain rule to obtain the implication

$$\begin{aligned} f(x, g(x)) = 0 \quad \Longrightarrow \quad O_{m\times n} &= \begin{pmatrix} D_x f|_{(x,g(x))} & D_y f|_{(x,g(x))} \end{pmatrix} \begin{pmatrix} I_n \\ Dg|_x \end{pmatrix} \\ &= D_x f|_{(x,g(x))} + D_y f|_{(x,g(x))} Dg|_x. \end{aligned}$$

Therefore, under the assumption that the $m \times m$ matrix $D_y f|_{(x,g(x))}$ is *invertible*, we can write this as

$$f(x, g(x)) = 0 \quad \Longrightarrow \quad Dg|_x = -(D_y f|_{(x,g(x))})^{-1} D_x f|_{(x,g(x))}.$$

This method is called *implicit differentiation.* It gives us information about the derivative of g in terms of x and $g(x)$. Of course, to make sense of this calculation we need to know that a solution $y = g(x)$ of $f(x, y) = 0$ *exists*. Such a result is provided in the following theorem. The above calculation suggests that we should assume invertibility of $D_y f$ at least at a reference point in $\mathcal{Z}$ (i.e. satisfying $f(x_0, y_0) = 0$); we then seek a solution near x_0.

6.46. THEOREM (Implicit function theorem). *Let $\Omega \subset \mathbb{R}^n \times \mathbb{R}^m$ be open, $f \in C^1(\Omega, \mathbb{R}^m)$ and $(x_0, y_0) \in \mathcal{Z}$ such that the matrix $D_y f|_{(x_0,y_0)} \in \mathbb{R}^{m\times m}$ is invertible. Then there exist an open neighborhood U of x_0, an open neighborhood V of y_0, and a C^1 function $g : U \to V$ so that*

$$\mathcal{Z} \cap (U \times V) = \{(x, g(x)) \,:\, x \in U\}.$$

In other words, $U \times V \subset \Omega$ and $f(x, g(x)) = 0$ for all $x \in U$. Moreover, for x in a neighborhood of x_0,

$$Dg|_x = -(D_y f|_{(x,g(x))})^{-1} D_x f|_{(x,g(x))}. \tag{6.13a}$$

In particular, $Dg|_{x_0}$ can be computed as

$$Dg|_{x_0} = -(D_y f|_{(x_0,y_0)})^{-1} D_x f|_{(x_0,y_0)}. \tag{6.13b}$$

PROOF. The proof is an application of the inverse function theorem, Theorem 6.39. Define a map $F : \Omega \to \mathbb{R}^n \times \mathbb{R}^m$ by

$$F(x, y) = (x, f(x, y)).$$

Then F is C^1 and the matrix of $DF|_{(x_0,y_0)}$ is the $(n + m) \times (n + m)$ block matrix

$$\begin{pmatrix} I_n & 0 \\ D_x f|_{(x_0,y_0)} & D_y f|_{(x_0,y_0)} \end{pmatrix}.$$

To see that $DF|_{(x_0,y_0)}$ is invertible it suffices to check that its nullspace is $\{0\}$. If $U = (u, v)^\intercal \in \mathbb{R}^n \times \mathbb{R}^m$ and $DF|_{(x_0,y_0)}U = 0$ we see that $u = I_n u = 0$ and thus $D_y f|_{(x_0,y_0)} v = 0$, hence $v = 0$ by the assumed invertibility of $D_y f|_{(x_0,y_0)}$. It follows that $DF|_{(x_0,y_0)}$ is invertible.

We can now conclude that F is locally invertible at (x_0, y_0), by Theorem 6.39. Consequently, there exist an open neighborhood U' of x_0 and an open neighborhood V of y_0 so that $U' \times V \subset \Omega$, $F(U' \times V) \subset \mathbb{R}^n \times \mathbb{R}^m$ is open and $F|_{U'\times V}$ is invertible with a C^1 inverse

$$G : F(U' \times V) \to U' \times V,$$

and G is of the form

$$G(x, w) = (x, G_2(x, w))$$

for (x, w) near $(x_0, 0)$. Let $U = \{x \in U' \,:\, (x, 0) \in F(U' \times V)\} \subset \mathbb{R}^n$. Then U is open, because $F(U' \times V)$ is open. Also, $x_0 \in U$. For $x \in U$ we can now define $g(x) = G_2(x, 0)$. Then $g(x) \in V$ for $x \in U$ and

$$(x, f(x, g(x))) = F(x, g(x)) = F(G(x, 0)) = (x, 0),$$

so $f(x, g(x)) = 0$ for all $x \in U$. Moreover, g is of class C^1 in U and satisfies (6.13). □

6.47. EXAMPLE. Let $n = m = 1$ and $f(x, y) = x^2 + y^2 - 1$. Then $\mathcal{Z}$ is the unit circle around the origin, which is locally a graph at every point with $y_0 \neq 0$. Coincidentally,

$$D_y f|_{(x,y)} = 2y \neq 0 \text{ if and only if } y \neq 0.$$

In this case an implicit function g can be determined explicitly: if say $(x_0, y_0) = (0, 1)$, then $g : (-1, 1) \to \mathbb{R}$ with

$$g(x) = \sqrt{1 - x^2}$$

is C^1 and satisfies $f(x, g(x)) = 0$.

6.48. EXAMPLE. Let $n = m = 1$ and $f(x, y) = x^2 - y^3$. Then $\mathcal{Z}$ is a cubic curve with a *cusp singularity* at the origin. In this case, $\mathcal{Z}$ is (globally) the graph of the function $g : \mathbb{R} \to \mathbb{R}$ with $g(x) = |x|^{2/3}$. However,

$$D_y f|_{(x,y)} = -3y^2.$$

so the implicit function theorem does not apply at the cusp $(x_0, y_0) = (0, 0) \in \mathcal{Z}$. This is consistent with the fact that g is not C^1 at zero.

6.49. Example. Let $n = m = 1$ and $f(x, y) = (y-x)(y+x)$. Then $\mathcal{Z}$ is locally the graph of a function at every point except for the origin, where it has a *self-intersection*.

6.3. Derivatives of higher order

In this section we introduce higher order derivatives for $\mathbb{R}$-valued functions defined in open sets of $\mathbb{R}^n$. We will give two applications, one to a local integrability result for vector fields (§6.3.2), and one to an approximation result, namely Taylor's theorem for multivariate functions (§6.3.4). The reader interested in the development of the theory of higher order differentiation in general normed spaces may want to consult the appendix §6.7 below.

6.3.1. Second order derivatives. We will limit consideration to multivariate real-valued functions defined in an open set $U \subset \mathbb{R}^n$. Assuming that $f : U \to \mathbb{R}$ is differentiable in U we have defined the first order derivative at $x \in U$ as a linear functional in $L(\mathbb{R}^n, \mathbb{R})$, and $Df|_x$ is given by $Df|_x[h] = (\nabla f(x))^\intercal h$. It is natural to define the second derivative as the first (Fréchet) derivative of $x \mapsto Df|_x$ if this function is indeed Fréchet differentiable; the second derivative is then an object in $L(\mathbb{R}^n, L(\mathbb{R}^n, \mathbb{R}))$. Since we can identify $L(\mathbb{R}^n, \mathbb{R})$ with $\mathbb{R}^n$ (by identifying the linear form $h \mapsto u^\intercal h$ with the vector u) it is more convenient to define the second derivative as the Fréchet derivative of the gradient:

6.50. Definition. Let $U \subset \mathbb{R}^n$ be open, $f : U \to \mathbb{R}$ be Fréchet differentiable and $a \in U$.

(i) We say that f has a second total derivative at a if the function $\nabla f : U \to \mathbb{R}^n$ is Fréchet differentiable at a. The derivative of ∇f at a is called the *Hessian of* f and its matrix with respect to the standard basis is the *Hessian matrix* $\mathrm{H}f|_a$.

(ii) We denote by $C^2(U)$ the class of functions $f : U \to \mathbb{R}$ such that ∇f is differentiable in U and $a \mapsto \mathrm{H}f|_a$ is continuous in U.

We recall that the gradient is given by $\nabla f(x) = \sum_{i=1}^n \partial_i f(x) e_i$. Thus its Fréchet derivative involves iterated partial derivatives

$$\partial_i(\partial_j f) = \partial_{x_i}(\partial_{x_j} f)$$

evaluated at a, here $i, j \in \{1, \dots, n\}$. We usually omit the parentheses and just write $\partial_i \partial_j f$. If f has a total second derivative as in Definition 6.50 we can use (6.6) to compute the Hessian matrix as the matrix with $\partial_j \partial_i f(a)$ in the jth entry of the ith row, i.e.

$$(\mathrm{H}f|_a) = (D\nabla f|_a) = \begin{pmatrix} \partial_1\partial_1 f(a) & \cdots & \partial_n\partial_1 f(a) \\ \vdots & \ddots & \vdots \\ \partial_1\partial_n f(a) & \cdots & \partial_n\partial_n f(a) \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

One expects the Hessian matrix to be symmetric because of the symmetry property of the difference operators Δ_h defined by

$$\Delta_h f(x) = f(x + h) - f(x).$$

By commutativity of addition we get $\Delta_h \Delta_k = \Delta_k \Delta_h$. Indeed,

$$\Delta_h\Delta_k f(x) = \Delta_k f(x + h) - \Delta_k f(x) = f(x + k + h) - f(x + h) - \big(f(x + k) - f(x)\big),$$
$$\Delta_k\Delta_h f(x) = \Delta_h f(x + k) - \Delta_h f(x) = f(x + h + k) - f(x + k) - \big(f(x + h) - f(x)\big).$$

One would like to apply this to $h = se_i$, $k = te_j$ since the existence of the mixed partial derivatives implies

$$\partial_i\partial_j f = \lim_{s\to 0}\lim_{t\to 0} \frac{\Delta_{se_i}\Delta_{te_j} f}{st}, \qquad \partial_j\partial_i f = \lim_{t\to 0}\lim_{s\to 0} \frac{\Delta_{te_j}\Delta_{se_i} f}{st}$$

and since $\Delta_{se_i}\Delta_{te_j} = \Delta_{te_j}\Delta_{se_i}$ we expect the differential operators ∂_i, ∂_j to commute. To prove such a result we have to be able to interchange the limits. The assumption of Fréchet differentiability of ∇f at a is crucial here:

6.51. THEOREM. *Let $f : U \to \mathbb{R}$ be differentiable in U and assume that ∇f is Fréchet differentiable at a. Then the second order partial derivatives satisfy*

$$\partial_i \partial_j f(a) = \partial_j \partial_i f(a)$$

for $i, j \in \{1, \dots, n\}$. Thus $\mathrm{H}f|_a$ is symmetric.

PROOF. We first note that the assumption of Fréchet differentiability of ∇f at a implies (in fact is equivalent to) the Fréchet differentiability of $\partial_i f$ at a, for $1 \leqslant i \leqslant n$. We examine the behavior of $\Delta_{te_i}\Delta_{te_j} f$ as $t \to 0$ and make use of the fact that this expression does not change if we switch e_i and e_j. Set $A = \mathrm{H}f|_a$ so that $e_i^\intercal A e_j = \partial_j \partial_i f(a)$. We claim

$$\Delta_{te_i}\Delta_{te_j} f(a) - t^2 e_i^\intercal A e_j = o(t^2) \tag{6.14}$$

as $t \to 0$. Because of $\Delta_{te_i}\Delta_{te_j} = \Delta_{te_j}\Delta_{te_i}$ this implies

$$t^2 |e_i^\intercal A e_j - e_j^\intercal A e_i| = o(t^2)$$

and after dividing by t^2 and letting $t \to 0$ we see that $e_i^\intercal A e_j = e_j^\intercal A e_i$. Hence the claim implies $\partial_j \partial_i f(a) = \partial_i \partial_j f(a)$.

It remains to check (6.14). We split the left hand side as $I(t) + II(t)$ where

$$\begin{aligned} I(t) &= \Delta_{te_i}\Delta_{te_j} f(a) - t\partial_i f(a + te_j) + t\partial_i f(a), \\ II(t) &= t\partial_i f(a + te_j) - t\partial_i f(a) - t^2 e_i^\intercal A e_j \,. \end{aligned}$$

Since $\partial_i f$ is Fréchet differentiable at a we get

$$\partial_i f(a + te_j) - \partial_i f(a) - te_i^\intercal A e_j \; = o(t).$$

and consequently $II(t) = o(t^2)$ as $t \to 0$.

It remains to show $I(t) = o(t^2)$. This is a bit more technical. Consider for β in a neighborhood of the interval $[0, 1]$ the function

$$t \mapsto G_t(\beta) = f(a + t\beta e_i + te_j) - f(a + t\beta e_i) - t\beta \partial_i f(a + te_j) + t\beta \partial_i f(a)$$

and check that

$$G_t(1) - G_t(0) = \Delta_{te_i}\Delta_{te_j} f(a) - t\partial_i f(a + te_j) + t\partial_i f(a).$$

By the mean value theorem $G_t(1) - G_t(0) = G_t'(\tau)$ with some $\tau = \tau(t) \in [0, 1]$. Hence it suffices to show that $|G_t'(\tau(t))| = o(t^2)$. We compute

$$G_t'(\tau(t)) = t\partial_i f(a + t\tau(t)e_i + te_j) - t\partial_i f(a + t\tau(t)e_i) - t\partial_i f(a + te_j) + t\partial_i f(a).$$

We use again the Fréchet differentiability of $\partial_i f$ and expand

$$\begin{aligned} \partial_i f(a + t\tau(t)e_i + te_j) &= \partial_i f(a) + t\tau(t)\partial_i^2 f(a) + t\partial_j \partial_i f(a) + o(t) \\ \partial_i f(a + t\tau(t)e_i) &= \partial_i f(a) + t\tau(t)\partial_i^2 f(a) + o(t) \\ \partial_i f(a + te_j) &= \partial_i f(a) + t\partial_j \partial_i f(a) + o(t). \end{aligned}$$

Using these in the expression for $G_t'(\tau(t))$ we get $G_t'(\tau(t)) = o(t^2)$ and thus $II(t) = o(t^2)$. This finishes the proof of (6.14). □

6.52. COROLLARY. *If $f \in C^2(U)$, then $\partial_i \partial_j f = \partial_j \partial_i f$ for every $i, j \in \{1, \dots, n\}$.*

It turns out that the mere existence of the two mixed partial derivatives is not enough to justify this:

6.53. Exercise. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} xy\frac{x^2-y^2}{x^2+y^2}, & \text{if } (x,y) \neq (0,0), \\ 0, & \text{if } (x,y) = (0,0). \end{cases}$$

Show that the mixed partial derivatives $\partial_x\partial_y f$ and $\partial_y\partial_x f$ both exist at every point in $\mathbb{R}^2$, but that $\partial_x\partial_y f(0,0) \neq \partial_y\partial_x f(0,0)$.

However, if one adds a stronger continuity assumption on one of the mixed derivatives $\partial_i\partial_j f$, $\partial_j\partial_i f$ it implies the existence of and the equality with the other. Note that in the following theorem no assumptions are made on the pure second order derivatives $\partial_i^2 f$, $\partial_j^2 f$ (in contrast to Theorem 6.51).

6.54. Theorem (Schwarz). *Let $U \subset \mathbb{R}^n$ be open. Fix $i, j \in \{1, \dots, n\}$, $i \neq j$. Let $f : U \to \mathbb{R}$ be such that $\partial_i f, \partial_j f$, $\partial_i\partial_j f$ exist at every point in U and $\partial_i\partial_j f$ is continuous at some point $a \in U$. Then $\partial_j\partial_i f(a)$ exists and*

$$\partial_j\partial_i f(a) = \partial_i\partial_j f(a).$$

Proof. We first consider the case $n = 2$, $i = 1$, $j = 2$ and then reduce to the general case.

Let f be as in the theorem and $a = (a_1, a_2) \in U$ and $(h_1, h_2) \in \mathbb{R}^2\backslash\{0\}$ such that $(a_1 + h_1, a_2 + h_2)$ is contained in an open ball around a that is contained in U. To show the existence of $\partial_2\partial_1 f$ we need to study the difference quotient

$$\frac{\partial_1 f(a_1, a_2 + h_2) - \partial_1 f(a_1, a_2)}{h_2}$$

for $h_2 \to 0$. This leads us to consider (as in the proof of Theorem 6.51) the quantity involving the "mixed second difference"

$$\begin{aligned} \mathcal{M}_h(a) &= \Delta_{h_2 e_2}[\Delta_{h_1 e_1} f](a) \\ &= (f(a_1 + h_1, a_2 + h_2) - f(a_1, a_2 + h_2)) - (f(a_1 + h_1, a_2) - f(a_1, a_2)). \end{aligned}$$

Define $g(t) = \Delta_{h_1 e_1} f(a_1, t) = f(a_1 + h_1, t) - f(a_1, t)$. Since $\partial_2 f$ exists at every point in U, the mean value theorem (Corollary 2.37) implies that there exists $\eta = \eta_h$ contained in the open interval with endpoints a_2 and $a_2 + h_2$ such that

$$\mathcal{M}_h(a) = g(a_2 + h_2) - g(a_2) = g'(\eta)h_2 = h_2(\partial_2 f(a_1 + h_1, \eta) - \partial_2 f(a_1, \eta))$$

Since $\partial_1\partial_2 f$ exists at every point in U, another application of the mean value theorem yields

$$\mathcal{M}_h(a) = h_1 h_2 \cdot \partial_1\partial_2 f(\xi, \eta),$$

where $\xi = \xi_{h_1,\eta}$ is in the open interval with endpoints a_1 and $a_1 + h_1$.

Let $\varepsilon > 0$. Since $\partial_1\partial_2 f$ is continuous at (a_1, a_2),

$$|\partial_1\partial_2 f(a) - \partial_1\partial_2 f(x)| \leqslant \varepsilon$$

whenever $\|a - x\|$ is small enough. Thus, for small enough h_1 and h_2 we have

$$|\partial_1\partial_2 f(a) - \frac{\mathcal{M}_h(a)}{h_1 h_2}| \leqslant \varepsilon. \tag{6.15}$$

Letting $h_1 \to 0$ and using that $\partial_1 f$ exists at every point we get for $h_2 \neq 0$

$$\begin{aligned} \lim_{h_1\to 0} \frac{\mathcal{M}_h(a)}{h_1 h_2} &= \lim_{h_1\to 0} \frac{\frac{f(a_1+h_1,a_2+h_2)-f(a_1,a_2+h_2)}{h_1} - \frac{f(a_1+h_1,a_2)-f(a_1,a_2)}{h_1}}{h_2} \\ &= \frac{\partial_1 f(a_1, a_2 + h_2) - \partial_1 f(a_1, a_2)}{h_2} \end{aligned}$$

and inequality (6.15) implies for sufficiently small h_2

$$\left|\partial_1\partial_2 f(a) - \frac{\partial_1 f(a_1, a_2 + h_2) - \partial_1 f(a_1, a_2)}{h_2}\right| \leqslant \varepsilon.$$

Since ε was arbitrary, $\partial_2\partial_1 f(a)$ exists and equals $\partial_1\partial_2 f(a)$.

Finally we generalize the above for functions of n variables, for any $n \geqslant 2$. Let e_i, e_j be different unit coordinate vectors and consider the function of two variables

$$F(s_1, s_2) = f(a + s_1 e_i + s_2 e_j)$$

for (s_1, s_2) near the origin. The assumption that $\partial_i\partial_j f$ is continuous at a implies that $\partial_1\partial_2 F$ is continuous at $(0, 0)$ and so by the above $\partial_2\partial_1 F(0) = \partial_1\partial_2 F(0)$. From this and the definitions of partial derivatives we conclude that $\partial_j\partial_i f(a)$ exists and is equal to $\partial_i\partial_j f(a)$. □

6.3.2. Local integrability of vector fields. We study an important application of Theorem 6.51. It is natural to ask when a vector-valued function $F : U \to \mathbb{R}^n$ on an open set $U \subset \mathbb{R}^n$ can be written as the gradient of a function $g : U \to \mathbb{R}$, i.e. we would like to know when there exists g such that

$$F = \nabla g.$$

Theorem 6.51 gives us necessary conditions. Namely if $F = \nabla g$ for some $g \in C^2$ it needs to satisfy the $\frac{n(n-1)}{2}$ *compatibility conditions*

$$\partial_j F_i = \partial_i F_j, \quad 1 \leqslant i < j \leqslant n. \tag{6.16}$$

Indeed, if $F_j = \partial_j g$, for $j = 1, \ldots, n$ then $\partial_i F_j = \partial_i\partial_j g = \partial_j\partial_i g = \partial_j F_i$. This leads us to the question whether these compatibility conditions are also sufficient. If U is convex, we can use the fundamental theorem of calculus to show this.

6.55. Theorem. *Let $U \subset \mathbb{R}^n$ be open and convex, let F be of class C^1, and suppose the compatibility condition* (6.16) *holds on U. Then $F = \nabla g$ on U, for some $g \in C^2(U)$.*

Proof. Fix a in U. By convexity of U the line segment between a and x is contained in U, for all $x \in U$. We can then define

$$g(x) = \int_0^1 (x - a)^\intercal F(a + t(x - a))dt$$

We remark that this is the standard calculus definition of a *line integral* of the vector field F along the line parametrized by $\gamma(t) = a + t(x - a)$ with $\gamma'(t) = x - a$.

We claim that $F = \nabla g$. Compute by differentiating under the integral sign (see Example 4.78),

$$\begin{aligned}
\partial_i g(x) &= \partial_i \int_0^1 \sum_{j=1}^n (x_j - a_j) F_j(a + t(x - a))dt \\
&= \int_0^1 \Big[F_i(a + t(x - a)) + \sum_{j=1}^n (x_j - a_j)\partial_i F_j(a + t(x - a))t\Big]\, dt \\
&= \int_0^1 \Big[F_i(a + t(x - a)) + \sum_{j=1}^n (x_j - a_j)\partial_j F_i(a + t(x - a))t\Big]\, dt
\end{aligned}$$

by the compatibility condition $\partial_i F_j = \partial_j F_i$. Observe using the chain rule that $\frac{d}{dt}\big(F_i(a+t(x-a))\big)$ equals $\sum_{j=1}^n (x_j - a_j)\partial_j F_i(a+t(x-a))$ so that the last displayed expression is equal to

$$\begin{aligned}&\int_0^1 F_i(a+t(x-a)) + t\frac{d}{dt}\big(F_i(a+t(x-a))\big)\,dt\\ &= \int_0^1 \frac{d}{dt}\big(tF_i(a+t(x-a))\big)\,dt = 1F_i(x) - 0F_i(a) = F_i(x).\end{aligned}$$

Here we have used, in the last displayed line, the product rule and then the fundamental theorem of calculus. We have shown $\partial_i g = F_i$, for all i and hence $\nabla g = F$. □

Theorem 6.55 is a special case of a general theory on differential forms in $\mathbb{R}^n$ (see e.g. Spivak [**71**] and Tu [**83**]). We remark that the theorem also holds under the weaker assumption that U is connected and in addition *simply connected* (i.e. every curve can be continuously contracted to a point in U). Otherwise it can fail, see Example 6.57. To prepare for this, we first make the following observation which follows directly from the fundamental theorem of calculus and the chain rule.

6.56. Observation. *Let U be an open set in $\mathbb{R}^n$ and let $\gamma : [0,1] \to U$ be a C^1 function. Then for any C^1 function g on U*

$$\int_0^1 \sum_{j=1}^n \partial_j g(\gamma(t))\gamma_j'(t)dt = \int_0^1 \frac{d}{dt}[g(\gamma(t))]dt = g(\gamma(1)) - g(\gamma(0)) \tag{6.17}$$

so the right hand side depends on the values of g at $\gamma(1)$ and $\gamma(0)$ but not on $\gamma(t)$ for $0 < t < 1$.

6.57. Example. Now consider the case $n = 2$, $U = \mathbb{R}^2\backslash\{0\}$ and let

$$F(x_1, x_2) = \big(\tfrac{-x_2}{x_1^2+x_2^2}, \tfrac{x_1}{x_1^2+x_2^2}\big).$$

We check that both $\partial_2 F_1$ and $\partial_1 F_2$ are equal to $\frac{x_2^2 - x_1^2}{(x_1^2+x_2^2)^2}$ on U. We now show that there is no $C^2(U)$ function g with $F = \nabla g$ on U. To see this define γ, $\tilde\gamma$ on $[0,1]$ by $\gamma(t) = (\cos(\pi t), \sin(\pi t))$ and $\tilde\gamma(t) = (\cos(-\pi t), \sin(-\pi t))$ and note that $\gamma(t), \tilde\gamma(t) \in U$, with $\gamma(0) = \tilde\gamma(0) = (1,0)$, and $\gamma(1) = \tilde\gamma(1) = (-1,0)$. Also

$$\begin{aligned}\int_0^1 F_1(\gamma(t))\gamma_1'(t) + F_2(\gamma(t))\gamma_2'(t)dt &= \int_0^1 (\sin^2(\pi t) + \cos^2(\pi t))\pi\,dt = \pi\\ \int_0^1 F_1(\tilde\gamma(t))\tilde\gamma_1'(t) + F_2(\tilde\gamma(t))\tilde\gamma_2'(t)dt &= \int_0^1 (\sin^2(-\pi t) + \cos^2(-\pi t))(-\pi)\,dt = -\pi\end{aligned}$$

If we had $\nabla g = F$ with $g \in C^2(U)$ then the two integrals would have to be equal, by (6.17).

We explore this example further. Let $\mathcal{V}_{\rm r} = \{(x_1, x_2) : x_1 > 0\}$ be the right half-plane. By Theorem 6.55, we find $g_{\rm r} : \mathcal{V}_{\rm r} \to \mathbb{R}$ with $\nabla g_{\rm r} = F$ and this function is unique up to additive constants. By integrating F with respect to x_1 and x_2 separately we compute that $g_{\rm r}(x) - \arctan(x_2/x_1)$ is constant on $\mathcal{V}_{\rm r}$. If in addition we require $g_{\rm r}(1,0) = 0$ we get

$$g_{\rm r}(x) = \arctan(x_2/x_1) \text{ for } x \in \mathcal{V}_{\rm r}.$$

Thus $g_{\rm r}(x)$ is the angle that the vector (x_1, x_2) makes with the positive x_1-axis.

Similarly we can repeat this on the upper and lower half-planes $\mathcal{V}_+ = \{(x_1, x_2) : x_2 > 0\}$, $\mathcal{V}_- = \{(x_1, x_2) : x_2 < 0\}$. We find that

$$g_+ : \begin{cases}\mathcal{V}_+ \to \mathbb{R}\\ x \mapsto \frac{\pi}{2} - \arctan(x_1/x_2)\end{cases} \qquad g_- : \begin{cases}\mathcal{V}_- \to \mathbb{R}\\ x \mapsto -\frac{\pi}{2} - \arctan(x_1/x_2)\end{cases}$$

are functions with $\nabla g_\pm = F$ on $\mathcal{V}_\pm$ and one checks (using $\cot(\pi - \alpha) = \tan(\alpha)$) that $g_{\rm r} = g_+$ on $\mathcal{V}_+ \cap \mathcal{V}_{\rm r}$ and $g_{\rm r} = g_-$ on $\mathcal{V}_- \cap \mathcal{V}_{\rm r}$. Thus we can define the angle function g on $\mathbb{R}^2\backslash\{(x_1, 0) : x_1 \leqslant 0\} = \mathcal{V}_+ \cup \mathcal{V}_{\rm r} \cup \mathcal{V}_-$ as a C^2 function such that g coincides with $g_{\rm r}$, g_+, g_- in $\mathcal{V}_{\rm r}$, $\mathcal{V}_+$, $\mathcal{V}_-$, respectively. Note that for $x_1 < 0$

we have $\lim_{x_2\to 0+} g(x_1,x_2)=\pi$ and $\lim_{x_2\to 0-} g(x_1,x_2)=-\pi$ so that g cannot be defined globally as a continuous function on $\mathbb{R}^2\backslash\{0\}$.

6.3.3. Higher order partial derivatives.

6.58. Definition. (i) Let $U\subset\mathbb{R}^n$ be an open set and $f:U\to\mathbb{R}$. Let $k\in\mathbb{N}$. If all partial derivatives of f up to order k exist, i.e. if for all $j\in\{1,\dots,k\}$ and $i_1,\dots,i_j\in\{1,\dots,n\}$, the $\partial_{i_1}\cdots\partial_{i_j}f$ exist and are continuous, then we write $f\in C^k(U)$ and say that f is k times differentiable.

(ii) If $f:U\to\mathbb{R}^m$ then we say that $f\in C^k(U)$ if the coordinate functions $f_1,\dots,f_m$ are in $C^k(U)$.

We note using applications of Theorem 6.32 that the definition for $k=1,2$ is consistent with our previous definitions of the classes $C^1(U)$, $C^2(U)$.

An iterative application of Corollary 6.52 gives

6.59. Corollary. *If $f\in C^k(U)$ and $\pi:\{1,\dots,k\}\to\{1,\dots,k\}$ is a bijection, then*

$$\partial_{i_1}\cdots\partial_{i_k}f=\partial_{i_{\pi(1)}}\cdots\partial_{i_{\pi(k)}}f$$

for all $i_1,\dots,i_k\in\{1,\dots,n\}$.

6.60. Proposition. *In what follows $U\subset\mathbb{R}^n$ is an open set.*

(i) If $f\in C^k(U)$, and $1\leqslant l\leqslant k$ then any partial derivative of order l of f belongs to $C^{k-l}(U)$.

(ii) If $a_1,\dots,a_N:U\to\mathbb{R}$ are C^k functions then their product $\prod_{i=1}^N a_i$ is a C^k function.

(iii) If $g:U\to\mathbb{R}^m$ is in $C^k(U)$, if V is an open set in $\mathbb{R}^m$ containing $g(U)$ and if $f:V\to\mathbb{R}^p$ is in $C^k(U)$ then $f\circ g:U\to\mathbb{R}^p$ belongs to $C^k(U)$.

Sketch of proof. (i) This follows by induction on k.

(ii) We note by iteration of the product rule

$$\partial_j\Big(\prod_{i=1}^N a_i\Big)|_x=\sum_{k=1}^N\partial_j a_k|_x\prod_{\substack{i\in\{1,\dots,N\}\\ i\neq k}}a_i|_x$$

Now argue by induction, using that in the induction step $\partial_j a_k\prod_{\substack{i\in\{1,\dots,N\}\\ i\neq k}}a_i$ belongs to $C^{l-1}(U)$ if $a_1,\dots,a_N\in C^l(U)$.

(iii) We may assume that $p=1$. We again use induction on k. Assuming the statement is true for $k-1$ we use the chain rule

$$\partial_l(f\circ g)|_x=\sum_{\nu=1}^m\partial_\nu f|_{g(x)}\partial_l g_\nu|_x.$$

We then rely on (i), (ii) to carry out the induction step. □

We can now formulate versions of the inverse and implicit function theorems in Euclidean spaces.

6.61. Theorem (Inverse function theorem, C^k version). *Let $\Omega\subset\mathbb{R}^n$ be open, $k\in\mathbb{N}$, and let $f:\Omega\to\mathbb{R}^n$ be of class C^k on Ω. Let $a\in\Omega$ and assume that $Df|_a$ is invertible. Then f is locally invertible at a in some open neighborhood $U\subset\Omega$ of a with $(f|_U)^{-1}$ of class C^k on $V=f(U)$.*

Proof. For $k=1$ this has been proved in Theorem 6.39 already. Assume $k>1$ and that the C^{k-1} version of the theorem holds.

Denote by $g=(f|U)^{-1}$ the local inverse, $g:V\to U$, then we have the formula $Dg|_{f(x)}=(Df|_x)^{-1}$ for $x\in U$. Hence, for $y\in V$

$$Dg|_y=(Df|_{g(y)})^{-1}.$$

Let $A(x)=(Df|_x)$. We use that the determinant of a matrix is a sum of products of matrix coefficients, $\det Df|_{g(y)}\neq 0$ and Cramer's rule for the inverse of a matrix. We see then that the matrix coefficients

of $A(g(y))^{-1}$ are sums of expressions of the form $\frac{1}{\det A(g(y))}\prod_{k=1}^{n} b_k(g(y))$ where the b_k are matrix coefficients of $A(g(y))$. By the induction hypothesis $g \in C^{k-1}(V)$. It follows from Proposition 6.60 that the matrix coefficients of $Dg|_y$ are C^{k-1} functions on V. Hence $g \in C^k(V)$. □

Similarly, one proves the following C^k version of the implicit function theorem.

6.62. THEOREM (Implicit function theorem). *Let $\Omega \subset \mathbb{R}^n \times \mathbb{R}^m$ be open, $f \in C^k(\Omega, \mathbb{R}^m)$ and let (x_0, y_0) be such that $f(x_0, y_0) = 0$ and $D_y f|_{(x_0,y_0)} \in \mathbb{R}^{m\times m}$ is invertible. Then there exist an open neighborhood U of x_0, an open neighborhood V of y_0, and a C^k function $g : U \to V$ so that for $(x, y) \in U \times V$ the equation $f(x, y) = 0$ holds if and only if $y = g(x)$.*

SKETCH OF PROOF. For $k = 1$ this was already proved in Theorem 6.46. For $k > 1$ we use induction. Also by Theorem 6.46,

$$Dg|_x = -(D_y f_{(x,g(x))})^{-1} D_x f|_{(x,g(x))}.$$

Note that the right hand side involves only x and $g(x)$ but no derivative of $g(x)$. We may thus argue as in the proof of Theorem 6.61 and use Cramer's rule and Proposition 6.60 to carry out the induction step. □

6.3.4. Taylor's theorem. We now discuss a version of Taylor's theorem for C^{k+1} functions of several variables (see Theorem 2.49 for the one-variable version). Given a vector $h \neq 0$, recall that the directional derivative at $a \in \mathbb{R}^n$ is given by

$$D_h f(a) = \sum_{j=1}^{n} h_j \partial_j f(a) = \langle h, \nabla f\rangle_a.$$

We also set $D_h^0 f = f$, define higher directional derivatives by $D_h^j = D_h D_h^{j-1}$, $j \leqslant k + 1$.

6.63. THEOREM (Taylor, first version). *Let $U \subset \mathbb{R}^n$ be open and convex, $f : U \to \mathbb{R}$ of class C^{k+1} in U and $a, a + h \in U$. Then*

$$f(a + h) = \sum_{j=0}^{k} \frac{D_h^j f(a)}{j!} + \mathcal{R}_k f(a, h)$$

where

$$\mathcal{R}_k f(a, h) = \frac{1}{k!}\int_0^1 (1 - s)^k D_h^{k+1} f(a + sh)ds = \frac{D_h^{k+1} f(\xi)}{(k + 1)!}$$

and ξ is a point on the open line segment connecting a and $a + h$ (i.e. ξ takes the form $\xi = a + \theta h$ for some $\theta \in (0, 1)$).

PROOF. An application of the standard Taylor formula for $s \mapsto g(s)$ (see Theorems 2.49, 2.50) gives

$$g(1) = \sum_{j=0}^{k} \frac{g^{(j)}(0)}{j!} + R_k g(0, 1)$$

where

$$R_k g(0, 1) = \frac{1}{k!}\int_0^1 (1 - s)^k g^{(k+1)}(s)ds = \frac{g^{(k+1)}(\theta)}{(k + 1)!}$$

for some $\theta \in (0, 1)$.

For $s \in [0, 1]$ (in fact for s in a slightly larger interval) define $g(s) = f(a + sh)$ and observe that $g'(s) = D_h f(a + sh)$, and similarly $g^{(j)}(s) = D_h^j f(a + sh)$. The assertion follows. □

We write out more explicitly the remainder terms for $k = 1$.

6.64. Corollary. *Let $U \subset \mathbb{R}^n$ be open and convex and let $a \in U$.*

(i) Let $f : U \to \mathbb{R}$ be in $C^2(U)$. Then for $x \in U$

$$f(x) = f(a) + (x-a)^\intercal \nabla f(a) + E(x,a) \tag{6.18}$$

with

$$E(x,a) = \frac{1}{2}(x-a)^\intercal A(x,a)(x-a)$$

where $A(x,a) = (a_{ij}(x,a))$ is the symmetric matrix with

$$a_{ij}(x,a) = \int_0^1 2(1-s)\partial_j\partial_i f(a+s(x-a))\,ds = \partial_j\partial_i f(a) + o(1)$$

as $x \to a$.

(ii) Let $f : U \to \mathbb{R}$ be in $C^3(U)$. Then for $x \in U$

$$f(x) = f(a) + (x-a)^\intercal \nabla f(a) + \frac{1}{2}(x-a)^\intercal \mathrm{H}f|_a (x-a) + O(\|x-a\|^3).$$

Proof of Corollary 6.64. We compute that

$$D_h^2 f(a+sh) = \sum_{j=1}^n \sum_{i=1}^n \partial_i\partial_j f(a+sh) h_i h_j = h^\intercal \mathrm{H}f|_{a+sh} h$$

and apply Theorem 6.63 with $k=1$. Then set $h = x-a$ and use that $\|\mathrm{H}f|_x - \mathrm{H}f|_a\| \to 0$ as $x \to a$.

Similarly, for (ii) use $D_h^3 f(a+sh) = \sum_{1\leqslant i,j,k\leqslant n} \partial_i\partial_j\partial_k f(a+sh) h_i h_j h_k$ which is an $O(\|h\|^3)$ term. Now apply Theorem 6.63 with $k=2$. □

In the case that $\mathrm{H}f(a)$ is invertible the remainder term in the preceding corollary also can be written as a quadratic form in a slightly perturbed coordinate system:

6.65. Theorem (Morse lemma). *Let U, a, and f be as in Corollary 6.64 and assume that $\mathrm{H}f|_a$ is invertible. There is a continuous $\mathbb{R}^n$-valued function $x \mapsto y(x)$, defined for x near a, such that*

$$y(x) = x - a + o(\|x-a\|),$$

and for x sufficiently close to a

$$f(x) - f(a) = (x-a)^\intercal \nabla f(a) + \frac{1}{2} y(x)^\intercal \mathrm{H}f|_a y(x).$$

Proof. We need to show that the remainder term $E(x,a) = \frac{1}{2}(x-a)^\intercal A(x,a)(x-a)$ in Taylor's formula (6.18) can be expressed as

$$E(x,a) = \frac{1}{2} y(x)^\intercal \mathrm{H}f|_a y(x)$$

for a suitable change of variable $x \mapsto y(x)$. Its definition will rely on the following:

Claim: Let A be a fixed invertible symmetric matrix. Then for every $\varepsilon > 0$ there exists a $\delta > 0$ so that for all symmetric matrices B with $\|B - A\| < \delta$ there exists an $M = \mathcal{M}(B) \in \mathbb{R}^{n\times n}$ with $\|M - I\| < \varepsilon$ so that $B = M^\intercal A M$. The function $B \mapsto \mathcal{M}(B)$, defined for $\|B - A\| < \delta$ is of class C^1 and satisfies $\mathcal{M}(A) = I$ (the identity matrix).

Taking the claim for granted we apply it to $A = \mathrm{H}f|_a$, and $B(x) = A(x,a)$ as in Corollary 6.64, which is close to $A = A(a,a)$ provided that x is close to a. Define $M(x) = \mathcal{M}(A(x,a))$, then M is continuous and $M(a) = I$. Set

$$y(x) = M(x)(x-a).$$

Then $y(x) = (x-a) + o(\|x-a\|)$ and

$$2E(x,a) = (x-a)^\intercal A(x,a)(x-a) = (x-a)^\intercal M(x)^\intercal A M(x)(x-a) = y(x)^\intercal A y(x)$$

and the assertion follows from Corollary 6.64.

It remains to prove the claim. We shall use the inverse function theorem in the form of Theorem 6.42. Let $X = \mathbb{R}^{n\times n}$, the n^2-dimensional space of real $n \times n$ matrices, and $Y = \mathrm{Sym(n)}$ the $\frac{n(n+1)}{2}$-dimensional space of symmetric $n \times n$ matrices. We consider the function $F : \mathbb{R}^{n\times n} \to \mathrm{Sym}(n)$ given by $F(M) = M^\intercal AM$, which satisfies $F(I) = A$. To compute the derivative of F expand, for small $H \in \mathbb{R}^{n\times n}$,

$$F(M + H) - F(M) = H^\intercal AM + M^\intercal AH + H^\intercal AH.$$

The quadratic term satisfies $\|H^\intercal AH\| = O(\|H\|^2) = o(\|H\|)$ as $\|H\| \to 0$. Thus $DF|_M : \mathbb{R}^{n\times n} \to \mathrm{Sym}(n)$ is given by $DF_M[H] = H^\intercal AM + M^\intercal AH$ and the map $M \to DF_M$ is continuous as a map from $\mathbb{R}^{n\times n}$ to $L(\mathbb{R}^{n\times n}, \mathrm{Sym}(n))$. Evaluate DF_M for $M = I$ to get

$$DF_I(H) = H^\intercal A + AH.$$

The map $DF|_I : \mathbb{R}^{n\times n} \to \mathrm{Sym}(n)$ is surjective since for any symmetric matrix S the matrix $H_S = \frac{1}{2}A^{-1}S$ satisfies $DF_I(H_S) = S$. Indeed,

$$AH_S = \tfrac{1}{2}AA^{-1}S = \tfrac{1}{2}S, \quad \text{and} \quad H_S^\intercal A = \tfrac{1}{2}S^\intercal (A^{-1})^\intercal A = \tfrac{1}{2}S^\intercal (A^\intercal)^{-1}A = \tfrac{1}{2}SA^{-1}A = \tfrac{1}{2}S$$

and thus $H_S^\intercal A + AH_S = S$. By Theorem 6.42 there exists a neighborhood V of A and a C^1 function $\mathcal{M} : V \to \mathbb{R}^{n\times n}$ satisfying $\mathcal{M}(A) = I$, and $\|\mathcal{M}(B) - I\| < \varepsilon$ for $B \in V$ such that $F(\mathcal{M}(B)) = B$. The claim is proved. □

6.66. Exercise. Assume $f \in C^k(U)$ for $k > 2$, $D^2 f|_a$ invertible for some $a \in U$. Then show that one can find $x \mapsto y(x)$ with the specifications in Theorem 6.65 so that y is of class C^{k-2} near a.

6.67. Exercise. One may also look into the dependence on the parameter a. By examining the above proof, show that the change of variable $x \mapsto y(x, a)$ in Theorem 6.65 also depends continuously on a.

Remark. The Morse lemma and its refinements are especially useful in situations where $\nabla f(a) = 0$ in which case we say that a is a *critical point* of f. If a is a *nondegenerate critical point,* i.e. when $\det \mathrm{H} f(a) \neq 0$ the Morse lemma says that $f(x) - f(a)$ is expressed near a as a quadratic form in the variable $y = y(x)$.

We now return to the general form of Taylor's formula and seek to explicitly compute the term $D_h^m f(a)$ in terms of the higher partial derivatives (just as we did in the case $m = 2$). As in the case $m = 2$ an m-fold application of the definition of D_h yields the following formula involving m summations.

$$D_h^m f(a) = \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h_{i_1} \dots h_{i_m} \partial_{i_1} \cdots \partial_{i_m} f(a). \tag{6.19}$$

Convince yourself by an induction proof that this formula is correct.

Multiindex notation. One would like to use Corollary 6.59 to combine equal terms in (6.19) with the same mixed derivatives. For this it is convenient to introduce *multiindex notation.* A *multiindex of order* k is a vector $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n = \{0, 1, 2, \dots\}^n$ such that $\sum_{i=1}^n \alpha_i = k$. We write $|\alpha| = \sum_{i=1}^n \alpha_i$. For every multiindex α we introduce the notation

$$\partial^\alpha f = \partial_1^{\alpha_1} \cdots \partial_n^{\alpha_n} f, \tag{6.20a}$$

where $\partial_i^{\alpha_i}$ is short for $\partial_i \cdots \partial_i$ (α_i times). For $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ we also write

$$x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n} \tag{6.20b}$$

and $\alpha! = \alpha_1! \cdots \alpha_n!$. Moreover, for $\alpha, \beta \in \mathbb{N}_0^n$, $\alpha \leqslant \beta$ means that $\alpha_i \leqslant \beta_i$ for every $i \in \{1, \dots, n\}$.
With this notation, we can quite succinctly state Taylor's theorem in $\mathbb{R}^n$ using partial derivatives.

6.68. THEOREM (Taylor, second version). *Let $U \subset \mathbb{R}^n$ be open and convex, $f \in C^{k+1}(U)$ and $x, x+h \in U$. Then there exists ξ on the open line segment between x and $x+h$ such that*

$$f(x+h) = \sum_{|\alpha|\leqslant k} \frac{\partial^\alpha f(x)}{\alpha!} h^\alpha + \sum_{|\alpha|=k+1} \frac{\partial^\alpha f(\xi)}{\alpha!} h^\alpha.$$

REMARK. Without multiindex notation the statement of this theorem would look more messy:

$$\sum_{|\alpha|\leqslant k} \frac{\partial^\alpha f(x)}{\alpha!} h^\alpha = \sum_{\substack{\alpha_1,\dots,\alpha_n \geqslant 0,\\ \alpha_1+\cdots+\alpha_n \leqslant k}} \frac{\partial_1^{\alpha_1} \cdots \partial_n^{\alpha_n} f(x)}{\alpha_1! \cdots \alpha_n!} h_1^{\alpha_1} \cdots h_n^{\alpha_n}.$$

PROOF OF THEOREM 6.63. We apply Theorem 6.68 and then (6.19) and the proof reduces to checking that

$$\sum_{i_1,\dots,i_m=1}^{n} \partial_{i_1} \cdots \partial_{i_m} f(x+sh) h_{i_1} \cdots h_{i_m} = \sum_{|\alpha|=m} \frac{m!}{\alpha!} \partial^\alpha f(x+sh) h^\alpha.$$

This follows because for a given $\alpha = (\alpha_1, \dots, \alpha_n)$ with $|\alpha| = m$ there are

$$\frac{m!}{\alpha!} = \frac{m!}{\alpha_1! \cdots \alpha_n!} = \binom{m}{\alpha_1} \binom{m-\alpha_1}{\alpha_2} \cdots \binom{m-\alpha_1-\cdots-\alpha_{n-1}}{\alpha_n}$$

many tuples $(i_1, \dots, i_m) \in \{1, \dots, n\}^m$ such that i appears exactly α_i times among the i_js. In other words, this is the number of ways to sort m pairwise different marbles into n numbered bins such that bin number i contains exactly α_i marbles. □

6.69. COROLLARY. *If $\Omega \subset \mathbb{R}^n$ is open and $f \in C^k(\Omega)$, then for every $x \in \Omega$,*

$$f(x+h) = \sum_{|\alpha|\leqslant k} \frac{\partial^\alpha f(x)}{\alpha!} h^\alpha + o(\|h\|^k) \quad \text{as } h \to 0.$$

PROOF. Let $x \in \Omega$ and $\delta > 0$ be small enough so that $U = B_\delta(x) \subset \Omega$. By Taylor's theorem we have for every h with $x + h \in U$ that

$$\begin{aligned} f(x+h) &= \sum_{|\alpha|\leqslant k-1} \frac{\partial^\alpha f(x)}{\alpha!} h^\alpha + \sum_{|\alpha|=k} \frac{\partial^\alpha f(x+\theta h)}{\alpha!} h^\alpha \\ &= \sum_{|\alpha|\leqslant k} \frac{\partial^\alpha f(x)}{\alpha!} h^\alpha + \sum_{|\alpha|=k} \frac{\partial^\alpha f(x+\theta h) - \partial^\alpha f(x)}{\alpha!} h^\alpha \end{aligned}$$

for some $\theta \in (0,1)$. Since $\partial^\alpha f$ is continuous for every $|\alpha| = k$, it holds that

$$|\partial^\alpha f(x+\theta h) - \partial^\alpha f(x)| \to 0 \quad \text{as } h \to 0.$$

Also $|h^\alpha| = |h_1|^{\alpha_1} \cdots |h_n|^{\alpha_n} \leqslant \|h\|^{\alpha_1+\cdots+\alpha_n} = \|h\|^{|\alpha|}$, so

$$\sum_{|\alpha|=k} \frac{\partial^\alpha f(x+\theta h) - \partial^\alpha f(x)}{\alpha!} h^\alpha = o(\|h\|^k).$$

□

6.70. EXERCISE. Provide a version of Theorem 6.68 with integral remainder.

6.4. Extreme values and optimization

6.4.1. Local extrema. Let $\Omega \subset \mathbb{R}^n$ be an open set and $f : \Omega \to \mathbb{R}$ a function.

6.71. Definition. We say that f has a *local maximum* at $a \in \Omega$ if there exists an open set $U \subset \Omega$ with $a \in U$ such that $f(a) \geqslant f(x)$ for all $x \in U$. In this case we call a a *local maximum point.* We say *strict local maximum (point)* if $f(a) > f(x)$ for all $x \in U$, $x \neq a$. We define the terms *local minimum (point), strict local minimum (point)* accordingly.

A *local extremum* is a local maximum or a local minimum. A *strict local extremum* is a strict local maximum or a strict local minimum.

6.72. Theorem. *Suppose the partial derivative $\partial_i f$ exists on Ω. Then, if f has a local extremum at $a \in \Omega$, then $\partial_i f(a) = 0$.*

Proof. Let $\delta > 0$ be such that $a + te_i \in \Omega$ for all $|t| \leqslant \delta$. Define $g : (-\delta, \delta) \to \mathbb{R}$ by $g(t) = f(a + te_i)$. By the chain rule, g is differentiable and $g'(t) = \partial_i f(a + te_i)$. Also, 0 is a local extremum of g so, by Theorem 2.34, $0 = g'(0) = \partial_i f(a)$. □

6.73. Corollary. *If f is differentiable at a and a is a local extremum point, then $\nabla f(a) = 0$.*

Remark. $\nabla f(a) = 0$ is not a sufficient condition for f to have a local extremum at a.

6.74. Definition. If $a \in \Omega$ is such that $\nabla f(a) = 0$, then we call a a *critical point* of f.

Recall from linear algebra: A matrix $A \in \mathbb{R}^{n\times n}$ is called *positive definite* if $\langle x, Ax\rangle > 0$ for all $x \in \mathbb{R}^n\backslash\{0\}$ and *positive semidefinite* if $\langle x, Ax\rangle \geqslant 0$ for all $x \in \mathbb{R}^n$. We also write $A > 0$ to express that A is positive definite and $A \geqslant 0$ to express that A is positive semidefinite. The terms *negative definite, negative semidefinite* are defined accordingly. A is *indefinite* if it is not positive semidefinite and not negative semidefinite. Every real symmetric matrix has real eigenvalues and there is an orthonormal basis of eigenvectors (spectral theorem). A real symmetric matrix is positive definite if and only if all eigenvalues are positive.

6.75. Theorem. *Let $f \in C^2(\Omega)$ and $a \in \Omega$ with $\nabla f(a) = 0$.*

(i) If $\mathrm{H}f|_a$ is positive definite, then f has a strict local minimum at a.
(ii) If $\mathrm{H}f|_a$ is negative definite, then f has a strict local maximum at a.
(iii) If $\mathrm{H}f|_a$ is indefinite, then f does not have a local extremum at a.

Remark. If $\mathrm{H}f|_a$ is only positive semidefinite or negative semidefinite, then we need more information to be able to decide whether or not f has local extrema at a.

Proof of Theorem 6.75. We write $A = \mathrm{H}f|_a$. By Corollary 6.64 there exists, for every $\varepsilon > 0$, a $\delta(\varepsilon) > 0$ such that for all y with $\|y\| \leqslant \delta(\varepsilon)$,

$$f(a + y) = f(a) + \tfrac{1}{2}\langle y, Ay\rangle + r(y) \quad \text{with } |r(y)| \leqslant \varepsilon\|y\|^2. \tag{6.21}$$

Regarding (i), let A be positive definite. Let $S = \{y \in \mathbb{R}^n : \|y\| = 1\}$. S is compact, so the continuous map $y \mapsto \langle y, Ay\rangle$ attains its minimum on S. That is, there exists $y_0 \in S$ such that

$$\langle y_0, Ay_0\rangle \leqslant \langle y, Ay\rangle$$

for all $y \in S$. Define $\alpha = \langle y_0, Ay_0\rangle$. Since $y_0 \neq 0$ and A is positive definite, $\alpha > 0$. Let $y \in \mathbb{R}^n, y \neq 0$. Then $\frac{y}{\|y\|} \in S$, so

$$\alpha \leqslant \langle \frac{y}{\|y\|}, A\frac{y}{\|y\|}\rangle = \frac{1}{\|y\|^2}\langle y, Ay\rangle.$$

Thus, $\langle y, Ay\rangle \geqslant \alpha\|y\|^2$ for all $y \in \mathbb{R}^n$. Now we set $\varepsilon = \frac{\alpha}{4}$. Then for $\|y\| \leqslant \delta(\alpha/4)$, $y \neq 0$,

$$\begin{aligned} f(a + y) &\geqslant f(a) + \tfrac{1}{2}\langle y, Ay\rangle - |r(y)| \geqslant f(a) + \tfrac{1}{2}\langle y, Ay\rangle - \tfrac{\alpha}{4}\|y\|^2 \\ &\geqslant f(a) + \tfrac{\alpha}{2}\|y\|^2 - \tfrac{\alpha}{4}\|y\|^2 = f(a) + \tfrac{\alpha}{4}\|y\|^2 > f(a). \end{aligned}$$

Therefore a is a local minimum point.

(ii) follows from (i) by replacing f by $-f$.

For (iii) let A be indefinite. We need to show that in every open neighborhood of a there exist points y', y'' such that

$$f(y'') < f(a) < f(y').$$

Since A is not negative semidefinite there exists $\xi \in \mathbb{R}^n$, $\|\xi\| = 1$, such that $\alpha = \langle \xi, A\xi \rangle > 0$. If $\varepsilon = \alpha/4$ and $|t| \leqslant \delta(\alpha/4)$ we have from (6.21)

$$f(a + t\xi) = f(a) + \tfrac{1}{2}\langle t\xi, At\xi \rangle + r(t\xi) \geqslant f(a) + \tfrac{1}{2}\alpha t^2 - \tfrac{\alpha}{4}\|t\xi\|^2$$

and since $\|\xi\| = 1$ we get, for $|t| \leqslant \delta(\alpha/4)$.

$$f(a + t\xi) \geqslant f(a) + \tfrac{\alpha}{4}t^2 > f(a)$$

Similarly since A is also not positive semidefinite, there exists $\eta \in \mathbb{R}^n$, $\|\eta\| = 1$, $\beta > 0$ such that $\langle \eta, A\eta \rangle = -\beta < 0$. The same argument gives for $|t| \leqslant \delta(\beta/4)$

$$f(a + t\eta) = f(a) - \tfrac{1}{2}\beta t^2 + r(t\eta) \leqslant f(a) - \tfrac{\beta}{4}t^2 < f(a).$$

We conclude that f does not have a local extremum at a. □

6.76. Examples.

(i) Let $f(x, y) = c + x^2 + y^2$ for $c \in \mathbb{R}$. Then

$$\mathrm{H}f|_0 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

is positive definite and 0 is a strict local minimum point of f (actually f has a global minimum at 0).

(ii) Let $f(x, y) = c + x^2 - y^2$ for $c \in \mathbb{R}$. Then

$$\mathrm{H}f|_0 = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

is indefinite and 0 is not a local extremum point of f.

(iii) Let $f_1(x, y) = x^2 + y^4$, $f_2(x, y) = x^2$, $f_3(x, y) = x^2 + y^3$. Then

$$\mathrm{H}f_i|_0 = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

is positive semidefinite, but f_1 has a strict local minimum at 0, f_2 has a (non-strict) local minimum at 0 and f_3 has no local extremum at 0.

6.4.2. Local extrema with constraints. A standard problem is to find local extrema of a function f of n variables in the presence of additional constraints given by equations $g(x) = 0$, where $g(x) = (g_1(x), \ldots, g_m(x))$, $i = 1, \ldots, m$. We shall assume $m < n$ and that the gradients ∇g_i are linearly independent; we then seek to find a necessary condition for local extrema of f on the surface given by

$$g_i(x) = 0, \quad i = 1, \ldots, m.$$

We say that f has a local maximum at a with respect to the constraint $g = 0$ if there is an open neighborhood U of a such that $f(a) \geqslant f(x)$ for all $x \in U$ that satisfy $g(x) = 0$. The definition of a local minimum under the constraint $g = 0$ is analogous. The following theorem gives a necessary condition for f to have a local extremum at a under the constraint $g = 0$.

6.77. THEOREM. *Let $a \in \mathbb{R}^n$ and let U be an open neighborhood of a. Let $m < n$, let $g = (g_1, \dots, g_m) : U \to \mathbb{R}^m$ be in $C^1(U; \mathbb{R}^m)$, $f : U \to \mathbb{R}$ be in $C^1(U)$ and assume that $g(a) = 0$ and $\operatorname{rank}(Dg|_a) = m$. If f has a local extremum at a under the constraint $g = 0$, then the vectors $\nabla f(a)$, $\nabla g_1(a), \dots, \nabla g_m(a)$ are linearly dependent, and there are $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ such that*

$$\nabla f(a) = \sum_{k=1}^{m} \lambda_k \nabla g_k(a).$$

The numbers $\lambda_1, \dots, \lambda_m$ are called *Lagrange multipliers.*

PROOF OF THEOREM 6.77. Without loss of generality we may assume, after possibly relabeling the variables, that the $m \times m$ matrix with entries $\partial_i g_j(a)$ with $i, j = 1, \dots, m$ has maximal rank m. By the implicit function theorem the equation $g(x) = 0$ can be solved near a by expressing the variables x_i, $i = 1, \dots, m$ as functions of $x'' = (x_{m+1}, \dots, x_n)$. Thus there is a $\delta > 0$ such that for

$$B_\delta = \{x = (x', x'') \in \mathbb{R}^n : |x_i - a_i| < \delta, i = 1, \dots, n\}$$

the equation $g(x) = 0$ for $x \in B_\delta$ is satisfied if and only if $x' = u(x'')$ for $x'' = (x_{m+1}, \dots, x_n)$. Here u is of class C^1 near a'', $u(a'') = a'$ and

$$g(u(x''), x'') = 0 \tag{6.22}$$

for x'' close to a''. The invertibility of the $m \times m$ matrix $(\partial_i g_j(a))_{i,j=1,\dots,m}$ also implies that the system of m linear equations

$$\sum_{k=1}^{m} \lambda_k \partial_i g_k(a) = \partial_i f(a), \quad i = 1, \dots, m \tag{6.23}$$

has a unique solution $(\lambda_1, \dots, \lambda_m)$.

By assumption the function

$$F : x'' \mapsto f(u(x''), x'')$$

has a local extremum at a'' and therefore, by Corollary 6.73, we have $\partial_\nu F(a'') = 0$ for $\nu = m+1, \dots, n$. By the chain rule this gives

$$\partial_\nu f(a) + \sum_{i=1}^{m} \partial_i f(a) \partial_\nu u_i(a'') = 0, \quad \nu = m+1, \dots, n. \tag{6.24}$$

Differentiating (6.22) with respect to x_ν, we obtain

$$\partial_\nu g_k(a) + \sum_{i=1}^{m} \partial_i g_k(a) \partial_\nu u_i(a'') = 0 \tag{6.25}$$

for $\nu = m+1, \dots, n$ and $k = 1, \dots, m$. Combining (6.24) and (6.23) we obtain, for $\nu = m+1, \dots, n$

$$\begin{aligned} \partial_\nu f(a) = -\sum_{i=1}^{m} \partial_i f(a) \partial_\nu u_i(a'') &= -\sum_{i=1}^{m} \Big(\sum_{k=1}^{m} \lambda_k \partial_i g_k(a) \Big) \partial_\nu u_i(a'') \\ &= -\sum_{k=1}^{m} \lambda_k \Big(\sum_{i=1}^{m} \partial_i g_k(a) \partial_\nu u_i(a'') \Big). \end{aligned}$$

Now use (6.25) to obtain

$$\partial_\nu f(a) = \sum_{k=1}^{m} \lambda_k \partial_\nu g_k(a), \quad \nu = m+1, \dots, n. \tag{6.26}$$

Combining (6.26) with (6.23) we get $\nabla f(a) = \sum_{k=1}^{m} \lambda_k \nabla g_k(a)$ as claimed. □

6.78. Example. For $x \in \mathbb{R}^n$ let $g(x) = x^\intercal x = \sum_{j=1}^n x_j^2$ and let $S = \{x : g(x) = 1\}$, the unit sphere centered at the origin. Consider the quadratic form

$$f(x) = x^\intercal Ax = \sum_{i,j=1}^n a_{ij} x_i x_j$$

where A is a real symmetric $n \times n$ matrix, for $x \in S$. The continuous function f has a maximum on the compact set S, which occurs at v, say. Theorem 6.77 tells us that $\nabla f(v)$ and $\nabla g(v)$ are linearly dependent and there is a $\lambda \in \mathbb{R}$ such that $\nabla f(v) = \lambda \nabla g(v)$. Compute $\nabla f(v) = 2Av$ and $\nabla g(v) = 2v$. Thus we have $Av = \lambda v$ with $v \in S$. This means that v is an eigenvector of A, with corresponding eigenvalue λ.

Remark. The existence of an eigenvector in $\mathbb{R}^n$ for real symmetric matrices is an important first step in the proof of the spectral theorem for such matrices.

6.4.3. Global minima and convexity. In applications it is often desirable to find a *global minimum* of a given function $f : \Omega \to \mathbb{R}$, i.e. to find $x_* \in \Omega$ such that $f(x_*) \leqslant f(x)$ for all $x \in \Omega$. We call such a point x_* a *global minimum point* of f. We say that f has a *strict global minimum* at x_* if $f(x_*) < f(x)$ for all $x \neq x_*$.

6.79. Example (Linear regression). Say we are given finitely many points

$$(x_1, y_1), \ldots, (x_N, y_N) \in \mathbb{R}^n \times \mathbb{R}.$$

Suppose for instance that these represent measurements or observations of some physical system. For example, x_i could represent a point in space and y_i the corresponding air pressure measurement. We are looking to discover a "hidden relation" between the x and y coordinates. That is, we are looking for a function $F : \mathbb{R}^n \to \mathbb{R}$ such that $F(x_i)$ is (at least roughly) y_i. One way this is done is *linear regression.* Here we search only among F that take the form

$$F_{a,b}(x) = \langle x, a \rangle + b$$

with some parameters $a \in \mathbb{R}^n, b \in \mathbb{R}$. That is, we are trying to "model" the hidden relation by an affine linear function. The task is now to find the parameters a, b such that $F_{a,b}$ "fits best" to the given data set. To make this precise we introduce the error function

$$E(a, b) = \sum_{i=1}^N (F_{a,b}(x_i) - y_i)^2.$$

The problem of linear regression is to find the parameters (a, b) such that $E(a, b)$ is minimal.

One approach to minimizing a function $f : \Omega \to \mathbb{R}$ is to solve the equation $\nabla f(x) = 0$, i.e. to find all critical points. By Corollary 6.73 we know that every minimum must be a critical point. However it is often difficult to solve that equation, so more practical methods are needed.

Gradient descent. Choose $x_0 \in \mathbb{R}^n$ arbitrary and let

$$x_{n+1} = x_n - \alpha_n \nabla f(x_n)$$

where $\alpha_n > 0$ is a small enough number to be determined later. The idea of this iteration is to keep moving into the direction where f decreases the fastest. Sometimes this simple process successfully converges to a minimum and sometimes it does not, depending on f, x_0 and α_n. What we can say from the definition is that, if $f \in C^1(\Omega)$ and $(x_n)_{n \in \mathbb{N}}$ converges, then the limit is a critical point of f. The following lemma gives some more hope.

6.80. Lemma. *Let $f \in C^1(\Omega)$. Then, for every $x \in \Omega$ there exists $\delta(x) > 0$ such that for $0 \leqslant \alpha < \delta(x)$,*

$$f(x - \alpha \nabla f(x)) \leqslant f(x).$$

PROOF. If $\nabla f(x) = 0$ there is nothing to show, so we may assume $\nabla f(x) \neq 0$. By the definition of Fréchet derivative,

$$f(x - \alpha\nabla f(x)) = f(x) + \langle \nabla f(x), -\alpha\nabla f(x)\rangle + o(\alpha) = f(x) - \alpha\|\nabla f(x)\|^2 + o(\alpha)$$

which is $\leqslant f(x)$ provided that $\alpha > 0$ is small enough. □

REMARK. Note that the smallness of α in this lemma depends on the point x. Also, this result is not enough to prove anything about the convergence of gradient descent.

We will see that gradient descent works well if f is a convex function.

6.81. DEFINITION. Let $E \subset \mathbb{R}^n$ be a convex set. A function $f : E \to \mathbb{R}$ is called *convex* if

$$f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y)$$

for all $x, y \in E$, $t \in [0,1]$. f is called *strictly convex* if

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y)$$

for all $x \neq y \in E$ and $t \in (0,1)$.

6.82. THEOREM. *Let $\Omega \subset \mathbb{R}^n$ be open and convex and $f \in C^1(\Omega)$. Then f is convex if and only if*

$$f(u+v) \geqslant f(u) + \langle \nabla f(u), v\rangle$$

for all $u, u+v \in \Omega$.

PROOF. ($\Rightarrow$) We assume that f is convex on Ω. Fix $u, u+v \in \Omega$. By convexity, for $t \in [0,1]$,

$$f(u+tv) = f((1-t)u + t(u+v)) \leqslant (1-t)f(u) + tf(u+v).$$

By definition of the derivative,

$$f(u+tv) = f(u) + t\langle \nabla f(u), v\rangle + r(t),$$

where $\lim_{t\to 0} \frac{r(t)}{t} = 0$. Thus,

$$f(u) + t\langle \nabla f(u), v\rangle + r(t) \leqslant (1-t)f(u) + tf(u+v)$$

which implies

$$f(u) + \langle \nabla f(u), v\rangle - f(u+v) \leqslant \frac{-r(t)}{t} \to 0 \quad \text{as } t \to 0.$$

Since the left hand side does not depend on t we get $f(u) + \langle \nabla f(u), v\rangle \leqslant f(u+v)$.

($\Leftarrow$) We assume $f(u+v) \geqslant f(u) + \langle \nabla f(u), v\rangle$ for all $u, u+v \in \Omega$. Let $x, y \in \Omega$, $t \in [0,1]$. Let $u = tx + (1-t)y$ and $v = x - u$. Then the assumption implies

$$f(x) \geqslant f(u) + \langle \nabla f(u), x - u\rangle.$$

On the other hand, letting $v = y - u$, the assumption implies

$$f(y) \geqslant f(u) + \langle \nabla f(u), y - u\rangle.$$

Therefore

$$\begin{aligned} tf(x) + (1-t)f(y) &\geqslant t(f(u) + \langle \nabla f(u), x-u\rangle) + (1-t)(f(u) + \langle \nabla f(u), y-u\rangle) \\ &= f(u) + \langle \nabla f(u), t(x-u) + (1-t)(y-u)\rangle = f(u) + \langle \nabla f(u), tx + (1-t)y - u\rangle. \end{aligned}$$

Recalling that $u = tx + (1-t)y$, we get

$$tf(x) + (1-t)f(y) \geqslant f(u) = f(tx + (1-t)y).$$

□

6.83. THEOREM. *Let $\Omega \subset \mathbb{R}^n$ be open and convex and $f \in C^2(\Omega)$. Then*

(i) f is convex if and only if $\mathrm{H}f|_x$ is positive semidefinite for all $x \in \Omega$,

(ii) f is strictly convex if $\mathrm{H}f|_x$ is positive definite for all $x \in \Omega$.

PROOF. Consider (i). Let f be convex. By Taylor's theorem, for $u, u+tv \in \Omega$,

$$f(u+tv) = f(u) + t\langle \nabla f(u), v\rangle + \tfrac{1}{2}t^2\langle \mathrm{H}f|_u v, v\rangle + o(t^2)$$

and by Theorem 6.82,

$$f(u+tv) \geqslant f(u) + t\langle \nabla f(u), v\rangle.$$

Combining these two pieces of information we obtain

$$\tfrac{1}{2}t^2\langle \mathrm{H}f|_u v, v\rangle + o(t^2) \geqslant 0.$$

This implies $\langle \mathrm{H}f|_u v, v\rangle \geqslant 0$ for all $v \in \mathbb{R}^n$, i.e. $\mathrm{H}f|_u$ is positive semidefinite for all $u \in \Omega$.

Conversely, assume that $\mathrm{H}f|_u$ is positive semidefinite for all $u \in \Omega$. By Taylor's theorem, for all $u, u+v \in \Omega$, there exists $\xi \in \Omega$ such that

$$f(u+v) = f(u) + \langle \nabla f(u), v\rangle + \tfrac{1}{2}\langle \mathrm{H}f|_\xi v, v\rangle \geqslant f(u) + \langle \nabla f(u), v\rangle.$$

Therefore f is convex by Theorem 6.82.

Regarding (ii) if $\mathrm{H}f|_\xi$ is positive definite for all $\xi \in \Omega$ we have in the last argument $\langle \mathrm{H}f|_\xi v, v\rangle > 0$ for all $v \in \mathbb{R}^n$ and we get $f(u+v) > f(u) + \langle \nabla f(u), v\rangle$ for all $v \neq 0$. Thus f is strictly convex. □

REMARK. For strictly convex f it is not necessarily true that $\mathrm{H}f|_x$ is positive definite for all x. Consider as an example $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^4$. Then $\mathrm{H}f|_x = f''(x) = 12x^2$ which is 0 at $x = 0$, but f is strictly convex.

6.84. THEOREM. *Let $\Omega \subset \mathbb{R}^n$ be open and convex and $f \in C^2(\Omega)$. Then*

(i) If f is convex, then every critical point of f is a global minimum point.
(ii) If f is strictly convex, then f has at most one critical point.

REMARKS. 1. Convex functions may have more than one critical point. For instance, the constant function $f \equiv 0$ is convex.

2. Conclusion (i) implies that if f is convex and gradient descent converges, then it converges to a global minimum.

PROOF OF THEOREM 6.84. (i): Let $\nabla f(x_*) = 0$. If f is convex, then $\mathrm{H}f$ is positive semidefinite on Ω, by Theorem 6.83. Then by Taylor's theorem, for every $x \in \Omega$ there exists $\xi \in \Omega$ such that

$$f(x) = f(x_*) + \underbrace{\langle \nabla f(x_*), x - x_*\rangle}_{=0} + \tfrac{1}{2}\underbrace{\langle \mathrm{H}f|_\xi (x - x_*), x - x_*\rangle}_{\geqslant 0} \geqslant f(x_*).$$

(ii): Let f be strictly convex. Let $x_1, x_2 \in \Omega$ be critical points of f. By (i), they are global minima. This implies $f(x_1) = f(x_2)$. If $x_1 \neq x_2$, then by strict convexity,

$$f(x_1) = \frac{f(x_1) + f(x_2)}{2} > f\Big(\frac{x_1 + x_2}{2}\Big).$$

This is a contradiction to x_1 being a global minimum. Therefore we must have $x_1 = x_2$. □

6.85. EXAMPLE. If $\|\cdot\|$ is a norm on $\mathbb{R}^n$, then the function $x \mapsto \|x\|$ is convex:

$$\|tx + (1-t)y\| \leqslant t\|x\| + (1-t)\|y\|$$

by the triangle inequality. Also, this function has a unique global minimum at $x = 0$.

6.86. LEMMA. *Let $I \subset \mathbb{R}$, $E \subset \mathbb{R}^n$ be convex and suppose that*

(i) $f : E \to I$ is convex, and
(ii) $g : I \to \mathbb{R}$ is convex and nondecreasing.

Then the function $h : E \to \mathbb{R}$ given by $h = g \circ f$ is convex.

PROOF. By convexity of f, $f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y)$. Since g is nondecreasing

$$g(f(tx + (1-t)y)) \leqslant g(tf(x) + (1-t)f(y)).$$

Since g is convex,

$$g(tf(x) + (1-t)f(y)) \leqslant tg(f(x)) + (1-t)g(f(y)).$$

Combine the two displayed inequalities to get

$$h(tx + (1-t)y) \leqslant th(x) + (1-t)h(y).$$

□

6.87. COROLLARY. *If $\|\cdot\|$ is a norm on $\mathbb{R}^n$, then the function $x \mapsto \|x\|^2$ is convex.*

6.88. EXAMPLE. Recall the error function from linear regression (Example 6.79):

$$E(a,b) = \sum_{i=1}^{N} (\langle a, x_i \rangle + b - y_i)^2$$

We claim that $E : \mathbb{R}^{n+1} \to \mathbb{R}$ is a convex function. We first rewrite $E(a,b)$ into a different form. Define an $N \times (n+1)$ matrix M and a vector $v \in \mathbb{R}^{n+1}$ by

$$M = \begin{pmatrix} x_{11} & \cdots & x_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{N1} & \cdots & x_{Nn} & 1 \end{pmatrix} \in \mathbb{R}^{N \times (n+1)}, \quad v = \begin{pmatrix} a_1 \\ \vdots \\ a_n \\ b \end{pmatrix} \in \mathbb{R}^{n+1},$$

where $x_i = (x_{i1}, \ldots, x_{in}) \in \mathbb{R}^n$ for $i = 1, \ldots, N$ and $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. Then

$$E(a,b) = E(v) = \sum_{i=1}^{N} ((Mv)_i - y_i)^2 = \|Mv - y\|^2,$$

where we work with the Euclidean norm $\|c\| = \left(\sum_{i=1}^{N} |c_i|^2 \right)^{1/2}$.

Let us rename variables and consider

$$E(x) = \|Mx - y\|^2$$

for $x \in \mathbb{R}^n$, $M \in \mathbb{R}^{N \times n}$, $y \in \mathbb{R}^N$. Let $F : \mathbb{R}^N \to \mathbb{R}$ be defined by $F(y) = \|y\|^2$ and $G : \mathbb{R}^n \to \mathbb{R}^N$ by $G(x) = Mx - y$. We have

$$\partial_i F(y) = 2y_i, \text{ so } DF|_y = 2y^\intercal \in \mathbb{R}^{1 \times N}.$$

and $DG|_x = M \in \mathbb{R}^{N \times n}$. Therefore, by the chain rule we obtain

$$DE|_x = 2(Mx - y)^\intercal M = 2(Mx)^\intercal M - 2y^\intercal M = 2x^\intercal M^\intercal M - 2y^\intercal M \in \mathbb{R}^{1 \times n}.$$

Therefore,

$$D^2 E|_x = (\partial_i DE|_x)_{i=1,\ldots,n} = (2(M^\intercal M)_i)_{i=1,\ldots,n} = 2M^\intercal M.$$

Notice that $M^\intercal M$ is positive semidefinite because

$$\langle M^\intercal M x, x \rangle = \langle Mx, Mx \rangle = \|Mx\|^2 \geqslant 0.$$

Therefore E is convex by Theorem 6.83.

6.89. EXAMPLES. Convex functions do not necessarily have a critical point. For instance the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x$ is convex, because $\mathrm{H}f|_x = f''(x) = 0$ for all $x \in \mathbb{R}$. But $\nabla f(x) = f'(x) = 1 \neq 0$ for all $x \in \mathbb{R}$.

It is also not enough to assume strict convexity. For instance, the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$ is strictly convex, because $f''(x) = e^x > 0$. But $f'(x) = e^x > 0$ for all $x \in \mathbb{R}$.

The examples motivate us to consider a stronger notion of convexity.

6.90. Definition. Let $\Omega \subset \mathbb{R}^n$ be convex and open. Let $f \in C^2(\Omega)$. We say that f is *strongly convex* if there exists $\beta > 0$ such that

$$\langle \mathrm{H}f|_x y, y\rangle \geqslant \beta \|y\|^2$$

for all $x \in \Omega$, $y \in \mathbb{R}^n$.

Remarks.

1. f is strongly convex if and only if there exists $\beta > 0$ such that $\mathrm{H}f|_x - \beta I$ is positive semidefinite for all $x \in \Omega$. This follows directly from the definition using that $\beta\|y\|^2 = \langle \beta I y, y\rangle$. The condition $\mathrm{H}f|_x - \beta I$ is positive semidefinite is equivalent to the smallest eigenvalue of $\mathrm{H}f|_x$ being $\geqslant \beta$. Yet another equivalent way of stating this is saying that the function $g(x) = f(x) - \frac{\beta}{2}\|x\|^2$ is convex. This is because $\mathrm{H}g|_x = \mathrm{H}f|_x - \beta I$.
2. If f is strongly convex, then f is strictly convex (by Theorem 6.83).
3. If f is strictly convex, then f is not necessarily strongly convex. For example consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$. For every $\beta > 0$ there exists $x \in \mathbb{R}$ such that $e^x < \beta$ because $e^x \to 0$ as $x \to -\infty$.

The following exercise shows that the assumption of strong convexity is not as restrictive as it may seem at first sight: strictly convex functions are strongly convex when restricted to compact sets.

6.91. Exercise. Suppose that $f \in C^2(\mathbb{R}^n)$ is strictly convex. Let $K \subset \mathbb{R}^n$ be compact and convex. Show that there exist $\beta_-, \beta_+ > 0$ such that

$$\beta_-\|y\|^2 \leqslant \langle \mathrm{H}f|_x y, y\rangle \leqslant \beta_+\|y\|^2$$

for all $x \in K$ and $y \in \mathbb{R}^n$. (In particular, f is strongly convex on K.)

6.92. Theorem. *Let $\Omega \subset \mathbb{R}^n$ be open and convex. Let $f \in C^2(\Omega)$. Then f is strongly convex if and only if there exists $\gamma > 0$ such that*

$$f(u+v) \geqslant f(u) + \langle \nabla f(u), v\rangle + \gamma\|v\|^2$$

whenever $u, u+v \in \Omega$.

Proof. ($\Rightarrow$) Let f be strongly convex on Ω and $\beta > 0$ be such that $g(x) = f(x) - \frac{\beta}{2}\|x\|^2$ is convex. Then, by Theorem 6.82,

$$g(u+v) \geqslant g(u) + \langle \nabla g(u), v\rangle = f(u) - \tfrac{\beta}{2}\|u\|^2 + \langle \nabla f(u) - \beta u, v\rangle$$

On the other hand,

$$g(u+v) = f(u+v) - \tfrac{\beta}{2}\|u+v\|^2.$$

Thus,

$$f(u+v) \geqslant f(u) + \langle \nabla f(u), v\rangle + \tfrac{\beta}{2}(\|u+v\|^2 - \|u\|^2 - 2\langle u, v\rangle) = f(u) + \langle \nabla f(u), v\rangle + \tfrac{\beta}{2}\|v\|^2.$$

The converse ($\Leftarrow$) follows in the same way from the converse direction of Theorem 6.82. □

6.93. Theorem. *Let $f \in C^2(\mathbb{R}^n)$ be strongly convex. Then for every $c \in \mathbb{R}$, the* sublevel set

$$S = \{x \in \mathbb{R}^n \,:\, f(x) \leqslant c\}$$

is bounded.

Proof. By Theorem 6.92 we have, for a suitable $\gamma > 0$,

$$f(x) \geqslant f(0) + \langle \nabla f(0), x\rangle + \gamma\|x\|^2.$$

Therefore, $\lim_{\|x\|\to\infty} f(x) = \infty$. Suppose that S is unbounded. Then there exists a sequence $(x_n)_{n\geqslant 1} \subset S$ such that $\lim_{n\to\infty} \|x_n\| = \infty$. But $f(x_n) \leqslant c$, which yields a contradiction. □

6.94. Theorem. *Let $f \in C^2(\mathbb{R}^n)$ be strongly convex. Then there exists a unique global minimum of f.*

PROOF. By the previous theorem, the set $S = \{x \in \mathbb{R}^n : f(x) \leqslant f(0)\}$ is bounded. Thus, there exists $R > 0$ such that $S \subset B_R = \{x \in \mathbb{R}^n : \|x\| \leqslant R\}$. B_R is compact, so f attains its minimum on B_R at some point $x_* \in B_R$. Then $f(x_*) \leqslant f(x)$ for all $x \in B_R$. It remains to show $f(x_*) \leqslant f(x)$ for all $x \notin B_R$. If $x \notin B_R$, then $x \notin S$, so $f(x) > f(0)$. Also, $0 \in B_R$, so $f(x_*) \leqslant f(0) < f(x)$. □

We now prove that gradient descent converges for (strongly) convex functions.

6.95. THEOREM. *Let $f \in C^2(\mathbb{R}^n)$ be strongly convex and $x_0 \in \mathbb{R}^n$. There exists $\delta > 0$ such that for $0 < \alpha < \delta$ the sequence (x_n) defined by*

$$x_{n+1} = x_n - \alpha \nabla f(x_n) \quad \text{for } n \geqslant 0,$$

converges to the global minimum x_* of f.

REMARK. We consider f defined on $\mathbb{R}^n$ only for convenience (the same is true for Theorems 6.93 and 6.94). Analogous results hold for functions on restricted domains.

PROOF OF THEOREM 6.95. Let $\alpha > 0$. Define

$$T(x) = x - \alpha \nabla f(x).$$

Then $x_{n+1} = T(x_n)$. We want T to be a contraction. For $R > 0$ define $B_R = \{x \in \mathbb{R}^n : \|x - x_*\| \leqslant R\}$ (where $\|\cdot\|$ stands for the standard Euclidean norm in $\mathbb{R}^n$). Let $R > 0$ be large enough such that $x_0 \in B_R$. By the contraction principle it suffices to prove the following CLAIM. If α is small enough, then T is a contraction of B_R.

Proof of claim. x_* is a global minimum point of f, so $\nabla f(x_*) = 0$. Thus, $T(x_*) = x_*$. We have

$$DT|_x = I - \alpha \mathrm{H}f|_x.$$

The operator norm of $\mathrm{H}f|_x$ is a continuous function of x which is bounded on the compact set B_R. Therefore there exists $\gamma > 0$ such that

$$\langle \mathrm{H}f|_x y, y \rangle \leqslant \gamma \|y\|^2$$

for all $y \in \mathbb{R}^n$ and $x \in B_R$. By strong convexity, there is $\beta > 0$ such that

$$\beta \|y\|^2 \leqslant \langle \mathrm{H}f|_x y, y \rangle \leqslant \gamma \|y\|^2.$$

This implies that the eigenvalues of $\mathrm{H}f|_x$ are contained in the interval $[\beta, \gamma]$, for all $x \in B_R$. Let $\alpha \leqslant \frac{1}{2\gamma}$. Then the eigenvalues of $I - \alpha \mathrm{H}f|_x$ are contained in

$$[1 - \tfrac{\gamma}{2\gamma}, 1 - \tfrac{\beta}{2\gamma}] = [\tfrac{1}{2}, 1 - \tfrac{\beta}{2\gamma}] \subset (0, 1).$$

Set $c = 1 - \frac{\beta}{2\gamma}$. By Example 5.18,

$$\|I - \alpha \mathrm{H}f|_x\|_{\mathrm{op}} \leqslant c < 1.$$

Therefore, $\|T(x) - T(y)\| \leqslant c\|x - y\|$ for all $x, y \in B_R$. It remains to show that $T(B_R) \subset B_R$. Let $x \in B_R$. Then since $T(x_*) = x_*$,

$$\|T(x) - x_*\| = \|T(x) - T(x_*)\| \leqslant c\|x - x_*\| \leqslant cR \leqslant R.$$

We have shown that T is a contraction on B_R and thus (x_n) converges to the unique fixed point of T which is x_*. □

6.5. Ordinary differential equations

In this section we study *initial value problems* of the form

$$\begin{cases} y'(t) = F(t, y(t)) \\ y(t_0) = y_0, \end{cases}$$

where $\Omega \subset \mathbb{R} \times \mathbb{R}$ is open, $(t_0, y_0) \in \Omega$ and $F \in C(\Omega)$ are given. Let I be an open interval that includes the point t_0 and let $y : I \to \mathbb{R}$ be a differentiable function. We say that y is a *solution* to the initial value problem if $(t, y(t)) \in \Omega$ for all $t \in I$, $y(t_0) = y_0$, and $y'(t) = F(t, y(t))$ for all $t \in I$. The equation $y'(t) = F(t, y(t))$ is a *first order ordinary differential equation.* We also write this differential equation in short form as

$$y' = F(t, y).$$

Geometric interpretation. At each point $(t, y) \in \Omega$ imagine a small line segment with slope $F(t, y)$. We are looking for a function such that its graph has the slope $F(t, y)$ at each point (t, y) on the graph of the function.

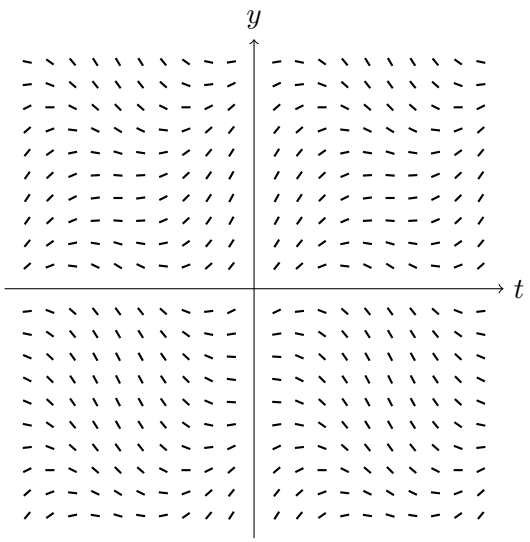


Figure 6.1. Visualization of $F(t, y)$.

6.96. Example. Consider the equation $y'(t) = t^{-1}y(t)$, for $t > 0$. The solutions of this equation are of the form $y(t) = ct$ for $c \in \mathbb{R}$.

6.97. Example. Sometimes we can solve initial value problems by computing an explicit expression for y. Consider initial value problems of the form

$$\begin{cases} y'(t) = f(t)g(y(t)), \\ y(t_0) = y_0 \end{cases} \tag{6.27}$$

where f is continuous near t_0 and g is continuous near y_0. This problem can be solved by the method of *separation of variables,* for t near t_0, provided that $g(y_0) \neq 0$.

Indeed, if A is such that $A'(y) = 1/g(y)$ for y near y_0 a solution $y(t)$ satisfies for t near t_0

$$A(y(t)) - A(y(t_0)) = \int_{t_0}^{t} \frac{y'(s)}{g(y(s))}\, ds = \int_{t_0}^{t} f(s)\, ds.$$

Since A is invertible near y_0 the solution is given by

$$y(t) = A^{-1}\Big(A(y_0) + \int_{t_0}^{t} f(s)\,ds\Big).$$

Consider for instance

$$\begin{cases} y'(t) = \frac{t}{y(t)} \\ y(t_0) = y_0 \end{cases}$$

for $t_0 > 0$, $y_0 > 0$. Then $y(t) = \sqrt{t^2 + y_0^2 - t_0^2}$. Note that if $y_0^2 - t_0^2 \geqslant 0$, then y is defined on $I = (0, \infty)$. But if $y_0^2 - t_0^2 < 0$, then y is only defined on the interval $I = (\sqrt{t_0^2 - y_0^2}, \infty)$ (in both cases the interval I contains the point t_0).

In general, however it is not easy to find a solution. It may also happen that the solution is not expressible in terms of elementary functions. Try for instance, to solve the initial value problem $y'(t) = e^{y(t)^2 t^2} \sin(t + y(t))$ with $y(1) = 5$ near $t = 1$.

6.5.1. The existence and uniqueness theorem. The main existence and uniqueness theorem shows that there is a unique solution without explicitly computing it.

6.98. Theorem (Picard–Lindelöf). *Let $\Omega \subset \mathbb{R} \times \mathbb{R}$ be open, $(t_0, y_0) \in \Omega$, $F \in C(\Omega)$. Let $a > 0$ and $b > 0$ be small enough such that*

$$R = \{(t, y) \in \mathbb{R}^2 : |t - t_0| \leqslant a, |y - y_0| \leqslant b\} \subset \Omega.$$

Let $M = \sup_{(t,y)\in R} |F(t, y)| < \infty$. Assume that there exists $\lambda > 0$ such that

$$|F(t, y) - F(t, u)| \leqslant \lambda |y - u| \tag{6.28}$$

for all $(t, y), (t, u) \in R$. Define $a_* = \min(a, b/M)$ and let $I = [t_0 - a_*, t_0 + a_*]$. Then the initial value problem

$$\begin{cases} y'(t) = F(t, y(t)), \\ y(t_0) = y_0. \end{cases} \tag{6.29}$$

has a unique solution $y : I \to \mathbb{R}$ in $C^1(I)$.

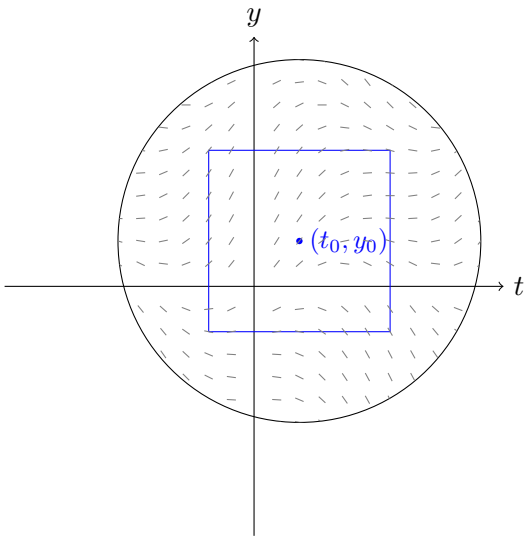


Figure 6.2. Visualization of $F(t, y)$.

By assuming additional differentiability assumptions on F we can also show higher regularity of the solution:

6.99. Corollary. *Let $\Omega \subset \mathbb{R}\times\mathbb{R}$ be open, $(t_0, y_0) \in \Omega$, $F \in C^k(\Omega)$. Then there exists an open interval $I \subset \mathbb{R}$ with $t_0 \in I$ and a unique differentiable function $y : I \to \mathbb{R}$ such that $(t, y(t)) \in \Omega$ for all $t \in I$ and y solves* (6.29). *Moreover, $y \in C^{k+1}(I)$.*

Proof. We first note that the continuity of $\partial_y F$ and the mean value theorem imply condition (6.28). If $F \in C^1(\Omega)$ and y is continuous we also see that $t \mapsto F(t, y(t))$ is continuous and thus y' is continuous, hence $y \in C^1$. By induction, using Proposition 6.60, we also see that y' is of class C^k if $F \in C^k(\Omega)$ hence $y \in C^{k+1}$. □

Remarks. 1. If F satisfies condition (6.28), we also say that F is *Lipschitz continuous in the second variable.* Note that the solution interval I guaranteed by the theorem is independent of the *Lipschitz constant* λ.

2. The condition (6.28) follows if F is differentiable in the second variable and

$$|\partial_y F(t, y)| \leqslant \lambda$$

for every $(t, y) \in R$ (by the mean value theorem).

It will be advantageous to reformulate the initial value problem (6.29) as an equivalent integral equation

$$y(t) = y_0 + \int_{t_0}^{t} F(s, y(s))\, ds. \tag{6.30}$$

More precisely, we have the following by the fundamental theorem of calculus (provide the details!)

6.100. Lemma.

(i) If $y \in C^1(I)$ solves (6.29) *on I then y solves the integral equation* (6.30).

(ii) Vice versa, if y is a continuous function solving (6.30) *on I then y is C^1 on I and solves* (6.29).

Proof of Theorem 6.98. Let $J = [y_0 - b, y_0 + b]$. It suffices to show that there exists a unique continuous function $y : I \to J$ such that

$$y(t) = y_0 + \int_{t_0}^{t} F(s, y(s))\, ds$$

(that is, y is a solution of the integral equation). Let

$$\mathcal{Y} = \{y : I \to J \,:\, y \text{ continuous on } I, \}$$

Every $y \in \mathcal{Y}$ is defined on $I = [t_0 - a_*, t_0 + a_*]$ and takes values in $J = [y_0 - b, y_0 + b]$. Thus, for $y \in \mathcal{Y}$, $t \mapsto F(t, y(t))$ is a well-defined continuous function on I. Define

$$Ty(t) = y_0 + \int_{t_0}^{t} F(s, y(s))\, ds.$$

We first check that $T : \mathcal{Y} \to \mathcal{Y}$. Let $y \in \mathcal{Y}$. Then Ty is a continuous function on I. It remains to show that $Ty(t) \in J$ for all $t \in I$. Recalling that $|F(t, y)| \leqslant M$ for all $(t, y) \in R$ we obtain

$$|Ty(t) - y_0| \leqslant \int_{t_0}^{t} |F(s, y(s))|\, ds \leqslant |t_0 - t|M \leqslant Ma_* \leqslant b,$$

where we used that $a_* = \min\{a, b/M\} \leqslant b/M$.

It remains to prove that T is a contraction with respect to a suitable metric on $\mathcal{Y}$, and that $\mathcal{Y}$ with this metric is complete. It is natural and tempting to try the usual supremum metric $d_\infty(g_1, g_2) = \sup_{t\in I} |g_1(t) - g_2(t)|$. Then $\mathcal{Y} \subset C(I)$ is closed, so $(\mathcal{Y}, d_\infty)$ is a complete metric space. However, T will not necessarily be a contraction on $\mathcal{Y}$ (in fact in order to establish a contraction property in this metric we would have to work with a smaller interval, depending on the Lipschitz constant).

Instead, we define the metric

$$d_*(g_1, g_2) = \sup_{t\in I} e^{-2\lambda|t-t_0|}|g_1(t) - g_2(t)|.$$

Then $d_*(g_1, g_2) \leqslant d_\infty(g_1, g_2) \leqslant e^{2\lambda a_*} d_*(g_1, g_2)$. In other words, d_* and d_∞ are equivalent metrics, in the sense that a Cauchy sequence with respect to d_∞ is also a Cauchy sequence with respect to d_*, and vice versa. We see that $(\mathcal{Y}, d_*)$ is still a complete metric space.

We now prove that $T : \mathcal{Y} \to \mathcal{Y}$ is a contraction with respect to d_*. Specifically we will show that

$$d_*(Tg_1, Tg_2) \leqslant \tfrac{1}{2} d_*(g_1, g_2). \tag{6.31}$$

Let us assume that $t \in [t_0, t_0 + a_*]$. For $g_1, g_2 \in \mathcal{Y}$, $t \in I$ we have by the Lipschitz condition (6.28),

$$\begin{aligned} |Tg_1(t) - Tg_2(t)| &= \Big| \int_{t_0}^{t} (F(s, g_1(s)) - F(s, g_2(s)))\, ds \Big| \\ &\leqslant \lambda \int_{t_0}^{t} |g_1(s) - g_2(s)|\, ds. \end{aligned}$$

Then

$$\begin{aligned} |Tg_1(t) - Tg_2(t)| &\leqslant \lambda \int_{t_0}^{t} |g_1(s) - g_2(s)|\, ds \leqslant \lambda \int_{t_0}^{t} e^{2\lambda(s-t_0)} d_*(g_1, g_2)\, ds \\ &= \lambda d_*(g_1, g_2) \tfrac{1}{2\lambda}(e^{2\lambda(t-t_0)} - 1) \leqslant \tfrac{1}{2} d_*(g_1, g_2) e^{2\lambda|t-t_0|}. \end{aligned}$$

We argue similarly for $t \in [t_0 - a_*, t_0]$ and obtain in this case

$$\begin{aligned} |Tg_1(t) - Tg_2(t)| &\leqslant \lambda \int_{t}^{t_0} |g_1(s) - g_2(s)|\, ds \leqslant \lambda \int_{t}^{t_0} e^{2\lambda(t_0-s)} d_*(g_1, g_2)\, ds \\ &= \lambda d_*(g_1, g_2) \tfrac{1}{2\lambda}(e^{2\lambda(t_0-t)} - 1) \leqslant \tfrac{1}{2} d_*(g_1, g_2) e^{2\lambda|t-t_0|}. \end{aligned}$$

Thus for all $t \in I = [t_0 - a_*, t_0 + a_*]$

$$e^{-2\lambda|t-t_0|}|Tg_1(t) - Tg_2(t)| \leqslant \tfrac{1}{2} d_*(g_1, g_2)$$

Taking the sup over $t \in I$ on the left hand side yields the contraction property (6.31).

By the Banach fixed point theorem, there exists a unique $y \in \mathcal{Y}$ such that $Ty = y$, i.e. by Lemma 6.100 a unique solution to our initial value problem (6.29). □

Remark. The proof is constructive. The proof of the Banach fixed point theorem tells us that we can compute the solution by the *method of successive approximation*. We form the sequence $(Y_n)_{n=0}^{\infty}$ with $Y_{n+1} = TY_n$ which will converge to the unique fixed point y which is the solution of the integral equation (and thus the solution of the initial value problem). Starting with the initial value y_0 at time $t = t_0$ we get the sequence $(Y_n)_{n\geqslant 0} \subset \mathcal{Y}$ defined by

$$\begin{aligned} Y_0(t) &= y_0 \\ Y_n(t) &= y_0 + \int_{t_0}^{t} F(s, Y_{n-1}(s))\, ds \quad \text{for } n = 1, 2, \dots \end{aligned} \tag{6.32}$$

Then $(Y_n)_{n\geqslant 0}$ converges with respect to d_* and therefore uniformly on I to the solution y. The method of successive approximation for our initial value problem is also called *Picard iteration*. See Exercise 6.152 below for more information about the speed of convergence of the Picard iteration.

6.101. Example. Consider the initial value problem

$$\begin{cases} y'(t) = \frac{e^t \sin(t+y(t))}{ty(t)-1}, \\ y(1) = 5. \end{cases} \tag{6.33}$$

Let $F(t,y) = \frac{e^t \sin(t+y)}{ty-1}$. We need to choose a rectangle R around the point $(1,5)$ where we have control over $|F(t,y)|$ and $|\partial_y F(t,y)|$. Thus we need to stay away from the set of (t,y) such that $ty-1=0$. Say,

$$R = \{(t,y) \,:\, |t-1| \leqslant \tfrac{1}{2},\, |y-5| \leqslant 1\}.$$

Then for $(t,y) \in R$:

$$ty - 1 \geqslant (1-\tfrac{1}{2})(5-1) - 1 = 1.$$

Also, $|e^t \sin(t+y)| \leqslant e^{3/2}$. Setting $M = e^{3/2}$, we obtain

$$|F(t,y)| \leqslant M \quad \text{for all } (t,y) \in R.$$

Compute

$$\partial_y F(t,y) = \frac{e^t \cos(t+y)}{ty-1} - t\frac{e^t \sin(t+y)}{(ty-1)^2}.$$

For $(t,y) \in R$ we estimate

$$|\partial_y F(t,y)| \leqslant \Big|\frac{e^t \cos(t+y)}{ty-1}\Big| + \Big|t\frac{e^t \sin(t+y)}{(ty-1)^2}\Big| \leqslant e^{3/2} + \tfrac{3}{2}e^{3/2} =: \lambda.$$

Then the number a_* from Theorem 6.98 is $a_* = \min(a, b/M) = \min(\frac{1}{2}, 1/e^{3/2}) = e^{-3/2}$. So the theorem yields the existence and uniqueness of a solution of the initial value problem (6.33) in the interval $I = [1-e^{-3/2}, 1+e^{-3/2}]$. We can also give the Picard iteration: 'Let $y_0(t) = 5$ and

$$y_n(t) = 5 + \int_1^t \frac{e^s \sin(s + y_{n-1}(s))}{s y_{n-1}(s) - 1}\, ds.$$

The sequence $(y_n)_{n\in\mathbb{N}}$ converges uniformly on I to the solution y.

6.102. Example. Sometimes one can extend solutions beyond the interval obtained from the Picard–Lindelöf theorem. Consider the initial value problem

$$\begin{cases} y'(t) = \cos(y(t)^2 - 2t^3) \\ y(0) = 1 \end{cases} \tag{6.34}$$

We claim that there exists a unique solution $y : \mathbb{R} \to \mathbb{R}$. To prove this it suffices to demonstrate the existence of a unique solution on the interval $[-L, L]$ for every $L > 0$. To do this we invoke the Picard–Lindelöf theorem. Set

$$R = \{(t,y) \in \mathbb{R}^2 \,:\, |t| \leqslant L, |y-1| \leqslant L\}.$$

Let $F(t,y) = \cos(y^2 - 2t^3)$. Then

$$|F(t,y)| \leqslant 1 \quad \text{for all } (t,y) \in \mathbb{R}^2.$$

We have $\partial_y F(t,y) = -2y\sin(y^2-2t^3)$, so $|\partial_y F(t,y)| \leqslant 2|y| \leqslant 2(L+1)$ for all $(t,y) \in R$. Then by Theorem 6.98, there exists a unique solution y_L to (6.34) on $I_L = [-L, L]$. Notice that if $L_2 \geqslant L_1$, then the solutions y_{L_1} and y_{L_2} must agree on the smaller interval I_{L_1}, by the uniqueness assertion in the theorem. This shows that there is a unique solution on $(-\infty, \infty)$.

6.103. Exercise. Determine an interval $I = [1-a_*, 1+a_*]$ on which the initial value problem

$$\begin{cases} y'(t) = e^{y(t)^2 t^2} \sin(t + y(t)), \\ y(1) = 5. \end{cases}$$

has a unique solution.

6.104. Example. If the Lipschitz condition (6.28) fails, then the initial value problem may have more than one solution. Consider

$$\begin{cases} y'(t) = |y(t)|^{1/2}, \\ y(0) = 0. \end{cases} \tag{6.35}$$

The function $y \mapsto |y|^{1/2}$ is not Lipschitz continuous in any neighborhood of 0: for $y > 0$ its derivative $\frac{1}{2}y^{-1/2}$ is unbounded as $y \to 0$. The function $y_1(t) = 0$ solves the initial value problem (6.35). The functions

$$y_2(t) = \begin{cases} t^2/4, & \text{if } t > 0 \\ 0, & \text{if } t \leqslant 0 \end{cases}, \quad y_3(t) = \begin{cases} 0, & \text{if } t > 0 \\ -t^2/4, & \text{if } t \leqslant 0 \end{cases}, \quad y_4(t) = \begin{cases} t^2/4, & \text{if } t > 0 \\ -t^2/4, & \text{if } t \leqslant 0 \end{cases}$$

also do. Note that the functions are of class C^1, but y_2, y_3, y_4 are not of class C^2 at the origin.

6.5.2. The Peano existence theorem. The existence of a solution still holds without the assumption (6.28). We will prove this as a consequence of the Arzelá-Ascoli theorem. Example 6.104 shows the solution may not be unique.

6.105. Theorem (Peano existence theorem). *Let $\Omega \subset \mathbb{R} \times \mathbb{R}$ be open, $(t_0, y_0) \in \Omega$, $F \in C(\Omega)$,*

$$R = \{(t, y) \,:\, |t - t_0| \leqslant a,\ |y - y_0| \leqslant b\} \subset \Omega.$$

Let $M = \sup_{(t,y)\in R} |F(t, y)| < \infty$. Define $a_* = \min(a, b/M)$ and let $I = [t_0 - a_*, t_0 + a_*]$. Then the initial value problem

$$\begin{cases} y'(t) = F(t, y(t)), \\ y(t_0) = y_0. \end{cases} \tag{6.36}$$

has a solution $y \in C^1(I)$.

Proof By Lemma 6.100 it suffices to produce a continuous solution to the integral equation

$$y(t) = y_0 + \int_{t_0}^{t} F(s, y(s))\, ds. \tag{6.37}$$

on the interval $[t_0 - a_*, t_0 + a_*]$. We restrict our attention to the interval $[t_0, t_0 + a_*]$, which we denote by I. The construction is similar on the other half, $[t_0 - a_*, t_0]$.

Let $\mathcal{P}$ be a partition $\mathcal{P} = \{t_0 < t_1 < \cdots < t_N = t_0 + a_*\}$ of $[t_0, t_0 + a_*]$. We let $\Delta\mathcal{P} = \max_{0\leqslant k\leqslant N-1}(t_{k+1} - t_k)$ denote the *maximal width* of $\mathcal{P}$. We try to build an approximate solution given as a continuous piecewise linear function. The function $y_\mathcal{P} : [t_0, t_0 + a_*] \to \mathbb{R}$ shall be defined so that $y_\mathcal{P}(t_0) = y_0$ and so that on every partition interval $[t_k, t_{k+1}]$ the function is given by $y_\mathcal{P}(t) - y_\mathcal{P}(t_k) = m_k(t - t_k)$ with $m_k = F(t_k, y_\mathcal{P}(t_k))$.

Claim 0. There is a unique continuous function $y_\mathcal{P}$ on $[t_0, t_0 + a_*]$ with $y_\mathcal{P}(t_0) = y_0$ such that the following properties hold.

(i) $y_\mathcal{P}(t) = y_\mathcal{P}(t_{k-1}) + F(t_{k-1}, y_\mathcal{P}(t_{k-1}))(t - t_{k-1})$ for $t_{k-1} \leqslant t \leqslant t_k$, $1 \leqslant k \leqslant N$

(ii) For $t_0 \leqslant t \leqslant t_0 + a_*$ we have $y_0 - M(t - t_0) \leqslant y_\mathcal{P}(t) \leqslant y_0 + M(t - t_0)$, in particular $|y_\mathcal{P}(t) - y_0| \leqslant Ma_* \leqslant b$.

Proof of Claim 0. We prove this by induction, establishing for each $1 \leqslant n \leqslant N$ the properties (i) and (ii) on $[t_0, t_n]$.

First define on $[t_0, t_1]$ $y_\mathcal{P}(t) = y_0 + F(t_0, y_0)(t - t_0)$. Then $y_\mathcal{P}(t_0) = y_0$ so that (i) holds for $t_0 \leqslant t \leqslant t_1$. Since $|F(t_0, y_0)| \leqslant M$ we also have (ii) for $t_0 \leqslant t \leqslant t_1$.

Now assume that $y_\mathcal{P}$ is already defined on $[t_0, t_n]$ and (i) and (ii) hold on the interval $[t_0, t_n]$. If $n = N$ we are done, so let's assume $n < N$. We define $y_\mathcal{P}(t) = y_\mathcal{P}(t_n) + F(t_n, y_\mathcal{P}(t_n))(t - t_n)$ for $t_n \leqslant t \leqslant t_{n+1}$. Clearly with the induction hypothesis on $[t_0, t_n]$ this makes $y_\mathcal{P}$ a continuous piecewise

linear function on $[t_0, t_{n+1}]$. By the induction hypothesis we have $y_0 - M(t-t_0) \leqslant y_{\mathcal{P}}(t) \leqslant y_0 + M(t-t_0)$ for $t_0 \leqslant t \leqslant t_n$, and now, since $|F(t_n, y_{\mathcal{P}}(t_n))| \leqslant M$ we get for $t_n \leqslant t \leqslant t_{n+1}$

$$y_{\mathcal{P}}(t) \leqslant y_{\mathcal{P}}(t_n) + M(t-t_n) \leqslant y_0 + M(t_n - t_0) + M(t-t_n) = y_0 + M(t-t_0)$$
$$y_{\mathcal{P}}(t) \geqslant y_{\mathcal{P}}(t_n) - M(t-t_n) \geqslant y_0 - M(t_n - t_0) - M(t-t_n) = y_0 - M(t-t_0)$$

so that in particular $(t, y_{\mathcal{P}}(t)) \in R$ for $t_0 \leqslant t \leqslant t_{n+1}$.

We have verified (i), (ii) on $[t_0, t_{n+1}]$ which finishes the induction.

Letting $n = N-1$ we have defined our polygonal function satisfying the properties (i) and (ii) on $[t_0, t_0 + a_*]$. This establishes the claim. □

Claim 1. For $t, t' \in [t_0, t_0 + a_*]$,

$$|y_{\mathcal{P}}(t) - y_{\mathcal{P}}(t')| \leqslant M|t - t'|.$$

Proof of Claim 1. Say $t' \in [t_k, t_{k+1}]$, $t \in [t_\ell, t_{\ell+1}]$, $k \leqslant \ell$. If $k = \ell$, then

$$|y_{\mathcal{P}}(t) - y_{\mathcal{P}}(t')| = |F(t_k, y_{\mathcal{P}}(t_k))(t-t')| \leqslant M|t-t'|.$$

If $k < \ell$, then

$$\begin{aligned}|y_{\mathcal{P}}(t) - y_{\mathcal{P}}(t')| &= |y_{\mathcal{P}}(t) - y_{\mathcal{P}}(t_\ell) + \sum_{j=k+1}^{\ell-1} (y_{\mathcal{P}}(t_{j+1}) - y_{\mathcal{P}}(t_j)) + y_{\mathcal{P}}(t_{k+1}) - y_{\mathcal{P}}(t')| \\ &\leqslant |y_{\mathcal{P}}(t) - y_{\mathcal{P}}(t_\ell)| + \sum_{j=k+1}^{\ell-1} |y_{\mathcal{P}}(t_{j+1}) - y_{\mathcal{P}}(t_j)| + |y_{\mathcal{P}}(t_{k+1}) - y_{\mathcal{P}}(t')| \\ &\leqslant M(t - t_\ell) + \sum_{j=k+1}^{\ell-1} M(t_{j+1} - t_j) + M(t_{k+1} - t') = M(t-t').\end{aligned}$$

and the claim is established. □

Now define

$$g_{\mathcal{P}}(t) = F(t_k, y_{\mathcal{P}}(t_k)) \quad \text{for } t \in (t_k, t_{k+1}].$$

Then $g_{\mathcal{P}}$ is a step function and $y'_{\mathcal{P}}(t) = g_{\mathcal{P}}(t)$ for $t \in (t_k, t_{k+1})$.

Let $\varepsilon > 0$. F is uniformly continuous on R, because R is compact (Theorem 4.77). Thus there exists $\delta = \delta(\varepsilon) > 0$ such that

$$|F(t,y) - F(t',y')| \leqslant \varepsilon \tag{6.38}$$

for all $(t,y), (t',y') \in R$ with $\|(t,y) - (t',y')\| \leqslant 2\delta$.

Claim 2. Suppose that $\Delta\mathcal{P} \leqslant \delta(\varepsilon)\min(1, M^{-1})$. Then we have for all $t \in [t_0, t_0 + a_*]$ that

$$y_{\mathcal{P}}(t) = y_0 + \int_{t_0}^{t} g_{\mathcal{P}}(s)\,ds \quad \text{and} \quad |g_{\mathcal{P}}(s) - F(s, y_{\mathcal{P}}(s))| \leqslant \varepsilon \text{ if } s \in (t_{k-1}, t_k).$$

Proof of Claim 2. We have for $t = t_k$:

$$\begin{aligned}y_{\mathcal{P}}(t_k) - y_0 &= y_{\mathcal{P}}(t_k) - y_{\mathcal{P}}(t_0) = \sum_{j=1}^{k} y_{\mathcal{P}}(t_j) - y_{\mathcal{P}}(t_{j-1}) \\ &= \sum_{j=1}^{k} F(t_{j-1}, y_{\mathcal{P}}(t_{j-1}))(t_j - t_{j-1}) = \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} g(s)\,ds = \int_{t_0}^{t_k} g_{\mathcal{P}}(s)\,ds.\end{aligned}$$

Similarly, for $t \in (t_k, t_{k+1})$:

$$y_{\mathcal{P}}(t) - y_{\mathcal{P}}(t_k) = F(t_k, y_{\mathcal{P}}(t_k))(t - t_k) = \int_{t_k}^{t} g_{\mathcal{P}}(s)\, ds.$$

Thus,

$$y_{\mathcal{P}}(t) = y_{\mathcal{P}}(t_k) + \int_{t_k}^{t} g_{\mathcal{P}}(s) ds = y_0 + \int_{t_0}^{t_k} g_{\mathcal{P}}(s)\, ds + \int_{t_k}^{t} g_{\mathcal{P}}(s)\, ds = y_0 + \int_{t_0}^{t} g_{\mathcal{P}}(s)\, ds.$$

Let $s \in (t_{k-1}, t_k)$. Then

$$|g_{\mathcal{P}}(s) - F(s, y_{\mathcal{P}}(s))| = |F(t_{k-1}, y_{\mathcal{P}}(t_{k-1})) - F(s, y_{\mathcal{P}}(s))|.$$

We have

$$|y_{\mathcal{P}}(t_{k-1}) - y_{\mathcal{P}}(s)| \leqslant M|t_{k-1} - s| \leqslant M(t_k - t_{k-1}) \leqslant M \cdot \Delta\mathcal{P} \leqslant \delta.$$

Also, $|t_{k-1} - s| \leqslant t_k - t_{k-1} \leqslant \Delta\mathcal{P} \leqslant \delta$. Thus,

$$\|(t_{k-1}, y_{\mathcal{P}}(t_{k-1})) - (s, y_{\mathcal{P}}(s))\| \leqslant 2\delta.$$

By (6.38),

$$|g_{\mathcal{P}}(s) - F(s, y_{\mathcal{P}}(s))| = |F(t_{k-1}, y_{\mathcal{P}}(t_{k-1})) - F(s, y_{\mathcal{P}}(s))| \leqslant \varepsilon.$$

□

Claim 3. Suppose that $\Delta\mathcal{P} \leqslant \delta(\varepsilon)\min(1, M^{-1})$. Then it holds for all $t \in [t_0, t_0 + a_*]$ that

$$|y_{\mathcal{P}}(t) - (y_0 + \int_{t_0}^{t} F(s, y_{\mathcal{P}}(s)) ds)| \leqslant \varepsilon a_*.$$

Proof of Claim 3. By Claim 2, the left hand side equals

$$\Big| \int_{t_0}^{t} (g_{\mathcal{P}}(s) - F(s, y_{\mathcal{P}}(s)))\, ds \Big| \leqslant \int_{t_0}^{t} |g_{\mathcal{P}}(s) - F(s, y_{\mathcal{P}}(s))|\, ds.$$

Claim 2 implies that this is no larger than $\varepsilon(t - t_0) \leqslant \varepsilon a_*$. □

Claim 3 says that $y_{\mathcal{P}}$ is almost a solution if the maximal width of the partition $\mathcal{P}$ is sufficiently small. In the final step we use a compactness argument to obtain an honest solution.

Claim 4. The set $\mathcal{F} = \{y_{\mathcal{P}} : \mathcal{P} \text{ partition}\} \subset C([t_0, t_0 + a_*])$ is relatively compact.

Proof of Claim 4. By Claim 1,

$$|y_{\mathcal{P}}(t) - y_{\mathcal{P}}(t')| \leqslant M|t - t'|.$$

This implies that $\mathcal{F}$ is equicontinuous. It is also bounded:

$$|y_{\mathcal{P}}(t)| \leqslant |y_{\mathcal{P}}(t_0)| + |y_{\mathcal{P}}(t) - y_{\mathcal{P}}(t_0)| \leqslant |y_0| + M|t - t_0| \leqslant |y_0| + Ma_*.$$

Thus the claim follows from the Arzelà-Ascoli theorem (Theorem 4.112). □

Proof of Theorem 6.105, conclusion. For $n \in \mathbb{N}$, choose a partition $\mathcal{P}_n$ with

$$\Delta\mathcal{P}_n \leqslant \delta(1/n)\min(1, M^{-1}).$$

By compactness of $\overline{\mathcal{F}}$, the sequence $(y_{\mathcal{P}_n})_{n\in\mathbb{N}} \subset \mathcal{F} \subset \overline{\mathcal{F}}$ has a convergent subsequence that converges to some limit $y \in C([t_0, t_0 + a_*])$. It remains to show that y is a solution to the integral equation (6.37). Let us denote that subsequence by $(y_n)_{n\in\mathbb{N}}$. By (uniform) continuity of F, we have that $F(s, y_n(s)) \to F(s, y(s))$ uniformly in $s \in [t_0, t]$ as $n \to \infty$. Thus,

$$\int_{t_0}^{t} F(s, y_n(s))\, ds \longrightarrow \int_{t_0}^{t} F(s, y(s))\, ds \quad \text{as } n \to \infty.$$

On the other hand, by Claim 3 we get

$$\left| y_n(t) - (y_0 + \int_{t_0}^{t} F(s, y_n(s))\, ds) \right| \leqslant \frac{a_*}{n} \longrightarrow 0 \quad \text{as } n \to \infty.$$

Therefore, y solves the integral equation (6.37). □

6.5.3. Systems of first-order ordinary differential equations. The theory for ordinary differential equations that we have developed turns out to be far more general. The proofs of the Picard–Lindelöf theorem and the Peano existence theorem can easily be extended to apply to systems of differential equations:

$$\begin{cases} y'(t) = F(t, y(t)), \\ y(t_0) = y_0 \end{cases}$$

for $F : \Omega \to \mathbb{R}^m$, $\Omega \subset \mathbb{R} \times \mathbb{R}^m$ open, $(t_0, y_0) \in \Omega$.

For the formulation of the following version for $\mathbb{R}^m$ valued functions we continue to use the Euclidean norm in $\mathbb{R}^m$.

6.106. Theorem (Picard–Lindelöf, vector-valued). *Let $E \subset \mathbb{R} \times \mathbb{R}^m$ be open, let $(t_0, y_0) \in E$, $F \in C(E; \mathbb{R}^m)$. Let $a > 0$ and $b > 0$ be small enough such that*

$$R = \{(t, y) \in \mathbb{R} \times \mathbb{R}^m : |t - t_0| \leqslant a, \|y - y_0\| \leqslant b\} \subset E.$$

Let $M = \sup_{(t,y)\in R} \|F(t, y)\| < \infty$. Assume that there exists $c \in (0, \infty)$ such that

$$\|F(t, y) - F(t, u)\| \leqslant c\|y - u\| \tag{6.39}$$

for all $(t, y), (t, u) \in R$. Define $a_* = \min(a, b/M)$ and let $I = [t_0 - a_*, t_0 + a_*]$. Then there exists a unique solution $y = (y_1, \ldots, y_m) : I \to \mathbb{R}^m$ to the initial value problem

$$\begin{cases} y'(t) = F(t, y(t)), \\ y(t_0) = y_0. \end{cases} \tag{6.40}$$

6.107. Exercise. Rework the proof of the Theorem 6.98 so that it applies to Theorem 6.106

6.108. Exercise. Formulate and prove a version of the Peano existence theorem (Theorem 6.105) that applies to systems of differential equations.

6.5.4. Higher order differential equations. The result for first order systems of differential equations can also be applied to higher order differential equations.

6.109. Theorem (Picard–Lindelöf, Higher order version).

Let $\Omega \subset \mathbb{R} \times \mathbb{R}^d$ be open, and let $y_0 = (y_{0,0}, \ldots, y_{0,d-1})$ and $(t_0, y_0) \in \Omega$. Let $a > 0$ and $b > 0$ be small enough such that

$$R = \{(t, y) \in \mathbb{R} \times \mathbb{R}^d : |t - t_0| \leqslant a, \|y - y_0\| \leqslant b\} \subset \Omega.$$

Let G be continuous on R and assume that there exists $c \in (0, \infty)$ such that

$$|G(t, y) - G(t, u)| \leqslant c\|y - u\| \tag{6.41}$$

for all $(t, y), (t, u) \in R$.

Then there exists $a_* > 0$ so that on $I = [t_0 - a_*, t_0 + a_*]$ there exists a unique solution $t \mapsto y(t)$ satisfying the initial value problem

$$y^{(d)}(t) = G(t, y(t), y'(t), \ldots, y^{(d-1)}(t)) \tag{6.42a}$$

$$y(t_0) = y_{0,0},\ y'(t_0) = y_{0,1},\ \ldots,\ y^{(d-1)}(t_0) = y_{0,d-1}. \tag{6.42b}$$

The idea is to reformulate this as a first order system of d equations by introducing "dummy" variables $Y_1, \dots, Y_{d-1}$ for the derivatives $y', \dots, y^{(d-1)}$.

Let us consider, with suitable vectors $Y = (Y_0, \dots, Y_{d-1})$ in $\mathbb{R}^d$ the function

$$F(t, Y_0, \dots, Y_{d-1}) = \big(Y_1, \dots, Y_{d-1}, G(t, Y_0, Y_1, \dots, Y_{d-1})\big) \tag{6.43}$$

Consider the initial value problem

$$\begin{cases} Y'(t) = F(t, Y(t)), \\ Y(t_0) = (y_{0,0}, \dots, y_{0,d-1}) \end{cases} \tag{6.44}$$

We then observe that if $Y(t) = (Y_0(t), Y_1(t), \dots, Y_{d-1}(t))$ satisfies (6.44) then $Y_j'(t) = Y_{j+1}(t)$ and thus $Y_0^{(j)}(t) = Y_j(t)$, for $j = 1, \dots, d-1$. Hence $Y_0^{(d)}(t) = G(t, Y_0(t), \dots, Y_0^{(d-1)}(t))$. Thus $y(t) = Y_0(t)$ satisfies the initial value problem (6.42a), (6.42b).

Vice versa if $y(t)$ satisfies the initial value problem (6.42a), (6.42b) then

$$Y(t) = \big(y(t), y'(t), \dots, y^{(d-1)}(t)\big)$$

satisfies (6.44).

Consequently, applying Theorem 6.106 to the system (6.44) gives the conclusion of Theorem 6.109.

6.6. Further exercises

6.110. Exercise. Let X, Y be normed vector spaces and $F : X \to Y$ a map. Let F be Fréchet differentiable at $a \in X$. Show that F is continuous at a.

6.111. Exercise. Let $\mathbb{R}^{n\times n}$ denote the space of real $n \times n$ matrices and define $F(A) = A^m$ where $m \in \mathbb{N}$. Show that F is Fréchet differentiable and compute $DF|_A$.

6.112. Exercise. Let $\mathbb{R}^{n\times n}$ be the space of $n \times n$ matrices with real entries. Let $\mathrm{Sym}(n)$ be the space of all symmetric matrices (i.e. satisfying $A = A^\intercal$). Let $A \in \mathrm{Sym}(n)$. Define

$$F : \mathbb{R}^{n\times n} \to \mathrm{Sym}(n), \quad F(M) = M^\intercal A M$$

Prove that F is Fréchet differentiable at M and compute the derivative.

6.113. Exercise. Let $X = C([0,1])$ be the Banach space of continuous functions on $[0,1]$ (with the supremum norm) and define a map $F : X \to X$ by

$$F(f)(s) = \int_0^s \cos(f(t)^2)\, dt, \quad s \in [0,1].$$

(i) Show that F is Fréchet differentiable and compute the Fréchet derivative $DF\,|_f$ for each $f \in X$.
(ii) Show that $F(X) = \{F(f) : f \in X\} \subset X$ is relatively compact.

6.114. Exercise. For all $k \in \mathbb{N}$ the trace function $\mathrm{tr} : \mathbb{R}^{k\times k} \to \mathbb{R}$ is given by $\mathrm{tr}(A) = \sum_{i=1}^k a_{ii}$. Let $A, B \in \mathbb{R}^{m\times n}$. Prove that

$$\sum_{i=1}^m \sum_{j=1}^n a_{ij} b_{ij} = \mathrm{tr}(AB^\intercal) = \mathrm{tr}(BA^\intercal) = \mathrm{tr}(A^\intercal B) = \mathrm{tr}(B^\intercal A)\,.$$

6.115. Exercise. Consider the determinant function

$$\det : \mathbb{R}^{n\times n} \to \mathbb{R}, \qquad A \mapsto \det(A).$$

(i) Let $n = 2$ and $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$. Then $\det(A) = a_{11}a_{22} - a_{12}a_{21}$. Prove that det is totally differentiable and that the derivative $D\det$ at A is given by

$$D\det|_A[H] = a_{22}h_{11} + a_{11}h_{22} - a_{21}h_{12} - a_{12}h_{21}.$$

(ii) Prove a similar formula for the differentiation of the determinant when acting on 3×3 matrices.

6.116. Exercise. We extend Exercise 6.115 to all $n \in \mathbb{N}$. See Exercise 6.114 for the definition of trace.

For $A \in \mathbb{R}^{n\times n}$ denote by $\operatorname{adj}(A) = \operatorname{cof}(A)^\intercal$ its adjugate matrix (the transpose of the cofactor matrix). Here $\operatorname{cof}(A)_{ij} = (-1)^{i+j} M_{ij}$ where the minor M_{ij} is the determinant of the matrix obtained from A by deleting its ith row and jth column.

Prove that $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$ is totally differentiable and that

$$D\det|_A[H] = \operatorname{tr}(\operatorname{adj}(A)H).$$

Compare with Exercise 6.115.

6.117. Exercise. Suppose that $(t, x, v) \mapsto \kappa(t, x, v)$ is a C^2 function on $[a, b] \times \mathbb{R}^2$, i.e. all partial derivatives up to order 2 are continuous. Define a linear functional acting on real-valued functions $t \mapsto x(t)$ in $C^1([a, b])$ by

$$\mathcal{J}(x) = \int_a^b \kappa(t, x(t), x'(t))\, dt\,.$$

Show that the Fréchet derivative at the function x is given as a linear operator acting on C^1 functions h by

$$\begin{aligned} D\mathcal{J}_x[h] = \int_a^b \Big(\frac{\partial\kappa}{\partial x}(t, x(t), x'(t)) - \frac{d}{dt}\Big[\frac{\partial\kappa}{\partial v}(t, x(t), x'(t))\Big]\Big)h(t)\, dt \\ + \frac{\partial\kappa}{\partial v}(b, x(b), x'(b))h(b) - \frac{\partial\kappa}{\partial v}(a, x(a), x'(a))h(a) \end{aligned}$$

6.118. Exercise. Repeat Exercise 6.117 with the space $C^1([a, b])$ replaced by the subspace

$$\mathcal{Y} = \{x \in C^1([a, b]) : x(a) = x(b) = 0\}$$

Compute $DJ_x : \mathcal{Y} \to \mathbb{R}$, and show that if $DJ_x = 0$ the function $t \mapsto x(t)$ satisfies the *Euler–Lagrange equation*

$$\frac{\partial\kappa}{\partial x}(t, x(t), x'(t)) = \frac{d}{dt}\Big[\frac{\partial\kappa}{\partial v}(t, x(t), x'(t))\Big]$$

6.119. Exercise. Let $U \subseteq \mathbb{R}^n$ be open and convex and $f : U \to \mathbb{R}$ differentiable such that $\partial_1 f(x) = 0$ for all $x \in U$.

(i) Show that the value of $f(x)$ for $x = (x_1, \dots, x_n) \in U$ does not depend on x_1.
(ii) Does (i) still hold if we assume that U is connected instead of convex? Give a proof or counterexample.

6.120. Exercise. One can show that the integral formulation of the mean value theorem (i.e. the fundamental theorem of calculus) in Theorem 6.23 remains true in the vector-valued category, for functions with values in a Banach space. This requires the notion of a Riemann integral for Y-valued functions, see Exercise 5.125.

Let X be a normed space, and let $U \subset X$ be open and convex. Let Y be a Banach space, let $f : U \to Y$ be differentiable in U. Then for every $x, y \in U$,

$$f(x) - f(y) = \int_0^1 Df|_{tx+(1-t)y}(x - y)\, dt.$$

First prove a weaker version of this statement: For every bounded linear functional $\lambda \in Y^*$,

$$\lambda\big(f(x) - f(y)\big) = \lambda\Big(\int_0^1 Df|_{tx+(1-t)y}(x - y)\, dt\Big).$$

Then use the Hahn-Banach theorem to derive the strong version above. Show that the Hahn-Banach theorem is not needed if Y is any finite dimensional normed space, or if Y is a Hilbert space, or if $Y = \ell^p$ for $1 \leqslant p \leqslant \infty$.

6.121. Exercise. This exercise generalizes the concept of partial derivatives. Let $X = X_1 \times \cdots \times X_n$ be a product of normed spaces and let Y be a normed space. Let $U \subset X$ be open, $a = (a_1, \ldots, a_n) \in U$ and let $f : U \to Y$. If the Y-valued function

$$x_j \to f(a_1, \ldots, a_{j-1}, x_j, a_{j+1}, \ldots, a_n)$$

is Fréchet differentiable at a_j we call this derivative the partial derivative $D_j f|_a$, it is an element of $L(X_j, Y)$.

Suppose that the partial derivatives $D_j f|_x$ exist for all $x \in U$, $j = 1, \ldots, n$, and that they are continuous at $x = a$. Prove that f is Fréchet differentiable at a and $Df|_a[h] = \sum_{j=1}^n D_j f|_a[h_j]$, for $h = (h_1, \ldots, h_n)$.

6.122. Exercise. Let X, Y, Z be Banach spaces, Ω an open subset of $X \times Y$, $(x_0, y_0) \in \Omega$. Let $f : \Omega \to Z$ be of class C^1 and assume that $f(x_0, y_0) = 0$ and that the partial derivative $D_2 f|_{(x_0,y_0)}$ is an isomorphism (i.e. invertible in $L(Y, Z)$). Show that the equation $f(x, y) = 0$ can be solved with a Y-valued C^1 function g defined near x_0, and $g(x_0) = y_0$, and derive a formula for $Dg|_x$. This is a generalization of the implicit function theorem 6.46.

6.123. Exercise. Let $R = [a, b] \times [c, d]$, and $f : R \to \mathbb{R}$ be continuously differentiable. Let $\alpha, \beta : [c, d] \to [a, b]$ be differentiable functions and for $x \in [c, d]$,

$$\mathcal{I}(x) = \int_{\alpha(x)}^{\beta(x)} f(t, x)\, dt.$$

Prove $\mathcal{I}$ is differentiable on $[c, d]$ and

$$\mathcal{I}'(x) = \int_{\alpha(x)}^{\beta(x)} \frac{\partial f}{\partial x}(t, x)\, dt + f(\beta(x), x)\, \beta'(x) - f(\alpha(x), x)\, \alpha'(x).$$

6.124. Exercise. A function $f : \mathbb{R}^n \backslash \{0\} \to \mathbb{R}$ is called *homogeneous of degree* $\alpha \in \mathbb{R}$ if $f(tx) = t^\alpha f(x)$ for all $t > 0$ and $x \in \mathbb{R}^n \backslash \{0\}$. Suppose that f is differentiable in $\mathbb{R}^n \backslash \{0\}$. Then show that f is homogeneous of degree α if and only if

$$\sum_{i=1}^n x_i \partial_i f(x) = \alpha f(x)$$

for all $x \in \mathbb{R}^n \backslash \{0\}$.

6.125. Exercise. Define $F : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$F(x, y) = (x^4 - y^4, e^{xy} - e^{-xy}).$$

(i) Compute the Jacobian of F.

(ii) Let $p_0 \in \mathbb{R}^2$ and $p_0 \neq (0, 0)$. Show that there exist open neighborhoods $U, V \subset \mathbb{R}^2$ of p_0 and $F(p_0)$, respectively and a function $G : V \to U$ such that $G(F(p)) = p$ for all $p \in U$ and $F(G(p)) = p$ for all $p \in V$.

(iii) Compute $DG\,|_{F(p_0)}$.

(iv) Is F a bijective map?

6.126. Exercise. Prove that there exists $\delta > 0$ such that for all square matrices $A \in \mathbb{R}^{n \times n}$ with $\|A - I\| < \delta$ (where I denotes the identity matrix) there is a $B \in \mathbb{R}^{n \times n}$ such that $B^2 = A$.

6.127. Exercise. Look at each of the following as an equation to be solved for $x \in \mathbb{R}$ in terms of parameters $y, z \in \mathbb{R}$. Notice that $(x, y, z) = (0, 0, 0)$ is a solution for each of these equations. For each one, prove that it can be solved for x as a C^1-function of y, z in a neighborhood of $(0, 0, 0)$.

(a) $\cos(x)^2 - e^{\sin(xy)^3 + x} = z^2$

(b) $(x^2 + y^3 + z^4)^2 = \sin(x - y + z)$

(c) $x^7 + ye^z x^3 - x^2 + x = \log(1 + y^2 + z^2)$

6.128. Exercise. Let X, Y be Banach spaces over $\mathbb{R}$, X finite dimensional. Let $x_0 \in X$, let Ω be an open set with $x_0 \in \Omega$. Let $f : \Omega \to Y$ be in $C^1(\Omega, Y)$, and assume that $Df|_{x_0}$ is injective. Then there is a neighborhood U of x_0, a neighborhood V of $f(x_0)$ and a C^1 function $g : V \to U$ such that $g(f(x)) = x$ for all $x \in U$.

6.129. Exercise. One may drop the assumptions of finite dimensionality in Theorem 6.42 and Exercise 6.128, with an adjustment of the hypothesis. Prove:

(i) Let X, Y be Banach spaces over $\mathbb{R}$. Let $x_0 \in X$, let Ω be a neighborhood of x_0, $f : \Omega \to Y$ be of class C^1, and assume that there exists $T \in L(Y, X)$ such that $Df|_{x_0} \circ T = \mathrm{id}_Y$. Then there is a neighborhood V of $f(x_0)$, and a C^1 function $g : V \to U$ such that $f(g(y)) = y$ for all $y \in V$.

(ii) Let X, Y be Banach spaces over $\mathbb{R}$. Let $x_0 \in X$, let Ω be an open set with $x_0 \in \Omega$. Let $f : \Omega \to Y$ be in $C^1(\Omega, Y)$, and assume that there exists $T \in L(Y, X)$ such that $T \circ Df|_{x_0} = \mathrm{id}_X$. Then there is a neighborhood U of x_0, a neighborhood V of $f(x_0)$ and a C^1 function $g : V \to U$ such that $g(f(x)) = x$ for all $x \in U$.

6.130. Exercise. For a function $f : [a, b] \to \mathbb{R}$ define

$$\mathcal{I}(f) = \int_a^b (1 + f'(t)^2)^{1/2}\, dt.$$

Let $\mathcal{A} = \{f \in C^2([a, b]) : f(a) = c, f(b) = d\}$. Determine $f_* \in \mathcal{A}$ such that

$$\mathcal{I}(f_*) = \inf_{f \in \mathcal{A}} \mathcal{I}(f).$$

What is the geometric meaning of $\mathcal{I}(f)$ and $\inf_{f \in \mathcal{A}} \mathcal{I}(f)$?

6.131. Exercise. Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be smooth functions (that is, all partial derivatives exist to arbitrary orders and are continuous). Show that for all multiindices $\alpha \in \mathbb{N}_0^n$,

$$\partial^\alpha (f \cdot g)(x) = \sum_{\beta \in \mathbb{N}_0^n \,:\, \beta \leqslant \alpha} \binom{\alpha}{\beta} \partial^\beta f(x) \partial^{\alpha - \beta} g(x)$$

for all $x \in \mathbb{R}^n$, where $\binom{\alpha}{\beta} = \frac{\alpha!}{\beta!(\alpha-\beta)!} = \frac{\alpha_1! \cdots \alpha_n!}{\beta_1! \cdots \beta_n! (\alpha_1 - \beta_1)! \cdots (\alpha_n - \beta_n)!}$.

6.132. Exercise. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be such that $\partial_1 \partial_2 f$ exists everywhere. Does it follow that $\partial_1 f$ exists? Give a proof or counterexample.

6.133. Exercise. Show that every continuous function $f : [a, b] \to [a, b]$ has a fixed point.

6.134. Exercise. Let X be a real Banach space. Let $B = \{x \in X : \|x\| \leqslant 1\}$ and $\partial B = \{x \in X : \|x\| = 1\}$. Show that the following are equivalent:

(i) every continuous map $f : B \to B$ has a fixed point
(ii) there exists no continuous map $r : B \to \partial B$ such that $r(b) = b$ for all $b \in \partial B$.

6.135. Exercise. Determine the local minima and maxima of the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x, y) = (4x^2 + y^2) e^{-x^2 - 4y^2}.$$

What can one say about global minima and maxima?

6.136. Exercise. Let $E \subset \mathbb{R}^n$ be open, $f : E \to \mathbb{R}$ and $x \in E$. Assume that for y in a neighborhood of 0 we have

$$f(x + y) = \sum_{|\alpha| \leqslant k} c_\alpha y^\alpha + o(\|y\|^k)$$

as $y \to 0$ and

$$f(x+y) = \sum_{|\alpha| \leqslant k} \widetilde{c}_\alpha y^\alpha + o(\|y\|^k)$$

as $y \to 0$. Show that $c_\alpha = \widetilde{c}_\alpha$ for all $|\alpha| \leqslant k$.

6.137. Exercise. Let $D = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\}$. Determine the maximum and minimum values of the function $f : D \to \mathbb{R}$, $f(x,y) = 4x^2 - 3xy$.

6.138. Exercise. Let f be a real-valued function defined in the interval $[-2,2]$ satisfying

$$\frac{f(b)-f(a)}{b-a} \leqslant \frac{f(c)-f(b)}{c-b} \leqslant A$$

whenever $-2 \leqslant a < b < c \leqslant 2$ (i.e., f is a convex function). Show that there is $C \geqslant 0$ such that for $|h| \leqslant 1$

$$\int_{-1}^{1} |f(x+h) + f(x-h) - 2f(x)|\, dx \leqslant Ch^2.$$

6.139. Exercise. Let $f \in C^2(\mathbb{R}^n)$ and suppose that the Hessian of f is positive definite at every point. Show that $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is an injective map.

6.140. Exercise. For $f(x,y,z) = x + y + z$ determine all maxima and minima under the constraints $x^2 + y^2 = 2$ and $x + z - 1 = 0$. Use the method of Lagrange multipliers.

6.141. Exercise. Let $f \in C^2(\mathbb{R}^n)$ be strongly convex. Show that $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is a diffeomorphism (that is, show that it is differentiable, bijective and that its inverse is differentiable).

6.142. Exercise. Let $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c$ with $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n, c \in \mathbb{R}$. Assume that A is symmetric and positive definite. Show that f has a unique global minimum at some point x_* and determine $f(x_*)$ in terms of A, b, c.

6.143. Exercise. Prove that the point x_* from Exercise 6.142 can be computed using gradient descent: that is, if $x_0 \in \mathbb{R}^n$ arbitrary and

$$x_{n+1} = x_n - \alpha \nabla f(x_n)$$

for $n = 0, 1, 2, \ldots$, then the sequence $(x_n)_{n\in\mathbb{N}}$ converges to x_* for all starting points $x_0 \in \mathbb{R}^n$, provided that α is chosen sufficiently small.

6.144. Exercise. Let $\mathcal{D} \subset \mathbb{R}^2$ be a finite set. Define a function $E : \mathbb{R}^3 \to \mathbb{R}$ by

$$E(a,b,c) = \sum_{x\in\mathcal{D}} (ax_1^2 + bx_1 + c - x_2)^2.$$

(i) Show that E is convex.
(ii) Does there exist a set $\mathcal{D}$ such that E is strongly convex? Proof or counterexample.

6.145. Exercise.
(a) Find a convex function that is not bounded from below.
(b) Find a strictly convex function that is not bounded from below.
(c) If a function is strictly convex and bounded from below, does it necessarily have a critical point? (Proof or counterexample.)

6.146. Exercise.
(a) Give an example of a convex function that is not continuous.
(b) Let $f : (a,b) \to \mathbb{R}$. Show that if f is convex, then f is continuous.

6.147. Exercise. Construct a strictly convex function $f : \mathbb{R} \to \mathbb{R}$ such that f is not differentiable at x for every $x \in \mathbb{Q}$.

6.148. Exercise. Let $f \in C^2(\mathbb{R}^n)$. Recall that we defined f to be *strongly convex* if there exists $\beta > 0$ such that $\langle D^2 f|_x y, y\rangle \geqslant \beta\|y\|^2$ for every $x, y \in \mathbb{R}^n$. Show that f is strongly convex if and only if there exists $\gamma > 0$ such that

$$f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y) - \gamma t(1-t)\|x-y\|^2$$

for all $x, y \in \mathbb{R}^n, t \in [0,1]$.
(Consequently, that condition can serve as an alternative definition of strong convexity, which is also valid if f is not C^2.)

6.149. Exercise. (See also Exercise 7.152 as motivation for this exercise.) Fix a function $\sigma \in C^1(\mathbb{R})$ and define for $x \in \mathbb{R}^n, W \in \mathbb{R}^{m\times n}, v \in \mathbb{R}^m$,

$$\mu(x, W, v) = \sum_{i=1}^{m} \sigma((Wx)_i)v_i$$

Given a finite set of points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N) \in \mathbb{R}^n \times \mathbb{R}\}$ define

$$E(W, v) = \sum_{j=1}^{N} (\mu(x_j, W, v) - y_j)^2.$$

Is E necessarily convex? (Proof or counterexample.)

6.150. Exercise. Let $(t_0, y_0) \in \mathbb{R}^2$, $c \in \mathbb{R}$ and define $Y_0(t) = y_0$,

$$Y_n(t) = y_0 + c\int_{t_0}^{t} sY_{n-1}(s)ds.$$

(i) Which initial value problem does Y solve?
(ii) For the case $t_0 = 0$ compute $Y_n(t)$ and $Y(t) = \lim_{n\to\infty} Y_n(t)$.

6.151. Exercise. Let F, (x_0, y_0), M, C, a_* be as in the Picard–Lindelöf theorem (Theorem 6.98). On $[x_0 - a_*, x_0 + a_*]$, the method of successive approximation produces a sequence Y_n, satisfying $Y_0(x_0) = y_0$, $Y_n(x) = y_0 + \int_{x_0}^{x} F(t, Y_{n-1}(t))dt$, for $n \geqslant 1$. The Banach fixed point argument in the proof of the theorem gives that Y_n converges uniformly to the solution of the initial value problem. The convergence rate for large n is better than predicted by the Banach fixed point argument. To see this prove by induction that

$$\max_{|x-x_0|\leqslant a_*} |Y_n(x) - Y_{n-1}(x)| \leqslant \frac{MC^{n-1}a_*^n}{n!}$$

for $n \geqslant 1$. The same phenomenon occurred also in Exercise 4.168 on Volterra integral equations.

6.152. Exercise. Consider the initial value problem

$$\begin{cases} y'(t) = e^{y(t)^2} - \frac{1}{ty(t)}, \\ y(1) = 1. \end{cases}$$

Find an interval $I = (1-h, 1+h)$ such that this problem has a unique solution y in I. Give an explicit estimate for h (it does not need to be best possible).

6.153. Exercise. Consider the initial value problem

$$\begin{cases} y'(t) = t + \sin(y(t)), \\ y(2) = 1. \end{cases}$$

Find the largest interval $I \subseteq \mathbb{R}$ containing $t_0 = 2$ such that the problem has a unique solution y in I.

6.154. Exercise. Let $a > 0$ and $b > 0$. Prove that there is a unique differentiable function f defined on $(-\infty, \infty)$ satisfying $f(0) = 0$ and

$$f'(x) = a - b|f(x)|^{3/2}$$

for all x. Also show that $\lim_{x\to\infty} f(x)$ exists and determine this limit.

6.155. Exercise. Let g be the solution on some interval $[0,T)$ to the problem:

$$g'(t) = g(t)^2\big(2+\sin(e^t+g(t))\big)$$
$$g(0)=1.$$

Show that $T<1$.

6.156. Exercise. Prove the following version of the Picard–Lindelöf theorem:

Let I be a compact interval and let $F\in C(I\times\mathbb{R}^m,\mathbb{R}^m)$. Assume that there exists $C\in(0,\infty)$ such that

$$|F(t,y)-F(t,u)|\leqslant C\|y-u\|$$

for all (t,y), $(t,u)\in I\times\mathbb{R}^m$. Let $y_0\in\mathbb{R}^m$, $t_0\in I$. Then there exists a unique solution $y\in C^1(I,\mathbb{R}^m)$ to the initial value problem

$$\begin{cases} y'(t)=F(t,y(t)),\\ y(t_0)=y_0.\end{cases}$$

6.157. Exercise. Let X be a Banach space and let $T:X\to X$ be a bounded linear operator.

(i) Prove that the series

$$\exp(T)=\sum_{k=0}^{\infty}\frac{1}{k!}T^k$$

converges in $L(X,X)$. Here, as before $T^0=\mathrm{id}_X$ and T^k is the k-fold composition $T\circ\cdot\circ T$. Show that $\|\exp(T)\|_{\mathrm{op}}\leqslant e^{\|T\|_{\mathrm{op}}}$.

(ii) Prove that $\exp(T)$ is invertible with $(\exp(T))^{-1}=\exp(-T)$.

(iii) Let $S,T\in L(X,X)$ with $ST=TS$. Prove that $\exp(S+T)=\exp(S)\exp(T)$.

(iv) Show that in general

$$\exp(tS)\exp(tT)-\exp(t(S+T))=t^2R_2(S,T)-t^3R_3(S,T)+O(t^4),\qquad \text{as } t\to 0$$

where $R_2(S,T)=\frac{1}{2}[S,T]$ where $[S,T]=ST-TS$ is the commutator of S and T. Also express $R_3(S,T)$ using commutator notation.

(v) Let $X=\mathbb{R}^n$ and let $T(x)=Ax$, where $A\in\mathbb{R}^{n\times n}$. If A is symmetric, what can you say about $\exp(T)$?

(vi) Define $T:\mathbb{R}^2\to\mathbb{R}^2$ by $T(x)=(\lambda x_1+x_2,\lambda x_2)$. Compute $\exp(T)$.

6.158. Exercise. Let $a_{j,k}\in\mathbb{R}$, $j,k=1,\dots n$. Solve the initial value problem

$$y_j'(t)=\sum_{k=1}^{n}a_{jk}y_k(t),\quad j=1,\dots,n$$
$$y_j(0)=u_j,\quad j=1,\dots n$$

using the matrix exponential function (Exercise 6.157).

6.159. Exercise. Let $t\mapsto a_{j,k}(t)$ be continuous functions on an interval I, $j,k=1,\dots,n$. Let $t_0\in I$ and $u=(u_1,\dots,u_n)\in\mathbb{R}^n$. Then there is a unique solution $t\mapsto y(t)=(y_1(t),\dots,y_n(t))$ of the initial value problem

$$y_j'(t)=\sum_{k=1}^{n}a_{jk}(t)y_k(t),\quad j=1,\dots n$$
$$y_j(t_0)=u_j,\qquad j=1,\dots,n.$$

6.160. EXERCISE. Let I be a compact interval and let $t \mapsto a_j(t)$, $j = 0, \dots, d-1$ be continuous functions on I. Let $t_0 \in I$ and $v_0, \dots, v_{d-1} \in \mathbb{R}$. Prove that there is a unique $y \in C^d(I)$ satisfying

$$y^{(d)}(t) = \sum_{k=0}^{d-1} a_k(t) y^{(k)}(t), \quad t \in I$$
$$y(t_0) = v_0, \quad y'(t_0) = v_1, \ \dots, \ y^{(d-1)}(t_0) = v_{d-1}.$$

6.161. EXERCISE. Let F be a smooth function on $\mathbb{R}^2$ (i.e. partial derivatives of all orders exist everywhere and are continuous) and suppose that the initial value problem $y' = F(t, y)$, $y(t_0) = y_0$ has a unique solution y on the interval $I = [t_0, t_0 + a]$ with y smooth on I. Let $h > 0$ be sufficiently small and define $t_k = t_0 + kh$ for integers $0 \leqslant k \leqslant a/h$.
Define a function y_h recursively by setting $y_h(t_0) = y_0$ and

$$y_h(t) = y_h(t_k) + (t - t_k) F(t_k, y_h(t_k))$$

for $t \in (t_k, t_{k+1}]$ for integers $0 \leqslant k \leqslant a/h$.

(i) From the proof of Peano's theorem (Theorem 6.105) it follows that $y_h \to y$ uniformly on I as $h \to 0$. Prove the following stronger statement: there exists a constant $C > 0$ such that for all $t \in I$ and $h > 0$ sufficiently small,

$$|y(t) - y_h(t)| \leqslant Ch.$$

(ii) Let $F(t, y) = \lambda y$ with $\lambda \in \mathbb{R}$ a parameter. Explicitly determine y, y_h and a value for C in part (i).

6.162. EXERCISE. Let us improve the approximation from Exercise 6.161. In the context of that exercise, define a piecewise linear function y_h^* recursively by setting $y_h^*(t_0) = y_0$ and

$$y_h^*(t) = y_h^*(t_k) + (t - t_k) G(t_k, y_h^*(t_k), h),$$

for $t \in (t_k, t_{k+1}]$ for integers $0 \leqslant k \leqslant a/h$, where

$$G(t, y, h) = \tfrac{1}{2}(F(t, y) + F(t + h, y + hF(t, y))).$$

Prove that there exists a constant $C > 0$ such that for all $t \in I$ and $h > 0$ sufficiently small,

$$|y(t) - y_h^*(t)| \leqslant Ch^2.$$

6.163. EXERCISE. Let $\lambda : \mathbb{C} \to \mathbb{C}$ be $\mathbb{C}$-linear.

(i) Show that there is $c \in \mathbb{C}$ such that $\lambda(z) = cz$ for all $z \in \mathbb{C}$.

(ii) Let $\Lambda_1(x_1, x_2) = \operatorname{Re} \lambda(x_1 + ix_2)$, $\Lambda_2(x_1, x_2) = \operatorname{Im} \lambda(x_1 + ix_2)$, and $\Lambda(x_1, x_2) = (\Lambda_1(x_1, x_2), \Lambda_2(x_1, x_2))$. Prove that $\Lambda : \mathbb{R}^2 \to \mathbb{R}^2$ is $\mathbb{R}$-linear, and given by $\Lambda(x) = Ax$ where the 2×2-matrix A is either 0 or satisfies

$$A = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} = \begin{pmatrix} r\cos\theta & -r\sin\theta \\ r\sin\theta & r\cos\theta \end{pmatrix}$$

with suitable $\alpha, \beta \in \mathbb{R}$, $\alpha^2 + \beta^2 > 0$, $r > 0$, $\theta \in \mathbb{R}$.

The linear transformations $\Lambda : \mathbb{R}^2 \to \mathbb{R}^2$ of this form are called *conformal*; they are composites of dilations and rotations.

6.164. EXERCISE. Recall the notion of complex differentiability (see Definition 3.66). Let $w \in \mathbb{C}$ and let f, g be complex differentiable at w. Then $f + g$ and fg are complex differentiable at w and $(f + g)'(w) = f'(w) + g'(w)$ and

$$(fg)'(w) = f'(w)g(w) + f(w)g'(w).$$

If $f(w) \neq 0$ then $1/f$ is differentiable at w and $(1/f)'(w) = -f'(w)(f(w))^{-2}$.

6.165. Exercise. Consider a complex-valued function f defined in a neighborhood of $w = w_1 + iw_2 \in \mathbb{C}$. For (x_1, x_2) near (w_1, w_2) define

$$F_1(x_1, x_2) = \text{Re } f(x_1 + ix_2), \quad F_2(x_1, x_2) = \text{Im } f(x_1 + ix_2),$$

and $F(x_1, x_2) = (F_1(x_1, x_2), F_2(x_1, x_2))$. Assume that f is complex differentiable at w.

Prove that the partial derivatives $\frac{\partial F_1}{\partial x_1}(w)$, $\frac{\partial F_1}{\partial x_2}(w)$, $\frac{\partial F_2}{\partial x_1}(w)$, $\frac{\partial F_2}{\partial x_2}(w)$ exist and satisfy the *Cauchy-Riemann differential equations*,

$$\frac{\partial F_1}{\partial x_1}(w) = \frac{\partial F_2}{\partial x_2}(w), \qquad \frac{\partial F_1}{\partial x_2}(w) = -\frac{\partial F_2}{\partial x_1}(w).$$

The complex derivative at w satisfies

$$f'(w) = \frac{\partial F_1}{\partial x_1}(w) + i\frac{\partial F_2}{\partial x_1}(w) = \frac{\partial F_2}{\partial x_2}(w) - i\frac{\partial F_1}{\partial x_2}(w).$$

6.166. Exercise. Assume that f, F are as in Exercise 6.165. Show that the following three statements are equivalent.

(i) f is complex differentiable at w.
(ii) $F = (F_1, F_2)$ is differentiable at (w_1, w_2) and the total derivative $DF|_{(w_1,w_2)}$ is either zero or a conformal linear transformation.
(iii) $F = (F_1, F_2)$ is differentiable at (w_1, w_2) and satisfies the Cauchy-Riemann differential equations $\frac{\partial F_1}{\partial x_1}(w) = \frac{\partial F_2}{\partial x_2}(w)$, $\frac{\partial F_1}{\partial x_2}(w) = -\frac{\partial F_2}{\partial x_1}(w)$.

6.167. Exercise. Define the Wirtinger derivatives ∂_z, $\partial_{\bar{z}}$ by

$$\partial_z = \frac{1}{2}\Big(\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2}\Big), \quad \partial_{\bar{z}} = \frac{1}{2}\Big(\frac{\partial}{\partial x_1} + i\frac{\partial}{\partial x_2}\Big),$$

Let f, F be as in Exercise 6.165 and assume that F is differentiable at (w_1, w_2). Show that f is complex differentiable at w if and only if $\partial_{\bar{z}} f(w) = 0$. Moreover, in this case $f'(w) = \partial_z f(w)$.

6.168. Exercise.

(i) Let $f(x_1 + ix_2) = x_1^2 + x_2^2 + i2x_1x_2$. Is f complex differentiable on $\mathbb{C}$?

(ii) Let $f(x_1 + ix_2) = x_1^2 - x_2^2 + i2x_1x_2$. Is f complex differentiable on $\mathbb{C}$?

(iii) For $n \in \mathbb{N}$ let $f_1(z) = (\text{Re}\,(z))^n$, $f_2(z) = (\text{Im}\,(z))^n$, $f_3(z) = (\overline{z})^n$ and $f_4(z) = z^n$. For $\nu = 1, 2, 3, 4$ determine where f_ν is complex differentiable.

(iv) Determine where $\cos z$, $\cos \bar{z}$, $\sin z$, $\sin \bar{z}$ are complex differentiable.

6.169. Exercise. The definition of Fréchet derivative also applies when X and Y are complex normed spaces. The Fréchet derivative is then defined as a bounded complex linear map. Show that various key theorems in this chapter can be extended to this setting (such as the mean value inequalities, the chain rule, the inverse function theorem ...). Which theorems cannot be extended?

6.7. Appendix: Higher order differentials

We sketch the basics of the theory of higher derivatives and differentials in normed spaces. Higher differentials will be defined as bounded multilinear maps which we will study first.

6.7.1. Multilinear maps. Let $X_1, \ldots, X_n$ and Y be vector spaces and let

$$f : X_1 \times \cdots \times X_n \to Y$$

be a function. This function (map) is called *n-linear* if it is linear in the kth variable, for $k = 1, \ldots, n$. This means that for every k and every $a_i \in X_i$, $i \neq k$ the function

$$x_k \to f(a_1, \ldots, a_{k-1}, x_k, a_{k+1}, \ldots, a_n) \in Y$$

is a linear function from X_k to Y. 1-linear functions are called linear, 2-linear functions are called bilinear and 3-linear functions are trilinear. A *multilinear* function or map is k-linear for some $k \in \mathbb{N}$ (typically $k \geqslant 2$).

We now assume that the X_k are normed spaces, and use the norm

$$\|(x_1, \dots, x_n)\|_{X_1 \times \cdots \times X_n} = \max_{1 \leqslant k \leqslant n} \|x_k\|_{X_k}$$

on the product $\prod_{j=1}^n X_j = X_1 \times \cdots \times X_n$. We have characterizations of continuity of multilinear maps which are generalizations of the corresponding theorem for linear maps (Theorem5.11).

6.170. THEOREM. *Let $X_1, \dots, X_k$, Y be normed spaces and let $f : X_1 \times \cdots \times X_k \to Y$ be a k-linear map. Then the following are equivalent.*

(i) f is continuous at every $a = (a_1, \dots, a_k) \in X_1 \times \cdots \times X_k$.

(ii) f is continuous at $(0, \dots, 0) \in X_1 \times \cdots \times X_k$.

(iii) $\sup \{\|f(x_1, \dots, x_k)\|_Y : \|x_j\|_{X_j} \leqslant 1,\ j = 1, \dots, k\} < \infty$.

(iv) There exists $C \geqslant 0$ such that for all $x = (x_1, \dots, x_k)$

$$\|f(x_1, \dots, x_k)\|_Y \leqslant C \prod_{j=1}^k \|x_j\|_{X_j}$$

for all $x \in \prod_{j=1}^k X_j$.

PROOF. *(i) $\Longrightarrow$ (ii)* There is nothing to show.

(ii) $\Longrightarrow$ (iii) If (ii) holds there is a $\delta > 0$ such that $\|f(x)\|_Y \leqslant 1$ provided that $\|x_i\|_{X_i} \leqslant \delta$ for $i = 1, \dots, k$. By using the homogeneity in every variable, this shows that if $\|x_j\|_{X_j} \leqslant 1$ for $j = 1, \dots, k$ then $\|f(x)\|_Y \leqslant \delta^{-k}$. This gives (iii).

(iii) $\Longrightarrow$ (iv). (iv) follows from (iii) by using the homogeneity in each variable.

(iv) $\Longrightarrow$ (i). We observe that with $h_j = x_j - a_j$, $h = (h_1, \dots, h_k)$ we can write the difference $f(a + h) - f(a)$ as a telescoping sum

$$\begin{aligned}
&\sum_{j=1}^k f(a_1 + h_1, \dots, a_j + h_j, a_{j+1}, \dots, a_k) - f(a_1 + h_1, \dots, a_{j-1} + h_{j-1}, a_j, \dots, a_k) \\
&= \sum_{j=1}^k f(a_1 + h_1, \dots, a_{j-1} + h_{j-1}, h_j, a_{j+1}, \dots a_k).
\end{aligned}$$

By the triangle inequality,

$$\|f(a + h) - f(a)\|_Y \leqslant C \sum_{j=1}^k \|h_j\|_{X_j} \prod_{i:i<j} \|a_i + h_i\|_{X_i} \prod_{i:k \geqslant i > j} \|a_i\|_{X_i}.$$

Therefore, if $\max_{1 \leqslant i \leqslant k} \|h_i\| \leqslant 1$,

$$\|f(a + h) - f(a)\|_Y \leqslant Ck(1 + \max\{\|a_i\|_{X_i}\})^{k-1} \|h\|_{X_1 \times \cdots \times X_k}$$

and thus f is continuous at a. □

6.171. DEFINITION. Let $X_1, \dots, X_k, Y$ be normed spaces.

(i) A k-linear map $f : X_1 \times \cdots \times X_k \to Y$ is *bounded* if any of the equivalent conditions in Theorem 6.170 holds.

(ii) We denote by $\mathcal{L}^k(X_1 \times \cdots \times X_k; Y)$ the set of bounded k-linear maps from $X_1 \times \cdots \times X_k$ to Y. We use the notation $\mathcal{L}^k(X^k; Y)$ if $X_1 = \cdots = X_k = X$.

(iii) We denote by $\mathcal{L}^k_{\text{sym}}(X^k, Y)$ the set of all $f \in \mathcal{L}^k(X^k; Y)$ satisfying

$$f(x_{\pi(1)}, \dots, x_{\pi(k)}) = f(x_1, \dots, x_k)$$

for all permutations π on k letters.

6.172. THEOREM. *Let $X_1, \dots, X_k$, Y be normed spaces. Then:*

(i) $\mathcal{L}^k(X_1 \times \cdots \times X_k; Y)$ is a normed space with norm

$$\|f\|_{\text{op}} = \sup\big\{\|f(x_1, \dots, x_k)\|_Y : \max_{1 \leqslant j \leqslant k} \|x_j\|_{X_j} \leqslant 1\big\}.$$

(ii) If Y is a Banach space then so is $\mathcal{L}^k(X_1 \times \cdots \times X_k; Y)$.

(iii) The space $\mathcal{L}^k_{\text{sym}}(X^k, Y)$ is a closed subspace of $\mathcal{L}^k(X^k, Y)$.

This is left as an exercise. The arguments in (i) and (ii) are similar to the linear case.

6.173. THEOREM. *Let $f \in \mathcal{L}^k(X_1, \dots, X_k; Y)$. Then f is Frechet differentiable at any $a = (a_1, \dots, a_k) \in X_1 \times \cdots, \times X_k$, and the derivative is given by*

$$Df|_a[(h_1, \dots, h_k)] = \sum_{j=1}^{k} f(a_1, \dots, a_{j-1}, h_j, a_{j+1}, \dots, a_k).$$

PROOF. For $k = 1$ this is just Example 6.5. For $k = 2$ we use the bilinearity to expand

$$f(a_1 + h_1, a_2 + h_2) = f(a_1, a_2) + f(h_1, a_2) + f(a_1, h_2) + f(h_1, h_2)$$

where

$$\|f(h_1, h_2)\| \leqslant C\|h_1\|_{X_1}\|h_2\|_{X_2} \leqslant C\|h\|^2_{X_1 \times X_2}$$

which gives $Df|_a[(h_1, h_2)] = f(h_1, a_2) + f(a_1, h_2)$ and establishes the case $k = 2$. For the general case argue by induction. □

6.174. EXAMPLES (Product rules). Theorem 6.173 can be used to prove *product rules.* Let X, Y_1, Y_2, Z be normed spaces and $B : Y_1 \times Y_2 \to Z$ a bounded bilinear map. Let $a \in X$ and f_1, f_2 be defined near $a \in X$, with values in Y_1, Y_2, respectively, let both be differentiable at a and let

$$F(x) = B(f_1(x), f_2(x)).$$

Show that F is differentiable at a and $DF|_a \in L(X, Z)$ is given by

$$DF|_a h = B(Df_1|_a h, f_2(a)) + B(f_1(a), Df_2|_a h), \quad h \in X.$$

This can be proved directly or by using the chain rule.

Special cases are:

(a) $Y_1 = Y_2 = \mathbb{R}$ or $\mathbb{C}$ and $B(u, v) = uv$ the standard product.

(b) $Y_1 = Y_2 = \mathbb{R}^n$, and $B(u, v) = \sum_{i=1}^n u_i v_i$, the standard scalar product. Here $\mathbb{R}^n$ can be replaced by an inner product space with B the given inner product.

(c) We may have $Y_1 = \mathbb{R}^{m \times n}$, $Y_2 = \mathbb{R}^{k \times m}$ and $Z = \mathbb{R}^{k \times n}$, so f_1 has values in the set of $m \times n$ matrices, f_2 values in the set of $k \times m$ matrices and $B(f_1(x), f_2(x))$ is the matrix product $f_2(x)f_1(x)$ which gives for each x a $k \times n$ matrix.

(d) Given normed spaces V_1, V_2, V_3, we may have $Y_1 = L(V_1, V_2)$, $Y_2 = L(V_2, V_3)$, $Z = L(V_1, V_3)$ and B the bilinear map assigning to two linear maps their composition in $L(V_1, V_3)$.

It will be advantageous to identify linear maps in $L(X, L(X, Y))$ with bilinear maps via a canonical isomorphism.

6.175. THEOREM. *Let $F \in L(X, L(X, Y))$.*

(i) For each $u \in X$, $v \in X$ define $\imath F(u, v) = F(u)[v] \in Y$ (which is the value of the linear operator $F(u) \in L(X, Y)$ when evaluated at v). Then $\imath F \in \mathcal{L}^2(X^2; Y)$, and

$$\|\imath F\|_{\mathcal{L}^2(X^2, Y)} = \|F\|_{L(X, L(X, Y))}.$$

(ii) For $G \in \mathcal{L}^2(X^2; Y)$ and $u \in X$ let $\jmath G(u) : X \to Y$ be defined by $\jmath G(u)[v] = G(u, v)$. Then $\jmath G(u) \in L(X, Y)$ for all $u \in X$. Moreover, $u \mapsto \jmath G(u)$ is a bounded linear map $X \to L(X, Y)$.

(iii) The maps $\imath : L(X, L(X,Y)) \mapsto \mathcal{L}^2(X^2;Y)$ *and* $\jmath : \mathcal{L}^2(X^2;Y) \to L(X, L(X,Y))$ *are isometric isomorphisms of normed spaces.* $\jmath$ *is the inverse of* $\imath$*.*

This is an exercise of applying the definitions of the various spaces involved, and their norms. The details are left to the reader.

6.176. Definition. Let $L_1(X,Y) = L(X,Y)$ and, for $k \geqslant 2$, $L_k(X,Y) = L(X, L_{k-1}(X,Y))$.

6.177. Definition. Inductively define maps

$$\imath_k : L_k(X,Y) \to \mathcal{L}^k(X^k, Y)$$

as follows.

(i) Let $\imath_1 : L(X,Y) \to L(X,Y)$ be the identity map.

(ii) Let $\imath_2 : L_2(X,Y) = L(X, L(X,Y)) \to \mathcal{L}^2(X^2;Y)$ be the isomorphism defined by $\imath_2 F(x_1,x_2) = F(x_1)[x_2]$, as in the statement of Theorem 6.175.

(iii) Let $k > 2$ and $x_1 \in X$, $F \in L_k(X,Y) = L(X, L_{k-1}(X,Y))$. Then $F(x_1) \in L_{k-1}(X,Y)$, $\imath_{k-1}[F(x_1)] \in \mathcal{L}^{k-1}(X^{k-1};Y)$ and we let $\imath_k F : X_1 \times \cdots \times X_k \to Y$ be the k-linear map with

$$\imath_k F : (x_1, \ldots, x_k) \mapsto \imath_{k-1}(F(x_1))[x_2, \ldots, x_k].$$

6.178. Theorem. *For each* $F \in L_k(X,Y)$ *the map* $\imath_k F$ *belongs to* $\mathcal{L}^k(X^k;Y)$ *and*

$$\imath_k : L_k(X,Y) \to \mathcal{L}^k(X^k, Y)$$

defines an isometric isomorphism of normed spaces.

The proof is by induction on k. We refer to $\imath_k$ as the *canonical isomorphism* from $L_k(X,Y)$ to $\mathcal{L}^k(X^k, Y)$.

6.7.2. Higher order derivatives.

6.179. Definition. Let Ω be an open set in X, $a \in X$. Assume $k \geqslant 2$ and that f is $k-1$ times differentiable in Ω with $D^j f|_x \in L_j(X,Y)$, for $1 \leqslant j \leqslant k-1$.

(i) We say that f is k-times differentiable at a if the map $x \mapsto D^{k-1} f|_x$ is differentiable and call

$$D^k f|_a = D D^{k-1} f|_a \in L(X, L_{k-1}(X)) = L_k(X,Y)$$

the kth order Fréchet derivative at a.

(ii) Using the canonical isomorphism $\imath_k : L_k(X,Y) \to \mathcal{L}^k(X^k;Y)$ in Definition 6.177 we define the *kth order differential* of f at a as the k-linear map

$$\mathbb{D}^k f|_a = \imath_k D^k f|_a \in \mathcal{L}^k(X^k;Y).$$

6.180. Lemma. *Let* X, Y *be normed vector spaces,* $\Omega \subset X$ *be open,* $a \in \Omega$*. Let* $f : \Omega \to Y$ *and* $k \geqslant 2$*. Then* f *is* k*-times differentiable at* a *if and only if* $\mathbb{D}^{k-1} f$ *is differentiable at* a*. In this case the derivative* $D\mathbb{D}^{k-1} f|_a \in L(X, \mathcal{L}^{k-1}(X^{k-1};Y))$ *is given as follows: For* $h \in X$*,* $D\mathbb{D}^{k-1} f|_a h \in \mathcal{L}^{k-1}(X^{k-1};Y)$ *is the bounded* $(k-1)$*-linear map with*

$$D\mathbb{D}^{k-1} f|_a h[u_1, \ldots, u_{k-1}] = \mathbb{D}^k f|_a [h, u_1, \ldots, u_{k-1}]\,.$$

Moreover,

$$\begin{aligned} \mathbb{D}^k f|_a [u_1, \ldots, u_k] &= D_{u_1} \cdots D_{u_k} f(a) \\ &= \frac{\partial}{\partial t_1} \cdots \frac{\partial}{\partial t_k} f(a + t_1 u_1 + \cdots + t_k u_k)\Big|_{t_1 = \cdots = t_k = 0}\,. \end{aligned}$$

Proof. If $\mathbb{D}^{k-1} f$ is defined in a neighborhood of a and differentiable at a then for x near a, $D^{k-1} f_x = \jmath_{k-1} \mathbb{D}^{k-1} f|_x$ where $\jmath_{k-1} : \mathcal{L}^{k-1}(X^{k-1};Y) \to L_{k-1}(X,Y)$ is the inverse of $\imath_{k-1}$. By Corollary 6.9, it is differentiable.

Vice versa, if $D^{k-1}f|_x$ is defined near a and f is k-times differentiable at a then $\mathbb{D}^{k-1}f|_x = \imath_{k-1}D^{k-1}f|_x$ for x near a. Again by Corollary 6.9 it is differentiable, and we get

$$D(\mathbb{D}^{k-1}f)|_a h = D\imath_{k-1}D^{k-1}f|_a = \imath_{k-1}(D^k f|_a h)\,.$$

Applying these $k-1$-linear maps to $(u_1,\dots,u_{k-1})$ we get

$$\begin{aligned} D(\mathbb{D}^{k-1}f)|_a h[u_1,\dots,u_{k-1}] &= \imath_{k-1}(D^k f|_a h)[u_1,\dots,u_{k-1}] \\ &= \imath_k \circ D^k f|_a[h,u_1,\dots,u_{k-1}] = \mathbb{D}^k f|_a[h,u_1,\dots,u_{k-1}] \end{aligned}$$

as claimed.

For the formula on the directional derivatives we argue by induction. For $k=1$ this is Theorem 6.12. Suppose $k \geqslant 2$, then by the formula above

$$\begin{aligned} \mathbb{D}^k f|_a[u_1,\dots,u_{k-1},u_k] &= D(\mathbb{D}^{k-1}f)|_a u_1[u_2,\dots,u_k] \\ &= \lim_{t\to 0} \frac{\mathbb{D}^{k-1}f|_{a+tu_1} - \mathbb{D}^{k-1}f|_a}{t}[u_2,\dots,u_k] && \text{(limit in } \mathcal{L}^{k-1}(X^{k-1};Y)) \\ &= \lim_{t\to 0} \frac{\mathbb{D}^{k-1}f|_{a+tu_1}[u_2,\dots,u_k] - \mathbb{D}^{k-1}f|_a[u_2,\dots,u_k]}{t} && \text{(limit in } Y) \\ &= \lim_{t\to 0} \frac{D_{u_2}\cdots D_{u_k}f(a+tu_1) - D_{u_2}\cdots D_{u_k}f(a)}{t} && \text{(by the induction hypothesis)} \\ &= D_{u_1}\cdots D_{u_k}f(a). \end{aligned}$$

□

A key result is:

6.181. Theorem. *Let $\Omega \subset X$ be open, $a \in \Omega$, and let f be k times differentiable at a. Then $\mathrm{D}^k f|_a$ belongs to $\mathcal{L}^k_{\mathrm{sym}}(X;Y)$.*

Proof. We first start with the case $k=2$ which is proved mostly similarly to Theorem 6.51; we replace the two coordinate directions in $\mathbb{R}^n$ there by arbitrary direction vectors $u, v \in X$. Consider the quantity

$$\mathcal{M}_{h,k}f(a) = f(a+h+k) - f(a+h) - f(a+k) + f(a)$$

which satisfies $\mathcal{M}_{h,k}f(a) = \mathcal{M}_{k,h}f(a)$. We set $h = tu$, $k = tv$ for fixed u, v and study this term for $t \to 0$. One proves that

$$\mathcal{M}_{tu,tv}f(a) - t^2 D_v D_u f(a) = o(t^2)$$

as $t \to 0$. As in the proof of Theorem 6.51 this implies $t^2\big|D_u D_v f(a) - D_v D_u f(a)\big| = o(t^2)$. This gives $D_u D_v f|_a = D_v D_u f|_a$, and by Lemma 6.180 the symmetry of $\mathrm{D}^2 f|_a$.

Split $\mathcal{M}_{tu,tv}f(a) - t^2 D_v D_u f(a) = I(t) + II(t)$ where

$$\begin{aligned} I(t) &= \mathcal{M}_{tu,tv}f(a) - tD_u f(a+tv) + tD_u f(a), \\ II(t) &= tD_u f(a+tv) - tD_u f(a) - t^2 D_v D_u f(a)\,. \end{aligned}$$

By the Fréchet differentiability of Df (and therefore of $D_u f$) at a,

$$D_u f(a+tv) - D_u f(a) - tD_v D_u f(a) = o(t)$$

and thus $II(t) = o(t^2)$ as $t \to 0$.

To show that $I(t) = o(t^2)$ define for β in a neighborhood of the interval $[0,1]$ the function

$$t \mapsto G_t(\beta) = f(a+t\beta u + tv) - f(a+t\beta u) - t\beta D_u f(a+tv) + t\beta D_u f(a)$$

which satisfies

$$G_t(1) - G_t(0) = \mathcal{M}_{tu,tv}f(a) - tD_u f(a+tv) + tD_u f(a).$$

By the mean value inequality

$$\|G_t(1) - G_t(0)\|_Y \leqslant \sup_{0\leqslant\beta\leqslant 1} \|G_t'(\beta)\|_Y$$

and we need to check that the right hand side is $o(t^2)$. Fix β and compute

$$G_t'(\beta) = tD_uf(a + t\beta u + tv) - tD_uf(a + t\beta u) - tD_uf(a + tv) + tD_uf(a).$$

By the Fréchet differentiability of D_uf at a

$$\begin{aligned} D_uf(a + t\beta u + tv) &= D_uf(a) + t\beta D_u^2 f(a) + tD_vD_uf(a) + o(t),\\ D_uf(a + t\beta u) &= D_uf(a) + t\beta D_u^2 f(a) + o(t),\\ D_uf(a + tv) &= D_uf(a) + tD_vD_uf(a) + o(t). \end{aligned}$$

Using these in the expression for $G_t'(\beta)$ we get $\sup_{0\leqslant\beta\leqslant 1} \|G_t'(\beta)\|_Y = o(t^2)$ and thus $II(t) = o(t^2)$. This finishes the proof of the theorem in the case $k = 2$.

We argue by induction and assume $k \geqslant 3$, and that $\mathbb{D}^{k-1}f|_x$ exists near a and belongs to $\mathcal{L}^{k-1}_{\text{sym}}(X;Y)$. Denote by $\mathfrak{S}_k$ the permutations on k letters. Let $\pi \in \mathfrak{S}_k$. By Lemma 6.180 we need to show that

$$D_{u_{\pi(1)}} \cdots D_{u_{\pi(k)}} f|_a = D_{u_1} \dots D_{u_k} f|_a. \tag{6.45}$$

We distinguish two cases.

Case 1: $\pi(k) = k$. In this case there is a $\varpi \in \mathfrak{S}_{k-1}$ with $\varpi(\pi(j)) = j$ for $j = 1, \dots, k-1$ and by applying the induction hypothesis, we get $D_{u_{\pi(1)}} \cdots D_{u_{\pi(k-1)}} f|_x = D_{u_1} \cdots D_{u_{k-1}} f|_x$ for x near a, and (6.45) follows.

Case 2: $\pi(k) \neq k$. There is a permutation $\sigma \in \mathfrak{S}_{k-1}$ such that $\sigma(\pi(k-1)) = k$. Then by the induction hypothesis,

$$\begin{aligned} &D_{u_{\pi(1)}} \cdots D_{u_{\pi(k)}} f|_a = D_{u_{\sigma\circ\pi(1)}} \dots D_{u_{\sigma\circ\pi(k-2)}} D_{u_k} D_{u_{\pi(k)}} f|_a\\ &= D_{u_{\sigma\circ\pi(1)}} \dots D_{u_{\sigma\circ\pi(k-2)}} D_{u_{\pi(k)}} D_{u_k} f|_a = D_{u_1} \dots D_{u_k} f|_a \end{aligned}$$

since $D_{u_k} D_{\pi(u_k)} f = D_{\pi(u_k)} D_{u_k}$ as shown before, and the last identity follows by Case 1. This finishes the proof of (6.45). □

6.182. Definition. f is of class C^k on Ω if f is k times differentiable in Ω and $x \mapsto \mathbb{D}^k f|_x$ is continuous as a function from Ω to $\mathcal{L}^k(X^k;Y)$. f is of class C^∞ if f is of class C^n for all $n \in \mathbb{N}$.

Remark. Let Ω be an open subset of $\mathbb{R}^n$. Definition 6.182 is essentially the same as Definition 6.22 for $k = 1$, and Definition 6.50 (ii) for $k = 1$ and $k = 2$. For $Y = \mathbb{R}$ one verifies using $\partial_{x_j} \mathbb{D}^{k-1} f|_x = \mathbb{D}_{k-1}(\partial_{x_j} f)|_x$ and induction that Definition 6.182 is consistent with Definition 6.58 for all $k \in \mathbb{N}$.

6.183. Exercise. Let X be a normed space, $U \subset X$ open. Let Y_j, $j = 1, \dots m$ be normed spaces, let $f_j \in C^k(U, Y_j)$, $j = 1, \dots, m$ and let $f = (f_1, \dots, f_m)$, $f : U \to Y = Y_1 \times \cdots \times Y_m$. Show that $f \in C^k(U, Y)$.

6.184. Exercise. Let $F : X_1 \times \cdots \times X_n \to Y$ be a bounded n-linear map. Prove that F is of class C^∞ and that $\mathbb{D}^k F = 0$ for $k > n$.

6.185. Theorem. *Let X, Y, Z be normed spaces, and let $U \subset X$, $V \subset Y$ be open sets. Let $g : U \to V$ and $f : V \to Z$ be of class C^n. Then $f \circ g : U \to Z$ is of class C^n.*

Proof. For $n = 0$ this is just the fact that compositions of continuous functions are continuous. We assume $n \geqslant 1$ and that the theorem is already proved for compositions of C^{n-1} functions. We use the chain rule (Theorem 6.8):

$$D(f \circ g)|_x = Df|_{g(x)} \circ Dg|_x =: \text{Comp}(x).$$

It suffices to prove that $\mathrm{Comp} : x \mapsto Df|_{g(x)} \circ Dg|_x$ is a C^{n-1} function from X to $L(X,Z)$. We write $\mathrm{Comp} = \mathcal{B} \circ \mathcal{A}$ where

$$\mathcal{B} : \begin{cases} L(X,Y) \times L(Y,Z) \to L(X,Z) \\ (S,T) \mapsto T \circ S \end{cases} \qquad \mathcal{A} : \begin{cases} X \to L(X,Y) \times L(Y,Z) \\ x \mapsto (Dg|_x, Df|_{g(x)}) \end{cases} .$$

$\mathcal{B}$ is a bilinear map and also bounded (in view of $\|T \circ S\|_{L(X,Z)} \leqslant \|S\|_{L(X,Y)}\|T\|_{L(Y,Z)}$). Thus $\mathcal{B}$ is of class C^∞ (see Exercise 6.184). The function $x \mapsto Dg|_x$ is of class C^{n-1} by assumption and $x \mapsto Df|_{g(x)}$ is a composition of two C^{n-1} functions, hence by the induction hypothesis of class C^{n-1}. This implies that $\mathcal{A}$ is of class C^{n-1}. Hence, again by the induction hypothesis $\mathrm{Comp} = \mathcal{B} \circ \mathcal{A}$ is a C^{n-1} function which implies that $f \circ g$ is a C^n function. □

6.186. Exercise. Let X, Y, Z be normed spaces, $U \subset X$ be open, $V \subset Y$ be open, let $f : U \to V$ be twice differentiable on U, and let $g : V \to Z$ be twice differentiable on V. Prove that for $a \in U$

$$\mathbb{D}^2(g \circ f)|_a[h,k] = \mathbb{D}g_{f(a)}\big[\mathbb{D}^2 f|_a[h,k]\big] + \mathbb{D}^2 g|_{f(a)}\big[\mathbb{D}f|_a h, \mathbb{D}f|_a k\big] .$$

6.187. Theorem. *Let X, Y be Banach spaces and let $\mathrm{Iso}(X,Y)$ be the open set of bounded linear bijective maps from $X \to Y$ which have a bounded inverse. The function*

$$\mathrm{Inv} : \begin{cases} \mathrm{Iso}(X,Y) \to \mathrm{Iso}(Y,X) \\ T \to T^{-1} \end{cases}$$

is of class C^∞.

Proof. By Theorem 6.36 Inv is differentiable and $D(\mathrm{Inv})|_T[H] = -T^{-1}HT^{-1}$. We argue by induction and assume that Inv is of class C^{n-1} for some $n \geqslant 2$.

For $S_1 \in L(Y,X)$, $S_2 \in L(Y,X)$ let $\mathcal{B}(S_1,S_2)$ be the linear map from $L(X,Y)$ to $L(Y,X)$ which maps $H \in L(X,Y)$ to $\mathcal{B}(S_1,S_2)[H] = -S_1 H S_2 \in L(Y,X)$. The map $\mathcal{B}(S_1,S_2)$ is clearly bounded since $\|\mathcal{B}(S_1,S_2)[H]\|_{L(Y,X)} \leqslant \|S_1\|_{L(Y,X)}\|S_2\|_{L(Y,X)}\|H\|_{L(X,Y)}$.

The map $\mathcal{B}$ is a bilinear map from $L(Y,X) \times L(Y,X)$ to $L(L(X,Y),L(Y,X))$. By the preceding inequality it is bounded as a bilinear map since

$$\|\mathcal{B}(S_1,S_2)\|_{L(L(X,Y),L(Y,X))} \leqslant \|S_1\|_{L(Y,X)}\|S_2\|_{L(Y,X)}.$$

Thus $\mathcal{B}$ is of class C^∞ (Exercise 6.184). Let $\mathcal{A} : L(X,Y) \to L(Y,X) \times L(Y,X)$ be given by $\mathcal{A}(T) = (T^{-1}, T^{-1})$. Then by the induction hypothesis $\mathcal{A}$ is of class C^{n-1}. Now $D\,\mathrm{Inv} = \mathcal{B} \circ \mathcal{A}$ so $D\,\mathrm{Inv}$ is of class C^{n-1} which implies that Inv is of class C^n. Since n was arbitrary, Inv is of class C^∞. □

6.188. Exercise. Let X, Y be Banach spaces. Show that the n-th differential of $\mathrm{Inv} : \mathrm{Iso}(X,Y) \to \mathrm{Iso}(Y,X)$ is given by

$$\mathbb{D}^n \mathrm{Inv}_T[H_1,\dots,H_n] = (-1)^n \sum_{\pi \in \mathfrak{S}_n} T^{-1}H_{\pi(1)}T^{-1}H_{\pi(2)}T^{-1}\cdots T^{-1}H_{\pi(n)}T^{-1} ,$$

where $H_1,\dots,H_n \in L(X,Y)$ and the sum is extended over all permutations of $\{1,\dots,n\}$.

We now derive a strengthening of the inverse function theorem in Banach spaces.

6.189. Definition. Let X, Y be normed spaces, U open in X, V open in Y. A C^n function $f : U \to V$ is a *C^n-diffeomorphism* if f is bijective and the inverse $f^{-1} : V \to U$ is of class C^n.

In the inverse function theorem (Theorem 6.39) the assumption of f being C^1 near a, with $Df|_a$ invertible shows that f is a C^1 diffeomorphism from a neighborhood U of a to a neighborhood of $f(a)$. The next theorem tells us that if in addition $f \in C^n$ then f (restricted to U) is a C^n diffeomorphism.

6.190. Theorem. *Let X, Y be Banach spaces, U open in X, V open in Y. If $f : U \to V$ is a C^1 diffeomorphism, and if f is of class C^n then f is a C^n diffeomorphism.*

PROOF. Let $g : V \to U$ be the inverse of f. Since $f(g(y)) = y$ for all $y \in V$ we get $Df|_{g(y)} Dg|_y = \mathrm{id}_Y$ for all $y \in V$. Hence $Dg|_y = (Df|_{g(y)})^{-1}$. In other words $Dg = \mathrm{Inv} \circ Df \circ g$ where $g : V \to U$, $Df : U \to \mathrm{Iso(X, Y)}$ and $\mathrm{Inv} : \mathrm{Iso}(X, Y) \to \mathrm{Iso}(Y, X)$. Now Inv is a C^∞ map, Df a C^{n-1} map, so if $1 \leqslant k \leqslant n-1$ and g is a C^k map then Dg is also a C^k map, hence g is a C^{k+1} map. It follows that g is of class C^n on V. □

With the above terminology we can formulate Taylor's theorem.

6.191. EXERCISE. Let X, Y be normed spaces, $\Omega \subset X$ be an open convex set and let $f : \Omega \to Y$ be of class C^{n+1}. Let $a, a + h \in \Omega$ and let $D_h f(x) = \frac{d}{dt} f(x + th)|_{t=0}$ be the directional derivative with respect to h, evaluated at x. Then

$$f(a + h) = f(a) + \sum_{j=1}^{n} \frac{D_h^j f(a)}{j!} + \mathcal{R}_n f(a, h)$$

where

$$\mathcal{R}_n f(a, h) = \int_0^1 \frac{(1 - s)^n}{n!} D_h^{n+1} f(a + st)\, ds.$$

Here, with $(h, \dots, h) \in X^k$,

$$D_h^k f[a] = \mathbb{D}^k f|_a \underbrace{[h, \dots, h]}_{k \text{ times}} .$$

and

$$\|\mathcal{R}_n f(a, h)\|_Y \leqslant \sup_{0 \leqslant t \leqslant 1} \|\mathbb{D}^{n+1} f(a + th)\|_{\mathrm{op}} \frac{\|h\|_X^{n+1}}{(n + 1)!} .$$

CHAPTER 7

Approximation of functions

Let X be a (semi)normed vector space of functions (e.g. continuous functions on $[0,1]$) and $V \subset X$ be some subspace of it (e.g. polynomials). In this chapter we study how an arbitrary function $f \in X$ can be 'approximated' by functions g in V. The quality of approximation is measured by the error in norm, i.e. $\|f-g\|$. A basic question is whether we can make $\|f-g\|$ arbitrarily small. Of course the answer depends on the choice of V and $\|\cdot\|$. More precisely, we are asking whether V is *dense* in X. Recall that $V \subset X$ is called *dense* if $\overline{V} = X$. That is, if for every $f \in X$ and every $\varepsilon > 0$ there exists a $g \in V$ such that $\|f-g\| \leqslant \varepsilon$. In this chapter we will study several examples.

Density is a qualitative statement. We are also interested in finer questions about the quality of approximation, such as best approximation and rates of convergence. While the space V is typically infinite dimensional, it will often take the form

$$V = \bigcup_{N=1}^{\infty} V_N,$$

where the $V_N \subset X$ are finite dimensional subspaces (e.g. V_N could be polynomials of degree at most N). In this case we are interested in the *approximation errors*

$$\operatorname{dist}(f, V_N) = \inf_{g \in V_N} \|f-g\|.$$

A compactness argument shows that this infimum is always a minimum if the subspace is finite dimensional:

7.1. Proposition. *Let $V \subset X$ be a finite dimensional subspace and $f \in X$. Then there exists a* best approximation *$g \in V$ of f:*

$$\operatorname{dist}(f, V) = \|f-g\|.$$

Proof. Since V is finite dimensional, the closed ball

$$B = \{g \in V \,:\, \|g\| \leqslant 2\|f\|\}$$

is compact (Corollary 5.28). Thus the continuous function $g \mapsto \|f-g\|$ assumes its minimum in B (Theorem 4.102): there exists $g_* \in B$ such that $\|f-g_*\| \leqslant \|f-g\|$ for all $g \in B$. It remains to consider $g \in V \backslash B$. But then $\|g\| > 2\|f\|$, so by the triangle inequality $\|f-g\| \geqslant \|g\| - \|f\| > \|f\|$. But also $\operatorname{dist}(f,V) \leqslant \|f\|$ since $0 \in V$, so in summary $\operatorname{dist}(f,V) < \|f-g\|$ for all $g \in V \backslash B$. □

This proposition is promising, but we should note that it does not provide a practical method of obtaining the best approximation. It is often quite difficult to determine best approximations. Moreover, best approximations are not necessarily uniquely determined.

7.2. Example. Let $X = \mathbb{R}^2$ with norm $\|(x_1,x_2)\| = \max(|x_1|,|x_2|)$ and $V = \mathbb{R} \times \{0\}$. Consider $f = (x_1,x_2) \in X \backslash V$ (i.e. $x_2 \neq 0$). Then for every $\delta \in [-|x_2|, |x_2|]$, the vector $g = (x_1+\delta, 0) \in V$ is a best approximation to f in V with respect to the maximum norm, with $\|f-g\| = |x_2|$.

In this chapter, we will see several examples of dense subspaces and best approximations.

Remark. We consider functions taking values in either $\mathbb{R}$ or $\mathbb{C}$ and if nothing else is specified, we assume that values are allowed to be complex. $\mathbb{K}$ denotes either $\mathbb{R}$ or $\mathbb{C}$.

7.1. Weierstrass' theorem

We begin with the classical problem of approximating functions in the space $X = C([a,b])$ of continuous functions on a compact interval by polynomials. Denote by $\mathcal{P}_n$ the $(n+1)$-dimensional vector space of polynomials of degree at most n and let

$$\mathcal{P} = \bigcup_{n\geqslant 0} \mathcal{P}_n$$

denote the vector space of all polynomials (with complex coefficients). Recall that the norm on $C([a,b])$ is the supremum norm:

$$\|f\|_\infty = \sup_{t\in[a,b]} |f(t)|$$

and that convergence in the supremum norm is equivalent to uniform convergence on $[a,b]$.

7.3. Theorem (Weierstrass approximation theorem). *For every $f \in C([a,b])$ there exists a sequence of polynomials that converges to f uniformly on $[a,b]$.*

In other words, $\mathcal{P}$ is dense in $C([a,b])$. There are many proofs of this theorem in the literature and we will mention several of them in this chapter. Here we shall use *Bernstein polynomials.* We give the proof for the interval $[a,b] = [0,1]$. This is no loss of generality (by composing functions with the linear change of variables $t \mapsto a + t(b-a)$, which maps $[0,1]$ to $[a,b]$ and preserves both continuity and the space of polynomials).

Let f be continuous on $[0,1]$. For $n = 0, 1, 2, \ldots$, define

$$B_n f(t) = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\binom{n}{k} t^k (1-t)^{n-k}. \tag{7.1}$$

$B_n f$ is a polynomial of degree at most n. We will show that $B_n f$ converges to f uniformly on $[0,1]$. By the binomial theorem,

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

Thus,

$$B_n f(t) - f(t) = \sum_{k=0}^{n} \big(f(k/n) - f(t)\big)\binom{n}{k} t^k (1-t)^{n-k}. \tag{7.2}$$

Let $\varepsilon > 0$. By uniform continuity of f we may choose $\delta > 0$ such that $|f(t) - f(s)| \leqslant \varepsilon/2$ for all $t, s \in [0,1]$ with $|t-s| \leqslant \delta$. Write the sum on the right hand side of (7.2) as $\mathrm{I}_n + \mathrm{II}_n$, where

$$\mathrm{I}_n = \sum_{\substack{0\leqslant k\leqslant n:\\ |\frac{k}{n}-t|<\delta}} \big(f(k/n) - f(t)\big)\binom{n}{k} t^k (1-t)^{n-k},$$

and

$$\mathrm{II}_n = \sum_{\substack{0\leqslant k\leqslant n:\\ |\frac{k}{n}-t|\geqslant\delta}} \big(f(k/n) - f(t)\big)\binom{n}{k} t^k (1-t)^{n-k}.$$

We separately estimate I_n, for all $n \geqslant 1$, and II_n, for sufficiently large n. For the term I_n we use $|f(\frac{k}{n}) - f(t)| \leqslant \frac{\varepsilon}{2}$ for $|\frac{k}{n} - t| < \delta$ and obtain, for all $n \geqslant 1$,

$$|\mathrm{I}_n| \leqslant \frac{\varepsilon}{2} \sum_{k=0}^{n} \binom{n}{k} t^k (1-t)^{n-k} = \frac{\varepsilon}{2}.$$

To estimate II_n for sufficiently large n we first compute the Bernstein polynomials for the monomials $1, t, t^2$.

7.4. Lemma. *Let $g_m(t) = t^m$. Then*

(i) $B_n g_0(t) = 1$ *for* $n \geqslant 0$.
(ii) $B_n g_1(t) = t$ *for* $n \geqslant 1$.
(iii) $B_n g_2(t) = t^2 + \frac{t-t^2}{n}$ *for* $n \geqslant 2$.

Proof. We have

$$B_n g_0(t) = \sum_{k=0}^{n} \binom{n}{k} t^k (1-t)^{n-k} = (t + (1-t))^n = 1$$

by the binomial theorem. Next, using $\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}$ for $k \geqslant 1$, we obtain for $n \geqslant 1$,

$$B_n g_1(t) = \sum_{k=0}^{n} \frac{k}{n} \binom{n}{k} t^k (1-t)^{n-k} = \sum_{k=1}^{n} \binom{n-1}{k-1} t^k (1-t)^{n-k}$$

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

To compute $B_n g_2$ for $n \geqslant 2$ we use that

$$\frac{k^2}{n^2}\binom{n}{k} = \frac{k}{n}\binom{n-1}{k-1} = \frac{n-1}{n}\frac{k-1}{n-1}\binom{n-1}{k-1} + \frac{1}{n}\binom{n-1}{k-1}$$

$$= \frac{n-1}{n}\binom{n-2}{k-2} + \frac{1}{n}\binom{n-1}{k-1}.$$

Thus,

$$B_n g_2(t) = \frac{n-1}{n} \sum_{k=2}^{n} \binom{n-2}{k-2} t^k (1-t)^{n-k} + \frac{1}{n} \sum_{k=1}^{n} \binom{n-1}{k-1} t^k (1-t)^{n-k}$$

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

□

As a consequence, we obtain the following:

7.5. Lemma. *For all $t \in [0,1]$,*

$$\sum_{k=0}^{n} \left(\frac{k}{n} - t\right)^2 \binom{n}{k} t^k (1-t)^{n-k} \leqslant \frac{1}{n}.$$

Proof. From Lemma 7.4,

$$\sum_{k=0}^{n} \left(\frac{k}{n} - t\right)^2 \binom{n}{k} t^k (1-t)^{n-k} = B_n g_2(t) - 2t B_n g_1(t) + t^2 B_n g_0(t)$$

$$= t^2 + \frac{t-t^2}{n} - 2t^2 + t^2 = \frac{t-t^2}{n}.$$

Since $t \in [0,1]$, we have $0 \leqslant t - t^2 = t(1-t) \leqslant 1$. □

Now we are ready to estimate the term II_n. First note that f is bounded, so there exists $C \geqslant 0$ such that $|f(x)| \leqslant C$ for all $x \in [0,1]$. Choose $N \in \mathbb{N}$ such that $2C\delta^{-2}N^{-1} \leqslant \varepsilon/2$. Then for all $n \geqslant N$,

$$|\mathrm{II}_n| \leqslant 2C \sum_{\substack{0 \leqslant k \leqslant n, \\ |\frac{k}{n} - t| \geqslant \delta}} \binom{n}{k} t^k (1-t)^{n-k} \leqslant 2C\delta^{-2} \sum_{k=0}^{n} \left(\frac{k}{n} - t\right)^2 \binom{n}{k} t^k (1-t)^{n-k}$$

$$= 2C\delta^{-2} \frac{t-t^2}{n} \leqslant 2C\delta^{-2} N^{-1} \leqslant \frac{\varepsilon}{2}.$$

In the second inequality, we have used $\delta^{-2}|\frac{k}{n}-t|^2 \geqslant 1$. Thus if $n \geqslant N$ and $t \in [0,1]$, then

$$|B_n f(t) - f(t)| \leqslant |\mathrm{I}_n| + |\mathrm{II}_n| \leqslant \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

This concludes the proof of Weierstrass' theorem. □

REMARK. Given Example 7.2, it is perhaps surprising that for the problem of uniform approximation by polynomials, the best approximation is unique; i.e. there is a unique $p_* \in \mathcal{P}_n$ such that

$$\operatorname{dist}(f, \mathcal{P}_n) = \max_{t\in[a,b]} |f(t) - p_*(t)|.$$

We shall return to this problem (for continuous real-valued f) in §7.6.

7.6. EXERCISE (Landau polynomials). Define

$$Q_n(t) = c_n(1-t^2)^n,$$

where $c_n = \left(\int_{-1}^1 (1-s^2)^n ds\right)^{-1}$ so that $\int_{-1}^1 Q_n(t)\,dt = 1$. Let $f \in C([-\frac{1}{2}, \frac{1}{2}])$. The sequence of *Landau polynomials* associated with f is defined by

$$L_n f(t) = \int_{-1/2}^{1/2} f(s) Q_n(t-s)\,ds.$$

(i) Verify that $L_n f$ is a polynomial of degree at most $2n$.
(ii) Let $\gamma \in (0, \frac{1}{2})$ and $I_\gamma = [-\frac{1}{2}+\gamma, \frac{1}{2}-\gamma]$. Prove that $L_n f$ converges to f uniformly on I_γ.
(iii) Use part (ii) to give another proof of the Weierstrass approximation theorem.

7.7. EXERCISE. Let $f : [0,1] \to \mathbb{R}$ be in $C^1([0,1])$ and let $A = \sup_{t\in[0,1]} |f'(t)|$. Let $(B_n f)$ be the sequence of Bernstein polynomials defined in (7.1). Show that there is a constant C (independent of f) such that for all $n \geqslant 1$,

$$\max_{t\in[0,1]} |f(t) - B_n f(t)| \leqslant CAn^{-1/2}.$$

7.2. The Stone-Weierstrass theorem

Many other classes of functions with an algebraic structure similar to polynomials can be used for the approximation of continuous functions. This is the subject of a substantial generalization of the Weierstrass approximation theorem due to Stone. The theorem concerns subsets of the space of continuous functions $C(K;\mathbb{K})$ on a compact metric space K taking values in $\mathbb{K}$, where $\mathbb{K}$ denotes either $\mathbb{R}$ or $\mathbb{C}$.

7.8. DEFINITION (Algebras of functions). A set $\mathcal{A}$ of functions $f : K \to \mathbb{K}$ is a *(sub)algebra* if for all $f, g \in \mathcal{A}$ and all $c \in \mathbb{K}$,

$$f + g \in \mathcal{A},\ cf \in \mathcal{A},\ f \cdot g \in \mathcal{A}.$$

7.9. DEFINITION. Let $\mathcal{A}$ be a set of (real- or complex-valued) functions on a set K.

(i) $\mathcal{A}$ *separates points* if for every pair (x, y) of points in K with $x \neq y$ there exists $f \in \mathcal{A}$ such that $f(x) \neq f(y)$.
(ii) $\mathcal{A}$ *vanishes nowhere* if for every $x \in K$ there exists $f \in \mathcal{A}$ such that $f(x) \neq 0$.

Of course, for these definitions it is not required that K is compact. However, it is an essential prerequisite for the following theorem.

7.10. THEOREM (Stone–Weierstrass for real-valued functions). *Let K be a compact metric space and $\mathcal{A} \subset C(K;\mathbb{R})$. Assume that $\mathcal{A}$ is a real algebra that separates points and vanishes nowhere. Then $\mathcal{A}$ is dense in $C(K;\mathbb{R})$ (that is, $\overline{\mathcal{A}} = C(K;\mathbb{R})$).*

7.11. EXERCISE. Let K be a compact metric space. Show that if a subset $\mathcal{A} \subset C(K;\mathbb{R})$ does not separate points then $\mathcal{A}$ is not dense. Also if $\mathcal{A}$ does not satisfy the condition of vanishing nowhere then $\mathcal{A}$ is not dense.

7.12. Exercise. Let $\mathcal{A}$ be the set of all polynomials with real coefficients, considered as functions on $[1,2]$. Show that $\mathcal{A}$ is dense in $C([1,2];\mathbb{R})$, but not an algebra.

For the proof of the Stone-Weierstrass theorem we first establish some preliminary lemmas.

7.13. Lemma. *For every $a>0$ there exists a sequence of polynomials (p_n) with real coefficients such that $p_n(0)=0$ for all n and $\sup_{y\in[-a,a]}\big|p_n(y)-|y|\big|\to 0$ as $n\to\infty$.*

For $a=1$ an example of such a sequence has already been given in §4.3.1. For general $a>0$ take $p_n(x)=aP_n(x/a)$ where P_n is as in Example 4.84. Alternatively, one can also apply Weierstrass' Theorem 7.3 to first produce a sequence of polynomials q_n that converges uniformly to $f(y)=|y|$ on $[-a,a]$. In particular $q_n(0)\to 0$. Then set $p_n(y)=q_n(y)-q_n(0)$ and the sequence (p_n) satisfies the required properties.

In the following lemmas let $\mathcal{A}$ be a subalgebra of $C(K;\mathbb{R})$ that separates points and vanishes nowhere. Observe that the closure $\overline{\mathcal{A}}$ also is a real subalgebra of $C(K;\mathbb{R})$ that separates points and vanishes nowhere.

7.14. Lemma. *If $f\in\overline{\mathcal{A}}$, then $|f|\in\overline{\mathcal{A}}$.*

Proof. Let $\varepsilon>0$ and $a=\max_{x\in K}|f(x)|$ (which exists and is finite, by compactness of K). By Lemma 7.13 there exist $N\in\mathbb{N}$ and $c_1,\dots,c_N\in\mathbb{R}$ such that

$$\sup_{y\in[-a,a]}\Big|\sum_{j=1}^{N}c_jy^j-|y|\Big|\leqslant\varepsilon.$$

Set

$$g(x)=\sum_{j=1}^{N}c_j[f(x)]^j.$$

Since $\overline{\mathcal{A}}$ is an algebra, we see that $g\in\overline{\mathcal{A}}$ and we have

$$\sup_{x\in K}\big|g(x)-|f(x)|\big|\leqslant\varepsilon.$$

Thus, $|f|$ can be uniformly approximated by functions in $\overline{\mathcal{A}}$. But $\overline{\mathcal{A}}$ is closed, so $|f|\in\overline{\mathcal{A}}$. □

7.15. Lemma. *If $f_1,\dots,f_m\in\overline{\mathcal{A}}$, then $\min(f_1,\dots,f_m)\in\overline{\mathcal{A}}$ and $\max(f_1,\dots,f_m)\in\overline{\mathcal{A}}$.*

Proof. It suffices to show the claim for $m=2$ (the general case then follows by induction). Let $f,g\in\overline{\mathcal{A}}$. We have

$$\min(f,g)=\frac{f+g}{2}-\frac{|f-g|}{2},\ \max(f,g)=\frac{f+g}{2}+\frac{|f-g|}{2}.$$

Since $\overline{\mathcal{A}}$ is an algebra, Lemma 7.14 implies that $\min(f,g),\max(f,g)\in\overline{\mathcal{A}}$. □

7.16. Lemma. *For every $x_0,x_1\in K$, $x_0\neq x_1$ and $c_0,c_1\in\mathbb{R}$ there exists $f\in\overline{\mathcal{A}}$ such that $f(x_i)=c_i$ for $i=0,1$.*

In other words, any two points in $K\times\mathbb{R}$ that could lie on the graph of a function in $\overline{\mathcal{A}}$ do lie on the graph of a function in $\overline{\mathcal{A}}$.

Proof. Since $\overline{\mathcal{A}}$ separates points and vanishes nowhere there exist $g,h_0,h_1\in\overline{\mathcal{A}}$ such that $g(x_0)\neq g(x_1)$ and $h_i(x_i)\neq 0$ for $i=0,1$. Set

$$u_0(x)=g(x)h_0(x)-g(x_1)h_0(x),$$
$$u_1(x)=g(x)h_1(x)-g(x_0)h_1(x)$$

Then $u_0(x_0)\neq 0$, $u_0(x_1)=0$ and $u_1(x_1)\neq 0$, $u_1(x_0)=0$. Now let

$$f(x)=\tfrac{c_0}{u_0(x_0)}u_0(x)+\tfrac{c_1}{u_1(x_1)}u_1(x).$$

Then $f(x_0) = c_0$ and $f(x_1) = c_1$ and, since $\overline{\mathcal{A}}$ is an algebra, we get $f \in \overline{\mathcal{A}}$. □

This lemma can be seen as a toy model of the full theorem: the statement can be generalized to finitely many points. So we can use this generalization to find a function in $\overline{\mathcal{A}}$ that matches a given function f in any given collection of finitely many points (see Exercise 7.151). Thus, if K were finite, we would already be done. If K is not finite, we need to exploit compactness. Let us now get to the details.

PROOF OF THEOREM 7.10, CONCLUSION. Fix $f \in C(K)$ and let $\varepsilon > 0$. We first establish the following:

Claim: For every $x \in K$ there exists $g_x \in \overline{\mathcal{A}}$ such that $g_x(x) = f(x)$ and $g_x(t) > f(t) - \varepsilon$ for $t \in K$.

To see this, let $y \in K$. By Lemma 7.16 there exists $h_y \in \overline{\mathcal{A}}$ such that $h_y(x) = f(x)$ and $h_y(y) = f(y)$. By continuity of h_y there exists an open ball B_y around y such that $|h_y(t) - f(t)| < \varepsilon$ for all $t \in B_y$. In particular,

$$h_y(t) > f(t) - \varepsilon$$

for all $t \in B_y$. Observe that $(B_y)_{y \in K}$ is an open cover of K. Since K is compact, we can find a finite subcover by $B_{y_1}, \dots, B_{y_m}$. Set

$$g_x = \max(h_{y_1}, \dots, h_{y_m})$$

and we get $g_x(t) > f(t) - \varepsilon$ for all $t \in K$. By Lemma 7.15, $g_x \in \overline{\mathcal{A}}$, and the claim is proved.

By continuity of g_x there exists an open ball U_x such that

$$|g_x(t) - f(t)| < \varepsilon$$

for $t \in U_x$. In particular,

$$g_x(t) < f(t) + \varepsilon.$$

$(U_x)_{x \in K}$ is an open cover of K that has a finite subcover $U_{x_1}, \dots, U_{x_n}$. Then let

$$h = \min(g_{x_1}, \dots, g_{x_n}).$$

By Lemma 7.15 we have $h \in \overline{\mathcal{A}}$. Also,

$$f(t) - \varepsilon < h(t) < f(t) + \varepsilon$$

for all $t \in K$. That is, $|f(t) - h(t)| < \varepsilon$ for all $t \in K$. Since $\varepsilon > 0$ was arbitrary this proves $f \in \overline{\mathcal{A}}$. □

For complex-valued functions the theorem requires one additional assumption.

7.17. THEOREM (Stone–Weierstrass for complex-valued functions). *Let K be a compact metric space and let $\mathcal{A} \subset C(K)$ be a complex algebra that separates points and vanishes nowhere. In addition, assume that $\mathcal{A}$ is closed under complex conjugation, that is $\overline{f} \in \mathcal{A}$ for all $f \in \mathcal{A}$. Then $\mathcal{A}$ is dense in $C(K)$ (that is, $\overline{\mathcal{A}} = C(K)$).*

PROOF. Let $\mathcal{A}^{\mathbb{R}} \subset \mathcal{A}$ be the subclass of real-valued functions in $\mathcal{A}$. $\mathcal{A}^{\mathbb{R}}$ is a real algebra of functions.

Note that $\operatorname{Re}(f) = \frac{1}{2}(f + \overline{f})$ and $\operatorname{Im}(f) = \frac{1}{2i}(f - \overline{f})$. Thus, by the assumption that $\mathcal{A}$ is closed under complex conjugation, $\mathcal{A}$ contains for every $f \in \mathcal{A}$ also its real and imaginary parts, and these then belong by definition to $\mathcal{A}^{\mathbb{R}}$. $\mathcal{A}^{\mathbb{R}}$ is a real subalgebra of $C(K, \mathbb{R})$. The assumption that $\mathcal{A}$ separates points implies that $\mathcal{A}^{\mathbb{R}}$ separates points (if $f(x_1) \neq f(x_2)$, then either $\operatorname{Re} f(x_1) \neq \operatorname{Re} f(x_2)$ or $\operatorname{Im} f(x_1) \neq \operatorname{Im} f(x_2)$). Similarly, the assumption that $\mathcal{A}$ vanishes nowhere implies that $\mathcal{A}^{\mathbb{R}}$ vanishes nowhere.

If $f \in C(K)$, then by the real version of Stone-Weierstrass (Theorem 7.10) there are sequences $(u_n)_{n \in \mathbb{N}}$, $(v_n)_{n \in \mathbb{N}}$ of real-valued functions in $\mathcal{A}^{\mathbb{R}}$ such that $u_n \to \operatorname{Re}(f)$ and $v_n \to \operatorname{Im} f$, uniformly on K. Then $(u_n + iv_n)$ is a complex-valued sequence of functions in $\mathcal{A}$ that converges uniformly to f. This implies $\overline{\mathcal{A}} = C(K)$. □

The additional assumption that $\mathcal{A}$ is closed under complex conjugates is crucial for the complex version, as the following example shows.

7.18. Example. Let K be any compact subset of the complex plane $\mathbb{C}$ such that $0 \notin K$ and K contains the unit circle $\{z : |z| = 1\}$. Let $\mathcal{A}$ be the set of polynomials with complex coefficients, as functions $\mathbb{C} \to \mathbb{C}$. Then $\mathcal{A}$ is a subalgebra of $C(K;\mathbb{C})$ that separates points and vanishes nowhere on K. Also, $\mathcal{A}$ is not closed under complex conjugation. We claim that $\mathcal{A}$ is not dense in $C(K;\mathbb{C})$. To see this, consider the linear functional $\Lambda : C(K;\mathbb{C}) \to \mathbb{C}$ defined by

$$\Lambda(f) = \int_0^{2\pi} f(e^{i\theta})e^{i\theta}d\theta.$$

Then $|\Lambda(f)| \leqslant 2\pi \sup_K |f|$, so Λ is a continuous linear functional. For $g_n(z) = z^n$, compute

$$\Lambda(g_n) = \int_0^{2\pi} e^{i(n+1)\theta}d\theta = 0, \quad n = 0, 1, 2, \dots$$

and, by linearity, this implies that $\Lambda(g) = 0$ for all $g \in \mathcal{A}$. By continuity of Λ also $\Lambda(g) = 0$ for all $g \in \overline{\mathcal{A}}$. However, if $h(z) = z^{-1}$ then $h \in C(K;\mathbb{C})$ (since $0 \notin K$) and we have $\Lambda(h) = 2\pi$, hence $h \notin \overline{\mathcal{A}}$.

7.19. Example (Trigonometric polynomials). A *trigonometric polynomial* is a 1-periodic function that takes the form

$$g(x) = \sum_{k=-N}^{N} c_k e^{2\pi ikx}$$

for some $N \in \mathbb{N}_0$ and $c_k \in \mathbb{C}$. We may view 1-periodic functions as functions on the torus group $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, which is a compact metric space (see Exercise 4.165). By the properties of the exponential function, the set of trigonometric polynomials forms a complex algebra that is closed under complex conjugation. Therefore we conclude from the Stone-Weierstrass theorem that trigonometric polynomials are dense in the space of continuous 1-periodic functions, $C(\mathbb{T})$. In other words, continuous 1-periodic functions can be uniformly approximated by trigonometric polynomials.

7.20. Example. A *Laurent polynomial* is a function $q : \mathbb{C} \to \mathbb{C}$ of the form

$$q(z) = \sum_{k=-N}^{N} c_k z^k$$

with $N \in \mathbb{N}_0, c_k \in \mathbb{C}$. When restricted to the unit circle $S^1 \subset \mathbb{C}$, Laurent polynomials are in one-to-one correspondence with trigonometric polynomials via the change of variables $z = e^{2\pi ix}$. In particular, Laurent polynomials form a dense subset of $C(S^1)$ (one can show this again from Stone-Weierstrass, or deduce it from the result for trigonometric polynomials).

7.21. Example. Consider $K = [0, \frac{1}{2}]$ and let $\mathcal{A}$ be the vector space of *cosine polynomials* of the form

$$x \mapsto \sum_{k=0}^{N} c_k \cos(2\pi kx).$$

Using the identity $\cos a \cos b = \cos(a + b) + \cos(a - b)$ we see that $\mathcal{A}$ is an algebra. The reader should check that the assumptions of the Stone-Weierstrass theorem are satisfied, and we see that $\mathcal{A}$ is dense in $C([0, \frac{1}{2}])$. (*Exercise:* Why can we not replace $[0, \frac{1}{2}]$ by $[0, 1]$ in this argument?)

Remark. A function in $C([0, \frac{1}{2}])$ can be uniquely extended to an even 1-periodic continuous function on $\mathbb{R}$ (and vice versa). Thus, Example 7.21 shows that even 1-periodic continuous functions can be uniformly approximated by cosine polynomials.

7.3. Orthogonality

In this section we will see that something remarkable happens when the norm is generated by an inner product: at least in finite dimensional subspaces, best approximations are always unique and there is a straightforward method to compute them, namely by *orthogonal projection*.

Throughout, let X be an inner product space. The inner product is denoted by $\langle \cdot, \cdot \rangle$ and the seminorm by $\|f\| = \langle f, f\rangle^{1/2}$.

7.22. Example. The main use case we have in mind is an inner product space of square-integrable functions, say on a bounded interval I, with the inner product

$$\langle f, g\rangle = \int_I f(x)\overline{g(x)}\,dx \tag{7.3}$$

and the corresponding L^2 *seminorm*

$$\|f\|_2 = \Big(\int_I |f(x)|^2 dx\Big)^{1/2}.$$

Note this is generally only a seminorm, since $\|f\|_2 = 0$ does not necessarily imply $f = 0$.

7.3.1. Orthogonal projection. Let $V \subset X$ be a finite dimensional subspace and $N = \dim V$. We are assuming throughout that the restriction of the inner product to V is *definite on* V (i.e. if $g \in V$ and $\|g\| = 0$ then $g = 0$). We are interested in computing for arbitrary $f \in X$, the best approximation error

$$\operatorname{dist}(f, V) = \inf_{g\in V} \|f - g\|$$

and in finding a (hopefully unique) best approximation $g_* \in V$.

To achieve this we will work with an *orthonormal basis* of V, i.e. a basis $\{\phi_1, \dots, \phi_N\}$ of V satisfying

$$\langle \phi_k, \phi_\ell\rangle = \begin{cases} 0 & \text{if } k \neq \ell \\ 1 & \text{if } k = \ell. \end{cases}$$

We recall from linear algebra that given an arbitrary basis $\{u_1, \dots, u_N\}$ of V, an orthonormal basis can always be constructed using the *Gram-Schmidt orthogonalization process.*

7.23. Proposition (Gram-Schmidt). *Let $\{u_1, \dots, u_N\}$ be a basis of V. There is an orthonormal basis $\{\phi_1, \dots, \phi_N\}$ of V that satisfies*

$$\operatorname{span}\{u_1, \dots, u_k\} = \operatorname{span}\{\phi_1, \dots, \phi_k\}, \qquad k = 1, \dots, N. \tag{7.4}$$

This basis can be inductively constructed as follows. Let $\phi_1 = u_1/\|u_1\|$, and for $m \geqslant 2$

$$\phi_m = \frac{v_m}{\|v_m\|} \text{ where } v_m = u_m - \sum_{j=1}^{m-1} \langle u_m, \phi_j\rangle \phi_j.$$

(Orthonormality of the new basis follows directly from its inductive construction.)

From now on let us fix an orthonormal basis $\{\phi_1, \dots, \phi_N\}$. A powerful advantage of orthonormal bases is that coefficients can be easily computed by taking inner products:

7.24. Proposition. *Let $g \in V$ be given by $g = \sum_{j=1}^N b_j\phi_j$ with $b_j \in \mathbb{K}$. Then for all $j = 1, \dots, N$*

$$b_j = \langle g, \phi_j\rangle.$$

Proof. By linearity of the inner product and the definition of an orthonormal basis,

$$\langle g, \phi_j\rangle = \langle \sum_{k=1}^N b_k\phi_k, \phi_j\rangle = \sum_{k=1}^N b_k\langle \phi_k, \phi_j\rangle = b_j.$$

□

The same calculation as in this short proof leads to the best approximation principle, and the definition of orthogonal projections:

7.25. Theorem (Unique best approximation). *For every $f \in X$ there is a unique best approximation to f in V, given by*

$$g_* = \sum_{j=1}^{N} \langle f, \phi_j \rangle \phi_j.$$

That is, for every $g \in V$, $\|f - g_*\| \leqslant \|f - g\|$ and equality holds if and only if $g = g_*$. Moreover,

$$\operatorname{dist}(f, V)^2 = \|f - g_*\|^2 = \|f\|^2 - \sum_{j=1}^{N} |\langle f, \phi_j \rangle|^2 \tag{7.5}$$

and for every $g \in V$ with $g = \sum_{j=1}^{N} b_j \phi_j$,

$$\|f - g\|^2 = \|f - g_*\|^2 + \sum_{j=1}^{N} |\langle f, \phi_j \rangle - b_j|^2. \tag{7.6}$$

7.26. Definition (Orthogonal projection). The map $P : X \to V$ assigning to each $f \in X$ its best approximation $g_* \in V$ is called *orthogonal projection*.

Orthogonal projection is a linear map, since it is given by

$$Pf = \sum_{j=1}^{N} \langle f, \phi_j \rangle \phi_j.$$

Remark. Observe that by the uniqueness claim in Theorem 7.25, Pf does not depend on the choice of the orthonormal basis $\{\phi_j\}$.

This can be interpreted geometrically by analogy to the classical Euclidean space $\mathbb{R}^n$: the orthogonal projection Pf is characterized by being the point in V that is closest to f and it is uniquely determined by this property (see Figure 7.1).

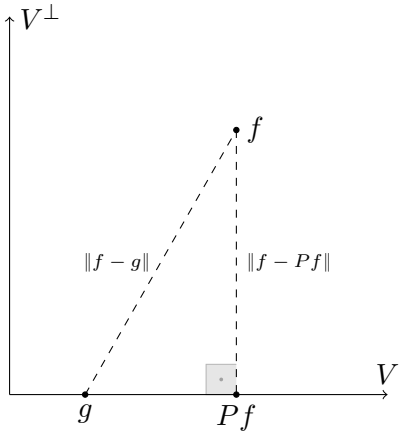


Figure 7.1. Pf is the orthogonal projection of f onto V.

Proof of Theorem 7.25. It suffices to check the identity (7.6). Indeed, this implies $\|f-g\| \geqslant \|f-g_*\|$ and that equality holds if and only if $g = g_*$. Then $\operatorname{dist}(f, V) = \|f - g_*\|$ and the second equality in (7.5) follows by substituting $g = 0$ in (7.6).

For $g \in V$ we calculate

$$\|f-g\|^2 = \langle f-g, f-g\rangle = \langle f,f\rangle - \overline{\langle g,f\rangle} - \langle g,f\rangle + \langle g,g\rangle.$$

Letting $g = \sum_{j=1}^N b_j\phi_j$ and $c_j = \langle f,\phi_j\rangle$ we compute using linearity of the inner product, and the definition of orthonormal bases:

$$\langle g,f\rangle = \sum_{j=1}^N b_j\langle \phi_j, f\rangle = \sum_{j=1}^N b_j\overline{c_j}.$$

Similarly,

$$\langle g,g\rangle = \sum_{j=1}^N\sum_{k=1}^N b_j\overline{b_k}\langle\phi_j,\phi_k\rangle = \sum_{j=1}^N |b_j|^2.$$

Thus,

$$\|f-g\|^2 = \langle f,f\rangle - \sum_{j=1}^N b_j\overline{c_j} - \sum_{j=1}^N \overline{b_j}c_j + \sum_{j=1}^N |b_j|^2$$

$$= \langle f,f\rangle - \sum_{j=1}^N |c_j|^2 + \sum_{j=1}^N |b_j-c_j|^2. \tag{7.7}$$

Applying this for $g = g_* = \sum_{j=1}^N c_j\phi_j$ (so $b_j = c_j$) gives

$$\|f-g_*\|^2 = \langle f,f\rangle - \sum_{j=1}^N |c_j|^2.$$

Plugging this into (7.7) gives (7.6). □

7.3.1.1. *Gram determinants.* While Theorem 7.25 makes computation of the approximation error straightforward if we know an orthonormal basis of V, it is not always convenient to construct an orthonormal basis. The Gram-Schmidt process is computationally expensive and may not always give good insight into a particular problem under consideration. We now consider the situation where an arbitrary basis $\{u_1,\dots,u_N\}$ of V is given and seek to express $\operatorname{dist}(f,V)$ and the best approximation in terms of these basis vectors.

7.27. Definition (Gram determinant). Let $m \in \mathbb{N}$. The *Gram determinant* associated with $f_1,\dots,f_m \in X$ is defined by

$$\mathcal{G}(f_1,\dots,f_m) = \det\begin{pmatrix} \langle f_1,f_1\rangle & \langle f_1,f_2\rangle & \dots & \langle f_1,f_m\rangle \\ \langle f_2,f_1\rangle & \langle f_2,f_2\rangle & \dots & \langle f_2,f_m\rangle \\ \vdots & \vdots & & \vdots \\ \langle f_m,f_1\rangle & \langle f_m,f_2\rangle & \dots & \langle f_m,f_m\rangle \end{pmatrix}.$$

We start with the following linear algebra fact.

7.28. Lemma. *Let $\{\phi_1,\dots,\phi_N\}$ be an orthonormal basis of $V \subset X$ and let $u_1,\dots,u_N$ be vectors in V. If $u_j = \sum_{i=1}^N a_{ij}\phi_i$, $j = 1,\dots N$ and if $A = (a_{ij}) \in \mathbb{K}^{N\times N}$, then*

$$\mathcal{G}(u_1,\dots,u_N) = \det(\overline{A}^{\mathsf{T}}A)$$

In particular, if $u_1,\dots,u_N$ are linearly independent, then $\mathcal{G}(u_1,\dots,u_N) > 0$.

Proof. We have

$$\langle u_k,u_l\rangle = \sum_{i=1}^N\sum_{j=1}^N a_{ik}\overline{a_{jl}}\langle\phi_i,\phi_j\rangle = \sum_{i=1}^N \overline{a_{il}}a_{ik} = (\overline{A}^{\mathsf{T}}A)_{lk}$$

and consequently $\mathcal{G}(u_1,\dots,u_N) = \det(\overline{A}^{\mathsf{T}}A)$. If $u_1,\dots,u_N$ are linearly independent, then A is invertible, and so is $\overline{A}^{\mathsf{T}}A$. The latter is then positive definite, and thus the Gram determinant is real and positive in this case. □

7.29. THEOREM. *Let $\{u_1,\dots,u_N\}$ be a basis of $V \subset X$. Then the following hold for each $f \in X$:*

(i)

$$\operatorname{dist}(f,V) = \sqrt{\frac{\mathcal{G}(u_1,\dots,u_N,f)}{\mathcal{G}(u_1,\dots,u_N)}}.$$

(ii) Let $A^{(j)}[f]$ be the $N \times N$ matrix with entries $A^{(j)}[f]_{ik} = \langle u_k, u_i\rangle$ for $k \neq j$ and $A^{(j)}[f]_{ij} = \langle f, u_i\rangle$. Then the best approximation to f in V is given by

$$Pf = \sum_{j=1}^{N} \beta_j u_j \text{ with } \beta_j = \frac{\det(A^{(j)}[f])}{\mathcal{G}(u_1,\dots,u_N)}.$$

PROOF. By the Gram-Schmidt procedure, V has an orthonormal basis $\{\phi_1,\dots,\phi_N\}$ and by Theorem 7.25 we have $\operatorname{dist}(f,V) = \|f - Pf\|$ where $Pf = \sum_{k=1}^{N}\langle f,\phi_k\rangle\phi_k$ is the best approximation to f in V. Expand $Pf = \sum_{j=1}^{N}\beta_j u_j$ with coefficients $\beta_j \in \mathbb{K}$, $j = 1,\dots,N$. Using $0 = \langle f - Pf, u_k\rangle = \langle f, u_k\rangle - \langle Pf, u_k\rangle$ for $k = 1,\dots,N$ we see

$$\sum_{j=1}^{N}\langle u_j, u_k\rangle\beta_j = \langle f, u_k\rangle, \quad k = 1,\dots,N. \tag{7.8}$$

Now let $d = \operatorname{dist}(f,V) = \|f - Pf\|$ and using again that $\langle f - Pf, Pf\rangle = 0$ we get

$$d^2 = \langle f - Pf, f - Pf\rangle = \langle f - Pf, f\rangle = \langle f, f\rangle - \sum_{j=1}^{N}\beta_j\langle u_j, f\rangle.$$

That is,

$$\sum_{j=1}^{N}\langle u_j, f\rangle\beta_j + d^2 = \langle f, f\rangle. \tag{7.9}$$

From (7.8), (7.9) we see that $(\beta_1,\dots,\beta_N,d^2)$ is a solution vector for the system of $N+1$ linear equations

$$\begin{aligned}
\langle u_1,u_1\rangle\beta_1 &+ \langle u_2,u_1\rangle\beta_2 + \dots + \langle u_N,u_1\rangle\beta_N &&= \langle f,u_1\rangle,\\
\langle u_1,u_2\rangle\beta_1 &+ \langle u_2,u_2\rangle\beta_2 + \dots + \langle u_N,u_2\rangle\beta_N &&= \langle f,u_2\rangle,\\
&\vdots\\
\langle u_1,u_N\rangle\beta_1 &+ \langle u_2,u_N\rangle\beta_2 + \dots + \langle u_N,u_N\rangle\beta_N &&= \langle f,u_N\rangle,\\
\langle u_1,f\rangle\beta_1 &+ \langle u_2,f\rangle\beta_2 + \dots + \langle u_N,f\rangle\beta_N &+ d^2 &= \langle f,f\rangle.
\end{aligned}$$

The determinant of the coefficient matrix of this linear system is equal to $\mathcal{G}(u_1,\dots,u_N)$, which is positive. Thus the solution vector of the linear system is unique and we have explicit formulas for it by Cramer's rule. The last component of this solution vector is equal to

$$d^2 = \frac{\mathcal{G}(u_1,\dots,u_N,f)}{\mathcal{G}(u_1,\dots,u_N)}.$$

We apply Cramer's rule for the other components and use that the last column of the coefficient matrix has zeros above the lowest entry. An expansion leads to the formula $\beta_j = \det(A^{(j)}[f])/\mathcal{G}(u_1,\dots,u_N)$, for $j = 1,\dots,N$. □

REMARK. The formula for $\operatorname{dist}(f,V)$ has a geometric interpretation. In Euclidean space the quantity $\sqrt{\mathcal{G}(u_1,\dots,u_N)}$ represents the N-dimensional volume of the parallelepiped determined by $u_1,\dots,u_N$. The quantity $\operatorname{dist}(f,V)$ is a height in the $N+1$-dimensional parallelepiped determined by $u_1,\dots,u_N,f$.

7.3.2. Orthonormal systems. We have already used orthonormal systems in a finite dimensional inner product space, in the approximation result of Theorem 7.25. We now extend the definition to systems of arbitrary cardinality, indexed by an index set $\mathcal{I}$ (usually countable).

7.30. DEFINITION. A family of vectors $(\phi_\alpha)_{\alpha\in\mathcal{I}}$ in X is called an *orthonormal system* if

$$\langle \phi_\alpha, \phi_\beta\rangle = \begin{cases} 0, & \text{if } \alpha \neq \beta, \\ 1, & \text{if } \alpha = \beta. \end{cases}$$

One sometimes relaxes the normalization assumption and considers *orthogonal systems* $(\phi_\alpha)_{\alpha\in\mathcal{I}}$ where it is assumed that $\|\phi_\alpha\| \neq 0$ for all α and $\langle \phi_\alpha, \phi_\beta\rangle = 0$ whenever $\alpha \neq \beta$. Of course, if $(\phi_\alpha)_{\alpha\in\mathcal{I}}$ is an orthogonal system then the normalized $\phi_\alpha/\|\phi_\alpha\|$ form an orthonormal system.

7.31. EXAMPLE (Disjoint support). Let $\phi_n(x) = \mathbb{1}_{[n,n+1)}$, $n = 0,\dots,N$. Then $(\phi_n)_{n=0}^N$ is an orthonormal system of functions defined on $[0,N)$ (with the inner product in (7.3)).

7.32. EXAMPLES (Trigonometric functions). The following are orthonormal systems of functions in $C(\mathbb{T})$ with the inner product in (7.3).

(i) $\phi_n(x) = e^{2\pi i n x}$, $n \in \mathbb{Z}$.

(ii) $\phi_n(x) = \sqrt{2}\cos(2\pi n x)$, $n = 1,2,\dots$.

(iii) $\phi_n(x) = \sqrt{2}\sin(2\pi n x)$, $n = 1,2,\dots$.

7.33. EXAMPLE. Let $M \in \mathbb{N}$ and denote $[M] = \{0,\dots,M-1\}$. Let X be the M-dimensional space of complex-valued functions on $[M]$. Define a scalar product on X by

$$\langle f, g\rangle = \frac{1}{M}\sum_{j\in[M]} f(j)\overline{g(j)}$$

and let

$$\phi_k(\ell) = e^{2\pi i k\ell/M}, \quad \ell \in [M].$$

Then one may check that the M functions $(\phi_k)_{k\in[M]}$ form an orthonormal system. Since X is M-dimensional, this system is an orthonormal basis.

7.34. EXERCISE (Rademacher functions). For $n = 0,1,\dots$ and $x \in [0,1)$ we define

$$\begin{aligned} r_0(x) &= 1 \\ r_n(x) &= \sum_{k=1}^{2^n} (-1)^{k-1}\mathbb{1}_{[\frac{k-1}{2^n},\frac{k}{2^n})}(x), \quad n = 1,2,\dots. \end{aligned} \tag{7.10}$$

Note that r_n is constant (with value 1 or -1) on dyadic intervals of length 2^{-n}, and r_n changes sign on two consecutive intervals of length 2^{-n}.

(i) Verify that for $x \in [0,1)$ such that $2^n x \notin \mathbb{Z}$ we have $r_n(x) = \operatorname{sgn}(\sin(2^n\pi x))$.

(ii) Show that $(r_n)_{n=0,1,2,\dots}$ is an orthonormal system of functions on $[0,1)$.

7.35. EXAMPLE (Orthogonal polynomials). Let w be a continuous positive function on $[a,b]$ such that $w(t) > 0$ for all $t \in [a,b]$. Given real valued polynomials q_n of exact degree $n \in \mathbb{N}_0$, that is, $q_n(t) = \sum_{k=0}^n a_{k,n}t^k$ with $a_{n,n} \neq 0$, we can apply the Gram-Schmidt orthogonalization procedure to (q_n), with the scalar product

$$\langle f, g\rangle_w = \int_a^b f(t)\overline{g(t)}w(t)dt.$$

This yields a sequence of orthogonal polynomials p_n, $n = 0, 1, 2, \dots$ of exact degree n, satisfying

$$\int_a^b p_n(t)p_m(t)w(t)\,dt \begin{cases} \neq 0, & \text{if } n = m \\ = 0, & \text{if } n \neq m \end{cases} \tag{7.11}$$

for all non-negative integers n, m. Let $c_n = \int_a^b |p_n(t)|^2 w(t)\,dt$ which is positive. Then the polynomials $c_n^{-1/2} p_n$ form an orthonormal system.

7.36. Examples (Orthogonal polynomials). Let $-\infty < a < b < \infty$. The condition on the weight in Example 7.35 can be relaxed: we may assume that w is continuous and positive on the open interval (a, b) and that the (possibly improper) integral $\int_a^b w(t)dt$ is finite.

An example on $(-1, 1)$ that frequently occurs in the literature is given by the inner product

$$\langle f, g\rangle_{\alpha,\beta} = \int_{-1}^{1} f(t)\overline{g(t)}\frac{dt}{(1-t)^\alpha(1+t)^\beta}$$

on $(-1, 1)$. Here we assume $\alpha < 1$, $\beta < 1$. The particularly useful special case with $\alpha = \beta = \frac{1}{2}$ will play a role in §7.7.

On $(-\infty, \infty)$ useful choices of weight functions are rapidly decreasing as $|x| \to \infty$. An example is $w(x) = e^{-x^2}$, with the corresponding inner product given by $\langle f, g\rangle_w = \int_{-\infty}^{\infty} f(x)\overline{g(x)}e^{-x^2}dx$. Then $\int_{-\infty}^{\infty} |p(x)|^2 w(x)dx < \infty$ for all polynomials. By the Gram-Schmidt procedure one can construct a sequence p_n of polynomials which is orthonormal with respect to $\langle \cdot, \cdot\rangle_w$.

Our first result about orthonormal systems is a consequence of Theorem 7.25.

7.37. Theorem (Bessel's inequality). *Let $(\phi_n)_{n=1,2,\dots}$ be an orthonormal system. Then*

$$\sum_{n=1}^{\infty} |\langle f, \phi_n\rangle|^2 \leqslant \|f\|^2. \tag{7.12}$$

Proof. Applying Theorem 7.25 to the vector spaces $V_N = \operatorname{span}(\phi_1, \dots, \phi_N)$ yields

$$\sum_{k=1}^{N} |\langle f, \phi_k\rangle|^2 \leqslant \|f\|^2$$

for all $N \in \mathbb{N}$, and letting $N \to \infty$ we obtain (7.12). □

Bessel's inequality can be extended to index sets that are not necessarily countable:

7.38. Theorem. *Let $(\phi_\alpha)_{\alpha\in\mathcal{I}}$ be an orthonormal system. Then*

$$\mathcal{S}(f) = \{\alpha \in \mathcal{I} : \langle f, \phi_\alpha\rangle \neq 0\} \tag{7.13}$$

is countable.

Omitting all zero terms in the α-sum we define $\sum_{\alpha\in\mathcal{I}} |\langle f, \phi_\alpha\rangle|^2 = \sum_{\alpha\in\mathcal{S}(f)} |\langle f, \phi_\alpha\rangle|^2$. By a combination of Theorems 7.37 and 7.38 we then get Bessel's inequality for arbitrary orthonormal systems,

$$\sum_{\alpha\in\mathcal{I}} |\langle f, \phi_\alpha\rangle|^2 \leqslant \|f\|^2. \tag{7.14}$$

Proof of Theorem 7.38. Let $E \subset \mathcal{I}$ with $\#E < \infty$. Applying Theorem 7.25 to $\operatorname{span}\{\phi_\alpha : \alpha \in E\}$ we get $\sum_{\alpha\in\mathcal{I}\cap E} |\langle f, \phi_\alpha\rangle|^2 \leqslant \|f\|^2$. In particular, if $E(b) = \{\alpha : |\langle f, \phi_\alpha\rangle| > b\}$ the inequality implies

$$b^2 \#(E(b) \cap E) \leqslant \|f\|^2$$

for all finite sets E and all $b > 0$. Taking the supremum over finite sets E we obtain the inequality $\#(E(b)) \leqslant b^{-2}\|f\|^2 < \infty$. Since $\mathcal{S}(f) = \bigcup_{n\in\mathbb{N}} E(n^{-1})$ we see that $\mathcal{S}(f)$ is countable. □

7.39. Corollary (Riemann–Lebesgue lemma). *Let $(\phi_\alpha)_{\alpha\in\mathcal{I}}$ be an orthonormal system. Assume that $f \in X$ and $\{\alpha : \langle f, \phi_\alpha\rangle \neq 0\}$ is an infinite set enumerated by $(\alpha_n)_{n=1}^\infty$. Then*

$$\lim_{n\to\infty} \langle f, \phi_{\alpha_n}\rangle = 0.$$

This is a consequence of $\sum_{n=1}^\infty |\langle f, \phi_{\alpha_n}\rangle|^2 < \infty$ which holds by Bessel's inequality.

In an inner product space X we would like to express every $f \in X$ in terms of the ϕ_α. This is only possible if the space of finite linear combinations of the ϕ_α is dense. This leads to the following definition.

7.40. Definition (Complete orthonormal system). An orthonormal system $(\phi_\alpha)_{\alpha\in\mathcal{I}}$ is called *complete* if the space of finite linear combinations of functions in (ϕ_α) is dense in X, i.e. for every $f \in X$ and every $\varepsilon > 0$ there exist $N \in \mathbb{N}$, $\alpha_1, \dots, \alpha_N \in \mathcal{I}$ and $c_1, \dots, c_N \in \mathbb{K}$ such that

$$\|f - \sum_{k=1}^N c_k \phi_{\alpha_k}\| < \varepsilon.$$

In an inner product space we can characterize completeness of orthonormal systems in various useful ways.

7.41. Theorem. *The following are equivalent.*

(i) The orthonormal system is complete in X (in the sense of Definition 7.40).

(ii) Parseval's identity

$$\|f\|^2 = \sum_{\alpha\in\mathcal{I}} |\langle f, \phi_\alpha\rangle|^2$$

holds for all $f \in X$.

(iii) For every $f \in X$ we have $\|f - \sum_{\alpha\in\mathcal{I}} \langle f, \phi_\alpha\rangle \phi_\alpha\| = 0$, with unconditional convergence.

Remarks. Condition (ii) is also equivalent to $\langle f, g\rangle = \sum_{\alpha\in\mathcal{I}} \langle f, \phi_\alpha\rangle \overline{\langle g, \phi_\alpha\rangle}$ for all $f, g \in X$, i.e. Parseval's equation for inner products. The equivalence with (ii) follows from the polarization identity in Proposition 5.71.

To elaborate on (iii), the index set $\mathcal{S}(f)$ for the nonzero Fourier coefficients (as in (7.13)) is countable (by Theorem 7.37) and unconditional convergence means: If A is a countable subset of $\mathcal{I}$ with $\mathcal{S}(f) \subset A$ and A is enumerated by $(\alpha_k)_{k=1}^\infty$ then $\|f - \sum_{k=1}^n \langle f, \phi_{\alpha_k}\rangle \phi_{\alpha_k}\|$ converges to 0 as $n \to \infty$. In a pre-Hilbert space (i.e. if the inner product has the positive-definite property), this means that $\sum_{k=1}^n \langle f, \phi_{\alpha_k}\rangle \phi_{\alpha_k}$ converges in X to the unique limit f, as $n \to \infty$.

Proof of Theorem 7.41. The proof follows from Theorem 7.25 if X is finite dimensional, so we may assume that X is infinite-dimensional and $\mathcal{I}$ is infinite. For (iii) $\Longrightarrow$ (i) there is nothing to show. It remains to prove (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii).

Proof of (i) $\Longrightarrow$ (ii). By Bessel's inequality $\|f\|^2 \geqslant \sum_{\alpha\in\mathcal{I}} |\langle f, \phi_\alpha\rangle|^2$. To prove the opposite inequality let $\varepsilon > 0$. By (i) there is a finite set $E \subset \mathcal{I}$ and $g = \sum_{\alpha\in E} b_\alpha \phi_\alpha$ such that $\|f - g\| < \varepsilon$. Let $V_E = \operatorname{span}(\{\phi_\alpha : \alpha \in E\})$. Using Theorem 7.25 we get

$$\|f\|^2 - \sum_{\alpha\in E} |\langle f, \phi_\alpha\rangle|^2 = \operatorname{dist}(f, V_E)^2 \leqslant \|f - g\|^2 < \varepsilon^2$$

and therefore

$$\sum_{\alpha\in\mathcal{I}} |\langle f, \phi_\alpha\rangle|^2 \geqslant \sum_{\alpha\in E} |\langle f, \phi_\alpha\rangle|^2 \geqslant \|f\|^2 - \varepsilon^2.$$

Since $\varepsilon > 0$ was arbitrary we get $\|f\|^2 \leqslant \sum_{\alpha\in\mathcal{I}} |\langle f, \phi_\alpha\rangle|^2$ which finishes the proof of the implication (i) $\Longrightarrow$ (ii).

Proof of (ii) $\Longrightarrow$ (iii). Let $f \in X$. By Theorem 7.38, $\mathcal{S}(f)$ is countable. Let A be an infinite countable subset of indices containing $\mathcal{S}(f)$ and let $(\alpha_k)_{k=1}^\infty$ be an enumeration of A. Given $\varepsilon > 0$

there exists, by (ii), an $N(\varepsilon) \in \mathbb{N}$ such that $0 \leqslant \|f\|^2 - \sum_{k=1}^{n} |\langle f, \phi_{\alpha_k}\rangle|^2 < \varepsilon^2$ for all $n \geqslant N(\varepsilon)$. Let $V_n = \operatorname{span}\{\phi_{\alpha_k} : 1 \leqslant k \leqslant n\}$. Then, by Theorem 7.25, we get for $n \geqslant N(\varepsilon)$

$$\Big\| f - \sum_{k=1}^{n} \langle f, \phi_{\alpha_k}\rangle \phi_{\alpha_k} \Big\|^2 = \operatorname{dist}(f, V_n)^2 = \|f\|^2 - \sum_{k=1}^{n} |\langle f, \phi_{\alpha_k}\rangle|^2 \leqslant \varepsilon^2,$$

and since f and $\varepsilon > 0$ were arbitrary this finishes the proof of the implication (ii) $\Longrightarrow$ (iii). □

7.3.2.1. *Consequences for Hilbert spaces.* In a Hilbert space we can add another characterization of completeness of orthonormal systems.

7.42. Theorem. *Let X be a Hilbert space. Then an orthonormal system $(\phi_\alpha)_{\alpha\in\mathcal{I}}$ is complete if and only if for every $f \in X$ we have the implication*

$$\langle f, \phi_\alpha\rangle = 0 \text{ for all } \alpha \in \mathcal{I} \quad \Longrightarrow \quad f = 0.$$

Proof. If the orthonormal system is complete, then Parseval's identity $\|f\|^2 = \sum_{\alpha\in\mathcal{I}} |\langle f, \phi_\alpha\rangle|^2$ holds by Theorem 7.41. Hence the hypothesis $\langle f, \phi_\alpha\rangle = 0$ for all $\alpha \in \mathcal{I}$ implies $\|f\| = 0$ and thus $f = 0$.

Vice versa assume that whenever $\langle g, \phi_\alpha\rangle = 0$ for all $\alpha \in \mathcal{I}$ then also $g = 0$. We prove completeness by proving the equivalent statement (iii) in Theorem 7.41. Fix $f \in X$. By Theorem 7.38 the set $\mathcal{S}(f) = \{\alpha : \langle f, \phi_\alpha\rangle \neq 0\}$ is at most countable. Let $A = \{\alpha_n : n \in \mathbb{N}\} \subset \mathcal{I}$ be an index set with $\mathcal{S}(f) \subset A$. By Bessel's inequality, $\sum_{k=1}^{\infty} |\langle f, \phi_{\alpha_k}\rangle|^2 < \infty$. Since by the Pythagorean theorem

$$\Big\| \sum_{k=m}^{n} \langle f, \phi_{\alpha_k}\rangle \phi_{\alpha_k} \Big\|^2 = \sum_{k=m}^{n} |\langle f, \phi_{\alpha_k}\rangle|^2$$

we see that the partial sums of the series $\sum_{k=1}^{\infty} \langle f, \phi_{\alpha_k}\rangle \phi_{\alpha_k}$ form a Cauchy sequence and since X is a Hilbert space the series converges. To verify that f is the limit consider

$$g = f - \sum_{k=1}^{\infty} \langle f, \phi_{\alpha_k}\rangle \phi_{\alpha_k}$$

and observe that for all $\alpha \in \mathcal{I}$

$$\langle g, \phi_\alpha\rangle = \langle f, \phi_\alpha\rangle - \sum_{k=1}^{\infty} \langle f, \phi_{\alpha_k}\rangle \langle \phi_{\alpha_k}, \phi_\alpha\rangle = \begin{cases} \langle f, \phi_\alpha\rangle - \langle f, \phi_\alpha\rangle = 0 & \text{if } \alpha \in \mathcal{S}(f), \\ \langle f, \phi_\alpha\rangle = 0 & \text{if } \alpha \notin \mathcal{S}(f), \end{cases}$$

i.e. $\langle g, \phi_\alpha\rangle = 0$ for all α, hence $g = 0$ by our assumption, and thus $f = \sum_{k=1}^{\infty} \langle f, \phi_{\alpha_k}\rangle \phi_{\alpha_k}$. □

7.43. Definition. A complete orthonormal system $\{\phi_\alpha : \alpha \in \mathcal{I}\}$ in a Hilbert space is called an *orthonormal basis* or *Hilbert basis.*

Remark. Note that the term (orthonormal) basis is used here in a sense different from the corresponding concept of an algebraic basis (also called *Hamel basis*) in linear algebra (*cf.* Example 8.35 in the next chapter).

7.44. Theorem. *Every Hilbert space contains an orthonormal basis.*

Proof. We use the axiom of choice in the form of Zorn's lemma (Theorem 1.8). The collection of orthonormal systems in the Hilbert space X is partially ordered by inclusion. Every totally ordered subset of orthonormal systems has an upper bound (take the union). Thus Zorn's lemma implies the existence of a maximal orthonormal system $\mathcal{O} = \{\phi_\alpha\}_{\alpha\in\mathcal{I}}$. We show that it is an orthonormal basis. If not then there exists by Theorem 7.42 an $f \in X$ such that $\langle f, \phi_\alpha\rangle = 0$ for all $\alpha \in \mathcal{I}$, and $f \neq 0$. But then $\mathcal{O} \cup \{f/\|f\|\}$ is an orthonormal system which properly contains $\mathcal{O}$, in contradiction to the maximality of $\mathcal{O}$. □

7.45. Exercise. Let $\mathcal{I}$ be an index set. Consider the space $\ell^2(\mathcal{I})$ of functions $f : \mathcal{I} \to \mathbb{C}$ such that $\|f\|_{\ell^2} = (\sum_{\alpha\in\mathcal{I}} |f(\alpha)|^2)^{1/2}$ is finite.

For $\alpha \in \mathcal{I}$ let $e_\alpha : \mathcal{I} \to \{0, 1\}$ be defined by

$$e_\alpha(\beta) = \begin{cases} 1 & \text{if } \beta = \alpha \\ 0 & \text{if } \beta \neq \alpha. \end{cases}$$

(i) Show that $\{e_\alpha\}_{\alpha\in\mathcal{I}}$ is an orthonormal basis of $\ell^2(\mathcal{I})$.

(ii) Let X be a Hilbert space and let $\{\phi_\alpha : \alpha \in \mathcal{I}\}$ be an orthonormal basis. Then the map

$$\imath : \begin{cases} X \to \ell^2(\mathcal{I}) \\ f \mapsto (\alpha \mapsto \langle f, \phi_\alpha\rangle) \end{cases}$$

assigning to each $f \in X$ its set of coefficients is an isometric isomorphism of Hilbert spaces.

7.46. Definition. A normed space is called *separable* if it has a countable dense subset.

7.47. Theorem. *Let X be a Hilbert space. The following conditions are equivalent.*

(i) X is separable.

(ii) X contains a countable orthonormal basis.

(iii) Every orthonormal basis of X is countable.

Proof. If X is finite-dimensional then every basis is finite and the linear combinations of basis vectors with coefficients in $\mathbb{Q} + i\mathbb{Q}$ form a countable dense subset.

In what follows we may assume that X is not finite-dimensional. If X is separable there is a countable dense set in X enumerated by $\{g_n\}_{n=1}^\infty$. There is an increasing function of indices $n \to k(n)$ so that $\dim \operatorname{span}\{g_1, \ldots, g_{k(n)}\} = n$ and $\dim\{g_1, \ldots, g_{k(n)-1}\} = n - 1$. Set $f_n = g_{k(n)}$ then $\{f_1, \ldots, f_n\}$ is a linearly independent set. Applying the Gram-Schmidt process to the f_n yields an orthonormal system $\{\phi_n\}$ with $\operatorname{span}(\{\phi_n : n \in \mathbb{N}\})$ dense in X and thus $\{\phi_n\}$ is an orthonormal basis of X.

Vice versa, let $\{\phi_n\}_{n\in\mathbb{N}}$ be a countable orthonormal basis of X. Then the finite linear combinations of basis vectors with coefficients in $\mathbb{Q} + i\mathbb{Q}$ form a countable dense set, so that X is separable.

Finally let X be separable, $G = \{g_n\}_{n\in\mathbb{N}}$ be a dense subset and $\{\phi_\alpha\}_{\alpha\in\mathcal{I}}$ be an orthonormal basis. For every α there is $n(\alpha) \in \mathbb{N}$ such that $\|g_{n(\alpha)} - \phi_\alpha\| < 1/2$. Since $\|\phi_\alpha - \phi_\beta\| = \sqrt{2}$ we see by the triangle inequality that $g_{n(\alpha)} \neq g_{n(\beta)}$ and since G is countable, so is $\mathcal{I}$. Thus every orthonormal basis is countable. □

To conclude this section we state an analog of the approximation theorem in 7.3 for a general closed (not necessarily finite-dimensional) subspace of a Hilbert space.

7.48. Theorem. *Let V be a closed subspace of a Hilbert space X and let*

$$V^\perp = \{h \in X : \langle g, h\rangle = 0 \text{ for all } g \in V\}.$$

Then the following holds.

(i) *Every $f \in X$ can be expressed uniquely as $f = g + h$ where $g \in V$ and $h \in V^\perp$.*

(ii) *Let for $f \in X$ denote by Pf the unique element in V such that $f - Pf \in V^\perp$. Then P defines a bounded linear operator on X, such that $P^2 = P$, $P = P^*$, range$(P) = V$ and $Ph = 0$ if and only if $h \in V^\perp$.*

(iii) *Pf is the unique element of V such that $\|f - Pf\| = \inf_{v\in V} \|f - v\|$. Moreover $\|Pf\| = \inf_{w\in V^\perp} \|f - w\|$.*

If $\{u_\alpha\}_{\alpha\in\mathcal{I}}$ is an orthonormal basis of V, then one can define $Pf = \sum_{\alpha\in\mathcal{I}} \langle f, u_\alpha\rangle u_\alpha$ (with unconditional convergence). The reader is invited to provide the details to complete the proof.

7.4. Trigonometric polynomials

This section is devoted to a particularly important example of an orthonormal system on $[0,1]$: the *trigonometric system* consists of the functions

$$e_k(x) = e^{2\pi ikx}, \quad k \in \mathbb{Z}.$$

Finite linear combinations of the functions (e_k) are called *trigonometric polynomials*, which we have already encountered in Example 7.19. Trigonometric polynomials are of the form

$$p(x) = \sum_{k=-N}^{N} c_k e^{2\pi ikx} \quad (x \in \mathbb{R}), \tag{7.15}$$

where $N \in \mathbb{N}$ and $c_k \in \mathbb{C}$. If c_N or c_{-N} is non-zero, then N is called the *degree* of p. The vector space of trigonometric polynomials of degree at most N will be denoted by $\mathcal{T}_N$, and the space of all trigonometric polynomials will be denoted by

$$\mathcal{T} = \bigcup_{N \geqslant 0} \mathcal{T}_N.$$

7.49. Observation. *$\{e_n\}_{n=-N}^{N}$ is an orthonormal basis of $\mathcal{T}_N$ with respect to the inner product $\langle p, q\rangle = \int_0^1 p(x)\overline{q(x)}\,dx$.*

Proof. Since $\mathcal{T}_N$ is a $(2N+1)$-dimensional vector space over $\mathbb{C}$, it suffices to show that the $(e_n)_{n=-N}^{N}$ form an orthonormal system, i.e. that $\langle e_k, e_\ell\rangle = 0$ if $k \neq \ell$ and $\langle e_k, e_\ell\rangle = 1$ if $k = \ell$. The case $k = \ell$ follows, because the integrand is then equal to 1 and the case $k \neq \ell$ follows from

$$\int_0^1 e^{2\pi i(k-\ell)x}\,dx = \tfrac{1}{2\pi i(k-\ell)}(e^{2\pi i(k-\ell)} - 1) = 0,$$

because $e^{2\pi in} = 1$ for all $n \in \mathbb{Z}$. □

In particular, by Proposition 7.24

$$c_n = \langle p, e_n\rangle = \int_0^1 p(x)e^{-2\pi inx}dx, \tag{7.16}$$

for every $p \in \mathcal{T}_N$ written in the form (7.15). The formula (7.16) makes sense also if p is replaced by a 1-periodic function that is not a trigonometric polynomial (say, Riemann integrable so that the integral is defined). This motivates the following definition.

7.50. Definition. For f Riemann integrable and 1-periodic define the nth *Fourier coefficient* by

$$\widehat{f}(n) = \langle f, e_n\rangle = \int_0^1 f(t)e^{-2\pi int}dt.$$

The operator S_N defined by

$$S_N f(x) = \sum_{n=-N}^{N} \widehat{f}(n)e^{2\pi inx}$$

is called the Nth *partial sum operator*.

We may view the space $\mathcal{T}$ of trigonometric polynomials as a subspace of the inner product space of Riemann integrable 1-periodic functions with the usual inner product $\langle f, g\rangle = \int_0^1 f(x)\overline{g(x)}\,dx$. This space comes with the mean-square seminorm (also called L^2 (semi)norm)

$$\|f\|_2 = \langle f, f\rangle^{1/2} = \Big(\int_0^1 |f(x)|^2\,dx\Big)^{1/2}.$$

The partial sum operator S_N is precisely the orthonormal projection onto the space $\mathcal{T}_N$. If the partial sums $(S_N f)_N$ converge in $\|\cdot\|_2$, then we also say that the *Fourier series*

$$\sum_{n=-\infty}^{\infty} \widehat{f}(n) e^{2\pi i n x} \tag{7.17}$$

converges in L^2 or that it *converges in the mean-square sense*. Note that this does *not* imply that the series (7.17) converges for a given value of $x \in \mathbb{R}$. It turns out that if f is Riemann integrable and 1-periodic, then $(S_N f)_N$ always converges to f in L^2.

By Theorem 7.41 this is equivalent to saying that the orthonormal system $(e_n)_{n\in\mathbb{Z}}$ is complete. Let us convince ourselves that this is the case.

7.51. Theorem (Completeness of trigonometric system). *The orthonormal system $(e_n)_{n\in\mathbb{Z}}$ is complete (in the inner product space of Riemann integrable 1-periodic functions).*

Proof. Let f be Riemann integrable and 1-periodic and let $\varepsilon > 0$. By Theorem 2.25 there exists $h \in C([0,1])$ with $h(0) = h(1) = 0$ so that

$$\|f - h\|_2 \leqslant \varepsilon/2$$

Since $h(0) = h(1)$, we may extend h to a 1-periodic function $h \in C(\mathbb{T})$. Since we have seen in Example 7.19 that $\mathcal{T}$ is dense in $C(\mathbb{T})$ (with respect to the uniform norm $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$), there exists $p \in \mathcal{T}$ such that

$$\|h - p\|_\infty \leqslant \varepsilon/2.$$

Using $\|g\|_2 \leqslant \|g\|_\infty$ we get

$$\|h - p\|_2 \leqslant \|h - p\|_\infty \leqslant \varepsilon/2.$$

Therefore,

$$\|f - p\|_2 \leqslant \|f - h\|_2 + \|h - p\|_2 \leqslant \varepsilon$$

as required. □

Theorem 7.41 then gives us convergence of $(S_N f)$ in $\|\cdot\|_2$ and Parseval's identity:

7.52. Theorem. *Let f be Riemann integrable and 1-periodic. Then*

(i) The Fourier series of f converges to f in L^2, i.e. $\|S_N f - f\|_2 \to 0$ as $N \to \infty$.

(ii) Parseval's identity holds:

$$\sum_{n\in\mathbb{Z}} |\widehat{f}(n)|^2 = \|f\|_2^2.$$

(iii) The Riemann-Lebesgue lemma holds: $\lim_{|n|\to\infty} \widehat{f}(n) = 0$.

Remark. There is nothing special about the period 1 in this section. By considering the orthonormal system $(L^{-1/2} e^{i\frac{2\pi}{L} n x})_{n\in\mathbb{Z}}$ one can obtain similar results for L-periodic functions.

7.4.1. Approximate units and Fejér means. The question of when the partial sums

$$S_N f(x) = \sum_{k=-N}^{N} \widehat{f}(k) e^{2\pi i k x} \tag{7.18}$$

of the Fourier series of f converge and in what sense they represent the function f is very subtle, and we will only scratch the surface in this text. It turns out that even for continuous f, the Fourier series may diverge at a point $x \in \mathbb{R}$. However, the *arithmetic means*

$$\sigma_N f(x) = \frac{1}{N+1} \sum_{n=0}^{N} S_n f(x) \tag{7.19}$$

of the S_n, also called *Fejér means*, converge uniformly to f. In other words, the Fourier series is Cesàro summable to f. The goal of this section is to prove this result (Theorem 7.59 below). In particular

this gives an alternative proof of the completeness of the trigonometric system (Theorem 7.51) using a concrete approximation procedure.

The first step to understand $S_N f$ and $\sigma_N f$ is to identify these objects as *convolutions* of f with certain *kernels*. We will then introduce the concept of *approximate units* and we will learn that convolutions with approximate units are particularly well-behaved. This is a powerful tool also outside of the theory of Fourier series. As we shall see, the convolution kernel of the partial sums $S_N f$ just barely fails to form an approximate unit, but the averaged convolution kernel of $\sigma_N f$ is an approximate unit.

Convolutions of periodic functions. Let f, g be Riemann integrable, 1-periodic functions.

7.53. Definition (Convolution). For two Riemann integrable 1-periodic functions f, g we define their *convolution* $f * g$ by

$$(f * g)(x) = \int_0^1 f(t)g(x-t)\,dt.$$

7.54. Theorem. *For Riemann integrable 1-periodic functions f, g the convolution $f * g$ is 1-periodic and continuous. Moreover convolution is commutative, i.e. $f * g = g * f$ for all f, g.*

Proof. The 1-periodicity of $f*g$ follows immediately from the 1-periodicity of g as we have $g(x+1-t) = g(x-t)$ for all x, t.

We first show the commutativity property. For $x \in [0,1]$,

$$\begin{aligned}(f * g)(x) &= \int_0^1 f(u)g(x-u)\,du = \int_{x-1}^x f(x-t)g(t)\,dt \\ &= \int_{x-1}^0 f(x-t)g(t)\,dt + \int_0^x f(x-t)g(t)\,dt \\ &= \int_x^1 f(x-(t-1))g(t-1)\,dt + \int_0^x f(x-t)g(t)\,dt \\ &= \int_x^1 f(x-t)g(t)\,dt + \int_0^x f(x-t)g(t)\,dt = (g * f)(x),\end{aligned}$$

where in the last step we used the 1-periodicity of f and g.

It remains to show that $f * g$ is continuous. We show that for every $\varepsilon > 0$ there exists $F \in C(\mathbb{T})$ such that $\sup_{x\in[0,1]} |(f * g)(x) - F(x)| < \varepsilon$. This establishes $f * g$ as a limit of continuous, 1-periodic functions, uniformly in $x \in \mathbb{R}$ and implies the continuity of the limiting function $f * g$.

Let M be such that $\|f\|_\infty \leqslant M$, $\|g\|_\infty \leqslant M$. Let $\varepsilon > 0$. By Theorem 2.25 there is a continuous function $h \in C([0,1])$ such that $h(0) = h(1) = 0$ and

$$\int_0^1 |f(t) - h(t)|\,dt < \frac{\varepsilon}{M}.$$

Extend h to a 1-periodic function $h \in C(\mathbb{T})$. Then h is uniformly continuous on $\mathbb{R}$. Let $F = h * g$, which is a 1-periodic function on $\mathbb{R}$, and satisfies

$$|F(x) - (f * g)(x)| = \Big| \int_0^1 (h(t) - f(t))g(x-t)\,dt \Big| \leqslant \int_0^1 |h(t) - f(t)||g(x-t)|\,dt \leqslant \int_0^1 \tfrac{\varepsilon}{M} M\,dt < \varepsilon.$$

Using $h * g = g * h$ we can show that F is continuous. Assume $|x - \tilde{x}| \leqslant \delta$ then

$$|F(x) - F(\tilde{x})| = \Big| \int_0^1 \big(h(x-t) - h(\tilde{x}-t)\big)g(t)\,dt \Big| \leqslant \int_0^1 |g(t)|\,dt \sup_{|w-\tilde{w}|\leqslant\delta} |h(w) - h(\tilde{w})|$$

and $\int_0^1 |g(t)|\,dt \leqslant M$. Thus the uniform continuity of h implies the uniform continuity of F. □

The partial sum $S_N f$ can be written in terms of a convolution: setting

$$D_N(x) = \sum_{n=-N}^{N} e^{2\pi i n x} \tag{7.20}$$

we obtain

$$S_N f(x) = \sum_{n=-N}^{N} \int_0^1 f(t) e^{-2\pi i n t} dt\, e^{2\pi i n x} = \int_0^1 f(t) \sum_{n=-N}^{N} e^{2\pi i n (x-t)} dt = (f * D_N)(x).$$

The sequence of functions $(D_N)_{N\in\mathbb{N}_0}$ is called *Dirichlet kernel.* We derive a more explicit formula.

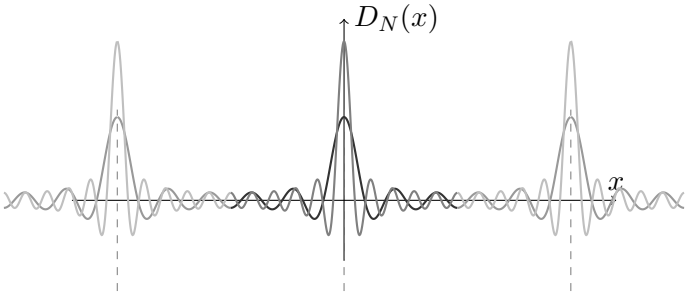


FIGURE 7.2. Dirichlet kernel D_N for $N \in \{5, 10\}$. Note the peak values $D_N(x) = 2N+1$ at $x \in \mathbb{Z}$ and the concentration near integer values of x as N increases.

7.55. LEMMA. *For $N = 0, 1, 2, \ldots$*

$$D_N(x) = \frac{\sin((2N+1)\pi x)}{\sin(\pi x)} \text{ if } x \notin \mathbb{Z}$$

and $D_N(x) = 2N + 1$ for $x \in \mathbb{Z}$.

PROOF. The assertion is clear for $x \in \mathbb{Z}$. For $x \notin \mathbb{Z}$ we compute using the geometric sum formula

$$\begin{aligned} D_N(x) &= \sum_{n=-N}^{N} e^{2\pi i n x} = e^{-2\pi i N x} \sum_{n=0}^{2N} e^{2\pi i n x} = e^{-2\pi i N x} \frac{e^{2\pi i (2N+1) x} - 1}{e^{2\pi i x} - 1} \\ &= \frac{e^{2\pi i (N+\frac{1}{2}) x} - e^{-2\pi i (N+\frac{1}{2}) x}}{e^{\pi i x} - e^{-\pi i x}} = \frac{\sin((2N+1)\pi x)}{\sin(\pi x)}. \end{aligned}$$

□

Similarly we can write the arithmetic means of the S_n as convolutions:

$$\sigma_N f = \mathcal{F}_N * f \text{ where } \mathcal{F}_N(x) = \frac{1}{N+1} \sum_{n=0}^{N} D_n(x). \tag{7.21}$$

The sequence of functions $(\mathcal{F}_N)_{N=0}^{\infty}$ is referred to as *Fejér kernel* and the arithmetic means of the partial sum operators, $\sigma_N f(x) = \mathcal{F}_N * f$, are called *Fejér means.*

7.56. LEMMA. *For $N = 0, 1, 2, \ldots$*

$$\mathcal{F}_N(x) = \sum_{k=-N}^{N} \Big(1 - \frac{|k|}{N+1}\Big) e^{2\pi i k x}$$

and $\sigma_N f(x) = \sum_{k=-N}^{N} \big(1 - \frac{|k|}{N+1}\big) \widehat{f}(k) e^{2\pi i k x}$. Moreover, $\mathcal{F}_N(x) = N + 1$ for $x \in \mathbb{Z}$ and

$$\mathcal{F}_N(x) = \frac{1}{2(N+1)} \frac{1 - \cos(2\pi(N+1)x)}{\sin(\pi x)^2} = \frac{1}{N+1} \Big(\frac{\sin(\pi(N+1)x)}{\sin(\pi x)}\Big)^2, \quad x \notin \mathbb{Z}. \tag{7.22}$$

PROOF. We write

$$\mathcal{F}_N(x) = \frac{1}{N+1}\sum_{n=0}^{N}\sum_{k=-n}^{n} e^{2\pi ikx} = \frac{1}{N+1}\sum_{k=-N}^{N}\sum_{n=|k|}^{N} e^{2\pi ikx} = \sum_{k=-N}^{N}\Big(1-\frac{|k|}{N+1}\Big)e^{2\pi ikx}$$

and thus $\sigma_N f(x)$ is equal to

$$\mathcal{F}_N * f(x) = \int_0^1 f(t)\sum_{k=-N}^{N}\Big(1-\frac{|k|}{N+1}\Big)e^{2\pi ik(x-t)}dt = \sum_{k=-N}^{N}\Big(1-\frac{|k|}{N+1}\Big)\widehat{f}(k)e^{2\pi ikx}.$$

Using that $2\sin(x)\sin(y) = \cos(x-y)-\cos(x+y)$ we compute for $x \notin \mathbb{Z}$

$$\begin{aligned} D_N(x) &= \frac{\sin(2\pi Nx+\pi x)}{\sin \pi x} = \frac{2\sin \pi x\,\sin(2\pi Nx+\pi x)}{2(\sin \pi x)^2} \\ &= \frac{\cos(2\pi Nx)-\cos(2\pi(N+1)x)}{2(\sin \pi x)^2}. \end{aligned}$$

Thus,

$$\sum_{n=0}^{N} D_n(x) = \frac{1}{2(\sin \pi x)^2}\sum_{n=0}^{N}\big(\cos(2\pi nx)-\cos(2\pi(n+1)x)\big) = \frac{1-\cos(2\pi(N+1)x)}{2(\sin \pi x)^2},$$

and the claimed formula for $\mathcal{F}_N$ now follows from $1-\cos 2\alpha = 2(\sin\alpha)^2$. □

As a consequence of this explicit formula we see that $\mathcal{F}_N(x) \geqslant 0$ for all $x \in \mathbb{R}$, which is not at all obvious from the initial definition.

Approximate units. We shall show the convergence of the Fejér means $\sigma_N f$ in $C(\mathbb{T})$ as a special case of a more general result about approximate units.

7.57. DEFINITION (Approximate unit). A sequence of 1-periodic continuous functions $(k_n)_n$ is called an *approximate unit* if for all 1-periodic continuous functions f the sequence $f * k_n$ converges uniformly to f on $\mathbb{R}$. That is,

$$\|f * k_n - f\|_\infty \to 0$$

as $n \to \infty$.

REMARK. The operation of convolution defines a kind of multiplication operation on the vector space $C(\mathbb{T})$. There is no *unit* for the convolution of functions in $C(\mathbb{T})$. More precisely, there exists no continuous function k such that $k * f = f$ for all continuous, 1-periodic f (this is the content of Exercise 7.155). An approximate unit is a sequence (k_n) so that the operators $f \mapsto k_n * f$ approximate the identity operator, in other words, k_n *approximates a unit*, $\lim_{n\to\infty} f * k_n = f$ for every $f \in C(\mathbb{T})$.

7.58. THEOREM. *Let $(k_n)_{n\in\mathbb{N}}$ be a sequence of 1-periodic continuous functions such that*

(i) $\int_{-1/2}^{1/2} k_n(t)\,dt = 1$.

(ii) $\sup_n \int_{-1/2}^{1/2} |k_n(t)|\,dt < \infty$.

(iii) For all $\delta \in (0, \frac{1}{2}]$, $\lim_{n\to\infty} \int_{\delta\leqslant|t|\leqslant 1/2} |k_n(t)|\,dt = 0$.

Then (k_n) is an approximate unit.

Assumption (iii) is a precise way to express the idea that the "mass" of k_n concentrates near the origin, as $n \to \infty$. In many applications k_n is nonnegative; then (ii) is implied by (i).

PROOF OF THEOREM 7.58. We set $M = \sup_n \int_{-1/2}^{1/2} |k_n(t)|\,dt$. Let $f \in C(\mathbb{T})$. By Theorem 4.77, f is bounded and uniformly continuous on $[-\frac{1}{2}, \frac{1}{2}]$. By periodicity, f is also bounded and uniformly continuous on all of $\mathbb{R}$. Let $\varepsilon > 0$. Then there exists $\delta > 0$ such that

$$|f(x-t)-f(x)| \leqslant \tfrac{\varepsilon}{2M} \tag{7.23}$$

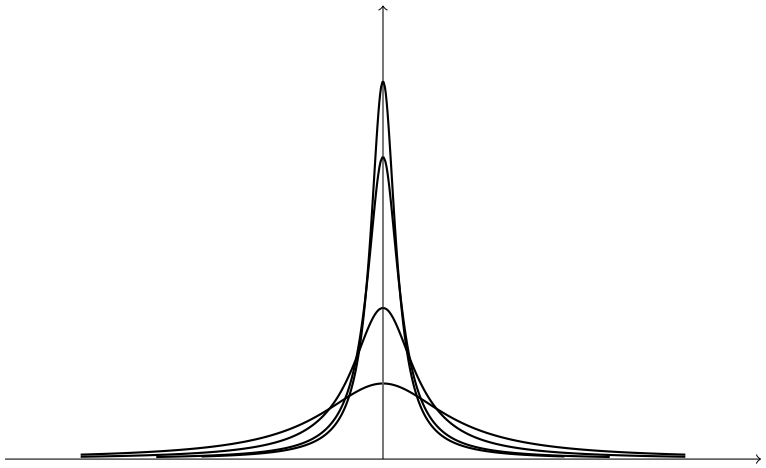


FIGURE 7.3. Approximation of unity

for all $|t| < \delta$, $x \in \mathbb{R}$. Using assumption (i),

$$(f * k_n)(x) - f(x) = \int_{-1/2}^{1/2} (f(x-t) - f(x)) k_n(t)\, dt = A + B,$$

where

$$A = \int_{|t| \leqslant \delta} (f(x-t) - f(x)) k_n(t)\, dt, \quad B = \int_{\frac{1}{2} \geqslant |t| \geqslant \delta} (f(x-t) - f(x)) k_n(t)\, dt.$$

By 7.23 and assumption (ii),

$$|A| \leqslant \frac{\varepsilon}{2M} \int_{|t| \leqslant \delta} |k_n(t)|\, dt \leqslant \tfrac{\varepsilon}{2}.$$

Since f is bounded there exists $C > 0$ such that $|f(x)| \leqslant C$ for all $x \in \mathbb{R}$. By assumption (iii) we may choose N to be large enough so that for all $n \geqslant N$,

$$\int_{\frac{1}{2} \geqslant |t| \geqslant \delta} |k_n(t)|\, dt \leqslant \tfrac{\varepsilon}{4C}.$$

Thus, if $n \geqslant N$,

$$|B| \leqslant 2C \int_{\frac{1}{2} \geqslant |t| \geqslant \delta} |k_n(t)|\, dt \leqslant \tfrac{\varepsilon}{2}.$$

This implies

$$|(f * k_n)(x) - f(x)| \leqslant \varepsilon$$

for $n \geqslant N$ and $x \in \mathbb{R}$. □

The Dirichlet kernel $(D_N)_N$ satisfies assumption (i) of the theorem:

$$\int_{-1/2}^{1/2} D_N(t)\, dt = \sum_{n=-N}^{N} \int_{-1/2}^{1/2} e^{2\pi i n t}\, dt = 1, \tag{7.24}$$

because the integral equals zero unless $n = 0$. It also satisfies assumption (iii). However, it does not satisfy (ii). In fact, we have $\int_0^1 |D_N(x)|\, dx \geqslant C \log(N+1)$ for a constant $C \in (0, \infty)$ (we will prove this in Lemma 7.77 below). On the other hand, the Fejér kernel $(\mathcal{F}_N)_N$ satisfies all three assumptions and thus forms an approximate unit:

7.59. THEOREM (Fejér). *For every $f \in C(\mathbb{T})$, the Fejér means $\sigma_N f$ converge uniformly to f as $N \to \infty$.*

PROOF. It suffices to verify the assumptions of Theorem 7.58. By (7.24),

$$\int_{-1/2}^{1/2} \mathcal{F}_N(t)\,dt = \frac{1}{N+1}\sum_{n=0}^{N}\int_{-1/2}^{1/2} D_n(t)\,dt = 1,$$

so assumption (i) is verified. Since $\mathcal{F}_N \geqslant 0$ by (7.22), this also shows assumption (ii). It remains to verify (iii). Let $\frac{1}{2} > \delta > 0$. By (7.22),

$$\mathcal{F}_N(x) \leqslant \frac{1}{(N+1)(\sin(\pi\delta))^2} \qquad \text{for } \delta \leqslant |x| \leqslant 1/2.$$

Thus,

$$\int_{\frac{1}{2}\geqslant|t|\geqslant\delta} \mathcal{F}_N(t)\,dt \leqslant \tfrac{1}{(N+1)(\sin(\pi\delta))^2},$$

which converges to 0 as $N \to \infty$. □

Fejér's theorem gives a direct proof of the completeness of the trigonometric system not relying on the Stone-Weierstrass theorem. In turn, Fejér's theorem can also be used to give a short proof of the original Weierstrass approximation theorem for polynomials.

7.60. EXERCISE. Give an alternative proof of Weierstrass' theorem (Theorem 7.3) by using Fejér's theorem and then approximating the resulting trigonometric polynomials by truncated Taylor expansions.

We proceed with an extension of Fejér's theorem.

7.61. THEOREM. *Let f be Riemann integrable and 1-periodic.*

(i) Let x_0 be such that the limit

$$\lim_{t\to 0+} \tfrac{1}{2}(f(x_0+t)+f(x_0-t)) = L$$

exists. Then the sequence $(\sigma_N f(x_0))_N$ converges to L.

(ii) Let J be a compact subinterval of $[0,1]$ and assume that f is continuous on J. Then for all $x \in J$, the sequence $(\sigma_N f(x))_N$ converges to $f(x)$, and the convergence is uniform on J.

PROOF. We use that the Fejér kernel is even and therefore $\int_0^{1/2} \mathcal{F}_N(t)\,dt = 1/2$. We thus have

$$\begin{aligned}\sigma_N f(x_0) - L &= \int_{-1/2}^{1/2} \mathcal{F}_N(t) f(x_0-t)\,dt - L\\ &= \int_0^{1/2} \mathcal{F}_N(t) f(x_0-t)\,dt + \int_0^{1/2} \mathcal{F}_N(t) f(x_0+t)\,dt - L\\ &= \int_0^{1/2} \mathcal{F}_N(t)\big[f(x_0-t)+f(x_0+t)-2L\big]\,dt.\end{aligned}$$

Let $\varepsilon > 0$. By assumption there exists $\delta = \delta(x_0,\varepsilon)$ such that

$$|f(x_0-t)+f(x_0+t)-2L| < \varepsilon$$

for $0 \leqslant t \leqslant \delta$. Let $M = \sup|f(x)| < \infty$. Since $\mathcal{F}_N \geqslant 0$, by the above,

$$|\sigma_N f(x_0) - L| \leqslant \varepsilon\int_0^{\delta} \mathcal{F}_N(t)\,dt + 4M\int_{\delta}^{1/2} \mathcal{F}_N(t)\,dt \leqslant \tfrac{\varepsilon}{2} + 4M\int_{\delta}^{1/2} \mathcal{F}_N(t)\,dt.$$

Since the Fejér kernel is an approximate unit (Theorem 7.59) there is an N_0 such that $4M\int_{\delta}^{1/2}|\mathcal{F}_N(t)|dt \leqslant \varepsilon/2$ for $N \geqslant N_0$. Putting the two estimates together we obtain the inequality $|\sigma_N f(x_0) - L| < \varepsilon$ for $N \geqslant N_0$.

Now let $J = [a, b]$ be a compact interval which does not contain any points of discontinuity of f. Let $\varepsilon > 0$. By uniform continuity there exists a $\delta > 0$ such that for all $x \in J$ and t with $x - t \in J$ and $|t| < \delta$ we have $|f(x - t) - f(x)| \leqslant \varepsilon/2$. The above argument gives

$$|\sigma_N f(x) - f(x)| \leqslant \tfrac{\varepsilon}{2} \int_0^\delta \mathcal{F}_N(t)\, dt + 4M \int_\delta^{1/2} \mathcal{F}_N(t)\, dt$$

for all $x \in J$. Continuing the argument as above we get (ii). □

7.4.2. Rates of approximation. In many cases one can make statements about the rates of convergence of an approximation procedure when the function f enjoys additional regularity (we have seen an example of such a phenomenon in Exercise 7.7 on the Bernstein polynomials).

7.62. Definition. For a bounded 1-periodic function f we define the *modulus of continuity* as a function of $\delta > 0$ by

$$\omega_f(\delta) = \sup_{|h| \leqslant \delta} \|f(\cdot + h) - f(\cdot)\|_\infty.$$

Here $\|\cdot\|_\infty$ is the usual sup-norm. A function on $\mathbb{R}$ is uniformly continuous if and only if

$$\lim_{\delta \to 0+} \omega_f(\delta) = 0,$$

and thus a 1-periodic function f belongs to $C(\mathbb{T})$ if and only if $\lim_{\delta \to 0+} \omega_f(\delta) = 0$.

7.63. Lemma.

(i) $\omega_f(\delta) \leqslant \omega_f(\delta')$ *if* $\delta \leqslant \delta'$.
(ii) $\omega_f(\delta_1 + \delta_2) \leqslant \omega_f(\delta_1) + \omega_f(\delta_2)$.
(iii) $\omega_{f+g}(\delta) \leqslant \omega_f(\delta) + \omega_g(\delta)$.
(iv) $\omega_f(R\delta) \leqslant \lceil R \rceil \omega_f(\delta)$ *for* $R \geqslant 1$ *and* $\delta > 0$.

Proof. (i), (ii), (iii): These follow from the definition of modulus of continuity.

(iv): This follows for $R = n \in \mathbb{N}$ from writing

$$f(x + nh) - f(x) = \sum_{j=0}^{n-1} \big(f(x + jh + h) - f(x + jh)\big),$$

and the bound $|f(x + jh + h) - f(x + jh)| \leqslant \omega_f(\delta)$ for $|h| \leqslant \delta$. If $R \notin \mathbb{N}$ use the monotonicity of ω_f to get $\omega_f(R\delta) \leqslant \omega_f(\lceil R \rceil \delta)$. □

The next result connects the modulus of continuity with the rate of convergence of the Fejér means $\sigma_N f$.

7.64. Theorem. *Let* $f \in C(\mathbb{T})$. *Then the following hold true.*

(i) For any $N \geqslant 1$,

$$\|\sigma_N f - f\|_\infty \leqslant \Big(\omega_f(\tfrac{1}{2N}) + \int_{(2N)^{-1}}^{1/2} \frac{\omega_f(t)}{4(N+1)t^2}\, dt\Big).$$

(ii) $\|\sigma_N f - f\|_\infty \leqslant (1 + 2\log(2N))\, \omega_f(\frac{1}{2N})$.
(iii) If $0 < \alpha < 1$ *and* $\omega_f(\delta) \leqslant A\delta^\alpha$ *for some* $A > 0$, *then* $\|\sigma_N f - f\|_\infty \leqslant AC_\alpha N^{-\alpha}$.

Proof. Since $\int_{-1/2}^{1/2} \mathcal{F}_N(t) dt = 1$ we may write

$$\sigma_N f(x) - f(x) = \int_{-1/2}^{1/2} \mathcal{F}_N(t)(f(x - t) - f(x))\, dt$$

We split the integral in two parts $I + II$ where $|t| < (2N)^{-1}$ in I and $(2N)^{-1} \leqslant |t| \leqslant 1/2$ in II. Use $\int_{-1/2}^{1/2} |\mathcal{F}_N(t)|\,dt = 1$ for I and $|\mathcal{F}_N(t)| \leqslant (N+1)^{-1}(\sin \pi t)^{-2} \leqslant (N+1)^{-1}(4t^2)^{-1}$ for II. Then

$$I \leqslant \omega_f(\tfrac{1}{2N}),$$

and

$$II \leqslant 2\int_{(2N)^{-1}\leqslant t\leqslant 1/2} \frac{\omega_f(t)}{(N+1)(4t^2)}\,dt.$$

Adding these terms yields part (i).

For (ii) we use from Lemma 7.63 the estimate $\omega_f(t) \leqslant \lfloor 2Nt+1\rfloor \omega_f(\frac{1}{2N})$ for $t > (2N)^{-1}$ and then

$$II \leqslant \omega_f(\tfrac{1}{2N}) \int_{(2N)^{-1}}^{1/2} \frac{\lfloor 2Nt+1\rfloor}{2(N+1)t^2}\,dt \leqslant \omega_f(\tfrac{1}{2N})\, 2\log(2N).$$

Regarding (iii), if $\sup_\delta \delta^{-\alpha}\omega_f(\delta) \leqslant A$ we get $I \leqslant A(2N)^{-\alpha}$ and

$$|II| \leqslant \frac{A}{4(N+1)} \int_{(2N)^{-1}}^{1/2} t^{\alpha-2}\,dt \leqslant 2\frac{A(2N)^{1-\alpha}}{4(1-\alpha)(N+1)}.$$

Hence, adding these terms,

$$\|\sigma_N f - f\|_\infty \leqslant \frac{2^{1-\alpha}}{1-\alpha} A N^{-\alpha}. \qquad \square$$

We shall now look closer into the problem of *best approximation* from the space of trigonometric polynomials of degree N. Here one aims to determine the asymptotic behavior of the distance of $f \in C(\mathbb{T})$ to the space of trigonometric polynomials of degree N, that is

$$\operatorname{dist}_\infty(f, \mathcal{T}_N) = \inf_{p\in\mathcal{T}_N} \|f - p\|_\infty.$$

Since $\operatorname{dist}_\infty(f, \mathcal{T}_N) \leqslant \|\sigma_N f - f\|_\infty$ upper bounds are given by Theorem 7.64. Part (ii) of that theorem yields $\operatorname{dist}_\infty(f, \mathcal{T}_N) \leqslant C\omega_f(\frac{1}{2N})\log N$ for large N, but the following theorem shows that this can be improved, removing the logarithmic factor.

7.65. Theorem (Jackson). *There exists a constant $C \in (0,\infty)$ such that for all $f \in C(\mathbb{T})$ and $N \geqslant 0$*

$$\operatorname{dist}_\infty(f, \mathcal{T}_N) \leqslant C\omega_f(\tfrac{1}{N+1}).$$

We shall first prove a few auxiliary facts. The idea of this theorem is to construct an approximation of the identity consisting of trigonometric polynomials $(\mathcal{K}_N)$ of degree roughly N so that $\|f - f*\mathcal{K}_N\|_\infty \leqslant C\omega_f(\frac{1}{N+1})$. In the following lemma we consider a family of approximative units with assumptions that are somewhat more restrictive than those in (i), (ii), (iii) of Theorem 7.58.

7.66. Lemma. *Let $(k_n)_{n\in\mathbb{N}}$ be a sequence of kernels satisfying the conditions*

(i) $\int_{-1/2}^{1/2} k_n(t)\,dt = 1$.

(ii) *There exists $\beta > 2$ and $A > 0$ such that*

$$|k_n(t)| \leqslant A\frac{n}{1+(n|t|)^\beta} \text{ for all } n \in \mathbb{N},\ |t| \leqslant 1/2$$

Then

$$\|f - f * k_n\|_\infty \leqslant C_\beta A \omega_f(\tfrac{1}{2n}),$$

where $C_\beta = 1 + \frac{2^{\beta+1}}{\beta-2}$.

PROOF. We write

$$|f * k_n(x) - f(x)| = \Big| \int_{-1/2}^{1/2} k_n(t)\big(f(x-t) - f(x)\big)\,dt\Big|$$
$$\leqslant A \int_{-1/2}^{1/2} \frac{n}{1+(n|t|)^\beta}\omega_f(|t|)\,dt \leqslant 2A \int_0^{1/2} \frac{n}{1+(nt)^\beta}\omega_f(t)\,dt$$

Observe

$$\int_0^{\frac{1}{2n}} \frac{n}{1+(nt)^\beta}\omega_f(|t|)\,dt \leqslant \tfrac{1}{2}\omega_f(\tfrac{1}{2n}).$$

Using Lemma 7.63 (iv) with $R = 2nt$ we have $\omega_f(t) \leqslant 4nt\,\omega_f(\frac{1}{2n})$ for $t \geqslant \frac{1}{2n}$ and may therefore estimate

$$\int_{\frac{1}{2n}}^{\frac{1}{2}} \frac{n}{1+(nt)^\beta}\omega_f(|t|)\,dt \leqslant \omega_f(\tfrac{1}{2n}) \int_{\frac{1}{2n}}^{\infty} 4n^{2-\beta}t^{1-\beta}\,dt = \tfrac{2^\beta}{\beta-2}\omega_f(\tfrac{1}{2n}).$$

□

Note that the condition $\beta > 2$ was crucial for the convergence of the last integral. For the Fejér kernel we have the estimate in (ii) only with $\beta = 2$, and this causes the additional logarithmic factor in part (ii) of Theorem 7.64. Hence, in order to proceed with the proof of Jackson's theorem, we need to construct kernels $\mathcal{K}_N$ satisfying the strengthened decay condition; we also want $\mathcal{K}_N$ to be trigonometric polynomials with controlled degree. The idea is to square the Fejér kernel

$$\mathcal{F}_N(x) = (N+1)^{-1}\Big(\frac{\sin \pi(N+1)x}{\sin \pi x}\Big)^2$$

and then appropriately normalize the square.

7.67. LEMMA. *For $N \geqslant 1$, define*

$$G_N(x) = \Big(\frac{\sin \pi(N+1)x}{\sin \pi x}\Big)^4.$$

Then

(i) G_N is an even trigonometric polynomial in $\mathcal{T}_{2N}$.

(ii) $\int_{-1/2}^{1/2} G_N(x)\,dx = B_N$ with $B_N = \frac{1}{3}(N+1)(2(N+1)^2+1)$.

(iii) The sequence of kernels $\mathcal{K}_N = B_N^{-1}G_N$ is an approximative unit of functions satisfying

$$|\mathcal{K}_N(x)| \leqslant C\frac{N+1}{1+((N+1)|x|)^4} \text{ for } |x| \leqslant 1/2.$$

PROOF. G_N is an even function. We use the formulas for the Fejér kernel (7.22)

$$G_N(x) = (N+1)^2[\mathcal{F}_N(x)]^2 = (N+1)^2\Big(\sum_{j=-N}^{N}\big(1-\frac{|j|}{N+1}\big)e^{2\pi ijx}\Big)^2 = \sum_{k=-2N}^{2N} c_k e^{2\pi ikx}$$

with suitable coefficients c_k, so that $G_N \in \mathcal{T}_{2N}$. Since $\int_{-1/2}^{1/2} e^{2\pi ikx}dx = 0$ for $k \neq 0$ we get

$$\int_{-1/2}^{1/2} G_N(x)\,dx = c_0 = (N+1)^2 \sum_{j=-N}^{N}\big(1-\frac{|j|}{N+1}\big)\big(1-\frac{|-j|}{N+1}\big) = (N+1)^2 + 2\sum_{\ell=1}^{N}\ell^2$$
$$= (N+1)^2 + 2\frac{N(N+1)(2N+1)}{6} = \frac{N+1}{3}\big(2(N+1)^2+1\big)$$

which is the formula for B_N. To show (iii) we use $|\sin \pi(N+1)x| \leqslant (N+1)\sin(\pi x)$ for $|x| \leqslant \frac{1}{\pi(N+1)}$ and $\sin \pi x \geqslant 2x$ for $0 \leqslant x \leqslant 1/2$ and get

$$|\mathcal{K}_N(x)| = \frac{|G_N(x)|}{B_N} \leqslant \begin{cases} \frac{3}{2}(N+1) & \text{for } |x| \leqslant \frac{1}{\pi(N+1)} \\ \frac{3}{2}\frac{(N+1)}{(2|x|(N+1))^4} & \text{for } \frac{1}{\pi(N+1)} < |x| < \frac{1}{2}\,. \end{cases}$$

which implies (iii). □

The family $(\mathcal{K}_N)_{N\geqslant 0}$ is also referred to as the *Jackson kernel* or *Jackson approximative unit*. Combining Lemmas 7.66 and 7.67 we can complete the proof of Jackson's theorem.

Proof of Theorem 7.65, conclusion. We consider separately the cases $N = 2L$ and $N = 2L+1$. By Lemma 7.66 (with $n = N+1$) and Lemma 7.67 we have

$$\operatorname{dist}_\infty(f, \mathcal{T}_{2L}) \leqslant \|f - f * \mathcal{K}_L\|_\infty \leqslant C\omega_f(\tfrac{1}{L+1})$$

and therefore

$$\operatorname{dist}_\infty(f, \mathcal{T}_{2L+1}) \leqslant \operatorname{dist}_\infty(f, \mathcal{T}_{2L}) \leqslant C\omega_f(\tfrac{1}{L+1}).$$

Now the claim follows by Lemma 7.63. □

Remark. For more on Jackson inequalities see [**54, 11, 15**].

For later applications we also note that for even functions we have a similar Jackson theorem for the distance of f to the space $\mathcal{T}_N^{\mathrm{ev}}$ of even trigonometric polynomials in $\mathcal{T}_N$ (which can be identified with the cosine polynomials in $\mathcal{T}_N$). This is an immediate consequence of

7.68. Lemma. *Let $f \in C(\mathbb{T})$ be even. Then* $\operatorname{dist}_\infty(f, \mathcal{T}_N) = \operatorname{dist}_\infty(f, \mathcal{T}_N^{\mathrm{ev}})$.

Proof. By definition of the infimum, $\operatorname{dist}_\infty(f, \mathcal{T}_N) \leqslant \operatorname{dist}_\infty(f, \mathcal{T}_N^{\mathrm{ev}})$. Let $p \in \mathcal{T}_N$ such that $\|f - p\|_\infty = \operatorname{dist}_\infty(f, \mathcal{T}_N)$. Then also $\|f(-\cdot) - p(-\cdot)\|_\infty = \|f - p\|_\infty = \operatorname{dist}_\infty(f, \mathcal{T}_N)$. Let $p^{\mathrm{ev}}(x) = \frac{1}{2}(p(x) + p(-x))$, so that $p^{\mathrm{ev}} \in \mathcal{T}_N^{\mathrm{ev}}$ and we have $\|f - p^{\mathrm{ev}}\|_\infty = \|\frac{f+f(-\cdot)}{2} - \frac{p+p(-\cdot)}{2}\|_\infty \leqslant \operatorname{dist}_\infty(f, \mathcal{T}_N)$. Hence $\operatorname{dist}_\infty(f, \mathcal{T}_N^{\mathrm{ev}}) \leqslant \operatorname{dist}_\infty(f, \mathcal{T}_N)$. □

While Jackson's theorem is about best uniform approximation from $\mathcal{T}_N$ (which is not necessarily realized by the partial sums $S_N f$), it can be used to obtain results about uniform convergence for Fourier series and the rate of convergence of $S_N f$ to f (see Theorem 7.74).

In approximation theory, results such as Theorems 7.64 and 7.65 are referred to as *direct (or Jackson-type) approximation results*. We now turn to *inverse (or Bernstein-type) approximation results* where assumptions on $\operatorname{dist}_\infty(f, \mathcal{T}_N)$ give information about the modulus of continuity.

7.69. Theorem. *Suppose $f \in C(\mathbb{T})$.*

(i) Setting $\mathcal{E}_N(f) = \operatorname{dist}_\infty(f, \mathcal{T}_N)$ we have

$$\omega_f(2^{-k}) \leqslant 2^{5-k}\mathcal{E}_0(f) + \sum_{\ell=0}^{k} 2^{5-\ell}\mathcal{E}_{2^{k-\ell}}(f).$$

(ii) Let $\alpha \in (0,1)$. There exists a constant $C_\alpha \in (0,\infty)$ such that

$$\operatorname{dist}_\infty(f, \mathcal{T}_N) \leqslant BN^{-\alpha} \text{ for all } N \geqslant 0 \implies \omega_f(\delta) \leqslant C_\alpha B\delta^\alpha \text{ for all } 0 < \delta \leqslant 1.$$

We need an important auxiliary result for the proof of Theorem 7.69 which relates the supremum norms of a trigonometric polynomial and its derivative. This is *Bernstein's inequality* for trigonometric polynomials.

7.70. Proposition. *For all $p \in \mathcal{T}_N$,*

$$\|p'\|_\infty \leqslant 4\pi(N+1)\|p\|_\infty.$$

The sharp form of Bernstein's inequality states that one can replace the constant $4\pi(N+1)$ with $2\pi N$ for all $p \in \mathcal{T}_N$ (the example $p(x) = e^{2\pi i N x}$ shows that this is optimal). We shall prove this later in Theorem 7.135. In many applications (such as in the proof of Theorem 7.69 below) one just needs that the constant is $O(N)$ as $N \to \infty$. We give a proof relying again on the Fejér kernel.

PROOF OF PROPOSITION 7.70. Let $p \in \mathcal{T}_N$, $p(x) = \sum_{k=-N}^{N} c_k e^{2\pi i k x}$. Then

$$p'(x) = 2\pi i \sum_{k=-N}^{N} k e^{2\pi i k x}.$$

We recall that the Fourier coefficients of the Fejér kernel $\mathcal{F}_N$ are given by $\widehat{\mathcal{F}_N}(k) = g_N(k)$ with $g_N(k) = (1 - |k|/(N+1))_+$, where $w_+ = \max(w, 0)$. We compute that

$$g_N(k-N-1) - g_N(k+N+1) = \begin{cases} 2 - k/(N+1) & \text{if } N < k \leqslant 2N+1 \\ k/(N+1) & \text{if } -N \leqslant k \leqslant N \\ -2 - k/(N+1) & \text{if } -2N-1 \leqslant k < -N \\ 0 & \text{if } |k| \geqslant 2N+2. \end{cases}$$

Observe that $g_N(k \mp (N+1))$ is the kth Fourier coefficient of $\mathcal{F}_N(x) e^{\pm 2\pi i (N+1) x}$. For $p \in \mathcal{T}_N$ we have $\widehat{p}(n) = 0$ for $|n| > N$ and thus only the range $-N \leqslant k \leqslant N$ is relevant. We get for $|k| \leqslant N$ and then also for all $k \in \mathbb{Z}$

$$\begin{aligned} \widehat{p'}(k) = 2\pi i k \widehat{p}(k) &= 2\pi i (N+1)\big(g_N(k-N-1) - g_N(k+N+1)\big)\widehat{p}(k) \\ &= 2\pi (N+1) \widehat{H_N}(k) \widehat{p}(k) \end{aligned}$$

where

$$H_N(x) = i(e^{2\pi i (N+1) x} - e^{-2\pi i (N+1) x}) \mathcal{F}_N(x) = -2 \sin(2\pi (N+1) x) \mathcal{F}_N(x).$$

Recall that $\mathcal{F}_N$ is nonnegative and $\int_0^1 \mathcal{F}_N(x)\, dx = 1$. Thus

$$\int_0^1 |H_N(x)|\, dx = 2 \int_0^1 \mathcal{F}_N(x) |\sin(2\pi (N+1) x)|\, dx < 2.$$

Hence,

$$\|p'\|_\infty = 2\pi (N+1) \|H_N * p\|_\infty \leqslant 2\pi (N+1) \int_0^1 |H_N|\, dx\, \|p\|_\infty \leqslant 4\pi (N+1) \|p\|_\infty. \qquad \square$$

We are now ready to prove the inverse approximation theorem.

PROOF OF THEOREM 7.69. Let $\varepsilon > 0$. By definition of $\mathcal{E}_N(f)$ and Proposition 7.1, there is a trigonometric polynomial p_N of degree at most N such that $\|f - p_N\|_\infty = \mathcal{E}_N(f)$. For $|h| \leqslant 2^{-k}$ we have

$$\begin{aligned} |f(x+h) - f(x)| &\leqslant |f(x+h) - p_{2^k}(x+h) - (f(x) - p_{2^k}(x))| + |p_{2^k}(x+h) - p_{2^k}(x)| \\ &\leqslant 2\|f - p_{2^k}\|_\infty + |h| \|p'_{2^k}\|_\infty. \end{aligned}$$

Now use that $p'_0 = 0$ and write p'_{2^k} as a collapsing sum $p'_{2^k} = \sum_{\ell=0}^{k-1} \big(p'_{2^{k-\ell}} - p'_{2^{k-\ell-1}}\big) + p'_1$. From this

$$|f(x+h) - f(x)| \leqslant 2\|f - p_{2^k}\|_\infty + 2^{-k} \Big(\sum_{\ell=0}^{k-1} \|p'_{2^{k-\ell}} - p'_{2^{k-\ell-1}}\|_\infty + \|p'_1 - p'_0\|_\infty \Big).$$

For the derivative terms we use Bernstein's inequality (Proposition 7.70) and get

$$\|p'_{2^{k-\ell}} - p'_{2^{k-\ell-1}}\|_\infty \leqslant 2^{k-\ell+5} \|p_{2^{k-\ell}} - p_{2^{k-\ell-1}}\|_\infty$$

and $\|p'_1 - p'_0\|_\infty \leqslant 2^5 \|p_1 - p_0\|_\infty$. Also

$$\|p_{2^{k-\ell}} - p_{2^{k-\ell-1}}\|_\infty \leqslant \|p_{2^{k-\ell}} - f\|_\infty + \|p_{2^{k-\ell-1}} - f\|_\infty \leqslant \mathcal{E}_{2^{k-\ell}}(f) + \mathcal{E}_{2^{k-\ell-1}}(f).$$

Hence

$$\omega_f(2^{-k}) \leqslant 2\mathcal{E}_{2^k}(f) + 2^{-k+5}\Big(\sum_{\ell=0}^{k-1} 2^{k-\ell}\mathcal{E}_{2^{k-\ell}}(f) + \mathcal{E}_0(f)\Big)$$

and we obtain the first estimate.

For the second estimate, assuming $\mathcal{E}_N(f) \leqslant BN^{-\alpha}$ we get

$$\omega_f(2^{-k}) \leqslant 2B2^{-k\alpha} \,+\, 2^{-k+5}\sum_{\ell=0}^{k-1} B2^{(k-\ell)(1-\alpha)} \leqslant C(\alpha)B2^{-k\alpha},$$

where we have used that $\alpha \in (0,1)$. Since $\omega_f(t) \leqslant \omega_f(2^{-k})$ for $t \in [2^{-k-1}, 2^{-k}]$ we obtain the asserted estimate for all $t \leqslant 1$. □

Combining the direct approximation theorem (Theorem 7.64) and the inverse approximation theorem (Theorem 7.69) we obtain an important consequence.

7.71. Corollary. *Let $\alpha \in (0,1)$. For any $f \in C(\mathbb{T})$ the following statements are equivalent.*

(i) $\operatorname{dist}_\infty(f, \mathcal{T}_N) = O(N^{-\alpha})$ *as* $N \to \infty$.

(ii) $\|\sigma_N f - f\|_\infty = O(N^{-\alpha})$ *as* $N \to \infty$.

(iii) $\omega_f(\delta) = O(\delta^\alpha)$ *as* $\delta \to 0+$.

Moreover, there exist positive constants C_1, C_2, C_3 such that

$$\sup_{N\geqslant 1} \operatorname{dist}_\infty(f, \mathcal{T}_N) \leqslant C_1 \sup_{N\geqslant 1} N^\alpha \|\sigma_N f - f\|_\infty \leqslant C_2 \sup_{0<\delta<1/2} \delta^{-\alpha}\omega_f(\delta) \leqslant C_3 \sup_{N\geqslant 1} \operatorname{dist}_\infty(f, \mathcal{T}_N).$$

Remarks. (i) There are analogues of this corollary for higher degrees of regularity (i.e. $\alpha \geqslant 1$). These involve higher order differences and moduli of continuity, or derivatives and the usual first order modulus of continuity (see Exercise 7.176). For more on this see DeVore–Lorentz [**15**, Ch. 7].

(ii) For $\alpha \in (0,1)$ the set of $f \in C(\mathbb{T})$ for which $\omega_f(\delta) = O(\delta^\alpha)$ is denoted in the literature by $C^{0,\alpha}(\mathbb{T})$ or $\operatorname{Lip}_\alpha(\mathbb{T})$ and referred to as the *Hölder class* of index α or *Lipschitz class* of index α. The spaces here are periodic versions of the space defined in Exercise 5.6.

(iii) The implication (i) $\Longrightarrow$ (iii) in the corollary fails for $\alpha = 1$. See Exercises 7.172, 7.174, for a replacement involving the *Zygmund class* defined via second differences.

7.72. Exercise. Let $\alpha \in (0,1)$ and consider the 1-periodic function f_α defined on $[-1/2, 1/2]$ by $f_\alpha(x) = (1-2|x|)^\alpha$. Show that $f_\alpha \in C^{0,\alpha}(\mathbb{T})$ but $f_\alpha \notin C^{0,\beta}(\mathbb{T})$ for $\beta > \alpha$.

7.5. More on convergence of Fourier series

We saw in Theorem 7.52 that the Fourier series of a continuous 1-periodic function converges in L^2. It is natural to ask whether we also get pointwise convergence, i.e. for a fixed $x \in \mathbb{R}$, does the Fourier series

$$\sum_{n=-\infty}^{\infty} \widehat{f}(n) e^{2\pi i n x}$$

converge? In Theorem 7.74 we will see that even uniform convergence holds provided that f satisfies a mild regularity condition. However, pointwise convergence may fail for functions that are only continuous (Theorem 7.76 below). On the other hand, if f is differentiable at a point x_0, then we can prove $S_N f(x_0) \to f(x_0)$ (Theorem 7.81 below). Finally, in §7.5.4 we will discuss conditions for *absolute convergence*, i.e. $\sum_{n\in\mathbb{Z}} |\widehat{f}(n)| < \infty$. Absolute convergence implies uniform convergence and in particular, pointwise convergence at every $x \in \mathbb{R}$.

7.5.1. Uniform convergence. We first prove an auxiliary lemma about the growth of the operator norms of the partial sum operators.

7.73. Lemma. *The operator norm $\|S_N\|_{\mathrm{op}}$ on $C(\mathbb{T})$ satisfies*

$$\|S_N\|_{\mathrm{op}} \leqslant 1 + \log((2N+1)\tfrac{\pi}{2}).$$

Proof. We have $S_N g = g * D_N$ and thus

$$|S_N g(x)| \leqslant \int_0^1 |D_N(t)|\,dt\,\|g\|_\infty.$$

We use the 1-periodicity, and the fact that $|D_N|$ is even in the following identity

$$\int_0^1 |D_N(t)|\,dt = \int_{-1/2}^{1/2} |D_N(t)|\,dt = 2\int_0^{1/2} |D_N(t)|\,dt = 2\int_0^{1/2} \frac{|\sin((2N+1)\pi t)|}{|\sin(\pi t)|}\,dt.$$

For $0 < s < \pi/2$ we have $\sin s \geqslant \frac{2}{\pi}s$ (prove this and draw a graph first). Hence we get

$$\int_0^1 |D_N(t)|dt \leqslant 2\int_0^{1/2} \frac{|\sin((2N+1)\pi t)|}{t}dt.$$

In the integrand we use $\sin((2N+1)\pi t) \leqslant (2N+1)\pi t$ for $0 < t \leqslant \frac{1}{(2N+1)\pi}$ and obtain that the last integral is bounded by

$$\int_0^{\frac{1}{(2N+1)\pi}} (2N+1)\pi\,dt + \int_{\frac{1}{(2N+1)\pi}}^{1/2} \frac{dt}{t} = 1 + \log(\tfrac{2N+1}{2}\pi).$$

□

The next result shows in particular that under a mild regularity condition on f we have uniform convergence of Fourier series.

7.74. Theorem. *There exists a constant $C \in (0,\infty)$ such that for all $f \in C(\mathbb{T})$*

$$\|S_N f - f\|_\infty \leqslant C\log(N+2)\,\omega_f(\tfrac{1}{N}).$$

In particular, if $\omega_f(\delta) = o(\frac{1}{\log \delta^{-1}})$ as $\delta \to 0$, then $S_N f$ converges uniformly to f.

Proof. Let $f \in C(\mathbb{T})$ and find a $p \in \mathcal{T}_N$ with $\mathrm{dist}_\infty(f, \mathcal{T}_N) = \|f - p\|_\infty$. Then using $S_N p = p$ and Lemma 7.73,

$$\begin{aligned}\|S_N f - f\|_\infty &= \|S_N(f-p) - (f-p)\|_\infty \leqslant \|S_N(f-p)\|_\infty + \|f-p\|_\infty \\ &\leqslant (1 + \|S_N\|_{\mathrm{op}})\|f-p\|_\infty \leqslant \big(2 + \log(\tfrac{2N+1}{2}\pi)\big)\|f-p\|_\infty.\end{aligned}$$

By Jackson's inequality (Theorem 7.65) there is a constant C_1 (independent of f) such that

$$\|f-p\|_\infty \leqslant C_1\omega_f(\tfrac{1}{N}).$$

□

7.75. Corollary. *If $f \in C^{0,\alpha}(\mathbb{T})$ for an $\alpha \in (0,1]$, then $(S_N f)$ converges uniformly to f.*

This follows from Theorem 7.74, because $f \in C^{0,\alpha}(\mathbb{T})$ means that $\omega_f(\delta) = O(\delta^\alpha)$. In the next section we will see that this is sharp in the sense that if f is only continuous, then the conclusion fails.

7.5.2. Failure of pointwise convergence.

7.76. THEOREM. *Let $x_0 \in \mathbb{R}$. There exists a 1-periodic continuous function f such that the sequence $(S_N f(x_0))_N \subset \mathbb{C}$ is unbounded.*

In particular, this proves that the Dirichlet kernel D_N is not an approximate unit. For the proof we will use the uniform boundedness principle (Theorem 5.35). We first show that the Dirichlet kernel (D_N) violates condition (ii) in Theorem 7.58. To do this we prove a lower bound for $\int_0^1 |D_N(t)|\,dt$ that matches, up to a multiplicative constant, the upper bounds obtained in Lemma 7.73.

7.77. LEMMA. *For $N \in \mathbb{N}_0$,*

$$\int_0^1 |D_N(x)|\,dx > \tfrac{4}{\pi^2}\log(N+1).$$

PROOF. As in the proof of Lemma 7.73,

$$\int_0^1 |D_N(x)|\,dx = \frac{|\sin(2\pi(N+\frac{1}{2})x)|}{|\sin(\pi x)|}\,dx \geqslant \tfrac{2}{\pi}\int_0^{1/2} \frac{|\sin((2N+1)\pi x)|}{x}dx$$

where for the last inequality we have used $|\sin(u)| \leqslant |u|$ for $u > 0$. Changing variables $t = (2N+1)\pi x$ we see that the right hand side of this display equals

$$\tfrac{2}{\pi}\int_0^{\pi\frac{2N+1}{2}} \frac{|\sin t|}{t}\,dt = \tfrac{2}{\pi}\sum_{k=0}^{2N}\int_{\frac{\pi}{2}k}^{\frac{\pi}{2}(k+1)} \frac{|\sin t|}{t}\,dt.$$

We have

$$\int_{k\frac{\pi}{2}}^{(k+1)\frac{\pi}{2}} \frac{|\sin t|}{t}\,dt \geqslant \tfrac{2}{\pi(k+1)}\int_{k\frac{\pi}{2}}^{(k+1)\frac{\pi}{2}} |\sin t|\,dt = \tfrac{2}{\pi(k+1)} \geqslant \tfrac{2}{\pi}\int_{k+1}^{k+2}\frac{1}{t}\,dt,$$

and hence

$$\tfrac{2}{\pi}\int_0^{\pi\frac{2N+1}{2}} \frac{|\sin(t)|}{t}\,dt > \tfrac{4}{\pi^2}\int_1^{2N+2}\frac{1}{t}\,dt = \tfrac{4}{\pi^2}\big(\log(N+1)+\log 2\big). \qquad \square$$

7.78. EXERCISE. The integrals $\|D_N\|_1 = \int_0^1 \frac{|\sin((2N+1)\pi x)|}{|\sin(\pi x)|}\,dx$ are referred to as *Lebesgue constants* for the partial sum operators. The lower bound in Lemma 7.77 can be upgraded to an asymptotic, thus improving on the upper bound in Lemma 7.73. Prove that there exists $(a_N)_N$ such that for $N \geqslant 1$,

$$\tfrac{4}{\pi^2}\log N \leqslant \|D_N\|_1 \leqslant (\tfrac{4}{\pi^2}+a_N)\log N \quad \text{ and } \lim_{N\to\infty} a_N = 0.$$

7.79. LEMMA. *Fix $x_0 \in \mathbb{R}$ and define a linear map $T_N = T_{N,x_0} : C(\mathbb{T}) \to \mathbb{C}$ by*

$$T_{N,x_0} f = S_N f(x_0).$$

For every $N \in \mathbb{N}$, $x_0 \in \mathbb{R}$, $T_{N,x_0} : C(\mathbb{T}) \to \mathbb{C}$ is a bounded linear functional and

$$\|T_{N,x_0}\|_{\mathrm{op}} = \int_0^1 |D_N(t)|dt.$$

PROOF. For every $f \in C(\mathbb{T})$ we have

$$|T_N f| = |f * D_N(x_0)| \leqslant \int_0^1 |f(x_0-t)D_N(t)|\,dt \leqslant \|f\|_\infty \int_0^1 |D_N(t)|\,dt.$$

Therefore, T_N is bounded and $\|T_N\|_{\mathrm{op}} \leqslant \int_0^1 |D_N|dt$. To prove the lower bound we test the functional T_N on

$$f_0(x) = \operatorname{sign}(D_N(x_0-x))$$

where $\operatorname{sign}(a)$ takes the values $1, 0, -1$ if $a > 0$, $a = 0$, $a < 0$, respectively. While f_0 is not a continuous function, it can be approximated by continuous functions: given $\varepsilon > 0$, Theorem 2.25 provides us with a function $g \in C(\mathbb{T})$ such that $\|g\|_\infty \leqslant 1$ and

$$\int_0^1 |f_0(t) - g(t)|dt \leqslant \frac{\varepsilon}{2N+1}.$$

We have

$$|T_N f_0| = |f_0 * D_N(x_0)| = \Big| \int_0^1 \operatorname{sgn}(D_N(t)) D_N(t) dt \Big| = \int_0^1 |D_N(t)|dt$$

Moreover,

$$|T_N g| \geqslant |T_N f_0| - |T_N(f_0 - g)|.$$

The error term $|T_N(f_0 - g)|$ can be estimated as follows:

$$\begin{aligned} |T_N(f_0 - g)| &\leqslant \int_0^1 |D_N(x_0 - t)||f_0(t) - g(t)|\, dt \\ &\leqslant \|D_N\|_\infty \int_0^1 |f_0(t) - g(t)|\, dt \leqslant (2N+1)\frac{\varepsilon}{2N+1} = \varepsilon, \end{aligned}$$

and so

$$\|T_N\|_{\mathrm{op}} \geqslant |T_N g| \geqslant \|D_N\|_1 - \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this implies $\|T_N\|_{\mathrm{op}} \geqslant \|D_N\|_1$. □

The same analysis gives the operator norm for S_N as a bounded operator from $C(\mathbb{T})$ to $C(\mathbb{T})$. Also it will later be useful to work with the subspace of even 1-periodic continuous functions denoted by $C^{\mathrm{ev}}(\mathbb{T})$. We ask the reader to check that convolutions of even functions are even, and, since D_N is an even function, S_N maps $C^{\mathrm{ev}}(\mathbb{T})$ to $C^{\mathrm{ev}}(\mathbb{T})$.

7.80. Lemma. *The operator $S_N : C(\mathbb{T}) \to C(\mathbb{T})$ has operator norm $\int_0^1 |D_N(x)|dx$. Moreover $S_N : C^{\mathrm{ev}}(\mathbb{T}) \to C^{\mathrm{ev}}(\mathbb{T})$ has the same operator norm.*

For this consult the proof of Lemma 7.79 and check that upper and lower bounds do not depend on x_0. Taking $x_0 = 0$ and using the evenness of D_N one also gets the second statement for the space $C^{\mathrm{ev}}(\mathbb{T})$.

Combining Lemma 7.77 and Lemma 7.79 we can now prove Theorem 7.76 as a direct consequence of the uniform boundedness principle (Theorem 5.35). Indeed, for every $x_0 \in \mathbb{R}$,

$$\|T_{N,x_0}\|_{\mathrm{op}} = \|D_N\|_1 \geqslant \frac{4}{\pi^2}\log(N) \tag{7.25}$$

and therefore

$$\sup_{N\in\mathbb{N}} \|T_{N,x_0}\|_{\mathrm{op}} = \infty.$$

By Theorem 5.35 there must exist an $f \in C(\mathbb{T})$ such that

$$\sup_{N\in\mathbb{N}} |T_{N,x_0} f| = \infty.$$

In other words, $(S_N f(x_0))_N$ is an unbounded sequence.

Remarks. (i) Continuous functions with divergent Fourier series can also be constructed explicitly. Indeed, the constructive proof of the uniform boundedness principle (Theorem 5.35) can be used to construct a concrete $f \in C(\mathbb{T})$ such that $\sup_N |S_N f(0)| = \infty$. For the exposition of a more explicit construction due to Dubois-Reymond see e.g. [**74**, Ch. 3, §2.2.].

(ii) The conclusion of Theorem 7.76 can be strengthened significantly: for every Lebesgue null set $A \subset \mathbb{T}$ there exists a continuous function whose Fourier series diverges on A. This is due to Kahane and Katznelson [**34**]. This is as far as it can be pushed: Carleson [**10**] proved that Fourier series of continuous functions converge *almost everywhere*, i.e. for every $f \in C(\mathbb{T})$ there is a Lebesgue null set A

such that $(S_N f(x))_N$ converges for every $x \in \mathbb{T}\backslash A$. Lacey and Thiele [**45**] gave an elegant alternative proof of Carleson's theorem.

7.5.3. Convergence at points of differentiability. While the Fourier series of a continuous function does not necessarily converge at every point, one can prove convergence at a specific point under an additional differentiability assumption.

7.81. THEOREM. *Let f be Riemann integrable and 1-periodic and let $x_0 \in \mathbb{R}$. Assume that f is differentiable at x_0. Then $S_N f(x_0) \to f(x_0)$ as $N \to \infty$.*

PROOF. By definition,

$$S_N f(x_0) = \int_0^1 f(x_0 - t) D_N(t)\, dt.$$

Also,

$$\int_0^1 D_N(t)\, dt = \sum_{n=-N}^{N} \int_0^1 e^{2\pi i n t}\, dt = 1.$$

Thus from Lemma 7.55,

$$\begin{aligned} S_N f(x_0) - f(x_0) &= \int_0^1 (f(x_0 - t) - f(x_0)) D_N(t) dt \\ &= \int_0^1 g(t) \sin(2\pi(N + \tfrac{1}{2})t) dt, \end{aligned} \tag{7.26}$$

where

$$g(t) = \frac{f(x_0 - t) - f(x_0)}{\sin(\pi t)}$$

for $t \notin \mathbb{Z}$. The differentiability of f at x_0 implies that $\lim_{t\to 0} g(t)$ exists and if we define $g(0) = g(1)$ as this limit we obtain a function that is continuous at x_0. Indeed,

$$\frac{f(x_0 - t) - f(x_0)}{\sin(\pi t)} = \frac{f(x_0 - t) - f(x_0)}{t} \frac{t}{\sin(\pi t)} \to f'(x_0) \frac{1}{\pi}$$

as $t \to 0$. Since f is Riemann integrable it follows that the set of discontinuities of g forms a Lebesgue null set and by Lebesgue's theorem (Theorem 4.129) we see that g is Riemann integrable and 1-periodic as well. Now splitting

$$\sin(2\pi(N + \tfrac{1}{2})t) = (2i)^{-1}(e^{2\pi i N t} e^{\pi i t} - e^{-2\pi i N t} e^{-\pi i t})$$

we see that $S_N f(x_0) - f(x_0)$ can be expressed in terms of Fourier coefficients of the functions $h_1(t) = g(t)e^{i\pi t}$ and $h_2(t) = g(t)e^{-i\pi t}$, namely

$$S_N f(x_0) - f(x_0) = (2i)^{-1}\big(\widehat{h_1}(N) - \widehat{h_2}(-N)\big).$$

By the Riemann-Lebesgue lemma, $\lim_{N\to\infty} \widehat{h_1}(N) = 0$, $\lim_{N\to\infty} \widehat{h_2}(-N) = 0$. Hence, $S_N f(x_0) \to f(x_0)$ as $N \to \infty$. □

7.82. EXERCISE. Let f be the 1-periodic function such that $f(x) = x$ for $x \in [0, 1)$. Compute the Fourier coefficient $\widehat{f}(n)$ for every $n \in \mathbb{Z}$ and use it to prove the formulas

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \qquad \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} = \frac{\pi^2}{8}.$$

7.83. EXERCISE. Show that

$$\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}.$$

7.84. Exercise. Let $\psi_1(x) = x - \frac{1}{2}$ for $x \in (0,1)$. Prove that

$$\psi_1(x) = -\frac{1}{\pi}\sum_{k=1}^{\infty}\frac{\sin(2\pi kx)}{k}, \quad x \in (0,1).$$

7.85. Exercise. Consider the 1-periodic Bernoulli functions

$$\psi_n(x) = b_n(x - \lfloor x \rfloor)$$

with the Bernoulli polynomials b_n defined recursively by (3.27). Starting with the result of Exercise 7.84 and integrating prove that for $\ell \geqslant 1$

$$\psi_{2\ell}(x) = (-1)^{\ell+1}\frac{2(2\ell)!}{(2\pi)^{2\ell}}\sum_{k=1}^{\infty}\frac{\cos(2\pi kx)}{k^{2\ell}}$$

$$\psi_{2\ell+1}(x) = (-1)^{\ell+1}\frac{2(2\ell+1)!}{(2\pi)^{2\ell+1}}\sum_{k=1}^{\infty}\frac{\sin(2\pi kx)}{k^{2\ell+1}}.$$

Let $B_n = b_n(0)$ be the Bernoulli numbers. Deduce that $0 = B_3 = B_5 = B_7 = \ldots$,

$$\sum_{k=1}^{\infty}\frac{1}{k^{2\ell}} = (-1)^{\ell+1}B_{2\ell}\frac{2^{2\ell-1}\pi^{2\ell}}{(2\ell)!}, \quad \ell = 1,2,\ldots$$

and that $B_{2\ell} > 0$ if ℓ is odd and $B_{2\ell} < 0$ if ℓ is even.

7.5.4. Absolute convergence. Let $A(\mathbb{T})$ be the space of all absolutely convergent Fourier series, i.e. all functions $f \in C(\mathbb{T})$ that can be written as $f(x) = \sum_{n\in\mathbb{Z}} a_n e^{2\pi inx}$ where $\sum_{n=-\infty}^{\infty}|a_n| < \infty$. Thus $A(\mathbb{T})$ is the range of the map

$$\mathcal{F}_{\mathbb{Z}} : \begin{cases} \ell^1 \to C(\mathbb{T}) \\ \mathcal{F}_{\mathbb{Z}}a(x) = \sum_{n=-\infty}^{\infty} a_n e^{2\pi inx} \end{cases}$$

Note that $\mathcal{F}_{\mathbb{Z}}$ is injective. Indeed, if $a = (a_n)_{n=-\infty}^{\infty} \in \ell^1$ and $f(x) = \sum_{n=-\infty}^{\infty} a_n e^{2\pi inx}$, then we can recover the a_n as the Fourier coefficients of f, $\widehat{f}(n) = a_n$. We are first interested in sufficient conditions on a function f to place it into the class $A(\mathbb{T})$.

For $f \in C(\mathbb{T})$ and $h > 0$ define

$$\Delta_h f(x) = f(x+h) - f(x-h).$$

7.86. Theorem. *Let $f \in C(\mathbb{T})$ satisfy $\sum_{k\geqslant 0} 2^{k/2}\|\Delta_{2^{-k}}f\|_2 < \infty$. Then $f \in A(\mathbb{T})$.*

Proof. First, $|\widehat{f}(n)| \leqslant \int_0^1 |f(x)|dx \leqslant \|f\|_2$ by the Cauchy-Schwarz inequality. We estimate, for $k \geqslant 3$ the dyadic blocks

$$A_k = \sum_{2^{k-3}\leqslant|n|<2^{k-2}} |\widehat{f}(n)|.$$

Note that $2\pi n2^{-k} \in [\pi/4, \pi/2]$ for $2^{k-3} \leqslant |n| < 2^{k-2}$ and thus $\sin(2\pi n2^{-k}) \geqslant 2^{-1/2}$ in this range. Moreover,

$$\widehat{f}(n)\sin(2\pi n2^{-k}) = \frac{1}{2i}\widehat{\Delta_{2^{-k}}f}(n).$$

Hence, by the Cauchy-Schwarz inequality and a subsequent application of Bessel's inequality

$$\begin{aligned} A_k &= \sum_{2^{k-3}\leqslant|n|<2^{k-2}} |\widehat{f}(n)| \leqslant 2^{(k-2)/2}\Big(\sum_{2^{k-3}\leqslant|n|<2^{k-2}} |\widehat{f}(n)|^2\Big)^{1/2} \\ &\leqslant 2^{(k-2)/2}\Big(\sum_{2^{k-3}\leqslant|n|<2^{k-2}} |\widehat{f}(n)2^{1/2}\sin(2\pi n2^{-k})|^2\Big)^{1/2} \leqslant 2^{(k-3)/2}\|\Delta_{2^{-k}}f\|_2. \end{aligned}$$

Summing over k this implies $\sum_{n\in\mathbb{Z}} |\widehat{f}(n)| < \infty$. □

An immediate corollary can be formulated for Hölder continuous functions $f \in C^{0,\alpha}(\mathbb{T})$, consisting of the 1-periodic functions satisfying $\omega_f(\delta) = O(\delta^\alpha)$.

7.87. Corollary. *Let $f \in C(\mathbb{T})$.*

(i) Suppose that $\sum_{k\geqslant 0} 2^{k/2}\|\Delta_{2^{-k}} f\|_\infty < \infty$. Then $f \in A(\mathbb{T})$.

(ii) Let $\varepsilon > 0$. Suppose that $\omega_f(\delta) = O(\delta^{1/2}(\log \frac{1}{\delta})^{-1-\varepsilon})$. Then $f \in A(\mathbb{T})$.

(iii) For $\alpha > \frac{1}{2}$, $C^{0,\alpha}(\mathbb{T}) \subset A(\mathbb{T})$.

Proof. Note that $\|\Delta_h f\|_2 \leqslant \|\Delta_h f\|_\infty$, and so (i) follows from Theorem 7.86. If $\omega_f(\delta) = O(\delta^\alpha(\log \frac{1}{\delta})^{-1-\varepsilon})$ then

$$\sum_{k\geqslant 0} 2^{k/2}\|\Delta_{2^{-k}} f\|_\infty \leqslant 2\sum_{k\geqslant 0} 2^{k/2}\omega_f(2^{-k}) \leqslant C\sum_{k\geqslant 0} k^{-1-\varepsilon} < \infty$$

and so (ii) follows from (i). If $f \in C^{0,\alpha}$ with $\alpha > 1/2$, the condition in (ii) is satisfied (in fact for all $\varepsilon > 0$) and we get (iii). □

Absolute convergence of the Fourier series implies uniform convergence. On the other hand, a uniformly convergent Fourier series may fail to converge absolutely . The reader may have noticed the gap between the assumptions needed in part (i) of Corollary 7.87 and the assumption $\|f(\cdot+h)-f\|_\infty = O(|h|^\varepsilon)$ which implies uniform convergence (*cf.* Theorem 7.74). We will now show that this gap cannot be closed and prove that Corollary 7.87 is sharp.

A useful tool for this is given by two sequences P_ℓ, Q_ℓ of complex polynomials of degree $2^\ell - 1$ which have the property that when restricted to the unit circle the maximums and L^2 norms are roughly equivalent.

7.88. Definition. The *Rudin–Shapiro polynomials* P_ℓ and the Rudin-Shapiro complementary polynomials Q_ℓ are defined by the following recursive relations.

$$P_0(x) = 1,\ Q_0(x) = 1$$
$$P_{\ell+1}(z) = P_\ell(z) + z^{2^\ell} Q_\ell(z)$$
$$Q_{\ell+1}(z) = P_\ell(z) - z^{2^\ell} Q_\ell(z)$$

The trigonometric versions of the Rudin–Shapiro and complementary polynomials are given by

$$p_\ell(x) = P_\ell(e^{2\pi i x}), \quad q_\ell(x) = Q_\ell(e^{2\pi i x}).$$

7.89. Lemma. *There exist sequences $(\alpha_n)_{n=0}^\infty$, $(\beta_n)_{n=0}^\infty$ such that $\alpha_n, \beta_n \in \{1, -1\}$ and*

(i) $P_\ell(z) = \sum_{n=0}^{2^\ell-1} \alpha_n z^n$

(ii) $Q_\ell(z) = \sum_{n=0}^{2^\ell-1} \beta_n z^n$.

(iii) For $|z| = 1$, $|P_\ell(z)|^2 + |Q_\ell(z)|^2 = 2^{\ell+1}$.

Proof. We prove these assertions by induction. The assertions are true for $\ell = 0$ and we have $\alpha_0 = 1 = \beta_0$. Assuming that the formulas in (i), (ii) hold for P_ℓ, Q_ℓ we obtain

$$P_{\ell+1}(z) = P_\ell(z) + z^{2^\ell} Q_\ell(z) = \sum_{n=0}^{2^\ell-1} \alpha_n z^n + \sum_{n=2^\ell}^{2^{\ell+1}-1} \beta_{n-2^\ell} z^n$$

and likewise

$$Q_{\ell+1}(z) = \sum_{n=0}^{2^\ell-1} \alpha_n z^n - \sum_{n=2^\ell}^{2^{\ell+1}-1} \beta_{n-2^\ell} z^n$$

so if we define for $2^\ell \leqslant n \leqslant 2^{\ell+1} - 1$, $\alpha_n = \beta_{n-2^\ell}$, $\beta_n = -\beta_{n-2^\ell}$ we get the asserted formula

For (iii) observe, again assuming the formula for P_ℓ, Q_ℓ,

$$\begin{aligned}|P_{\ell+1}(z)|^2 + |Q_{\ell+1}(z)|^2 &= |P_\ell(z) + z^{2^\ell} Q_\ell(z)|^2 + |P_\ell(z) - z^{2^\ell} Q_\ell(z)|^2 \\ &= 2|P_\ell(z)|^2 + 2|z|^{2^{\ell+1}} |Q_\ell(z)|^2\end{aligned}$$

as the two mixed terms $\pm 2\operatorname{Re}(P_\ell(z) z^{2^\ell} Q_\ell(z))$ drop out. Hence for $|z| = 1$,

$$|P_{\ell+1}(z)|^2 + |Q_{\ell+1}(z)|^2 = 2 \cdot 2^{\ell+1} = 2^{\ell+2},$$

by the induction hypothesis. □

7.90. Definition. The sequence α_n in Lemma 7.89 is called the *Golay–Rudin–Shapiro* sequence. For an explicit formula see Exercise 7.178.

7.91. Theorem. *There exist functions $f \in C(\mathbb{T}) \backslash A(\mathbb{T})$ such that*

$$\Big(\sum_{k=0}^{\infty} \big[2^{k/2} \|\Delta_{2^{-k}} f\|_\infty \big]^q \Big)^{1/q} < \infty \tag{7.27}$$

for all $q > 1$, in particular $f \in C^{0,\frac{1}{2}}(\mathbb{T})$.

Proof. Set p_ℓ as in Definition 7.88 and

$$g_\ell = p_\ell - p_{\ell+1}.$$

Thus $g_\ell \in \mathcal{T}_{2^{\ell+1}-1}$. From Lemma 7.89 we get $\|p_\ell\|_\infty \leqslant 2^{(\ell+1)/2}$ and therefore $\|g_\ell\|_\infty \leqslant 4 \cdot 2^{\ell/2}$. Define

$$f(x) = \sum_{\ell=1}^{\infty} 2^{-\ell} (\ell+1)^{-1} g_\ell(x).$$

We have uniform convergence of the series so that $f \in C(\mathbb{T})$. For the Fourier coefficients of f we get

$$\widehat{f}(n) = 2^{-\ell}(\ell+1)^{-1}\alpha_n, \text{ if } 2^\ell \leqslant n \leqslant 2^{\ell+1} - 1$$

and $\widehat{f}(n) = 0$ if $n < 2$. Since $|\alpha_n| = 1$ we get $\sum_{k=2^\ell}^{2^{\ell+1}-1} |\widehat{f}(n)| = (\ell+1)^{-1}$ which leads to $\sum_n |\widehat{f}(n)| = \infty$ and thus $f \notin A(\mathbb{T})$.

We show that (7.27) holds for all $q > 1$. For every k we decompose f into a high frequency part which involves Fourier coefficients $\widehat{f}(n)$ for $n \geqslant 2^k$ and a low frequency part, with the Fourier coefficients for $2 \leqslant n < 2^k$. That is, for every $k > 0$ we have $f = f^{k,\text{high}} + f^{k,\text{low}}$ where

$$f^{k,\text{high}} = \sum_{m=0}^{\infty} \frac{2^{-k-m}}{k+m+1} g_{k+m}(x), \quad f^{k,\text{low}} = \sum_{l=1}^{k} \frac{2^{-k+l}}{k-l+1} g_{k-l}(x).$$

By Minkowski's inequality, $(\sum_{k=0}^{\infty} [2^{\frac{k}{2}} \|\Delta_{2^{-k}} f\|_\infty]^q)^{\frac{1}{q}} \leqslant I + II$ where

$$I = \Big(\sum_{k=0}^{\infty} \big[2^{k/2} \|\Delta_{2^{-k}} f^{k,\text{high}}\|_\infty \big]^q \Big)^{1/q}, \qquad II = \Big(\sum_{k=0}^{\infty} \big[2^{k/2} \|\Delta_{2^{-k}} f^{k,\text{low}}\|_\infty \big]^q \Big)^{1/q}.$$

We use the bound $\|\Delta_{2^{-k}} g\|_\infty \leqslant 2\|g\|_\infty$ and Minkowski's inequality (in the form of the triangle inequality in ℓ^q for infinite sums) to estimate

$$\begin{aligned} I &\leqslant \sum_{m=0}^{\infty} \Big(\sum_{k=0}^{\infty} \big[2^{k/2} \frac{2^{-k-m}}{k+m+1} 2 \|g_{k+m}\|_\infty \big]^q \Big)^{1/q} \\ &\leqslant 2^3 \sum_{m=0}^{\infty} \Big(\sum_{k=0}^{\infty} \big[2^{k/2} \frac{2^{-k-m}}{k+m+1} 2^{\frac{k+m}{2}} \big]^q \Big)^{1/q} \leqslant 2^3 \sum_{m=0}^{\infty} 2^{-m/2} \Big(\sum_{k=0}^{\infty} (k+m+1)^{-q} \Big)^{1/q} \leqslant C(q) \end{aligned}$$

with $C(q) > 1$ if $q > 1$. For the term II we use the mean value theorem and then the Bernstein-type Proposition 7.70 to estimate

$$\begin{aligned}\|\Delta_{2^{-k}} g_{k-m}\|_\infty &\leqslant 2^{1-k}\|g'_{k-m}\|_\infty \\ &\leqslant 2^6 2^{-k}2^{k-m}\|g_{k-m}\|_\infty \leqslant 2^8 2^{-m}2^{\frac{k-m}{2}}.\end{aligned}$$

Consequently,

$$\begin{aligned}II &\leqslant \sum_{m=1}^{\infty}\Big(\sum_{k=m}^{\infty}\big[2^{k/2}\|\Delta_{2^{-k}}\frac{2^{-k+m}}{k-m+1}g_{k-m}\|_\infty\big]^q\Big)^{1/q} \leqslant \\ &\leqslant 2^8\sum_{m=1}^{\infty}\Big(\sum_{k=m}^{\infty}\big[2^{k/2}\frac{2^{-k+m}}{k-m+1}2^{-m}2^{\frac{k-m}{2}}\big]^q\Big)^{1/q} \leqslant 2^8\sum_{m=1}^{\infty}2^{-m/2}\Big(\sum_{k=m}^{\infty}(k-m+1)^{-q}\Big)^{1/q} \leqslant C(q).\end{aligned}$$

It follows that f satisfies (7.27) for $q > 1$. This also implies $\|\Delta_{2^{-k}} f\|_\infty = O(2^{-k/2})$ and thus $f \in C^{0,\frac{1}{2}}(\mathbb{T})$. □

Inspecting the proof of Theorem 7.91 one can also obtain the following variant.

7.92. Exercise. Show that there exists an $f \in C(\mathbb{T})\backslash A(\mathbb{T})$ such that

$$\omega_f(\delta) \leqslant C\delta^{1/2}(\log\tfrac{1}{\delta})^{-1}$$

for all $\delta \in (0, \frac{1}{2})$.

7.6. Best uniform approximation by polynomials

Denote by $\mathcal{P}_N^{\mathbb{R}}$ the vector space of *real* polynomials of degree at most N. We are asking for the best approximation of a real-valued function $f \in C([a,b];\mathbb{R})$ by a polynomial in $\mathcal{P}_N^{\mathbb{R}}$ in the supremum norm $\|g\|_\infty = \sup_{t\in[a,b]}|g(t)|$. Set

$$E_N(f) = \operatorname{dist}(f, \mathcal{P}_N^{\mathbb{R}}) = \inf_{q\in\mathcal{P}_N^{\mathbb{R}}}\|f-q\|_\infty .$$

Our goal in this section is to prove the following theorem about existence and *uniqueness* of a best approximation.

7.93. Theorem. *For every $f \in C([a,b];\mathbb{R})$ there exists a unique $p \in \mathcal{P}_N^{\mathbb{R}}$ such that*

$$E_N(f) = \|f-p\|_\infty.$$

The existence of such a p has already been proved (Proposition 7.1). To establish uniqueness we will first need to get some structural properties for the error function $\mathcal{E} = f - p$ which were proved by Chebyshev. The simplest of these is:

7.94. Lemma. *Let $f \in C([a,b];\mathbb{R})$, $p \in \mathcal{P}_N^{\mathbb{R}}$ such that $\|f-p\|_\infty = E_N(f)$ and $\mathcal{E} = f - p$. Then there exist $t_1, t_2 \in [a,b]$ with $t_1 \neq t_2$ such that*

(i) $|\mathcal{E}(t_1)| = |\mathcal{E}(t_2)| = E_N(f)$
(ii) $\mathcal{E}(t_1) = -\mathcal{E}(t_2)$.

Proof. The existence of t_1 with $|\mathcal{E}(t_1)| = E_N(f)$ follows since $|f-p|$ is a continuous function, and without loss of generality we may assume $\mathcal{E}(t_1) > 0$ (otherwise consider $-f$ in place of f). We argue by contradiction and assume that $\mathcal{E}(t) > -E_N(f)$ for all $t \in [a,b]$. Let $\mu = \min_{[a,b]}\mathcal{E}(t)$ so that $\mu > -E_N(f)$ and set $c = \mu + E_N(f) > 0$. Then $-E_N(f) + c \leqslant f(t) - p(t) \leqslant E_N(f)$ for all $t \in [a,b]$. The polynomial $q(t) = p(t) + \frac{c}{2}$ satisfies

$$-E_N(f) + \tfrac{c}{2} \leqslant f(t) - q(t) \leqslant E_N(f) - \tfrac{c}{2}$$

and hence $\|f-q\|_\infty < E_N(f)$. But since $q \in \mathcal{P}_N^{\mathbb{R}}$ this is a contradiction. □

We need a more sophisticated version of this lemma, for which we make a definition.

7.95. Definition. Let $\mathcal{E} \in C([a,b];\mathbb{R})$ and $\ell \geqslant 2$. A set of ℓ ordered points

$$t_0 < t_1 < \cdots < t_{\ell-1}$$

belonging to $[a,b]$ is called an *ℓ-alternant* for $\mathcal{E}$, if

(i) $|\mathcal{E}(t_j)| = \|\mathcal{E}\|_\infty$ for $j = 0, \ldots, \ell-1$,
(ii) $\mathcal{E}(t_i) = -\mathcal{E}(t_{i+1})$ for $i = 0, \ldots, \ell-2$.

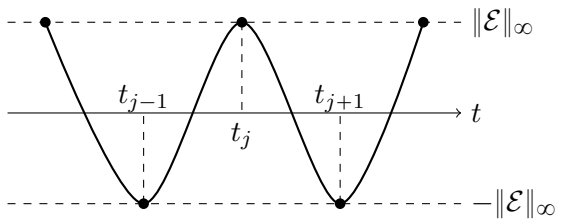


Figure 7.4. Illustration of ℓ-alternants.

7.96. Theorem (Chebyshev's alternant theorem). *Let $f \in C([a,b];\mathbb{R})$. A polynomial $p \in \mathcal{P}_N^{\mathbb{R}}$ is a best approximation to f from $\mathcal{P}_N^{\mathbb{R}}$ if and only if there exists an $(N+2)$-alternant for the error function $\mathcal{E} = f - p$.*

Proof. ($\Leftarrow$) Let $t_0 < \cdots < t_{N+1}$ be the points of an $(N+2)$-alternant for $\mathcal{E} = f - p$ and assume that p is *not* a best approximation. Then there is $q \in \mathcal{P}_N^{\mathbb{R}}$ such that

$$|f(t_i) - q(t_i)| \leqslant \|f - q\|_\infty < \|f - p\|_\infty = |f(t_i) - p(t_i)|, \quad i = 0, \ldots, N+1.$$

We consider $q - p \in \mathcal{P}_N^{\mathbb{R}}$. Since $q - p = (f - p) - (f - q)$ we see from the display that $q(t_i) - p(t_i)$ has the same sign as $f(t_i) - p(t_i)$. Since the sign of $f(t_i) - p(t_i)$ is opposite to the sign of $f(t_{i+1}) - p(t_{i+1})$ we see that $q - p$ has at least $N+1$ zeros which means $q - p = 0$. This gives a contradiction.

($\Rightarrow$) Assume that $p \in \mathcal{P}_N^{\mathbb{R}}$ is a best approximation. If $E_N(f) = 0$, then every choice of $N+2$ points in $[a,b]$ is an $(N+2)$-alternant, so assume that $E_N(f) > 0$. Let

$$t_0 = \min\{t \in [a,b] \,:\, |\mathcal{E}(t)| = E_N(f)\}.$$

Without loss of generality we may assume that $\mathcal{E}(t_0) > 0$ (otherwise consider $-f$ instead of f). By Lemma 7.94, $f - p$ has a 2-alternant. We argue by contradiction and assume that there exists no $(N+2)$-alternant. Then let $k \in \{1, \ldots, N\}$ such that $f - p$ has a $(k+1)$-alternant but not a $(k+2)$-alternant.

Form a partition $a = t_0 < t_1 < \cdots < t_M = b$ of $[a,b]$ such that for each partition interval

$$|\mathcal{E}(s) - \mathcal{E}(t)| < \tfrac{1}{10} E_N(f), \text{ for } s,t \in [t_{i-1}, t_i], \quad i = 1, \ldots, M. \tag{7.28}$$

We single out all partition intervals which contain a point t with $|\mathcal{E}(t)| = E_N(f)$. Label such an interval as a $+$interval if it contains a point t with $\mathcal{E}(t) = E_N(f)$ and as a $-$interval if it contains a point t with $\mathcal{E}(t) = -E_N(f)$. By (7.28) no interval is both a $+$interval and a $-$interval. Moreover, also by (7.28), the two neighboring intervals of a $+$interval cannot be $-$intervals and the two neighboring intervals of a $-$interval cannot be $+$intervals.

We label these $\pm$intervals as I_ν with the labeling indicating that I_ν is to the left of $I_{\nu+1}$ for all ν. We sort these $\pm$intervals into collections of intervals $\mathcal{G}_1, \ldots, \mathcal{G}_m$ with the first group $\mathcal{G}_1 = \{I_1, \ldots, I_{\kappa_1}\}$ consisting of $+$intervals, the second group $\mathcal{G}_2 = \{I_{\kappa_1+1}, \ldots, I_{\kappa_2}\}$ consisting of $-$intervals, etc. That is, $\mathcal{G}_j = \{I_{\kappa_{j-1}+1}, \ldots, I_{\kappa_j}\}$ consists of $+$intervals if j is odd, and consists of $-$intervals if j is even (and $\kappa_0 = 0$). Since $I_{\kappa_j} \cap I_{\kappa_j+1} = \varnothing$ we may choose $\xi_j \in [a,b]$ with

$$\max I_{\kappa_j} < \xi_j < \min I_{\kappa_j+1}.$$

Note that $1 \leqslant k \leqslant m \leqslant N$. Define

$$R(t) = \prod_{j=1}^{m} (\xi_j - t).$$

Then $R \in \mathcal{P}_N^{\mathbb{R}}$ and the zeros of R are the points ξ_j. Also, $R(t) > 0$ if t belongs to an interval in $\mathcal{G}_j$ with j odd, and $R(t) < 0$ if t belongs to an interval in $\mathcal{G}_j$ with j even. This means that R and $\mathcal{E}$ have the same sign in all intervals that belong to $\bigcup_{j=1}^{m} \mathcal{G}_j$.

We now argue as in Lemma 7.94 and aim to prove that for sufficiently small $\delta > 0$ the polynomial $q = p + \delta R$ satisfies $\|f - q\|_\infty < E_N(f)$ which will give a contradiction (by the definition of best approximation).

First note that $\max_{[t_{i-1},t_i]} \mathcal{E}(t) < E_N(f)$ if the partition interval $[t_{i-1}, t_i]$ is not a $+$interval and not a $-$interval. Thus there exists a $\rho > 0$ such that

$$|\mathcal{E}(t)| \leqslant E_N(f) - \rho \text{ for all } t \in \overline{[a,b] \setminus \cup_\nu I_\nu}.$$

Let $\delta > 0$ satisfy

$$\delta < \tfrac{1}{\|R\|_\infty} \min(\rho, \tfrac{1}{10} E_N(f))$$

On any $\pm$interval we have $|\mathcal{E}(t)| \geqslant \frac{9}{10} E_N(f) \geqslant 9\delta \|R\|_\infty$. Since $\mathcal{E}(t)$ and $R(t)$ have the same sign on each $\pm$interval, we get $|\mathcal{E}(t) - \delta R(t)| = |\mathcal{E}(t)| - \delta |R(t)|$ there.

Since the zeros of $R(t)$ do not lie in any of the intervals I_ν there exists an $\alpha > 0$ such that $|R(t)| \geqslant \alpha$ for all $t \in \cup_\nu I_\nu$, and thus for those t

$$\begin{aligned} |f(t) - q(t)| &= |f(t) - p(t) - \delta R(t)| = |\mathcal{E}(t) - \delta R(t)| \\ &= |\mathcal{E}(t)| - \delta |R(t)| \leqslant E_N(f) - \delta\alpha. \end{aligned}$$

On the other hand for $t \in [a,b] \setminus \cup_\nu I_\nu$,

$$|f(t) - q(t)| = |f(t) - p(t) - \delta R(t)| \leqslant |\mathcal{E}(t)| + \delta |R(t)| \leqslant E_N(f) - \rho + \delta \|R\|_\infty.$$

Hence,

$$\|f - q\|_\infty \leqslant E_N(f) - \min\left\{\delta\alpha, \rho - \delta\|R\|_\infty\right\} < E_N(f)$$

which gives the desired contradiction, so $\mathcal{E}$ must have an $(N+2)$-alternant. □

We now apply the alternant theorem to prove the uniqueness of the best approximation:

Proof of Theorem 7.93. Existence follows from Proposition 7.1. For the uniqueness let p and q be two best approximations to f from $\mathcal{P}_N^{\mathbb{R}}$. Let $r(t) = \frac{p+q}{2}$. Then

$$\|f - r\|_\infty = \|\tfrac{1}{2}(f-p) + \tfrac{1}{2}(f-q)\|_\infty \leqslant \tfrac{1}{2}\|f-p\|_\infty + \tfrac{1}{2}\|f-q\|_\infty = E_N(f)$$

which establishes r as a best approximation to f from $\mathcal{P}_N^{\mathbb{R}}$. By Theorem 7.96 there is an $(N+2)$-alternant $\{t_0, \ldots, t_{N+1}\}$ for $f - r$. We have for each t_i

$$\begin{aligned} E_N(f) = |f(t_i) - r(t_i)| &= \left|\tfrac{1}{2}\big(f(t_i) - p(t_i)\big) + \tfrac{1}{2}\big(f(t_i) - q(t_i)\big)\right| \\ &\leqslant \tfrac{1}{2}\big|f(t_i) - p(t_i)\big| + \tfrac{1}{2}\big|f(t_i) - q(t_i)\big| \leqslant E_N(f) \end{aligned}$$

showing that all terms in the last display are equal. In particular $f(t_i) - p(t_i)$ and $f(t_i) - q(t_i)$ have the same sign, and from the display and $\|f-p\|_\infty = \|f-q\|_\infty = E_N(f)$ it follows that $|f(t_i) - p(t_i)| = |f(t_i) - q(t_i)| = E_N(f)$. This together with the statement about the signs implies that

$$f(t_i) - p(t_i) = f(t_i) - q(t_i), \quad i = 0, \ldots, N+1.$$

Hence $p - q \in \mathcal{P}_N^{\mathbb{R}}$ is equal to 0 at $t_0, \ldots, t_{N+1}$ which yields $p = q$. □

REMARKS. (i) A similar approach, also using alternants, shows an existence and uniqueness result for approximation of $C(\mathbb{T};\mathbb{R})$ by real-valued trigonometric polynomials.

(ii) For other approaches to the uniqueness result, see DeVore–Lorentz [**15**, Ch.3]. In particular it can be shown that Theorem 7.93 remains true for complex-valued $f \in C([a,b])$ and uniform best approximation from the space $\mathcal{P}_N$ of polynomials of degree at most N with complex coefficients.

7.7. Chebyshev polynomials

We now consider the system of *Chebyshev polynomials* T_n which has several remarkable features. It is most convenient to define them by a recursion formula:

$$\begin{aligned} &T_0(t) = 1, \qquad T_1(t) = t \\ &T_n(t) = 2tT_{n-1}(t) - T_{n-2}(t), \quad n \geqslant 2. \end{aligned} \tag{7.29}$$

One reason why Chebyshev polynomials are useful is that they provide a relation between algebraic and trigonometric polynomials. The key to this relation is the following

7.97. LEMMA. *For all $n = 0, 1, 2, \ldots,$ and $0 \leqslant \vartheta \leqslant \pi$,*

$$T_n(\cos\vartheta) = \cos(n\vartheta)$$

and

$$T_n'(\cos\vartheta) = \frac{n\sin(n\vartheta)}{\sin\vartheta}, \quad T_n''(\cos\vartheta) = \frac{n\sin(n\vartheta)\cos\vartheta - n^2\cos(n\vartheta)\sin\vartheta}{(\sin\vartheta)^3}\,.$$

PROOF. This formula for T_n is proved by induction, and the claim is clear for $n = 0$ and $n = 1$. In order to establish the induction step we must prove

$$\cos(n\vartheta) = 2\cos\vartheta\cos((n-1)\vartheta) - \cos((n-2)\vartheta)$$

which is an immediate consequence of the relation $2\cos(a)\cos(b) = \cos(a+b) + \cos(a-b)$.

By differentiation, $-n\sin(n\vartheta) = \frac{d}{d\vartheta}[T_n(\cos\vartheta)] = T_n'(\cos\vartheta)(-\sin\vartheta)$ which leads to the asserted formula for T_n', and differentiating again we obtain the formula for T_n''. □

REMARKS. (i) Every $p \in \mathcal{P}_N$ can be written as a linear combination of the standard monomials t^k, and alternatively as a linear combination of the T_n, i.e.

$$p(t) = \sum_{k=0}^{N} a_k t^k = \sum_{n=0}^{N} b_n T_n(t).$$

We can substitute $t = \cos(2\pi x)$ where $x \in [0, 1/2]$ and we get the cosine polynomial

$$p(\cos 2\pi x) = \sum_{k=0}^{N} a_k(\cos(2\pi x))^k = \sum_{n=0}^{\infty} b_n \cos(2\pi n x),$$

by Lemma 7.97. This correspondence between (algebraic) polynomials and cosine polynomials is used to derive theorems about polynomials from (often easier) theorems about trigonometric polynomials.

(ii) In this context it is also useful to note that the functions $\vartheta \mapsto T_n'(\cos\vartheta)/n = \sin(n\vartheta)/\sin\vartheta$ behave like the Dirichlet kernel (see Lemma 7.55), in particular we have for $x \in [0,1]$

$$D_N(x) = \frac{T_{2N+1}'(\cos\pi x)}{2N+1}.$$

It will be convenient to use the scalar product

$$\langle f, g\rangle = \int_{-1}^{1} f(t)\overline{g(t)}\frac{dt}{\sqrt{1-t^2}}. \tag{7.30}$$

7.98. Proposition (Orthogonality of Chebyshev polynomials).

The Chebyshev polynomials T_n form an orthogonal system with respect to the scalar product (7.30). *Specifically for all nonnegative integers n, m,*

$$\int_{-1}^{1} T_n(t)T_m(t)\frac{dt}{\sqrt{1-t^2}} = \begin{cases} 0 & \text{if } m \neq n, \\ \pi/2 & \text{if } m = n > 0, \\ \pi & \text{if } m = n = 0. \end{cases}$$

Proof. Changing variables $t = \cos\vartheta$ shows that the claimed orthogonality property is equivalent to

$$\int_0^{\pi} \cos(n\vartheta)\,\cos(m\vartheta)\,d\vartheta = \begin{cases} 0 & \text{if } m \neq n, \\ \pi/2 & \text{if } m = n > 0, \\ \pi & \text{if } m = n = 0, \end{cases}$$

and this follows using $\cos(n\vartheta)\cos(m\vartheta) = \frac{1}{2}\big(\cos((m+n)\vartheta) - \cos((n-m)\vartheta)\big)$. □

The following lemma gives information about the zeros and extrema of the Chebyshev polynomials.

7.99. Lemma (Properties of Chebyshev polynomials). *The following properties hold.*

(i) $2^{1-n}T_n$ *is a polynomial of degree n with leading coefficient equal to* 1.
(ii) T_n *has n simple zeros in $(-1,1)$ at the points $t_k = \cos(\frac{2k+1}{2n}\pi)$ for $k = 0, \dots, n-1$.*
(iii) $|T_n(t)| \leqslant 1$ *for all $t \in [-1,1]$.*
(iv) T_n *has $n+1$ points with either a maximum or minimum in $[-1,1]$. They are given by $v_j = \cos(\frac{j}{n}\pi)$, $j = 0, \dots, n$, and $T_n(v_j) = (-1)^j$. In particular T_n has extrema at ± 1.*
(v) $T_n'(t_k) = (-1)^k n \sin(\frac{2k+1}{2n}\pi)^{-1} = (-1)^k n(1-t_k^2)^{-1/2}$.
(vi) $T_n''(t_k) = (-1)^k n t_k (1-t_k^2)^{-3/2}$.
(vii) $\max_{t\in[-1,1]} |T_n'(t)| = T_n'(1) = n^2$.

Proof. (i) is true for $n = 0, 1$ and for $n \geqslant 2$ it easily follows from the recursion relation. Regarding (ii), to see that $t_k = \cos(\frac{2k+1}{2n}\pi)$ is a zero of T_n we just need to notice that

$$T_n(t_k) = \cos(n \arccos t_k) = \cos(\frac{2k+1}{2}\pi) = 0.$$

Since we get n zeros in $(-1,1)$ these are all zeros of the nth degree polynomial T_n.

The inequality $|T_n(t)| \leqslant 1$ for $-1 \leqslant t \leqslant 1$ is immediate from $T_n(t) = \cos(n \arccos t)$ and from this it is also clear that $T_n(\cos(\frac{j}{n}\pi)) = \cos(j\pi) = (-1)^j$. These are the only extreme points since T_n' cannot have more than $n-1$ zeros in $(-1,1)$.

From Lemma 7.97 we have $T_n'(\cos\vartheta) = \frac{n\sin(n\vartheta)}{\sin\vartheta}$ for $0 < \vartheta < \pi$. By induction on n based on $\sin n\vartheta = \sin(n-1)\vartheta\cos\vartheta + \cos(n-1)\vartheta\sin\vartheta$ one verifies $|\sin(n\vartheta)| \leqslant n\sin\vartheta$ for $0 \leqslant \vartheta \leqslant \pi$. From this, $|T_n'(\cos\vartheta)| < n^2$ for $0 < \vartheta < \pi$ and $T_n'(1) = \lim_{\vartheta\to 0} n\sin(n\vartheta)/\sin\vartheta = n^2$. Similarly $T_n'(-1) = (-1)^{n-1}n^2$. Evaluating the derivatives at the zeros t_k of T_n gives

$$T_n'(t_k) = n\sin(n\tfrac{2k+1}{2n}\pi)(1-\cos^2(\tfrac{2k+1}{2n}\pi))^{-1/2} = (-1)^k\frac{n}{\sin(\frac{2k+1}{2n}\pi)}$$

and similarly we get (v), (vii). Similarly (vi) follows from the formula for T_n'' in Lemma 7.97. □

7.100. Definition. The polynomials with leading coefficient 1 (i.e. of the form $t^n + \sum_{l=0}^{n-1} b_l t^l$) are called *monic polynomials*. Let $\mathcal{P}_n^{\text{mon}}$ denote the class of monic polynomials of degree n.

Among all the n degree monic polynomials the polynomials $2^{1-n}T_n$ have minimal sup-norm on $[-1,1]$. In other words, among all node polynomials associated to n distinct nodes in $[-1,1]$ the node polynomial associated to the Chebyshev nodes $\{\cos(\frac{2j+1}{2n})\}_{j=0}^{n-1}$ has minimal sup-norm.

7.101. THEOREM (Extremal property of Chebyshev polynomials). *For $n \geqslant 1$, we have for all $P \in \mathcal{P}_n^{\mathrm{mon}}$*

$$\max_{t\in[-1,1]} |2^{1-n}T_n(t)| \leqslant \max_{t\in[-1,1]} |P(t)|$$

PROOF. By Lemma 7.99, $\max_{t\in[-1,1]} 2^{1-n}|T_n(t)| = 2^{1-n}$ (and the maximum occurs at the $n+1$ extreme points $v_k = \cos(k\pi/n)$, $k = 0, \dots, n$, including the boundary points ± 1).

We argue by contradiction and assume that there is a polynomial $P \in \mathcal{P}_n^{\mathrm{mon}}$ such that $\max_{t\in[-1,1]} |P(t)| < 2^{1-n}$ and consider $R(t) = 2^{1-n}T_n(t) - P(t)$; then $R \in \mathcal{P}_{n-1}$. Evaluating R at the v_k yields

$$R(v_k) = 2^{1-n}T_n(v_k) - P(v_k) = (-1)^k\big(2^{1-n} - (-1)^k P(v_k)\big).$$

By our assumption on P we have $|P(v_k)| < 2^{1-n}$ so that $R(v_k)$ and $R(v_{k+1})$ have opposite signs, and therefore R must have a zero in (v_k, v_{k+1}), $k = 0, \dots, n-1$. As R is of degree $\leqslant n-1$ and has n zeros we conclude that R is the zero polynomial which gives the desired contradiction. □

The Chebyshev polynomials can be used to find the best approximation of t^{n+1} on $[-1, 1]$ by polynomials in $\mathcal{P}_n$.

7.102. THEOREM. *Show that*

$$\min_{p\in\mathcal{P}_n} \max_{t\in[-1,1]} |t^{n+1} - p(t)| = 2^{-n}$$

where the unique minimizer is given by $p_*(t) = t^{n+1} - 2^{-n}T_{n+1}$.

PROOF. Let $f(t) = t^{n+1}$. Since $f(t) - p_*(t) = 2^{-n}T_{n+1}$ is monic we get $p_* \in \mathcal{P}_n$. Also $\|f - p_*\|_\infty = 2^{-n}\|T_{n+1}\|_\infty = 2^{-n}$. We show that $f - p_*$ is a minimizer, and then by Theorem 7.93 it is a unique minimizer. By Theorem 7.96 it suffices to show that $\mathcal{E} = f-(f-p_*) = 2^{-n}T_{n+1}$ has an $(n+2)$-alternant. But this is the content of Lemma 7.99 (iv). □

Chebyshev's polynomials can be used to get a version of Jackson's theorem for polynomials. Define for $f \in C([-1, 1])$,

$$\mathrm{dist}_\infty(f, \mathcal{P}_N) = \inf_{q\in\mathcal{P}_N} \max\{|f(t) - q(t)| : t \in [-1, 1]\}$$
$$\omega_f(\delta) = \sup\{|f(x) - f(x')| : x, x' \in [-1, 1],\ |x - x'| \leqslant \delta\}$$

7.103. THEOREM. *There exists a constant C such that for all $f \in C([-1, 1])$ we have*

$$\mathrm{dist}_\infty(f, \mathcal{P}_N) \leqslant C\omega_f(\tfrac{1}{N+1}).$$

PROOF. Define, for $x \in [-1/2, 1/2]$, $g(x) = f(\cos(2\pi x))$. Then g is an even function on $[-\frac{1}{2}, \frac{1}{2}]$ which extends to a 1-periodic function in $C^{\mathrm{ev}}(\mathbb{T})$. By Lemma 7.68 and Proposition 7.1 there exists an even trigonometric polynomial p such that $\|g - p\|_\infty = \mathrm{dist}_\infty(g, \mathcal{T}_N^{\mathrm{ev}})$ and by Jackson's theorem $\|g-p\| \leqslant C\omega_g(\frac{1}{N+1})$ where $\omega_g(r) = \sup\{|g(x+h)-g(x)| : |h| \leqslant r, x \in \mathbb{R}\}$. Also $\|g-p\|_\infty \leqslant 8C\omega_g(\frac{1}{8(N+1)})$. Now note that if $|x - x'| \leqslant \frac{1}{8(N+1)}$ and $t = \cos(2\pi x)$, $t' = \cos(2\pi x')$ then

$$|t - t'| \leqslant 2\pi|x - x'| \leqslant \frac{2\pi}{8(N+1)} < \frac{1}{N+1}$$

which shows that $\omega_g(\frac{1}{8(N+1)}) \leqslant \omega_f(\frac{1}{N+1})$.

We can write $p(x) = \sum_{k=0}^N a_k \cos 2\pi kx$, for $x \in [-1/2, 1/2]$ and setting $q(t) = p(\frac{1}{2\pi}\arccos t)$ yields

$$q(t) = \sum_{k=0}^N a_k \cos(k \arccos t) = \sum_{k=0}^N a_k T_k(t).$$

Thus $q \in \mathcal{P}_N$ and

$$\sup_{t\in[-1,1]} |f(t) - q(t)| = \sup_{x\in\mathbb{R}} |g(x) - p(x)| \leqslant 8C\omega_g(\tfrac{1}{8(N+1)}) \leqslant 8C\omega_f(\tfrac{1}{N+1})$$

which is what we wanted to show. □

There is some inefficiency in this proof which lies in the estimate $|t-t'| \leqslant 2\pi|x-x'|$. If x, x' are close to 0 or close to $1/2$ (i.e. if t, t' are close to 1 or -1) then this estimate can be improved. This is the reason why there is no equivalent of the Bernstein inverse theorem for polynomials on the full interval $[-1,1]$. The inverse results on best approximation still imply a regularity in the interior of our interval:

7.104. Exercise. Let $f \in C([-1,1])$ and let, for a compact interval $I \subset (-1,1)$

$$\omega_f(\delta; I) = \sup\{|f(x) - f(x')| : x, x' \in I,\ |x-x'| \leqslant \delta\}.$$

Let $0 < \alpha < 1$. There exists, for every compact $I \subset (-1,1)$ a constant $C_I > 0$ such that for all $f \in C([-1,1])$ satisfying $\sup_{N \geqslant 0} N^\alpha \mathrm{dist}(f, \mathcal{P}_N) \leqslant A$ we also have $\omega_f(\delta; I) \leqslant C_I A \delta^\alpha$.

Remark. The notation T_n is widely used in the literature because of the French transliteration of the Cyrillic letter Ч (Cha), as in Tchebychev (one of more than ten transliterations of the name Chebyshev found in the literature).

7.8. Interpolation of functions

In this section we study a different way to approximate functions. Instead of minimizing a certain (semi)norm quantifying the error, we aim to match the values of a target function f *exactly* at a finite collection of specified points often called *nodes*. For example, let $f \in C([a,b])$ and suppose $t_0 < \cdots < t_N$ in $[a,b]$. Can we find a polynomial $p \in \mathcal{P}_N$ such that

$$p(t_j) = f(t_j) \quad \text{for all } j = 0, \ldots, N\ ?$$

As we will see, the answer is yes: the polynomial p is uniquely determined and is called the *Lagrange interpolation polynomial*. We will also study variants where in addition also the values of p' are prescribed (Hermite interpolation), as well as a version for trigonometric polynomials.

It is then natural to ask how well these interpolation polynomials represent the function f away from the interpolation nodes (t_j). As we shall see, there are some unwelcome surprises here. The error is generally only well-behaved if f satisfies very stringent regularity requirements (such as analyticity).

Interpolation has applications to *numerical integration*: a natural approach to numerically approximate the integral $\int_a^b f(t)\,dt$ is to choose a large N and suitable nodes $t_0 < \cdots < t_N$ in $[a,b]$ and then instead evaluate $\int_a^b p(t)\,dt$, where p is the corresponding interpolation polynomial.

7.8.1. Lagrange interpolation. Given $N+1$ nodes $t_0 < t_1 < \cdots < t_N$ in $[a,b]$ and given complex numbers $y_0, \ldots, y_N$ we would like to find a polynomial $p \in \mathcal{P}_N$ which takes the value y_k at t_k, for $k = 0, \ldots, N$. In what follows we use the *elementary Lagrange polynomials* (also known as *Lagrange factors*).

$$\ell_k(t) = \prod_{\substack{j=0,\ldots N \\ j \neq k}} \frac{t - t_j}{t_k - t_j}. \tag{7.31}$$

One verifies that each ℓ_k is a polynomial of exact degree N, and

$$\ell_k(t_j) = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases} \tag{7.32}$$

The following alternative description is often useful. Given distinct nodes t_j, $j = 0, \ldots, N$ we call

$$\Omega(t) = (t - t_0) \cdots (t - t_N) \tag{7.33}$$

the associated *node polynomial*.

7.105. LEMMA. *Given distinct nodes* $\{t_k : k = 0, \dots, N\}$ *with node polynomial* Ω *we have*

$$\Omega'(t_k) = \prod_{j \neq k} (t_k - t_j)$$

and

$$\ell_k(t) = \frac{\Omega(t)}{(t - t_k)\Omega'(t_k)}, \quad k = 0, \dots, N.$$

PROOF. By the Leibniz rule,

$$\Omega'(t) = \sum_{l=0}^{N} \big(\prod_{j \neq l} (t - t_j) \big).$$

We evaluate this for $t = t_k$ and use $\prod_{j \neq l}(t_k - t_j) = 0$ for $l \neq k$. The claimed formula fo $\Omega'(t_k)$ follows, and then the stated formula for the ℓ_k is immediate. □

REMARK. For another way to write this, see Exercise 7.179.

7.106. THEOREM. *Given* $t_0 < \cdots < t_N$ *the polynomial*

$$P(t) = \sum_{k=0}^{N} y_k \ell_k(t)$$

is the unique $P \in \mathcal{P}_N$ *satisfying*

$$P(t_k) = y_k, \quad k = 0, \dots, N.$$

PROOF. First, $P \in \mathcal{P}_N$, and P has the desired property, because

$$P(t_m) = \sum_{k=0}^{N} y_k \ell_k(t_m) = y_m \ell_m(t_m) = y_m.$$

To prove the uniqueness, we let $P(t) = \sum_{j=0}^{N} a_j t^j$. P satisfies $P(t_k) = y_k$ for $k = 0, \dots, N$ if and only if $(a_0, \dots, a_N)$ solves the linear system of $N + 1$ equations

$$a_0 + a_1 t_k + a_2 t_k^2 + \cdots + a_N t_k^N = y_k, \quad k = 0, \dots, N.$$

The system of equations has a unique solution for every set of data $(y_k)_{k=0}^N$ if and only if the coefficient matrix is invertible (which is true if and only if it has nonzero determinant). The determinant of the coefficient matrix is the *Vandermonde determinant*

$$\det \begin{pmatrix} 1 & t_0 & t_0^2 & \dots & t_0^N \\ 1 & t_1 & t_1^2 & \dots & t_1^N \\ 1 & t_2 & t_2^2 & \dots & t_2^N \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_N & t_N^2 & \dots & t_N^N \end{pmatrix} = \prod_{\substack{i,j=0,\dots,N \\ i>j}} (t_i - t_j) \tag{7.34}$$

and indeed it is nonzero. This proves the uniqueness. □

7.107. EXERCISE. Prove formula (7.34) for the Vandermonde determinant by induction and elementary row and column operations.

7.8.2. Other types of interpolation.

Hermite interpolation. We briefly discuss a variant of Lagrange interpolation where not only the values are prescribed at the nodes but also their derivatives.

7.108. THEOREM (Hermite interpolation). *Given $n+1$ nodes $t_0, \dots, t_n$ and data $(y_0, \dots, y_n)$, $(v_0, \dots, v_n)$ there exists a unique polynomial $p \in \mathcal{P}_{2n+1}$*

$$P(t_j) = y_j, \quad P'(t_j) = v_j, \quad j = 0, \dots, n.$$

P is given by

$$P(t) = \sum_{j=0}^{n} y_j A_j(t) + \sum_{j=0}^{n} v_j B_j(t) \tag{7.35}$$

where (with ℓ_j as in (7.31))

$$A_j(t) = \big(1 - 2(t-t_j)\ell_j'(t_j)\big)(\ell_j(t))^2\,, \tag{7.36a}$$

$$B_j(t) = (t-t_j)(\ell_j(t))^2\,. \tag{7.36b}$$

P is called the Hermite interpolation polynomial associated with the data $(t_j; y_j, v_j)_{j=0}^n$

PROOF. We first show uniqueness. If P_1, P_2 solve the interpolation problem then $P = P_1 - P_2$ is a polynomial in $\mathcal{P}_{2n+1}$ with $n+1$ double zeros and thus must be identically zero.

To verify the identity (7.35) we observe (using (7.32)) that $A_j(t_k) = 0$ if $k \neq j$ and $A_k(t_k) = 1$, moreover $B_j(t_k) = 0$ for all k. For the derivatives we compute

$$\begin{aligned} A_j'(t) &= -2\ell_j'(t_j)\ell_j(t) + 2\big(1 - 2(t-t_j)\ell_j'(t_j)\big)\ell_j(t)\ell_j'(t) \\ B_j'(t) &= \ell_j(t)^2 + 2(t-t_j)\ell_j(t)\ell_j'(t) \end{aligned}$$

and we get $A_j'(t_k) = 0$ for all k and $B_j'(t_k) = 0$ when $j \neq k$ and $B_j'(t_j) = 1$. This implies that the polynomial $\sum_{j=0}^n y_j A_j + v_j B_j \in \mathcal{P}_{2n+1}$ satisfies the interpolation problem. □

Trigonometric interpolation. It is sometimes useful to also consider interpolation by trigonometric polynomials; in fact this will turn out to be useful in proving results about polynomials. We first formulate an analogue of Theorem 7.106.

7.109. THEOREM. *Let $x_0, \dots, x_{2N}$ be $2N+1$ different nodes in $[0,1)$ and let $y_0, \dots, y_{2N}$ be given. Then there exists a unique $p \in \mathcal{T}_N$ such that $p(x_j) = y_j$, for $j = 0, \dots, 2N$. It is given by*

$$p(x) = \sum_{k=0}^{N} y_k \widetilde{\ell}_k(x) \text{ with } \widetilde{\ell}_k(x) = \prod_{\substack{i=0,\dots,2N \\ i\neq k}} \frac{\sin(\pi(x-x_i))}{\sin(\pi(x_k - x_i))}\,.$$

Applying the uniqueness statement in this theorem we get

7.110. COROLLARY. *Any $p \in \mathcal{T}_N$ with $2N+1$ distinct zeros in $[0,1)$ must be identically 0.*

PROOF OF THEOREM 7.109. Note that part (i) is a special case of the uniqueness assertion in part (ii). It is easily checked that the given formula gives a trigonometric interpolation polynomial of degree at most N, for the data (x_j, y_j). It remains to prove uniqueness.

Let $p(x) = \sum_{k=-N}^{N} c_k e^{2\pi i k x}$ satisfy the equations $p(x_j) = y_j$, for $j = 0, \dots, 2N$. This is a system of $2N+1$ equations for the vector $(c_{-N}, \dots, c_N)$. The coefficient matrix of this linear system is given by

$$A(x_0, \dots, x_{2N}) = \begin{pmatrix} e^{2\pi i(-N)x_0} & e^{2\pi i(1-N)x_0} & \dots & e^{2\pi i N x_0} \\ e^{2\pi i(-N)x_1} & e^{2\pi i(1-N)x_1} & \dots & e^{2\pi i N x_1} \\ \vdots & \vdots & & \vdots \\ e^{2\pi i(-N)x_{2N}} & e^{2\pi i(1-N)x_{2N}} & \dots & e^{2\pi i N x_{2N}} \end{pmatrix}$$

and its determinant is $e^{-2\pi iN(x_0+x_1+\cdots+x_{2N})}V$ where V is the Vandermonde determinant with respect to the distinct numbers $e^{2\pi ix_j}$, $j=0,\dots,2N$. Thus

$$\det A(x_0,\dots,x_{2N}) = e^{-2\pi iN\sum_{j=0}^{2N}x_j}\prod_{\substack{j,k=0,\dots,2N\\ j>k}}\left(e^{2\pi ix_j}-e^{2\pi ix_k}\right)\neq 0,$$

and we obtain the uniqueness of the coefficients c_k, $-N\leqslant k\leqslant N$, and therefore the uniqueness of the trigonometric interpolation polynomial. □

A particularly nice formula is available for equidistant nodes.

7.111. Theorem (Trigonometric interpolation with equidistant nodes). *For $N\in\mathbb{N}$ consider the equidistant nodes $\{\frac{j}{2N+1} : j=0,\dots,2N\}$. Then the trigonometric interpolation polynomial $p\in\mathcal{T}_N$ satisfying $p(\frac{l}{2N+1})=y_l$ for $l=0,\dots,2N$ is given by*

$$p(x)=\sum_{k=-N}^{N}c_ke^{2\pi ikx}\ \text{ with } c_k=\frac{1}{2N+1}\sum_{j=0}^{2N}y_je^{-2\pi i\frac{jk}{2N+1}},\quad k=-N,\dots,N.$$

Proof. In view of the uniqueness statement in Theorem 7.109 we just have to check the interpolation property for p as in the display. We compute

$$p(\tfrac{l}{2N+1})=\sum_{k=-N}^{N}c_ke^{2\pi ik\frac{l}{2N+1}}=\frac{1}{2N+1}\sum_{j=0}^{2N}y_j\sum_{k=-N}^{N}e^{2\pi ik\frac{l-j}{2N+1}}$$

and $\sum_{k=-N}^{N}e^{2\pi ik\frac{l-j}{2N+1}}$ is equal to $2N+1$ if $l=j \mod N$ and equal to 0 if $l\neq j \mod N$ (this came up in Example 7.33 before). Hence $p(\frac{l}{2N+1})=y_l$. □

7.8.3. Error estimates. Given $f\in C([a,b])$ and $(N+1)$ nodes $t_0<t_1<\cdots<t_N$ in $[a,b]$ we define the associated *Lagrange interpolation polynomial* $L_Nf(t)$ as in Theorem 7.106, taking for y_k the values of f at t_k, i.e.

$$L_Nf(t)=\sum_{k=0}^{N}f(t_k)\ell_k(t) \tag{7.37}$$

Notice that L_Nf is a polynomial of degree at most N and L_N is a bounded linear operator on $C([a,b])$. By Theorem 7.106 L_Nf is the unique polynomial of degree at most N whose values at $t_0,\dots,t_N$ coincide with the values of f at those points.

We are now interested in an estimate for the error in Lagrange interpolation,

$$f(t)-L_Nf(t).$$

The following result can be used if $f\in C^{N+1}([a,b])$ (for an analogous result in Hermite interpolation see Exercise 7.181). Note the formal similarity of the statement with Taylor's theorem.

7.112. Theorem. *Let $f\in C^{N+1}([a,b])$. Given nodes $t_0<t_1<\cdots<t_N$ in $[a,b]$ let L_Nf be the Lagrange interpolation polynomial associated to these nodes as in (7.37). Then for each $t\in[a,b]$ there exists $\xi\in(a,b)$ such that*

$$f(t)-L_Nf(t)=\frac{f^{(N+1)}(\xi)}{(N+1)!}\prod_{j=0}^{N}(t-t_j).$$

Proof. Note that both sides are zero if t is one of the t_k. We may thus assume that $t\notin\{t_0,\dots,t_N\}$.

We set $q(t) = \prod_{j=0}^{N}(t-t_j)$ which is a polynomial of degree $N+1$ satisfying $q^{(N+1)} = (N+1)!$. Fix t and consider on $[a,b]$ the auxiliary function

$$s \mapsto h(s) = f(s) - L_N f(s) - \frac{f(t) - L_N f(t)}{q(t)} q(s)$$

which belongs to $C^{N+1}([a,b])$. Notice that $h(t_k) = 0$ for $k = 0, \dots, N$ and moreover $h(t) = 0$. This means that h has in $[a,b]$ at least $N+2$ distinct zeros. By Rolle's theorem the derivative h' has at least $N+1$ distinct zeros. Continuing with applications of Rolle's theorem we see for $\ell = 1, \dots N$ that $h^{(\ell)}$ has at least $N+2-\ell$ distinct zeros in $[a,b]$, in particular $h^{(N)}$ has two distinct zeros in $[a,b]$. By another application of Rolle's theorem $h^{(N+1)}$ has a zero $\xi \in (a,b)$. Now $(L_N f)^{(N+1)} = 0$ (since $L_N f$ is a polynomial of degree at most N) and therefore

$$0 = h^{(N+1)}(\xi) = f^{(N+1)}(\xi) - \frac{f(t) - L_N f(t)}{q(t)} (N+1)!$$

which gives the assertion. □

We also record the special case of Chebyshev nodes.

7.113. Theorem (Chebyshev interpolation). *Let $N \in \mathbb{N}$, let $t_{k,N} = \cos(\frac{2k+1}{2N+2}\pi)$ be the Chebyshev nodes, i.e. the $N+1$ zeros of T_{N+1} in $(-1,1)$. For $f \in C([-1,1])$ let $\mathcal{L}_N f$ be the Lagrange interpolation polynomial with $\mathcal{L}_N f(t_{k,N}) = f(t_{k,N})$. Then*

If $f \in C^{N+1}([-1,1])$, then for every $t \in [-1,1]$ there exists a $\xi \in (-1,1)$ such that

$$f(t) - \mathcal{L}_N f(t) = \frac{f^{(N+1)}(\xi)}{2^N (N+1)!} T_{N+1}(t).$$

Moreover,

$$\max_{t\in[-1,1]} |f(t) - \mathcal{L}_N f(t)| \leqslant \frac{1}{2^N (N+1)!} \max_{t\in[-1,1]} |f^{(N+1)}(t)|.$$

Proof. Note that $2^{-N} T_{N+1}(t) = \prod_{j=0}^{N}(t - t_{j,N})$. Then apply Theorem 7.112. Since $|T_{N+1}(t)| \leqslant 1$, the last claim follows. □

7.8.4. Uniform convergence. We may also ask what happens in Theorem 7.112 when $N \to \infty$. Here we choose for every N a set of nodes

$$a = t_{0,N} < \cdots < t_{N,N} = b$$

and denote by $\ell_{k,N}$ the associated elementary Lagrange polynomials. Assuming that the width $\max_k |t_{k,N} - t_{k+1,N}|$ of the resulting partition of $[a,b]$ tends to 0 as $N \to \infty$, one might hope that

$$L_N f = \sum_{k=0}^{N} f(t_{k,N}) \ell_{k,N} \tag{7.38}$$

converges uniformly on $[a,b]$ to f, at least if f is sufficiently smooth. Under the strong assumption of analyticity (which is a growth condition on the derivatives as $N \to \infty$) one can use Theorem 7.112 to show such a result:

7.114. Exercise. Suppose that f is analytic at the origin (see Definition 3.61). Show that there exists $r > 0$ so that for all choices of nodes in $[-r,r]$ with $\max_k |t_{k,N} - t_{k+1,N}| \to 0$ the sequence $(L_N f)$ in (7.38) converges uniformly to f on $[-r,r]$.

For example, if f is represented by a Taylor series with infinite convergence radius, such as the exponential function, cos and sin, we can deduce uniform convergence of the interpolation polynomials on a compact interval $[a,b]$ (provided that the width of the partition tends to 0).

One might hope that uniform convergence holds under less severe regularity assumptions. It turns out that such convergence results depend on the choice of nodes, and that the seemingly natural choice of equidistant nodes results in Lagrange interpolation polynomials that are badly behaved away from the interpolation nodes; in particular uniform convergence may fail even for smooth functions. This is known as *Runge's phenomenon* (see Figure 7.5). Runge gave the example

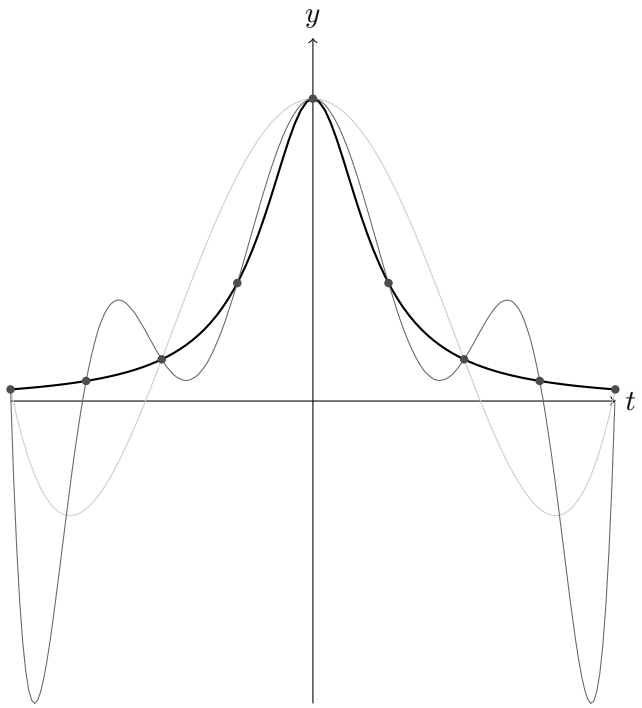


FIGURE 7.5. Runge's phenomenon: the function $f(t) = \frac{1}{1+t^2}$ is not well-represented by its Lagrange interpolation polynomials for equidistant nodes on the interval $[-5, 5]$ away from the interpolation nodes (here degrees 4 and 8 are shown).

$$f(t) = (1 + t^2)^{-1}$$

with the choice of $N + 1$ equidistant nodes in the interval $[-5, 5]$. By the above exercise, the sequence $L_N f$ converges for t near 0. But $L_N f$ does not converge uniformly to f on $[-5, 5]$. Indeed, $\max_{[-5,5]} |L_N f - f| \to \infty$ as $N \to \infty$ (we will not prove this here).

To get some insight into this phenomenon, let us compute the operator norm of L_N as an operator $C([a, b]) \to C([a, b])$.

7.115. THEOREM. *The operator norm of $L_N : C([a, b]) \to C([a, b])$ is given by*

$$\|L_N\|_{\mathrm{op}} = \sup_{t\in[a,b]} \sum_{k=0}^{N} |\ell_{k,N}(t)|$$

PROOF. The asserted upper bound for $\|L_N\|_{\mathrm{op}}$ follows from

$$|L_N f(x)| \leqslant \sup_{u\in[a,b]} |f(u)| \sum_{k=0}^{N} |\ell_{k,N}(x)| \leqslant \Big(\sup_{u\in[a,b]} \sum_{k=0}^{N} |\ell_{k,N}(u)| \Big) \|f\|_\infty.$$

For the lower bound fix $t_* \in [a, b]$. Then let $y_{k,N} = 1$ if $\ell_{k,N}(t_*) \geqslant 0$ and let $y_{k,N} = -1$ if $\ell_{k,N}(t_*) < 0$. Now define f on $[a, b]$ as the unique continuous piecewise linear function with the

property $f(t_{k,N}) = y_{k,N}$. Define $f(s) = f(t_{0,N})$ for $a \leqslant s < t_{0,N}$ and $f(s) = f(t_{N,N})$ for $t_{N,N} < s \leqslant b$ (this only applies when $t_{0,N} > a$ or $t_{N,N} < b$). Then $\|f\|_\infty = 1$ and

$$|L_N f(t_*)| = \Big| \sum_{k=0}^{N} f(t_{k,N}) \ell_{k,N}(t_*) \Big| = \sum_{k=0}^{N} |\ell_{k,N}(t_*)|$$

which shows that $\|L_N\|_{\mathrm{op}} \geqslant \sum_{k=0}^{N} |\ell_{k,N}(t_*)|$. Since t_* was arbitrary we get the asserted lower bound. □

The operator norms $\|L_N\|_{\mathrm{op}}$ are also called *Lebesgue constants* for the polynomial interpolation method under consideration.

Equidistant nodes. As a first example we consider the natural special case of equidistant nodes for the polynomial interpolation, i.e.

$$t_{k,N} = a + \tfrac{k}{N}(b-a), \qquad k = 0, \dots, N. \tag{7.39}$$

We would like to know whether the sequence of polynomials $L_N f$ converges uniformly (i.e. converges in $C([a,b])$). For a given function f, this may be difficult to determine. However, by the uniform boundedness principle (Theorem 5.35), if we can show that the operator norms $\|L_N\|_{\mathrm{op}}$ are unbounded in N, then there exists an $f \in C([a,b])$ so that $\|L_N f\|_\infty$ is unbounded; in particular $L_N f$ cannot converge uniformly. As we will see, in the equidistant case the operator norms grow rapidly (in fact exponentially) in N. This can be viewed as another manifestation of Runge's phenomenon (see Figure 7.5). Roughly speaking, Lagrange interpolation polynomials extrapolate badly to points away from the interpolation nodes and may oscillate wildly, particularly near the endpoints, as the degree increases.

7.116. Theorem. *Consider the equidistant nodes in* (7.39) *and define L_N with respect to these nodes. Then there exists $N_0 \in \mathbb{N}$ so that for all $N \geqslant N_0$,*

$$\|L_N\|_{\mathrm{op}} \geqslant \tfrac{1}{50}\, N^{-2} 2^N \tag{7.40}$$

Proof. We evaluate the expression for the operator norm from Theorem 7.115 at a distinguished point, $s_0 = a + \frac{1}{2N}(b-a)$ (the midpoint of the first partition interval) to obtain the lower bound

$$\|L_N\|_{\mathrm{op}} \geqslant \sum_{k=0}^{N} |\ell_{k,N}(s_0)|.$$

Let us evaluate the polynomials $\ell_{k,N}$ at s_0:

$$\begin{aligned} |\ell_{k,N}(s_0)| &= \prod_{\substack{j=0,\dots N \\ j\neq k}} \frac{|s_0 - t_j|}{|t_k - t_j|} = \prod_{\substack{j=0,\dots N \\ j\neq k}} \frac{|a + \frac{1}{2N}(b-a) - a - \frac{j}{N}|}{|a + \frac{k}{N}(b-a) - a - \frac{j}{N}(b-a)|} \\ &= \prod_{\substack{j=0,\dots N \\ j\neq k}} \frac{|2j-1|}{|2k-2j|} = \frac{1}{2k-1} \frac{1 \cdot 3 \cdots (2N-1)}{2^N \prod_{j=0}^{k-1}(k-j) \prod_{j=k+1}^{N}(j-k)}. \end{aligned}$$

Since $1 \cdot 3 \cdots (2N-1) = \frac{(2N)!}{2^N N!}$, $\prod_{j=0}^{k-1}(k-j) = k!$ and $\prod_{j=k+1}^{N}(j-k) = (N-k)!$,

$$|\ell_{k,N}(s_0)| = \frac{1}{2k-1} \frac{(2N)!}{2^{2N} N! k! (N-k)!}.$$

We now use Stirling's formula (Theorem 3.116), $n! = n^n e^{-n} \sqrt{2\pi n}(1+u_n)$ where $|u_n| \leqslant C/n$ for a constant $C \in (0,\infty)$. Hence there is $n_0 \geqslant 4$ such that $|u_n| \leqslant 1/2$ for $n \geqslant n_0$. We apply Stirling's formula for the four cases of n being $2N$, N, $N-k$ and k and obtain

$$|\ell_{k,N}(s_0)| = \frac{1}{2k-1} \frac{1}{\pi\sqrt{2k(N-k)}} N^N k^{-k} (N-k)^{-(N-k)} \frac{1+u_{2N}}{(1+u_N)(1+u_k)(1+u_{N-k})}.$$

Let us fix a special k, namely $k_\circ = \lfloor N/2 \rfloor$. Then

$$N^N k_\circ^{-k_\circ}(N-k_\circ)^{-(N-k_\circ)} = \begin{cases} 2^N & \text{if } N \text{ is even} \\ (\frac{N}{\sqrt{N^2-1}})^N 2^N & \text{if } N \text{ is odd.} \end{cases}$$

Use that for $N > 2n_0$

$$\frac{1+u_{2N}}{(1+u_N)(1+u_{k_\circ})(1+u_{N-k_\circ})} \geqslant \frac{1/2}{(3/2)^3} = c_0 > 0.$$

This gives

$$|\ell_{k_\circ,N}(s_0)| \geqslant c_1 \frac{1}{2k_\circ - 1} \frac{1}{\sqrt{k_\circ(N-k_\circ)}} 2^N,$$

where $c_1 = c_0 \pi^{-1} 2^{-1/2} > \frac{1}{50}$. Finally $\frac{1}{2k_\circ - 1} \geqslant N^{-1}$, $\frac{1}{\sqrt{k_\circ(N-k_\circ)}} \geqslant N^{-1}$ and the asserted lower bound follows. □

Remark. The lower bound can be somewhat improved, but not by that much (note that in the proof above we just relied on a bound from below for the term $\ell_{k,N}(s_0)$ for $k = \lfloor N/2 \rfloor$ but did not consider the entire sum $\sum_{k=0}^N |\ell_{k,N}(s_\circ)|$). The sharp asymptotic is $\|L_N\|_{\mathrm{op}} \sim \frac{2^{N+1}}{eN(\log N)}$ as $N \to \infty$ (obtained by Turetskii [**84**], Schönhage [**68**], with different accuracy in the error estimates). For more details see Trefethen-Weideman [**82**].

Lower bounds for projection operators. The property that there exists $f \in C([a,b])$ such that $(L_N f)$ does not converge uniformly on $[a,b]$ can be obtained in much greater generality, indeed for *arbitrary* sets of nodes. In order to prove such a result, we first take a detour to *trigonometric interpolation* considered already in §7.8.2, where the possible failure of uniform convergence is linked directly to the possible failure of uniform convergence of the partial sum operators (see §7.5.2). A key tool is given by a lower bound for the operator norms of general bounded projections to the subspaces $\mathcal{T}_N$.

7.117. Definition. Let X be a normed space and $V \subset X$ a closed subspace. A linear operator $P : X \to X$ is called a *bounded projection* onto V if P is bounded, the range of P is V and $Pg = g$ for all $g \in V$.

7.118. Examples.

(i) If X is a Hilbert space and V a closed subspace, then the orthogonal projection onto V is a bounded projection (with operator norm equal to 1). See Theorem 7.48.

(ii) Let $X = C(\mathbb{T})$ and $V = \mathcal{T}_N$. The partial sum operator $S_N : C(\mathbb{T}) \to \mathcal{T}_N$ is a bounded projection. The Fejér means σ_N are *not* projections (if $N \geqslant 1$).

Let $N \geqslant 0$ and let $(x_{j,N})_{j=0,\dots,2N}$ be $2N+1$ distinct nodes in $[0,1)$. For $f \in C(\mathbb{T})$ let $\mathcal{L}_N f$ be the unique polynomial in $\mathcal{T}_N$ satisfying $\mathcal{L}_N f(x_{j,N}) = f(x_{j,N})$ for $j = 0, \dots, 2N$. By the uniqueness statement in Theorem 7.109 the bounded linear operator $\mathcal{L}_N$ is the identity on $\mathcal{T}_N$ and thus a projection to $\mathcal{T}_N$.

(iii) Let $X = C([a,b])$ and $V = \mathcal{P}_N$. The Lagrange polynomial interpolation operators L_N as defined in §7.8 are bounded projections $C([a,b]) \to \mathcal{P}_N$.

In the following result we temporarily forget about interpolation and derive a general result about projections to $\mathcal{T}_N$.

7.119. Theorem. *Let P be a bounded projection operator from $C(\mathbb{T})$ to $\mathcal{T}_N$. Then its operator norm satisfies*

$$\|P\|_{\mathrm{op}} \geqslant \|S_N\|_{\mathrm{op}}.$$

Proof. Let $\tau_h f(x) = f(x-h)$. Define a bounded operator A by

$$Af(x) = \int_{-1/2}^{1/2} \tau_{-h} P \tau_h f \, dh.$$

Note that $\|\tau_{\pm h}\|_{\mathrm{op}} = 1$, and thus $\|\tau_{-h} P \tau_h\|_{\mathrm{op}} \leqslant \|P\|_{\mathrm{op}}$. We have

$$\|Af\|_\infty \leqslant \int_{-1/2}^{1/2} \|\tau_{-h} P \tau_h f\|_\infty \, dh \leqslant \int_{-1/2}^{1/2} \|P\|_{\mathrm{op}} \|f\|_\infty \, dh = \|P\|_{\mathrm{op}} \|f\|_\infty$$

and thus $\|A\|_{\mathrm{op}} \leqslant \|P\|_{\mathrm{op}}$. Let $e_n(x) = e^{2\pi i n x}$. We now show that

$$Ae_n = \begin{cases} e_n & \text{if } |n| \leqslant N\,, \\ 0 & \text{if } |n| > N\,. \end{cases} \tag{7.41}$$

Indeed by linearity (7.41) yields that $f = S_N f$ whenever f is a trigonometric polynomial, and since the trigonometric polynomials are dense in $C(\mathbb{T})$ and A and S_N are continuous we get $A = S_N$ as operators on $C(\mathbb{T})$.

To establish (7.41) we first note that $\tau_h e_n = e^{2\pi i n h} e_n$. For $|n| \leqslant N$ we have $Pe_n = e_n$ (as P is a projection onto $\mathcal{T}_N$). We thus get $\tau_{-h} P \tau_h e_n = e_n$ for $|n| \leqslant N$ and all h and after integrating we obtain (7.41) for $|n| \leqslant N$.

Now assume $|n| > N$. We then have $Pe_n = \sum_{k=-N}^{N} a_{n,k} e_k$ with coefficients $a_{n,k} \in \mathbb{C}$ and compute

$$\tau_{-h} P \tau_h e_n(x) = e^{-2\pi i h n} \tau_{-h} P e_n(x) = e^{-2\pi i h n} \sum_{k=-N}^{N} a_{n,k} e^{2\pi i k (x+h)}$$

Hence

$$Ae_n(x) = \sum_{k=-N}^{N} a_{n,k} e^{2\pi i k x} \int_{-1/2}^{1/2} e^{2\pi i k h} e^{-2\pi i n h} dh = 0$$

by the orthogonality property $\langle e_n, e_k \rangle = 0$, $n \neq k$ where e_n and e_k are considered functions of the variable h. This proves (7.41) also for $|n| > N$ and we are done. □

Remark. The proof relies on an averaging identity of Faber [**20**] (also in more generalized settings in later work by Marcinkiewicz, Berman, Kharshiladze, Lozinski).

7.120. Corollary. *Let $N \geqslant 0$ and let $(x_{j,N})_{j=0,\dots,2N}$ be $2N+1$ distinct nodes in $[0,1)$. For $f \in C(\mathbb{T})$ let $\mathcal{L}_N f$ be the unique polynomial in $\mathcal{T}_N$ satisfying $\mathcal{L}_N f(x_{j,N}) = f(x_{j,N})$ for $j = 0, \dots, 2N$.*

Then $\mathcal{L}_N$ is a bounded operator on $C(\mathbb{T})$, with operator norm greater than $\frac{4}{\pi^2}\log(N+1)$. There exists $g \in C(\mathbb{T})$ such that $\|\mathcal{L}_N g\|_\infty$ is unbounded; in particular $\mathcal{L}_N g$ does not converge uniformly.

Proof. As noted in Example 7.118 (ii) the operator $\mathcal{L}_N$ is a projection. By Lemma 7.77 we have $\|S_N\| > \frac{4}{\pi^2}\log(N+1)$ and the lower bound for $\mathcal{L}_N$ follows from Theorem 7.119. The second statement follows from the uniform boundedness principle. □

We now consider similar lower bounds for the Lagrange interpolation operators associated to polynomials $\sum_{k=0}^{N} a_k t^k$ for $t \in [a,b]$, with *any choice of $N+1$ nodes in $[a,b]$*. We shall show that there exist functions $f \in C([a,b])$ such that the sequence $L_N f$ is not uniformly bounded and therefore does not converge uniformly; this is known as *Faber's theorem*. The idea is to use the substitution

$$t = \tau(x) = \frac{a+b}{2} + \frac{b-a}{2} \cos 2\pi x$$

with $0 \leqslant x \leqslant 1/2$. Note that the function $x \to \tau(x)$ is a continuous bijective map from $[0, \frac{1}{2}]$ to $[a,b]$ and its inverse is continuous. As a consequence we see that the map $f \to f \circ \tau$ maps $C([a,b])$ to $C([0,\frac{1}{2}])$, with operator norm 1 and this map is linear and bijective. It is also bounded, with operator norm 1 and so is its inverse. Functions in $C([0,\frac{1}{2}])$ can be uniquely extended to even functions on $[-1/2, 1/2]$ and those can be further extended in a unique way to a function in $C^{\mathrm{ev}}(\mathbb{T})$, the space of

continuous 1-periodic functions satisfying $f(t) = f(-t)$. We have thus established an isomorphism of Banach spaces $\Phi : C([a,b]) \to C^{\mathrm{ev}}(\mathbb{T})$. The inverse is given by

$$\Phi^{-1}g(t) = g\big(\tfrac{1}{2\pi}\arccos(\tfrac{2t}{b-a} - \tfrac{a+b}{b-a})\big),$$

for $g \in C^{\mathrm{ev}}(\mathbb{T})$.

We examine the action of Φ on $\mathcal{P}_N$. For $p(t) = \sum_{k=0}^N a_k t^k$ we have

$$\Phi p(x) = \sum_{k=0}^N a_k(\tau(x))^k, \text{ and } \tau(x)^k = \sum_{j=0}^N \gamma_{jk}\cos(2\pi jx), \quad x \in [0,1/2],$$

by the binomial formula. Here $(\gamma_{jk})_{j,k=0,\dots,N}$ is a real $(N+1)\times(N+1)$ matrix, which by the above must be invertible. Again extending $\tau(x)^k$ as a 1-periodic even function we see that for $p \in \mathcal{P}_N$ we have $\Phi p \in \mathcal{T}_N^{\mathrm{ev}}$, the $N+1$ dimensional subspace of even trigonometric polynomials in $\mathcal{T}_N$, which can also be identified with the space consisting of the linear combinations $x \mapsto \sum_{j=0}^N b_j\cos(2\pi jx)$.

7.121. Theorem. *Let $N \geqslant 0$ and let $(t_{j,N})_{j=0,\dots,N}$ be $N+1$ distinct nodes in $[a,b]$. For $f \in C([a,b])$ let L_N be the Lagrange interpolation operator associated to these nodes, with $L_N f$ the unique $p \in \mathcal{P}_N$ satisfying $p(t_{j,N}) = f(t_{j,N})$ for $j = 0,\dots,N$.*

Then L_N is a bounded linear operator on $C([a,b])$, and the operator norm has the lower bound

$$\|L_N\|_{\mathrm{op}} \geqslant c\log(N+2)$$

for some $c > 0$.

We can use the uniform boundedness principle to deduce *Faber's theorem*:

7.122. Corollary. *There exists $g \in C([a,b])$ such that $(\|L_N g\|_\infty)_{N\in\mathbb{N}}$ is unbounded and thus $L_N g$ does not converge uniformly.*

Proof of Theorem 7.121. Let $\Phi : C([a,b]) \to C^{\mathrm{ev}}(\mathbb{T})$ be as above and for $f \in C^{\mathrm{ev}}(\mathbb{T})$ define $Ł_N : C^{\mathrm{ev}}(\mathbb{T}) \to \mathcal{T}_N^{\mathrm{ev}}$ by

$$Ł_N f(x) = \Phi L_N \Phi^{-1} f.$$

Let $x_{j,N} = \tau^{-1}(t_{j,N}) \in [0,1/2]$. The linear operator $Ł_N$ is bounded. It assigns to $f \in C^{\mathrm{ev}}(\mathbb{T})$ the unique even trigonometric polynomial P satisfying $P(x_{j,N}) = f(x_{j,N})$. The operator $Ł_N$ is a bounded projection from $C^{\mathrm{ev}}(\mathbb{T})$ to $\mathcal{T}_N^{\mathrm{ev}}$. Because of the restriction to even functions the averaging operator A in the proof of Theorem 7.119 is not applicable but a similar, slightly more complicated operator can be used.

We let S_N be the partial sum operator for the Fourier series and begin by recalling that S_N maps $C^{ev}(\mathbb{T})$ into itself. From Lemma 7.80 and Lemma 7.77

$$\|S_N\|_{C^{\mathrm{ev}}(\mathbb{T})\to C^{\mathrm{ev}}(\mathbb{T})} = \int_0^1 |D_N(x)|\,dx > \frac{4}{\pi^2}\log(N+1).$$

We now define

$$\mathcal{A}f(x) = \int_{h=-1/2}^{1/2} \tau_{-h}(I - Ł_N)(\tau_h + \tau_{-h})f(x)\,dh$$

and claim that

$$\mathcal{A}f = f - S_N f \text{ for } f \in C^{\mathrm{ev}}(\mathbb{T}) \tag{7.42}$$

Let $u_0(x) = 1$, $u_j(x) = \sqrt{2}\cos(2\pi jx)$, $j \in \mathbb{N}$. Then $\{u_j : j = 0,\dots,N\}$ is an orthonormal basis of $\mathcal{T}_N^{\mathrm{ev}}$ with respect to the scalar product $\langle f,g\rangle = 2\int_0^{1/2} f(x)\overline{g(x)}\,dx$. Moreover, the span of the $\{u_j\}_{j=0}^\infty$ is dense in $C^{\mathrm{ev}}(\mathbb{T})$; for this either see Example 7.21 or use the Fejér kernel. Therefore it suffices to show that $\mathcal{A}u_j = 0$ for $j = 0,\dots,N$ and $\mathcal{A}u_j = u_j$ for $j \geqslant N+1$.

Note that $\tau_h u_0 = u_0$ and

$$2^{-1/2}\tau_{\pm h}u_j(x) = \cos(2\pi j(x \mp h)) = \cos(2\pi jx)\cos(2\pi jh) \pm \sin(2\pi jx)\sin(2\pi jh)$$

and thus $(\tau_h + \tau_{-h})u_j(x) = 2^{1/2}u_j(x)u_j(h)$. Since $I - Ł_N = 0$ on $\mathcal{T}_N^{\mathrm{ev}}$ this implies $\mathcal{A}u_j = 0$ for $j = 0, \dots, N$.

For $j \geqslant N+1$ we have $Ł_N(\tau_h + \tau_{-h})u_j = 0$ and therefore,

$$\mathcal{A}u_j(x) = u_j(x) + \int_{-1/2}^{1/2} \tau_{-2h}u_j(x)dh, \quad j \geqslant N+1.$$

Note that for $j > 0$

$$2^{-\frac{1}{2}}\int_{-\frac{1}{2}}^{\frac{1}{2}} \tau_{-2h}u_j(x)dh = \int_{-\frac{1}{2}}^{\frac{1}{2}} \cos(2\pi x + 4\pi h)dh = \frac{1}{4\pi}(\sin(2\pi(x+1)) - \sin(2\pi(x-1))) = 0$$

and thus we get $\mathcal{A}u_j = u_j$ for $j \geqslant N+1$. By linearity we obtain $\mathcal{A}f = f - S_N f$, first for all even trigonometric polynomials and then by density for all $f \in C^{\mathrm{ev}}(\mathbb{T})$.

From (7.42) we see that (using the operator norms $C^{\mathrm{ev}}(\mathbb{T}) \to C^{\mathrm{ev}}(\mathbb{T})$)

$$\|I - S_N\|_{\mathrm{op}} = \|\mathcal{A}\|_{\mathrm{op}} \leqslant 2\|I - Ł_N\|_{\mathrm{op}}.$$

Thus

$$\|Ł_N\|_{\mathrm{op}} \geqslant \frac{\|S_N\|_{\mathrm{op}} - 3}{2} > \frac{2}{\pi^2}\log(N+1) - \frac{3}{2}$$

and since $\|Ł_N\|_{\mathrm{op}} \geqslant 1$ this gives the asserted lower bound. Since Φ and Φ^{-1} have operator norms 1 we see that $\|L_N\|_{\mathrm{op}} = \|Ł_N\|_{\mathrm{op}}$ which finishes the proof. □

Remark. For the case of equidistant nodes (for polynomials) one should compare the general logarithmic lower bound for L_N with the actual exponential lower bound in Theorem 7.116, so in this instance the lower bound in Theorem 7.121 is hugely inefficient. On the other hand in its generality (for *all* choices of nodes) Theorem 7.121 is for large N sharp up to a multiplicative constant. In §7.8.4 we study families of nodes that give especially good results for interpolation and for which the operator norms of the associated Lagrange interpolation operators grow only logarithmically.

Remark. Very strong pointwise divergence results were proved by Grünwald [**26**] and by Marcinkiewicz [**51**]. They constructed continuous functions f on $[-1,1]$ for which the sequence of the associated Lagrange polynomials for the Chebyshev nodes diverges for *every* $t \in [-1,1]$; see Natanson [**56**, II.3]. Such a result cannot hold for arbitrary choices of nodes (for example, a fixed node might show up in all collections of nodes and then convergence holds at this node). However, Erdős and Vertesi [**19**] proved that for every sequence L_N, with any choice of distinct nodes $(t_{k,N})_{k=0}^N$ in $[a,b]$, one can construct a function $f \in C([a,b])$ such that $L_N f(t)$ diverges for all $t \in [a,b]$ in the complement of a Lebesgue null set.

Interpolation with Chebyshev nodes. We shall now analyze the Lebesgue constants for the interpolation operator with Chebyshev nodes, i.e. the operator norms of $\mathcal{L}_N$ as an operator on $C([-1,1])$ with nodes

$$t_{k,N} = \cos\Big(\frac{2k+1}{2N+2}\pi\Big) \in (-1,1), \tag{7.43}$$

the $N+1$ zeros of T_{N+1}. To use Theorem 7.115 we need formulas for the elementary Lagrange polynomials.

7.123. Lemma. *The associated elementary Lagrange polynomials are given by*

$$\ell_{k,N}(t) = \frac{(-1)^{k-1}T_{N+1}(t)}{(N+1)(t - t_{k,N})}(1 - t_{k,N}^2)^{1/2}.$$

Setting $\vartheta_{k,N} = \arccos t_{k,N} = \frac{2k+1}{2N+2}\pi$,

$$\ell_{k,N}(\cos\vartheta) = (-1)^k \frac{\cos(N+1)\vartheta}{2N+2}\Big(\cot\frac{\vartheta+\vartheta_{k,N}}{2} - \cot\frac{\vartheta-\vartheta_{k,N}}{2}\Big).$$

PROOF. For the first formula we use Lemma 7.105 and that the node polynomial Ω is given by $2^{-N}T_{N+1}$, and $T'_{N+1}(t_{k,N}) = (-1)^{k-1}(1-t_{k,N}^2)^{-1/2}$. In the ϑ variables we get

$$\ell_{k,N}(\cos\vartheta) = (-1)^k \frac{\cos(N+1)\vartheta}{(N+1)(\cos\vartheta - \cos\vartheta_{k,N})}\sin\vartheta_{k,N}.$$

Use trigonometric identities

$$\frac{\sin\vartheta_{k,N}}{\cos\vartheta - \cos\vartheta_{k,N}} = -\frac{\sin(\frac{\vartheta+\vartheta_{k,N}}{2} - \frac{\vartheta-\vartheta_{k,N}}{2})}{2\sin\frac{\vartheta+\vartheta_{k,N}}{2}\sin\frac{\vartheta-\vartheta_{k,N}}{2}} = \frac{1}{2}\Big(\cot\frac{\vartheta+\vartheta_{k,N}}{2} - \cot\frac{\vartheta-\vartheta_{k,N}}{2}\Big)$$

to obtain the second formula. □

For the Lebesgue constants we already know $\|\mathcal{L}_N\|_{\mathrm{op}} \geqslant c\log(N+1)$. For positive convergence results discussed in §7.8.4 below, we shall need a matching upper bound and also a result for $\sum_{k=0}^N \ell_{k,N}(t)$ at the endpoint $t = 1$.

7.124. PROPOSITION. *There exist constants $C, C_1 \in (0,\infty)$ such that the following hold for all $N \geqslant 1$:*

(i)

$$\sum_{k=0}^N |\ell_{k,N}(1)| = \frac{1}{N+1}\sum_{k=0}^N \cot(\tfrac{2k+1}{4N+4}\pi) \geqslant \frac{2}{\pi}\log(N+1) - C.$$

(ii)

$$C_1^{-1}\log(N+1) \leqslant \|L_N\|_{\mathrm{op}} \leqslant C_1\log(N+1).$$

PROOF. (i) From Theorem 7.115

$$\|\mathcal{L}_N\|_{\mathrm{op}} = \sup_{\vartheta\in[0,\pi]} \frac{|\cos((N+1)\vartheta)|}{2N+2}\sum_{k=0}^N \Big|\cot\frac{\vartheta+\frac{2k+1}{2N+2}\pi}{2} - \cot\frac{\vartheta-\frac{2k+1}{2N+2}\pi}{2}\Big|$$

For $t = 1$, corresponding to $\vartheta = 0$ we get from Lemma 7.123 and $\cot\vartheta = -\cot(-\vartheta)$.

$$\sum_{k=0}^N |\ell_{k,N}(1)| = \frac{1}{N+1}\sum_{k=0}^N \cot(\tfrac{2k+1}{4N+4}\pi).$$

Observe that $|\cot s - 1/s|$ is bounded for $0 \leqslant s \leqslant \pi/2$ (use $s\cos s - \sin s = O(s^3)$ for small s). Thus

$$\begin{aligned}\frac{1}{N+1}\sum_{k=0}^N \cot(\tfrac{2k+1}{4N+4}\pi) &= \sum_{k=0}^N \frac{4}{(2k+1)\pi} + O(1) = \frac{2}{\pi}\sum_{k=0}^N \frac{1}{k+\frac{1}{2}} + O(1)\\ &= \frac{2}{\pi}\sum_{k=0}^N \frac{1}{k+1} + O(1) = \log(N+1) + O(1).\end{aligned}$$

(ii) The previous calculation also shows the lower bound $\|L_N\|_{\mathrm{op}} \geqslant \frac{2}{\pi}\log(N+1) - C$ and thus with a constant $c > 0$, $\|L_N\|_{\mathrm{op}} \geqslant c\log(N+1)$ (the latter result we already know from Theorem 7.121).

We now prove an upper bound. Set as above $\vartheta_{k,N} = \frac{2k+1}{2N+2}\pi$. Let $\vartheta \in [0,\pi]$. Then there exists an integer $l \in [0,N]$ such that $|\vartheta - \frac{2l+1}{2N+1}\pi| \leqslant \frac{\pi}{2N+1}$. Use $|\sin(\frac{\vartheta-\vartheta_{l,N}}{2})| \geqslant \frac{1}{\pi}|\vartheta - \vartheta_{l,N}|$ and from $\cos((N+1)\vartheta_{l,N}) = 0$ also $|\cos((N+1)\vartheta| \leqslant (N+1)|\vartheta - \vartheta_{l,N}|$. This leads to

$$\frac{|\cos((N+1)\vartheta)||\cot(\frac{\vartheta-\vartheta_{l,N}}{2})|}{2N+2}| \leqslant \frac{(N+1)\pi}{2N+2} = \frac{\pi}{2}.$$

Also for $|k-l| \geqslant 1$, we have $\frac{1}{2}|\vartheta - \frac{2k+1}{2N+2}\pi| \geqslant \frac{|k-l|}{2(N+1)}\pi - \frac{\pi}{4N+4}$ and thus

$$|\cot(\tfrac{\vartheta-\vartheta_{k,N}}{2})| \leqslant \frac{1}{\sin(\frac{2|k-l|\pi-\pi}{4(N+1)})} \leqslant \tfrac{2N+2}{2|k-l|-1} + C_3.$$

Hence

$$\frac{1}{2N+2} \sum_{\substack{0\leqslant k\leqslant N \\ k\neq l}} |\cot \tfrac{\vartheta-\vartheta_{k,N}}{2}| \leqslant C_4 \log(N+1).$$

Similarly

$$\frac{1}{2N+2} \sum_{0\leqslant k\leqslant N} |\cot \tfrac{\vartheta+\vartheta_{k,N}}{2}| \leqslant \frac{1}{2N+2} \sum_{0\leqslant k\leqslant N} |\cot \tfrac{\vartheta_{k,N}}{2}| \leqslant C_5 \log(N+1)$$

and combining these estimates we obtain our logarithmic upper bound. □

7.125* Exercise. Show that $\sum_{k=0}^{N} |\ell_{k,N}(t)|$ takes a global maximum at $t = 1$ and thus

$$\|L_N\|_{\mathrm{op}} = \frac{1}{N+1} \sum_{k=0}^{N} \cot(\tfrac{2k+1}{4N+4}\pi) = \frac{2}{\pi}\log(N+1) + O(1).$$

The upper bound in the proposition can be used to prove convergence results, in analogy with Theorem 7.74. For this we need to use Jackson's theorem for polynomials on $[-1,1]$ (Theorem 7.103).

7.126. Exercise. For the Lagrange interpolation operators with Chebyshev nodes, prove

$$\|L_N f - f\|_\infty \leqslant C \log(N+2)\, \omega_f(\tfrac{1}{N}).$$

In particular, $L_N f \to f$ uniformly if f is a Hölder continuous function on $[-1,1]$.

Hermite–Fejér interpolation with Chebyshev nodes. As we have seen, choosing Chebyshev nodes dramatically improves the convergence behavior of Lagrange interpolation polynomials when compared to equidistant nodes and uniform convergence only requires a relatively modest amount of regularity on the target function f. But is there an interpolation procedure that guarantees uniform convergence for *all* continuous f? A result of Fejér says that the answer is yes, but the caveat is that we need to use Hermite interpolation polynomials, i.e. also prescribe the derivatives of interpolation polynomials, which increases the degree of the polynomials. The idea here is that prescribing the derivatives and controlling their size will reduce the potentially large oscillation of interpolation polynomials (as seen in Figure 7.5).

Let $f \in C([-1,1])$ and let $t_{k,N}$ with $k = 0, \dots, N$ again denote the Chebyshev nodes as defined in (7.43). For given derivative data

$$v_N = (v_{k,N})_{k=0,\dots,N}$$

we then define the *Hermite–Fejér interpolation polynomial* $H_{N,v_N} f$ as the unique polynomial $p \in \mathcal{P}_{2N+1}$ that satisfies

$$p(t_{k,N}) = f(t_{k,N}), \quad p'(t_{k,N}) = v_{k,N}, \quad k = 0, \dots, N. \tag{7.44}$$

Recall that uniqueness follows from Theorem 7.108. The most basic case of Hermite–Féjer interpolation occurs when all the derivatives at the nodes are set to zero:

7.127. Theorem (Fejér). *For every $f \in C([-1,1])$, the Hermite–Fejér interpolation polynomials $H_{N,0} f$ converge to f uniformly on $[-1,1]$.*

This result continues to hold for certain nonzero values of the derivative data $\mathbf{v}_N$ and it is natural to ask what the precise growth condition on $\mathbf{v}_N$ that is needed. The appropriate condition involves the quantity

$$\gamma_N(v_N) = \frac{\log(N+1)}{N+1} \max_{k=0,\dots,N} |v_{k,N}|(1-t_{k,N}^2)^{\frac{1}{2}}. \tag{7.45}$$

Note that the presence of the factor $(1-t_{k,N}^2)^{1/2}$ in the definition of $\gamma_N(v_N)$ means that a size condition on $\gamma_N(v_N)$ implies a growth restriction on the $v_{k,N}$ which are weaker for the nodes near the boundary points ± 1. Assuming that $H_{N,v_N}f$ is uniformly bounded, Bernstein inequality for polynomials (see Theorem 7.136 below) gives $\max_k |v_{k,N}|(1-t_{k,N}^2)^{\frac{1}{2}} = O(N)$ which implies $\gamma_N(v_N) = O(\log(N+1))$ as $N \to \infty$. To prove convergence one needs to assume the somewhat stronger assumption $\gamma_N(v_N) \to 0$ as $N \to \infty$. We will prove the following result, which in particular implies Theorem 7.127.

7.128. Theorem (Fejér).

(i) If $\lim_{N\to\infty} \gamma_N(v_N) = 0$*, then* $H_{N,v_N}f$ *converges to* f *uniformly on* $[-1,1]$ *for all* $f \in C([-1,1])$*.*

(ii) For every nonnegative sequence $(\alpha_N)_N$ *with* $\limsup_{N\to\infty} \alpha_N > 0$ *there exists derivative data* $(v_N)_N$ *such that* $\gamma_N(v_N) \leqslant \alpha_N$ *and for all* $f \in C([-1,1])$*,* $H_{N,v_N}f$ *does not converge uniformly to* f*.*

We first need to make explicit the general Hermite polynomials associated to the Chebyshev nodes.

7.129. Lemma. *Let* $N \geqslant 0$ *and* $(y_k)_{k=0,\dots,N}$*,* $(v_k)_{k=0,\dots,N}$ *given. Then the unique polynomial* $p \in \mathcal{P}_{2N+1}$ *with* $p(t_{k,N}) = y_k, p'(t_{k,N}) = v_k$ *for all* $k = 0,\dots,N$ *is given by*

$$p(t) = \sum_{k=0}^{N} \big(y_k A_{k,N}(t) + v_k B_{k,N}(t)\big), \quad \textit{where}$$

$$A_{k,N}(t) = (1-t_{k,N}t)\Big(\frac{T_{N+1}(t)}{(N+1)(t-t_{k,N})}\Big)^2, \tag{7.46a}$$

$$B_{k,N}(t) = \frac{1-t_{k,N}^2}{t-t_{k,N}}\Big(\frac{T_{N+1}(t)}{N+1}\Big)^2. \tag{7.46b}$$

In particular, $A_{k,N}(t) \geqslant 0$ *for all* $t \in [-1,1]$*. Moreover,* $B_{k,N}(1) = (1+t_{k,N})(N+1)^{-2} > 0$*.*

Proof. We use the formulas for $A_{k,N}$, $B_{k,N}$ in (7.36) from Theorem 7.108. Indeed (7.46b) is immediate from (7.36b) and Lemma 7.123. From Lemma 7.97 we get the formulas $T'_{N+1}(t_k) = (-1)^k(N+1)(1-t_{k,N}^2)^{-1/2}$ and $T''_{N+1}(t_k) = (-1)^k(N+1)t_{k,N}(1-t_{k,N}^2)^{-3/2}$. Using the formula for the elementary Lagrange polynomials in Lemma 7.123 and Taylor expanding T_{N+1} about $t_{k,N}$ gives

$$\ell_{k,N}(t) = 1 + \frac{T''_{N+1}(t_{k,N})}{2T'_{N+1}(t_{k,N})}(t-t_{k,N}) + O((t-t_{k,N})^2)$$

so that $\ell'_{k,N}(t_{k,N}) = \frac{T''_{N+1}(t_{k,N})}{2T'_{N+1}(t_{k,N})} = \frac{1}{2}t_{k,N}(1-t_{k,N}^2)^{-1}$. Hence

$$\begin{aligned} A_{k,N}(t) &= \big(1-2(t-t_{k,N})\ell'_{k,N}(t_{j,N})\big)(\ell_{k,N}(t))^2 \\ &= \Big(1-2(t-t_{k,N})\frac{t_{k,N}}{2(1-t_{k,N}^2)}\Big)\frac{T_{N+1}(t)^2}{(N+1)^2(t-t_{k,N})^2} \end{aligned}$$

which simplifies to (7.46a). Since all the $t_{k,N}$ lie in $(-1,1)$ the expression $1-t_{k,N}t$ is positive in $(-1,1)$ and thus $A_{k,N}(t) \geqslant 0$ for $t \in [-1,1]$. □

Proof of Theorem 7.128. By Lemma 7.129,

$$H_{N,\mathbf{v}_N}f(t) = \sum_{k=0}^{N} \big(f(t_{k,N})A_{k,N}(t) + v_{k,N}B_{k,N}(t)\big).$$

We first handle the part depending linearly on f,

$$F_N f(t) = \sum_{k=0}^{N} f(t_{k,N})A_{k,N}(t).$$

Let us prove that $\|F_N\|_{\mathrm{op}} = 1$ and that $(F_N f)$ converges uniformly to f. Applying the Hermite interpolation formula in Theorem 7.108 to the constant function 1 (and zero derivative data) gives

$$\sum_{k=0}^{N} A_{k,N}(t) = F_N(1) = 1.$$

It is now crucial that the functions $A_{k,N}$ are nonnegative on $[-1,1]$, so we may estimate

$$\|F_N f\|_\infty \leqslant \|f\|_\infty \sup_{t\in[-1,1]} \sum_{k=0}^{N} |A_{k,N}(t)| = \|f\|_\infty \sup_{t\in[-1,1]} \Big| \sum_{k=0}^{N} A_{k,N}(t)\Big| = \|f\|_\infty.$$

Hence $\|F_N\|_{\mathrm{op}} \leqslant 1$ and since $F_N 1 = 1$ we get $\|F_N\|_{\mathrm{op}} = 1$.

To prove uniform convergence let $f \in C([-1,1])$ and $\varepsilon > 0$. Then there is $\delta > 0$ such that $|f(t) - f(t')| < \varepsilon/2$ provided that $|t - t'| \leqslant \delta$. Since $F_N(1) = 1$,

$$F_N f(t) - f(t) = \sum_{k=0}^{N} f(t_{k,N}) A_{k,N}(t) - f(t) \sum_{k=0}^{N} A_{k,N}(t) = \sum_{k=0}^{N} (f(t_{k,N}) - f(t)) A_{k,N}(t).$$

Let $\mathcal{I}(t) = \{k : 0 \leqslant k \leqslant N, |t - t_{k,N}| \leqslant \delta\}$. We then split $F_N f(t) - f(t) = I(t) + II(t)$ where

$$I(t) = \sum_{k\in\mathcal{I}(t)} (f(t_{k,N}) - f(t)) A_{k,N}(t),$$
$$II(t) = \sum_{k:k\notin\mathcal{I}(t)} (f(t_{k,N}) - f(t)) A_{k,N}(t).$$

By the positivity of the $A_{k,N}$,

$$|I(t)| \leqslant \frac{\varepsilon}{2} \sum_{k\in\mathcal{I}(t)} |A_{k,N}(t)| = \frac{\varepsilon}{2} \sum_{k\in\mathcal{I}(t)} A_{k,N}(t) = \frac{\varepsilon}{2}.$$

Now we use (7.46a). Since $\|T_{N+1}\|_\infty \leqslant 1$ we have $A_{k,N}(t) \leqslant 2\delta^{-2}(N+1)^{-2}$ for $k \notin \mathcal{I}(t)$. Hence

$$|II(t)| \leqslant \|f\|_\infty \sum_{k\notin\mathcal{I}(t)} 2\delta^{-2}(N+1)^{-2} \leqslant \|f\|_\infty 2\delta^{-2}(N+1)^{-1}.$$

We can find $N_\varepsilon \in \mathbb{N}$ so that $\|f\|_\infty 2\delta^{-2}(N+1)^{-1} < \varepsilon/2$ for $N \geqslant N_\varepsilon$. Then combining the two bounds we get

$$\|F_N f - f\|_\infty < \varepsilon \text{ for } N \geqslant N_\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the uniform convergence follows. The polynomials F_N are precisely the Hermite-Fejér polynomials $H_{N,0}f$ for zero derivative data used in Theorem 7.127 and thus we have proved that theorem. We now consider the polynomials

$$q_N(v_N, t) = \sum_{k=0}^{N} v_{k,N} B_{k,N}(t).$$

Then $q_N(v_N, \cdot) \to 0$ uniformly implies that for every $f \in C([-1,1])$ the sequence $H_{N,v_N} f$ converges to f uniformly. Vice versa, if $H_{N,v_N} f$ converges uniformly to f for some $f \in C([-1,1])$, then $q_N(v_N, \cdot) \to 0$ uniformly.

We thus show (i) by showing $\|q_N(v_N,\cdot)\|_\infty \to 0$ if $\lim_{N\to\infty}\gamma_N(v_N)=0$. For the following estimate we assume $N \geqslant 1$ and initially that t is not one of the $t_{k,N}$.

$$\begin{aligned}|q_N(v_N,t)| &= \Big|\sum_{k=0}^N v_{k,N}B_{k,N}(t)\Big| = \Big|\sum_{k=0}^N v_{k,N}\frac{1-t_{k,N}^2}{t-t_{k,N}}\Big(\frac{T_{N+1}(t)}{N+1}\Big)^2\Big| \\ &\leqslant \sum_{k=0}^N \frac{|v_{k,N}|(1-t_{k,N}^2)^{\frac12}}{N+1}\frac{|T_{N+1}(t)|}{(N+1)|t-t_{k,N}|}(1-t_{k,N}^2)^{\frac12}|T_{N+1}(t)| \\ &\leqslant \gamma_N(v_N)\frac{1}{\log(N+1)}\sum_{k=0}^N|\ell_{k,N}(t)| \leqslant \gamma_N(v_N)\frac{\|L_N\|_{\mathrm{op}}}{\log(N+1)}\end{aligned}$$

Here we have used the formula for $B_{k,N}$ in Lemma 7.129, the estimate $\|T_{N+1}\|_\infty \leqslant 1$, the formula for the elementary Lagrange polynomials in Lemma 7.123 and the expression of the operator norm in Theorem 7.115. The stated estimate remains true at the nodes, by continuity. By Proposition 7.124, $\|L_N\|_{\mathrm{op}}/\log(N+1)$ is bounded as $N\to\infty$ and thus the assumption $\lim_{N\to\infty}\gamma_N(v_N)=0$ implies that $q_N(v_N,\cdot)\to 0$ uniformly on $[-1,1]$.

Let α_N be a given sequence of nonnegative numbers such that $\limsup \alpha_N > 0$. Then there is $c_1>0$ and a strictly increasing sequence of integers $j(n)$, with $j(1)\geqslant 2$ such that $\alpha_{j(n)}\geqslant c_1$ for all $n\in\mathbb{N}$. By Proposition 7.124, $\sum_{k=0}^N|\ell_k(1)|\geqslant c_2\log(N+1)$ for some $c_2>0$.

Now define $v_{k,N}=0$ if N is not one of the $j(n)$, $n\in\mathbb{N}$. For $N=j(n)$ set

$$v_{k,j(n)} = \frac{\alpha_{j(n)}(j(n)+1)}{\log(j(n)+1)(1-t_{k,j(n)}^2)^{1/2}}.$$

Then $\gamma_N(v_N)\leqslant\alpha_N$ for all N. Note that the quantity $v_{k,N}(1-t_{k,N}^2)^{1/2}$ does not depend on k for this choice. For $N=j(n)$ and $t=1$ we may reverse the chain of inequalities above and obtain

$$\begin{aligned}|q_{j(n)}(v_{j(n)},1)| &= \Big|\sum_{k=0}^{j(n)} v_{k,j(n)}B_{k,j(n)}(1)\Big| = \sum_{k=0}^{j(n)}|v_{k,j(n)}||B_{k,j(n)}(1)| \\ &= \sum_{k=0}^{j(n)}\frac{|v_{k,j(n)}|(1-t_{k,j(n)}^2)^{\frac12}}{j(n)+1}\frac{|T_{j(n)+1}(1)|}{(j(n)+1)|1-t_{k,j(n)}|}(1-t_{k,j(n)}^2)^{\frac12}|T_{j(n)+1}(1)| \\ &= \alpha_{j(n)}\frac{1}{\log(j(n)+1)}\sum_{k=0}^{j(n)}|\ell_{k,j(n)}(1)| \geqslant c_1c_2>0.\end{aligned}$$

Consequently $q_N(v_N,t)$ does not converge uniformly to 0 and hence $(H_{N,v_N}f)$ does not converge uniformly to f, for any $f\in C([-1,1])$. □

REMARK. The Hermite-Fejér polynomials interpolate at the $N+1$ Chebyshev nodes, but are of degree $2N+1$. It is natural to ask whether this degree can be lowered (while still keeping the uniform convergence result). By Corollary 7.120 we cannot lower the degree to N. For given small $\delta>0$, Bernstein gave a linear interpolation method which assigns to $f\in C([-1,1])$ a linear operator $\mathcal{B}_{N,\delta}$ such that $\mathcal{B}_{N,\delta}f$ is a polynomial of degree at most $N(1+\delta)$ which interpolates f at the Chebyshev nodes $t_{k,N}$ and such that $\mathcal{B}_{N,\delta}f$ converges uniformly to f for all $f\in C([-1,1])$. See Natanson [**56**, IV.4].

7.8.5. Mean-square convergence for Gaussian interpolation. We now consider a sequence of non-zero polynomials (p_n), of exact degree n which are *orthogonal* with respect to the scalar product

$\langle f, g\rangle_w = \int_a^b f(t)g(t)w(t)\,dt$:

$$\int_a^b p_n(t)p_m(t)w(t)dt = 0 \ \text{ if } n \neq m \tag{7.47}$$

for all non-negative integers n, m (*cf.* Examples 7.35, 7.36). We first observe a crucial property about the zeros of such polynomials (generalizing the situation for Chebyshev polynomials in Lemma 7.99).

7.130. THEOREM. *Let $Q \in \mathcal{P}_{N+1}\backslash\{0\}$ and let Q satisfy $\langle Q, p\rangle_w = 0$ for all $p \in \mathcal{P}_N$. Then*

(i) Q has $N+1$ simple zeros $t_0, \ldots, t_N$ which lie in the open interval (a, b).

(ii) $Q = c\Omega$ where $c \neq 0$ and $\Omega(t) = \prod_{k=0}^N (t - t_k)$.

PROOF. For (i) we argue by contradiction and assume that Q has less than $N+1$ different zeros in (a, b) (not counted with multiplicity). If Q does not change sign in (a, b) then $\int_a^b Q(t)w(t)dt \neq 0$, contradicting that Q and 1 are orthogonal. Assume that Q changes sign in (a, b) exactly k times, with $1 \leqslant k \leqslant N$, and let $\tau_1 < \cdots < \tau_k$ be the points where Q changes its sign. Let $q_k(t) = \prod_{j=1}^k (t - \tau_j)$, a real polynomial of degree k which by assumption is orthogonal to Q. Observe that by the definition of the τ_j the function q_k changes sign when Q changes sign, i.e. Qq_k is either nonnegative in (a, b) or nonpositive in (a, b) (and of course nonzero away from the τ_j). This implies $\int_a^b Q(t)q_k(t)w(t)dt \neq 0$, a contradiction. This proves (i).

The space of polynomials of degree N which are orthogonal to all polynomials of degree $\leqslant N-1$ is one-dimensional and thus generated by a polynomial Q. Since t_0, ..., t_N are the simple zeros of Q we must have $Q(t) = c\prod_{k=0}^N (t - t_k)$ for some $c \neq 0$. This proves (ii). □

We now take the zeros of orthogonal polynomials as the nodes in our Lagrange interpolation polynomials.

7.131. DEFINITION. We are given the sequence of orthogonal polynomials $p_n \in \mathcal{P}_n$ satisfying (7.47). Let $t_{0,N}, \ldots, t_{N,N}$ be the $N+1$ simple zeros of p_{N+1} in (a, b) as in Theorem 7.130. The Lagrange interpolation operator L_N with respect to *Gaussian nodes* is given by

$$L_N f(t) = \sum_{k=0}^N f(t_{k,N})\ell_{k,N}(t).$$

We have the following mean-square convergence result which is due to Erdős and Turán [**18**].

7.132. THEOREM (Erdős–Turán). *Let $f \in C([a, b])$ and, for $N \geqslant 1$ let $L_N f \in \mathcal{P}_N$ be the Lagrange interpolation polynomial with respect to the Gaussian nodes $(t_{k,N})_{k=0}^N$. Then*

$$\lim_{N\to\infty} \Big(\int_a^b |L_N f(t) - f(t)|^2 w(t)dt \Big)^{1/2} = 0.$$

Before proving this theorem we note a crucial orthogonality property of the elementary Lagrange polynomials with Gaussian nodes.

7.133. LEMMA. *Let $N \geqslant 1$. The elementary Lagrange polynomials $\ell_{j,N}$, $j = 0, \ldots, N$ associated with the zeros of p_{N+1} satisfy the orthogonality relations*

$$\int_a^b \ell_{j,N}(t)\ell_{k,N}(t)w(t)dt = 0, \quad \text{for } j \neq k,\ 0 \leqslant j, k \leqslant N.$$

Moreover,

$$\sum_{j=0}^N \int_a^b [\ell_{j,N}(t)]^2 w(t)dt = \int_a^b w(t)dt.$$

PROOF. We first note that by Theorem 7.130 the polynomial $\Omega_{N+1}(t) = \prod_{j=0}^{N}(t - t_{j,N})$ is a multiple of p_{N+1} and satisfies

$$\int_a^b \Omega_{N+1}(t)p(t)w(t)dt = 0$$

for all $p \in \mathcal{P}_N$. If $j \neq k$ the function $t \mapsto \Omega_{N+1}(t)(t - t_{j,N})^{-1}(t - t_{k,N})^{-1}$ is a polynomial in $\mathcal{P}_{N-1}$ and thus orthogonal to Ω_{N+1}. We use the formula from Lemma 7.105 for the elementary Lagrange polynomials $\ell_{k,N}$, $k = 0, \dots, N$ and obtain for $j \neq k$

$$\int_a^b \ell_{j,N}(t)\ell_{k,N}(t)w(t)dt = \frac{1}{\Omega'_{N+1}(t_{j,N})\Omega'_{N+1}(t_{k,N})}\int_a^b \Omega_{N+1}(t)\frac{\Omega_{N+1}(t)}{(t - t_{j,N})(t - t_{k,N})}w(t)dt = 0$$

which shows the mutual orthogonality of the $\ell_{j,N}$ for fixed $N \geqslant 1$ and $j = 0, \dots, N$. Moreover, writing down the interpolation polynomial for $f(t) = 1$ reproduces the function 1 and we get

$$1 = \sum_{j=0}^{N} \ell_{j,N}(t)$$

for $t \in [a, b]$. Hence by the Pythagorean theorem,

$$\int_a^b w(t)dt = \int_a^b \Big|\sum_{j=0}^{N} \ell_{j,N}(t)\Big|^2 w(t)dt = \sum_{j=0}^{N}\int_a^b (\ell_{j,N}(t))^2 w(t)dt. \qquad \square$$

PROOF OF THEOREM 7.132. We use the norm notation $\|g\|_2 = (\int_a^b |g(t)|^2 w(t)dt)^{1/2}$. Let $M = \int_a^b w(t)dt < \infty$. Let $\varepsilon > 0$. By the Weierstrass approximation theorem there is a polynomial Q such that

$$\max\big\{|f(t) - Q(t)| : t \in [a, b]\big\} < \frac{\varepsilon}{2M^{1/2}}.$$

Then

$$\|f - Q\|_2^2 \leqslant \frac{\varepsilon^2}{4M}\int_a^b w(t)dt = \frac{\varepsilon^2}{4}.$$

Let $N_0(\varepsilon)$ be the degree of Q. Note that $L_N Q = Q$ for $N \geqslant N_0(\varepsilon)$ which we assume now. By the triangle inequality

$$\|f - L_N f\|_2 \leqslant \|f - Q\|_2 + \|Q - L_N f\|_2 \leqslant \frac{\varepsilon}{2} + \|L_N Q - L_N f\|_2.$$

To estimate the second term we get

$$\begin{aligned}
\|L_N Q - L_N f\|_2^2 &= \int_a^b \Big|\sum_{k=0}^{N} \big(Q(t_{k,N}) - f(t_{k,N})\big)\ell_{k,N}(t)\Big|^2 w(t)dt \\
&= \sum_{k=0}^{N}\int_a^b |Q(t_{k,N}) - f(t_{k,N})|^2 |\ell_{k,N}(t)|^2 w(t)dt \leqslant \frac{\varepsilon^2}{4M}\sum_{k=0}^{N}\int_a^b |\ell_{k,N}(t)|^2 w(t)dt \\
&= \frac{\varepsilon^2}{4M}\int_a^b w(t)dt = \frac{\varepsilon^2}{4}.
\end{aligned}$$

We have proved $\|L_N Q - L_N f\|_2 < \varepsilon/2$ for $N \geqslant N_0(\varepsilon)$, and thus $\|f - L_N f\|_2 < \varepsilon$ for $N \geqslant N_0(\varepsilon)$. Since $\varepsilon > 0$ was arbitrary, this completes the proof. $\square$

In particular, for Chebyshev interpolation we get:

7.134. Corollary. *Let $N \in \mathbb{N}$, let $t_{k,N} = \cos(\frac{2k+1}{2N+2}\pi)$ be the Chebyshev nodes, i.e. the $N+1$ zeros of T_{N+1} in $(-1,1)$. For $f \in C([-1,1])$ let $\mathcal{L}_N f$ be the Lagrange interpolation polynomial with $\mathcal{L}_N f(t_{k,N}) = f(t_{k,N})$. Then*

$$\lim_{N\to\infty} \int_{-1}^{1} |f(t) - \mathcal{L}_N f(t)|^2 \frac{dt}{\sqrt{1-t^2}} = 0.$$

This is an immediate consequence of Theorem 7.132 and the orthogonality properties of the Chebyshev polynomials.

7.9. Bernstein and Markov inequalities

In this section we prove some classical inequalities for trigonometric and algebraic polynomials in their sharp form. Interpolation identities will play a crucial role in some of the proofs. We first give the sharp form of Bernstein's inequality for trigonometric polynomials, improving on the constant in Proposition 7.70.

7.135. Theorem (Bernstein's inequality). *For all $p \in \mathcal{T}_N$,*

$$\|p'\|_\infty \leqslant 2\pi N \|p\|_\infty.$$

Equality is achieved for $f(x) = e^{2\pi i N x}$ and its real and imaginary parts.

Proof. Without loss of generality we may assume $N \geqslant 2$. We first assume that p is real valued, $\|p\|_\infty = 1$ and that p' is maximal at 0. Note that we then have $p''(0) = 0$. We argue by contradiction and assume that there is $R > 1$ such that

$$\|p'\|_\infty = p'(0) = 2\pi R N > 2\pi N \|p\|_\infty.$$

Consider the function

$$\Phi(x) = R\sin(2\pi N x) - p(x)$$

which satisfies $\Phi'(0) = 0$ and $\Phi''(0) = 0$. To achieve the desired contradiction it suffices to prove that Φ has $2N+1$ different zeros. Indeed, by Corollary 7.110 this would imply $\Phi = 0$ and therefore $p(x) = R\sin(2\pi N x)$ which gives a contradiction.

To show that Φ has $2N+1$ different zeros, let us analyze the zeros of Φ, Φ' and Φ''.

Let $\xi_k = \frac{2k-1}{4N}$, $k = 1, \dots, 2N$ and $\xi_0 = \xi_{2N} - 1$ and note that

$$\Phi(\xi_k) = R\sin(\tfrac{2k-1}{2}\pi) - p(\xi_k) = R(-1)^{k-1} - p(\xi_k).$$

Since $\|p\|_\infty = 1$ and $R > 1$ we see that $\Phi(\xi_k) > 0$ if k is odd and $\Phi(\xi_k) < 0$ if k is even. Therefore, by the intermediate value theorem there exist $\eta_k \in (\xi_{k-1}, \xi_k)$ such that $\Phi(\eta_k) = 0$, $k = 1, \dots 2N$.

Setting $\eta_0 = \eta_{2N} - 1$ we see by Rolle's theorem that there exist $\tau_j \in (\eta_{j-1}, \eta_j)$, $j = 1, \dots, 2N$ such that $\Phi'(\tau_j) = 0$. Recall that $\Phi'(0) = 0$, and so if all of the τ_j are different from 0 then Φ' has $2N+1$ zeros, and thus Φ is constant, and then necessarily identically 0; so we get the contradiction. Thus we may assume that 0 is among the points $\{\tau_j : j = 1, \dots, 2N\}$. Let $\tau_0 = \tau_{2N} - 1$. Since $\Phi'(\tau_j) = 0$ for $j = 0, \dots, 2N$ it is clear that Φ'' must have a zero in each interval (τ_{j-1}, τ_j), $j = 1, \dots, 2N$, and by our assumption 0 is not one of them. But we know $\Phi''(0) = 0$. Hence it follows that $\Phi'' \in \mathcal{T}_N$ has at least $2N+1$ different zeros in $(\tau_0, \tau_{2N}] = (\tau_0, \tau_0 + 1]$ which implies that $\Phi'' = 0$. Thus Φ' is constant, and therefore 0, which also implies that $\Phi = 0$ which yields a contradiction.

Summarizing we have proved that if $\|p\|_\infty = 1$, p' is maximal at 0 then $p'(0) \leqslant 2\pi N$. If p' is maximal at some x_0 we apply the previous situation to $q(x) = p(x_0 + x)$ and deduce that $p'(x_0) \leqslant 2\pi N$. This shows the asserted inequality for all real-valued polynomials of degree at most N, with $\|p\|_\infty = 1$.

It remains to prove the assertion for complex-valued polynomials. Assume that $|p'(x)|$ is maximal at x_0. We find an $\alpha \in \mathbb{R}$ such that $e^{-i\alpha} p'(x_0)$ is real and nonnegative. Then we may write

$$e^{-i\alpha} p'(x) = p_1(x) + i p_2(x),$$

where p_1 and p_2 are real valued trigonometric polynomials of degree at most N, with $p_2'(x_0) = 0$. By the above we can estimate

$$|p'(x_0)| = p_1(x_0) \leqslant 2\pi N \|p_1\|_\infty \leqslant 2\pi N \|p_1 + ip_2\|_\infty = 2\pi N \|p\|_\infty$$

and we are done. □

We get a related theorem for polynomials on the interval $[-1, 1]$.

7.136. THEOREM (Bernstein's inequality for polynomials). *For a polynomial $q \in \mathcal{P}_N$ set $\|q\|_\infty = \max_{t\in[-1,1]} |q(t)|$. Then*

$$|q'(t)| \leqslant \frac{N\|q\|_\infty}{(1-t^2)^{1/2}}, \quad \textit{for all } t \in [-1, 1].$$

PROOF. We consider the trigonometric polynomial $p(x) = q(\cos 2\pi x)$ and apply Theorem 7.135 to get $\sup_{x\in[0,1]} |p'(x)| \leqslant 2\pi N \sup_{x\in[0,1]} |p(x)|$, which gives

$$|2\pi| \sin 2\pi x||q'(\cos 2\pi x)| \leqslant 2\pi N \max_{x\in[0,1]} |q(\cos 2\pi x)|$$

and setting $t = \cos 2\pi x$ with $|\sin 2\pi x| = (1-t^2)^{1/2}$ yields the asserted inequality. □

Bernstein's inequality for polynomials is quite efficient, except for t very close to the boundary of the interval $[-1, 1]$. In order to control the behavior of a polynomial at the boundary one may use *Schur's inequality*

7.137. THEOREM (Schur). *Let $q \in \mathcal{P}_N$. Then*

$$\max_{t\in[-1,1]} |q(t)| \leqslant (N+1) \max_{t\in[-1,1]} |q(t)|(1-t^2)^{1/2}.$$

Equality is achieved for $q = T_{N+1}'$.

PROOF. We first consider $q(t) = T_{N+1}'(t)$ which has maximum $(N+1)^2$. We have

$$|(N+1)T_{N+1}'(t)(1-t^2)^{1/2}| = (N+1)^2 |\sin((N+1)\arccos t)|$$

and the right hand side is bounded by $(N+1)^2$, and if we evaluate at the points $t_k = \cos(\frac{2k+1}{2N+2}\pi)$, $k = 0, \dots, N$ i.e. the zeros of the Chebyshev polynomial T_{N+1} then we get $(N+1)^2$ on the right hand side. Thus equality is achieved for $q = T_{N+1}'$.

To show the inequality we normalize and assume the right hand side is

$$\max_{t\in[-1,1]} |q(t)|(1-t^2)^{1/2} = 1.$$

We thus need to prove $\|q\|_\infty \leqslant N+1$. Let $t_0 = \cos(\frac{\pi}{2N+2})$ and $t_N = -t_0$. We separately estimate $q(t)$ for $|t| \leqslant t_0$, $t_0 < t \leqslant 1$, and $-1 \leqslant t < t_N = -t_0$.

For $|t| \leqslant t_0 = \cos\frac{\pi}{2N+2}$ we have (by the normalization assumption and the inequality $\sin s \geqslant \frac{2}{\pi}s$ for $0 \leqslant s \leqslant \pi/2$)

$$|q(t)| \leqslant (1-t^2)^{-1/2} \leqslant (1-t_0^2)^{-1/2} = \left|\sin(\tfrac{\pi}{2N+2})\right|^{-1} \leqslant N+1.$$

Now let $t_0 < t \leqslant 1$. Any Lagrange interpolation polynomial of degree N applied to a polynomial reproduces the polynomial; we apply this with the nodes t_k. Therefore we get from Theorem 7.106 (with $\Omega(t) = \prod_{j=0}^N (t - t_j) = 2^N T_{N+1}(t)$)

$$q(t) = \sum_{k=0}^N q(t_k)\ell_k(t) = \sum_{k=0}^N q(t_k) \frac{T_{N+1}(t)}{T_{N+1}'(t_k)(t-t_k)}. \tag{7.48}$$

Observe that $T'_{N+1}(t_k) = (N+1)\sin((N+1)\arccos t_k)(1-t_k^2)^{-1/2} = (-1)^k(N+1)(1-t_k^2)^{-1/2}$. The last display therefore gives the estimate

$$(N+1)|q(t)| \leqslant \sum_{k=0}^{N} |q(t_k)|(1-t_k^2)^{1/2}\Big|\frac{T_{N+1}(t)}{t-t_k}\Big| \leqslant \sum_{k=0}^{N}\Big|\frac{T_{N+1}(t)}{t-t_k}\Big|$$

and since $T_{N+1}(t)/(t-t_k)$ is positive for $t > t_0$, we see using (7.48) for the polynomial T'_{N+1} that

$$(N+1)|q(t)| \leqslant \sum_{k=0}^{N}\frac{T_{N+1}(t)}{t-t_k} = T'_{N+1}(t), \quad t_0 < t < 1.$$

Now $T''_{N+1}(t) > 0$ for $t_0 < t < 1$ (see Lemma 7.99 (vii)). Thus T'_{N+1} is increasing on $(t_0, 1)$ and $T'_{N+1}(1) = (N+1)^2$, so we obtain the bound $|q(t)| \leqslant N+1$ for $t_0 \leqslant t \leqslant 1$.

Applying the same to $q(-t)$ yields $|q(t)| \leqslant N+1$ for $-1 \leqslant t \leqslant -t_1 = t_N$ and the proof is complete. □

One can combine the two previous theorems to get *Markov's inequality* .

7.138. Theorem (A. Markov). *On the interval* $[-1,1]$*, the inequality*

$$\|q'\|_\infty \leqslant N^2\|q\|_\infty\,.$$

holds for all $q \in \mathcal{P}_N$*. Equality is achieved for* $q = T_N$*, the Chebyshev polynomial.*

Proof. For the Chebyshev polynomial we have $\|T_N\|_\infty = 1$ and $T'_N(1) = N^2$, from Lemma 7.99.

Let q be any polynomial of degree N. By Schur's inequality applied to the $N-1$ degree polynomial q' we get

$$\|q'\| \leqslant N \max_{t\in[-1,1]} |q'(t)|(1-t^2)^{1/2}$$

and by Bernstein's inequality for polynomials,

$$|q'(t)| \leqslant N(1-t^2)^{-1/2}\|q\|_\infty.$$

Combining the two gives Markov's inequality. □

Remark. Bernstein's inequality can be iterated to give sharp inequalities for higher derivatives

$$\max_{[0,1]} |p^{(k)}(x)| \leqslant (2\pi N)^k \max_{[0,1]} |p(x)|, \quad p \in \mathcal{T}_N.$$

In contrast, iterating A. Markov's inequality gives

$$\max_{t\in[-1,1]} |q^{(k)}(t)| \leqslant C_N \max_{t\in[-1,1]} |q(t)|, \quad q \in \mathcal{P}_N.$$

with $C_N = N^2(N-1)^2\cdots(N-k+1)^2$ but this is not sharp. A result by V. Markov says that C_N can be replaced with $T_N^{(k)}(1) = C_N/(\prod_{j=1}^{k}(2j-1))$ which is optimal (since we get equality for $q = T_N$). For much more on polynomial inequalities see the monograph by Borwein and Erdélyi [**8**] and, regarding the Markov brothers' inequalities, the article by Shadrin [**69**].

7.10. The Müntz approximation theorem

Let $\Lambda = (\lambda_n)_{n\in\mathbb{N}_0}$ be a given sequence of increasing nonnegative real numbers

$$0 \leqslant \lambda_0 < \lambda_1 < \lambda_2 < \cdots \qquad \text{with } \lim_{j\to\infty}\lambda_j = \infty. \tag{7.49}$$

We shall consider the approximation of a continuous function f on the interval $[0,1]$ by *Müntz polynomials* which are defined as linear combinations of the monomial functions x^{λ_j}, i.e.

$$P_n(x) = \sum_{j=0}^{n} a_j x^{\lambda_j} \tag{7.50}$$

for some $n \in \mathbb{N}_0$. For fixed N we let $\Lambda_N = \{\lambda_0, \ldots, \lambda_N\}$ and denote by $\mathcal{P}(\Lambda_N)$ the space of functions of the form (7.50) with $n \leqslant N$. The space of all Müntz polynomials is denoted by

$$\mathcal{P}(\Lambda) = \bigcup_{N \geqslant 0} \mathcal{P}(\Lambda_N).$$

Our main goal in this section is to prove a beautiful theorem of Müntz [**53**] which characterizes the sets Λ of exponents for which all continuous functions on $[0,1]$ can be uniformly approximated by Müntz polynomials. Specifically, it states that $\mathcal{P}(\Lambda)$ is dense in $C([0,1])$ if and only if $\lambda_0 = 0$ and $\sum_{j=1}^{\infty} \lambda_j^{-1}$ diverges. Note that typically $\mathcal{P}(\Lambda)$ does not form an algebra of functions, so that this result usually falls outside the scope of the Stone-Weierstrass theorem (Theorem 7.17).

7.10.1. The main theorems. For a continuous function f on $[0,1]$ we denote the uniform best approximation error by

$$\operatorname{dist}_\infty(f, \mathcal{P}(\Lambda)) = \inf_{p \in \mathcal{P}(\Lambda)} \|f - p\|_\infty,$$

where as usual, $\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|$. Note that since the functions $x \to x^{\lambda_j}$ have a zero at $x = 0$ if $\lambda_j > 0$, constant functions can be uniformly approximated by Müntz polynomials only when $0 \in \Lambda$, i.e. with our enumeration, when $\lambda_0 = 0$. The monomials will play a special role and we define for $\mu \geqslant 0$ and $x \in [0,1]$:

$$f_\mu(x) = x^\mu.$$

The Müntz theorem for uniform approximation is as follows:

7.139. Theorem. *With Λ as in (7.49):*

(i) $\mathcal{P}(\Lambda)$ is dense in $C([0,1])$ (with respect to the uniform norm) if and only if $\lambda_0 = 0$ and

$$\sum_{j=1}^{\infty} \lambda_j^{-1} = \infty.$$

(ii) If $\sum_{j=1}^{\infty} \lambda_j < \infty$ and $\mu \geqslant 0$, then $\operatorname{dist}_\infty(f_\mu, \mathcal{P}(\Lambda)) > 0$ if and only if $\mu \notin \Lambda$.

Note that $\mathcal{P}(\Lambda)$ is dense in $C([0,1])$ if and only if for every $f \in C([0,1])$,

$$\operatorname{dist}_\infty(f, \mathcal{P}(\Lambda)) = 0.$$

7.140. Example. Let $(p_n)_{n \geqslant 1}$ be the sequence of all prime numbers (in increasing order). Let

$$\Lambda = \{0\} \cup \{p_n \,:\, n \in \mathbb{N}\}.$$

Then $\mathcal{P}(\Lambda)$ is dense in $C([0,1])$. This is a consequence of Theorem 7.139 and Example 3.87.

The main tool in proving Theorem 7.139 is a similar result for approximation in the L^2 norm $\|f\|_2 = (\int_0^1 |f(x)|^2 dx)^{1/2}$. The best approximation error in L^2 is denoted by

$$\operatorname{dist}_2(f, \mathcal{P}(\Lambda)) = \inf_{p \in \mathcal{P}(\Lambda)} \|f - p\|_2.$$

7.141. Theorem. *Let Λ be as in (7.49). Then:*

(i) We have $\operatorname{dist}_2(f, \mathcal{P}(\Lambda)) = 0$ for all continuous functions f on $[0,1]$ if and only if $\sum_{j=1}^{\infty} \lambda_j^{-1} = \infty$.

(ii) If $\sum_{j=1}^{\infty} \lambda_j < \infty$ and $\mu \geqslant 0$, then $\operatorname{dist}_2(f_\mu, \mathcal{P}(\Lambda)) > 0$ if and only if $\mu \notin \Lambda$.

We first show that the uniform approximation result Theorem 7.139 follows from Theorem 7.141, also using the Weierstrass approximation theorem (Theorem 7.3).

PROOF THAT THEOREM 7.141 IMPLIES THEOREM 7.139. First note that the condition $\lambda_0 = 0$ is necessary. Indeed, if λ_0 (and therefore all λ_j) were positive then we would have $p(0) = 0$ for all $p \in \mathcal{P}(\Lambda)$ and thus for $f_0(x) = 1$, $\operatorname{dist}_\infty(f_0, \mathcal{P}(\Lambda)) = 1$.

Next, we have $\|f - p\|_2 \leqslant \|f - p\|_\infty$ and thus

$$\operatorname{dist}_\infty(f, \mathcal{P}(\Lambda)) \geqslant \operatorname{dist}_2(f, \mathcal{P}(\Lambda)).$$

Hence the necessity of divergence of $\sum_{j=1}^\infty \lambda_j^{-1}$ and part (ii) of Theorem 7.139 follow from the corresponding statements in Theorem 7.141 (note that for the 'only if' part of (ii) there is nothing to show, because if $\mu \in \Lambda$, then $f_\mu \in \mathcal{P}(\Lambda)$.).

Now let us assume that $\lambda_0 = 0$ and $\sum_{j=1}^\infty \lambda_j^{-1} = \infty$. Also let $M \in \mathbb{N}$ be such that $\lambda_j > 1$ for $j \geqslant M$. Let $f \in C([0,1])$ and $\varepsilon > 0$. By the Weierstrass approximation theorem there is a (standard) polynomial

$$q(x) = \sum_{j=0}^K b_j x^j.$$

such that $\|f - q\|_\infty < \varepsilon/2$. Let $A = 2(1 + \sum_{j=0}^K |b_j|)$. It now suffices to find, for $j = 0, \dots, K$, Müntz polynomials $p_j \in \mathcal{P}(\Lambda)$ such that

$$\sup_{x \in [0,1]} |x^j - p_j(x)| < \frac{\varepsilon}{A}. \tag{7.51}$$

Indeed, we can then set $p(x) = \sum_{j=0}^K b_j p_j(x)$. Then $p \in \mathcal{P}(\Lambda)$, and

$$|q(x) - p(x)| \leqslant \sum_{j=0}^K |b_j||x^j - p_j(x)| < \varepsilon/2$$

for all $x \in [0,1]$, so $\|f - p\|_\infty < \varepsilon$. Since $\varepsilon > 0$ was arbitrary and $p \in \mathcal{P}(\Lambda)$, it follows that $\operatorname{dist}_\infty(f, \mathcal{P}(\Lambda)) = 0$ as required. Thus it suffices to show (7.51).

To show (7.51), note that the constant function $f_0(x) = 1$ belongs to $\mathcal{P}(\Lambda)$, by assumption that $\lambda_0 = 0$. Thus we may assume $j \geqslant 1$ in what follows.

We let

$$\widetilde{\Lambda} = \{\lambda_k - 1 : \lambda_k \in \Lambda, k \geqslant M\}.$$

Then

$$\sum_{k \geqslant M} (\lambda_k - 1)^{-1} \geqslant \sum_{k \geqslant M} \lambda_k^{-1} = \infty.$$

Applying Theorem 7.141 with the set of exponents $\widetilde{\Lambda}$ to the function $x \mapsto jx^{j-1}$, there exists $\tilde{p}_j \in \mathcal{P}(\widetilde{\Lambda})$ such that

$$\Big(\int_0^1 |jx^{j-1} - \tilde{p}_j(x)|^2 \Big)^{1/2} < \frac{\varepsilon}{A}.$$

Note that for all polynomials $\tilde{p} \in \mathcal{P}(\widetilde{\Lambda})$ the integrated polynomials $p(x) = \int_0^x \tilde{p}(t)dt$ belong to $\mathcal{P}(\Lambda)$. Set

$$p_j(x) = \int_0^x \tilde{p}_j(t)dt.$$

Then for all $x \in [0,1]$, $j = 1, \dots, K$,

$$\begin{aligned} |x^j - p_j(x)| &= \Big| \int_0^x jt^{j-1}dt - \int_0^x \tilde{p}_j(t)dt \Big| = \Big| \int_0^x (jt^{j-1} - \tilde{p}_j(t))dt \Big| \\ &\leqslant \int_0^x |jt^{j-1} - \tilde{p}_j(t)|dt \leqslant \Big(\int_0^1 |jt^{j-1} - \tilde{p}_j(t)|^2 dt \Big)^{1/2} < \frac{\varepsilon}{A}. \end{aligned}$$

Here we have used the Cauchy-Schwarz inequality. Thus (7.51) is proved and the proof is complete. □

7.10.2. The mean-square error for monomials. The main result in this section is an explicit expression for the error $\operatorname{dist}_2(f_\mu, \mathcal{P}(\Lambda_N))$ where $f_\mu(x) = x^\mu$, $\Lambda_N = \{\lambda_0, \dots, \lambda_N\}$. $\mu \in \Lambda_N$ implies $\operatorname{dist}_2(f_\mu, \mathcal{P}(\Lambda_N)) = 0$. Otherwise we have:

7.142. THEOREM. *For $\mu \notin \Lambda_N$*

$$\operatorname{dist}_2(f_\mu, \mathcal{P}(\Lambda_N)) = \frac{1}{\sqrt{2\mu+1}} \prod_{k=0}^{N} \frac{|\mu - \lambda_k|}{\mu + \lambda_k + 1}$$

We shall use Theorem 7.29. The main algebraic tool is a determinant identity due to Cauchy which we will prove first.

7.143. LEMMA. *Let $n \geqslant 1$, and $a_1, \dots, a_n$, $b_1, \dots, b_n$ be positive numbers. Let*

$$D_n(a_1, \dots, a_n; b_1, \dots, b_n) = \det \begin{pmatrix} \frac{1}{a_1+b_1} & \cdots & \frac{1}{a_1+b_n} \\ \vdots & & \vdots \\ \frac{1}{a_n+b_1} & \cdots & \frac{1}{a_n+b_n} \end{pmatrix}$$

and let $F_1(a_1, b_1) = \frac{1}{a_1+b_1}$ and, for $n \geqslant 2$,

$$F_n(a_1, \dots, a_n; b_1, \dots, b_n) = \frac{\prod\limits_{\substack{i,k=1,\dots n \\ i>k}} (a_i - a_k)(b_i - b_k)}{\prod\limits_{i,k=1,\dots,n} (a_i + b_k)}.$$

Then $D_n(a_1, \dots, a_n; b_1, \dots, b_n) = F_n(a_1, \dots, a_n; b_1, \dots, b_n)$ for all $n \in \mathbb{N}$.

PROOF. The assertion is correct for $n = 1$. Both expressions are equal to $\frac{1}{a_1+b_1}$. Now assume $n \geqslant 2$. From the definition,

$$F_n(a_1, \dots, a_n; b_1, \dots, b_n) = \frac{1}{a_n + b_n} \Big(\prod_{k=1}^{n-1} \frac{(a_n - a_k)(b_n - b_k)}{(a_n + b_k)(b_n + a_k)} \Big) F_{n-1}(a_1, \dots a_{n-1}; b_1, \dots, b_{n-1}). \tag{7.52}$$

Now consider the determinants. Note that $\frac{1}{a_i+b_k} - \frac{1}{a_n+b_k} = \frac{a_n - a_i}{(a_i+b_k)(a_n+b_k)}$. By subtracting the last row from the previous ones we see that $D_n(a_1, \dots, a_n; b_1, \dots, b_n)$ is equal to

$$\det \begin{pmatrix} \frac{a_n-a_1}{(a_1+b_1)(a_n+b_1)} & \frac{a_n-a_1}{(a_1+b_2)(a_n+b_2)} & \cdots & \frac{a_n-a_1}{(a_1+b_{n-1})(a_n+b_{n-1})} & \frac{a_n-a_1}{(a_1+b_n)(a_n+b_n)} \\ \frac{a_n-a_2}{(a_2+b_1)(a_n+b_1)} & \frac{a_n-a_2}{(a_2+b_2)(a_n+b_2)} & \cdots & \frac{a_n-a_2}{(a_2+b_{n-1})(a_n+b_{n-1})} & \frac{a_n-a_2}{(a_2+b_n)(a_n+b_n)} \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{a_n-a_{n-1}}{(a_{n-1}+b_1)(a_n+b_1)} & \frac{a_n-a_{n-1}}{(a_{n-1}+b_2)(a_n+b_2)} & \cdots & \frac{a_n-a_{n-1}}{(a_{n-1}+b_{n-1})(a_n+b_{n-1})} & \frac{a_n-a_{n-1}}{(a_{n-1}+b_n)(a_n+b_n)} \\ \frac{1}{a_n+b_1} & \frac{1}{a_n+b_2} & \cdots & \frac{1}{a_n+b_{n-1}} & \frac{1}{a_n+b_n}, \end{pmatrix}$$

which is equal to

$$\frac{1}{a_n + b_n} \Big(\prod_{k=1}^{n-1} \frac{a_n - a_k}{a_n + b_k} \Big) \cdot \widetilde{D}_n(a_1, \dots a_{n-1}; b_1, \dots, b_n),$$

where

$$\widetilde{D}_n(a_1, \dots a_{n-1}; b_1, \dots, b_n) = \det \begin{pmatrix} \frac{1}{a_1+b_1} & \frac{1}{a_1+b_2} & \cdots & \frac{1}{a_1+b_{n-1}} & \frac{1}{a_1+b_n} \\ \frac{1}{a_2+b_1} & \frac{1}{a_2+b_2} & \cdots & \frac{1}{a_2+b_{n-1}} & \frac{1}{a_2+b_n} \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{1}{a_{n-1}+b_1} & \frac{1}{a_{n-1}+b_2} & \cdots & \frac{1}{a_{n-1}+b_{n-1}} & \frac{1}{a_{n-1}+b_n} \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix}.$$

We repeat the same argument with the columns. Observe $\frac{1}{a_i+b_j}-\frac{1}{a_i+b_n}=\frac{b_n-b_j}{(a_i+b_j)(a_i+b_n)}$ and subtracting the last column from the previous ones one sees that $\widetilde{D}(a_1,\dots,a_{n-1};b_1,\dots,b_n)$ is equal to

$$\det\begin{pmatrix} \frac{b_n-b_1}{(a_1+b_1)(a_1+b_n)} & \frac{b_n-b_2}{(a_1+b_2)(a_1+b_n)} & \cdots & \frac{b_n-b_{n-1}}{(a_1+b_{n-1})(a_1+b_n)} & \frac{1}{a_1+b_n} \\ \frac{b_n-b_1}{(a_2+b_1)(a_2+b_n)} & \frac{b_n-b_2}{(a_2+b_2)(a_2+b_n)} & \cdots & \frac{b_n-b_{n-1}}{(a_2+b_{n-1})(a_2+b_n)} & \frac{1}{a_2+b_n} \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{b_n-b_1}{(a_{n-1}+b_1)(a_{n-1}+b_n)} & \frac{b_n-b_2}{(a_{n-1}+b_2)(a_{n-1}+b_n)} & \cdots & \frac{b_n-b_{n-1}}{(a_{n-1}+b_{n-1})(a_{n-1}+b_n)} & \frac{1}{a_{n-1}+b_n} \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

which in turn is equal to

$$\Big(\prod_{k=1}^{n-1}\frac{b_n-b_k}{a_k+b_n}\Big)D_{n-1}(a_1,\dots,a_{n-1};b_1,\dots,b_{n-1})\,.$$

Combining the two formulas for D_n and $\widetilde{D}_n$ yields

$$D_n(a_1,\dots,a_n;b_1,\dots,b_n) = \frac{1}{a_n+b_n}\Big(\prod_{k=1}^{n-1}\frac{(a_n-a_k)(b_n-b_k)}{(a_n+b_k)(b_n+a_k)}\Big)D_{n-1}(a_1,\dots a_{n-1};b_1,\dots,b_{n-1}). \tag{7.53}$$

Comparing (7.52) and (7.53) we see that F_n and D_n satisfy the same recursive formula, and since $F_1=D_1$ we see that $F_n=D_n$ for all $n\in\mathbb{N}$. □

Proof of Theorem 7.142. We use the formula

$$\operatorname{dist}_2(f_\mu,\mathcal{P}(\Lambda_N))^2=\frac{\mathcal{G}(f_{\lambda_1},\dots,f_{\lambda_N},f_\mu)}{\mathcal{G}(f_{\lambda_1},\dots f_{\lambda_N})}$$

from Theorem 7.29. For the inner products in these Gram determinants observe that

$$\langle f_\lambda,f_\mu\rangle=\int_0^1 x^{\lambda+\mu}dx=\frac{1}{\lambda+\mu+1}.$$

By Lemma 7.143 with $a_j=b_j=\lambda_j+\frac{1}{2}$, we get

$$\mathcal{G}(f_{\lambda_1},\dots,f_{\lambda_N})=\frac{\prod_{i>k}(\lambda_i-\lambda_k)^2}{\prod_{i,k}(\lambda_i+\lambda_k+1)}$$

and applying it again with $a_j=b_j=\lambda_j+\frac{1}{2}$ for $j\leqslant N$ and $a_{N+1}=b_{N+1}=\mu+\frac{1}{2}$ we also get

$$\mathcal{G}(f_{\lambda_1},\dots,f_{\lambda_N},f_\mu)=\frac{\prod_{i>k}(\lambda_i-\lambda_k)^2}{\prod_{i,k}(\lambda_i+\lambda_k+1)}\cdot\frac{\prod_{k=0}^N(\mu-\lambda_k)^2}{(2\mu+1)\prod_{k=0}^N(\mu+\lambda_k+1)^2}.$$

Dividing the determinants yields

$$\operatorname{dist}_2(f_\mu,\mathcal{P}(\Lambda_N))^2=\frac{1}{2\mu+1}\prod_{k=0}^N\frac{(\mu-\lambda_k)^2}{(\mu+\lambda_k+1)^2}$$

and we obtain the claimed formula. □

Proof of Theorem 7.141. We begin with the observation that

$$\operatorname{dist}_2(f,\mathcal{P}(\Lambda))=\lim_{N\to\infty}\operatorname{dist}_2(f,\mathcal{P}(\Lambda_N)). \tag{7.54}$$

Indeed, $\operatorname{dist}_2(f,\mathcal{P}(\Lambda))\leqslant\operatorname{dist}_2(f,\mathcal{P}(\Lambda_{N+1}))\leqslant\operatorname{dist}_2(f,\mathcal{P}(\Lambda_N))$ so $\lim_{N\to\infty}\operatorname{dist}_2(f,\mathcal{P}(\Lambda_N))$ exists and is not smaller than $\operatorname{dist}_2(f,\mathcal{P}(\Lambda))$. If $h\in\mathcal{P}(\Lambda)$, then $h\in\mathcal{P}(\Lambda_M)$ for some M and therefore $\|f-h\|\geqslant$

$\operatorname{dist}_2(f, \mathcal{P}(\Lambda_M))$. Taking the infimum over $h \in \mathcal{P}(\Lambda_M)$ we get $\operatorname{dist}_2(f, \mathcal{P}(\Lambda)) \geqslant \liminf_{N\to\infty} \operatorname{dist}_2(f, \mathcal{P}(\Lambda_N))$. Altogether (7.54) follows.

We now prove the theorem and start with the error term $\operatorname{dist}_2(f_\mu, \mathcal{P}(\Lambda))$ for $f_\mu(x) = x^\mu$. If $\mu \in \Lambda$, then $\operatorname{dist}_2(f_\mu, \mathcal{P}(\Lambda)) = 0$, so let us assume $\mu \notin \Lambda$. By (7.54) and Theorem 7.142 we have then

$$\operatorname{dist}_2(f_\mu, \mathcal{P}(\Lambda)) = \lim_{N\to\infty} \frac{1}{\sqrt{2\mu+1}} \prod_{k=0}^{N} \frac{|\mu - \lambda_k|}{\mu + \lambda_k + 1} \tag{7.55}$$

Now let $\mu \notin \Lambda$ and let M be such that $\lambda_k > \mu$ for $k \geqslant M$. We write

$$\frac{\mu + \lambda_k + 1}{\lambda_k - \mu} = 1 + \frac{2\mu + 1}{\lambda_k - \mu}$$

and we see that $\sum_{j=1}^{\infty} \lambda_j^{-1}$ is convergent if and only if $\sum_{j=M}^{\infty} \frac{2\mu+1}{\lambda_j - \mu}$ is convergent. By Theorem 3.80, the series $\sum_{j=M}^{\infty} \frac{2\mu+1}{\lambda_j-\mu}$ is convergent if and only if the product $\prod_{j=M}^{\infty} \frac{\mu+\lambda_j+1}{\lambda_j-\mu}$ is convergent (and then to a real number greater than 1). The convergence of $\prod_{j=M}^{\infty} \frac{\mu+\lambda_j+1}{\lambda_j-\mu}$ to a positive number holds if and only if

$$\lim_{N\to\infty} \prod_{j=M}^{N} \frac{\lambda_j - \mu}{\mu + \lambda_j + 1} = \prod_{j=M}^{\infty} \frac{|\lambda_j - \mu|}{\mu + \lambda_j + 1}$$

exists and is positive. Finally, observe that for $\mu \notin \Lambda$ we have $\sum_{j=0}^{M-1} \frac{|\lambda_j-\mu|}{2\mu+1} > 0$ (if there are exponents $\lambda_j \leqslant \mu$). To summarize, $\sum_j \lambda_j^{-1} < \infty$ is equivalent to $\lim_{N\to\infty} \prod_{k=0}^{N} \frac{|\mu-\lambda_k|}{\mu+\lambda_k+1} > 0$ and by (7.55) this is equivalent to $\operatorname{dist}_2(f_\mu, \mathcal{P}(\Lambda)) > 0$. Vice versa, $\sum_j \lambda_j^{-1} = \infty$ if and only if

$$\lim_{N\to\infty} \prod_{k=0}^{N} \frac{|\mu - \lambda_k|}{\mu + \lambda_k + 1} = 0,$$

which by (7.55) is equivalent to $\operatorname{dist}_2(f_\mu, \mathcal{P}(\Lambda)) = 0$. This finishes the proof of part (ii) and also shows the necessity of the condition $\sum_j \lambda_j^{-1} = \infty$ in part (i).

For the sufficiency in part (i) let us assume $\sum_j \lambda_j^{-1} = \infty$. Let f be a continuous function on $[0, 1]$ and $\varepsilon > 0$. By the Weierstrass approximation theorem we find a polynomial $q = \sum_{k=0}^{K} a_k f_k$ (where $f_k(x) = x^k$) and $\|f - q\|_\infty < \varepsilon/2$. Then also

$$\|f - q\|_2 \leqslant \|f - q\|_\infty < \varepsilon/2.$$

We have already seen above that $\operatorname{dist}_2(f_k, \mathcal{P}(\Lambda)) = 0$ for $k = 0, \dots K$. Thus we may find a Müntz polynomial $p \in \mathcal{P}(\Lambda)$ so that $\|q - p\|_2 < \varepsilon/2$. Then

$$\|f - p\|_2 \leqslant \|f - q\|_2 + \|q - p\|_2 < \varepsilon$$

and since $\varepsilon > 0$ was arbitrary this finishes the proof of the theorem. □

7.144. Exercise. Let $0 \leqslant a \leqslant b$, let $\Lambda = (\lambda_j)$ be a strictly increasing sequence of nonnegative numbers, and let $\mathcal{P}(\Lambda)$ be the associated space of Müntz polynomials.
(i) Show that if $a = 0$, then $\mathcal{P}(\Lambda)$ is dense in $C([0, b])$ if and only if

$$\sum_{j=1}^{\infty} \lambda_j = \infty \text{ and } \lambda_0 = 0.$$

(ii) In the statement (i) can one replace the interval $[0, b]$ with the interval $[-b, b]$?
(iii) Show that if $0 < a < b$ and $\sum_j \lambda_j^{-1} = \infty$, then $\mathcal{P}(\Lambda)$ is dense in $C([a, b])$.

Remark. Clarkson and Erdős [**12**] showed that in part (iii) of Exercise 7.144 the converse is also true, that is if $\sum_j \lambda_j < \infty$ then $\mathcal{P}(\Lambda)$ is not dense in $C([a, b])$.

7.11. Further exercises

7.145. Exercise. Let f be smooth on $[0,1]$ (that is, arbitrarily often differentiable).

(i) Let $\varepsilon > 0$ and p a polynomial such that $|f'(x) - p(x)| \leqslant \varepsilon$ for all $x \in [0,1]$. Construct a polynomial q such that $|f(x) - q(x)| \leqslant \varepsilon$ for all $x \in [0,1]$.

(ii) Prove that there exists a sequence of polynomials $(p_n)_n$ such that $(p_n^{(k)})_n$ converges uniformly on $[0,1]$ to $f^{(k)}$ for all $k = 0, 1, 2, \ldots$.

7.146. Exercise. Let $f \in C([0,1])$ and $\mathcal{A} \subset C([0,1])$ be dense. Suppose that

$$\int_0^1 f(x)\overline{a(x)}dx = 0$$

for all $a \in \mathcal{A}$. Show that $f = 0$.

7.147. Exercise. Let $f \in C([-1,1])$ and $a \in [-1,1]$. Show that for every $\varepsilon > 0$ there exists a polynomial p such that $p(a) = f(a)$ and $|f(x) - p(x)| < \varepsilon$ for all $x \in [-1,1]$.

7.148. Exercise. Suppose $f \in C([1,\infty))$ and $\lim_{x\to+\infty} f(x) = a$. Show that f can be uniformly approximated on $[1,\infty)$ by functions of the form $g(x) = p(1/x)$, where p is a polynomial.

7.149. Exercise. For a function $f : [a,b] \to \mathbb{R}$ the modulus of continuity on $[a,b]$ is defined by

$$\omega_f(\delta) = \sup\{|f(x) - f(x')| : x, x' \in [a,b], |x - x'| \leqslant \delta\}.$$

Verify for this definition the properties stated in Lemma 7.63.

7.150. Exercise. Let $(B_n f)$ be the sequence of Bernstein polynomials associated with $f \in C([0,1])$. Show that there is a constant $C \in (0,\infty)$ so that for all $f \in C([0,1])$,

$$\sup_{x\in[0,1]} |B_n f(x) - f(x)| \leqslant C\omega_f(n^{-1/2}).$$

7.151. Exercise (Stone–Weierstrass for finite sets). Let K be a finite set and $\mathcal{A}$ a family of functions on K that is an algebra (i.e. closed under taking finite linear combinations and products), separates points and vanishes nowhere. Give a purely algebraic proof that $\mathcal{A}$ must then already contain every function on K. (That means your proof is not allowed to use the concept of an inequality. In particular, you are not allowed to use any facts about metric spaces such as the Stone–Weierstrass theorem.)

7.152. Exercise (Uniform approximation by neural networks). Let $\sigma(t) = e^t$ for $t \in \mathbb{R}$. Fix $n \in \mathbb{N}$ and let $K \subset \mathbb{R}^n$ be a compact set. As usual, let $C(K)$ denote the space of real-valued continuous functions on K. Define a class of functions $\mathcal{N} \subset C(K)$ by saying that $\mu \in \mathcal{N}$ iff there exist $m \in \mathbb{N}$, $W \in \mathbb{R}^{m\times n}$, $v, b \in \mathbb{R}^m$ such that

$$\mu(x) = \sum_{i=1}^{m} \sigma((Wx)_i + b_i)v_i \text{ for all } x \in K.$$

Prove that $\mathcal{N}$ is dense in $C(K)$.

Remark. This is a special case of a result of Cybenko [**14**]. As a real-world motivation for this problem, note that a function $\mu \in \mathcal{N}$ can be interpreted as a neural network with a single hidden layer, see Figure 7.6. Consequently, in this problem you are asked to show that every continuous function can be uniformly approximated by neural networks of this form.

7.153. Exercise. Let $f \in C([a,b])$ and let $x_0, \ldots, x_m$ be distinct points in $[a,b]$. Show that there is a sequence of polynomials P_N, with $P_N(x_k) = f(x_k)$ for $k = 0, \ldots, m$ so that P_N converges uniformly on $[a,b]$ to f, as $N \to \infty$.

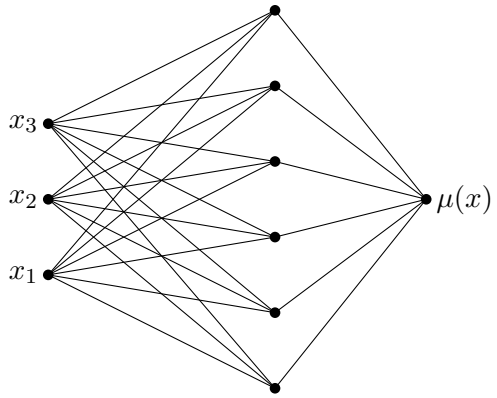


FIGURE 7.6. Visualization of μ when $n = 3$ and $m = 6$.

7.154. EXERCISE. *Sine polynomials* are functions of the form $x \mapsto \sum_{k=1}^{N} \sin(2\pi kx)$ (with $N \in \mathbb{N}$). Prove that a continuous function f on $[0, \frac{1}{2}]$ can be uniformly approximated on $[0, \frac{1}{2}]$ by sine polynomials if and only if $f(0) = 0$.

7.155. EXERCISE. Show that there exists no continuous 1-periodic function g such that $f * g = f$ holds for all continuous 1-periodic functions f.

7.156. EXERCISE. Find a sequence of continuous functions $(f_n)_n$ on $[0, 1]$ and a continuous function f on $[0, 1]$ such that $\|f_n - f\|_2 \to 0$, but $f_n(x)$ does not converge to $f(x)$ for any $x \in [0, 1]$.

7.157. EXERCISE. Let $(X, \langle \cdot, \cdot \rangle)$ be a pre-Hilbert space and $\widetilde{X}$ its completion as a Hilbert space (*cf.* Exercise 5.119).

Show that if $\{\phi_\alpha\}_{\alpha \in \mathcal{I}}$ is a complete orthonormal system in X and $\imath : X \mapsto \widetilde{X}$ the canonical isometry, then $\{\imath(\phi_\alpha) : \alpha \in \mathcal{I}\}$ is a Hilbert basis of $\widetilde{X}$.

7.158. EXERCISE. Let f be the 1-periodic function such that $f(x) = |x|$ for $x \in [-1/2, 1/2]$. Determine explicitly a sequence of trigonometric polynomials $(p_N)_N$ such that $p_N \to f$ uniformly as $N \to \infty$.

7.159. EXERCISE. Let f, g be continuous, 1-periodic functions.

(i) Show that $\widehat{f * g}(n) = \widehat{f}(n)\widehat{g}(n)$. (Recall Definition 7.53.)

(iii) If f is continuously differentiable, prove that $\widehat{(f')}(n) = 2\pi i n \widehat{f}(n)$. (Recall Definition 7.50.)

(iv) Let $y \in \mathbb{R}$ and set $f_y(x) = f(x + y)$. Show that $\widehat{f}_y(n) = e^{2\pi i n y}\widehat{f}(n)$.

(v) Let $m \in \mathbb{Z}$, $m \neq 0$ and set $f_m(x) = f(mx)$. Show that $\widehat{f_m}(n) = \widehat{f}(\frac{n}{m})$ if m divides n and $\widehat{f_m}(n) = 0$ otherwise.

7.160. EXERCISE. Let f be 1-periodic and k times continuously differentiable. Prove that there exists a constant $c > 0$ such that

$$|\widehat{f}(n)| \leqslant c|n|^{-k} \quad \text{for all } n \in \mathbb{Z}.$$

7.161. EXERCISE. Prove that

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

7.162. EXERCISE. Establish orthogonality properties of the system consisting of 1, $\cos(2\pi kx)$, $k \in \mathbb{N}$, $\sin(2\pi lx)$, $l \in \mathbb{N}$. If $p \in \mathcal{T}_N$ is written in the form $p(x) = a_0 + \sum_{n=1}^{N} a_n \cos(2\pi nx) + \sum_{n=1}^{N} b_n \sin(2\pi nx)$ prove that for $1 \leqslant k \leqslant N$,

$$a_0 = \int_{-\frac{1}{2}}^{\frac{1}{2}} p(x)\,dx, \quad a_k = 2\int_{-\frac{1}{2}}^{\frac{1}{2}} p(x)\cos(2\pi kx)\,dx, \; b_k = 2\int_{-\frac{1}{2}}^{\frac{1}{2}} p(x)\sin(2\pi kx)\,dx.$$

7.163. Exercise. Let f be 1-periodic and continuous.

(i) Suppose that $\widehat{f}(n) = -\widehat{f}(-n) \geqslant 0$ holds for all $n \geqslant 0$. Prove that

$$\sum_{n=1}^{\infty} \frac{\widehat{f}(n)}{n} < \infty.$$

(ii) Show that there does not exist a 1-periodic continuous function f such that

$$\widehat{f}(n) = \frac{\operatorname{sgn}(n)}{\log|n|} \quad \text{for all } |n| \geqslant 2.$$

Here $\operatorname{sgn}(n) = 1$ if $n > 0$ and $\operatorname{sgn}(n) = -1$ if $n < 0$.

7.164. Exercise. Let $p \in \mathcal{T}_N$. Show that

$$\int_0^1 p(x)dx = \frac{1}{N+1} \sum_{k=0}^{N} p\big(\frac{k}{N+1}\big)$$

7.165. Exercise. Let $f : \mathbb{R} \to \mathbb{C}$ be a 1-periodic continuous function. Let $a \in \mathbb{R}\backslash\mathbb{Q}$.

(i) Show that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} f(x+ka) = \int_0^1 f(t)\,dt,$$

uniformly on $\mathbb{R}$.

(ii) Prove that (i) holds with uniform convergence on $\mathbb{R}$.

(iii) What can one say about the limit if $a \in \mathbb{Q}$?

7.166. Exercise. Let f be continuous on $[0,1]$.

(i) Show that $\lim_{n\to\infty} \int_0^1 f(x)\sin(2\pi nx)\,dx = 0$.

(ii) Does the limit

$$\lim_{n\to\infty} \int_0^1 f(x)|\sin(2\pi nx)|\,dx$$

exist? If yes, determine its value.

7.167. Exercise. Show that if $f \in C(\mathbb{T})$ and if $\operatorname{dist}_\infty(f, \mathcal{T}_N) = O(1/N)$, then the modulus of continuity satisfies for small δ

$$\omega_f(\delta) \leqslant C\delta \log(\tfrac{1}{\delta}),$$

where $C \in (0,\infty)$ is a constant independent of f and δ.

7.168. Exercise. The *de la Vallée-Poussin means* are defined for a 1-periodic function f by

$$V_N f(x) = 2\sigma_{2N+1} f(x) - \sigma_N f(x).$$

The de la Vallée-Poussin means combine the reproducing property of the partial sum operators S_N and the convergence properties of the Fejér means $\sigma_N f$ (see (7.19)).

(i) Show that $V_N f \in \mathcal{T}_{2N}$. Write $V_N = k_N * f$ and compute the kernel k_N and its Fourier coefficients.

(ii) Show $V_N p = p$ for all $p \in \mathcal{T}_N$.

(iii) Prove $\|f - V_N f\|_\infty \leqslant 4\operatorname{dist}_\infty(f, \mathcal{T}_N)$.

(iv) Prove $\|f - V_N f\|_\infty \leqslant C\omega_f(\frac{1}{N+1})$ for a constant $C \in (0,\infty)$ not depending on f and N.

7.169. Exercise (Rogosinski singular integrals). Let $f \in C(\mathbb{T})$. For $0 \leqslant \theta \leqslant 1$ define the *Rogosinski singular integrals* by

$$R_{N,\theta} f(x) = \frac{1}{2}\Big(S_N f(x + \tfrac{\theta}{2N+1}) + S_N f(x - \tfrac{\theta}{2N+1})\Big).$$

The *Rogosinski regularization* of the partial sum operator S_N is given by the case $\theta = 1/2$. Show that

(i) $R_{N,\theta} f = \mathcal{R}_{N,\theta} * f$ where for $x \neq \pm\frac{\theta}{2N+1}$

$$\mathcal{R}_{N,\theta}(x) = \cos(\pi\theta)\, \mathcal{R}^{\mathrm{sing}}_{N,\theta}(x) + \sin(\pi\theta)\, \mathcal{R}^{\mathrm{reg}}_{N,\theta}(x)$$

with

$$\mathcal{R}^{\mathrm{sing}}_{N,\theta}(x) = \tfrac{1}{2}\sin((2N+1)\pi x)\Big(\frac{1}{\sin(\pi(x+\frac{\theta}{2N+1}))} + \frac{1}{\sin(\pi(x-\frac{\theta}{2N+1}))}\Big),$$
$$\mathcal{R}^{\mathrm{reg}}_{N,\theta}(x) = \tfrac{1}{2}\cos((2N+1)\pi x)\Big(\frac{1}{\sin(\pi(x+\frac{\theta}{2N+1}))} - \frac{1}{\sin(\pi(x-\frac{\theta}{2N+1}))}\Big).$$

Determine the value of $\mathcal{R}_{N,\theta}(\pm\frac{\theta}{2N+1})$.

(ii) Prove that for $|x| \leqslant 1/2$, $N \geqslant 1$,

$$|\mathcal{R}_{N,\theta}(x)| \leqslant C\frac{N}{1+N|x|}$$

moreover, for $\theta = 1/2$,

$$|\mathcal{R}_{N,1/2}(x)| \leqslant C\frac{N}{1+N^2x^2},$$

where $C \in (0,\infty)$ is a constant not depending on x, N, θ.

(iii) For the operator norms on $C(\mathbb{T})$, show

$$\|\mathcal{R}_{N,\theta}\|_{\mathrm{op}} \geqslant c(\theta)\log N$$

with constants $c(\theta) > 0$ for $\theta \in [0,1]\backslash\frac{1}{2}$. Moreover, for all $\theta \in [0,1]$,

$$\|R_{N,\theta}\|_{\mathrm{op}} \leqslant C(1 + |\cos\pi\theta|\log N).$$

(iv) Show that there exists a constant $C \in (0,\infty)$ such that

$$\|R_{N,\theta} f - f\|_\infty \leqslant C(1 + |\cos\pi\theta|\log N)\,\omega_f(\tfrac{1}{N+1}).$$

(v) Let (θ_N) be a sequence of numbers in $[0,1]$. Prove that the following are equivalent:

(a) For all $f \in C(\mathbb{T})$, $R_{N,\theta_N} f$ converges to f uniformly as $N \to \infty$.

(b) $\sup_{N\geqslant 1} |\theta_N - \frac{1}{2}|\log N < \infty$.

7.170. Exercise (Higher order modulus of continuity). For $f \in C(\mathbb{T})$, $r \in \mathbb{N}$, define the difference operator Δ_h by

$$\Delta_h f(x) = f(x+h) - f(x),$$

let $\Delta_h^1 = \Delta_h$ and for $r \geqslant 2$, $\Delta_h^r f = \Delta_h[\Delta_h^{r-1} f]$. Define the *modulus of continuity of order* $r \geqslant 1$ by

$$\omega_{r,f}(\delta) = \sup_{|h|\leqslant\delta} \|\Delta_h^r f\|_\infty.$$

Prove:

(i) $\Delta_h^r f(x) = \sum_{k=0}^r \binom{r}{k}(-1)^{r-k} f(x+kh)$.

(ii) $\Delta_{nh}^r f(x) = \sum_{k_1=0}^{n-1}\cdots\sum_{k_r=0}^{n-1} \Delta_h^r f(x+k_1h+\cdots+k_rh)$.

(iii) For $\lambda > 1$, $\omega_{r,f}(\lambda\delta) \leqslant (\lambda+1)^r \omega_{r,f}(\delta)$.

(iv) If $f \in C^r(\mathbb{T})$ (i.e. $f, \ldots, f^{(r)}$ exist and are continuous), then $\omega_{r,f}(\delta) = O(\delta^r)$.

7.171. Exercise. For continuous functions on the real line define the difference operators $\Delta_h f(x) = f(x+h) - f(x)$. Let $\alpha > 1$. Determine the class of functions f such that for all $x \in \mathbb{R}$ we have $\Delta_h f(x) = o(|h|)$ as $|h| \to 0$.

7.172. Exercise. Consider the Jackson approximate unit $(\mathcal{K}_N)$ defined in Lemma 7.67. Let $\omega_{2,f}(\delta)$ be the second order modulus of continuity, i.e.

$$\omega_{2,f}(\delta) = \sup_{|h|\leqslant\delta} \|f(\cdot - h) + f(\cdot + h) - 2f(\cdot)\|_\infty. \tag{7.56}$$

(i) Prove

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \mathcal{K}_N(y) f(x-y)dy - f(x) = \int_0^{\frac{1}{2}} \mathcal{K}_N(y)\Delta_y^2 f(x-y)dy.$$

(ii) Prove that $\operatorname{dist}_\infty(f, \mathcal{T}_N) \leqslant C\omega_{2,f}(\frac{1}{N+1})$ for a constant $C \in (0,\infty)$.

(iii) Prove that the following statements are equivalent for any given $\alpha \in (0,2)$:

(1) $\operatorname{dist}_\infty(f, \mathcal{T}_N) = O(N^{-\alpha})$ as $N \to \infty$.
(2) $\|\mathcal{K}_N * f - f\|_\infty = O(N^{-\alpha})$ as $N \to \infty$.
(3) $\omega_{2,f}(\delta) = O(\delta^\alpha)$ as $\delta \to 0+$.

7.173. Exercise. For $\ell \in \mathbb{N}$ let

$$\mathcal{K}_{\ell,N} = \frac{1}{\beta_{\ell,N}}\Big(\frac{\sin \pi(N+1)x}{\sin \pi x}\Big)^{2\ell},$$

where the constant $\beta_{\ell,N}$ is defined such that $\int_{-\frac{1}{2}}^{\frac{1}{2}} \mathcal{K}_{\ell,N}(y)dy = 1$.

(i) Prove that for each ℓ there exist $c_1, c_2 > 0$ such that $c_1 N^{2\ell-1} \leqslant \beta_{\ell,N} \leqslant c_2 N^{2\ell-1}$.

(ii) Prove that $\int_{-\frac{1}{2}}^{\frac{1}{2}} y^k \mathcal{K}_{\ell,N}(y)dy \leqslant C_\ell N^{-k}$, $k = 0, 1, \dots, 2\ell - 2$.

(iii) Prove that

$$\int_{-1/2}^{1/2} \big(f(x) + (-1)^{\ell+1}\Delta_y^\ell f(x)\big)\mathcal{K}_{\ell,N}(y)dy$$

belongs to $\mathcal{T}_{\ell N}$.

(iv) Prove that $\operatorname{dist}_\infty(f, \mathcal{T}_N) \leqslant C_\ell \omega_{\ell,f}(\frac{1}{N+1})$ for all $N \geqslant 0$.

(v) Let $\ell \in \mathbb{N}$ and $0 < \alpha < \ell$. Then the condition $\sup_{N\in\mathbb{N}_0} N^\alpha \operatorname{dist}_\infty(f, \mathcal{T}_N) < \infty$ holds if and only if $\sup_{\delta>0} \delta^{-\alpha}\omega_{\ell,f}(\delta) < \infty$.

7.174. Exercise (Zygmund class). Let $f \in C(\mathbb{T})$ and let $\omega_{2,f}$ be as in (7.56).

(i) Show that for $\alpha \in (0,1)$ the condition $\omega_f(\delta) = O(\delta^\alpha)$ is equivalent to $\omega_{2,f}(\delta) = O(\delta^\alpha)$.

(ii) Let $\Lambda_1(\mathbb{T})$ be the space of $C(\mathbb{T})$ functions f satisfying $\omega_{2,f}(\delta) = O(\delta)$; it is called the *Zygmund class*. Prove that $\Lambda_1(\mathbb{T})$ is a Banach space, with the norm

$$\|f\|_{\Lambda_1} = \|f\|_\infty + \sup_{0<\delta\leqslant 1} \delta^{-1}\omega_{2,f}(\delta).$$

7.175. Exercise (Zygmund vs. Lipschitz). There are functions $f \in C(\mathbb{T})$ belonging to the Zygmund class $\Lambda_1(\mathbb{T})$ from Exercise 7.174 that are not Lipschitz continuous, i.e. do not satisfy $\omega_f(\delta) = O(\delta)$. Consider

$$f(x) = \sum_{k=1}^{\infty} \frac{\sin(2\pi k x)}{k^2}.$$

(i) Show that f is differentiable in $(0,1)$.

(ii) Show that $\lim_{x\to 0+} f'(x) = \infty$ and conclude that $\omega_f(\delta)$ is not $O(\delta)$ as $\delta \to 0$.

(iii) Show that f belongs to the Zygmund class.

7.176. Exercise.

(i) Let f be 1-periodic and $f, f' \in C(\mathbb{T})$. Prove that there exists a constant $C \in (0,\infty)$ independent of f such that for $N \geqslant 1$,

$$\operatorname{dist}_\infty(f, \mathcal{T}_N) \leqslant CN^{-1}\operatorname{dist}_\infty(f', \mathcal{T}_N).$$

(ii) Let f be 1-periodic and assume that $f, f', \dots, f^{(k)}$ are continuous. Then for $N > k$,

$$\operatorname{dist}(f, \mathcal{T}_N) \leqslant C_k N^{-k} \omega_{f^{(k)}}(N^{-1}),$$

where $C_k \in (0, \infty)$ is a constant only depending on k.

7.177. Exercise. Let $f \in C([a,b])$ and let $\alpha \in (0,1)$. Suppose that for all $N \geqslant 1$,

$$\inf_{p \in \mathcal{P}_N} \|f - p\|_\infty = O(N^{-\alpha}).$$

Prove that the modulus of continuity satisfies

$$\omega_f(\delta) = O(\delta^{\alpha/2}).$$

7.178. Exercise. Consider the Golay–Rudin–Shapiro sequence of coefficients α_n in the Rudin-Shapiro polynomials $P_\ell(z) = \sum_{n=0}^{2^\ell - 1} \alpha_n z^n$, see Definition 7.88 and Lemma 7.89. Prove that

$$\alpha_n = (-1)^{\lambda_n},$$

where λ_n is defined as follows. Let $n = \sum_{k \geqslant 0} \nu_k(n) 2^k$ be the unique binary expansion of n, with $\nu_k(n) \in \{0,1\}$. Then $\lambda_n = \sum_{k \geqslant 0} \nu_k(n)\nu_{k+1}(n)$ which is the number of occurrences of two consecutive digits 1 in the binary expansion. For example, for $47 = 1 + 2 + 4 + 8 + 32$ (i.e. 101111 in the binary notation) we have $\lambda_{47} = 3$ and $\alpha_{47} = -1$.

7.179. Exercise (Barycentric interpolation formula). Show: Given distinct nodes $\{t_j\}_{j=0}^N$ the associated elementary Lagrange polynomials satisfy the formula

$$\ell_j(t) = \frac{\frac{c_j}{t - t_j}}{\sum_{k=0}^N \frac{c_k}{t - t_k}} \quad \text{with } c_j = \frac{1}{\prod_{k:k \neq j}(t_j - t_k)}$$

The unique polynomial satisfying $p(t_k) = y_k$, $k = 0, \dots, N$ is given by

$$p(t) = \frac{\sum_{j=1}^N \frac{c_j y_j}{t - t_j}}{\sum_{j=1}^N \frac{c_j}{t - t_j}}.$$

7.180. Exercise (Newton interpolation formula). Show that given nodes $t_0, \dots, t_n$ the unique interpolation polynomial $p \in \mathcal{P}_n$ with $p(t_j) = y_j$ can be written as

$$p(t) = \beta_0 + \beta_1(t - t_0) + \beta_2(t - t_0)(t - t_1) + \cdots + \beta_n(t - t_0)\cdots(t - t_{n-1}),$$

where the constants β_k can be computed by evaluating successively at $t_0, \dots t_n$. Define β_k recursively.

Remark. The formula is practical in a situation where one node x_{n+1} is added, and only one coefficient β_{n+1} needs to be computed. It is not practical if one works with given sets of nodes but many data sets of values.

7.181. Exercise (Error estimates for Hermite interpolation). Theorem 7.112 has an analogue for Hermite interpolation: Let $f : [a,b] \to \mathbb{R}$ be in $C^{2n+2}([a,b])$ and let the $n+1$ nodes $(t_\nu)_{\nu=0}^n$ belong to $[a,b]$. Let P be the Hermite interpolation polynomial for the data $(t_\nu, f(t_\nu), f'(t_\nu))_{\nu=0}^n$. Prove that for each $t \in [a,b]$ there exists a $\tau \in [a,b]$ such that

$$f(t) - P(t) = \frac{f^{(2n+2)}(\tau)}{(2n+2)!} \prod_{\nu=0}^n (t - t_\nu)^2.$$

7.182. Exercise. In numerical integration one seeks to approximate the value of $\int_a^b f(t)w(t)dt$ using *quadrature formulas* i.e. expressions of the form $I(f) = \sum_{k=0}^N \beta_k f(t_k)$ such that the formula is exact for polynomials in $\mathcal{P}_N$. This is closely related to Lagrange interpolation.

Prove: Let $t_0 < \cdots < t_N$ be $N+1$ nodes in $[a,b]$. There exist unique numbers $\beta_0, \ldots, \beta_N$ such that for all polynomials $P \in \mathcal{P}_N$

$$\int_a^b P(t)w(t)dt = \sum_{k=0}^{N} \beta_k P(t_k).$$

The constants β_k are given by

$$\beta_k = \int_a^b \ell_k(t)w(t)dt = \int_a^b \frac{\Omega(t)}{(t-t_k)\Omega'(t_k)} w(t)dt,$$

with the Lagrange polynomials ℓ_k as in (7.31), and $\Omega(t) = \prod_{k=0}^{N}(t-t_k)$. The coefficients satisfy $\sum_{k=0}^{N} \beta_k = \int_a^b w(t)dt$.

7.183. Exercise. Gauss noticed that the result in Exercise 7.182 may be improved in the sense that the formula also works for some polynomials of degree $2N+1$ if we take as nodes the zeroes of a polynomial of degree $N+1$ which is orthogonal to all $p \in \mathcal{P}_N$. Here orthogonality is defined with respect to the $\langle f, g\rangle_w = \int_a^b f(t)\overline{g(t)}w(t)dt$ on $C([a,b])$.

Prove: Let $Q \in \mathcal{P}_{N+1}$, not identically zero, and let Q satisfy $\langle Q, p\rangle_w = 0$ for all $p \in \mathcal{P}_N$. Let $\{t_k\}_{k=0}^{N}$ be the zeroes of Q (which by Theorem 7.130 are simple and lie in (a,b)). Then the *Gaussian quadrature formula*

$$\int_a^b P(t)w(t)dt = \sum_{k=0}^{N} \beta_k P(t_k)$$

holds for all $P \in \mathcal{P}_{2N+1}$. Here

$$\beta_k = \int_a^b \frac{\Omega(t)}{(t-t_k)\Omega'(t_k)} w(t)dt = \frac{\int_a^b \prod_{j\neq k}(t-t_j)^2 w(t)dt}{\prod_{j\neq k}(t_k-t_j)^2} > 0, \quad k = 0, \ldots, N.$$

7.184. Exercise. Consider a sequence p_n of polynomials of degree n, which are mutually orthogonal with respect to the inner product $\langle f, g\rangle_w$. Let $(t_{k,N})_{k=0}^{N}$ be the (simple) zeros of p_{N+1} in (a,b). For $f \in C([a,b])$ let

$$\mathcal{I}_N[f] = \sum_{k=0}^{N} \beta_{k,N} f(t_{k,N}) \text{ with } \beta_{k,N} = \frac{\int_a^b \prod_{j\neq k}(t-t_{j,N})^2 w(t)dt}{\prod_{j\neq k}(t_{k,N}-t_{j,N})^2}.$$

Then $\lim_{N\to\infty} \mathcal{I}_N[f] = \int_a^b f(t)w(t)dt$.

7.185. Exercise.

(i) Given an interval $[a,b]$ find the unique polynomial p with leading coefficient 1 that minimizes the supremum norm of p.

(ii) Find versions of the polynomial inequalities by Bernstein, Schur and Markov that apply to the interval $[a,b]$.

7.186. Exercise. For the Chebyshev polynomials (T_n), prove that

$$T_n(t) = \sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k \binom{n}{2k} t^{\,n-2k} (1-t^2)^k.$$

7.187. Exercise (Chebyshev differential equation). Show that the Chebyshev polynomials (T_n) satisfy

$$(1-t^2)T_n''(t) - tT_n'(t) + n^2 T_n(t) = 0.$$

7.188. Exercise. Show that the Chebyshev polynomials (T_n) satisfy

$$T_n^{(k)}(1) = \frac{n^2(n-1)^2 \cdots (n-k+1)^2}{1 \cdot 3 \cdots (2k-1)}.$$

7.189. Exercise (Chebyshev generating function).

(i) For $|r| < 1$ and $\vartheta \in \mathbb{R}$ prove that

$$\sum_{n=0}^{\infty} r^n \cos(n\vartheta) = \frac{1 - r\cos\vartheta}{1 - 2r\cos\vartheta + r^2}.$$

(ii) Use (i) to show that

$$\sum_{n=0}^{\infty} r^n T_n(t) = \frac{1 - rt}{1 - 2rt + r^2}, \quad t \in [-1, 1].$$

We say that $r \mapsto \frac{1-rt}{1-2rt+r^2}$ is the *generating function* for the Chebyshev polynomials.

7.190. Exercise. For $n \in \mathbb{N}_0$ the polynomials

$$U_n(t) = \frac{1}{n+1} T'_{n+1}(t) = \frac{\sin(n+1)\vartheta}{\sin\vartheta}, \quad t = \cos\vartheta,$$

are called *Chebyshev polynomials of the second kind.* Verify the recursion $U_0(t) = 1$, $U_1(t) = 2t$ and

$$U_n(t) = 2tU_{n-1}(t) - U_{n-2}(t), \quad n \geqslant 2$$

and the orthogonality relations

$$\int_{-1}^{1} U_n(t)U_m(t)\sqrt{1-t^2}\,dt = \begin{cases} 0 & \text{if } n \neq m \\ \pi/2 & \text{if } n = m \end{cases}$$

7.191. Exercise. Let $N \geqslant 1$ and let $t_{k,N} = \cos(\frac{2k+1}{2N+2}\pi)$, $k = 0, \ldots, N$, i.e. the zeros of the Chebyshev polynomial T_{N+1}. Prove the discrete orthogonality relations

$$\frac{1}{N+1} \sum_{k=0}^{N} T_m(t_{k,N})T_n(t_{k,N}) = \begin{cases} 0, & \text{if } m \neq n, \\ 1/2, & \text{if } m = n \neq 0, \\ 1, & \text{if } m = n = 0. \end{cases}$$

7.192. Exercise (Legendre polynomials). Define

$$p_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n}[(1-t^2)^n]$$

for $n = 0, 1, \ldots$. Prove that

(i) p_n is a polynomial of exact degree n. Determine the leading coefficient.

(ii) $\int_{-1}^{1} p_n(t)p_m(t)dt = 0$ if $n \neq m$.

(iii) $\int_{-1}^{1} [p_n(t)]^2 dt = \frac{2}{2n+1}$.

7.193. Exercise. Let X_N be a sequence of finite-dimensional subspaces of $C([a,b])$, such that $X_N \subset X_{N+1}$ for all $N \in \mathbb{N}$. Let $P_N : C([a,b]) \to X_N$ be a projection. Let $f \in C([a,b])$ and assume that $\lim_{N\to\infty} \operatorname{dist}_\infty(f, X_N)\|P_N\|_{\mathrm{op}} = 0$. Prove that $\lim_{N\to\infty} \|P_N f - f\|_\infty = 0$.

7.194. Exercise. Let $\mathcal{L}_N f$ be the trigonometric interpolation polynomial associated with the equidistant nodes $x_j = \frac{j}{2N+1}$, and values $y_j = f(x_j)$, $j = 0, \ldots, 2N$. For $f \in C(\mathbb{T})$ prove

$$\lim_{N\to\infty} \Big(\int_0^1 |\mathcal{L}_N f - f|^2 dx \Big)^{1/2} = 0$$

(i) Show that

$$\|\mathcal{L}_N\|_{\mathrm{op}} \approx \log N$$

for large N, in the sense that $c_1 \log N \leqslant \|\mathcal{L}_N\|_{\mathrm{op}} \leqslant c_2 \log N$ for $N > 1$ with positive constants c_1, c_2.

(ii) Show that there is $C \geqslant 1$ such that for all $N \geqslant 2$

$$\|\mathcal{L}_N f - f\|_\infty \leqslant C \log N\, \omega_f(\tfrac{1}{N+1}).$$

Note: More can be said about the asymptotics, namely $\|\mathcal{L}_N\|_{\mathrm{op}} \sim \frac{2}{\pi}\log N$ as $N \to \infty$ (see [**17**]).

7.195. Exercise. Let $\Lambda \subset \mathbb{Z}$ be finite and let $\mathcal{T}(\Lambda)$ be the space of trigonometric polynomials with frequencies in Λ, i.e. the trigonometric polynomials of the form $\sum_{k\in\Lambda} c_k e^{2\pi ikx}$. Let S_Λ be the associated orthogonal projection operator,

$$S_\Lambda f(x) = \sum_{k\in\Lambda} \widehat{f}(k) e^{2\pi ikx}.$$

(i) Let P be any bounded projection operator from $C(\mathbb{T})$ to $\mathcal{T}(\Lambda)$. Prove

$$\|P\|_{\mathrm{op}} \geqslant \|S_\Lambda\|_{\mathrm{op}}$$

for the $C(\mathbb{T}) \to C(\mathbb{T})$ operator norm.

(ii) Prove that

$$\|S_\Lambda\|_{\mathrm{op}} = \int_0^1 \Big| \sum_{k\in\Lambda} e^{2\pi ikx} \Big| \, dx.$$

Remarks. (i) A conjecture attributed to Littlewood [**28**] stated that for arbitrary Λ one has the lower bound $\|S_\Lambda\|_{\mathrm{op}} \geqslant c\log(\#(\Lambda))$, matching the lower bound for the partial sum operators for the Fourier series. Cohen [**13**] showed that always $\|S_\Lambda\|_{\mathrm{op}} \to \infty$ as $\#\Lambda \to \infty$, with additional quantitative lower bounds. Littlewood's conjecture was later confirmed independently by Konyagin [**40**] and by McGee, Pigno and Smith [**48**] (see the exposition of this very subtle result in DeVore–Lorentz [**15**, Ch. 9.6]).

(ii) A stronger conjecture, also by Littlewood, is that for given cardinality $\#\Lambda = N$ the operator norms $\|S_\Lambda\|_{\mathrm{op}}$ are minimal when Λ is an arithmetic progression. This is still open.

(iii) The actual size of $\|S_\Lambda\|_{\mathrm{op}}$ can be much larger than $\log N$. For example, let

$$\Lambda_N^{\mathrm{lac}} = \{1, 2, 4, \ldots, 2^{N-1}\}.$$

Then $\|S_{\Lambda_N^{\mathrm{lac}}}\|_{\mathrm{op}} \geqslant cN^{1/2}$, for some $c > 0$. This is essentially largest possible, since by the Cauchy-Schwarz inequality we always have $\|S_\Lambda\|_{\mathrm{op}} \leqslant (\#\Lambda)^{1/2}$. The same lower bound holds for more general $\{\kappa_j\}_{j=1}^N$ such that κ satisfies the lacunarity condition $\kappa_{j+1}/\kappa_j \geqslant C > 1$ (see DeVore–Lorentz [**15**, Ch. 9.5]). For many other choices of Λ the behavior of $\|S_\Lambda\|_{\mathrm{op}}$ appears to be a mystery.

7.196. Exercise (Dyadic intervals). For non-negative integers j, k with $0 \leqslant j < 2^k$ we define

$$I_{k,j} = [2^{-k}j, 2^{-k}(j+1)) \subset [0,1).$$

The interval $I_{k,j}$ is called a *dyadic interval* and k is called its *generation.* We denote by $\mathcal{D}_k$ the set of all dyadic intervals of generation k and by $\mathcal{D} = \bigcup_{k\geqslant 0} \mathcal{D}_k$ the set of all dyadic intervals on $[0,1)$. Each dyadic interval $I \in \mathcal{D}$ with $|I| = 2^{-k}$ can be split in the middle into its *left child* and *right child,* which are again dyadic intervals that we denote by I_ℓ and I_r, respectively.

(i) Prove that two dyadic intervals are either disjoint or contained in each other. That is, for every $I, J \in \mathcal{D}$ at least one of the following is true: $I \cap J = \varnothing$ or $I \subset J$ or $J \subset I$. Moreover, for every $k \geqslant 0$ the dyadic intervals of generation k are a partition of $[0,1)$.

(ii) Let $J \subset [0,1]$ be any interval. Show that there exists $I \in \mathcal{D}$ such that $|I| \leqslant |J|$ and $3I \supset J$. (Here $3I$ denotes the interval with three times the length of I and the same center as I.)

7.197. Exercise (Haar system). For each dyadic interval $I \in \mathcal{D}$ (see Exercise 7.196) we define the Haar wavelet associated with I by

$$\psi_I = |I|^{-1/2}(\mathbb{1}_{I_\ell} - \mathbb{1}_{I_r})$$

If I has length 2^{-N} we refer to its reciprocal 2^N as the *Haar frequency* of ψ_I. The countable set of functions given by

$$\mathcal{H} = \{\mathbb{1}_{[0,1)}\} \cup \{\psi_I \,:\, I \in \mathcal{D}\}$$

is called the *Haar system* on $[0,1)$.

Take the standard inner product $\langle f, g\rangle = \int f(x)\overline{g(x)}\,dx$ on the space of Riemann integrable functions on $[0,1]$. Prove that the Haar system $\mathcal{H}$ on $[0,1)$ is an orthonormal system .

7.198. Exercise. For a function f on $[0,1]$ and an interval $I \subset [0,1]$ of length $|I|$ we write $\mathrm{av}_I(f) = \frac{1}{|I|}\int_I f$ to denote the *average* or the *mean* of f over I. For $n = 0,1,2,\dots$ define the linear operator $\mathbb{E}_n$ by $\mathbb{E}_n f(x) = \mathrm{av}_I f$ if $x \in I$ and $I \in \mathcal{D}_n$. That is, $\mathbb{E}_n f = \sum_{I\in\mathcal{D}_n} \mathrm{av}_I(f)\mathbb{1}_I$. For $k = 0,1,2,\dots$ define the projection operator $\mathbb{D}_k$ by $\mathbb{D}_k f(x) = \sum_{I\in\mathcal{D}_k}\langle f, \psi_I\rangle\psi_I(x)$, for $x \in [0,1)$. Prove that for f piecewise continuous on $[0,1)$ we have

$$\mathbb{D}_k f(x) = \mathbb{E}_{k+1} f(x) - \mathbb{E}_k f(x), \quad x \in [0,1).$$

7.199. Exercise. Let f be continuous in $[0,1]$. Prove that

$$\sum_{k=0}^{n-1}\sum_{I\in\mathcal{D}_k}\langle f, \psi_I\rangle\psi_I \to f - \mathbb{E}_0 f \quad \text{uniformly on } [0,1) \text{ as } n \to \infty.$$

Moreover, prove that for any collections $\widetilde{\mathcal{D}}_n \subset \mathcal{D}_n$

$$\sum_{I\in\widetilde{\mathcal{D}}_n}\langle f, \psi_I\rangle\psi_I \to 0 \quad \text{uniformly on } [0,1) \text{ as } n \to \infty.$$

Prove that the Haar system is complete in the space of Riemann integrable functions on $[0,1]$, in the sense of Definition 7.40, and we have for such functions

$$\|f\|_2^2 = \Big|\int_0^1 f\Big|^2 + \sum_{I\in\mathcal{D}}|\langle f, \psi_I\rangle|^2.$$

7.200. Exercise. Let (r_n) denote the Rademacher functions as defined in (7.10).

(i) Show that every r_n for $n \geqslant 1$ can be written as a finite linear combination of Haar wavelets and determine the coefficients of this linear combination.

(ii) Show that the orthonormal system on $[0,1]$ given by $(r_n)_n$ is not complete.

7.201. Exercise (Superorthogonality). Let (r_n) denote the Rademacher functions as defined in (7.10).

(i) Let $m \in \mathbb{N}$, $j_1 \leqslant j_2 \leqslant \cdots \leqslant j_{2m-1} \leqslant j_{2m}$ and assume that there is at least one $k \in \{1,\dots,m\}$ such that $j_{2k-1} < j_{2k}$. Then

$$\int_0^1 \prod_{\nu=1}^{2m} r_{j_\nu}(x)\,dx = 0.$$

This is referred to as the *superorthogonality property* of the Rademacher system.

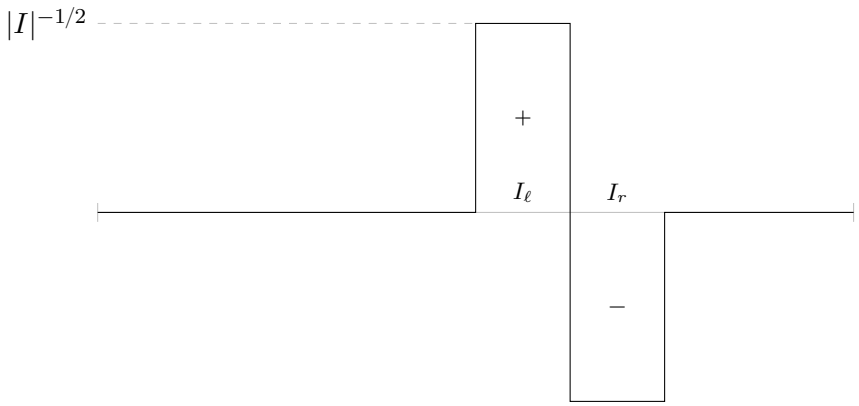


Figure 7.7. A Haar wavelet ψ_I.

(ii) Show using (i) that for $m = 1, 2, 3, \dots$ there is a constant $B_m \in (0, \infty)$ such that for all $N \geqslant 1$, and $a_1, \dots, a_N \in \mathbb{C}$ we have

$$\Big(\int \Big| \sum_{n=1}^{N} a_n r_n(x) \Big|^{2m} dx \Big)^{1/2m} \leqslant B_m \Big(\sum_{n=1}^{N} |a_n|^2 \Big)^{1/2}.$$

REMARKS. 1. See Pierce [**61**] for more on superorthogonality properties and their role in modern analysis, see also [**24**].

2. The inequality in (ii) continues to hold when $2m$ is replaced by an arbitrary $p \in (0, \infty)$. These inequalities and their reverse variants are known as *Khinchine's inequalities.*

CHAPTER 8

Baire's theorem and consequences

8.1. The Baire category theorem

In analysis, one often encounters "pathological" examples – think of Weierstrass' nowhere differentiable function (Example 3.40), or approximations that fail to converge uniformly (see Ch. 7). It is natural to ask how "common" such examples are. In this chapter we will prove *Baire's theorem*, a basic property of complete metric spaces that can often be used to show that pathological examples are in a sense quite common, or "generic". Baire's theorem has many interesting applications. We will discuss a few of the most common ones.

Let (X, d) be a metric space. Given $A \subset X$ we let A^o denote the interior of A. Note that A is dense if and only if $A \cap U \neq \varnothing$ for all nonempty open sets $U \subset X$.

8.1. Definition. A set $A \subset X$ is called *nowhere dense* if its closure has empty interior. In other words, if $(\overline{A})^o = \varnothing$. Equivalently, A is nowhere dense if and only if $\overline{A}$ contains no nonempty open set.

8.2. Observation. *(i) A subset of a nowhere dense set is nowhere dense.*

(ii) The closure of a nowhere dense set is nowhere dense.

8.3. Lemma. *Let $A \subset X$.*

(i) Assume that A is closed; then A has empty interior if and only if $A^{\complement} = X \backslash A$ is open and dense.

(ii) A is nowhere dense if and only if $A^{\complement}$ contains an open dense set.

(iii) A is nowhere dense if and only if A is contained in a closed set with empty interior.

Proof. (i) We use that a set A is closed if and only if $A^{\complement}$ is open. And A has empty interior if and only if $A^{\complement}$ is dense.

(ii) If A is nowhere dense then every ball contains an element in $(\overline{A})^{\complement}$ which is thus dense. So $(\overline{A})^{\complement}$ is an open dense set contained in $A^{\complement}$.

Vice versa if O is open and dense and contained in $A^{\complement}$, then A is contained in the closed set $O^{\complement}$ and thus $\overline{A} \subset O^{\complement}$. No $x \in \overline{A}$ can be an interior point of $\overline{A}$ as this would contradict the denseness of O; thus A is nowhere dense.

(iii) follows from (i) and (ii). □

8.4. Example. The Cantor middle third set introduced in Example 4.142,

$$\mathfrak{C} = [0,1] \backslash \bigcup_{\ell=0}^{\infty} \bigcup_{k=0}^{3^\ell - 1} \left(\tfrac{3k+1}{3^{\ell+1}}, \tfrac{3k+2}{3^{\ell+1}}\right).$$

is a closed subset of $[0, 1]$ and has empty interior. Therefore, it is nowhere dense.

8.5. Lemma. *Suppose $A_1, \ldots, A_n \subset X$ are nowhere dense sets. Then $\bigcup_{k=1}^n A_k$ is nowhere dense.*

Proof. Without loss of generality let $n = 2$. We need to show that $\overline{A_1} \cup \overline{A_2}$ has empty interior. Equivalently, setting $U_k = (\overline{A_k})^{\complement}$ for $k = 1, 2$, we show that $U_1 \cap U_2$ is dense. Let $U \subset X$ be a nonempty open set. Then $V_1 = U \cap U_1$ is open and nonempty, because U_1 is dense. Since U_2 is also dense, $V_1 \cap U_2 = U \cap (U_1 \cap U_2)$ is nonempty, so $U_1 \cap U_2$ is dense. □

However, countable unions of nowhere dense sets are not necessarily nowhere dense sets.

8.6. Example. Enumerate the rationals as $\mathbb{Q} = \{q_1, q_2, \dots\}$. For every $k = 1, 2, \dots$, the set $A_k = \{q_k\}$ is nowhere dense in $\mathbb{R}$. But $\mathbb{Q} = \bigcup_{k=1}^{\infty} A_k \subset \mathbb{R}$ is not nowhere dense (it is dense!).

8.7. Definition. A set $A \subset X$ is called *meager* (or of *first category*) in X if it is a countable union of nowhere dense sets. A is called *comeager* (or *residual*) if $A^{\complement}$ is meager. A set that is not meager is called *of second category*.

A set that is both, comeager and not meager is called *generic*. If a property holds on a generic set, one also says that it holds *generically* (in the Baire category sense).

The above example shows that $\mathbb{Q} \subset \mathbb{R}$ is meager. In fact, every countable subset of $\mathbb{R}$ is meager (because single points are nowhere dense in $\mathbb{R}$).

By definition, countable unions of meager sets are meager. The choice of the word "meager" suggests that meager sets are somehow "small" or "negligible". But how "large" can meager sets be? For example, can X be meager? That is, can we write the entire metric space X as a countable union of nowhere dense subsets? The Baire category theorem will show that the answer is no, *if X is complete*.

8.8. Theorem (Baire category theorem). *In a complete metric space, meager sets have empty interior. Equivalently, countable intersections of open dense sets are dense.*

8.9. Corollary. *Let X be a complete metric space. Then the following hold.*

(i) Let $A \subset X$ be a meager set. Then $A \neq X$. In other words, X is not a meager subset of itself.

(ii) A subset of X is comeager if and only if it is generic.

8.10. Example. The conclusion of the Baire category theorem fails if we drop the assumption that X is complete. Consider $X = \mathbb{Q}$ with the metric inherited from $\mathbb{R}$ (so $d(p, q) = |p - q|$). Then X is a meager subset of itself because it is countable and single points are nowhere dense in X (X has no isolated points). But the interior of X is nonempty, because X is open in X.

8.11. Example. Not every set with empty interior is meager: consider the irrational numbers $A = \mathbb{R}\backslash\mathbb{Q}$. A has empty interior, because $A^{\complement} = \mathbb{Q}$ is dense. It is not meager, because otherwise $\mathbb{R} = A \cup A^{\complement}$ would be meager, contradicting the Baire category theorem.

8.12. Exercise. The definition of Lebesgue null set is another notion of "smallness", see Definition 4.121. Find an example of a subset of $\mathbb{R}$ that is both a comeager set and a Lebesgue null set. This implies in particular that Lebesgue null sets are not necessarily meager and meager sets are not necessarily Lebesgue null sets.

For the proof of the Baire category theorem we will need the following lemma.

8.13. Lemma. *Let X be complete and let $A_1 \supset A_2 \supset \cdots$ be a decreasing sequence of nonempty closed sets in X such that*

$$\operatorname{diam}(A_n) = \sup_{x,y \in A_n} d(x, y) \longrightarrow 0$$

as $n \to \infty$. Then $\bigcap_{n=1}^{\infty} A_n$ is nonempty.

Proof. For every $n \geqslant 1$ we choose $x_n \in A_n$. Then $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, because for all $n \geqslant m$ we have $d(x_n, x_m) \leqslant \operatorname{diam}(A_m) \to 0$ as $m \to \infty$. Since X is complete, there exists $x \in X$ such that $\lim_{n\to\infty} x_n = x$. Let $N \in \mathbb{N}$. Then A_N contains the sequence $(x_n)_{n\geqslant N}$ and since A_N is closed, it must also contain the limit of this sequence, so $x \in A_N$. This proves $x \in \bigcap_{N=1}^{\infty} A_N$. □

Proof of Theorem 8.8. Let $(U_n)_{n\in\mathbb{N}}$ be open dense sets. We need to show that $\bigcap_{n=1}^{\infty} U_n$ is dense. Let $U \subset X$ be open and nonempty. It suffices to show that $U \cap \bigcap_{n=1}^{\infty} U_n$ is nonempty. Since U_1 is open and dense, $U \cap U_1$ is open and nonempty. Choose a closed ball $\overline{B}(x_1, r_1) \subset U \cap U_1$ with $r_1 \in (0, 1)$. Then $B(x_1, r_1) \cap U_2$ is open and nonempty (because U_2 is dense), so we can choose a closed ball $\overline{B}(x_2, r_2) \subset B(x_1, r_1) \cap U_2$ with $r_2 \in (0, \frac{1}{2})$. Iterating this process, we obtain a sequence of closed balls $(\overline{B}(x_n, r_n))_n$ such that $\overline{B}(x_n, r_n) \subset B(x_{n-1}, r_{n-1}) \cap U_n$ and $r_n \in (0, \frac{1}{n})$. By Lemma 8.13

there exists a point x contained in $\bigcap_{n=1}^{\infty} \overline{B}(x_n, r_n)$. Since $\overline{B}(x_n, r_n) \subset U \cap U_n$ for all $n \geqslant 1$, we have $x \in U \cap \bigcap_{n=1}^{\infty} U_n$. □

The Baire category theorem has many interesting consequences, some of which we will discuss in the remaining sections of this chapter.

8.2. Nowhere differentiable continuous functions

We have seen in Example 3.40 a construction of a continuous function on an interval that is *nowhere differentiable* (see also Exercise 3.41). It turns out that this behavior of continuous functions is typical (i.e. generic in the sense of Baire category).

8.14. Theorem. *Let $\mathcal{A} \subset C([0,1])$ be the set of all functions that are differentiable at least at one point in $[0,1]$. Then $\mathcal{A}$ is meager.*

Proof. For $n \in \mathbb{N}$ we define $\mathcal{A}_n$ as the set of all $f \in C([0,1])$ such that there exists $t \in [0,1]$ such that

$$\Big|\frac{f(t+h) - f(t)}{h}\Big| \leqslant n$$

holds for all $h \in \mathbb{R}$ with $t + h \in [0,1]$. Then

$$\mathcal{A} \subset \bigcup_{n=1}^{\infty} \mathcal{A}_n.$$

It suffices to show that each $\mathcal{A}_n$ is nowhere dense.

We first prove that $\mathcal{A}_n$ is closed. Let $(f_k)_{k\in\mathbb{N}} \subset \mathcal{A}_n$ be a sequence that converges to some $f \in C([0,1])$. We show that $f \in \mathcal{A}_n$. Indeed, by assumption, there exists $(t_k)_{k\in\mathbb{N}} \subset [0,1]$ such that

$$\Big|\frac{f_k(t_k+h) - f_k(t_k)}{h}\Big| \leqslant n$$

holds for all $k \geqslant 1$ if $t_k + h \in [0,1]$. By the Bolzano-Weierstrass theorem, there exists a subsequence $t_{\kappa(k)}$ (with $\kappa(k)$ strictly increasing) which converges to some $t_* \in [0,1]$. Then, by continuity of f,

$$\Big|\frac{f(t_*+h) - f(t_*)}{h}\Big| = \lim_{k\to\infty} \Big|\frac{f_{\kappa(k)}(t_{\kappa(k)}+h) - f_{\kappa(k)}(t_{\kappa(k)})}{h}\Big| \leqslant n.$$

Therefore, $f \in \mathcal{A}_n$ and $\mathcal{A}_n$ is closed.

We now show that $\mathcal{A}_n$ has empty interior, i.e. that $C([0,1])\backslash\mathcal{A}_n$ is dense. We need to show that every $f \in C([0,1])$ can be approximated by functions in $\mathcal{A}_n^{\complement}$. Fix f and let $\varepsilon > 0$. Let p be a differentiable function with continuous derivative such that $\|f - p\|_\infty < \varepsilon/2$ (by the Weierstrass theorem, we can find a polynomial with this property). Let $\|p'\|_\infty = B$. Let $N > (2n + B)\varepsilon^{-1}$ and let s be the continuous piecewise linear function with $s(\frac{k}{N}) = 0$ for $k = 0, \dots, N$ and $s(\frac{k}{N} + \frac{1}{2N}) = \frac{\varepsilon}{2}$, and s is linear on the intervals $[\frac{\ell-1}{2N}, \frac{\ell}{2N}]$, $\ell = 1, \dots 2N$. Note that $\|s\|_\infty = \varepsilon/2$ and the slopes of the linear functions are $N\varepsilon > 2n + B$. Thus for every fixed $t \in [0,1]$

$$\sup\Big\{\Big|\frac{s(t+h) - s(t)}{h}\Big| : t + h \in [0,1]\Big\} = N\varepsilon > 2n + B.$$

And, for $g = p + s$ we have $\|f - g\|_\infty \leqslant \|f - p\|_\infty + \|s\|_\infty < \varepsilon$ and

$$\Big|\frac{g(t+h) - g(t)}{h}\Big| \geqslant \Big|\frac{s(t+h) - s(t)}{h}\Big| - \Big|\frac{p(t+h) - p(t)}{h}\Big| \geqslant \Big|\frac{s(t+h) - s(t)}{h}\Big| - B$$

and therefore, for each $t \in [0,1]$,

$$\sup\Big\{\Big|\frac{g(t+h) - g(t)}{h}\Big| : t + h \in [0,1]\Big\} > 2n.$$

Hence $g \notin \mathcal{A}_n$. □

The Baire category theorem implies that $\mathcal{A}$ has empty interior. In other words, the set of nowhere differentiable functions $C([0,1])\backslash\mathcal{A}$ is dense. In this sense, it is "generic" behavior for continuous functions to be nowhere differentiable. In particular, we can conclude that there exists a continuous nowhere differentiable function on $[0,1]$ without actually constructing such a function.

8.3. Sets of continuity

It is natural to ask which subsets of X arise as the continuity set of some function on X. For instance, does there exist a function $f : \mathbb{R} \to \mathbb{R}$ that is continuous precisely at rational x and discontinuous everywhere else?

8.15. Definition. Let X, Y be metric spaces and $f : X \to Y$ a map. The set

$$C_f = \{x \in X \,:\, f \text{ is continuous at } x\} \subset X$$

is called the *set of continuity of* f. Similarly, $D_f = X\backslash C_f$ is called the *set of discontinuity of* f.

8.16. Example. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = 1$ if x is rational and $f(x) = 0$ if x is irrational. Then $C_f = \varnothing$.

8.17. Example. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x$ if x is rational and $f(x) = 0$ if x is irrational. Then $C_f = \{0\}$.

8.18. Example. Recall Thomae's function from Exercises 1.84, 4.53, extended to the full real line: set $f(0) = 1$ and $f(x) = 1/q$ if $x = \frac{p}{q}$, where $p \in \mathbb{Z}$, $q \in \mathbb{N}$ and p, q are coprime. If $x \notin \mathbb{Q}$, set $f(x) = 0$. Then $C_f = \mathbb{R}\backslash\mathbb{Q}$.

8.19. Definition. Let X be a metric space. A set $A \subset X$ is called an F_σ *set* if it is a countable union of closed sets. A set $G \subset X$ is called a G_δ *set* if it is a countable intersection of open sets.

These names are historically motivated. The F in F_σ is for *fermé* which is French for *closed.* On the other hand, the G in G_δ is for *Gebiet* which is German for *region.* Moreover the subscript δ is for the German *Durchschnitt* which means *intersection* while the subscript σ indicates *Summe* (sum) which stands here for *union.*

8.20. Examples. 1. Every open set is a G_δ set and every closed set is an F_σ set.
2. Let $x \in X$. Then $\{x\}$ is a G_δ set: it is the intersection of the open balls $B(x, 1/n)$.
3. $\mathbb{Q} \subset \mathbb{R}$ is an F_σ set, because $\mathbb{Q} = \bigcup_{q\in\mathbb{Q}}\{q\}$ (a countable union of closed sets).

8.21. Lemma. *Let X be a metric space.*
(i) A subset A of X is a G_δ set if and only if $A^\complement$ is an F_σ set.
(ii) An F_σ set in X is either meager or has nonempty interior.
(iii) A G_δ set in X is either comeager or not dense.

Proof. (i) If $A = \bigcup_n A_n$, then $A^\complement = \bigcap_n A_n^\complement$, so the claim follows.

(ii) Let $A \subset X$ be an F_σ set with empty interior. Then it is a countable union of closed sets with empty interior and therefore meager.

(iii) Let E be a G_δ set which is dense. Then $E^\complement$ is an F_σ set with empty interior, and thus $E^\complement$ is meager. Hence E is comeager. □

8.22. Theorem. *Let X and Y be metric spaces and $f : X \to Y$ a map. Then $C_f \subset X$ is a G_δ set and D_f is an F_σ set.*

Proof. Let $f : X \to Y$ be given. It suffices to show that C_f is a G_δ set. We recall the notion of oscillation $\operatorname{osc}_f(x)$ from Definition 4.49; we have $x \in C_f$ if and only if $\operatorname{osc}_f(x) = 0$. Hence if $W_n = \{x \in X \,:\, \operatorname{osc}_f(x) < 1/n\}$ then $C_f = \bigcap_{n\in\mathbb{N}} W_n$. The sets W_n are open, by Lemma 4.51, and thus C_f is a G_δ set. □

As a sample application of the Baire category theorem we now answer one of our previous questions negatively:

8.23. Example. $\mathbb{Q} \subset \mathbb{R}$ is not a G_δ set. Consequently, there exists no function $f : \mathbb{R} \to \mathbb{R}$ such that $C_f = \mathbb{Q}$. To see this, we argue by contradiction. If $\mathbb{Q}$ were a G_δ set, then $\mathbb{R}\backslash\mathbb{Q}$ would be an F_σ set with empty interior, thus would be meager by Lemma 8.21 (ii). But then $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R}\backslash\mathbb{Q})$ would be meager too, in contradiction to the Baire category theorem.

Remark. It is natural to ask if the converse of Theorem 8.22 is true in the following sense: given a G_δ set $G \subset X$, can we find a function $f : X \to \mathbb{R}$ such that $C_f = G$? This cannot hold in general: suppose X contains an isolated point, that is X contains an open set of the form $\{x\}$. Then necessarily $x \in C_f$, but x is not necessarily contained in every possible G_δ set. However, this turns out to be the only obstruction: if X contains no isolated points, then for every G_δ set $G \subset X$ one can find $f : X \to \mathbb{R}$ such that $C_f = G$. For a very short proof of this, see Kim [**36**].

8.4. Baire functions

Consider the *Dirichlet function* D

$$D(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases} \tag{8.1}$$

It is natural to ask whether D is the pointwise limit of a sequence of continuous functions. The answer turns out to be no.

8.24. Definition. Let X be a metric space. A function $f : X \to \mathbb{R}$ is a *Baire-1 function* if there is a sequence of continuous functions $f_n : X \to \mathbb{R}$ such that $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in X$.

A Baire-1 function does not have to be continuous everywhere in X. However the following theorem shows that it will be continuous on a comeager set.

8.25. Theorem. *Let X be a complete metric space and let $f : X \to \mathbb{R}$ be a Baire-1 function. Then the set*

$$C_f = \{x : f \text{ is continuous at } x\}$$

is a dense G_δ set, and thus comeager.

In the proof of this theorem we shall apply the Baire category theorem twice. The first application is used in the following lemma.

8.26. Lemma. *Let Y be a complete metric space, and let $f_n : Y \to \mathbb{R}$ be continuous functions on Y converging pointwise to f. Then for every $\alpha > 0$ there is an $N \in \mathbb{N}$ and an open ball B such that*

$$|f_n(x) - f(x)| \leqslant \alpha$$

for all $n \geqslant N$ and all $x \in B$.

Proof. Let

$$A_n = \{x \in Y : |f_n(x) - f_k(x)| \leqslant \alpha \text{ for all } k > n\}.$$

First observe that A_n is a closed set. Indeed, the set $E_{n,k} = \{x : |f_n(x) - f_k(x)| \leqslant \alpha\}$ is closed since $|f_n - f_k|$ is continuous and we have $A_n = \bigcap_{k=n+1}^{\infty} E_{n,k}$.

Since f_n converges pointwise to f, we have $|f_n(x) - f(x)| \leqslant \alpha$ for $x \in A_n$, and $Y = \bigcup_{n=1}^{\infty} A_n$. By Baire's theorem there exists $n \in \mathbb{N}$ such that A_n is not nowhere dense and since A_n is closed there is an open ball $B \subset A_n$. The claim follows. □

Proof of Theorem 8.25, conclusion. As in the previous section we use $\operatorname{osc}_f(x)$ as in definition 4.49. Consider the sets $W_M = \{x : \operatorname{osc}_f(x) < 1/M\}$ which are open by Lemma 4.51 and we have $C_f = \bigcap_{M=1}^{\infty} W_M$. We show that W_M is dense in X. Let B_0 be an open ball. We need to show that its intersection with W_M is not empty. Let B_1 be an open ball such that $\overline{B_1} \subset B_0$. We apply Lemma

8.26 with the complete metric space $\overline{B}_1$, and $\alpha = (4M)^{-1}$ and find an open ball B_2 (which is an open ball in X contained in B_0) and $n \in \mathbb{N}$ such that $|f_n(x) - f(x)| \leqslant (4M)^{-1}$ for all $x \in B_2$. Since f_n is continuous we find an open ball $B_3 \subset B_2$ such that $|f_n(x) - f_n(y)| < (4M)^{-1}$ for all $x, y \in B_3$. Hence

$$|f(x) - f(y)| \leqslant |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| \leqslant \tfrac{3}{4M} < \tfrac{1}{M}$$

for $x, y \in B_3$ and therefore $\operatorname{osc}_f(x) < 1/M$ for all $x \in B_3$. Thus $B_3 \subset W_M$.

We have identified the G_δ set C_f as a countable intersection of dense open sets, hence it is a dense set by Baire's theorem, and thus a comeager set. □

REMARK. Baire considered a hierarchy of increasing classes that we refer to as *Baire-n classes*: One defines the *Baire-0 class* as the class of continuous functions on X. Then for $n \geqslant 1$ the Baire-n class consists of the pointwise limits of sequences of functions in the Baire-$(n-1)$ class.

As an illustration consider an enumeration $\{r_n\}$ of the rational numbers and define the function D_n so that $D_n(r_k) = 1$ for $1 \leqslant k \leqslant n$ and $D_n(x) = 0$ elsewhere. Verify that the functions D_n are Baire-1 and that D_n converges pointwise to the Dirichlet function in (8.1). This identifies D as a Baire-2 function which by Theorem 8.25 is not Baire-1. Alternatively one can also use the formula $D(x) = \lim_{j\to\infty}(\lim_{m\to\infty}(\cos(j!\pi x))^{2m})$ to show that D is Baire-2.

8.5. The Banach-Steinhaus theorem

We revisit the *uniform boundedness principle* from Theorem 5.35, but we now get a somewhat stronger conclusion by Banach and Steinhaus (1927) based on the Baire category ideas.

8.27. THEOREM (Banach–Steinhaus). *Let X, Y be normed spaces and $\mathcal{F} \subset L(X,Y)$ be a family of bounded linear operators. Then the following holds.*

(i) If $\sup_{T\in\mathcal{F}} \|Tx\|_Y < \infty$ for all x in a nonmeager (i.e. second category) set, then $\sup_{T\in\mathcal{F}} \|T\|_{\mathrm{op}} < \infty$.

(ii) Suppose $\sup_{T\in\mathcal{F}} \|T\|_{\mathrm{op}} = \infty$. Then $\{x \in X : \sup_{T\in\mathcal{F}} \|Tx\|_Y = \infty\}$ is comeager.

PROOF. Statements (i) and (ii) are equivalent, so it suffices to prove (i).

Assume that $\sup_{T\in\mathcal{F}} \|Tx\|_Y < \infty$ for all x in a nonmeager set. Define

$$A_n = \big\{x \in X \,:\, \sup_{T\in\mathcal{F}} \|Tx\|_Y \leqslant n\big\}.$$

A_n is a closed set: if $(x_k)_{k\in\mathbb{N}} \subset A_n$ is a sequence with $x_k \to x \in X$, then since T is continuous, $\|Tx\|_Y = \lim_{k\to\infty} \|Tx_k\|_Y \leqslant n$ for all $T \in \mathcal{F}$, so $x \in A_n$.

The assumption $\sup_{T\in\mathcal{F}} \|Tx\|_Y < \infty$ for all x in a nonmeager set implies that there exists $n_0 \in \mathbb{N}$ such that A_{n_0} has nonempty interior. This means that there exists $x_0 \in A_{n_0}$ and $\varepsilon > 0$ such that

$$\overline{B}(x_0, \varepsilon) \subset A_{n_0}.$$

Let $x \in X$ be such that $\|x\|_X \leqslant \varepsilon$. Then for all $T \in \mathcal{F}$,

$$\|Tx\|_Y = \|T(x_0 - x) - Tx_0\|_Y \leqslant \|T(x_0 - x)\|_Y + \|Tx_0\|_Y \leqslant 2n_0.$$

Now we use the usual scaling trick: let $x \in X$ satisfy $\|x\|_X = 1$. Thus for all $T \in \mathcal{F}$

$$\|Tx\|_Y = \varepsilon^{-1}\|T(\varepsilon x)\|_Y \leqslant 2\varepsilon^{-1} n_0.$$

This implies

$$\sup_{T\in\mathcal{F}} \|T\|_{\mathrm{op}} = \sup_{T\in\mathcal{F}} \sup_{\|x\|_X=1} \|Tx\|_Y \leqslant 2\varepsilon^{-1} n_0 < \infty. \qquad \square$$

8.28. THEOREM (Banach–Steinhaus uniform boundedness principle). *Let X be a Banach space and Y a normed vector space. Let $\mathcal{F} \subset L(X,Y)$ be a family of bounded linear operators. Then the following holds. Suppose that $\sup_{T\in\mathcal{F}} \|T\|_{\mathrm{op}} = \infty$. Then the set $\{x : \sup_{T\in\mathcal{F}} \|Tx\|_Y < \infty\}$ has empty interior, and the set $\{x : \sup_{T\in\mathcal{F}} \|Tx\|_Y = \infty\}$ is generic and dense.*

PROOF. The set

$$\mathcal{A} = \big\{x : \sup_{T\in\mathcal{F}} \|Tx\|_Y = \infty\big\}$$

is comeager by part (ii) of Theorem 8.27. By the Baire category theorem it is also generic. Its complement, the meager set $\mathcal{A}^{\complement} = \{x : \sup_{T\in\mathcal{F}} \|Tx\|_Y < \infty\}$ is a linear proper subspace of X (by the linearity of $T \in \mathcal{F}$ and the triangle inequality). We conclude by observing that a proper linear subspace of X has empty interior and its complement is dense. □

REMARK. Theorem 8.28 is a stronger version of Theorem 5.35, but unlike the proof of Theorem 5.35 the proof of Theorem 8.28 is nonconstructive since it involves the Baire category theorem in its proof.

8.29. EXAMPLE (Failure of uniform convergence of Lagrange interpolation polynomials). Consider sequences of Lagrange interpolation polynomials $L_N f$ associated to given distinct nodes $(x_{k,N})_{k=0}^N$ in $[a, b]$, and satisfying $L_N f(x_{k,N}) = f(x_{k,N})$. In Theorem 7.121 it was shown that the $C([a, b])$-operator norm satisfies $\sup \|L_N\|_{\mathrm{op}} = \infty$ and by the uniform boundedness principle this implies the existence of a function $f \in C([a, b])$ for which $\sup_N \sup_{t\in[a,b]} |L_N f(t)| = \infty$; in particular uniform convergence fails. The Baire category version of the uniform boundedness principle yields failure of uniform convergence for a generic dense family of $C([a, b])$ functions.

8.30. EXAMPLE (Divergence of Fourier series at a point). Recall that for a 1-periodic continuous function $f : \mathbb{R} \to \mathbb{C}$ we defined the partial sums of its Fourier series by

$$S_N f(x) = \sum_{n=-N}^{N} \widehat{f}(n) e^{2\pi i n x} = (f * D_N)(x), \tag{8.2}$$

where $\widehat{f}(n) = \int_0^1 f(t)e^{-2\pi i t n} dt$ and $D_N(x) = \sum_{n=-N}^{N} e^{2\pi i x n} = \frac{\sin(2\pi(N+\frac{1}{2})x)}{\sin(\pi x)}$ is the Dirichlet kernel (see §7.4). We have seen in §7.5.2 using the uniform boundedness principle that for every fixed $x_0 \in \mathbb{R}$ there exists a 1-periodic continuous function f such that $\sup_N |S_N f(x_0)| = \infty$, in particular the Fourier series of f diverges at x_0. Using Theorem 8.27 a stronger result follows, namely the divergence of the Fourier series of a $C(\mathbb{T})$ function at a point x_0 is generic in the Baire category sense. In particular, there must be a generic dense family of functions in $C(\mathbb{T})$ for which the Fourier series diverges at x_0.

A strengthening of this example can be obtained by working with the linear functionals $\Lambda_{n,j}(f) = S_N f(x_j)$, where $N \in \mathbb{N}_0$ and $\{x_j : j \in \mathbb{N}\}$ is a countable dense set of points in $\mathbb{T}$:

8.31. EXERCISE. Prove that there is a comeager set $\mathcal{E} \subset C(\mathbb{T})$ such that

$$\Gamma_f = \{x \in \mathbb{T} : \sup_n |S_n f(x)| = \infty\}$$

is a dense subset of $\mathbb{T}$ for every $f \in \mathcal{E}$.

A further substantial strengthening of this divergence result is contained in Exercise 8.63 below.

8.6. Open mappings

We present three results from functional analysis where Baire's theorem is an essential tool. They are all related to the open mapping theorem in §8.6.1 below. We first discuss a striking special case of that theorem which is called the *theorem of the inverse operator.*

8.32. THEOREM. *Let X and Y be Banach spaces and let $T : X \to Y$ be a bounded linear operator which is also bijective. Then its inverse $T^{-1} : Y \to X$ is also a bounded operator.*

If X and Y are finite-dimensional, this result is immediate as all linear operators on finite-dimensional normed spaces are bounded. However, no useful general conclusion about the operator norm of the inverse can be made. Imagining a passage from a finite-dimensional situation to an infinite-dimensional situation, possibly by a limiting process, one should expect that some quantitative information about the inverses of the approximating operators is needed. Yet in Theorem 8.32 no such

quantitative information enters, so this seems a rather surprising result. The assumption that both the spaces X and Y are complete is crucial for the proof of this theorem. We postpone the proof of Theorem 8.32 to §8.6.1 where it will be deduced from the more general *open mapping theorem*.

8.33. Example. We consider the *Fourier transform* on the sequence space ℓ^1, i.e. map

$$\mathcal{F}_{\mathbb{Z}} : \begin{cases} \ell^1 \to C(\mathbb{T}) \\ a = (a_n)_{n\in\mathbb{Z}} \mapsto \mathcal{F}_{\mathbb{Z}}a \text{ with } \mathcal{F}_{\mathbb{Z}}a(x) = \sum_{n=-\infty}^{\infty} a_n e^{-2\pi i n x} \end{cases}$$

which assigns to $a \in \ell^1(\mathbb{Z})$ the 1-periodic continuous function with Fourier coefficients a_n. This is an injective map since we can recover

$$a_n = \int_0^1 (\mathcal{F}_{\mathbb{Z}}a)(t) e^{2\pi i n t} dt.$$

We have already seen in §7.5.4 that this map is not surjective and now recast the proof using the theorem of the inverse operator. We argue by contradiction and assume that $\mathcal{F}_{\mathbb{Z}}$ is surjective, hence bijective and by Theorem 8.32 it has then a bounded inverse. That means we have an inequality $\|a\|_{\ell^1} \leqslant C\|\mathcal{F}_{\mathbb{Z}}a\|_\infty$ for some C. Let (α_n) be the Golay-Rudin-Shapiro sequence (see Definitions 7.90, 7.88) and consider this inequality with $a = a^{(\ell)}$ where $a^{(\ell)}(n) = \alpha_n$ for $0 \leqslant n \leqslant 2^\ell - 1$, and $a^{(\ell)}(n) = 0$ for $n \geqslant 2^\ell$ and for $n < 0$. Then $\mathcal{F}_{\mathbb{Z}}a^{(\ell)} = p_\ell$, the Rudin–Shapiro trigonometric polynomial. We have $\|p_\ell\|_\infty \leqslant 2^{(\ell+1)/2}$ from Lemma 7.89 and since $\|a^{(\ell)}\|_{\ell^1} = 2^\ell$ we get $2^\ell \leqslant C2^{(\ell+1)/2}$ for all $\ell \in \mathbb{N}$. This gives a contradiction for $2^\ell > 2C^2$.

The following examples show that the assumption that X and Y are complete is necessary for Theorem 8.32.

8.34. Example. Let $X = \mathfrak{c}_0$, the space of sequences $a = (a_n)_{n\in\mathbb{N}}$ with

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

We take the ℓ^∞-norm $\|a\| = \sup_k |a_k|$ on $\mathfrak{c}_0$ which makes $\mathfrak{c}_0$ a subspace of the Banach space ℓ^∞. Thus $\mathfrak{c}_0$ is a Banach space (see Example 5.3). Let Y be the space of sequences $b = (b_k)_{k\in\mathbb{N}}$ such that

$$kb_k = o(1) \quad \text{as} \quad k \to \infty.$$

This is a dense subspace of X. We also take the ℓ^∞-norm on Y so that Y is not complete. Let Ta be the sequence with

$$(Ta)_k = k^{-1}a_k.$$

Then $T : X \to Y$ and $\|Ta\|_\infty \leqslant \|a\|_\infty$ so that T is continuous. Also, T is bijective, and T^{-1} is given by $(T^{-1}b)_n = nb_n$. Let $e^{(\ell)}$ be the sequence with $e^{(\ell)}_\ell = 1$ and $e^{(\ell)}_k = 0$ if $k \neq \ell$. Then $e^{(\ell)} \in Y$ and $\|e^{(\ell)}\|_\infty = 1$. But $T^{-1}e^{(\ell)} = \ell e^{(\ell)}$ and thus $\|T^{-1}e^{(\ell)}\|_\infty = \ell$ and we see that T^{-1} is unbounded. This means that the assumption of completeness of Y is necessary in Theorem 8.32.

8.35. Example. Let X be a real infinite-dimensional normed space. Then there exists an unbounded linear operator $A : X \to X$.

To go ahead we use Zorn's lemma (Theorem 1.8) which implies that every vector space X has a basis $\mathcal{B}$ in the sense of linear algebra, also known as a *Hamel basis*. That is, there is a (maximal) collection $\mathcal{B} = \{u_\alpha : \alpha \in \mathcal{I}\}$ of vectors in X so that the set $\mathcal{B}$ is linearly independent and every $v \in X$ is a unique linear combination of a *finite* subset of $\mathcal{B}$. To see this, consider the family $\mathfrak{S}$ consisting of collections of linearly independent subsets of X. $\mathfrak{S}$ is partially ordered by inclusion. Chains of linearly independent subsets have their union as an upper bound. By Zorn's lemma there exists a maximal element in $\mathfrak{S}$. One can show that it is a Hamel basis (the reader should provide details as an exercise).

Here we consider the situation that X is a normed space. Given a Hamel basis $\mathcal{B} = \{u_\alpha : \alpha \in \mathcal{I}\}$ and a set $\{c_\alpha\}_{\alpha\in\mathcal{I}}$ of real numbers we may then uniquely define a linear map $A : X \to X$ so that $Au_\alpha = c_\alpha u_\alpha$.

If we impose the condition that $\sup_\alpha |c_\alpha| = \infty$ we see that $\sup_\alpha \|Au_\alpha\|/\|u_\alpha\| = \sup_\alpha |c_\alpha| = \infty$ and thus A is not a bounded operator.

8.36. Example. We use Example 8.35 to show that the assumption of completeness of X is needed in Theorem 8.32. Let Y be an infinite-dimensional Banach space with norm $\|\cdot\|_Y$ and use the product norm $\|(v_1, v_2)\|_{Y\times Y} = \max\{\|v_1\|_Y, \|v_2\|_Y\}$ on $Y \times Y$. By Example 8.35 there is an unbounded linear map $A : Y \to Y$. Let

$$X = \operatorname{graph}(A) = \{(v, Av) : v \in Y\} \subset Y \times Y$$

and take on X the norm inherited from $Y \times Y$. Define

$$T : \begin{cases} X \to Y \\ (v, Av) \mapsto v \end{cases}$$

Then T is a linear bijective map and $\|Tv\|_Y = \|v\|_Y \leqslant \max\{\|v\|_Y, \|Av\|_Y\} = \|(v, Av)\|_X$ so that T is a bounded operator with $\|T\|_{\mathrm{op}} = 1$. On the other hand $\|T^{-1}v\|_X = \max\{\|v\|_Y, \|Av\|_Y\}$ and because of the unboundedness of A the inequality $\|T^{-1}v\|_X \leqslant C\|v\|_Y$ fails for suitable v, no matter how large C is.

8.6.1. The open mapping theorem.

8.37. Definition. Let X, Y be metric spaces and $f : X \to Y$. The function f is called an open map if $f(U)$ is open in Y whenever U is open in X.

8.38. Theorem (Open mapping theorem). *Let X and Y be Banach spaces and let $T : X \to Y$ be a bounded linear operator which is also surjective. Then T is an open map.*

We first deduce the theorem of the inverse operator.

Proof of Theorem 8.32. If $T : X \to Y$ is a linear bijective bounded operator, Theorem 8.38 shows that for every open set O in X we have $T(O)$ is open in Y. But $T(O) = (T^{-1})^{-1}(O)$, the inverse image of O under the map T^{-1}. This means that the linear map $T^{-1} : Y \to X$ is continuous, i.e. a bounded operator in $L(Y, X)$. □

We now give the proof of the open mapping theorem.

Proof of Theorem 8.38. We will prove the theorem in four steps, verifying Claims 1-4 below.

Claim 1: Let $G \subset X$ be an open set in X such that $0 \in G$. Then the closure $\overline{T(G)}$ of $T(G)$ contains an interior point.

To see this, observe that 0 is an interior point of G so we can find an open ball $B(0, \varepsilon)$ with $B(0, \varepsilon) \subset G$. Then $\bigcup_{n=1}^\infty nB(0, \varepsilon) = X$ since $nB(0, \varepsilon) = B(0, n\varepsilon)$. Therefore also $X = \bigcup_{n=1}^\infty nG$ where $nG = \{nx : x \in G\}$. Hence

$$Y = \bigcup_{n\geqslant 1} T(nG) = \bigcup_{n\geqslant 1} nT(G)$$

by linearity of T. By Baire's theorem applied to the complete metric space Y the sets $nT(G)$ are not all nowhere dense, so there exists $n_0 \in \mathbb{N}$ such that $\overline{n_0T(G)} = n_0\overline{T(G)}$ has an interior point, say y_0. This also means that $n_0^{-1}y_0$ is an interior point of $\overline{T(G)}$, thus there is an open set contained in $\overline{T(G)}$.

Claim 2: Let $G \subset X$ be an open set in X such that $0 \in G$. Then 0 is an interior point of $\overline{T(G)}$.

To see this, first consider $X \times X$ as a Banach space with norm $\|(x_1, x_2)\| = \max\{\|x_1\|_X, \|x_2\|_X\}$ and observe that the linear map

$$\mathrm{Diff} : \begin{cases} X \times X \to X, \\ (x_1, x_2) \mapsto x_1 - x_2, \end{cases}$$

is continuous. Since G is an open neighborhood of the origin in X there is a neighborhood U of this origin such that $U - U = \mathrm{Diff}^{-1}(U) \subset G$ is an open neighborhood of the origin. Therefore

$$\overline{T(G)} \supset \overline{TU - TU} \supset \overline{TU} - \overline{TU}.$$

Now by the claim above (applied with U in place of G), there is an open neighborhood V_0 of 0 in Y such that $V_0 \subset \overline{T(U)}$, and we obtain that $V_0 - V_0 \subset \overline{TG}$. Since $V_0 - V_0$ is an open neighborhood of the origin in Y we have proved Claim 2.

Claim 3: For every open neighborhood U of 0 in X there exists a neighborhood V of the origin in Y such that $V \subset T(U)$. Fix U, and let B_r be the ball of radius r centered at the origin in X. We find $r > 0$ such that $B_r \subset U$. By Claim 2, we find $\delta > 0$, such that $B_Y(0, 2\delta) \subset \overline{TB_r}$. Then for all $k \in \mathbb{N}$, $B_Y(0, \frac{2\delta}{2^k})$ is contained in $\frac{1}{2^k}\overline{TB_r} = \overline{2^{-k}TB_r} = \overline{TB_{2^{-k}r}}$. We observed

$$B_Y(0, \tfrac{2\delta}{2^k}) \subset \overline{TB_{2^{-k}r}}$$

We shall now show that $B_Y(0, \delta) \subset TB_r$. Fix $y \in TB_r$. By the displayed containment for $k = 1$ we find $x_1 \in B_{r/2}$ such that $\|y - Tx_1\|_Y < \delta/2$. We claim that for $n = 1, 2, 3, \dots$ there are points $x_n \in B_{2^{-n}r}$ such that $\|y - \sum_{j=1}^n Tx_j\| < 2^{-n}\delta$, and $x_j \in B_{2^{-n}r}$. We have shown this for $n = 1$.

Assume this is true for some $n \in \mathbb{N}$. We use the displayed containment for $k = n + 1$, so $B_Y(0, 2^{-n}\delta) \subset \overline{TB_{2^{-n-1}r}}$. Thus we get $y - \sum_{j=1}^n Tx_j \in \overline{TB_{2^{-n-1}r}}$. Using the definition of closure we find a point in $TB_{2^{-n-1}r}$ which has distance less than $2^{-n-1}\delta$ to $y - \sum_{j=1}^n Tx_j$. In other words we find x_{n+1} with $\|x_{n+1}\| < 2^{-n-1}r$ such that $\|y - \sum_{j=1}^{n+1} Tx_j\|_Y < 2^{-n-1}\delta$ which proves the assertion for $n + 1$.

Now the sequence $\sum_{j=1}^n x_j$ converges since $\|x_j\| < 2^{-j}r$ and X is complete. Let $x = \sum_{j=1}^\infty x_j$. Then $\|x\| < r$ and $Tx = \lim_{n\to\infty} \sum_{j=1}^n Tx_j = y$ and $y \in TB_r \subset TU$ as claimed.

Claim 4: Let O be an open set in X. Then $T(O)$ is open in Y.

Let $y \in T(O)$ so that $y = Tx$ with $x \in O$. There exists an open neighborhood U of $0 \in X$ such that $x + U \subset O$. Then

$$y + T(U) = Tx + T(U) \subset T(O).$$

By Claim 3 there exists a neighborhood V of $0 \in Y$ such that $V \subset T(U)$. Hence $y + V \subset T(O)$. Since $y + V$ is an open neighborhood of y we see that y is an interior point of $T(O)$. We have shown that $T(O)$ is open. □

8.6.2. The closed graph theorem.

8.39. Definition. For a linear map $T : X \to Y$ we define

$$\mathrm{graph}(T) = \{(x, Tx) \in X \times Y : x \in X\}$$

If norms $\|\cdot\|_X$, $\|\cdot\|_Y$ are given on X, Y we set $\|(x, Tx)\|_{\mathrm{graph}(T)} = \max\{\|x\|_X, \|Tx\|_Y\}$.

We observe that if $T : X \to Y$ is continuous then $\mathrm{graph}(T)$ is a closed subset of $X \times Y$. Indeed if $(x, y) \in \overline{\mathrm{graph}(T)}$ and $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} Tx_n = y$ the continuity of T implies $y = Tx$; so graph(T) is closed.

It is significantly easier to check that the graph of T is closed; one has to verify that if $x_n \to x$ and $Tx_n \to y$ then $y = Tx$ (but one does not have to show convergence of Tx_n as it is assumed here). The following result tells us that for linear maps on complete spaces the two notions of continuity and closed graph are equivalent.

8.40. Theorem (Closed graph theorem). *Let X, Y be Banach spaces, and $T : X \to Y$ a linear map. Then T is continuous on X if and only if* $\mathrm{graph}(T)$ *is closed in $X \times Y$.*

PROOF. The 'only if' direction was just observed above (without a linearity assumption on T and without completeness assumptions on X and Y). Now assume that $\operatorname{graph}(T)$ is closed in $X \times Y$. The linear map

$$P : \begin{cases} \operatorname{graph}(T) \to X \\ (x, Tx) \mapsto x \end{cases}$$

is bijective and a bounded operator. Hence by Theorem 8.32 its inverse is continuous, and therefore we get the inequality

$$\|P^{-1}x\|_{X\times Y} = \max\{\|x\|_X, \|Tx\|_Y\} \leqslant C\|x\|_X$$

for some $C \geqslant 0$. This implies the continuity of T. □

8.41. EXAMPLE. The assumption of linearity in Theorem 8.40 is crucial. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

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

Then f is a bijective map on $\mathbb{R}$ with a closed graph in $\mathbb{R} \times \mathbb{R}$ but f is not continuous at 0. A closed graph theorem for nonlinear functions can be proved under a compactness assumption of the target space Y, see Exercise 4.184.

The completeness assumptions on both X and Y are needed in Theorem 8.40; this follows from

8.42. EXERCISE. Revisit Examples 8.34 and 8.36 and show that in both examples the unbounded operator $T^{-1} : Y \to X$ has a closed graph.

8.43. EXERCISE. Let $Y = C([0,1])$ and let $X = C^1([0,1])$ be the subspace of functions which have a continuous derivative on $[0,1]$, both with the norm $\|f\| = \sup_{t\in[0,1]} |f(t)|$. Show that the operator of differentiation $\frac{d}{dx} : X \to Y$ is an unbounded linear operator which has a closed graph.

We finish this section with an application of the closed graph theorem. In this example a purely algebraic condition for certain linear operators on Hilbert spaces leads to an analysis implication about boundedness.

8.44. DEFINITION. Let X be an inner product space and $T : X \to X$ be a linear map. We say that T is *self-adjoint* if $\langle Tx, y\rangle = \langle x, Ty\rangle$ for all $x, y \in X$.

8.45. THEOREM (Hellinger–Toeplitz). *Let H be a Hilbert space and $T : H \to H$ a self-adjoint linear map. Then T is a bounded operator.*

PROOF. To show continuity via the closed graph theorem we need to verify that if $x_n \to x$ in H and $Tx_n \to y$ in H then $y = Tx$. Set $w = Tx - y$ and observe that

$$\langle w, Tx\rangle = \langle Tw, x\rangle = \lim_{n\to\infty} \langle Tw, x_n\rangle = \lim_{n\to\infty} \langle w, Tx_n\rangle = \langle w, y\rangle$$

i.e. $0 = \langle w, Tx - y\rangle = \|Tx - y\|^2$. Hence $y = Tx$. □

8.7. Further exercises

8.46. EXERCISE. Let X be a complete metric space. Prove that countable unions of meager sets in X are meager and countable intersections of comeager sets in X are comeager.

8.47. EXERCISE. We define the subset $A \subset \mathbb{R}$ as follows: $x \in A$ if and only if there exists $c > 0$ such that

$$|x - j2^{-k}| \geqslant c2^{-k}$$

holds for all $j \in \mathbb{Z}$ and integers $k \geqslant 0$. Show that A is meager and dense.

8.48. EXERCISE. Let (X, d) be a complete metric space without isolated points. Prove that X cannot be countable.

8.49. Exercise. Let $(f_n)_{n=1}^{\infty}$ be a sequence of continuous functions on $[a,b]$, and assume that (f_n) is pointwise bounded, i.e. $\sup_n |f_n(x)| < \infty$ for all $x \in [a,b]$. Show that there is an open interval $I \subset [a,b]$ such that $\sup_{x\in I} \sup_n |f_n(x)| < \infty$, i.e. (f_n) is uniformly bounded on I.

8.50. Exercise.

(i) Show that if X is a normed vector space and $U \subset X$ a proper subspace, then U has empty interior.

(ii) Let X be an infinite dimensional Banach space. Prove that X cannot have a countable Hamel basis.

8.51. Exercise. Let X be the real vector space of polynomials $P : \mathbb{R} \to \mathbb{R}$ with real coefficients. Prove that there exists no norm $\|\cdot\|$ on X such that $(X, \|\cdot\|)$ is a Banach space.

8.52. Exercise. Prove that the space of polynomials is a meager subspace of $C([a,b])$.

8.53. Exercise. Consider $X = C([-1,1])$ (with the usual norm $\|f\|_\infty = \sup_{t\in[-1,1]} |f(t)|$). Let

$$A_+ = \{f \in X \,:\, f(t) = f(-t) \text{ for all } t \in [-1,1]\},$$
$$A_- = \{f \in X \,:\, f(t) = -f(-t) \text{ for all } t \in [-1,1]\}.$$

(i) Show that A_+ and A_- are meager.

(ii) Is $A_+ + A_- = \{f + g \,:\, f \in A_+, g \in A_-\}$ meager?

8.54. Exercise. Let $(B_n)_{n\geqslant 1}$ be an increasing sequence of positive real numbers with $\lim_{n\to\infty} B_n = \infty$ and $B_{n+1} - B_n = o(B_n)$ as $n \to \infty$. Let $f : (0,\infty) \to \mathbb{R}$ be a continuous function with the property that $\lim_{n\to\infty} f(B_n t) = 0$ for every $t > 0$. Show that $\lim_{x\to\infty} f(x) = 0$.

8.55. Exercise. Show that there is a continuous function f on $(0,\infty)$ such that $\lim_{n\to\infty} f(2^n t) = 0$ for all $t > 0$ and such that for every $L \in \mathbb{R}$ there is a sequence ξ_n with $\lim_{n\to\infty} \xi_n = \infty$ and $\lim_{n\to\infty} f(\xi_n) = L$. In particular this shows that one cannot omit the assumption $B_{n+1} - B_n = o(B_n)$ in Exercise 8.54.

8.56. Exercise. A function $f \in C([a,b])$ is called *nowhere monotonic* if f is not monotonic on every open subinterval of $[a,b]$. Show that the nowhere monotonic functions are a comeager subset of $C([a,b])$.

8.57. Exercise. Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth function so that for every $x \in \mathbb{R}$ there exists $n \geqslant 0$ with $f^{(n)}(x) = 0$. Prove that f is a polynomial.

8.58. Exercise. Let $f : [0,1] \to \mathbb{R}$ be differentiable at every $x \in [0,1]$. Prove that the derivative f' is continuous on a dense G_δ set in $[0,1]$.

8.59. Exercise. Let $f \in C(\mathbb{T})$ and let $(S_N)_{N\geqslant 0}$ be the sequence of partial sum operators for the Fourier series of f, defined by $S_N f(x) = \sum_{k=-N}^{N} \widehat{f}(k) e^{2\pi i k x}$.

(i) Prove that for $f \in C(\mathbb{T})$,

$$\lim_{N\to\infty} \frac{\sup_{x\in\mathbb{T}} |S_N f(x)|}{\log N} = 0.$$

(ii) Let $(\gamma_n)_{n\geqslant 2}$ be a nonnegative sequence such that $\lim_{N\to\infty} \frac{\gamma_N}{\log N} = 0$. Show that there exists an $f \in C(\mathbb{T})$ such that $\limsup_{N\to\infty} \frac{|S_N f(0)|}{\gamma_N} = \infty$. Is the collection of $C(\mathbb{T})$-functions with this property comeager? Is it dense?

8.60. Exercise. Let X be a Banach space, Y a normed space. For every $m \in \mathbb{N}$ let $\mathfrak{T}^m = (T_n^m)_{n=1}^{\infty}$ be a sequence of bounded operators $T_n^m : X \to Y$ such that for every $m \in \mathbb{N}$ there exists $x_m \in X$ with $\limsup_{n\to\infty} \|T_n^m x_m\|_Y = \infty$. Show that there exists $x \in X$ such that $\limsup_{n\to\infty} \|T_n^m x\|_Y = \infty$ for every $m \in \mathbb{N}$. This is called the *principle of condensation of singularities.*

8.61. Exercise. Let X, Y, Z be normed spaces. A bilinear map $B : X \times Y \to Z$ is called *separately continuous* if the linear maps $x \mapsto B(x,y)$ are continuous for all $y \in Y$ and the linear maps $y \mapsto B(x,y)$ are continuous for all $x \in X$.

Prove: If at least one of X, Y is a Banach space then every separately continuous bilinear map $B : X \times Y \to Z$ is continuous.

8.62. Exercise (Kuratowski–Ulam). Let (X_1, d_1), (X_2, d_2) be complete metric spaces and assume that X_2 is separable (i.e. contains a countable dense set). Let A be a dense G_δ set in $(X_1 \times X_2, d)$ where $d((x_1, x_2), (y_1, y_2)) = d_1(x_1, y_1) + d_2(x_2, y_2)$. For $x_1 \in X_1$ and $A \subset X_1 \times X_2$ let

$$\pi_{x_1}(A) = \big\{x_2 \in X_2 : (x_1, x_2) \in A\big\}.$$

Then there exists a dense G_δ set E in (X_1, d_1) such that for all $x_1 \in E$ the vertical slices $\pi_{x_1}(A)$ are dense G_δ sets in (X_2, d_2).

8.63. Exercise (Kahane). Let S_N be the partial sum operators for the Fourier series in (8.2). Prove that there exists a dense G_δ set E of functions in $C(\mathbb{T})$ such that for each $f \in E$ we have

$$\limsup_{N\to\infty} |S_N f(x)| = \infty$$

for all x in a dense G_δ set $A_f \subset \mathbb{T}$.

Hints for selected exercises

1.10: Consider the collection $\mathfrak{P}$ of pairs $(Y, \leqslant_Y)$ where Y is a subset of X and $\leqslant_Y$ is a well-order on Y. Define a partial order on $\mathfrak{P}$ by setting $(Y, \leqslant_Y) \preceq (Z, \leqslant_Z)$ if Y, Z are subsets of X with $Y \subset Z$ and if the restriction of $\leqslant_Z$ to $Y \times Y$ is equal to $\leqslant_Y$. Use Zorn's lemma.
1.52: Distinguish the cases $\ell < d$, $\ell = d$, $\ell > d$. (The answer may or may not depend on the coefficients of p, q.)
2.67: Consider $F_n(t) = \frac{1}{n!} t^n (1-t)^n$ for $0 \leqslant t \leqslant 1$ and, under the assumption that $\pi^2 = p/q$, examine the integrals $p^n \int_0^1 F_n(t) \sin(\pi t)\, dt$.
2.93: Consult the proof of Theorem 2.11.
2.98: The trigonometric identity $\cos 2\alpha = 2\cos^2 \alpha - 1$ might be helpful.
1.97: Establish the recursion formula $1 - 2a_{n+1} = (1 - 2a_n)^2$.
2.113: Use induction on n to prove it.
2.115: For $G_n \leqslant A_n$ use Jensen's inequality with the convex function $t \mapsto -\log t$ on $(0, \infty)$.
2.116: Consider $\log(\sqrt[n]{G_n})$.
2.117: If $r_1 < r_2$, apply Jensen's inequality with the convex function $t \mapsto t^{r_2/r_1}$ on $(0, \infty)$.
2.118: Show that the function $f(x) = -x\sqrt{\log(1/x)}$ is convex on $(0, 1)$.
2.120: Prove by induction that $\frac{d^n}{(dx)^n}[f(g(x))]$ is a linear combination of terms of the form

$$f^{(m_1+\cdots+m_n)} \circ g(x) \prod_{j=1}^{n} \big(g^{(j)}(x)\big)^{m_j}$$

where $(m_1, \ldots, m_n)$ are n-tuples of nonnegative integers satisfying $\sum_{j=1}^n j m_j = n$.
2.125: Assume $f'(a) < \eta < f'(b)$; consider $g(x) = f(x) - \eta x$ and show that $g$must have a local extremum in (a, b).
2.128: Use the Cauchy mean value theorem.
2.133: Use Taylor's formula to estimate for $h > 0$, $M_1 \leqslant hM_2 + M_0/h$.
2.136: Examine the derivative of $f(x)e^{-G(x)}$.
2.141: For this, prove using integration by parts

$$\mathcal{J}(\lambda) = \Big(\frac{i}{\lambda}\Big)^N \int_I e^{i\lambda\phi(x)} \mathcal{L}^N[\chi](x)\, dx,$$

where

$$\mathcal{L}[\chi](x) = \frac{d}{dx}\Big(\frac{\chi(x)}{\phi'(x)}\Big) = \frac{\chi'(x)}{\phi'(x)} - \frac{\chi(x)\phi''(x)}{(\phi'(x))^2}$$

and $\mathcal{L}^N[\chi] = \mathcal{L}[\mathcal{L}^{N-1}[\chi]]$. Argue by induction that $\mathcal{L}^N[\chi]$ is a C^∞-function.
3.12: Let $p > 0$. Show that for $m, n \in \mathbb{N}$

$$\int_{m+1}^{n+1} x^{-p} dx \leqslant \sum_{k=m+1}^{n} k^{-p} \leqslant \int_m^n x^{-p} dx.$$

3.14: Use $\log n = \int_1^n x^{-1} dx = \sum_{k=1}^{n-1} \int_k^{k+1} x^{-1} dx$ and $\frac{1}{k+1} \leqslant \int_k^{k+1} x^{-1} dx \leqslant \frac{1}{k}$.

3.22: Examine the partial sum $\sum_{k=0}^{(M+N)n-1} b_k$ for large n. It is useful to use the result of Exercise 3.14.
3.41: Examine $h^{-1}(f(x+h)-f(x))$ for $h=\pm 2^{-k}$ with $k\in\mathbb{N}$.
3.47: We need a bound for $\sum_{k=1}^{n}\cos(kx)$. One can use the trigonometric formula $\sin((k+\frac{1}{2})x)-\sin((k-\frac{1}{2})x)=2\sin(kx)\cos\frac{x}{2}$. Alternatively one may use Euler's identity (see Fact 3.68 below) and complex geometric sums to show this.
3.75: Use Abel summation and consider the Cauchy product of the series $\sum_{n=0}^{\infty} a_n z^n$ and $\sum_{n=0}^{\infty} b_n z^n$.
3.90: Consider $\log\binom{n+\alpha}{n}$ and use (3.21) as in Theorems 3.80, 3.84, as well as Exercise 3.14.
3.94: Distinguish the cases $p<0$ and $0\leqslant p<1$.
3.103: Use Exercises 3.88 and 3.102.
3.104: $3=\sqrt{1+2\cdot 4}=\sqrt{1+2\sqrt{1+3\cdot 5}}$
3.108: Letting $f(t)=\sum_{k=0}^{\infty} B_k\frac{t^k}{k!}$ examine the product $e^{tx}f(t)$.
3.120: Use telescoping sums.
3.132: Consider Kummer's test with $b_n=(n-1)\log(n-1)$ for $n\geqslant 2$.
3.133: Raabe's test can be applied here.
3.146: By Telescoping and Taylor series of logarithm: $1-2\log(2)$.
3.147: Telescopes to $\frac{1}{k\cdot k!}$ after $\frac{1}{n(n+1)\cdots(n+k)}=\frac{1}{k}\Big(\frac{1}{n(n+1)\cdots(n+k-1)}-\frac{1}{(n+1)(n+2)\cdots(n+k)}\Big)$.
3.148: (iii) Recall Leibniz' formula from Example 3.74.
3.152: First compute $\lim_{N\to\infty}\frac{1}{N^{1+q}}\sum_{k=1}^{N}k^q$.
3.153: (iii) If $\sum_{k=0}^{\infty}c_k$ is Cesàro summable, then verify that $\frac{c_n}{n+1}$ tends to 0.
3.155: (i) Write an explicit formula for $s_n-\sigma_n$ in terms of the a_k. (ii) Estimate the difference between $\sum_{k=1}^{n}a_k$ and $\sum_{k=1}^{n}a_k r^k$ for $r=1-\frac{1}{n}$.
3.156: Let $\lambda>1$(but close to 1). Verify the splitting

$$\sum_{k=0}^{n}a_k=\frac{\lfloor\lambda n\rfloor+1}{\lfloor\lambda n\rfloor-n}\sum_{k=0}^{\lfloor\lambda n\rfloor}\Big(1-\frac{k}{\lfloor\lambda n\rfloor+1}\Big)a_k$$
$$-\frac{n+1}{\lfloor\lambda n\rfloor-n}\sum_{k=0}^{n}\Big(1-\frac{k}{n+1}\Big)a_k-\frac{\lfloor\lambda n\rfloor+1}{\lfloor\lambda n\rfloor-n}\sum_{n<k\leqslant\lfloor\lambda n\rfloor}\Big(1-\frac{k}{\lfloor\lambda n\rfloor+1}\Big)a_k.$$

3.159: Use Taylor's formula with remainder R_n(Theorem 2.49) and prove the inequality $0\leqslant R_n(0,x)\leqslant(x/\rho)^{n+1}f(\rho)$.
3.163: Estimate $s_n(x)$ for $n\leqslant\frac{1}{x}$ and for $n>\frac{1}{x}$ separately.
3.168: For (ii), use the asymptotics in (i). Do Exercise 3.48 first.
4.67: (i) Use the contraction principle in the space $C([-1,1])$.
4.141: It might help to try this first for the sequence $1/n$.
4.144: It helps to work with $\delta=2^{-L}$, where $L\in\mathbb{N}$.
4.146: It may help to answer his first with δ of the form 2^{-n}.
4.159: Use Exercise 4.155.
4.171: Use Taylor's formula, expanding $0=f(\beta)$ about the point x.
4.179: Use the Arzelà–Ascoli theorem.
4.183: Assume that X is not sequentially compact and construct an unbounded continuous function on X.
5.102: The bound for symmetric matrices (see Example 5.18) can be applied to $A^{\intercal}A$.
5.103: (iv) First do Exercise 5.102.
5.112: Mimic the proof of Arzelà–Ascoli.
5.121: Use the parallelogram identity.

5.122: Use a result from the Hahn–Banach theory.
6.27: Consider overlapping open balls along a continuous path connecting two given points in U.
6.91: Consider the minimal eigenvalue of $D^2 f|_x$ as a function of x.
6.116: Recall from linear algebra that $A\,\mathrm{adj}(A) = (\det A)I$.
6.124: Consider the function $t \mapsto g(t) = f(tx) - t^\alpha f(x)$.
6.138: Show first that if $x_0 < x_1 < \cdots < x_N$ and $x_i - x_{i-1} = h$ then

$$\sum_{i=1}^{N-1} |f(x_{i+1}) - 2f(x_i) + f(x_{i-1})| \leqslant C'|h|.$$

6.161: (i) The left hand side is zero if $t = t_0$. Use Taylor expansion to study how the error changes as t increases from t_k to t_{k+1}.
7.7: Examine the proof of the Weierstrass approximation theorem via Bernstein polynomials.
7.60: (Solution sketch) By a change of variable it suffices to consider the interval $[a, b] = [-1/4, 1/4]$. Extend the function f defined on $[-1/4, 1/4]$ to a continuous function F on $[-1/2, 1/2]$ so that $F(x) = f(x)$ on $[-1/4, 1/4]$ and $F(-1/2) = F(1/2)$. We can extend F to a continuous 1-periodic function on $\mathbb{R}$. Let $\varepsilon > 0$. By Fejér's theorem we can find a trigonometric polynomial

$$p(x) = a_0 + \sum_{k=1}^{N} [a_k \cos(2\pi kx) + b_k \sin(2\pi kx)]$$

so that $\max_{x\in\mathbb{R}} |F(x) - p(x)| < \varepsilon/2$. Now the Taylor series for cos and sin converge uniformly on every compact interval. Thus we can find an algebraic polynomial P so that

$$\max_{x\in[-1/2,1/2]} |p(x) - P(x)| < \varepsilon/2.$$

Combine the two estimates (and use that $f = F$ on $[-1/4, 1/4]$) to obtain

$$\max_{|x|\leqslant 1/4} |f(x) - P(x)| \leqslant \max_{|x|\leqslant 1/2} |F(x) - P(x)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

7.78: Let $\varepsilon > 0$ be small. For the upper bound split the integral

$$\|D_N\|_1 = 2\int_0^{1/2} \frac{|\sin((2N+1)\pi t)|}{|\sin(\pi t)|} dt$$

into an integral over $[0, \varepsilon_0]$ and an integral over $[\varepsilon_0, \frac{1}{2}]$. The lower bound $\sin \pi t \geqslant 2t$, valid in $[0, \frac{1}{2}]$, improves on $[\varepsilon_0, \frac{1}{2}]$.
7.146: Show that $\int_0^1 |f(x)|^2 dx = 0$.
7.151: Take a close look at the proof of the Stone–Weierstrass theorem.
7.155: Use the Riemann–Lebesgue lemma.
7.156: Let $f = 0$ and build f_n so that $\|f_n\|_\infty = 1$ and the size of the support of f_n shrinks to zero as $n \to \infty$(to guarantee L^2 convergence), but its location slides around in $[0, 1]$ to avoid pointwise convergence.
7.160: What can you say about the Fourier coefficients of $f^{(k)}$?
7.161: Let $f(x) = x$ for $|x| \leqslant 1/2$ and evaluate its Fourier series at $x = 1/(2\pi)$.
7.165: (i) Consider first the case where f is a 1-periodic trigonometric polynomial.
7.166: (i) Use the Riemann-Lebesgue lemma. (ii) The answer is yes and the limit equals $\frac{2}{\pi}\int_0^1 f(x)dx$(prove it!).
7.170: Use induction and the mean value theorem.

7.175:

$$\sum_{k=1}^{\infty} k^{-1}\cos(2\pi kx) = \sum_{k=1}^{N} k^{-1} + \sum_{k=1}^{N} k^{-1}(\cos(2\pi kx)-1) + \sum_{k=N+1}^{\infty} \cos(2\pi kx)$$

and show that the second and third terms are $O(1)$. The first term is $\approx \log N$.

7.177: Use the same argument as in the proof of the inverse approximation theorem (Theorem 7.69) for trigonometric functions, but use Markov's inequality (Theorem 7.138) in place of Bernstein's inequality.

7.180: Show the recursion formula

$$\beta_0 = y_0, \quad \beta_k = \frac{y_k - \beta_0 - \sum_{1\leqslant j\leqslant k-1} \beta_j \prod_{\nu=0}^{j-1}(x_k - x_\nu)}{\prod_{\nu=0}^{k-1}(x_k - x_\nu)}\,.$$

7.183: Consider $P \in \mathcal{P}_{2N+1}$. For the formula for β_k write $P = \Omega R_1 + R_2$ with $R_1, R_2 \in \mathcal{P}_N$. For the second formula consider the Gaussian interpolation formula for the nonnegative polynomials $u_k(t) = \prod_{j\neq k}(t-t_j)^2$.

7.184: The argument is similar to the one used for the Erdős–Turán theorem (Theorem 7.132).

7.186: Use the binomial expansion of $(\cos\vartheta + i\sin\vartheta)^n$.

7.188: From Exercise 7.187 one gets by successive differentiation

$$(1-t^2)T_n^{(m+2)}(t) - (2m+1)tT_n^{(m+1)}(t) + (n^2-m^2)T_n^{(m)}(t) = 0.$$

and therefore a two term recursion formula for $T_n^{(k)}(1)$. Deduce $(2k+1)T_n^{(k+1)}(1) = (n^2-k^2)T_n^{(k)}(1)$.

7.192: (i) $(-1)^n \frac{(2n)!}{2^n(n!)^2}$. (ii), (iii): Repeatedly integrate by parts.

7.195: Examine the proof of Theorem 7.119.

8.50: For (ii), use (i) and the Baire category theorem.

8.51: Use Exercise 8.50.

8.57: Let $\mathcal{I}$ be the set of open intervals I so that $f|_I$ is a polynomial. Use that the complete space $\mathbb{R}\setminus \cup\, \mathcal{I}$ is not meager, writing it as a union of sets $A_n = \{x : f^{(n)}(x) = 0\}$. Be careful with relative topologies.

8.62: Write $A = \bigcap_n \Omega_n$ where Ω_n are open and dense. Let $(B_\nu)_{\nu=1}^{\infty}$ be the collection of balls in X_2 with rational radii which are centered in a countable dense set. Show that

$$E_{n,\nu} = \{x_1 \in X_1 : \pi_{x_1}(\Omega_n) \cap B_\nu \neq \varnothing\}$$

is open and dense in X_1, and $E = \bigcap_{n,\nu} E_{n,\nu}$ is a dense G_δ set. For every $x_1 \in E$ the slices $\pi_{x_1}(\Omega_n)$ are open and dense in X_2. Conclude that for $X_1 \in E$ the set $\pi_{x_1}(A)$ is a dense G_δ set in X_2.

8.63: Consider $\Omega_{M,N} = \bigcup_{n\geqslant N}\{(f,x) \in C(\mathbb{T})\times\mathbb{T} : |S_N f(x)| > M\}$ and $A = \bigcap_{M,N}\Omega_{M,N}$. Show that $A = \{(f,x) \in C(\mathbb{T})\times\mathbb{T} : \limsup_{n\to\infty} |S_n f(x)| = \infty\}$. Show that $\Omega_{M,N}$ is open and dense in $C(\mathbb{T})\times\mathbb{T}$, and hence A is a dense G_δ set in $C(\mathbb{T})\times\mathbb{T}$. Then apply the Kuratowski–Ulam theorem in Exercise 8.62.

Index

L^2 (semi)norm, 286
L^2 convergence, 296
L^2 norm, 197
L^p norm, 212
$\omega_f(\delta)$ (modulus of continuity), 302, 347

Abel summable, 115
Abel summation, 106
Abel's theorem, 115
Abelian group, 37
absolute convergence (normed space), 182
absolute convergence (products), 118
absolute convergence (series), 92
absolute value, 48, 76
accumulation point, 145
adjoint operator, 201
algebra (of functions), 282
algebraic completeness, 162
algebraically complete, 21
almost convergent (Lorentz), 208
alternant, 316
alternating series test, 107
AM-GM inequality, 192
analytic function, 112
annihilator, 206
anti-symmetric, 10
antiderivative, 63
approximate unit, 299
Archimedean property, 18
arcsine function, 68
arithmetic mean, 83
arithmetic mean (function), 83
arithmetic progression, 83
Arzelà–Ascoli theorem, 164
ascending factorial, 131
Assouad dimension, 179
average, 356
axiom of choice, 12

Baire category theorem, 360
Baire-1 function, 363
Baire-n classes, 364
Banach fixed point theorem, 151
Banach limit functional, 207
Banach limit set, 208
Banach space, 181
Banach–Steinhaus theorem, 188, 364
barycentric interpolation formula, 352
Bernoulli formulas, 128
Bernoulli functions ψ_k, 126
Bernoulli numbers B_n, 125
Bernoulli polynomials $b_n(x)$, 125
Bernoulli's inequality, 26
Bernstein polynomial, 280
Bernstein's inequality (polynomials), 340
Bernstein's inequality (trigonometric polynomials), 305, 339
Bernstein-type (inverse) approximation theorem, 305
Bessel's inequality, 291
best approximation, 279, 303
Big-O (sequence), 30
bijective, 11
binary sequence, 17
binomial coefficient, 19, 120
binomial formula, 19
Bolzano–Weierstrass property, 157
Bolzano–Weierstrass theorem, 158
Borel's lemma, 153
Borel's theorem, 111
bounded (metric space), 151
bounded function, 32, 80, 151
bounded linear functional, 185
bounded linear map, 183
bounded projection, 328
bounded range, 151
bounded sequence, 26
box counting dimension, 170

Cantor diagonalization sequence, 159
Cantor set, 171

Cantor's power set theorem, 16
Cantor–Schröder–Bernstein theorem, 14
cardinality, 20
Cartesian product, 8, 11
Cauchy condensation test, 94
Cauchy criterion, 29
Cauchy criterion (infinite product), 117
Cauchy mean value theorem, 61
Cauchy product, 98
Cauchy sequence, 28, 146
Cauchy–Hadamard formula, 108
Cauchy–Schwarz inequality, 198
Cauchy-Riemann differential equations, 271
Cesàro sum, 137
chain, 12
chain rule, 59, 218
characterizations of compactness (metric spaces), 158
Chebyshev differential equation, 353
Chebyshev generating function, 354
Chebyshev interpolation, 325
Chebyshev polynomials, 318
Chebyshev polynomials (second kind), 354
Chebyshev's alternant theorem, 316
choice function, 12
closed ball $\overline{B}(x_0, r)$, 144
closed graph theorem, 368
closed interval, 21
closed set, 21, 144
closure, 144
closure of a set, 144
codomain, 11
comeager (set), 360
commutative group, 37
compact (metric space), 153
comparison of ratio and root test, 134
comparison of series and integrals, 94
Comparison test (nested radicals), 122
compatibility conditions (vector field), 238
complement (set), 8
complete orthonormal system, 292
completeness (metric space), 146
completion (metric space), 174
complex analytic function, 114
complex conjugate, 75
complex differentiable, 114
complex number, 75
condensation of singularities, 370
conditionally convergent (series), 97
conformal (linear transformation), 270
connected component, 149
connected set, 148
content zero set, 166
continued square-root, 122
continuity, 31
continuous, 147
continuously differentiable, 65, 222
contraction, 151
contraction mapping theorem, 151
convergence ($x_n \to x$), 146
convergence (in metric spaces), 146
convergence (nested radical), 121
convergence (sequence), 22
convergence (series), 91
convergence of series (normed space), 182
convex, 203
convex (function), 249
convex body, 203
convex combination, 203
convex function, 84
convolution, 297
coordinate functions, 220
coprime, 9
cosine, 69
cosine polynomials, 285
countable, 16
covering number $\mathcal{N}(E, \delta)$, 170
critical point, 245
cut, 40

Darboux's theorem, 86
de la Vallée-Poussin means, 349
de Morgan's law, 8
decreasing sequence, 25
Dedekind complete (ordered field), 18
Dedekind cut, 40
degenerate interval, 21
degree (of trigonometric polynomial), 295
degree (polynomial), 25
dense, 172
dense (subset), 279
derivative, 58, 215
descending factorial, 131
diameter (metric space), 151
diffeomorphism (C^n), 277
difference (sets), 8
differentiable, 58, 215
differential, 216
Dini's uniform convergence theorem, 156

direct (Jackson-type) approximation theorem, 305
directional derivative, 219
Dirichlet function, 53, 363
Dirichlet kernel, 298
discrete orthogonality (Chebyshev polynomials), 354
divergence (series), 91
domain, 11
doubly indexed series, 100
doubly infinite series, 92
dual space X', 185
dyadic interval, 355

empty set, 8
endpoint (interval), 21
entire function, 114
equicontinuous, 163
equisummable, 212
equivalence class, 10
equivalence relation, 10
equivalent norms, 186
Euclidean metric, 143
Euclidean space, 196
Euler's formula, 76
Euler's identity, 115
Euler's number e, 25, 67
Euler's theorem (irrationality of e), 68
Euler–Lagrange equation, 264
Euler–Mascheroni constant, 94
evaluation map, 213
ex falso quodlibet, 8
exponential function, 67, 114
extended real line, 21
extremal property, Chebyshev polynomials, 320

F-sigma set (F_σ), 362
Féjer means, 296
Faà di Bruno's formula, 85
Faber's theorem, 329, 330
factorial, 19
family, 11
Faulhaber formulas, 128
Fejér means, 298
Fejér kernel, 298
Fibonacci numbers, 138
field, 36
finite intersection property, 154
fixed point, 151
floor function ($\lfloor x \rfloor$), 146
Fourier coefficient $\widehat{f}(n)$, 295
Fourier series, 296
Fourier transform on ℓ^1, 366
Fréchet derivative, 215
Fréchet differentiable, 215
Fréchet's theorem, 212
Fresnel integrals, 88
Fubini's theorem (series), 102
function, 10
function of class C^∞, 65
function of class C^k, 65
fundamental theorem of algebra, 162
fundamental theorem of calculus, 63

G-delta set (G_δ), 362
Gâteaux differentiable, 219
Gamma function, 74
Gauss' test, 135
Gaussian error function, 74
Gaussian integral, 73
Gaussian nodes, 337
Gaussian quadrature formula, 353
generalized geometric mean, 193
generating function, 354
generic set (Baire), 360
geometric mean, 83
geometric mean (function), 84
geometric series, 91
gliding hump method, 188
global minimum, 248
global minimum/maximum, 248
Golay–Rudin–Shapiro sequence, 314
gradient $\nabla f(a)$, 220
Gradient descent, 248
Gram determinant, 288
Gram-Schmidt orthogonalization, 286

Hölder class, 307
Hölder continuous ($C^{0,s}([a,b])$, 182
Hölder dual exponent, 192
Hölder's inequality, 192
Haar system, 355
Haar wavelet, 355
Hahn-Banach extension theorem, 204
half-open interval, 21
Hamel basis, 366
harmonic mean (function), 84
harmonic series, 94
Heine's theorem, 154
Heine–Borel property, 153

Heine–Borel theorem, 158, 187
Hermite interpolation, 323
Hermite interpolation polynomial, 323
Hermite–Fejér interpolation polynomial, 333
Hermitian matrix, 213
Heron's method, 177
Herschfeld's theorem, 123
Hessian, 235
Hessian matrix, 235
higher order differential, 274
Hilbert basis, 293
Hilbert space, 200
Hilbert–Schmidt norm, 210
homeomorphic, 156
homeomorphically equivalent, 156
homeomorphism, 156
homogeneous (function), 265
Horner's scheme, 106
hypergeometric function, 140
hypergeometric series, 140

identity map, 11
image, 11
imaginary numbers, 75
imaginary part, 75
imaginary unit, 75
implicit differentiation, 233
implicit function theorem, 234, 241
increasing function, 33
indefinite (matrix), 245
indexed family, 11
induction, 8
infimum, 18
initial value problem, 254
injective, 11
inner product $\langle \cdot, \cdot \rangle$, 196
inner product space, 196
integer, 17
integration by parts, 64, 106
integration by substitution, 64
interchanging order of summation, 100
interior point, 21, 144
intermediate value theorem, 33, 149
intersection, 11
intersection (sets), 8
interval, 21
inverse (Bernstein-type) approximation theorem, 305
inverse element, 37
inverse function theorem, 227
inverse function theorem, C^p version, 240
inverse image ($f^{-1}(A)$), 147
irrationality of π (Lambert's theorem), 70
irrationality of $\sqrt{2}$, 17
irrationality of e (Euler's theorem), 68
isometry (metric spaces), 148
isomorphic (normed spaces), 226
isomorphism (normed spaces), 225
isomorphism (ordered fields), 39
isomorphism of ordered fields, 18

Jackson kernel, 305
Jackson-type (direct) approximation theorem, 305
Jacobian matrix, 220
Jensen's inequality, 84

Khinchine's inequalities, 357
Kummer's test, 135

l'Hôpital's rule, 86
Lagrange factors, 321
Lagrange interpolation polynomial, 324
Lagrange multipliers, 247
Lagrange polynomials, 321
Lambert's theorem (irrationality of π), 70
Landau polynomials, 282
Landau's inequality, 87
Laurent polynomial, 285
least upper bound axiom, 18
Lebesgue constants (Chebyshev interpolation), 331
Lebesgue constants (partial sum operators), 309
Lebesgue constants (polynomial interpolation), 327
Lebesgue null set, 166
left inverse, 11
left-continuous, 31
Legendre polynomials, 354
Leibniz rule, 85
Leibniz test, 107
Leibniz' formula, 105, 116
length (polygon), 62
limit (sequence), 22, 146
limit inferior, 27
limit superior, 27
limiting method, 190
line segment, 203
linear map, 183
linear operator, 183

linear regression, 248
linearization, 216
Lipschitz class, 307
Lipschitz condition, 231
Lipschitz condition (Df), 229
Lipschitz continuous, 182, 256
Little-o (sequence), 30
local inverse, 227
local maximum/minimum/extremum, 245
local maximum/minimum/extremum point, 245
local minimum point, 61
locally compact, 188
logarithm (log, ln), 67
logarithm with base a, $\log_a$, 67
lower bound, 12
lower Minkowski dimension $\underline{\dim}_{\mathrm{M}} E$, 170
lower Riemann–Darboux integral, 52
lower sum (Riemann), 51

Müntz polynomials, 342
Machin's formula, 77
map, 10
Markov's inequality, 341
meager (set), 360
mean value theorem, 61, 221
mean value theorems (integrals), 56
Mean-square convergence, 296
metric, 143
metric entropy function, 170
metric space, 143
Minkowski dimension, 170
Minkowski functional, 203
Minkowski's inequality, 193
modulo, 10
modulus of continuity, 302, 347
modulus of continuity (higher order), 350
monic polynomials, 319
monotone function, 33
monotone sequence, 25
Morse lemma, 242
multiindex notation, 243
multilinear, 272

n-tuple, 11
natural logarithm (log, ln), 67
natural number, 8
natural numbers $\mathbb{N}$ (not including 0), 9
natural numbers $\mathbb{N}_0$ (including 0), 9
negative (semi)definite (matrix), 245
nested radical, 121
nested square-root, 122
neural network, 347
neutral element, 37
Newton interpolation formula, 352
Newton iteration, 230
Newton iteration (modified), 230
Newton's method, 177
Newton–Kantorovich iteration, 232
node polynomial, 321
non-increasing sequence, 25
nondegenerate critical point, 243
nonnegative integers $\mathbb{N}_0$, 9
norm, 181
normed vector space, 181
nowhere dense (set), 359
nowhere differentiable function, 103

one-sided derivative, 58
one-sided limit, 30
one-to-one, 11
onto (function), 11
open ball $B(x_0, r)$, 144
open interval, 21
open mapping theorem, 367
open set, 21, 144
ordered field, 17, 38
ordered pair, 8
ordinary differential equation, 254
orthogonal, 198
orthonormal basis, 286
orthonormal basis (Hilbert space), 293
orthonormal system, 290
oscillation at a point, osc_f, 150

Pólya's criterion (nested radicals), 124
parallelogram identity, 199
Parseval's identity, 292
partial derivative, 219
partial order, 10
partial sum operator (Fourier series), 295
partial sums, 91
partition (of an interval), 51
Pascal's triangle, 20
path, 149
path connected, 149
path equivalent, 149
Peano existence theorem, 259
permutation, 131
pi (π), 68

Picard iteration, 257
Picard–Lindelöf theorem, 255
piecewise continuous, 81, 197
Pochhammer function, 131
Pochhammer symbols, 131
pointwise bounded, 163
pointwise convergence, 79
polar coordinates, 76
polarization identities, 199
polynomial, 25
positive (semi)definite (matrix), 245
positive definite (inner product), 196
positive half-line, 21
power series, 107
power set, 8
pre-Hilbert space, 196
preimage, 11
prime number, 9
primitive, 63
product (sets), 8
product metric, 144
product rule, 59, 218
product rules, 273
proof, 7
Pythagorean theorem, 115, 199

quadrature formulas, 352
quasi-Assouad dimension, 179

Raabe's test, 95
Rademacher functions, 290
radius of convergence, 108
range, 11
ratio test, 93
rational cuts, 46
rational numbers, 17
real analytic function, 112
real numbers, 40
real part, 75
rearrangement (series), 96
rectifiable, 62
refinement (partition), 51
reflexive, 10, 213
regula falsi, 176
regular infinite matrix, 190
relation, 10
relative topology, 145
relatively compact, 153
residual (set), 360
restriction of a function to a set, 145
Richardson extrapolation, 130
Riemann ζ-function, 94
Riemann integrable, 52
Riemann integrable (vector-valued function), 214
Riemann integral (vector valued), 214
Riemann sum, 56
Riemann–Darboux integrals, 52
Riemann–Lebesgue lemma, 292
Riesz representation theorem (Hilbert spaces), 200
right inverse, 11
right-continuous, 31
Rogosinski regularization (partial sum operator), 349
Rogosinski singular integrals, 349
Rolle's theorem, 61
root test, 93
Rudin–Shapiro polynomials, 313

scalar product, 196
Schröder-Bernstein theorem, 14
Schur's inequality (polynomials), 340
Schur's test, 184
Schwarz second derivative, 86
Schwarz' theorem, 237
second mean value theorem for integrals, 87
self-adjoint, 369
selfadjoint, 213
seminorm, 181
separable, 294
separately continuous (bilinear map), 370
sequence, 11
sequence space c_0, 144
sequential ℓ^p space, 192
sequential continuity, 31, 148
sequentially compact, 157
series, 91
sesquilinear property, 196
set of first category (Baire), 360
set of second category (Baire), 360
shift invariant (linear functional), 207
shift operator, 207
shifted factorial, 131
simply connected, 239
sine, 69
sine polynomial, 348
singleton set, 21
smooth bump function, 111
smooth function, 65

space of k-times continuously differentiable functions $C^k([a,b])$, 178
space of bounded sequences ℓ^∞, 143
space of continuous functions $C([a,b])$, 143
space of linear operators $L(X,Y)$, 183
squeeze theorem, 22
step function, 52
Stone–Weierstrass theorem, 282, 284
strict global minimum/maximum, 248
strict local maximum/minimum, 245
strictly convex (function), 249
strictly decreasing sequence, 25
strictly increasing function, 33
strictly monotone function, 33
strongly convex (function), 252
subfield, 39
sublinear functional, 202
subsequence, 26
subsequential limits, 27
subset, 8
subset notation $A \subset B$, 8
subset notation $A \subsetneq B$ (proper), 8
successive approximation, 152, 257
successor, 8
summability method, 115
summation by parts, 106
superorthogonality property, 356
supremum, 18
supremum metric, 143, 151
supremum norm, 181
surjective, 11
symmetric, 10
symmetric convex body, 203

tagged partition, 214
Tauberian theorem, 138
Taylor series, 109
Taylor's theorem (one variable), 65
Taylor's theorem (several variables), 241
tertium non datur, 7
theorem of the inverse operator, 365
Thomae's function, 32, 150, 362
Tonelli–Fubini theorem (series), 102
topology, 144
torus group, $\mathbb{T}$, 176
total derivative, 220
total order, 10
totally bounded, 157
totally ordered, 10
trace (of a matrix), 263
transitive, 10
trapezoid rule, 128
triangle inequality (complex numbers), 76
triangle inequality (series), 92
triangle inequality., 143
trigonometric functions, 115
trigonometric polynomial, 285
trigonometric polynomials, 295
trivial metric, 145

unbounded interval, 21
unconditional convergence, 96
unconditional convergence (normed spaces), 182
uncountable, 16
uniform boundedness principle, 188, 364
uniform convergence, 78
uniform convergence (metric space valued), 150
uniform convergence (sequence), 78
uniform convergence (series), 102
uniformly bounded, 163
uniformly Cauchy, 79
uniformly continuous, 154
union, 11
union (sets), 8
unit circle, S^1, 76
unit interval, 21
universal mapping property (completions), 175
upper Assouad spectrum, 179
upper bound, 12
upper box counting dimension, 170
upper Minkowski dimension $\overline{\dim}_{\mathrm{M}} E$, 170
upper Riemann–Darboux integral, 52
upper sum (Riemann), 51

van der Waerden function, 105
Vandermonde determinant, 322
Viète's formula, 125
Volterra integral equation, 176

Wallis product, 119
Weierstrass test, 102
Weierstrass' approximation theorem, 280
Weierstrass' function, 103
well-order, 14
Wirtinger derivatives, 271

Young's inequality, 193

Zeta-function (ζ), 94
Zorn's lemma, 13
Zygmund class, 351

Bibliography

[1] Fernando Albiac, Nigel J. Kalton. Topics in Banach space theory. Grad. Texts in Math., 233 Springer, 2016.

[2] George E. Andrews; Richard Askey, Ranjan Roy. Special functions. Encyclopedia Math. Appl., 71 Cambridge University Press, 1999.

[3] Tom Apostol. Calculus. Vol. 1: One-Variable Calculus, with an Introduction to Linear Algebra. Vol. II: Multi-Variable Calculus and Linear Algebra with Applications to Differential Equations and Probability. 2nd ed. John Wiley & Sons, 1991.

[4] Tom M. Apostol. Mathematical analysis. Second edition Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1974.

[5] Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. Fund. Math. 3 (1922) 133–181; also in Oevres avec des Commentaires, vol. 2, Editions scientifiques de Pologne, Warsaw, 1979.

[6] Patrick Billingsley. Notes: Van der Waerden's continuous nowhere differentiable function. Amer. Math. Monthly 89 (1982), no.9, 691.

[7] Nicolas Bourbaki. *General Topology. Chapters 1–4*. Elements of Mathematics. Springer-Verlag, Berlin, 1995.

[8] Peter Borwein, Tamás Erdélyi. Polynomials and polynomial inequalities. Grad. Texts in Math., 161 Springer-Verlag, New York, 1995.

[9] Andrew Bruckner, Judith Bruckner, Brian Thomson. Can one visualize a continuous nowhere differentiable function? Amer. Math. Monthly 130 (2023), no. 3, 214–221.

[10] Lennart Carleson, On convergence and growth of partial sums of Fourier series. Acta Math. 116 (1966), 135–157.

[11] E.W. Cheney. Introduction to approximation theory. McGraw-Hill Book Co., New York-Toronto-London, 1966.

[12] J.A. Clarkson, P. Erdős. Approximation by polynomials. Duke Math. J.10 (1943), 5–11.

[13] Paul J. Cohen. On a Conjecture of Littlewood and Idempotent Measures. Amer. J. Math. 82, no. 2, pp. 191-212, 1960.

[14] George Cybenko. Approximation by superpositions of a sigmoidal function. Math. Control Signals Systems 2 (1989), no.4, 303–314.

[15] Ronald A. DeVore, George G. Lorentz. Constructive approximation. Grundlehren Math. Wiss., 303. Springer-Verlag, Berlin, 1993.

[16] Jean Dieudonné. Foundations of modern analysis. Pure Appl. Math., Vol. X. Academic Press, New York-London, 1960.

[17] V. K. Dzyadyk, S. Yu. Dzyadyk and A. S. Prypik. Asymptotic behavior of Lebesgue constants in trigonometric interpolation, Ukrainian Mathematical 33 (1981), 553–559.

[18] Paul Erdős, Paul Turán. On interpolation. I. Quadrature- and mean-convergence in the Lagrange-interpolation. Ann. of Math., II. Ser. 38, 142-155 (1937).

[19] P. Erdős, P. Vertesi. On the almost everywhere divergence of Lagrange interpolatory polynomials for arbitrary systems of nodes. Acta Mat. Acad. sci. Hungar. 36 (1980) 71–89, 38 (1981) 263.

[20] Georg Faber, Über die interpolatorische Darstellung stetiger Funktionen, Jahresber. Deutsch. Math.-Verein., 23 (1914), 192–210.

[21] Kenneth Falconer. Fractal geometry. Mathematical foundations and applications. Third edition John Wiley & Sons, Ltd., Chichester, 2014.

[22] Gerald B. Folland. Real Analysis: Modern techniques and their applications. Second edition. Pure Appl. Math. (N. Y.) Wiley-Intersci. Publ. John Wiley & Sons, Inc., New York, 1999.

[23] Jonathan M. Fraser. Assouad dimension and fractal geometry. Cambridge Tracts in Math., 222 Cambridge University Press, Cambridge, 2021.

[24] Philip T. Gressman, Lillian B. Pierce, Joris Roos, and Po-Lam Yung. A new type of superorthogonality. Proc. Amer. Math. Soc. 152 (2024), no. 2, 665–675.

[25] Karl-Goswin Grosse-Erdmann. Universal families and hypercyclic operators. Bull. Amer. Math. Soc. (N.S.) 36 (1999), no. 3, 345–381.

[26] Géza Grünwald. Über Divergenzerscheinungen der Lagrangeschen Interpolationspolynome stetiger Funktionen. Annals of Math. 37, No. 4 (1936), 908–918.
[27] G.H. Hardy and E.M. Wright. An introduction to the theory of numbers. Oxford University Press, Oxford, 2008.
[28] G. H. Hardy and J. E. Littlewood. A new proof of a theorem on rearrangements. Journal of the London Mathematical Society, vol. 23 (1948), pp. 163–168.
[29] Paul R. Halmos. Naive set theory. The University Series in Undergraduate Mathematics D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto-London-New York, 1960.
[30] Hans Hahn. Über Folgen linearer Operationen. Monatsh. Math. Phys. 32 (1922), 3–88.
[31] A. Herschfeld. On infinite radicals. Amer. Math. Monthly 42, no. 7, 419–429, 1935.
[32] Lars Hörmander. The analysis of linear partial differential operators, I. Distribution theory and Fourier analysis. Second edition, Springer Study Ed. Springer-Verlag, Berlin, 1990.
[33] Jean-Pierre Kahane, Baire's category theorem and trigonometric series. J. Anal. Math. 80 (2000), 143–182.
[34] Jean-Pierre Kahane, Yitzhak Katznelson. Sur les ensembles de divergence des séries trigonométriques, Studia Math. 26 (1966), 305–306.
[35] Jean-Pierre Kahane, Hervé Queffélec. Ordre, convergence et sommabilité de produits de séries de Dirichlet. Ann. Inst. Fourier (Grenoble) 47 (1997), no. 2, 485–529.
[36] Sung Soo Kim. A Characterization of the Set of Points of Continuity of a Real Function. Amer. Math. Monthly 106 (1999), no.3, 258–259.
[37] Konrad Knopp. Theorie and Anwendung der unendlichen Reihen (German) Fünfte berichtigte Auflage. Die Grundlehren der mathematischen Wissenschaften, Band 2 Springer-Verlag, Berlin-New York, 1964.
[38] Konrad Knopp. Infinite sequences and series. Translated by Frederick Bagemihl. Dover Publications, Inc., New York, 1956.
[39] A.N. Kolmogorov, S.V. Fomīn. Introductory real analysis. Translated from the second Russian edition and edited by Richard A. Silverman. Corrected reprinting. Dover Publications, Inc., New York, 1975.
[40] S. V. Konyagin, On a problem of Littlewood. Math. USSR-Izv. 18 (1982), no. 2, 205–225.
[41] Thomas W. Körner. Fourier analysis. Cambridge University Press, Cambridge, 1988.
[42] Thomas W. Körner. A Companion to Analysis: A Second First and First Second Course in Analysis. Grad. Stud. Math., 62 American Mathematical Society, Providence, RI, 2004.
[43] Thomas W. Körner. Besicovitch via Baire. Studia Math.158 (2003), no.1, 65–78.
[44] Rainer Kress, Numerical analysis. Grad. Texts in Math., 181 Springer-Verlag, New York, 1998.
[45] Michael Lacey and Christoph Thiele. A proof of boundedness of the Carleson operator. Math. Res. Lett. 7 (2000), no. 4, 361–370.
[46] Edmund Landau. Foundations of Analysis. The arithmetic of whole, rational, irrational and complex numbers. Translation by F. Steinhardt of the German original *Grundlagen der Analysis*. Chelsea 1960.
[47] Jonathan W. Lewin. A simple proof of Zorn's lemma. American Math. Monthly 98, 4, (1991), 353-354.
[48] O. Carruth McGehee, Louis Pigno, and Brent Smith, Hardy's Inequality and the L^1 Norm of Exponential Sums. Ann. Math. 113 (1981), no. 3, 613–618.
[49] Józéf Marcinkiewicz. Collected papers. Edited by Antoni Zygmund. With the collaboration of Stanislaw Lojasiewicz, Julian Musielak, Kazimierz Urbanik and Antoni Wiweger. Instytut Matematyczny Polskiej Akademii Nauk Państwowe Wydawnictwo Naukowe, Warsaw, 1964.
[50] Mar]Mar
[51] Józéf Marcinkiewicz. Sur la divergence des polynomes d'interpolation. Acta Sci. Math. (Szeged) 8:2-3(1936-37), 131-135. See also Collected papers 1964.
[52] Elliott Mendelson. Number systems and Foundations of Analysis. Dover Publications, 2008. Republication of the 1973 Academic Press edition.
[53] C. H. Müntz. Über den Approximationssatz von Weierstraß. H.A. Schwarz Festschrift, 303–312, Berlin, 1914.
[54] I.P.Natanson. Constructive function theory. Vol. I. Uniform approximation. Translated from the Russian by Alexis N. Obolensky. Frederick Ungar Publishing Co., New York, 1964.
[55] I.P. Natanson. Constructive function theory. Vol. II. Approximation in mean. Translated from the Russian by John R. Schulenberger. Frederick Ungar Publishing Co., New York, 1965.
[56] I.P. Natanson, Constructive function theory. Vol. III. Interpolation and approximation quadratures. Translated from the Russian by John R. Schulenberger. Frederick Ungar Publishing Co., New York, 1965.
[57] Niven, Ivan, A simple proof that π is irrational, Bull. Amer. Math. Soc., 153 (1947), 509.
[58] Niven, Ivan. A proof of divergence of $\sum \frac{1}{p}$. Amer. Math. Monthly 78 (3) (1971), 272–273.
[59] George Pólya and Gabor Szegö. Aufgaben und Lehrsätze aus der Analysis, I. 1924. English Edition (Translation by Dorothee Aeppli): Problems and Theorems in Analysis I, Springer, 1972.
[60] George Pólya and Gabor Szegö. Aufgaben und Lehrsätze aus der Analysis, II. English Translation by Dorothee Aeppli: Problems and Theorems in Analysis II, Springer, 1972.
[61] Lillian B. Pierce. On superorthogonality. J. Geom. Anal. 31 (2021), no. 7, 7096–7183.

[62] Manya Raman-Sundström. A pedagogical history of compactness, Amer. Math. Monthly 122, no. 17 (2015), 619–635.
[63] A. C. M. van Rooij and W. H. Schikhof. A second course on real functions. Cambridge University Press, Cambridge-New York, 1982.
[64] Maxwell Rosenlicht. Integration in finite terms. Amer. Math. Monthly 79, no. 9 (1972), 963-972.
[65] Walter Rudin. Principles of mathematical analysis. Third edition. Internat. Ser. Pure Appl. Math. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, 1976.
[66] Walter Rudin. Real and complex analysis. Third edition. McGraw-Hill Book Co., New York, 1987.
[67] Carl Runge. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. Zeitschrift für Mathematik und Physik, 46:224–243, 1901.
[68] Arnold Schönhage. Fehlerfortpflanzung bei Interpolation. Numerische Mathematik 3:62–71, 1961.
[69] Aleksei Shadrin, Twelve proofs of the Markov inequality.(English summary)Approximation theory: a volume dedicated to Borislav Bojanov, 233–298. Prof. Marin Drinov Academic Publishing House, Sofia, 2004.
[70] Alan D. Sokal. A really simple elementary proof of the uniform boundedness theorem. Amer. Math. Monthly 118(2011), no.5, 450–452.
[71] Michael Spivak. Calculus, Fourth edition. Publish or Perish Inc. 2003.
[72] Michael Spivak. Calculus on manifolds. A modern approach to classical theorems of advanced calculus. W. A. Benjamin, Inc., New York-Amsterdam, 1965.
[73] Elias M. Stein, Some results in harmonic analysis in $\mathbb{R}^n$, for $n \to \infty$. Bull. Amer. Math. Soc. (N.S.)9 (1983), no.1, 71–73.
[74] Elias M. Stein, Rami Shakarchi. Fourier Analysis. An introduction. Princeton Lect. Anal. 1, Princeton University Press, 2003.
[75] Elias M. Stein, Rami Shakarchi. Complex Analysis. Princeton Lect. Anal. 2, Princeton University Press, 2003.
[76] Elias M. Stein, Rami Shakarchi. Real Analysis. Measure theory, integration and Hilbert spaces Princeton Lect. Anal. 3, Princeton University Press, 2005.
[77] Elias M. Stein, Rami Shakarchi. Functional analysis. Introduction to further topics in analysis. Princeton Lect. Anal. 4, Princeton University Press, 2011.
[78] Louis Sucheston, Banach limits. Amer. Math. Monthly 74 (1967), 308–311.
[79] Terence Tao, Analysis, I. Third edition Texts Read. Math., 37 Hindustan Book Agency, New Delhi, 2015.
[80] Gérald Tenenbaum. Introduction to analytic and probabilistic number theory. Cambridge Stud. Adv. Math., 46 Cambridge University Press, Cambridge, 1995.
[81] Christoph Thiele. On the first chapter. Bonn 2017. arXiv:1706.08905. Updated as: A formal system of mathematics based on definitions (2020).
[82] L.N. Trefethen, J.A.C. Weideman. Two results on polynomial interpolation in equally spaced points. J. Approx. Theory 65 (1991), no. 3, 247–260.
[83] Loring W. Tu. An introduction to manifolds. Universitext, Springer, New York, 2011.
[84] A. H. Turetskii. The bounding of polynomials prescribed at equally distributed points. Proceedings of the Vitebsk Pedagogical Institute 3:117–127, 1940.
[85] Antoni Zygmund, Trigonometric series. Vol. I, II. Cambridge Math. Lib. Cambridge University Press, Cambridge, 2002.